ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Regras de associações entre as características dos acidentes de trânsito em rodovias federais brasileiras por meio de aprendizado de máquina

Ramon Batista de Araújo
Universidade Federal de Minas Gerais, Departamento de Engenharia de Transportes e Geotecnia. Belo Horizonte, Brasil.

Marcelo Franco Porto
Universidade Federal de Minas Gerais, Departamento de Engenharia de Transportes e Geotecnia. Belo Horizonte, Brasil.

Renata Maria Abrantes Baracho Porto
Universidade Federal de Minas Gerais, Departamento de Tecnologia do Design, da Arquitetura e do Urbanismo. Belo Horizonte, Brasil.

Recibido: 24.03.2022. Aceptado: 06.08.2024.

Resumo

Acidentes de trânsito são considerados um sério problema de saúde pública que, somado ao expressivo número de mortos e feridos, evidencia a necessidade de uma análise mais profunda das causas de acidentes. O objetivo dessa pesquisa consiste em identificar regras de associações entre as causas de acidentes e as características viário-ambientais e veiculares em rodovias federais brasileiras, comparando as técnicas de aprendizado de máquina *Apriori*, *Eclat*, *FP-Growth* e *FP-Max*. A metodologia propõe uma tabulação de dados de variáveis categóricas, utilizando-se de um método misto para coleta e transformação dos dados, por meio de um procedimento dentro de um contexto real em um estudo de caso. Através dos resultados foi possível realizar a comparação entre algoritmos e concluir que *Apriori*, *FP-Growth* e *Eclat* apresentam o mesmo desempenho, com índices de suporte e quantidade de características similares e o *FP-Max* apresentou resultado mais preciso. São apresentadas regras de associações entre o sexo do condutor, dia, horário, características da via e do veículo, e as causas de acidentes. É proposto um método para compreender as causas dos acidentes, contribuindo com pesquisadores e gestores na tomada de decisões e aprimoramento de políticas públicas na segurança viária.

PALAVRAS-CHAVES: ACIDENTES. SEGURANÇA VIÁRIA. REGRAS DE ASSOCIAÇÃO. APRENDIZADO DE MÁQUINA. ALGORITMOS.

Rules of Associations Between the Characteristics of Traffic Accidents on Brazilian Federal Highways Using Machine Learning

Abstract

Traffic accidents are considered a serious public health problem and the significant number of deaths highlights the need for a deeper analysis of the causes of accidents. The objective of this research was to identify rules of association between the causes of accidents and the

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

characteristics of road-environmental and vehicular on Brazilian federal highways. The machine learning techniques Apriori, Eclat, FP-Growth and FP-Max were compared. The methodology proposes a data table of categorical variables, in a mixed method for data collection and transformation. A case study was carried out within a real context. The comparison between algorithms concludes that Apriori, FP-Growth and Eclat present the same performance, with similar support indexes and a number of characteristics. The FP-Max in reverse, provides a more accurate result. The study presents associations between the driver's gender, day, time, road and vehicle characteristics, and the causes of accidents. It proposes a method to understand the causes of accidents, aiding researchers and managers in decision-making and improving public policies for road safety.

KEYWORDS: ACCIDENTS. ROAD SAFETY. ASSOCIATION RULES. MACHINE LEARNING. ALGORITHMS.

Introdução

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 por volta de 65 mil acidentes nas rodovias federais brasileiras em 2021. Destes, aproximadamente 72 mil pessoas ficaram feridas e 5.396 mortes foram registradas, isto é, cerca de 15 mortes por dia. Tais acidentes acarretam perdas incalculáveis em vários aspectos assim como custos para a economia brasileira (BRASIL, 2021). Em um estudo realizado sobre acidentes em rodovias federais brasileiras no período de 2014, foi constatado que esses eventos geram um custo para a sociedade de 12,8 bilhões de reais, sendo 62% associados a danos pessoais e 38% a danos materiais e perdas de carga (IPEA, 2020).

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. 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 gerando avanços na mineração de dados e aprendizado de máquina nos últimos anos.

De acordo com Cunto (2008), o conhecimento extraído a partir de diferentes modelos de análise estatística dos dados de acidentes auxilia engenheiros de segurança a tomarem decisões. Técnicas como *machine learning* (aprendizado de máquina) podem ser utilizadas para encontrar padrões ocultos e criar regras de associações entre os atributos de banco de dados de acidentes de trânsito. Atnafu & Kaur (2017) afirmam que é importante realizar uma análise cautelosa dos dados de acidentes para identificar a natureza do mesmo, assim, torna-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 contexto, essa pesquisa visa responder às seguintes questões: (i) Existem regras de associações entre as causas dos acidentes e as características viário-ambientais e veiculares em dados de acidentes das rodovias federais brasileiras? (ii) Qual o algoritmo que melhor identifica essas associações?

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Fundamentação teórica conceitual

Esta pesquisa teve como objetivo principal identificar regras de associação entre as causas de acidentes e as características dos veículos, utilizando dados de acidentes obtidos no site da PRF (Polícia Rodoviária Federal) do Brasil. Os dados abrangem todas as causas e tipos de acidentes ocorridos entre janeiro de 2017 e fevereiro de 2020. Além disso, foram consideradas as características dos veículos constantes no Registro Nacional de Veículos Automotores (Renavam), disponibilizados pelo Ministério da Infraestrutura do Brasil.

Para fundamentar a pesquisa, realizou-se uma revisão bibliográfica do tema, abrangendo a importância de diversos assuntos, como os acidentes de trânsito no Brasil e no mundo, e os fatores relacionados aos acidentes, tais como rodovias, usuários, meio ambiente e veículos. A revisão inclui estudos sobre as características dos veículos brasileiros, aprendizado de máquina e algoritmos de regras de associações.

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) também relatam que os fatores que contribuem com a causa dos acidentes de trânsito, em maior ou menor grau, são o condutor, 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.

O elevado número de acidentes é considerado um desafio para 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. 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.

Os veículos e os acidentes

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.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

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. Além disso, 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, quantitativa e qualitativa, os quais dificultam as análises por métodos estatísticos (Kumar & Toshniwal, 2015). Esses dados possibilitam o uso de novas tecnologias, as quais visam contribuir com pesquisadores e gestores a extrair conhecimento, auxiliando-os em tomadas de decisões.

Portanto, 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 especificidades dos veículos como idade, marca e potência do motor. Assim, para identificar esses padrões, utilizou-se das técnicas de *machine learning*, as quais proporcionam a melhor compreensão de dados e, ainda, a verificação de independência entre as características e as causas dos acidentes, para tanto, empregando a análise de correspondência multivariada, assim como Kaur (2015), o qual utilizou de técnicas como ANOVA e MANOVA para comparar os algoritmos *Apriori* e *FP-Growth* em regras de associação para detecções de doenças hepáticas.

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 e, assim, técnicas como *machine learning* são ferramentas utilizadas para analisar tais dados.

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.

Para Zhang (2020), o aprendizado de máquina é um subcampo da inteligência artificial que possibilita a construção de um modelo matemático com base em dados de uma 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); Deekshitha *et al.* (2019); Figueira *et al.* (2017); Amorim (2019); Kumar *et al.* (2017); Li *et al.* (2017); Meng *et al.* (2019); Nandurge & Dharwadkar (2017); Silva *et al.* (2019); Soares *et al.* (2018). Os algoritmos e tipos de aprendizagem de máquina são divididos pela literatura em aprendizagem supervisionada, semi-supervisionada ou aprendizagem por reforço e aprendizagem não supervisionada.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Regras de associação

De acordo com Borgelt & Kruse (2002), regras de associações 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.

Desta forma, 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, 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.

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, 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. Sendo assim, para selecionar regras do conjunto de todas as regras possíveis, um suporte mínimo é fixado.

Seguindo os preceitos da literatura, os valores de suporte mínimo utilizados estão descritos na Tabela 1 de acordo com os trabalhos relacionados a esse estudo.

Tabela 1. Valores de suporte mínimo utilizado por autores de trabalhos relacionados

Referência	Mínimo valor de suporte utilizado
Ali & Hamed (2018)	0,4
Atnafu & Kaur (2017)	-
Daher *et al.* (2016)	0,4
Deekshitha *et al.* (2019)	-
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
Soares *et al.* (2018)	0,1
Xi *et al.* (2016)	0,4

Fonte: elaborado pelo autor.

Trabalhos relacionados

Estudos utilizando técnicas de *machine learning* têm sido frequentemente realizados na área de transportes, 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. Essas técnicas são utilizadas por pesquisadores para prever, classificar e encontrar associações e padrões ocultos em dados de acidentes rodoviários no Brasil e no mundo.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

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 em 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, também 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, também, um algoritmo de regras de associações *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.

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.

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.

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,

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

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) também 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 associações. Amorim (2019) afirma que o assunto é muito estudado nas demais partes do mundo e que poderia, ainda, ser mais bem 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, também, a potência do motor.

O *Apriori* é o algoritmo mais utilizado para encontrar regras de associações em acidentes de trânsito, onde se mostrou eficaz, conforme estudos apresentados. Entretanto, 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

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Apriori, *Tree-projection*, *FP-Growth* e *Eclat*, demonstrando suas vantagens e desvantagens.

Kaur (2015) identificou regras de associações 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 analisada 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, assim, a importância do seu estudo para o desenvolvimento da discussão científica a respeito do tema proposto.

