



**UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DEPARTAMENTO DE COMPUTAÇÃO E ELETRÔNICA
COLEGIADO DO CURSO DE ENGENHARIA DA COMPUTAÇÃO
ENGENHARIA DA COMPUTAÇÃO**

Gustavo Fardin Monti

**ANÁLISE DE DADOS DE ACIDENTES DE TRÂNSITO DA POLÍCIA RODOVIÁRIA
FEDERAL NO PERÍODO DE 2017 A 2019**

São Mateus, ES

2022

Gustavo Fardin Monti

**ANÁLISE DE DADOS DE ACIDENTES DE TRÂNSITO DA POLÍCIA RODOVIÁRIA
FEDERAL NO PERÍODO DE 2017 A 2019**

Projeto de Graduação apresentado ao Colegiado do Curso de Engenharia de Computação do Departamento de Computação e Eletrônica da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Bacharel em Engenharia de Computação

Universidade Federal do Espírito Santo – UFES
Departamento de Computação e Eletrônica
Colegiado do Curso de Engenharia da Computação

Orientador: Prof. Dr. Silvia das Dores Rissino.

São Mateus, ES

2022

Gustavo Fardin Monti

**ANÁLISE DE DADOS DE ACIDENTES DE TRÂNSITO DA POLÍCIA RODOVIÁRIA
FEDERAL NO PERÍODO DE 2017 A 2019**

Projeto de Graduação apresentado ao Colegiado do Curso de Engenharia de Computação do Departamento de Computação e Eletrônica da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Bacharel em Engenharia de Computação

BANCA EXAMINADORA

Profa. Dra. Silvia das Dores Rissino.
Dpto. de Computação e Eletrônica / UFES

Profa. Dra. Luciana Lee
Dpto. de Computação e Eletrônica / UFES

Prof. Dr. Oberlan Christo Romao
Dpto. de Computação e Eletrônica / UFES

São Mateus, ES

2022

AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer a Deus, pela oportunidade, pela força e pelas bençãos nesta jornada. Agradeço a meus pais, pela confiança, pelo apoio e por me ajudarem a chegar onde estou hoje. Sem eles, nada disso seria possível. Agradeço a minha família por se preocuparem comigo e sempre terem ajudado. Gostaria de agradecer a Ana Luiza, minha namorada, que tem me ajudado a evoluir e crescer com cada dia que passo ao seu lado. Gostaria de agradecer a minha orientadora por me dar a oportunidade de iniciar em um projeto de pesquisa e ter me guiado nesse caminho para produzir esta monografia. Agradeço a meus amigos Brian e Igor pelo apoio, risos e pela presença, apesar da distância. Agradeço aos meus colegas de curso pelo pelas amizades e laços que construímos, dia a dia, ao longo desses anos da graduação.

“A mind needs books like a sword needs a whetstone.”

Tyrion Lannister

RESUMO

Com dimensões continentais, o Brasil é um país que tem uma grande malha rodoviária, a qual ocupa papel de destaque no país com relação a integração, movimento e transporte de cargas, mas em função da extensão e quantidade de veículos em movimento, a malha rodoviária apresenta um quantitativo de acidentes que merece estudos. Este estudo tem por objetivo encontrar padrões entre atributos da base de dados de acidentes da Polícia Rodoviária Federal referente aos anos de 2017 até 2019, agrupados por pessoas, para melhor compreender fatores que influenciam o perfil das vítimas de acidente de trânsito. Aplica-se o processo de Descoberta de Conhecimento na base de dados abertos de acidentes. Para realizar as atividades deste estudo, utiliza-se a linguagem R e ferramenta RStudio. As atividades de pré-processamento e análise exploratória foram realizadas em conjunto. Na etapa de mineração de dados, utiliza-se o algoritmo Apriori, que resultou em um conjunto de regras de associação para cada valor das colunas de interesse (Sexo, Estado Físico e Tipo Envolvido), assim gerando aproximadamente 50 regras por par atributo-valor com confiança maior que 0,8 e lift maior que 1,2. Com a aplicação da descoberta de conhecimento sob a base de dados da Polícia Rodoviária Federal, utilizando a linguagem R como ferramenta, gerou-se regras de associação como uma forma de conhecimento. Ao interpretar as regras, encontra-se: Uma parte significante dos dados é composto por indivíduos masculinos, ilesos que são condutores; Passageiros estão associados ao sexo feminino e lesões leves; Pedestres estão associados a acidentes mais graves; Testemunhas são associadas a Valores ausentes.

Palavras chave: Dados Abertos; Mineração de Dados; Regras de Associação; Perfil de Acidentados.

ABSTRACT

With its continental dimensions, Brazil is a country with a great highway network, responsible for integrating the country by allowing movement and transport of merchandise. However, due to its large extension and high traffic of vehicles, the highway network has an abundant number of accidents that inspires research. The objective of this study was to find patterns between attributes of the road-traffic accident database grouped by individuals of the Federal Highway Police corresponding to the years of 2017 to 2019 to understand the factors that influence the profile of a road traffic accident victim. The process of Knowledge Discovery in Databases was applied to the public database. The R programming language and the RStudio tool were used to execute the necessary steps for this study. The steps of Pre-processing and Exploratory Analysis were executed simultaneously. In the data mining step, the Apriori algorithm was used, resulting in a set of association rules for each value of the columns of interest (Sex, Physical State, and Involvement), thereby producing approximately 50 rules per attribute-value pair with confidence larger than 0,8 and lift larger than 1,2. The results obtained by applying the Apriori algorithm were rules with confidence greater than 0,8 and lift greater than 1,2. Using the R programming language, we applied the process of Knowledge Discovery to the Federal Highway Police data set in order to generate association rules. By interpreting the rules, our findings are: A significant part of the dataset is comprised of male, uninjured drivers; Passengers are associated with female individuals who sustained minor injuries; Pedestrians are associated with more severe injuries; Witnesses are associated with missing values.

Keywords: Open Data; Data Mining; Association Rules; Road Traffic Accident Victim Profiles.

LISTA DE ILUSTRAÇÕES

Figura 1 - Ilustração de etapas do processo de KDD.....	17
Figura 2 - Processo das etapas de CRISP-DM.....	19
Figura 3 - Ilustração do processo da metodologia SEMMA.....	20
Figura 4 - Diagrama de ferramentas e pacotes utilizados.....	37
Figura 5 - Estado Físico dos acidentados agrupados por ano.....	47
Figura 6 - Classificação dos acidentes agrupados por ano.....	48
Figura 7 - Mapa de calor de distribuição acidentes de trânsito para o período de 2017-2019..	49
Figura 8 - Mapa de calor de numero de acidentes por data.....	50
Figura 9 - Horário de acidentes agrupados em intervalos de 30 minutos.....	51
Figura 10 - Wordclouds de marca de veículo agrupados por ano.....	52
Figura 11 - Sexo de acidentados por agrupado por ano.....	53
Figura 12 - Tipo de acidentados agrupado por ano.....	53
Figura 13 - Comparação de diagramas de caixas de idade de acidentados com e sem pré- processamento.....	54
Figura 14 - Grafo para o conjunto de regras extraídas para "sexo=Masculino".....	70
Figura 15 - Grafo para o conjunto de regras extraídas para "sexo=Feminino".....	71
Figura 16 - Grafo para o conjunto de regras extraídas para "sexo=Não Informado".....	72
Figura 17 - Grafo para o conjunto de regras extraídas para "estado_fisico=Ileso".....	73
Figura 18 - Grafo para o conjunto de regras extraídas para "estado_fisico=Lesões Leves"....	74
Figura 19 - Grafo para o conjunto de regras extraídas para "estado_fisico=Lesões Graves"....	75
Figura 20 - Grafo para o conjunto de regras extraídas para "estado_fisico=Óbito".....	76
Figura 21 - Grafo para o conjunto de regras extraídas para "estado_fisico=Não Informado" ..	77
Figura 22 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Condutor".....	78
Figura 23 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Passageiro".....	79
Figura 24 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Pedestre".....	80
Figura 25 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Testemunha"....	81

LISTA DE EQUAÇÕES

Equação 1 - Suporte para um conjunto de itens.....	27
Equação 2 - Suporte para uma regra de associação.....	27
Equação 3 - Confiança para uma regra de associação.....	27
Equação 4 - Lift para uma regra de associação.....	28
Equação 5 - Confiança inversa mínima.....	44

LISTA DE TABELAS

Tabela 1 - Resultados de análise exploratória inicial elaborado de acordo com Pearson (2018)	39
Tabela 2 - Contabilização de identificadores para a base de dados.....	47
Tabela 3 - Municípios com maior quantidade de acidentes por ano.....	49
Tabela 4 - Rodovias com maior quantidade de acidentes por ano.....	50
Tabela 5 - Medidas obtidas para idade pré-processada através do diagrama de caixas.....	54
Tabela 6 - Métricas para as regras de associação para o atributo sexo.....	55
Tabela 7 - Métricas para as regras de associação para o atributo estado físico.....	57
Tabela 8 - Métricas para as regras de associação para o atributo tipo envolvido.....	59

LISTA DE QUADROS

Quadro 1 - Exemplo de um conjunto de transações de supermercado.....	27
Quadro 2 -Setup Experimental.....	37

SUMÁRIO

1	INTRODUÇÃO.....	12
1.1	CONSIDERAÇÕES GERAIS.....	12
1.2	OBJETIVO GERAL.....	14
1.3	OBJETIVOS ESPECÍFICOS.....	14
1.4	ORGANIZAÇÃO DO TRABALHO.....	14
2	REFERENCIAL TEÓRICO.....	15
2.1	MINERAÇÃO DE DADOS.....	15
2.1.1	Motivações.....	15
2.1.2	Metodologias.....	16
2.1.3	Evolução e Atualidade.....	21
2.2	REGRAS DE ASSOCIAÇÃO.....	25
2.2.1	Introdução e conceitos.....	25
2.2.2	Métricas.....	27
2.2.3	Algoritmos Comuns.....	28
2.3	TRABALHOS RELACIONADOS.....	30
3	METODOLOGIA.....	32
3.1	FERRAMENTAS UTILIZADAS.....	32
3.1.1	Ambiente de Desenvolvimento.....	32
3.1.2	Documentação.....	32
3.1.3	Processamento e Transformação de Dados.....	33
3.1.4	Visualização de Dados.....	34
3.1.5	Mineração de Regras de Associação.....	35
3.1.6	Visualização de Regras de Associação.....	36
3.1.7	Setup Experimental.....	36
3.1.8	Diagrama de Uso De Ferramentas.....	37
3.2	EXPERIMENTOS.....	37
3.2.1	Análise de Domínio.....	37
3.2.2	Seleção de Conjunto de Dados.....	38
3.2.3	Limpeza e Pré-processamento.....	39
3.2.4	Redução e Projeção de Dados.....	42
3.2.5	Seleção de Método de Mineração.....	42

3.2.6	Análise Exploratória de Mineração.....	43
3.2.7	Mineração de Dados.....	43
3.2.8	Interpretação de Resultados.....	44
4	RESULTADOS E DISCUSSÕES.....	46
4.1	ANÁLISE EXPLORATÓRIA.....	46
4.1.1	Identificadores.....	46
4.1.2	Estado Físico.....	47
4.1.3	Classificação de Acidentes.....	48
4.1.4	Distribuição Geográfica.....	48
4.1.5	Dados Temporais.....	50
4.1.6	Marca de Veículos.....	52
4.1.7	Informações de Acidentados.....	52
4.2	REGRAS DE ASSOCIAÇÃO.....	54
4.2.1	Perfis de Acidentados Por Sexo.....	55
4.2.2	Análise de Perfil por Estado Físico.....	56
4.2.3	Análise de Acidentes Por Tipo de Envolvido.....	59
5	CONSIDERAÇÕES FINAIS.....	62
6	TRABALHOS FUTUROS.....	63
	REFERÊNCIAS.....	64
	APÊNDICE A – GRAFOS DE REGRAS DE ASSOCIAÇÃO EXTRAÍDAS	69
	A.1 GRAFOS PARA O ATRIBUTO SEXO.....	70
	A.2 GRAFOS PARA O ATRIBUTO ESTADO FÍSICO.....	73
	A.3 GRAFOS PARA O ATRIBUTO TIPO ENVOLVIDO.....	78

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES GERAIS

As dimensões continentais do Brasil fazem com que o país possua uma malha rodoviária grande, importante e responsável pela integração, movimentação e transporte de cargas produzidas internamente ou oriundas de importação e exportação. Segundo o Ministério da Infraestrutura, o transporte rodoviário é responsável por mais de 65% da movimentação de cargas (BRASIL, 2019a). A malha rodoviária brasileira federal possui 75.8 mil km, sendo que 64.4 mil km são pavimentadas e 10.4 mil km não são pavimentadas (BRASIL, 2019b). Neste contexto, de extensão, movimentação e transporte de cargas há uma quantidade de veículos em movimento, o que possibilita um quantitativo de acidentes que merecem estudos.

Estima-se que anualmente 50 milhões de pessoas são feridas e 1.2 milhões morrem como consequência de acidentes, com alta prevalência de sequelas físicas e distúrbios psicológicos pós acidentes em todos os níveis de gravidade (MESQUITA FILHO, 2012). Além do impacto físico e psicológico sobre as vítimas e suas famílias, há também um impacto econômico significante. Com uma demanda abundante de atendimentos provenientes de acidentes automobilísticos, a prevenção de acidentes se torna um ato importante da gestão de saúde pública dado a redução de gastos diretos no Sistema Único de Saúde (MASSAÚ E ROSA, 2016). Para reduzir o volume de acidentes, melhores diretrizes de políticas públicas devem ser adotadas, entre as quais, pode se destacar campanhas educativas, melhorias nas estruturas de gestão, fiscalização, legislativas e na infraestrutura viária, através de intervenções de engenharia visando identificar pontos críticos (CARVALHO, 2016).

Os serviços públicos que lidam com acidentes geram um volume de dados cada vez maior como consequência dos avanços nos meios de coleta e armazenamento de dados. A integração de múltiplas fontes de dados permite subsidiar o planejamento e execução de ações de segurança viária, contudo a falta de apoio político dificulta a sustentabilidade deste processo dado a dificuldade de acesso às bases de dados (ABULATIF, 2018). Encorajados pela adoção de legislação que incentiva o compartilhamento de informações em conjunto com a adoção de práticas de governo aberto pelo EUA, os executivos do Governo Brasileiro adotaram diretrizes de disponibilização de informação, preferencialmente utilizando a

Internet, com o intuito de gerar novos conhecimentos, serviços e produtos (RIBEIRO E ALMEIDA, 2011).

Devido aos fatos citados, e com a aprovação da Lei 12.527 (BRASIL, 2011), também conhecida como lei de Acesso a Informação, criou-se o Portal Brasileiro de Dados Abertos, que atualmente possuí 10.268 conjuntos de dados disponibilizados, de vastas áreas da administração pública que podem ser acessados, utilizados, modificados e compartilhados por qualquer pessoa (BRASIL, 2021). Além disto, a Polícia Rodoviária Federal (PRF) anualmente disponibiliza, através do seu portal, conjuntos de dados abertos referentes a acidentes e infrações de trânsito. A partir dos dados disponibilizados, é interessante considerar a possibilidade de gerar conhecimento que ajude compreender os fatores que influenciam e causam acidentes de trânsito. Para isso, é necessário recorrer a metodologias próprias para a análise de grandes volumes de dados.