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 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. Principais estudos relacionados sobre *machine learning* e acidentes rodoviários

Referência	País	Período	Algoritmos										
			Árvore de	*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		*			*						

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Referência	País	Período	Algoritmos										
			Árvore de	*Apriori*	CART	*Clustering*	*Eclat*	*FP-Growth*	*Naive Bayes*	Redes Neurais	Regressão	SVM	FP-Max
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, observa-se que o *Apriori*, é o algoritmo mais influente na área de análise de padrões em acidentes de trânsito. Entretanto, existem outros algoritmos pouco aplicados na área, mas utilizados em outras áreas de estudo e, principalmente, no Brasil, os quais também devem ser estudados. Portanto, 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.

Método

Essa pesquisa utilizou um método misto para coleta, transformação dos dados e análise dos resultados, tendo sido realizado um estudo de caso em um contexto real. O

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

desenvolvimento inclui as seguintes etapas: procedimentos de obtenção dos dados; análise, interpretação dos resultados e descobertas, com base em Jung (2004), para atender ao objetivo principal de identificar regras de associações 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.

Sabe-se que os procedimentos concomitantes apresentam uma análise abrangente do problema, uma vez que convergem dados qualitativos e quantitativos, onde dispõe de um procedimento de dados maior para analisar diferentes questões. Nessa estratégia, coletam-se dados qualitativos e quantitativos, de forma simultânea, e integra-se as informações na interpretação dos resultados (CRESWELL et al., 2007). A Figura 1 apresenta a estratégia transformada concomitante de métodos mistos, adaptada para essa pesquisa.

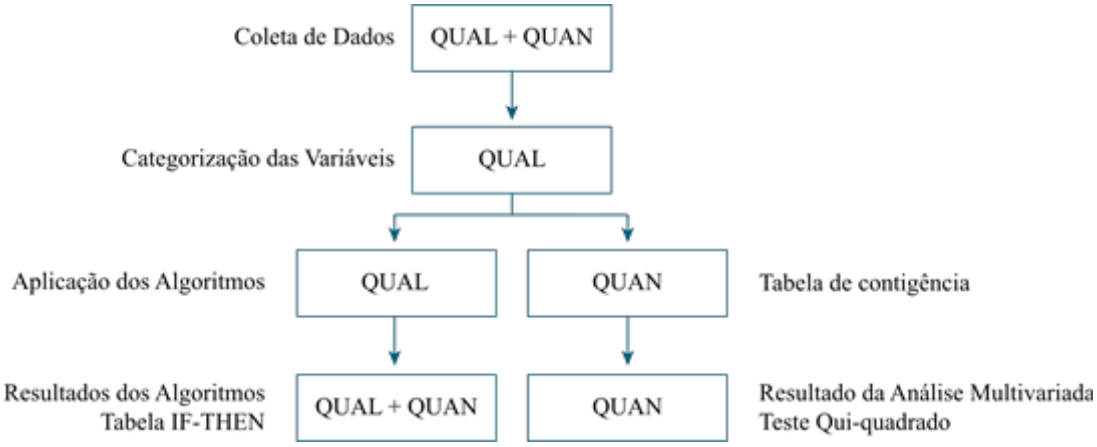

Figura 1. Método misto de análise. *Fonte: elaborado pelo autor*.

Os dados contendo os acidentes rodoviários e as características de veículos são dados qualitativos e quantitativos que foram coletados simultaneamente. Em sequência, os dados quantitativos foram categorizados, transformando-se em qualitativos visando reduzir a amplitude dos dados quantitativos e, assim, proporcionando um melhor resultado do método de regras de associações, evitando *overfitting*, ou seja, o ajuste de um modelo muito complexo aos dados.

Antes da aplicação dos algoritmos, visando compreender como é a relação entre as variáveis, criou-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.

Para aplicação dos algoritmos de regras de associação são necessários dados qualitativos, 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 como quais regras são mais pertinentes, isto

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Figura 2. *Framework* da pesquisa. *Fonte: elaborado pelo autor.*

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

é, com mais de um item e melhor *support*. Além disso, foram comparados resultados quantitativos, como qual o algoritmo com maior número de regras *IF-THEN* e qual obtiveram melhores resultados de *support, confidence* e *lift*.

Framework da pesquisa

Um *framework* é uma maneira de representar o conhecimento e informações úteis para a compreensão de um estudo (Minsky, 1974). As etapas do método desse estudo estas estão apresentadas no *framework* da Figura 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 associações entre as causas dos acidentes e as características viário-ambientais e veiculares em dados de acidentes das rodovias federais brasileiras? (ii) Qual o algoritmo que melhor identifica essas associações? Assim, por meio destes questionamentos, inicia-se a primeira etapa da metodologia, a obtenção dos dados.

Base de dados

Sabe-se que uma fonte dos dados insegura pode introduzir incerteza e, assim, impactar a veracidade de um conjunto de dados. Logo, dados duvidosos e incorretos podem afetar o desempenho do aprendizado de máquina provendo regras de associações equivocadas (L'heureux *et al*., 2017). Sendo assim, foram necessários dados confiáveis de acidentes de trânsito e de características dos veículos.

Para essa metodologia, no banco de dados de acidentes deve constar atributos que contenham: o momento que o acidente ocorreu, (dia, mês, ano e o horário); atributos das características das vias, 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, para a criação de regras de associações, são necessários dados como a causa do acidente. Os dados utilizados nessa pesquisa foram obtidos de duas fontes distintas e, com a obtenção destes, foi possível realizar uma análise precisa para a criação dos modelos de *machine learning*.

Análise exploratória

Dentre as 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.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

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.

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

Na etapa de transformação para variável qualitativa criou-se faixas de grupos dos dados quantitativos, transformando as variáveis quantitativas e qualitativas em apenas variáveis qualitativas, onde a integração desses dados ocorreu durante a fase de análise.

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

Devido às variáveis serem qualitativas categóricas, realizou-se análise de correspondência multivariada, onde criou-se uma tabela de contingência com a contagem das observações por variável, realizando-se, em seguida, a análise estatística multivariada pelo teste qui-quadrado na tabela de contingência. Assim, 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 associações. Com os dados transformados, aplica-se os algoritmos *Apriori*, *Eclat*, *FP-Growth* e *FP-Max*.

Aplicação dos algoritmos

Utilizando a linguagem Python e a biblioteca Mlxtend e pyECLAT construiu-se cada modelo de aprendizado de máquina dos algoritmos *Apriori*, *Eclat*, *FP-Growth* e *FP-Max*. Com a aplicação dos algoritmos, obteve-se uma tabela *IF-THEN*, isto é, regras de associações concludentes de uma condição, em outros termos, SE condição ENTÃO conclusão. Tais regras são criadas com variáveis qualitativas, que seriam as regras de associações, e variáveis quantitativas, que são os índices de *support, confidence* e *lift*. Então, cria-se um relatório com representações gráficas dos resultados e comparam-se os algoritmos que obtiveram maior quantidade de regras com melhores índices.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Como o algoritmo *Eclat* tem como resultado somente o índice de suporte, optou-se por utilizar esse índice para comparação entre os algoritmos. Sendo assim, criou-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, executou-se uma análise de variância multivariada (MANOVA) na média de índices de suporte e tamanho de itens e, desta forma, foi possível testar estatisticamente a igualdade entre médias, assim como realizaram Kaur (2015) e Hunyadi (2011).

Após, verificou-se quais regras obtiveram maior quantidade de características (*IF*) com melhores índices de suporte, categorizando as regras de associações pertinentes para tomadas de decisões e políticas públicas.

Desenvolvimento e resultados

Aplicou-se a metodologia descrita anteriormente nos dados de acidentes das rodovias federais brasileiras e das características dos veículos, explicando todas as etapas e decisões tomadas durante o tratamento e a análise dos dados. Por fim, discutiram-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.

Obtenção dos dados com variáveis qualitativas e quantitativas

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, esta pesquisa optou por coletar dados abertos dos acidentes de trânsito e concatená-los com dados das características dos veículos constantes no Registro Nacional de Veículos Automotores (Renavam), concedidos pelo Ministério da Infraestrutura.

Os dados de acidentes foram obtidos no site da PRF 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 incluírem a marca do veículo, diferentemente dos bancos de dados dos anos anteriores.

É importante ressaltar que 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, os quais foram utilizados como base nesta pesquisa e podem ser consultados em BRASIL (2017), conforme Tabela 3.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Tabela 3. 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.

Variável	Descrição
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.
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).

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

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.

Tabela 4. 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.

Análise exploratória das variáveis

As bibliotecas utilizadas para desenvolvimento do estudo e análise exploratória foram 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*.

A primeira etapa é o pré-processamento. Foram importados, separadamente, os bancos de dados dos acidentes dos anos de 2017, 2018, 2019 e 2020 e, em seguida, os dados das características dos veículos.

Com o banco de dados de acidentes importados Antes de concatena-los, foi verificado as informações de cada conjunto de dados e realizado a conferência de valores únicos. O banco de dados com os acidentes de 2017 apresentaram 349.067 observações e 37

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

atributos. Já o do ano de 2018, apresentaram 324.809 e 37 atributos. No conjunto de dados do ano de 2019, apresentou-se 331.666 observações e 37 atributos. Já o último conjunto de acidentes, do ano 2020, que não contempla todo período, conforme explicado anteriormente, o banco de dados conta com 93.731. Como todos os bancos de dados de acidentes possuem as mesmas colunas, porém somente diferentes números de observações podem ser concatenados sem nenhuma obstrução. Após concatená-los, obteve-se 1.099.273 observações e 37 atributos.