Com a disponibilização de dados e a utilização de metodologias que permitem extrair informações, é possível realizar estudos quanto as suas aplicações para garantir a segurança no trânsito para a população. Através de uma revisão de literatura, destaca-se a importância da Mineração de Dados, que pode ser utilizada para identificar padrões que auxiliam na tomada de decisão de prevenção de acidentes, contribuindo para a redução dos gastos com a Saúde no Trânsito (GODOI E GUIMARÃES, 2014).

Através da revisão de literatura realizado na plataforma dos periódicos CAPES, notou-se a carência de estudos que utilizam dados dos últimos 5 anos da Polícia Rodoviária Federal.

Dados de acidente de trânsito são disponibilizados anualmente pela Polícia Rodoviária Federal, e com o passar do tempo, imagina-se que há melhorias na qualidade dos dados disponibilizados. Ao analisar estes dados, pode ser que as dificuldades encontradas em estudos passados ainda persistam ou tenha sido solucionadas.

Há também a possibilidade de ocorrer mudanças ou persistência nos padrões de acidentes que ocorrem com base nas estratégias adotadas para reduzir a quantidade de acidentes. Por isso, é do interesse de todos continuar o acompanhamento através da análise dos dados de acidentes.

Outro fator observado foi a recorrência na análise de subdomínios no conjunto de dados, onde muitas vezes os estudos se limitam a uma região ou a uma rodovia. É evidente que ao diminuir o escopo da análise, os resultados encontrados possuem melhor interpretabilidade quando leva-se em consideração os fatores regionais, entretanto, identificar padrões que valem para toda malha rodoviária federal também tem seu valor.

1.2 OBJETIVO GERAL

Este trabalho tem por objetivo encontrar padrões entre atributos da base de dados de acidentes da Polícia Rodoviária Federal (PRF) referente aos anos de 2017 até 2019, para compreender fatores que influenciam o perfil das vítimas de acidente de trânsito. Através deste estudo, busca-se complementar e atualizar a bibliografia existente utilizando dados atualizados.

1.3 OBJETIVOS ESPECÍFICOS

- Estudar as etapas de descoberta de conhecimento em bases de dados;
- Estudar a linguagem de programação R;
- Estudar técnicas de mineração de dados para a aplicação em dados de acidentes de trânsito;
- Aplicar as etapas da descoberta de conhecimento na base de dados da Polícia Rodoviária Federal utilizando a linguagem de programação R;
- Identificar e discutir padrões entre os dados de acidentes de trânsito;
- Gerar resultados reproduzíveis e que possam ser facilmente explorados.

1.4 ORGANIZAÇÃO DO TRABALHO

No Capítulo 2, realiza-se uma revisão bibliográfica, analisando conceitos da Mineração de Dados, A Metodologia de Descoberta de Conhecimento e Regras de associação. Em sequência, no Capítulo 3, descreve-se as técnicas empregadas a este estudo. No Capítulo 4, apresenta-se os resultados em cada etapa do processo de descoberta de conhecimento. No Capítulo 5, apresenta-se a conclusão e por último no Capítulo 6, sugere-se possíveis trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo será apresentada uma revisão bibliográfica referente ao estudo da mineração de dados. Desta forma será feita uma introdução a conceitos de mineração de dados, metodologias utilizadas e algoritmos de mineração por regras de associação.

2.1 MINERAÇÃO DE DADOS

2.1.1 Motivações

A análise exploratória de dados não é algo novo. Para modelar dados multivariados, é comum recorrer a técnicas tradicionais da estatística como regressões, análise discriminante e *Naive Bayes* (SMYTH, PREGIBON & FALOUTSOS, 2002). Podemos brevemente listar os diferentes períodos que contribuíram para chegar na mineração de dados e comparar suas características (TUFFÉRY, 2011):

- Estatística Tradicional (Até 1950): Análise de centenas de observações; Algumas variáveis definidas com algum protocolo (*sampling, experimental design, etc*); Suposições firmes em relação as distribuições envolvidas; Desenvolvimento teórico de modelos e comparação com os dados; Métodos probabilísticos e estatísticos; Utilizado em laboratórios.
- Análise de dados (1960 – 1980): Análise de milhares de observações; Dezenas de variáveis envolvidas; Construção de tabelas de indivíduos x variáveis; Uso da computação e representação visual.
- Mineração de dados (+1990): Milhões de indivíduos; Centenas/milhares de variáveis analisadas; Uso elevado de variáveis não numéricas (textuais ou imagens); Suposições fracas sobre as distribuições envolvidas; Coleta de dados posterior ao estudo; Presença de dados anormais e erros; Aplicado no mercado.

Podemos dizer que a mineração de dados surgiu como consequência de três grandes avanços técnicos: As melhorias de capacidade armazenamento computacional – Como consequência do desenvolvimento de *data warehouses*, arquiteturas distribuídas e avanços computacionais; A disponibilização (crescente) e evolução de pacotes de algoritmos estatísticos e de mineração de dados; Mudanças na visão de como decisões devem ser

tomadas – Uso de análise e de mineração de dados em processos de produção e não em estudos unitários. Como consequência destes avanços, Surgiu-se a possibilidade de processar qualquer tipo de dado (mesmo que anormal, ou até dados textuais). Um último ponto que vale destacar é que as bases de dados começaram ser desenhadas em torno dos requisitos de negócio, e esses conjuntos de dados possuem um potencial até então não explorado (TUFFÉRY, 2011).

O termo mineração de dados se popularizou entre a comunidade acadêmica durante o final dos anos 80. Naquela época a definição e escopo da palavra ainda estava em aberto, e pode-se argumentar que este é o caso até hoje. A mineração de dados pode ser brevemente definida como o conjunto de mecanismos e técnicas, implementados em software, que permitem extrair informação oculta a partir de dados. Uma definição que se popularizou e é comumente adotada é a de Fayaad que enxerga a mineração de dados como um subprocesso dentro do processo de descoberta de conhecimento em bases de dados (COENEN, 2011).

A literatura sobre mineração de dados acaba sendo uma fonte de confusão quanto a definição do termo (AZEVEDO E SANTOS, 2008). O interesse crescente acabou tornando a mineração de dados em uma *buzzword* obscura. Apesar das diferentes definições, é de comum acordo que o campo de mineração de dados combina análise estatística, aprendizado de máquina e gerenciamento de bancos de dados para extrair conhecimento de grandes bases de dados (THURAISINGHAM, 2000).

2.1.2 Metodologias

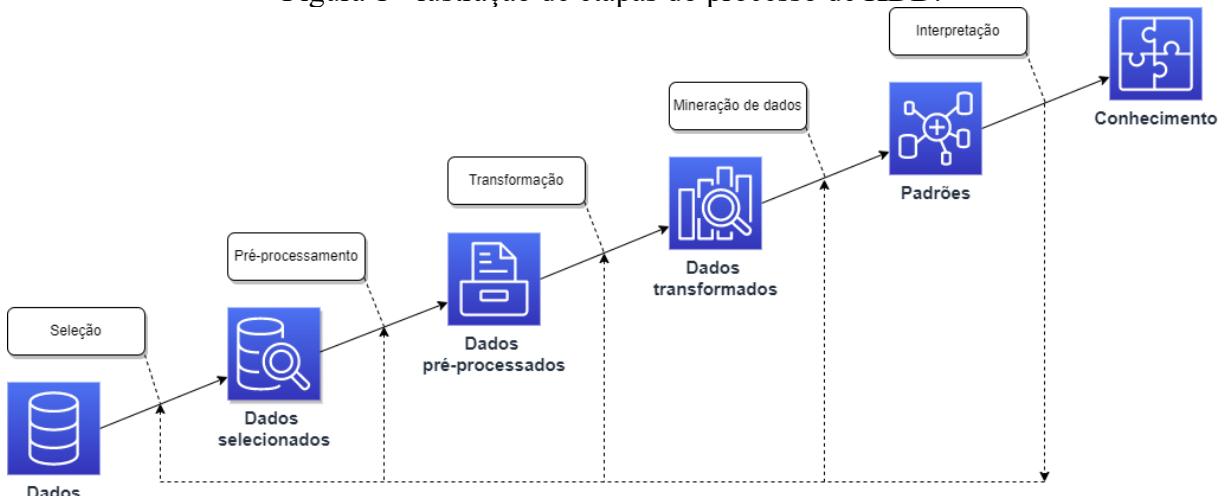
Para realizar a mineração de grandes bases de dados, podemos recorrer a metodologias que descrevem o processo de mineração passo a passo, como um projeto a ser desenvolvido. Diferentes metodologias foram propostas para guiar o processo de análise de grandes volumes de dados.

O processo de Descoberto de Conhecimento em Base de dados (*Knowledge Discovery in Databases* - KDD), como sendo o processo não trivial de identificar padrões válidos, novos, potencialmente úteis, e acima de tudo, comprehensíveis, no conjunto de dados. O KDD é um processo interativo e iterativo, e envolve vários passos que podem ser resumidos como sendo (FAYYAD, PIATETSKY-SHAPIRO E SMYTH, 1996):

1. Entender o domínio de aplicação, conhecimento prévio relevante e identificar o objetivo do processo;

2. Seleção e/ou criação de um conjunto de dados para realizar o processo de descoberta de conhecimento;
3. Limpeza de dados e pré-processamento – Isso inclui tratamento de: ruídos; valores ausentes; análise de informação de sequência de tempo;
4. Redução e projeção dos dados para encontrar variáveis úteis de acordo com o objetivo da tarefa;
5. Alinhar os objetivos do processo para escolher um método de mineração – pode ser classificação, regressão, agrupamento, ou sumarização;
6. Análise exploratória para entender o que é desejado da mineração e escolher o algoritmo(s) de mineração;
7. Mineração de dados – inclui procurar padrões de interesse dentro de uma representação ou conjuntos de representações que podem ser regras ou árvores de classificação, regressão ou agrupamentos;
8. Interpretação dos padrões minerados – pode incluir a visualização dos padrões minerados ou dos dados;
9. Agir sobre o conhecimento descoberto – Aplicar o conhecimento descoberto em algum outro sistema ou simplesmente documentá-lo.

Figura 1 - Ilustração de etapas do processo de KDD.



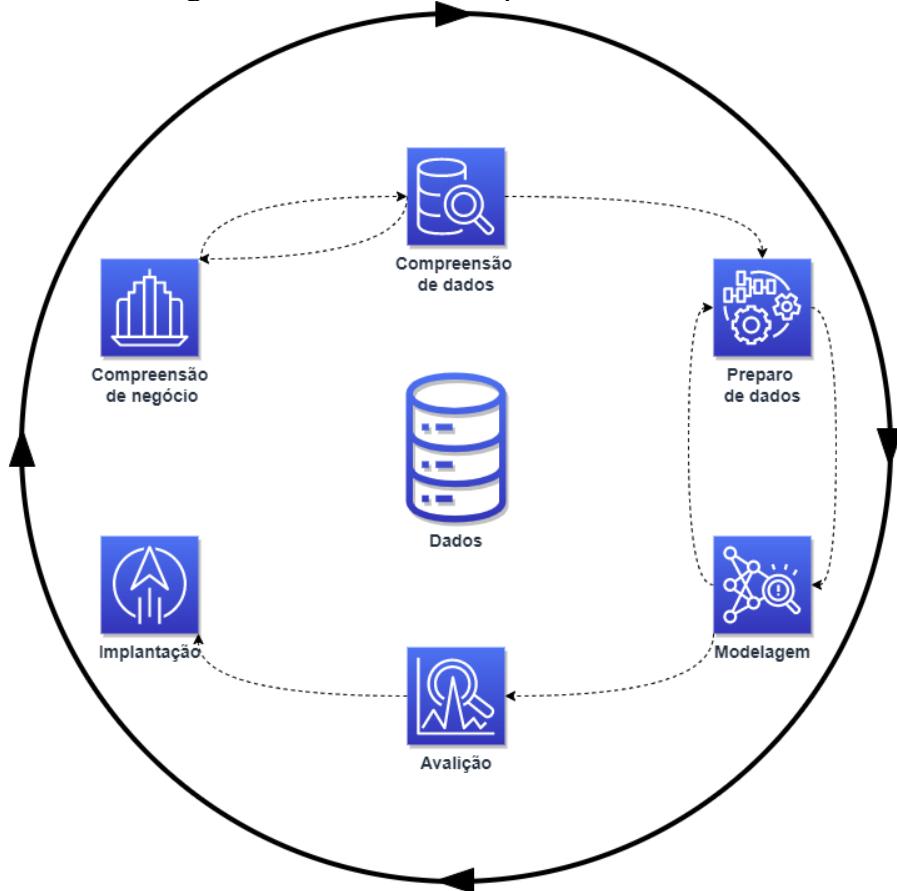
Fonte: Adaptado De Fayyad, Piatetsky-Shapiro E Smyth (1996).

Há ainda outras alternativas para abordar o processo de mineração de dados. O Processo Padrão Inter-Indústrias para Mineração de Dados (*Cross Industry Standard Process for Data Mining* – CRISP-DM) é uma metodologia que descreve o ciclo de vida de um projeto de mineração de dados, contendo etapas de compreensão de negócio, compreensão de dados, preparo de dados, modelagem, avaliação e implantação (WIRTH & HIPP, 2000).

O modelo de referência CRISP-DM demonstra um apanhado do ciclo de vida de um projeto de mineração de dados. Através do círculo externo na ilustração, nota-se a natureza cíclica da mineração de dados em si. A mineração de dados não termina quando o projeto está implantando, pois o aprendizado daquele processo pode ser utilizado para direcionar novos projetos. A baixo, destacamos e caracterizamos as fases (WIRTH & HIPP, 2000):

- Compreensão de negócio: Entender os objetivos e requisitos do projeto do ponto de vista de negócio, convertendo o conhecimento para um plano de projeto.
- Compreensão de dados: Inclui a coleta inicial de dados, se familiarizar com os dados, obter *insights* iniciais e/ou detectar subconjuntos interessantes para formular hipóteses sobre informações ocultas nos dados.
- Preparo de dados: Envolve todas as atividades para construir o conjunto de dados final que é alimentado para a ferramenta(s) de modelagem. Esta fase normalmente é realizada múltiplas vezes, e não necessariamente segue alguma ordem. As tarefas podem incluir seleção de tabelas, linhas ou atributos, além da limpeza de dados e criação/transformação de atributos.
- Modelagem: Várias técnicas de modelagem podem ser aplicadas e os seus parâmetros podem ser calibrados de forma experimental. Um mesmo problema pode ser modelo com múltiplas técnicas. Normalmente é na etapa de modelagem que surge a necessidade de ajustar o preparo de dados.
- Avaliação: Uma vez que pelo menos um modelo já foi elaborado, é necessário avaliar que o modelo encontrado está de acordo e atinge os objetivos de negócio.
- Implantação: Uma vez que o modelo foi criado, é necessário organizar a apresentar o conhecimento obtido de uma forma que um cliente pode utilizar. Isto pode ser gerar um relatório ou implementar um processo de mineração cíclico.

Figura 2 - Processo das etapas de CRISP-DM.



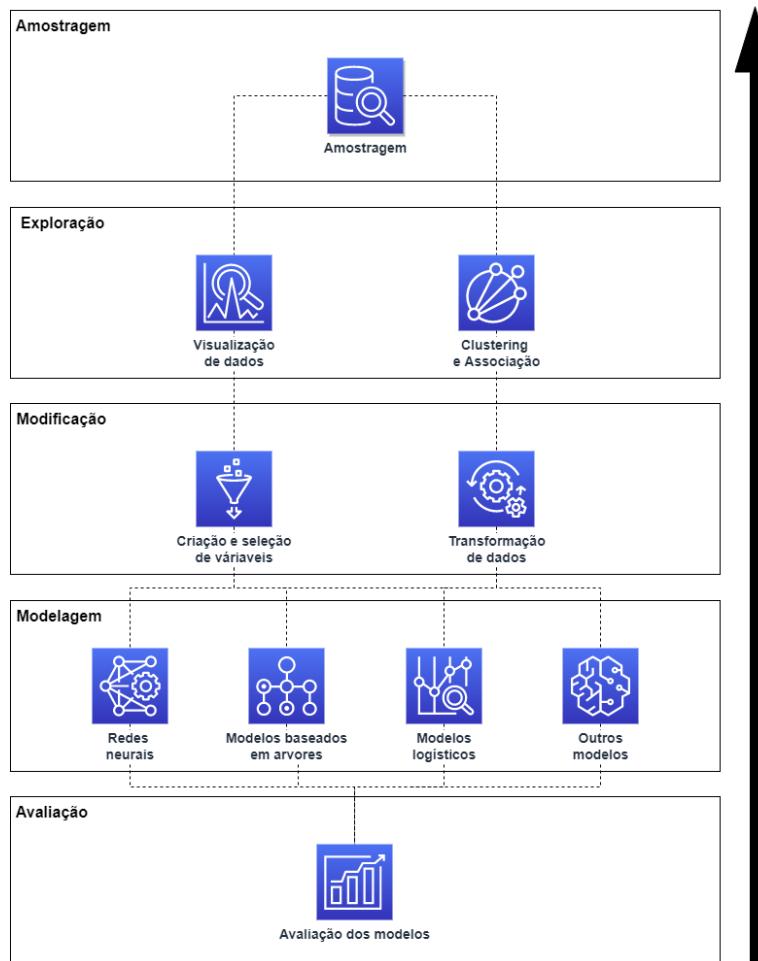
Fonte: Adaptado De Wirth E Hipp (2000).

A metodologia SEMMA, desenvolvida pelo instituto SAS, descreve o processo de Mineração de Dados com etapas de amostragem, exploração, modificação, modelagem e avaliação de um grande volume de dados, com o objetivo de descobrir padrões desconhecidos (SAS INSTITUTE, 2017).

- Amostragem: Criação de uma ou mais tabelas dentro da ferramenta SAS. A amostra deve ser grande o suficiente para conter informações significantes, mas pequena o suficiente que seu processamento não se torne um problema. Pode ser que a amostragem não seja necessário e o conjunto de dados completo seja usado.
- Exploração: Obter conhecimento sobre as relações que existem dentro dos dados, identificar padrões não esperados e anomalias. Este conhecimento ajuda a conhecer os dados e permite formular ideias para continuar o processo. A tarefa pode ser desenvolvida através da visualização de dados, criação de agrupamentos e associação entre variáveis.

- Modificação: É o processo de transformar os dados e suas variáveis para alimentar o processo de modelagem. Variáveis podem ser selecionadas, removidas ou criadas conforme necessário.
- Modelagem: A modelagem é realizada através de ferramentas analíticas que permitem procurar uma combinação de dados que é capaz de prevê um resultado desejado de forma confiável. Para isso é possível utilizar redes neurais, modelos de árvores de decisão, modelos logísticos ou outras técnicas de modelagem estatística.
- Avaliação: Avaliação dos resultados encontrados levando em conta o quanto útil são os resultados e o quanto confiáveis eles são.

Figura 3 - Ilustração do processo da metodologia SEMMA.



Fonte: Adaptado De Sas Institute (2017).

Um paralelo pode ser feito comparando as diferentes metodologias voltadas para mineração de dados. Ao comparar as metodologias SEMMA e CRISP-DM, é possível visualizá-las como sendo guias de implementações do processo de KDD em sistemas reais (AZEVEDO E

SANTOS, 2008). Apesar das técnicas terem uma quantidade de fases diferentes, ao analisar a definição de cada etapa, a necessidade de realizar as tarefas são inerentes ao processo.

2.1.3 Evolução e Atualidade

Antes de começar a tratar sobre os avanços em mineração de dados, precisamos primeiro introduzir a evolução de arquiteturas para organização de dados. Utilizaremos então este tópico para compreender como as técnicas de organização de dados evoluíram ao longo do tempo, e como isto nos permite a entender como essas mudanças influenciaram a mineração de dados.

Data warehouses surgiram para permitir que líderes de negócios conseguissem gerar *insights* analíticos organizando dados de banco de dados transacionais em *warehouses* centralizados. Os dados são escritos com um esquema definido (*schema-on-write*), o que garante que os dados estavam otimizados para consumo. *Data warehouses* normalmente armazenam os dados em formatos de arquivo proprietários que são mais otimizados, mas podem se tornar um problema – *vendor lock-in* (ZAHARIA ET. AL., 2021). *Data warehouses* podem ser destacados como sendo a fundação de sistemas de suporte de decisão. É possível realizar a mineração de dados sem um *data warehouse*, mas as características de um *data warehouse* aumentam a probabilidade de sucesso de um projeto de mineração de dados. A integração de dados; disponibilização de dados detalhados, agregados, históricos e metadados contribuem para otimizar o processo de mineração (INMON, 1996).

Com o passar do tempo esses tipos de sistemas começaram a sofrer com problemas de desempenho, ao qual podemos atribuir os motivos de acoplamento entre armazenamento e processamento com recursos *on-premises*; crescimento rápido dos dados e introdução de dados não estruturados como vídeo, áudio e documentos de texto (ZAHARIA ET. AL., 2021).

Podemos dizer que o termo *big data* surgiu como resposta a massa de dados que vem sendo gerada por máquinas, dispositivos, soluções de nuvem, etc. Os principais desafios que surgem com *big data* são a captura, armazenamento, processamento, análise, filtragem, busca, compartilhamento e segurança de dados. O termo normalmente é definido em torno dos três Vs: Velocidade (de crescimento e mudança dos dados gerados); Variedade (de diferentes e múltiplos formatos) e Volume (grande quantidade de dados que são gerados a cada segundo). Dada as oportunidades potenciais para descobertas, *big data* pode ser uma chave para inovação, competitividade e produtividade. Há necessidade de utilizar tecnologias e

metodologias avançadas para garantir desempenho, confiabilidade e escalabilidade para lidar com os problemas de modelagem e mineração de *big data* (SOWMYA & SUNEETHA, 2017). No geral, o processamento de *big data* é voltado para a extração ou mineração de conhecimento de quantidades massivas de dados. O processamento pode ser de três tipos. O processamento em *batch* é onde os dados são primeiro armazenados e depois processados. O processamento por *streaming* possuí dados armazenados após a realização do processamento. O processamento híbrido é composto por quatro camadas. A camada em *batch* possuí dados imutáveis que crescem continuamente. A camada de serviço é utilizada para disponibilizar visões sobre os dados *batch*. A camada de velocidade serve para compensar atualizações de alta latência. Por último a camada de combinação é utilizada para sincronização dos dados (MIOSLAVSKAYA E TOLSTOY, 2016).

Em resposta ao conceito de *big data*, podemos introduzir o conceito de *data lakes*. Um *data lake* é um repositório de armazenamento com grande capacidade de escalabilidade, capaz de armazenar dados crus (sem processamento) em conjunto um motor de processamento que pode disponibilizar os dados. *Data lakes* são construídos para conseguir lidar com grandes volumes de dados não estruturados que chegam rapidamente. Note que o termo *data lake* tem sido fortemente associado com produtos que suportam Hadoop no mercado. O *data lake* pode servir como um local barato para armazenar grandes volumes de dados e conduzir uma análise preliminar sobre os dados, realizando atividades de estruturação de dados e produzindo como saída dados analisados que podem alimentar um sistema analítico (MIOSLAVSKAYA E TOLSTOY, 2016).

Os *data lakes* surgiram como uma segunda geração de plataformas de análise de dados: Estes sistemas possuem um custo de armazenamento reduzido, utilizando formatos genéricos e normalmente abertos, como Apache Parquet e ORC. Diferente dos *data warehouses*, o *schema* do *data lake* é definido em momento de leitura (*schema-on-read*) que fornece agilidade e tem baixo custo de armazenamento. O Apache Hadoop foi uma das tecnologias que auxiliaram para dar início a este padrão de arquitetura, utilizando o Hadoop File System como meio barato de armazenamento. Plataformas em nuvem se tornaram a principal forma de hospedar *data lakes*, como o uso do Amazon Simple Storage Service (S3), Azure Data Lake Store (ADLS) e Google Cloud Storage (GCS). Estas plataformas possuem recursos como georeplicação, alta durabilidade dados, e disponibilidade de *storage* de arquivamento (ZAHARIA ET. AL., 2021).

Podemos relacionar as características de grandes volumes de dados e natureza do tipo de dado que um *data lake* é utilizado para armazenar. Vemos que com o conceito de *data lake*,

temos uma estrutura voltada para armazenar e disponibilizar grandes volumes de dados de forma escalável.

Grandes volumes de dados já são minerados por governos através de redes sociais, blog, e outras fontes de informação para identificar atividades suspeitas e predizer eventos futuros. A demanda por algoritmos de aprendizado de máquina e mineração de dados escaláveis e paralelos tem aumentado significativamente. O NIMBLE é uma plataforma construída em cima do Hadoop MapReduce para habilitar a rápida implementação de algoritmos de mineração de dados e aprendizado de máquina. O Apache Mahout é uma biblioteca de implementações de algoritmos de mineração de dados e aprendizado de máquina. Outras alternativas são o BC-PDM, PEGASUS, Giraph e Graph Lab. Há entretanto uma incompatibilidade entre as demandas de gerenciamento de *big data* e o que os sistemas tradicionais de dados podem oferecer (CHE, SAFRAN & PENG, 2013).

Técnicas comuns de aprendizado de dados tradicional como *Naive Bayes*, *Support Vector Machines* e *Singular Value Decomposition* podem ser implementadas com MapReduce. Entretanto, estas técnicas podem não ser aplicáveis a *big data analytics* devido à diferença de conjuntos de dados de treino, dificuldade para paralelizar as tarefas de treino e restrições de memória. Algumas técnicas tradicionais de mineração de dados como *clustering* podem ser adaptados para analisar amostras de dados, mas outras como o *K-Nearest Neighbour* requerem manter todos os dados de treino de memória e não são aplicáveis para *big data analytics* (WANG, 2017).

Os principais desafios de e problemas e desafio relacionados a *big data mining* atualmente são (CHE, SAFRAN E PENG, 2013):

- Variedade: A modelagem dos dados através de um único modelo pode gerar resultados insatisfatórios. Alguns domínios que são objetivos de estudos possuem múltiplas entidades e relações complexas entre elas, dificultando a modelagem.
- Escalabilidade: Dado o grande volume requer alta escalabilidade nas ferramentas de gerenciamento e mineração de dados. Duas possíveis soluções para este problema são o uso de computação em nuvem e suporte a interação avançada para usuários – ambos com o objetivo de permitir e otimizar a navegação de grandes volumes de dados para mineração.
- Velocidade: Há obrigação de realizar processamentos e tarefas de mineração em um curto intervalo de tempo, principalmente para *streams* de dados, evitando que o resultado perca seu valor com o passar do tempo.

- Acurácia e Confiabilidade: Enquanto no passado os dados consumidos pelos algoritmos de mineração eram curados e tratados, atualmente nem todos os dados são modelados ou validados. Há necessidade de monitorar, mensurar e comparar a evolução dos dados com métricas de confiabilidade.
- Privacidade: É um problema que existe desde quando a mineração de dados começou a ser aplicada sobre dados do mundo real. Há necessidade de melhorias de regulamentação quanto ao acesso de dados pessoais.
- Interatividade: A melhora de interatividade é necessário para otimizar o processo de mineração seja para que o usuário decida qual subconjunto dos dados devem ser minerados ou identifique e avalie as relações complexas existentes nos dados.
- Mineração de lixo: Grande parte dos dados disponibilizados possuem informações datadas, corrompidas que precisam ser analisadas e tratadas (ou “recicladas”) no futuro próximo.

Uma desvantagem dos *data lakes* é a necessidade de garantir a qualidade e governança de dados se torna um processo posterior. Os dados de um *data lake* normalmente são extraídos para um *data warehouse* de consumo. Atualmente, os principais problemas com o qual *data lakes* sofrem são (ZAHARIA ET. AL., 2021):

- Confiabilidade: Integrar *data lakes* e *data warehouses* é uma tarefa difícil, cara, e requer esforço constante para realizar rotinas de extração – que são possíveis fontes de inconsistências de dados e bugs.
- Desatualização de dados: Caso a extração não ocorra com certa frequência, é comum ter dados disponíveis no *data lake* mas não no *data warehouse*.
- Suporte limitado para análises avançadas: Nenhum dos principais *frameworks* de aprendizado de máquina integra bem com soluções de *data warehouses* disponíveis. Os usuários podem tentar realizar esses processos utilizando os arquivos em formato aberto do *data lake*, mas o usuário perde as garantias de integridade do *data warehouse*.
- Custo total continuo: Além de pagar pelas extrações contínuas, usuários também pagam o armazenamento do dado copiado para o *data warehouse*, que normalmente está em um formato proprietário o que aumenta os custos de migração.

Data lakehouses são um novo tipo de sistema proposto para resolver o problema de gerenciamento de dados de *data lakes*, onde algumas soluções como *Delta Lake* e *Apache*

Iceberg propõem implementar a visão transacional de dados (normalmente feito no *data warehouse*) como transações ACID e versionamento de dados diretamente em cima dos arquivos do *data lake*, reduzindo a quantidade de etapas de extração. Com a evolução dos *frameworks* de aprendizado de máquina, também há possibilidade para realizar as análises avançadas diretamente em cima dos arquivos armazenados em formatos abertos. Bibliotecas de aprendizado de máquina como Tensorflow e MLlib podem ler formatos de *data lakes*, logo, podem facilmente ser integrados ao *lakehouse*. Há também necessidade dos *lakehouses* implementarem melhorias de performance em cima dos formatos abertos de arquivos, com uma camada de metadados para armazenar dados auxiliares sobre os arquivos para otimizar o desempenho de *queries* de SQL em cima destes arquivos (ZAHARIA ET. AL., 2021).

Percebe-se que com o passar do tempo, a forma de armazenar e organizar os dados mudou para lidar com os problemas de escalabilidade. Como consequência, a forma que os dados são utilizados para produzir conhecimento também muda.