Com o conjunto de dados de acidentes importado e concatenado, iniciamos o pré-processamento. Foi determinada a remoção dos registros nulos e duplicados, pois há um grande volume de registros, o que nos permite removê-los sem prejudicar o resultado dos algoritmos. Após a remoção dos valores ausentes, independe da variável, o conjunto de dados de acidentes foi reduzido de 1.099.273 para 830.291 registros. Em seguida, eliminamos os valores duplicados da variável id (Tabela 3), reduzindo o conjunto de dados de 830.291 para 830.290 registros, o que indica que restou apenas um valor duplicado após a remoção dos valores ausentes. Com isso, é necessário entender melhor cada variável para realizar uma exploração mais detalhada dos atributos.

As variáveis de identificação do acidente e da pessoa envolvida não foram utilizadas nesse estudo e, portanto, 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 associações. A próxima variável analisada foi o dia da semana, onde percebeu-se que os acidentes ocorrem em dias próximos ao final de semana, sendo 49,47% dos acidentes ocorrendo entre sexta a domingo, conforme Figura 3.

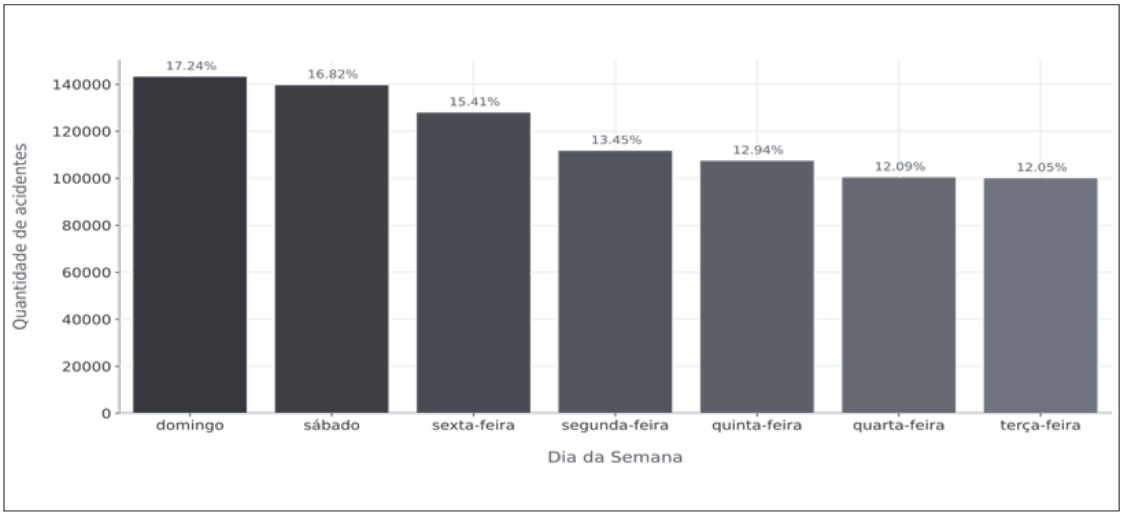

Figura 3. Quantidades de acidentes por dia da semana. *Fonte: elaborado pelo autor.*

Outro atributo analisado foi o horário, o qual apresenta 1.430 valores únicos, onde a maioria dos acidentes ocorreu no período de 17h às 19h. Esse período de apenas 2 horas do dia corresponde a, aproximadamente, 14,68% dos acidentes, os quais aconteceram, principalmente, às 18h30, horários denominados como horários de pico da tarde, então, mais um fator com potencial significativo para o algoritmo de regras de associação.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Essa análise apresenta dados interessantes, onde, por exemplo, o estado em que mais ocorreram acidentes no período de janeiro de 2017 a fevereiro de 2020 em rodovias federais foi Minas Gerais, seguido por Paraná e Santa Catarina. 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. 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 (75,71%). 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, também 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. Assim, embora represente apenas 33,80% da malha rodoviária, a região registra 63,12% dos acidentes. Contudo, é importante ressaltar que mais de 75% da frota de automóveis está concentrada no Sul e Sudeste do país. Portanto, uma análise mais abrangente, incluindo dados de contagem de tráfego em rodovias e outros fatores socioeconômicos, é necessária para uma compreensão completa desses números. No entanto, essa análise está além do escopo do presente estudo.

O banco de dados da PRF determina se uma causa de acidente foi principal ou não, com dois valores únicos. Portanto, para uma análise mais criteriosa a respeito da causa principal do acidente, optou-se por considerar somente as causas principais, removendo registos com mais de uma causa, assim, o banco de dados de acidentes reduziu de 830.290 para 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 5.

Tabela 5. 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

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Nº	Causa principal do acidente	Quantidade de acidentes
14	Mal súbito	5.839
15	Restrição de visibilidade	5.437
Nº	**Causa principal do acidente**	**Quantidade de acidentes**
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.

Percebeu-se uma 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, bem como a desobediência às normas de trânsito pelo condutor.

Referente ao atributo ordem do tipo de acidente, este apresenta 11 valores únicos, no entanto, não tem muita significância nessa pesquisa, assim como o tipo de acidente. Tal atributo não é utilizado na modelagem, pois se tem interesse nas características anteriores ao ocorrido do acidente, e o tipo de acidente é resultado deste, ou seja, atributos pós-acidente. Destes, 479.564 foram com vítimas feridas, 117.568 sem vítimas e 66.052 com vítimas fatais. Mesmo não se utilizando a 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, percebe-se que a maioria dos acidentes ocorrem durante o dia, seguido pela noite e com poucos registros ao anoitecer e amanhecer, conforme Figura 4. Além disso, para criação do modelo, ou utiliza-se o atributo horário ou a fase do dia, assim, evitando-se redundância, desta forma, como a fase do dia possui menor número de registros únicos, optou-se por utilizá-la. Nessa característica, observou-se que mais da metade dos acidentes, isso é, 57,67% dos acidentes ocorrem durante pleno dia, fator devido ao número de veículos que transitam durante o dia é superior aos demais horários.

Outro atributo importante, referente ao meio em que o acidente ocorreu, são as condições meteorológicas, as quais contêm 10 valores únicos. Entretanto, há observações registradas como ignorado, as quais foram removidas e, assim, as condições em que ocorreram mais acidentes foram a céu aberto com 54,41%, nublado com 38,37% e chuvoso com 13,69%.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

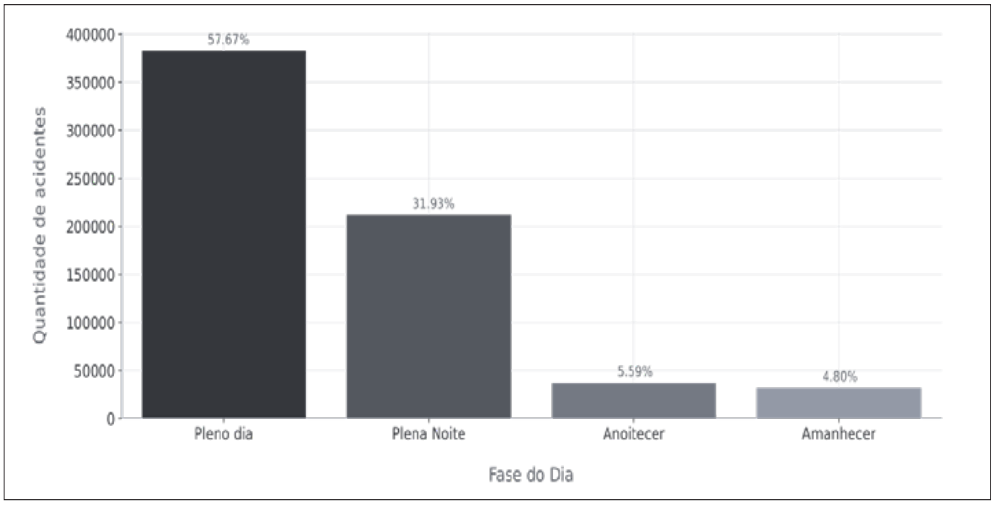

Figura 4. Acidentes por fase do dia. *Fonte: elaborado pelo autor.*

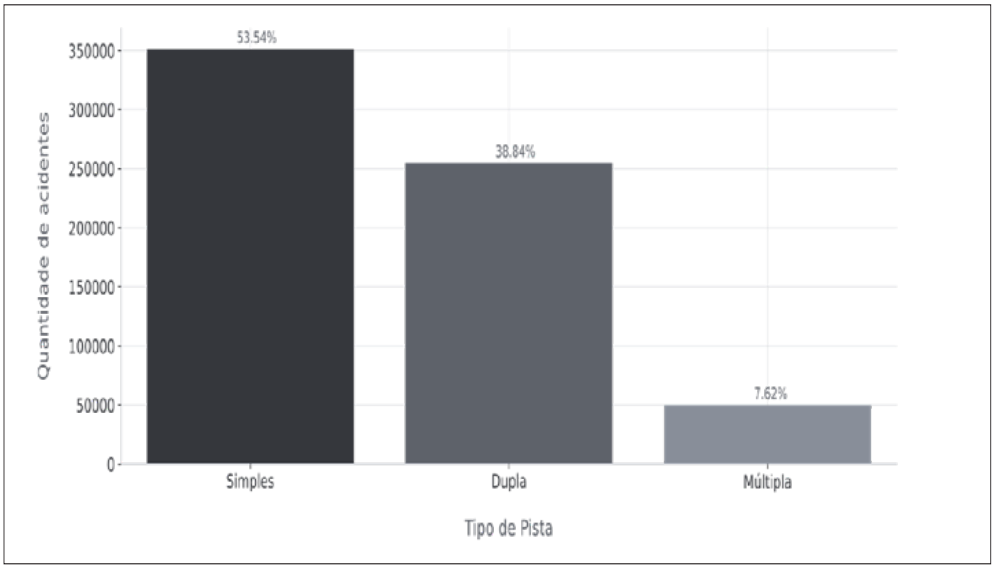

Figura 5. Quantidade de acidentes de trânsito por tipo de pista. *Fonte: elaborado pelo autor.*