Nota-se a existência de ambiguidade e confusão entre as definições dos termos referentes a análise de grandes bases de dados utilizados pela bibliografia existente. A descoberta de conhecimento é o conceito de mais alto nível, que engloba termos como *data analytics* (inteligência analítica ou análise de dados) entre outros, todos com o objetivo de descobrir ou produzir conhecimento novo a partir de dados. *Data analytics* é um campo interdisciplinar que tem o propósito de apoiar a tomada de decisões através da análise de dados existentes. Dentro de *data analytics* existem outras disciplinas como *big data analytics* ou *business intelligence* (BI). Quando falamos de *big data*, nos referimos a uma parte de *big data analytics*. Apesar de estarem em um mesmo nível hierárquico, O BI atua em cima de *Date Warehouses* e *Data Marts* (dados estruturados), *Big Data Analytics* pode trabalhar com *data warehouses*, mas também há alternativas com plataformas Hadoop (DEDIĆ & STANIER, 2017).

2.2 REGRAS DE ASSOCIAÇÃO

Neste tópico, revisa-se algumas definições básicas sobre as regras de associação, como elas surgiram, e um exemplo prático. Em seguida, define-se as métricas utilizadas neste estudo. Para finalizar, introduz-se algoritmos de regras de associação mais comuns.

2.2.1 Introdução e conceitos

De forma resumida, os principais métodos de mineração de dados são classificação, regressão, agrupamento, sumarização entre outros. A sumarização é definida como os métodos que envolvem encontrar uma descrição compacta para um subconjunto de dados. As regras de sumarização, mais conhecidas como regras de associação, são citadas como exemplo de aplicação da sumarização (FAYYAD, PIATETSKY-SHAPIRO E SMYTH, 1996).

O objetivo do uso de regras de associação é para encontrar os conjuntos de valores combinados mais comuns dentro de um conjunto de valores de uma base de dados. Uma regra pode ser enxergada como uma expressão do tipo: Se CONDIÇÃO, então RESULTADO. Em uma regra de associação, um item nunca pode estar no lado esquerdo e lado direito da regra (TUFFÉRY, 2011).

A análise de cesta de dados pode ser utilizada para introduzir os conceitos necessários para minerar regras de associação a partir de dados transacionais. Um conjunto de itens que mais é que um conjunto de atributos binários (ausente ou presente na transação). O conjunto de itens representa possíveis valores em uma transação. Uma transação pode ser definida como sendo um subconjunto dos itens. Uma regra é uma implicação do tipo $X \Rightarrow Y$ onde tanto X e Y são subconjuntos não vazios de itens. X é chamado de antecedente da regra ou *Left hand side* (LHS) que se traduz para lado esquerdo da regra. Y é chamado de consequente da regra ou *Right hand side* (RHS) que se traduz para lado direito da regra (HAHSLER, GRÜN E HORNIK, 2005).

Quadro 1 - Exemplo de um conjunto de transações de supermercado

ID de transação	Itens
1	Leite, Pão
2	Pão, Manteiga
3	Cerveja
4	Leite, Pão, Manteiga
5	Pão, Manteiga

Fonte: Adaptado De Hahsler, Grün E Hornik (2005).

O quadro 1 ilustra um exemplo de um conjunto de transações para o domínio de supermercados, cujo conjunto de itens é dado por $I = \{\text{Leite, Pão, Manteiga, Cerveja}\}$. Um exemplo para uma regra poderia ser $\{\text{Leite, pão}\} \Rightarrow \{\text{Manteiga}\}$. O significado desta regra é que se leite e pão são comprados, clientes compram manteiga junto. Para selecionar regras interessantes dentro de um conjunto possível de regras, restrições podem ser impostas sobre

as métricas das regras (HAHSLER, GRÜN E HORNIK, 2005). No próximo tópico, analisa-se as principais métricas utilizadas para mensurar a qualidade de regras de associação.

2.2.2 Métricas

O suporte pode ser calculado tanto para um conjunto de itens quanto para uma regra de associação, conforme apresentado na Equação (1). O suporte de um conjunto é definido como sendo a proporção das transações que contém o conjunto de itens (HAHSLER, GRÜN E HORNIK, 2005). O suporte de um subconjunto pode ser calculado dividindo a frequência (número de transações em que ele ocorre) do item nos dados transacionais pelo número de transações (HAHSLER, 2020b).

$$supp(X) = \frac{Frequência(X)}{Número\ de\ Transações} \quad (1)$$

Para uma regra, o suporte é calculado dividindo a frequência que o antecedente e consequente acontecem em conjunto pelo número de transações. O suporte pode ser interpretado como a proporção das transações que contém todos itens de uma regra (HAHSLER, GRÜN E HORNIK, 2005). Outra forma de interpretar o suporte de uma regra é como sendo a probabilidade (P) de uma transação selecionada aleatoriamente conter todos os itens da regra (HAHSLER, 2020b). Esta definição para o suporte de uma regra é representado na Equação (2).

$$supp(X \Rightarrow Y) = supp(X \cup Y) = P(X \cap Y) \quad (2)$$

A confiança de uma regra é dada pelo suporte da regra dividido pelo suporte do antecedente, conforme representado na Equação (3) (HAHSLER, GRÜN E HORNIK, 2005). A confiança é uma forma de calcular a proporção de vezes em que quando o antecedente de uma regra ocorre em uma transação, o consequente da regra também ocorre. A confiança é direcional, logo é diferente de (HAHSLER, 2020b).

$$conf(X \Rightarrow Y) = \frac{supp(X \Rightarrow Y)}{supp(X)} \quad (3)$$

O lift é conhecido como sendo uma métrica adicional para melhor filtrar as regras de associação geradas. Esta métrica é calculada através da divisão do suporte da regra pelo produto do suporte do antecedente com o suporte do consequente, conforme apresentado na Equação (4). Ela pode ser interpretada como uma medida de quantas vezes X e Y acontecem em conjunto comparado ao esperado caso eles fossem estatisticamente independentes (HAHSLER, GRÜN E HORNIK, 2005). A confiança não é direcional, logo ao trocar o antecedente com o consequente e calcular a confiança, obtém-se o mesmo resultado (HAHSLER, 2020b).

$$lift(X \Rightarrow Y) = lift(Y \Rightarrow X) = \frac{conf(X \Rightarrow Y)}{supp(Y)} = \frac{conf(Y \Rightarrow X)}{supp(X)} = \frac{P(X \cap Y)}{P(X)P(Y)} \quad (4)$$

2.2.3 Algoritmos Comuns

O problema de descobrir regras de associação pode ser dividido em dois subproblemas. Primeiro, deve-se encontrar o conjunto de itens que possuem um valor de suporte mínimo (minsup), chamados de grandes conjuntos de itens. Em seguida, os grandes conjuntos de itens podem ser utilizados para gerar o conjunto de regras desejado e o valor de confiança das regras podem ser calculados, preservando apenas as regras que possuem confiança superior que o valor de confiança mínimo (minconf). (AGRAWAL & SRIKANT, 1994).

O Apriori é um algoritmo de mineração de regras de associação proposto por Srikant e Agrawal em 1994. O algoritmo possui duas fases para encontrar regras de associação: Geração de candidatos e poda. Na fase de geração de candidatos, o algoritmo utiliza conhecimento prévio da última iteração, onde o conjunto de k itens é utilizado para explorar o conjunto de k + 1 itens. Isto se deve ao motivo que se um conjunto X possuí um valor de suporte mínimo, então todos os subconjuntos de X também possuem sete valor de suporte mínimo. Após gerar todos os k + 1 candidatos, uma passagem é feita sobre conjunto de transações para calcular o suporte destes novos candidatos. Então, na fase de poda, os conjuntos de itens k – 1 que não atendem o valor de suporte mínimo são eliminados (ROBU & DOS SANTOS, 2019).

O *Equivalence Class Transformation Algorithm* (ECLAT) é um algoritmo de mineração de padrões frequentes, capaz de minerar padrões frequentes de maneira eficiente,

através de uma busca em profundidade. Todas as transações que pertencem a um conjunto de itens são agrupados na mesma linha. Após realizar a intersecção dos conjuntos frequentes de k itens, o conjunto $k + 1$ é gerado. Este processo é repetido até não se encontrar mais conjuntos frequentes de itens. Uma vantagem deste algoritmo é que ele não precisa ler a base de dados inteira múltiplas vezes. Após ler a base de dados uma vez, os conjuntos $k + 1$ podem ser encontrado pela intersecção dos conjuntos k . Uma vantagem deste algoritmo é que ele possui tempo de acesso reduzido, entretanto, uma desvantagem é que o algoritmo não utiliza a confiança da regra como métrica de restrição de regras (ROBU & DOS SANTOS, 2019).

O algoritmo *FP-growth* utiliza uma abordagem de dividir para conquistar para extrair regras de associação. O algoritmo armazena dados de padrões frequentes comprimidos através de uma *Frequent Pattern Tree*. A principal vantagem deste algoritmo é que o conjunto de dados é lido apenas duas vezes. Ele é composto por dois passos (LIU ET. AL, 2021):

1. Ler a base de dados para criar uma lista FP ordenada (*frequent pattern list*) e contabilizar o suporte do conjunto de transações ordenando por suporte, removendo itens que não possuem suporte mínimo.
2. Construir uma árvore FP (*frequent pattern tree*) com raiz vazia, inserindo transações cada vez maiores, relendo a base de dados. No caso que o prefixo de uma transação já existe na árvore, um contador é incrementado para o nó.

Estes algoritmos possuem implementações em múltiplos *frameworks*, ferramentas e pacotes de desenvolvimento. Algoritmos de regras de associação tradicionais não são capazes de extrair regras em um tempo satisfatório ou de forma eficiente. Com base nisso, no restante deste tópico realiza-se uma breve análise sobre alguns exemplos de implementações para estes algoritmos voltados para o Apache Spark, um *framework* voltado para processamento de grandes volumes de dados.

O algoritmo Apriori é muito eficiente para encontrar conjuntos frequentes em grandes conjuntos de dados. Entretanto, o desempenho do algoritmo é influenciado pelo crescimento da dimensionalidade do conjunto de dados. O Apache Spark é capaz de processar grandes volumes de dados, que pode melhorar muito a velocidade da execução do algoritmo Apriori. Além de poder utilizar o algoritmo Apriori em Spark e também a implementação MRAApriori em Hadoop, que é capaz de executar aplicações até 100x mais rápido em memória, e 10x mais rápido em disco, uma versão melhorada do algoritmo também pode ser implementada para obter resultados ainda melhores (GAO, KHANDELWAL & LIU, 2019).

O uso de algoritmos de mineração de conjuntos de itens frequentes implementados em Hadoop MapReduce possuem operações custosas de I/O de disco. O Spark pode ser considerado um *framework* mais eficiente para implementar algoritmos tão iterativos através do uso de computação em memória e o uso de *resilient distributed datasets* (RDD). O Algoritmo ECLAT pode ser redesenhadado e implementado em Spark com múltiplas variantes, cada uma com abordagens e heurísticas diferentes. Estas variantes são capazes de produzir resultados que superam o YAFIM, uma implementação *Spark* do Apriori, principalmente com valores pequenos de suporte mínimo (SINGH ET. AL., 2019).

O algoritmo FP-Growth possui uma implementação com paralelismo implementado em Spark, chamada SSPFP. O algoritmo é capaz de lidar com dados de *streaming*, separando eles em *batches* (através de RDDs) e atualizando a árvore em memória dinamicamente. O primeiro passo do algoritmo gera listas FP em através dos RDDs, em seguida ele gera uma Árvore FP local baseado nas listas, depois ele gera uma árvore FP composta por todas as árvores locais, e por último gera regras de associação utilizando a árvore. O algoritmo é escala de acordo com a quantidade de nós disponíveis para o Spark, porém o crescimento não é linear, e a etapa de montar a árvore final consome uma quantidade de tempo elevada (LIU ET. AL, 2021).

2.3 TRABALHOS RELACIONADOS

Com a disponibilização de dados e a utilização de metodologias que permitem extrair informações, é possível realizar estudos quanto as suas aplicações para garantir a segurança no trânsito para a população. Através de uma revisão de literatura, destaca-se a importância da Mineração de Dados, que pode ser utilizada para identificar padrões que auxiliam na tomada de decisão de prevenção de acidentes, contribuindo para a redução dos gastos com a Saúde no Trânsito (GODOI E GUIMARÃES, 2014).

Utilizando Mineração de Dados, bases da Secretaria de Justiça e Segurança Pública e do Sistema Único de Saúde municipal de Cuiabá foram utilizadas para gerar conhecimento quanto à caracterização de vítimas de acidentes utilizando regras de associação (GALVÃO, 2009). No estudo supracitado, utilizou-se a ferramenta WEKA para aplicar o algoritmo Apriori, com o intuito de verificar regras e padrões de associação quanto ao perfil do acidentado, características do acidente e informações sobre atendimento médico seguindo um processo adaptado da Mineração de Dados. A aplicação do algoritmo Apriori possibilita gerar

regras de associação que indicam conhecimento compreensível quanto a caracterização das vítimas de acidentes em Cuiabá, sendo a Mineração de Dados uma poderosa ferramenta para análise de dados e também como ferramenta para auxiliar no processo de tomada de decisões.

Outro estudo utilizou dados abertos de boletins de ocorrências de rodovias federais brasileiras geradas pela Polícia Rodoviária Federal para a aplicação da Mineração de Dados com o propósito de extrair e analisar associações entre atributos além de discutir a viabilidade da aplicação deste processo (COSTA, BERNARDINI E FILHO, 2014). Os autores aplicaram técnicas de aprendizado supervisionado (J48 e PART) para fins de classificação de causas de acidentes com a ferramenta WEKA. Os modelos foram utilizados para extrair regras de decisão, que foram selecionadas buscando considerar regras que do ponto de vista dos autores poderiam ser utilizadas pelas autoridades e pela população para diminuir a quantidade de acidentes em rodovias federais. No estudo também empregou-se o algoritmo Apriori para gerar 38 regras de associação com confiança maior que 0,8 e duas regras com confiança maior que 0,9. O principal desafio encontrado neste trabalho, foi a disponibilização dos dados na forma que são coletados em sua fonte, algo que dificulta o processo de Mineração de Dados devido a baixa qualidade e padronização dos dados.

Dados da Polícia Rodoviária Federal, referente a acidentes na BR-381 no período de 2008 até 2012, foram utilizados para criar regras de associação em trechos críticos (REIS, SILVA E MAIA, 2015). Os trechos críticos foram levantados através do uso de equações propostas pelo Departamento Nacional de Infraestrutura de Transportes (DNIT). Os passos da metodologia CRISP-DM foram seguidos para a aplicação do algoritmo Apriori. Conjuntos de regras de associação foram analisados com base na separação do sexo do condutor (12 regras para cada), buscando regras com altos valores de confiança. As regras foram apresentadas através de uma tabela, com dados complementares como a gravidade e tipo de pista do acidente. A análise de resultados deu-se através de uma discussão e comparação sobre as diferentes regras encontradas em conjunto com os resultados de estudos anteriores.

3 METODOLOGIA

Neste capítulo será apresenta-se as etapas realizadas ao longo do desenvolvimento do trabalho, assim como as ferramentas, pacotes e técnicas utilizadas para manipular e processar os dados.

3.1 FERRAMENTAS UTILIZADAS

Neste tópico serão apresentadas as bibliotecas e ferramentas utilizadas para o desenvolvimento prático do projeto. Em cada tópico, introduz-se brevemente cada ferramenta. Em seguida, descreve-se como a ferramenta foi utilizada no estudo.

3.1.1 Ambiente de Desenvolvimento

O RStudio é um *Integrated Development Environment* (IDE), ou seja, um ambiente de desenvolvimento integrado para a linguagem de Programação R. O RStudio é um projeto de código aberto que combina componentes (como o console, editor de código, gráficos, histórico, documentação, controle de pacotes, etc.) do R em um único *workbench* (ALLAIRE, 2012). Esta ferramenta foi utilizada como ambiente de desenvolvimento para testar, executar e documentar todas as etapas deste estudo. As demais ferramentas listadas, que são pacotes de desenvolvimento que complementam a linguagem R, foram utilizadas e gerenciadas através deste ambiente. Para a realização deste estudo, utilizou-se a versão 1.2.5042 do RStudio.

3.1.2 Documentação

O Knitr é uma ferramenta com o propósito de geração dinâmica de documentos em R para produção de relatórios. O pacote permite que o usuário consiga documentar o código que foi executado em conjunto com a saída do terminal, além de eventuais mensagens, *warnings* e erros (XIE, 2022). A ferramenta Knitr foi utilizada para exportar a análise executada em conjunto com descrição dos diferentes trechos de código em um documento dinâmico. Com o uso deste pacote, disponibilizou-se os trechos documentados, visualizados e discutidos desde o primeiro passo do estudo. Para a realização deste estudo, utilizou-se a versão 1.28 do Knitr.

3.1.3 Processamento e Transformação de Dados

O projeto R é um ambiente de software gratuito para computação estatística e produção de gráficos. A linguagem R fornece uma grande variedade de técnicas estatísticas e gráficas, além de ser extremamente extensível. O conjunto de ferramentas fornecido facilita a manipulação e análise de dados, realização de cálculos, e visualização gráfica. Pacotes podem ser utilizados para estender as funcionalidades do R (THE R FOUNDATION, 2022). As ferramentas de manipulação de arquivos, funções de manipulação e transformação de dados, além de várias outras funcionalidades, utilizadas neste estudo, são funcionalidades básicas do R. A linguagem foi utilizada para implementar grande parte das manipulações de dados necessário para produzir a análise dos dados deste estudo. Deste a importação dos dados originais, até a visualização das regras de associação, utiliza-se funcionalidades da base que o projeto R fornece. Note que pacotes adicionais estendem as funcionalidades do R base. Para a realização deste estudo, utilizou-se a versão 3.6.2 do R.

O Dplyr é uma gramática de manipulação de dados, que fornece um conjunto consistente de verbos que auxiliam vários desafios de manipulação de dados como: Mutação – para criar transformações em cima de variáveis existentes; Seleção – Para selecionar subconjuntos de variáveis; Filtragem – para selecionar pontos de dados com valores específicos; Sumarização – Para resumir de múltiplos valores e Ordenação – para alterar a ordenação dos conjuntos de dados (WICKHAM ET. AL., 2022). Este pacote foi utilizado durante a realização de pré-processamentos da base de dados, principalmente para selecionar atributos específicos, além de filtrar linhas de dados baseados em valores específicos. Para a realização deste estudo, utilizou-se a versão 0.8.4 do Dplyr.

O pacote Magrittr oferece um conjunto de operadores que deixam códigos mais legíveis. O pacote implementa estruturas de sequências de operações de dados, organizando as transformações de esquerda para direita em vez de utilizar funções aninhadas. Isto minimiza a necessidade de utilizar definições de funções e variáveis locais. Além disto, o operador implementa torna fácil alterar o código quando necessário (BACHE & WICKHAM, 2022). Este pacote foi utilizado em algumas transformações de dados durante a realização do estudo. Isto foi feito para simplificar a legibilidade e manutenção do código. Além disto, operações que podem parecer muito complexas sem os operadores do Magrittr se tornam muito mais legíveis. Para a realização deste estudo, utilizou-se a versão 1.5 do Magrittr.

A estrutura de dados do tipo tempo hora pode ser difícil de manipular na linguagem R. O Lubridate facilita a manipulação de valores relacionados a tempo e hora. O pacote

implementa a conversão de datas e horas em diferentes formatos, extração de componentes como dia da semana ou hora, além de utilidades para lidar com fusos horários (GROLEMUND & WICKHAM, 2011). Este pacote foi utilizado durante o pré-processamento dos dados do tipo data e hora. Através dele, criou-se colunas novas como colunas de dia, mês, ano, etc. O uso do pacote possibilitou preparar estes dados para que eles pudessem ser utilizados para visualização e exploração dos dados, além do uso na criação de regras de associação. Para a realização deste estudo, utilizou-se a versão 1.7.4 do Lubridate.

3.1.4 Visualização de Dados

O ggplot2 é um sistema para criação de gráficos de forma declarativa. O usuário fornece ao pacote o conjunto de dados, quais variáveis devem ser utilizadas e como elas devem ser mapeadas para gerar o gráfico desejado (WICKHAM, 2016). O uso do pacote ggplot2 foi essencial para a análise exploratória do conjunto de dados. Através do uso dos diversos tipos de gráficos como gráficos de barras, gráficos de linhas, diagramas de caixa, etc., foi possível analisar o conjunto de valores dentro das variáveis e sua distribuição para a base de dados. Estes gráficos foram utilizados para obter um conhecimento prévio antes de analisar as regras de associação. Assim, foi possível levantar hipóteses sobre o conjunto de dados antes mesmo de produzir regras de associação, descartando ou confirmando as mesmas conforme observado na análise dos resultados de mineração. Para a realização deste estudo, utilizou-se a versão 3.3.0 do ggplot2.

A biblioteca gráfica Plotly, para R, possibilita criar gráficos interativos que atendem padrões de publicação. O Plotly R é um pacote livre de código aberto e seu código pode ser visualizado no repositório Github (PLOTLY, 2022). Para melhorar a qualidade e interatividade dos gráficos construídos com o ggplot2, o pacote Plotly para R foi utilizado para renderizar estes gráficos no documento criado com o Knitr. Gráficos estáticos ganharam opções de filtros e informação complementares com o passar do mouse. Isso melhorou o processo de análise tornando o mesmo mais interativo, e também gerou um resultado final satisfatório visualmente. Para a realização deste estudo, utilizou-se a versão 4.9.2 do Plotly.

O pacote sp fornece um conjunto de classes e métodos para manipular dados espaciais em R. As estruturas de dados implementadas incluem pontos, linhas, polígonos e malhas. O pacote fornece métodos e funções para criar estas estruturas a partir de matrizes e *data*

frames, além de métodos de coerção para realizar a transformação dos dados (PEBESMA & BIVAND, 2005). O pacote SP foi utilizado como complemento. Através dele, preparou-se os dados de coordenadas de acidentes para plotagem com o pacote Leaflet. Para a realização deste estudo, utilizou-se a versão 1.3-2 do sp.

O Leaflet é uma biblioteca construída com Javascript de código aberto para construção de mapas interativos. O pacote em R possibilita integrar e controlar mapas utilizando esta biblioteca. Algumas vantagens do Leaflet em R são a criação de mapas direto pelo console R, incorporação em documentos knitr ou R Markdown, e a integração com dados espaciais dos pacotes sp e sf (RSTUDIO, 2022a). Para explorar os dados de coordenadas disponibilizados na base de dados usou-se o pacote Leaflet. O Leaflet possibilita criar visualização como mapas de calor, mapas de distribuição e mapas coropléticos interativos. Apesar de não ser o foco principal do estudo, o pacote possibilitou analisar a distribuição dos acidentes pelo território nacional. Com este pacote, construiu-se diferentes tipos de mapas que foram disponibilizados para análise. Para a realização deste estudo, utilizou-se a versão 2.0.3 do Leaflet.

O pacote wordcloud2 fornece uma interface HTML5 para produzir nuvens de palavras com fins de visualização de dados. O pacote utiliza uma biblioteca Javascript para produzir os mapas de palavras (CRAN, 2018). A análise de dados textuais com grandes quantidades de valores distintos é um desafio. Para conseguir obter algum conhecimento, mesmo que mínimo, utilizou-se deste pacote para produzir um mapa de palavras utilizando o modelo de veículos acidentes. Para a realização deste estudo, utilizou-se a versão 2.6 do Wordcloud2.

O Pacote DT para a linguagem R é uma interface para a biblioteca de JavaScript de DataTables. Os tipos de dados, como matrizes ou *data frames*, podem ser visualizados em páginas que utilizam HTML para implementar ferramentas de ordenação, paginação e filtragem sobre os dados tabulares (RSTUDIO, 2022b). Esta biblioteca foi utilizada para disponibilizar os dados pré-processados e agregados em um formato tabular, que poderia ser facilmente buscado e filtrado pelo leitor da documentação. Além disto, as tabelas utilizadas ao longo deste trabalho foram derivadas das tabelas construídas com este pacote. Para a realização deste estudo, utilizou-se a versão 0.13 do DT.

3.1.5 Mineração de Regras de Associação

O pacote arules fornece uma infraestrutura básica para a criação e manipulação de conjuntos de dados voltados para a análise de conjuntos de itens e regras de associação. O pacote possuí interfaces para implementações rápidas em C de algoritmos populares como Apriori e ECLAT. Os algoritmos podem ser utilizados para minerar conjuntos de itens frequentes e regras de associação (HAHSLER, GRÜN E HORNIK, 2005). Uma vez que os dados haviam sido previamente explorados, tratados e filtrados, utilizou-se do pacote arules para produzir múltiplos conjuntos de regras de associação. Os algoritmos e parâmetros utilizados para a produção de regras de associação são destacados no tópico 3.2.7. Para a realização deste estudo, utilizou-se a versão 1.6-4 do arules.

3.1.6 Visualização de Regras de Associação

O pacote arulesViz implementa múltiplas técnicas para realização da exploração de regras de associação. Estas visualizações podem ser utilizadas para explorar um conjunto de regras de associação. O pacote fornece funções de visualização como gráficos de dispersão, gráficos de matriz agrupadas, grafos de regras, entre outros (HAHSLER & CHELLUBOINA, 2017). Este pacote foi utilizado para realização da exploração das regras de associação geradas. Além disto, algumas visualizações, como por exemplo o grafo, ajudam a encontrar partes de itens que são compartilhados entre múltiplas regras. Além disto, utilizou-se também do grafo de dispersão para demonstrar as métricas das regras geradas de forma interativa. Para a realização deste estudo, utilizou-se a versão 1.3.3 do arulesViz.

3.1.7 Setup Experimental

Todo o processo de desenvolvimento foi feito utilizando a máquina local do autor, no qual é que utilizou o seguinte hardware para executar códigos na linguagem R. As configurações são apresentadas na Tabela 1.

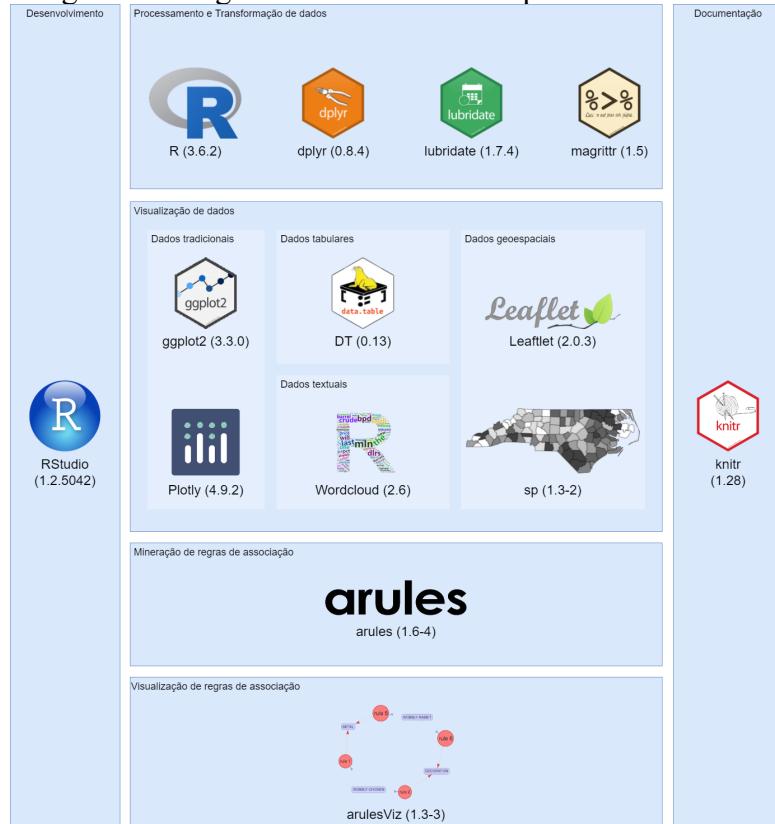
Quadro 2 -Setup Experimental

Processador	GPU	Memória	Sistema Operacional
I7-7700HQ	GTX 1050 TI	16 GB	Windows 10 x64 (build 18362)

Fonte: Elaborada pelo autor.

3.1.8 Diagrama de Uso De Ferramentas

Figura 4 - Diagrama de ferramentas e pacotes utilizados.



Fonte: Elaborado pelo autor.

3.2 EXPERIMENTOS

Neste tópico será abordado a forma com que os experimentos foram guiados, tendo como base os conceitos de análise técnica, coletados no referencial teórico e fazendo uso das bibliotecas e ferramentas citadas anteriormente. Os subtópicos foram organizados de acordo com as etapas destacadas pelo processo do KDD.

3.2.1 Análise de Domínio

O trabalho teve início com o levantamento bibliográfico dos trabalhos nacionais mais atuais envolvendo mineração de dados aplicado a bases de acidentes de trânsito. Esta análise foi utilizada para obter conhecimento prévio quanto ao domínio de estudo. O resultado deste

estudo encontra-se no tópico de Fundamentação Teórica deste trabalho. O estudo iniciou com uma análise da mineração de dados de forma genérica. Analisou-se o contexto histórico de mineração de dados e sua evolução. Após este estudo, com base em trabalhos anteriores, o tema de regras de associação foi aprofundado com o intuito de utilizar este tipo de técnicas no experimento.

3.2.2 Seleção de Conjunto de Dados

Atualmente, a Polícia Rodoviária Federal tem sob sua responsabilidade a segurança viária e a prevenção e repressão qualificada ao crime em mais de 71 mil quilômetros de rodovias e estradas federais em todos os estados brasileiros (POLÍCIA RODOVIÁRIA FEDERAL, 2021a). Seguindo as práticas de política de Dados Abertos governamental, a Polícia Rodoviária Federal, disponibiliza através do Portal de Dados Abertos, múltiplas bases de dados referentes a acidentes e infrações. Há três fontes de dados principais segundo o anuário de 2020: Dados de acidentes; Dados de fiscalização, infrações e criminalidade; Notas de revisão para contemplar possíveis retificações nos dados (POLÍCIA RODOVIÁRIA FEDERAL, 2021b).

Os dados de acidentes de trânsito são disponibilizados na forma: agrupado por pessoa com todos as causas e tipos, onde toda a sequência de motivos causadores do acidente são listadas; Agrupado por pessoa, onde todas as pessoas envolvidas no acidente e seus respectivos dados são listadas como registro, tendo múltiplos registros por acidente no caso de envolvimento de múltiplas pessoas mas com causa simples, onde a principal causa do acidente é destacada; Agrupado por ocorrência, onde há apenas um registro para cada acidente, com informações que resumem as características do acidente e de seus envolvidos. Além destas divisões, os dados também são separados por ano, com início no ano de 2007. Todas estas bases são disponibilizadas no formato CSV.