O atributo sentido da via possui dois valores, crescente e decrescente, e é relevante quando utilizado juntamente com br e km, no entanto, uma vez que não serão utilizados estes atributos, também não se 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, sendo importante sua utilização nos modelos dessa pesquisa, visto que possui características da via. Entretanto, o atributo traçado da via possui registros denominado como não informado e, por esta razão, optou-se por removê-los. A Figura 5 exibe o número de acidentes por tipo de pista, onde se percebe que o número de acidentes em pista simples é superior ao somatório do número de acidentes em pista dupla e múltipla, isto é, 53,54% dos acidentes ocorrem em pista simples.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Até o momento, após todas exclusões, o atributo uso do solo apresentou 338.816 acidentes em áreas rurais e 239.157 em áreas urbanas. Embora os acidentes em perímetro urbano possam apresentar características distintas dos acidentes em áreas rurais, essa avaliação não se enquadra no escopo inicial deste estudo. Da mesma forma, a variável de identificação do veículo também foi removida por não ser significativa para este trabalho.

Para este estudo, é fundamental considerar a influência do tipo de veículo envolvido nos acidentes. O banco de dados contém registros de 21 tipos diferentes de veículos, mas para esta análise, foram selecionados apenas os veículos do tipo automóvel. Assim, foi necessário criar um novo banco de dados, filtrando os registros para incluir somente automóveis, o que resultou em 265.928 observações de um total de 663.184 processados até o momento. O atributo ano de fabricação de veículo foi utilizado para definir a idade do automóvel, pois se entende que, como os registros das ocorrências são em anos diferentes, não é relevante quando o veículo foi fabricado, mas sim a idade que o veículo tinha naquele no 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* da Figura 6 (a).

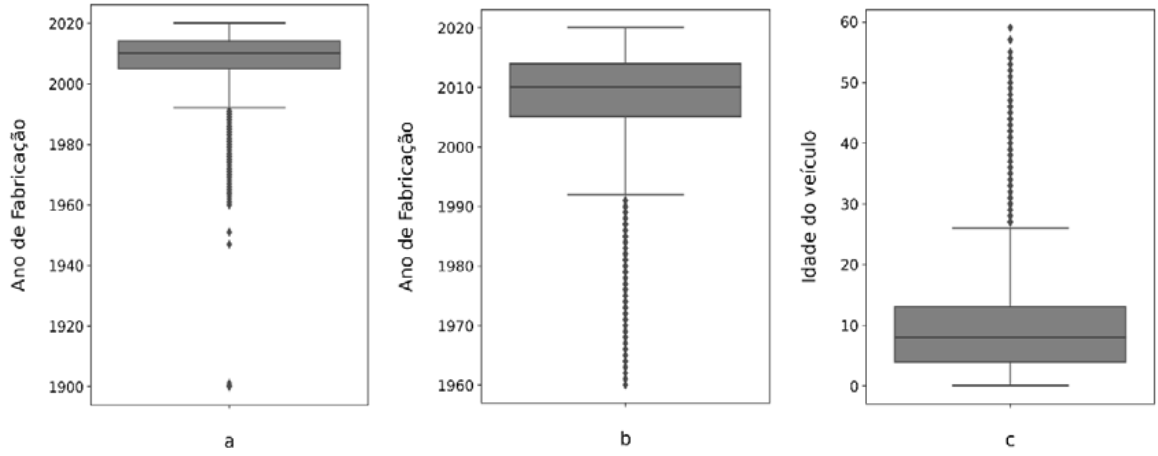

Figura 6. (a) *Boxplot* do ano de fabricação dos automóveis. (b) *Boxplot* do ano de fabricação dos automóveis de 1956 a 2020. *(c) Boxplot* da idade do veículo. *Fonte: elaborado pelo autor.*

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*, conforme Figura 6 (b). Já, para encontrar a idade do veículo, criou-se uma nova variável com a subtração do ano em que o acidente ocorreu pelo ano de fabricação do veículo, obtendo o *boxplot* inverso ao do ano de fabricação após 1956, representado pela Figura 6 (c).

No atributo tipo de envolvido, que apresenta 4 valores únicos (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. Além disso, este estudo procura saber a influência das características do condutor na gravidade do acidente e, diante disso, o novo banco de dados, que inicialmente continha 265.928 registros

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

de automóveis, foi reduzido para 171.493 registros após a filtragem dos dados dos condutores.

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 na Figura 7 (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 (170.692) 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* da Figura 7 (b).

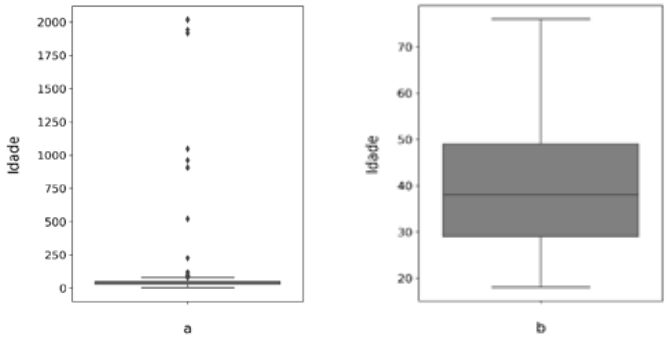

Figura 7. (a) Boxplot das idades dos condutores. (b) Boxplot das idades dos condutores entre 18 e 76 anos. *Fonte: elaborado pelo autor.*

A variável sexo do envolvido possui 3 valores únicos, sendo 139.932 casos registrados como masculino, 30.749 como feminino e 11 classificados como ignorado. Conforme 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, assim como as demais variáveis como latitude, longitude, regional, delegacia e unidade operacional, pois não se pretende utilizar a localização exata e nem de onde foi realizado o boletim de ocorrência, resultando em um banco de dados com 169.680 observações.

Com o banco de dados de acidentes tratado e preparado para receber dados das características dos veículos, é necessário realizar o tratamento dos dados que contém a potência do motor.

Antes de concatenar os bancos de dados de características de veículos e acidentes, analisou-se 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. É importante ressaltar que existem veículos da mesma marca e

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

modelo com potências diferentes, logo, optou-se por agrupá-los pela média da potência. O resultado pode ser visto na Figura 8.

Conforme ilustrado nas Tabelas 3 e 4, ambos os conjuntos de dados compartilham as colunas marca e ano de fabricação. Portanto, essas colunas foram utilizadas para realizar a junção dos bancos de dados. No conjunto de dados de veículos, que inicialmente continha 482.312 registros, foi feito um processo de ordenação por potência e remoção de duplicados, resultando em 238.480 registros únicos por marca e ano. Em seguida, os dados dos veículos foram unidos com os dados de acidentes com base nas colunas marca e ano, gerando um banco de dados combinado com 171.490 registros, mantendo os valores em comum e dos dados de acidentes. Entretanto, como poderia haver veículos sem a referida potência no banco de dados das características dos veículos, optou-se por remover os valores ausentes, mantendo apenas os valores comuns entre os conjuntos, estabelecendo um novo banco de dados.

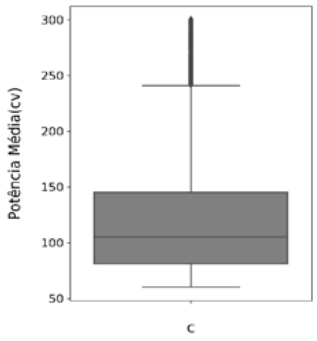

Figura 8. Boxplot das potências médias dos veículos. *Fonte: elaborado pelo autor.*

Após a junção, decidiu-se remover a informação do modelo do veículo, mantendo apenas a marca, para simplificar e otimizar a variável para a aplicação do modelo. Optou-se por reter somente as 20 marcas mais envolvidas em acidentes. Com essas modificações, o banco de dados tratado e explorado finalizou com 126.545 observações e 13 variáveis.

Portanto, o processo de limpeza e filtragem dos dados começou com um banco de dados contendo 1.099.273 registros de acidentes. Inicialmente, foram removidos os registros com valores ausentes, reduzindo o total para 830.291. Em seguida, os registros duplicados foram eliminados, resultando em 830.290 entradas. Focando apenas nos acidentes envolvendo automóveis, o número de registros foi reduzido para 265.928. Após a filtragem para incluir apenas condutores (excluindo passageiros, pedestres e cavaleiros), o total caiu para 171.493. Considerando apenas os condutores com idades entre 18 e 76 anos, o conjunto foi reduzido para 170.692. Finalmente, registros com informações incompletas sobre o sexo do condutor foram descartados, resultando em um banco de dados de acidentes com 169.680 observações. Após a junção com o conjunto de dados de veículos, obteve-se um conjunto de dados, com 126.545 observações, que foi utilizado para a aplicação dos algoritmos, conforme detalhado na Tabela 6.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Tabela 6. Tipos de variáveis

Nº	Variável	Tipo de variável
1	Data inversa	quantitativa
2	dia da semana	qualitativa
3	horário	quantitativa
4	Causa do acidente	qualitativa
5	Fase do dia	qualitativa
6	Condição meteorológica	qualitativa
7	Tipo de pista	qualitativa
8	Traçado da via	qualitativa
9	marca	qualitativa
10	idade	quantitativa
11	sexo	qualitativa
12	Idade do veículo	quantitativa
13	potência	quantitativa

Fonte: elaborado pelo autor.