Os arquivos de dados obtidos através do site da Polícia Rodoviária Federal foram reunidos em uma pasta com intuito de organizar os arquivos do estudo. Utilizando a ferramenta RStudio, um documento R *Notebook* foi criado e salvo na mesma pasta. Utilizando este documento carregou-se os dados CSV para o ambiente de trabalho com o intuito de explorar os dados e implementar as etapas do KDD. Disponibilizou-se no Github o conjunto de dados resultante do processo de seleção (e processamento) de bases de dados.

3.2.3 Limpeza e Pré-processamento

Para início das atividades de pré-processamento, primeiro é necessário realizar uma análise exploratória inicial, buscando compreender o número de atributos, identificar quais deles possuem valores ausentes, quais são seus tipos de dados e quais são os valores mais comuns. Realizou-se a análise exploratória inicial com base em uma função descrita por Pearson (2018) com o objetivo de gerar um *data frame* que para cada coluna da base relata o seu nome; tipo de dado correspondente; quantidade de valores distintos; valor mais repetido, quantidade de vezes que aparece e percentagem que representa dentro do conjunto de valores; quantidade de valores ausentes e percentual dentro do conjunto de valores. Os resultados desta análise inicial, disponibilizados através da Tabela 1, auxiliaram a decidir como os seguintes passos deveriam ser realizados.

Tabela 1 - Resultados de análise exploratória inicial elaborado de acordo com Pearson (2018)

Atributo	Tipo	Valores Distintos	Moda	Quant. Moda	Quant. Moda (%)	Ausentes	Ausentes (%)
id	integer	89563	31281	75	0.0003669	0	0
pesid	integer	204395	1	1	4,89E-03	1	4,89E-03
data_inversa	factor	365	2017-12-23	1280	0.006262	0	0
dia_semana	factor	7	domingo	35393	0.1732	0	0
horario	factor	1358	18:00:00	3133	0.01533	0	0
uf	factor	27	MG	28820	141	0	0
br	integer	116	101	30702	0.1502	324	0.001585
km	factor	8533	1	730	0.003572	324	0.001585
municipio	factor	1835	CURITIBA	2782	0.01361	0	0
causa_acidente	factor	23	Falta de Atenção à Condução	81410	0.3983	0	0
tipo_acidente	factor	16	Colisão traseira	46203	226	0	0
classificacao_acidente	factor	3	Com Vítimas Feridas	131012	641	0	0
fase_dia	factor	4	Pleno dia	115251	0.5639	0	0
sentido_via	factor	3	Crescente	108519	0.5309	324	0.001585
condicao_meteorologica	factor	10	Céu Claro	111110	0.5436	0	0
tipo_pista	factor	3	Simples	110398	0.5401	0	0
tracado_via	factor	10	Reta	125779	0.6154	19442	0.09512

uso_solo	factor	2	Não	118851	0.5815	0	0
id_veiculo	integer	144985	74540	63	0.0003082	4	1,96E-02
tipo_veiculo	factor	25	Automóvel	96359	0.4714	20	9,79E-02
marca	factor	5783	NA	9184	0.04493	9184	0.04493
ano_fabricacao_veiculo	integer	67	2011	17660	0.0864	10068	0.04926
tipo_envolvido	factor	6	Condutor	144786	0.7084	1	4,89E-03
estado_fisico	factor	5	Ileso	103303	0.5054	10528	0.05151
idade	integer	144	NA	17213	0.08421	17213	0.08421
sexo	factor	4	Masculino	147150	0.7199	12063	0.05902
ilesos	integer	2	1	103303	0.5054	0	0
feridos_leves	integer	2	0	138793	679	0	0
feridos_graves	integer	2	0	185680	0.9084	0	0
mortos	integer	2	0	198148	0.9694	0	0
latitude	factor	75707	-12	203	0.0009932	0	0
longitude	factor	75385	-49,27348	174	0.0008513	0	0
regional	factor	27	SR-MG	28721	0.1405	0	0
delegacia	factor	173	DEL7/1	9151	0.04477	0	0
uop	factor	81	UOP01/MG	13628	0.06667	9812	0.04801

Fonte: Elaborado pelo autor.

Através do *data frame* gerado, foram identificados a existência de valores ausentes e também strings que eram utilizadas para preencher valores ausentes. A importação da base de dados para o ambiente foi ajustada para substituir estas strings por valores ausentes.

Com base no *data frame* criado, e com uma análise dos valores dos atributos, optou-se por tratar os atributos: “id”; “id_veiculo”; “pesid”; “estado_fisico”; “classificacao_acidente”; “uf”; “municipio”; “latitude” e “longitude”; “br”; “km”; “sentido_via”; “dia_semana”; “data_inversa”; “horario”; “fase_dia”; “condicao_metereologica”; “causa_acidente”; “tipo_acidente”; “ano_fabricacao_veiculo”; “marca”; “tipo_veiculo”; “tracado_via”; “tipo_pista”; “uso_solo”; “idade”; “sexo”; “tipo_envolvido”. Os demais atributos não analisados foram considerados de pouca importância para a análise através de regras de associação.

Como tratamento de pré-processamento para a análise de variáveis, utilizou-se:

- A filtragem ou substituição de valores ausentes.
- A detecção e filtragem de *outliers* quando cabível.

- Discretização de valores contínuos em intervalos quando necessário.

Elaborou-se funções de pré-processamento voltadas para cada coluna da base de dados que precisava ser tratada. A primeira substitui os valores ausentes por “Não Informado” para os atributos “estado_físico”; “br”; “sentido_via”; “ano_fabricação”; “marca”; “tipo_veiculo”; “tracado_via”; “tipo_envolvido”; “sexo” - onde o valor “Ignorado” também foi substituído.

Devido à ocorrência de ruídos nos valores de idades, criou-se uma função de tratamento. Separou-se as idades como valores possíveis da faixa de 0 até 107, os valores fora desta faixa foram tratados como ausentes. Casos entre 0 até 3 anos de idade com o tipo envolvido Condutor também foram tratados como ausente.

Outra função elaborada foi a de preparo de valores numéricos (conversão de vírgulas para pontos decimais), e divisão em intervalos numéricos, para os atributos “km” - intervalos de 10 quilômetros e “idade” - onde foi necessário criar o atributo “faixa_etaria”, com intervalos de 10 anos.

Segundo o manual de referência do pacote arules, o método “coerce” é utilizando quando um *data frame* é fornecido como conjunto de transações para transformar o *data frame* em transações, requerendo apenas que todas as colunas do *data frame* possuam dados categóricos ou lógicos (HAHSLER ET AL., 2022). Logo, foi necessário preparar as colunas de “data_inversa”; “data_dia”; “data_mes”; “data_ano” e “data_semana” já que essas eram as únicas colunas que não foram transformados em valores categóricos (*factors*). Uma função foi estabelecida para tratar a coluna de “data_inversa” onde foram criadas quatro novas colunas: “data_dia” - referente ao dia numérico do mês; “data_mes” - referente ao mês do ano; “data_ano” - referente ao ano do acidente e “data_semana” - referente a semana do mês, que foi calculada com base em divisão dos dias do mês em intervalos de 7 dias.

O último tipo de função de pré-processamento criada foi para reduzir o tamanho da strings do atributo “causa_acidente”. Outra função similar foi elaborada para a coluna “uso_solo” onde substitui-se os valores “Sim” e “Não” por “Rural” e “Urbano”, visando a interpretabilidade dos resultados.

Para identificar o sucesso dos processos de pré-processamento, criou-se diversos tipos de gráficos interativos como:

- Gráficos de barras;
- Gráficos de linhas;
- Gráficos de dispersão;
- Gráficos do tipo boxplot (diagrama de caixa);

- Mapas coropléticos;
- Mapas geográficos de calor;
- Mapas de calendário de calor.

Os resultados destas visualizações foram renderizados como figuras e apresentados no tópico 4.1. As visualizações também estão disponibilizados através do repositório do Github disponibilizado, em conjunto com todo o código para a sua elaboração.

3.2.4 Redução e Projeção de Dados

A base de dados foi preparada conforme sugerido por Hashler, Grün e Hornik (2005), sendo necessário realizar a redução de dados através da seleção de colunas. Criou-se *três data frames* distintos contendo uma seleção de colunas para os dados referentes aos três anos estudados. Cada *data frame* foi utilizado para analisar um atributo de interesse referente a pessoa acidentada.

Para a etapa de mineração, foram agrupados os dados dos três anos em um único *data frame*. Os atributos selecionados para gerar o *data frame* utilizado para encontrar as regras foram: "dia_semana"; "data_semana"; "data_mes"; "data_ano"; "uf"; "br"; "km"; "municipio"; "causa_acidente"; "tipo_acidente"; "fase_dia"; "condicao_metereologica"; "tipo_pista"; "tracado_via"; "uso_solo"; "tipo_veiculo"; "estado_fisico"; "faixa_etaria"; "sexo"; "tipo_envolvido". A mesma seleção foi usada para os três conjuntos de regras, porém, cada uma associada a uma variável diferente, para cada atributo possuir seu próprio *data frame*.

3.2.5 Seleção de Método de Mineração

Foi identificado uma grande quantidade de valores distintos para as colunas, logo, concluiu-se que o método de mineração mais adequado seria através de regras de associação. Além disto, a decisão sobre o método a ser utilizado foi tomada com base na bibliografia.

Vale destacar que nesta etapa também considerou-se o uso de árvores de decisão, mas em uma primeira tentativa utilizando o pacote RPart, a alternativa foi descartada devido ao baixo desempenho computacional com o volume de dados utilizado.

3.2.6 Análise Exploratória de Mineração

Conforme dito anteriormente, três subconjuntos de dados foram preparados para produzir e analisar regras de associação. Através destes conjuntos de dados, regras de associação foram geradas de forma genérica para encontrar atributos que poderiam ser melhor explorados. O resultado desta etapa de experimentação foi a seleção dos três atributos: “sexo”; “estado_fisico” e “tipo_envolvido”.

A seleção de uma confiança mínima foi feita após identificar que utilizar uma confiança mínima de 1 (ou seja, 100%) pode gerar resultados muitos específicos, requerendo ajustar o suporte para um valor muito baixo. Encontrou-se que seria melhor trabalhar com um valor mínimo de 0.8 (ou seja, 80%) de confiança.

Outro fator que influência a quantidade de regras geradas é o valor de suporte mínimo. Através deste parâmetro é possível ajustar a quantidade e o escopo das regras geradas. Para encontrar o valor ideal, realizou-se um processo iterativo de gerar regras e analisar os resultados através de gráficos. Encontrou-se que escolher um valor de suporte mínimo que gere em torno de 50 regras possibilita obter resultados interessantes que ainda sejam interpretáveis graficamente.

3.2.7 Mineração de Dados

O algoritmo Apriori foi selecionado, sendo este proposto originalmente por Agrawal e Srikant (1994). O pacote arules foi a implementação selecionada para produzir as regras de associação utilizando a linguagem R.

Para as três análises (“sexo”; “estado_fisico” e “tipo_envolvido”), adotou-se um valor de confiança mínima (confmin) de 0,8 com tamanho mínimo de regra de 5 e máximo de 10 e tempo máximo de 100 (segundos). Para cada atributo, o algoritmo foi executado várias vezes com reajuste de parâmetros produzindo conjuntos de regras diferentes. Através do ajuste de parâmetros, um conjunto de regras foi produzido para cada valor possível daquele atributo. Ajustou-se o suporte mínimo (supmin) de forma iterativa, buscando gerar um conjunto próximo de 50 regras. Após gerar as regras, elas são reordenadas de forma decrescente por suporte (com lift como critério de desempate) e filtradas para ter lift maior que 1,2.

Para cada conjunto de regras geradas, qualquer um dos atributos não sendo aquele do RHS são por padrão pertencentes ao LHS. Definiu-se o RHS como sendo a variável de interesse em conjunto com o valor, assim separando as regras para cada valor.

- Sexo: gerou-se 3 conjuntos de regras: “sexo=Masculino”; “sexo=Feminino” e sexo=“Não Informado”.
- Estado físico: Gerou-se 5 conjuntos de regras: “estado_fisico=Ileso”; “estado_fisico=Lesões Leves”; “estado_fisico=Lesões Graves”; “estado_fisico=Óbito” e “estado_fisico=Não informado”.
- Tipo Envolvido: Gerou-se 6 conjuntos de regras: “tipo_envolvido=Condutor”; “tipo_envolvido=Passageiro”; “tipo_envolvido=Pedestre”; “tipo_envolvido=Testemunha”; “tipo_envolvido=Calaveiro” e “tipo_envolvido=Não Informado”.

Para seleção das regras utilizadas, optou-se por ordenar as regras por suporte, seguido por lift. É possível fazer o contrário, mas desejou-se encontrar regras mais abrangentes.

3.2.8 Interpretação de Resultados

Para a interpretação dos conjuntos de regras geradas utilizando o pacote arules, o pacote arulesViz estende o pacote arules apresentando várias técnicas de visualização para regras de associações (HAHSLER, 2020a). Para interpretar as regras foram analisadas as métricas retornadas pelo pacote arules, grafos interativos, gráficos de dispersão interativos e tabelas interativas, todos gerados com o pacote arulesViz. Além disso, a métrica “Confiança Inversa Mínima” foi definida através da Equação 5, que foi calculada invertendo a ordem das regras para calcular a confiança. O objetivo desta métrica é medir o quanto abrangente uma regra é para transações com o mesmo consequente.

$$\text{Confiança Inversa Mínima} = \text{conf}(Y \Rightarrow X) = \frac{\text{supp}(Y \Rightarrow X)}{\text{supp}(Y)} = \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(Y)} \quad (5)$$

Os itens antecedentes foram selecionados para discussão contabilizando o número de vezes que o par atributo-valor ocorreram dentro do conjunto de regras, e selecionando os 6 atributos que mais ocorrem. Para agilizar o processo de contabilização manual, valores de

regra que ocorrem com frequência menor que 6 regras foram desconsiderados para a contagem. Em caso de empate de frequência, todos os valores com a mesma frequência foram selecionados.

Para analisar as regras, optou-se por utilizar os grafos de regras para visualizar as relações entre os conjuntos de regras gerados. Além disto, gerou-se diferentes gráficos que permitem conferir as métricas de cada regra e a qualidade das regras geradas.

4 RESULTADOS E DISCUSSÕES

Neste tópico, apresenta-se e discute-se os resultados obtidos através da aplicação do processo de descoberta de conhecimento na base de dados da Polícia Rodoviária Federal. Primeiro aborda-se os resultados da etapa de análise exploratória e pré-processamento de dados. Em seguida, apresenta-se as regras de associação geradas.

Alguns atributos não foram discutidos/apresentados devido à ocorrência de uma grande quantidade de valores únicos, algo que dificulta a apresentação e discussão em um formato textual. Além disto, atributos com poucos valores também foram removidos deste tópico para tornar a apresentação de resultados mais objetiva.