Após compreender e tratar cada variável dos bancos de dados de acidentes e das características dos veículos observa-se a existência de variáveis qualitativas e quantitativas. Assim, a próxima etapa da pesquisa foi 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). Assim, criaram-se faixas de grupos das variáveis quantitativas, ou seja, categorias para aplicação dos algoritmos *de machine learning*.

Transformação das variáveis em qualitativa

São as 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 de 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. Entretanto, 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. Sendo assim, 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 e, entre 17 e 19 horas, segundo Resende & Souza (2009). Desta forma, obteve-se 36.456 horários de pico e 900.089 horários normais.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

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 no histograma da Figura 9.

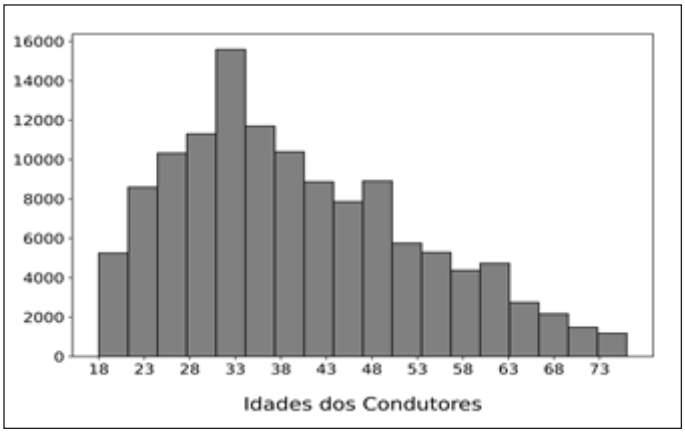

Figura 9. Histograma da Idade dos condutores. *Fonte: elaborado pelo autor.*

Categorizando-se as variáveis, idade do veículo, tem-se a estatística do histograma e *boxplot* da Figura 10.

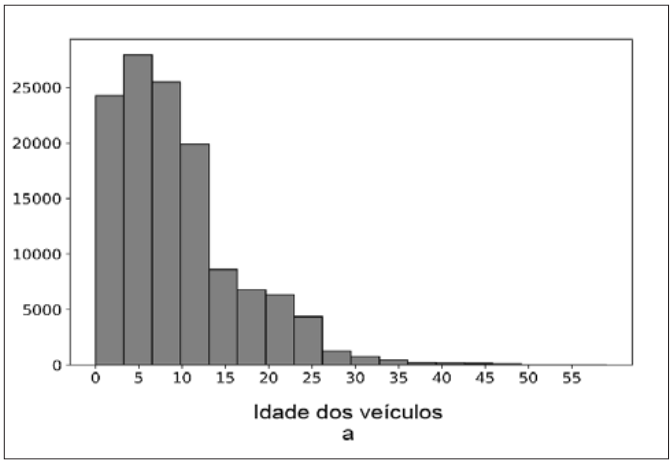

Figura 10. Histograma da idade dos veículos. *Fonte: elaborado pelo autor.*

Para criar categorias utilizou-se a 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. Em seguida, considerou-se a soma da média em cada faixa até os 30 anos. Para critério de criação das categorias de faixas para a potência do motor foi utilizado o desvio padrão e, com isso, as variáveis quantitativas foram categorizadas e transformadas em qualitativas. Assim, as variáveis, após tratamento, apresentam as propriedades que constam na Tabela 7 e estão preparadas para criação dos modelos de regras de associação.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Tabela 7. *Variáveis selecionadas e categorizadas para modelagem*

Nº	Variável	Fator	Qtd. Valores Únicos	Tipo de variável
1	Dia da Semana	Meio Ambiente	7	qualitativa
2	Feriado	Meio Ambiente	2	qualitativa
3	Horário de Pico	Meio Ambiente	2	qualitativa
4	Fase do Dia	Meio Ambiente	4	qualitativa
5	Condição Meteorológica	Meio Ambiente	8	qualitativa
6	Tipo de Pista	Via	3	qualitativa
7	Traçado da Via	Via	9	qualitativa
8	Faixa Etária	Usuário	12	qualitativa
9	Sexo	Usuário	2	qualitativa
10	Marca	Veículo	20	qualitativa
11	Idade do Veículo	Veículo	7	qualitativa
12	Potência	Veículo	8	qualitativa
13	Causa do Acidente	Todos	24	qualitativa
Soma valores únicos			108	

Análise de correspondência multivariada

Nesta etapa criou-se uma tabela de contingência com as contagens das características por causa de acidente. Para criação dessa tabela, foi realizada a contagem de acidentes por característica, por tipo de causa de acidente, utilizando a função *Crosstab* da biblioteca Pandas da linguagem Python. Dessa forma, obteve-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.

Têm-se variáveis quantitativas e para identificar a independência das variáveis criadas, foi feita uma análise de correspondência multivariada através do teste qui-quadrado com as seguintes hipóteses: Hipótese nula (H_o) = Independência das variáveis, isto é, independente da causa as características aparecem com a mesma frequência; e Hipótese alternativa (H_1) = Dependência das variáveis.

Por meio da análise dos dados, notou-se que o p-valor é zero, logo, existem evidências para se rejeitar a hipótese nula (H_o), que é a hipótese de independência. Isso significa que, dependendo da causa do acidente, a frequência das características é diferente e, portanto, sabe-se que a frequência das características não é a mesma para todas as causas de acidentes.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Aplicação dos algoritmos e resultados das regras de associação

Sabendo que a quantidade de características dos acidentes é dependente da causa do acidente, aplicam-se os algoritmos de aprendizado de máquina na tabela de variáveis categóricas. Com essa tabela 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. Portanto, têm-se dados de variáveis categóricas, logo se transforma 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*.

Apriori

O algoritmo *Apriori* apresentou 1.010 itens com a estatística de suporte conforme Figura 11.

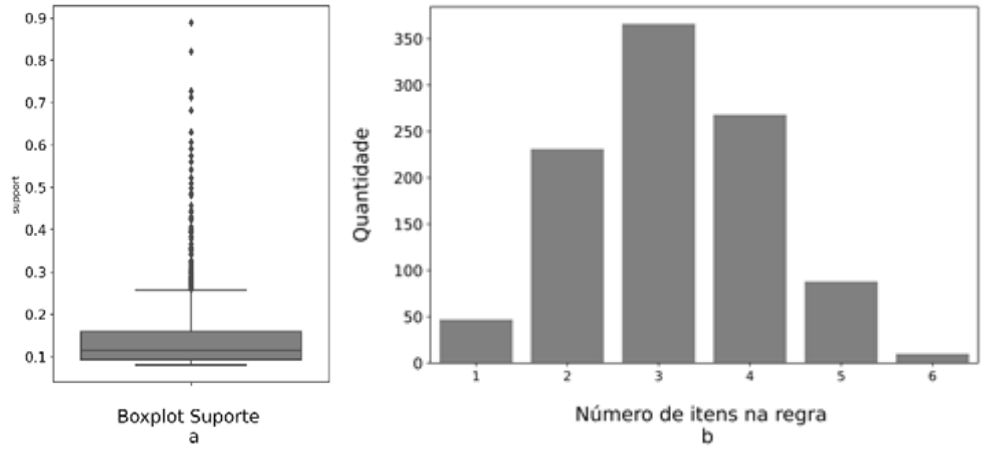

Figura 11. (a) Boxplot do suporte (b) Histograma da quantidade de características.
Fonte: elaborado pelo autor.

Nota-se 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 Figura 12.

Criando a tabela *IF-THEN*, nesse cenário, através da métrica de confiança, obtém-se 759 regras. Por meio dos resultados, 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 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. Um resumo das 5 regras ordenado pelos valores de lift estão demonstrados a seguir na Tabela 8.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

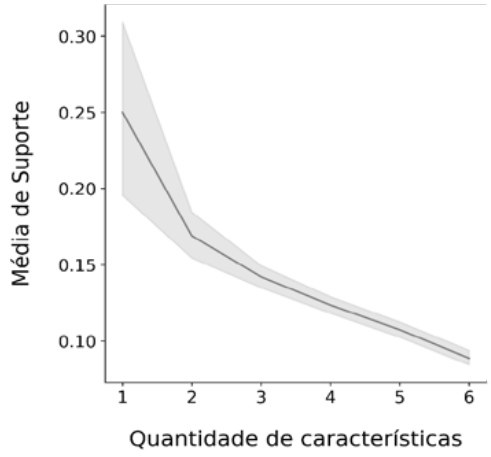

Figura 12. Média e desvio padrão dos suportes por quantidade de características. *Fonte: elaborado pelo autor.*

Tabela 8. 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**
{'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
{'Masculino', 'Plena Noite', 'Automóvel de 86 a 111 cv', 'Dia Normal'}	**{'Horário Normal'}**	**0,1137**	**0,7116**	**0,0926**	**0,8144**	**1,1443**

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.