4.1 ANÁLISE EXPLORATÓRIA

Neste tópico, avalia-se individualmente os atributos da base de dados. Cada atributo foi analisado individualmente, sendo pré-processado conforme descrito no capítulo 3. Algumas observações são feitas a partir das visualizações construídas com o intuito de adquirir um conhecimento prévio para a análise de regras de associação.

4.1.1 Identificadores

Há três tipos de identificadores no conjunto de dados, cada um com uma funcionalidade:

- Identificador de acidentes: Este identificador é denotado pelo atributo “id”. A sua função consiste em diferenciar diferentes ocorrências de acidentes.
- Identificador de veículos: cada veículo recebe um identificador único denotado por “id_veiculo”. Desta forma, é possível diferenciar veículos envolvidos em um acidente.
- Identificador de pessoas: O terceiro e último identificador na base de dados é chamado de “pesid”. Ele é utilizado para identificar unicamente uma pessoa acidentada.

Utilizando a linguagem R, foram contabilizados a quantidade de acidentes, veículos acidentados e pessoas acidentadas ao longo dos anos. Para contabilizar acidentes, basta calcular a quantidade de valores distintos para o atributo “id”. O mesmo pode ser feito para veículos através do atributo “id_veiculo”. Para a quantidade de pessoas, basta contabilizar a

quantidade de linhas no *data frame*, já que cada linha representa uma pessoa. O resultado desta contagem é exposto na tabela 2

Tabela 2 - Contabilização de identificadores para a base de dados

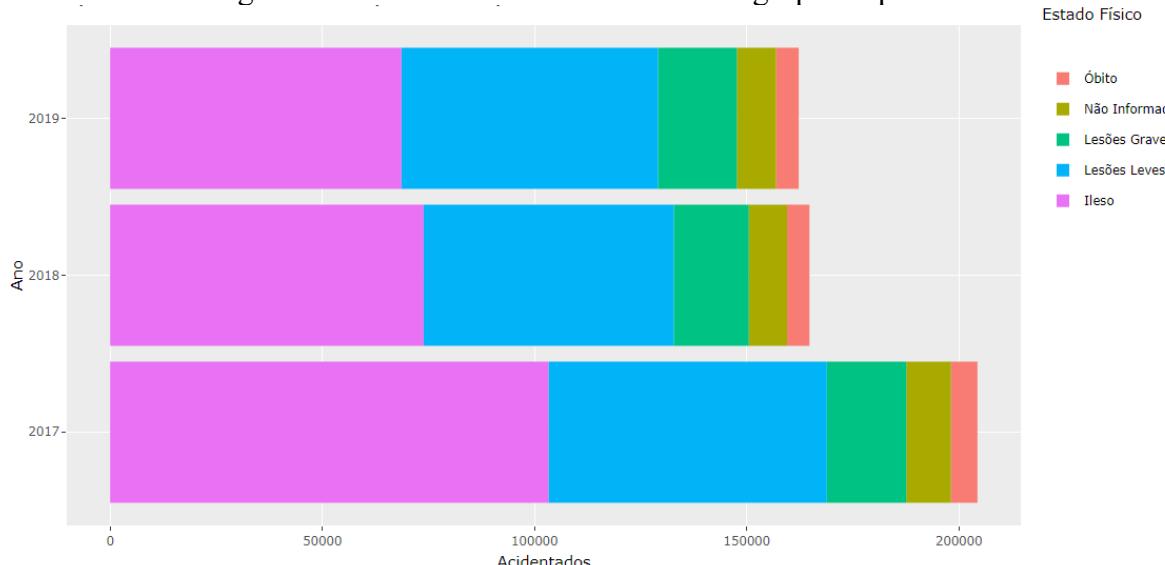
Ano	Acidentes	Veículos Acidentados	Pessoas Acidentadas
2017	89.563	114.981	204.395
2018	69.295	114.473	164.802
2019	67.446	112.051	162.273
Total	226.304	371.505	531.470

Fonte: Elaborado pelo autor.

4.1.2 Estado Físico

O atributo “estado_físico” contém o estado físico da pessoa após o acidente. Na figura 5, exibe-se a alteração de ocorrências para os diferentes valores de estado físico. Analisando a figura, é possível afirmar que não há quantidade significante de valores ausentes. Nota-se que os casos seguem uma ordem de quantidade (decrescente) para os valores: “Ileso”, “Lesões Leves”, “Lesões Graves”, “Não Informado” e “Óbito”. Percebe-se que a quantidade de casos com pessoas ilesas diminuiu com o passar dos anos. Enquanto isso, a quantidade dos outros estados físicos permaneceram com valores semelhantes.

Figura 5 - Estado Físico dos acidentados agrupados por ano.



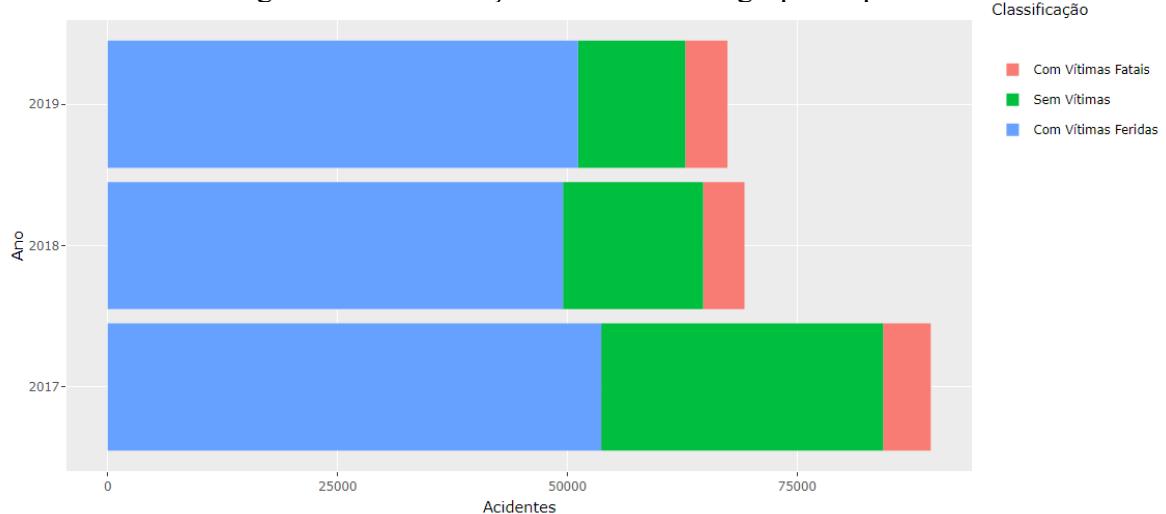
Fonte: Elaborado pelo autor.

4.1.3 Classificação de Acidentes

A classificação de um acidente é definida pelo estado físico mais grave das pessoas envolvidas no acidente. Acidentes sem vítimas são os que os estados físicos dos envolvidos são todos ilesos; Acidentes com vítimas feridas são acidentes que envolveram pessoas levemente ou gravemente feridas; e acidentes com vítimas fatais são acidentes que envolveram pessoas com que foram a óbito.

A figura 6 contabiliza a classificação de cada acidente registrado. Note que o número de observações no gráfico são pequenos, já que a granularidade do gráfico é por acidente. Ao analisar a figura, nota-se que a quantidade de pessoas em acidentes com classificação sem vítimas foi a que mais caiu com o passar dos anos. A quantidade de pessoas em acidentes com vítimas feridas também caiu em 2018, mas aumentou novamente em 2019. Já as pessoas em acidentes com vítimas fatais caiu, mas pouco, e depois aumentou.

Figura 6 - Classificação dos acidentes agrupados por ano.



Fonte: Elaborado pelo autor.

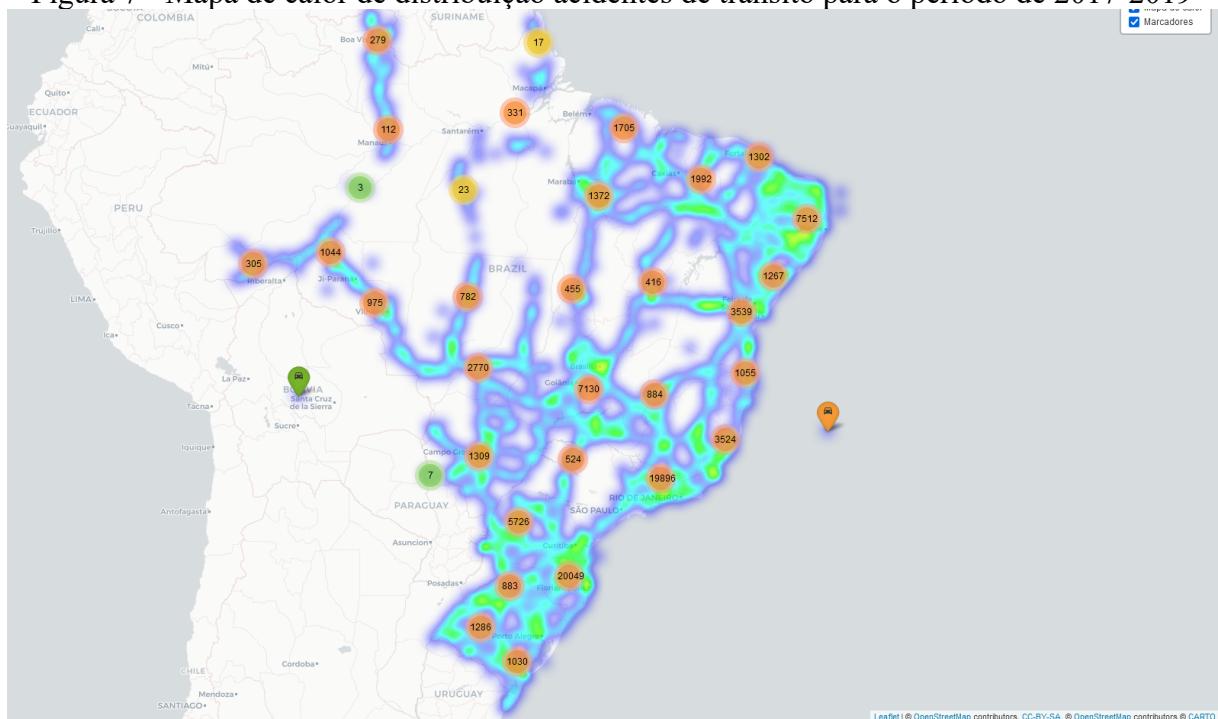
4.1.4 Distribuição Geográfica

Neste tópico analisa-se os atributos do conjunto que referenciam informações geográficas sobre o acontecimento dos acidentes. Os seguintes atributos foram utilizados para montar as análises: “uf”; “municipio”; “latitude”; “longitude”; “br”, “km” e “sentido_via”.

Uma análise foi construída a partir da listagem dos pontos de cada acidente. Utilizando o pacote Leaflet, construiu-se um mapa interativo onde os acidentes foram listados como

pontos em um mapa. Uma imagem do mapa pode ser vista através da figura 7. Ao se distanciar do mapa, os acidentes são agrupados em círculos com números que representam a quantidade de acidentes. Nota-se que a quantidade de acidentes no mapa é relativamente pequena. Isto ocorreu devido à presença de ruído nos dados. Isto pode ser observado na figura, com pontos de acidente fora do território nacional. Cada ponto de acidente é listado com uma cor de acordo com a gravidade do acidente, verde sendo sem vítimas, laranja com vítimas feridas e vermelho com vítimas fatais.

Figura 7 - Mapa de calor de distribuição acidentes de trânsito para o período de 2017-2019



Fonte: Elaborado pelo autor.

Para a análise do município de acidente, contabilizou-se o número de acidentes por município. Para os três anos, listou-se os cinco municípios com maior quantidade de acidentes totais. Os resultados foram disponibilizados na tabela 3.

Tabela 3 - Municípios com maior quantidade de acidentes por ano

Município	Número de acidentes por ano			
	2017	2018	2019	Total
CURITIBA	1.245	1.019	1.094	3.358
BRASILIA	1.093	867	1.090	3.050
SAO JOSE	949	874	819	2.642
GUARULHOS	865	755	717	2.337
PALHOCA	770	688	606	2.064

Fonte: Elaborado pelo autor.

Para a análise da rodovia federal do acidente (BR) contabilizou-se a quantidade de acidentes por rodovia. Para os três anos, listou-se as cinco BRs com maior quantidade de acidentes totais. Os resultados foram disponibilizados através da tabela 4.

Tabela 4 - Rodovias com maior quantidade de acidentes por ano

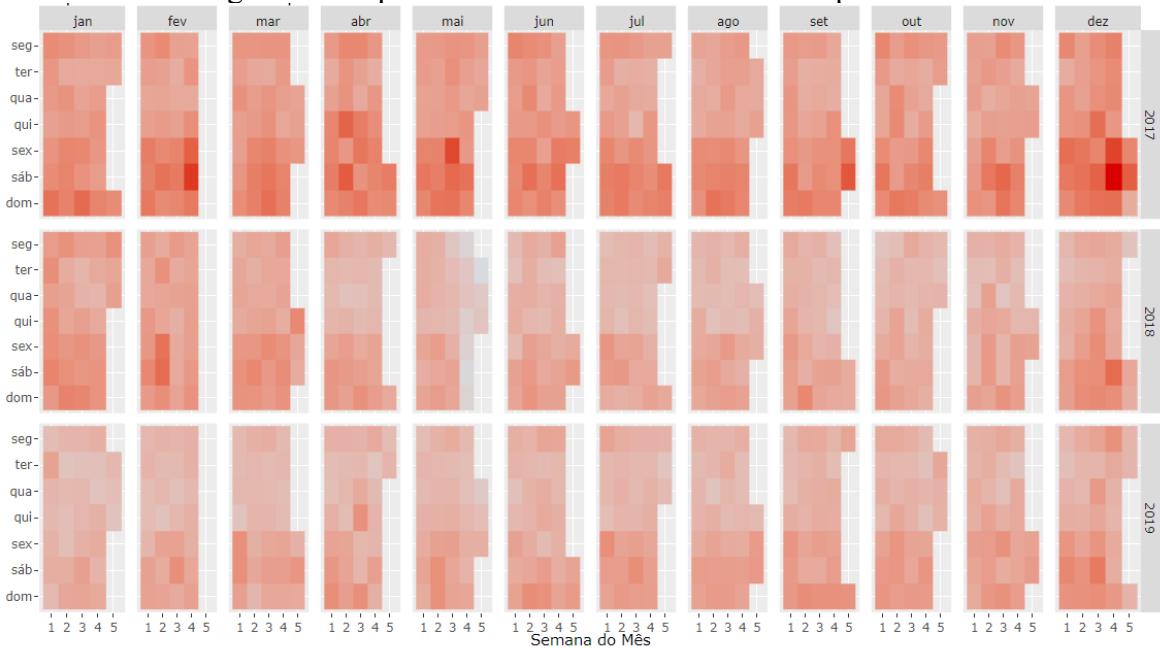
BR	Número de acidentes por ano			
	2017	2018	2019	Total
101	13.964	11.212	11.355	36.531
116	13.454	9.972	9.604	33.030
381	4.989	3.550	3.274	11.813
40	4.159	3.220	3.244	10.623
153	3.719	2.834	2.611	9.164

Fonte: Elaborado pelo autor.

4.1.5 Dados Temporais

Neste tópico, apresenta-se os resultados da análise das informações temporais referentes ao acidente.