No entanto, essa base de dados não apresentou nenhuma causa de acidente como consequência. Com o objetivo de balancear os dados, isto é, equilibrar a quantidade de características, o algoritmo foi aplicado em mais dois cenários: um que excluiu os atributos 'Dia Normal' e 'Horário Normal', mantendo apenas 'Feriado' e 'Horário de Pico'; e outro que omitiu essas características, além de não considerar o sexo 'Masculino'.

Assim, no 2º cenário, com mínimo de confiança de 0,08, percebeu-se novas características, tais como a ingestão de álcool envolvida ao sexo masculino e automóveis da marca VM. Ainda sobre veículos, outra característica presente nesse cenário foi o envolvimento

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

de acidentes em reta com 'Automóvel de 14 a 20 anos' de idade e condutores do sexo masculino. Outro fator importante também presente é a condição meteorológica 'Céu Claro'. (Tabela 9)

Tabela 9. *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
{'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**
{'Automóvel de 14 a 20 anos', 'Reta'}	**{'Masculino'}**	**0,0963**	**0,8200**	**0,0835**	**0,8673**	**1,0576**

Esse cenário apresentou 103 THEN, sendo todas do sexo 'Masculino'. Sendo assim, aplicou-se o algoritmo em um 3º cenário (Sem 'Dia Normal', 'Horário Normal' e 'Masculino'), como na Tabela 10.

Tabela 10. *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**

No terceiro cenário, identificaram-se apenas quatro regras relevantes. Destaca-se a forte associação entre a condição meteorológica 'Sol' e a fase do dia 'Pleno dia'. Outras três regras interessantes relacionam acidentes por falta de distância de segurança em pistas retas durante o dia. Ao excluir as características 'Dia Normal', 'Horário Normal' e 'Masculino', e dado a ausência das antônimas dessas características, elas são implicitamente consideradas. Uma regra relevante incluindo essas características excluídas é {'Dia Normal', 'Horário Normal' e 'Masculino'} + {'Pleno dia', 'Não guardar distância de segurança', 'Reta'}. Assim, os mesmos cenários foram aplicados com o algoritmo FP-Growth.

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 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*.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Sendo assim, entende-se que o algoritmo *FP-Growth* da biblioteca mlxtend é similar ao algoritmo *Apriori* da mesma biblioteca. Portanto, foram aplicados os algoritmos *FP-Max* e *Eclat* para fins de comparação.

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 na Figura 13.

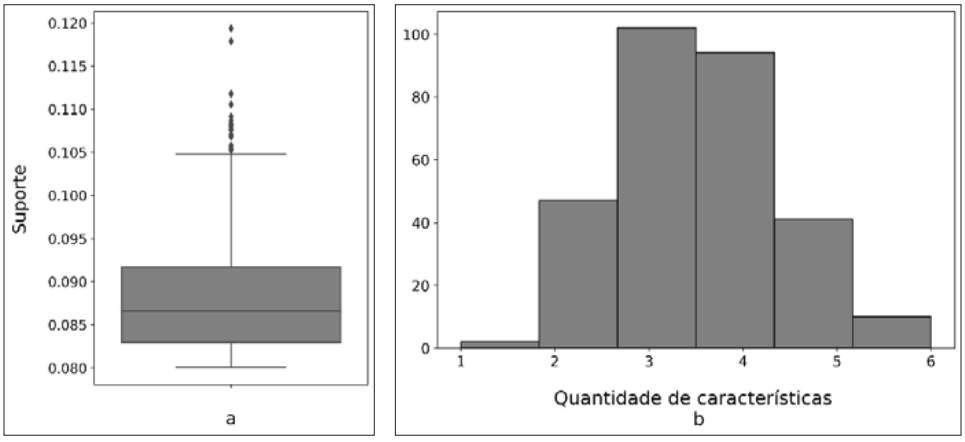

Figura 13. (a) Boxplot do suporte (b) Histograma da quantidade de características.
Fonte: elaborado pelo autor.

Os índices de suporte variam de 0,08 a 0,11, com uma média de 0,08 e um desvio padrão baixo. A média de características por item é de 3,5236. Notavelmente, ao contrário dos algoritmos Apriori e FP-Growth, a média de suporte aumenta com um maior número de características, como ilustrado no gráfico da Figura 14.

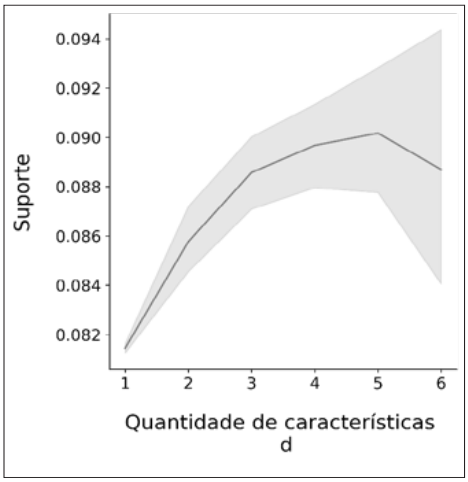

Figura 14. Média e desvio padrão dos suportes por quantidade de características.
Fonte: elaborado pelo autor.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

A tabela IF-THEN foi criada usando apenas o índice de suporte, pois não tinha valores de lift e confiança (ambos nulos). Ela contém um total de 394 regras, e na Tabela 11 é possível ver um resumo das 10 principais regras ordenadas pelo suporte.

Tabela 11. *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	-	-

O algoritmo FP-Max destacou regras com diversas características, diferentemente do Apriori e FP-Growth. Identificou associações significativas como 'Horário Normal', 'Masculino', 'Automóvel de 86 a 111 cv', 'Simples' e 'Dia Normal', validando sua relevância no banco de dados. Além disso, 'Dia Normal' e 'Horário Normal' estiveram presentes em todas as principais associações. Em seguida, o algoritmo foi aplicado em dois outros cenários para comparação, sendo o segundo cenário sem 'Dia Normal' e 'Horário Normal', com um mínimo de confiança de 0,08. A Tabela 12 apresenta os resultados do segundo cenário.

Tabela 12. *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	-	-
{'Céu Claro', 'Masculino'}	{'Reta', 'Automóvel de 86 a 111 cv'}	-	-	0,1445	-	-

No segundo cenário, que possui 298 regras, são evidenciadas associações relevantes, incluindo características como 'Automóvel de 86 a 111 cv', 'Céu Claro', 'Masculino' e

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

'Reta', similarmente ao que é observado no Apriori e FP-Growth. No entanto, este cenário não inclui a marca do automóvel e a idade, ao contrário do que é identificado pelo Apriori e FP-Growth.

Para o terceiro cenário (Sem 'Dia Normal', 'Horário Normal' e 'Masculino'), com um mínimo de suporte de 0,05 (Tabela 13).

Tabela 13. *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	-	-

Neste cenário, foram identificadas 28 regras, sendo a principal característica a "Falta de Atenção à Condução", conforme a Tabela X indica.

O algoritmo FP-Max, menos explorado na área, produziu uma regra de associação notável, demonstrando a relação entre os fatores {'Dia Normal' 'Horário Normal' 'Masculino'} + {'Reta', 'Céu Claro', 'Pleno dia'} ☐ 'Falta de Atenção à Condução'. Por fim, o último algoritmo aplicado foi o Eclat, por meio da biblioteca pyeclat.

Eclat

O algoritmo Eclat, em contraste com Apriori, FP-Growth e FP-Max, não faz parte da biblioteca mlxtend. Optou-se por utilizar a biblioteca pyEclat para aplicá-lo, estabelecendo um suporte mínimo de 0,08 e um máximo de 6 características. Como resultado, o Eclat identificou 644 itens, com uma média de suporte de 0,1595 e 2,4953 características. As estatísticas comparativas dos algoritmos podem ser visualizadas na Figura 15.

Neste algoritmo, foi estabelecido um limite máximo de combinações, sendo escolhido o valor 6 para manter consistência com os outros algoritmos que serão avaliados. Nota-se que, no caso deste algoritmo, os valores máximos e mínimos de suporte coincidiram precisamente com os do Apriori e FP-Growth. As distribuições entre quartis demonstraram diferenças, conforme evidenciado na Figura 16, que possui um perfil semelhante ao do Apriori e FP-Growth.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

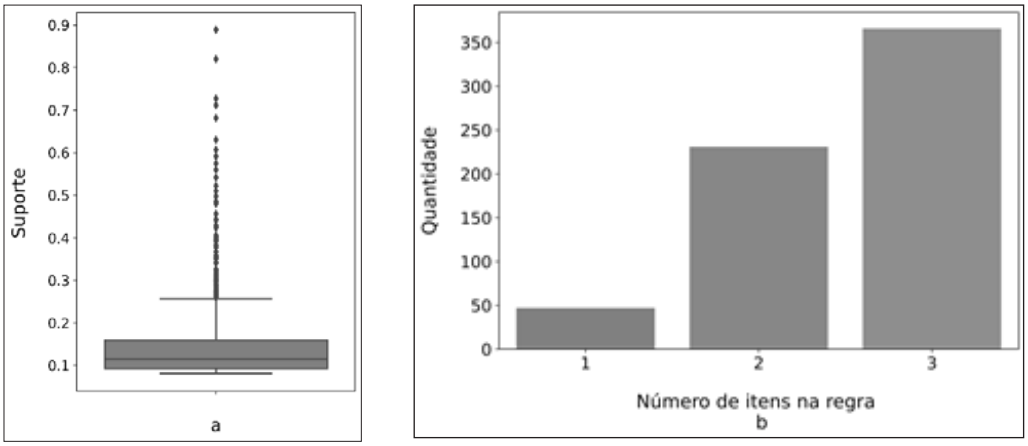

Figura 15. (a) Boxplot do suporte (b) Histograma da quantidade de características.
Fonte: elaborado pelo autor.

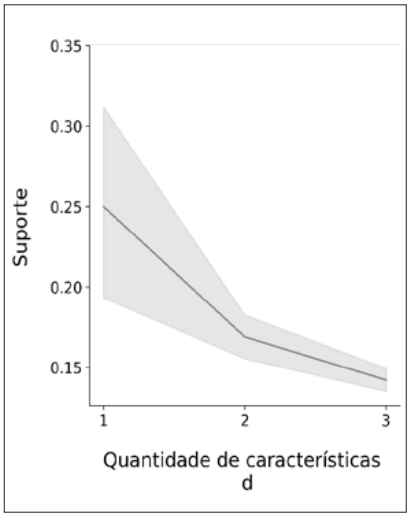

Figura 16. Média e desvio padrão dos suportes por quantidade de características.
Fonte: elaborado pelo autor.

Para obter a tabela IF-THEN do algoritmo Eclat, utilizou-se a biblioteca mlxtend com os resultados do pyEclat, já que este último não a gera diretamente.

A tabela IF-THEN foi criada com base apenas no índice de suporte, pois não foram obtidos valores de lift e confiança (ambos nulos). Ela contém um total de 394 regras, sendo possível visualizar um resumo das 5 principais regras ordenadas pelo suporte, conforme apresentado na Tabela 14.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Tabela 14. *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	-	-
{' Masculino'}	{'Reta '}	-	-	0,5596	-	-
{'Reta '}	{' Masculino'}	-	-	0,5596	-	-

No que diz respeito ao "IF", o Eclat apresentou apenas itens com uma característica. Mesmo assim, revelou associações interessantes como 'Reta' e 'Masculino'. No entanto, para uma avaliação mais aprofundada, optou-se por utilizar os resultados diretos do pyEclat, considerando apenas os índices de suporte. Os 10 principais itens com 3 características estão listados na Tabela 15.

Tabela 15. *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

Na tabela, observa-se a presença de associações entre 'Masculino', 'Dia Normal' e 'Horário Normal', semelhante aos outros algoritmos. Além disso, características como 'Céu Claro', 'Reta' e 'Dia Normal' também demonstram importância para análises.

No segundo cenário (Sem 'Dia Normal' e 'Horário Normal') com uma confiança mínima de 0,08, o algoritmo resultou em 53 regras, associando novamente 'Reta', 'Masculino', 'Simples' e 'Céu Claro'. Devido à presença de apenas uma característica, a Tabela 16 exibe os 10 principais itens sem a aplicação do IF-THEN.

Tabela 16. *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

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

item	Suporte
{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

No segundo cenário, foi possível observar a associação entre a característica de pista 'Reta' com o sexo 'Masculino' e a 'Falta de Atenção à Condução', bem como 'Automóvel de 86 a 111' e 'Céu Claro'. Dado que o sexo masculino estava presente em todos os casos, foi aplicado o terceiro cenário.

No terceiro cenário (Sem 'Dia Normal', 'Horário Normal' e 'Masculino') com um suporte mínimo de 0,05, com um suporte mínimo de 0,05, foram identificadas 64 regras com resultados serem semelhantes ao do segundo cenário.

Os resultados revelam associações relevantes, como {Pleno dia, Reta, Falta de Atenção à Condução}, que também foram identificadas em outros algoritmos. Esta regra é significativa para pesquisadores, engenheiros de trânsito e gestores.

Para comparar os algoritmos, foram utilizadas estatísticas descritivas, considerando suporte e quantidade de características por item, pois esses dados estavam disponíveis em todos os algoritmos. Além disso, uma análise de variância multivariada foi realizada para a comparação entre os algoritmos.

Comparação dos algoritimos

Foi conduzida uma análise de variância multivariada (MANOVA) para avaliar a diferença significativa entre as médias dos índices de suporte e tamanho das características. Os dados foram normalizados utilizando a biblioteca scipy.stats.norm do Python. Posteriormente, foi verificada a normalidade dos dados por meio do teste de Shapiro-Wilk, com a hipótese nula (H0) de que os dados seguem uma distribuição normal e a hipótese alternativa (H1) de que não seguem.

Os resultados do teste de Shapiro-Wilk indicaram que, após a normalização, os dados apresentaram uma distribuição normal, com um p-valor superior a 0,05. Isso significa que não há evidências para rejeitar a hipótese nula, permitindo a utilização do modelo MANOVA para a análise de variância. Na análise de correspondência de variância, os índices de suporte e a quantidade de características foram considerados como variáveis resposta.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Tabela 17. *Análise Multivariada de Variância*

Interceppt	**Pr > F**
Wilk's Lambda	0.8631
Pillai's trace	0.8631
Hotelling-Lawley trace	0.8631
Roy's gratestes	0.8631

Fonte: elaborado pelo autor.

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 estatísticas descritivas do estudo. Apesar das diferentes estatísticas de suporte e quantidade de características entre os algoritmos, todos apresentaram regras de associação relevantes nos três cenários estudados. Isso respondeu às perguntas de pesquisa e alcançou os objetivos do trabalho.

O estudo também contribuiu ao comparar diferentes algoritmos em diversas áreas de estudo, utilizando a linguagem Python em vez da tradicional ferramenta WEKA. Além disso, analisou e categorizou regras de associação em uma base de dados de acidentes brasileiros, incluindo as características dos veículos. Essa abordagem complementou os trabalhos relacionados e proporcionou uma compreensão mais aprofundada das causas dos acidentes, beneficiando engenheiros de segurança e gestores na formulação de políticas públicas.

Considerações finais

O principal objetivo deste trabalho consistiu em comparar algoritmos de aprendizado de máquina para fins de identificação das regras de associações 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. Esta pesquisa visou analisar independências das características dos acidentes e suas causas, bem como obter regras de associações por meio de algoritmos de aprendizado de máquina não supervisionados e, por fim, compará-los de modo a relacionar as regras de associações pertinentes para tomadas de decisões e políticas públicas.

O estudo se baseou em dados abertos de acidentes fornecidos pela Polícia Rodoviária Federal (PRF) do Brasil. A disponibilidade desses dados é crucial na era do *Big Data Analytics*, sendo fundamental para pesquisas em cidades inteligentes e no campo de transporte, contribuindo para o aprimoramento da segurança viária. Vale destacar que os dados referentes às características dos veículos e detalhes específicos dos acidentes não estão disponíveis de forma pública. Essa acessibilidade é considerada uma importante conquista desse estudo, visto que outros pesquisadores podem utilizar o mesmo método de obtenção dos dados.

O estudo empregou uma abordagem que considerou variáveis qualitativas. Foi utilizada a linguagem Python, uma plataforma de código aberto amplamente adotada por

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

cientistas de dados, para realizar a análise exploratória e tratamento dos dados. Esse método permitiu o pré-processamento e limpeza eficaz dos dados, incluindo a remoção de registros duplicados, ausentes e irrelevantes, contribuindo para a confiabilidade dos resultados. Além disso, a escolha do Python oferece a vantagem de poder ser replicado em outros estudos sem a necessidade de aquisição de software pago.

Durante a exploração dos atributos, foram gerados relatórios com representações gráficas dos dados dos acidentes, revelando indicadores relevantes para a análise das rodovias federais brasileiras. Foram identificados registros inconsistentes, incluindo outliers e dados faltantes, levantando questionamentos sobre os métodos de registro de acidentes pela PRF. Apesar das dificuldades no pré-processamento dos dados, o estudo obteve sucesso na aplicação dos algoritmos Apriori, Eclat, FP-Growth e FP-Max, com resultados significativos de lift, suporte e confiança.

Na análise dos resultados, foi constatado que os algoritmos Apriori, FP-Growth e Eclat demonstraram desempenho semelhante, exibindo índices de suporte e quantidade de características parecidos. À medida que a quantidade de características aumentava, o índice de suporte diminuía. Por outro lado, o FP-Max, ao propor uma métrica de suporte mais alta para um maior número de características, obteve um resultado oposto, fornecendo uma precisão maior. No entanto, tanto o FP-Max quanto o Eclat não geraram índices de lift e confiança para esse conjunto de dados.

Além dos índices de suporte e confiança significativos, os algoritmos revelam associações que suscitam novas reflexões, especialmente no que se refere à segurança viária. Por exemplo, quando um condutor do sexo masculino opera um veículo em um dia que não é feriado, fora do horário de pico, em uma estrada reta, essas características estão correlacionadas com acidentes causados por falta de distância de segurança. Da mesma forma, quando um condutor do sexo masculino dirige em um dia não feriado, fora do horário de pico, em uma reta, sob céu claro durante o dia, isso está associado a acidentes causados por falta de atenção na condução. Embora essas associações sejam intrigantes, o estudo ressalta a importância de que as características das regras de associação mais significativas sejam aquelas relacionadas a um maior número de acidentes, conforme indicado pelas métricas no relatório com representações gráficas dos acidentes. Portanto, para análises mais precisas, sugere-se em futuros estudos a aplicação desta metodologia em conjuntos de dados mais amplos e balanceados, nos quais cada variável possua uma quantidade equivalente de características.

O estudo atingiu com o objetivo de comparar algoritmos de aprendizado de máquina para identificar regras de associações entre as características viário-ambientais e veiculares em rodovias federais brasileiras.

Agradecimentos

O presente trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

Referências bibliográficas

» Ali, F. M. N., & Hamed, A. A. M. (2018). Usage of Apriori and clustering algorithms in WEKA tools to mine dataset of traffic accidents. *Journal of Information and Telecommunication, 2*(2), 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, 47*(4), 718–731.

» Amorim, B. S. P. (2019). *Uso de aprendizado de máquina para classificação de risco de acidentes em rodovias* (Dissertação de Mestrado em Ciência da Computação). Universidade Federal de Campina Grande, Campina Grande.

» Angolini, A. C. (2005). *Romi-Isetta: O pequeno pioneiro*. DBA Editora.

» Atnafu, B., & Kaur, G. (2017). Survey paper on analyzing and predicting the nature of road traffic accidents using data mining techniques in Maharashtra, India. *International Journal of Engineering Technology Science and Research, 53*(1), 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* (2nd ed.). In *Methods in molecular biology* (Vol. 1107). https://doi.org/10.1007/978-1-62703-748-8_7

» Blows, S., Ivers, R. Q., Woodward, M., Connor, J., Ameratunga, S., & Norton, R. (2003). Vehicle year and the risk of car crash injury. *Injury Prevention, 9*(4), 353–356. https://doi.org/10.1136/ip.9.4.353

» Borgelt, C., & Kruse, R. (2002). Induction of association rules: A priori implementation. *Proceedings of the 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* (pp. 273–276). Abu Dhabi. https://doi.org/10.1109/INNOVATIONS.2012.6207746

» Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. *Planalto*. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

» 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. *Planalto*. https://www.planalto.gov.br/ccivil_03/portaria/dlg6-2020.htm

» Brasil (2018). *Avaliação das políticas públicas de transportes: Segurança nas rodovias federais*. Ministério dos Transportes, Portos e Aviação, Brasília.

» Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica, 29*, 89–98. https://doi.org/10.31449/inf.v29i1.21

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

» Confederação Nacional dos Transportes – CNT. (2020). *Anuário da malha rodoviária*. https://anuariodotransporte.cnt.org.br/2020/Rodoviario/1-3-1-1-1-/Malha-rodovi%C3%A1ria-total

» 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, 3*(2), 139–157. https://doi.org/10.5380/atoz.v3i2.41346

» Creswell, J. W. (2007). *Projeto de pesquisa: Métodos qualitativo, quantitativo e misto* (2a ed.; L. de Oliveira da Rocha, Trad.). 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, 81*, 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, 111*, 43–55. https://doi.org/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), 6*(5), 6682–6685. https://doi.org/10.13140/RG.2.2.34583.52646

» Figueira, A. C., Pitombo, C. S., Oliveira, P. T. M., & 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, 5*(2), 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, 1*(2), 144–150. https://doi.org/10.4103/2249-4863.104987

» Hegland, M. (2007). The Apriori algorithm–a tutorial. In *Advances in data mining: Applications and theories* (pp. 79–94). World Scientific. https://doi.org/10.1142/9789812709066_0006

» Hunyadi, D. (2011). Performance comparison of Apriori and FP-Growth algorithms in generating association rules. In *The 5th European Computing Conference (ECC-11)*, Paris.

» 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. *Transportation Research Part F: Traffic Psychology and Behaviour, 27*, 81–89. https://doi.org/10.1016/j.trf.2014.09.008

» Jung, C. (2004). *Metodologia para pesquisa & desenvolvimento: Aplicada a novas tecnologias, produtos e processos*. Axcel Books do Brasil.

» Kaur, G. (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, 6*(3), 2244–2255.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

» Kumar, S., & Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data, 2*(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, 8*, 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, 5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

» 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)*, London.

» Martín, L., Baena, L., Garach, L., López, G., & de Oña, J. (2014). Using data mining techniques to improve road safety in Spanish roads. *Procedia - Social and Behavioral Sciences, 160*, 607–614. https://doi.org/10.1016/j.sbspro.2014.12.174

» 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* (pp. 3549–3559). Nanjing.

» Ministério da Infraestrutura (2020). *Frota de veículos – 2020*. https://www.gov.br/infraestrutura/pt-br/assuntos/transito/conteudo-denatran/frota-de-veiculos-2020

» Minsky, M. L. (1974). A framework for representing knowledge. *Massachusetts Institute of Technology A.I. Laboratory*.

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

» Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. The MIT Press.

» 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* (pp. 1504–1509). 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* (pp. 2350–2357). 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, 19*, 213–229. https://doi.org/10.1007/s10796-016-9673-7

» Polícia Rodoviária Federal – PRF (2020). *Dados abertos – acidentes*. https://portal.prf.gov.br/dados-abertos-acidentes. Acesso em 12 ago. 2020.

» Polícia Rodoviária Federal – PRF (2023). *Dados abertos – acidentes (agrupados por ocorrência em 2021)*. https://portal.prf.gov.br/dados-abertos-acidentes

» 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.

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

» 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.

» Shanti, S., Ramani, D. R. G., Shanthi, S., & Ramani, R. G. (2011). Classification of vehicle collision patterns in road accidents using data mining algorithms. *International Journal of Computer Applications, 35*(12), 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 ações governamentais para a redução de mortes e lesões no trânsito. *Revista Transporte y Territorio, 19*, 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* (pp. 758–763), 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, 5*, 1–12.

» Tyagi, A., Kumar, A., Gandhi, A., & Mueller, K. (2018). Road accidents in the UK (analysis and visualization). In *IEEE VIS 2018*.

» World Health Organization. (2018). *Global status report on road safety 2018*. Geneva: World Health Organization. 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, 137*, 680–687. https://doi.org/10.1016/j.proeng.2016.01.305

» Zhang, X. (2020). *A matrix algebra approach to artificial intelligence*. https://doi.org/10.1007/978-981-15-2770-8

Ramon Batista de Araújo / ramonbatista2006@gmail.com
Engenheiro Mecânico (Pontifícia Universidade Católica de Minas Gerais), Especialista em Engenharia de Produção (Pontifícia Universidade Católica de Minas Gerais), Cientista de Dados (Pontifícia Universidade Católica de Minas Gerais) e Discente no Departamento de Engenharia de Transportes e Geotecnia da Universidade Federal de Minas Gerais.

Marcelo Franco Porto / marceloporto@ufmg.br
Professor do Departamento de Engenharia de Transportes e Geotecnia - ETG da Escola de Engenharia da Universidade Federal de Minas Gerais - UFMG. Subcoordenador do Programa de Pós-Graduação em Geotecnia e Transportes (2021-2023). Membro titular do COMPUR -Conselho Municipal de Políticas Urbanas, PBH, representante do setor técnico (universidades). Chefe do Departamento de Engenharia de Transportes e Geotecnia - ETG / UFMG (2014 - 2016 e 2016 - 2018). Membro do Programa de Pós-graduação Geotecnia e Transportes da Escola de Engenharia da UFMG. Coordenador do Nucletrans - Núcleo Ensino e Pesquisa em Transportes da Escola de Engenharia da

ISSN 1852-7175
Transporte y Territorio 31 (julio-diciembre, 2024): [163-204]
doi: 10.34096/rtt.i31.11246

UFMG (CNPq). Coordenador do Laboratório Transcolar (CNPq) do FNDE/MEC. Doutor em Tratamento da Informação Espacial pela Pontifícia Universidade Católica de Minas Gerais, mestre em Ciências da Computação pelo DCC / UFMG, especialista em Gestão Estratégica pelo CEPEAD / UFMG, possui graduação em Engenharia Elétrica pela UFMG e graduação Superior em Tecnologia de Processamento de Dados pela Universidade FUMEC. Desenvolve pesquisas na área de Mobilidade Urbana, Sistemas Inteligentes de Transportes (Intelligent Transportation Systems) e Cidades Inteligentes (Smart Cities) além de Modelagem e Tratamento da Informação Espacial em engenharia - BIM (Building Information Modeling). Trabalha também com projetos rodoviários e ferroviários.

Renata Maria Abrantes Baracho Porto / renatabaracho@ufmg .br
Professora Associada da Universidade Federal de Minas Gerais - UFMG, Coordenadora do Programa de Pós-Graduação em Ambiente Construído e Patrimônio Sustentável da Escola de Arquitetura - PACPS/UFMG. Departamento de Tecnologia do Design, da Arquitetura e do Urbanismo? TAU/EA/UFMG. Pós doutorado / Visiting Scholar na University of South Florida - USF/USA. Visiting Researcher, Universitá della Svizzera Italiana- USI, Suíça Doutora em Ciência da Informação pela UFMG com PDEE na The Pennsylvania State University- USA, Mestre em Ciência da Computação DCC-UFMG, possui graduação em Arquitetura e Urbanismo e em Ciência da Computação. Presidente da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação - ANCIB (2014-2016). Member of International Institute of Informatics and Systemics - IIIS - International Federation for Systems Research - IFSR, IEEE, ICOM, ICOMOS. Professora do Programa de Pós-Graduação em Gestão e Organização do Conhecimento? PPGGOC-UFMG e do PACPS/UFMG. Áreas de atuação: Cidades Inteligentes, Smart City, Smart Building, Smart Life, Building Information Modeling/BIM, Sistemas de Informação, Modelagem, Ciência da Informação, com ênfase em Recuperação e Representação da Informação.