Figura 8 - Mapa de calor de numero de acidentes por data



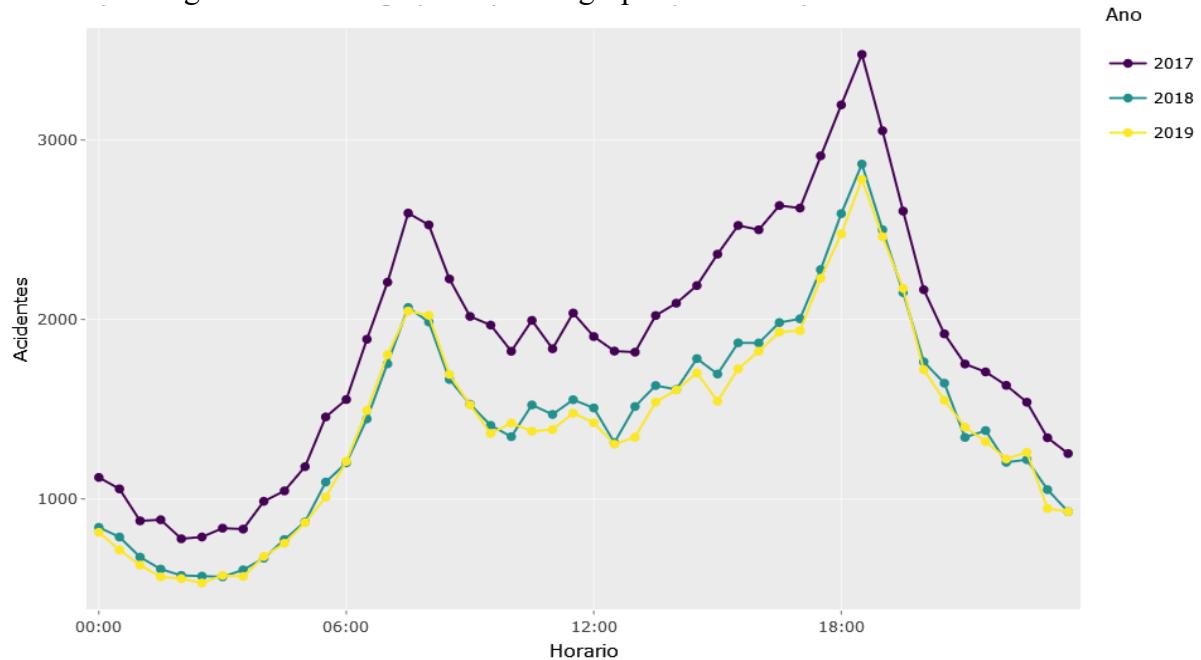
Fonte: Elaborado pelo autor.

A primeira análise foi realizada através do atributo de “data_inversa”, onde os dados foram formatados e utilizados para produzir um calendário de calor, que destaca a maior quantidade de acidentes para determinadas épocas do ano. O mapa de calor pode ser utilizado

para comparar os anos lado a lado. Analisando a figura 8, percebe-se que os acidentes aumentam em quantidade em finais de semanas e próximo de feriados.

Outra análise foi feita utilizando a variável “horario” do conjunto de dados. Com o agrupamento dos horários de acidentes em intervalos de 30 em 30 minutos, criou-se um gráfico para destacar como os acidentes são distribuídos ao longo do dia para os três anos. Os padrões para os três anos são relativamente semelhantes, como picos em início (07:30h) e fim (18:30h) de horário comercial.

Figura 9 - Horário de acidentes agrupados em intervalos de 30 minutos



Fonte: Elaborado pelo autor.

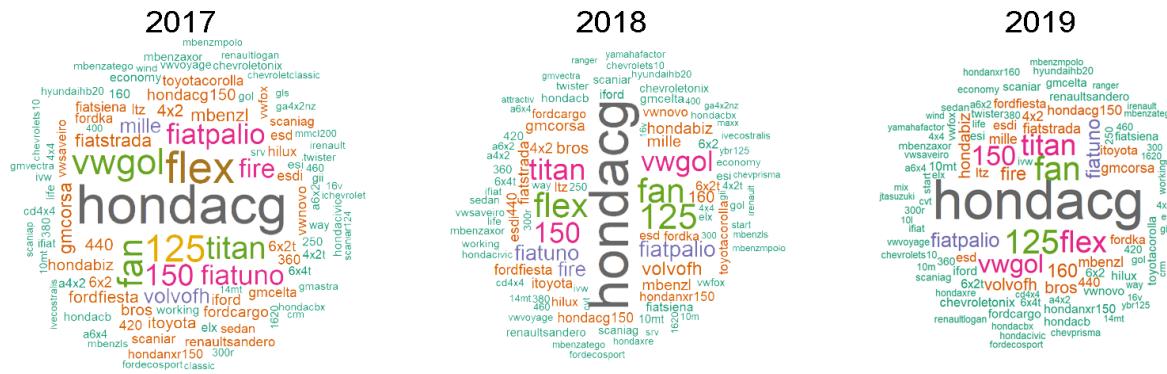
Vale destacar que o atributo “fase_dia” foi analisado, porém os resultados eram semelhantes a análise de horário. Há quatro valores para este atributo: Plena Noite, Pleno Dia, Anoitecer e Amanhecer. No geral os acidentes concentram-se em pleno dia e plena noite, com uma quantidade de extremamente pequena em torno de amanhecer e anoitecer.

Gráficos para a condição meteorológica também foram produzidos, mas devido a alta quantidade de valores distintos, optou-se por não apresentar o gráfico. Os principais valores para os acidentes são: Céu claro, Nublado, Chuva, Sol, Garoa/Chuvisco e Ignorando. Outros três valores que acontecem em quantidades desprezíveis são Nevoeiro/Neblina, Vento, Granizo e Neve. Os valores foram apresentados por ordem de maior quantidade de ocorrência.

4.1.6 Marca de Veículos

O atributo “marca” identifica o fabricante e modelo do veículo. Este atributo possui uma quantidade muito grande de valores distintos. Analisar um atributo com tantos valores distintos torna-se inviável para realizar com gráficos de barras, então optou-se por utilizar um *Wordcloud*. As 100 palavras mais repetidas dentro do conjunto de marcas/modelos de veículos por ano foram selecionadas. As palavras mais centralizadas e com tamanho de fonte maior possuem maior quantidade de repetições.

Figura 10 - Wordclouds de marca de veículo agrupados por ano



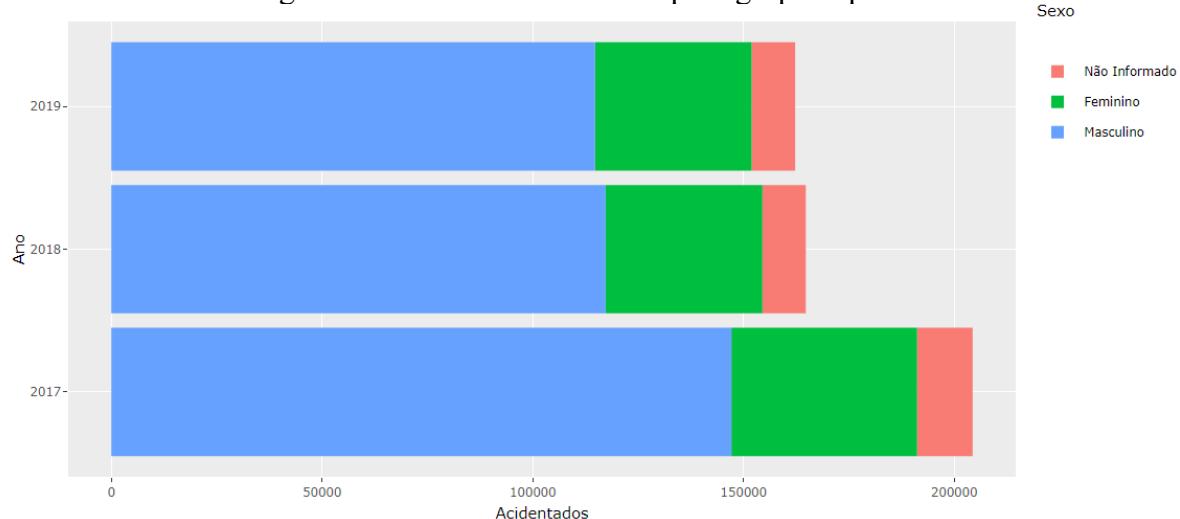
Fonte: Elaborado pelo autor.

Ao analisar a figura 10, a predominância é de veículos populares é evidente, onde a maioria das palavras que mais ocorrem podem ser relacionadas a modelo de motocicletas. Ao se distanciar do centro da figura, nota-se que as palavras associadas a carros populares começam a se destacar. Por último, nas bordas da figura, e portanto possuindo menor frequência, palavras relacionadas a carros não tão populares e caminhões surgem.

4.1.7 Informações de Acidentados

O atributo “sexo” informa o sexo do acidentado. A contabilização do sexo dos acidentados foi realizada para o conjunto de dados. Ao analisar a figura 11, nota-se uma predominância de envolvidos do sexo masculino independente do ano do acidente.

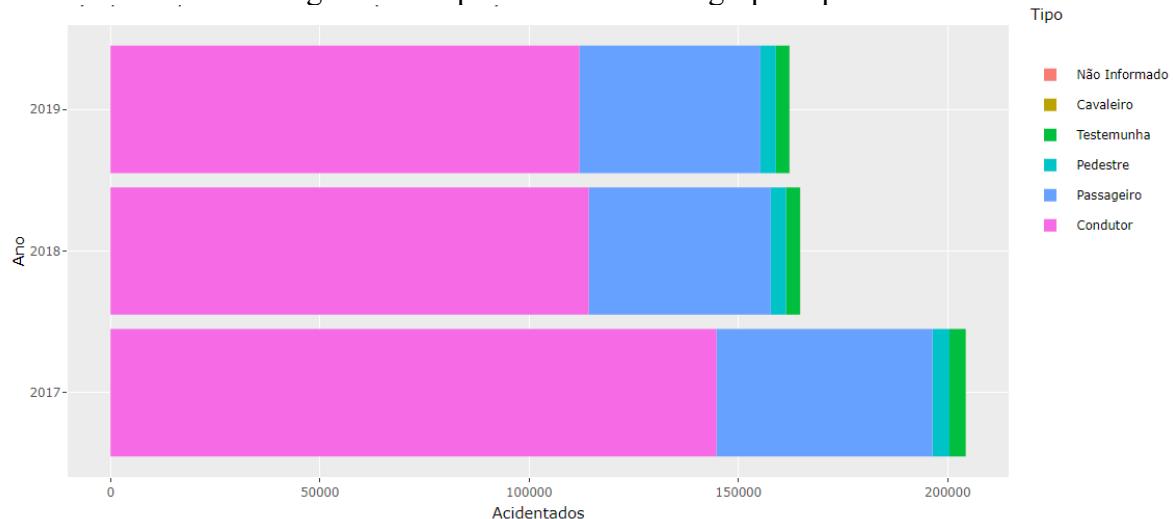
Figura 11 - Sexo de acidentados por agrupado por ano



Fonte: Elaborado pelo autor.

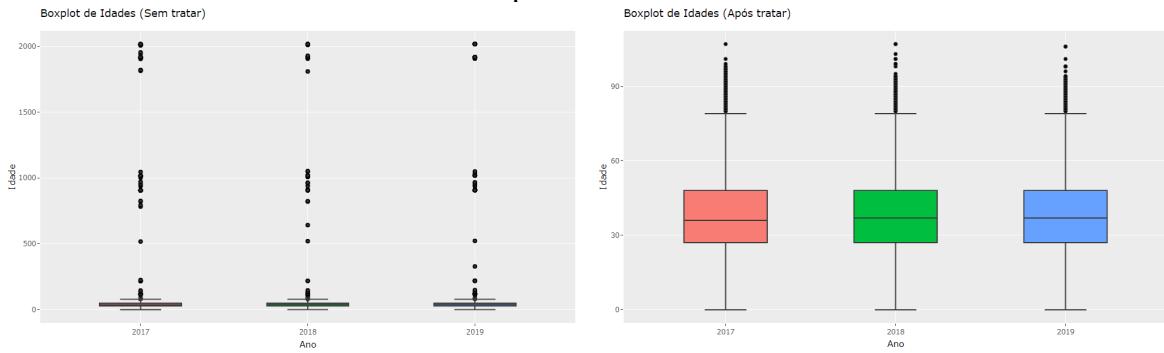
O atributo “tipo_envolvido” relata o envolvimento da pessoa acidentada. Um gráfico de barras foi elaborado para analisar a distribuição deste atributo. Pela figura 12, percebe-se que a maioria dos envolvidos são condutores, seguidos por passageiros. Pedestres e testemunhas ocorrem, porém em uma quantidade menor. Há uma quantidade desprezível de ocorrências com cavaleiros e não informados que não é visível na figura.

Figura 12 - Tipo de acidentados agrupado por ano



Fonte: Elaborado pelo autor.

Figura 13 - Comparação de diagramas de caixas de idade de acidentados com e sem pré-processamento



Fonte: Elaborado pelo autor.

Por último, apresenta-se o atributo de idade. Através da figura, nota-se a importância das atividades de pré-processamento para remoção de ruídos. Vale destacar o potencial de gerar regras tendenciosas caso usado o conjunto de idades não pré-processados. Para o conjunto tratado, as seguintes medidas de tendência central e distribuição foram obtidas para os anos a partir do gráfico de caixa da figura 13:

Tabela 5 - Medidas obtidas para idade pré-processada através do diagrama de caixas

Ano	Idade					
	Mínimo	1º Quartil	Mediana	3º Quartil	Limite Superior	Máximo
2017	0	27	36	48	79	107
2018	0	27	37	48	79	107
2019	0	27	37	48	79	106

Fonte: Elaborado pelo autor.

Nota-se que após pré-processar as idades, os resultados da comparação das métricas da tabela 5 são praticamente iguais para os três anos. Após o tratamento, as idades foram agrupadas em faixas etárias de 10 anos para uso na extração de regras de associação.

4.2 REGRAS DE ASSOCIAÇÃO

Neste tópico descreve-se as atividades realizadas para a etapa de mineração de dados. Após a geração do conjunto de regras, apresenta-se as métricas obtidas para o conjunto de regras gerados. Para interpretação dos resultados, aplica-se métodos de visualização de regras de associação para destacar itens comuns entre conjuntos de regras para discussão. Os resultados discutidos foram derivados a partir da exploração destes conjuntos de regras, através de uma análise visual combinada com uma inspeção manual. O conjunto de grafos

utilizados para analisar as regras geradas são disponibilizados no Apêndice A – Grafos de regras de Associação extraídas. O conjunto de regras completos foi disponibilizado através do repositório do Github.

4.2.1 Perfis de Acidentados Por Sexo

O atributo possui três valores possíveis para um acidentado: “Masculino”; “Feminino” e “Não Informado”. Logo, após a criação do *data frame* foram gerados três conjuntos de regras de associação para a análise de cada valor. As características de cada conjunto são destacadas através da tabela 6.

Tabela 6 - Métricas para as regras de associação para o atributo sexo

RHS	Suporte (RHS)	N. de Regras	Suporte Min.	Suporte Max.	Conf. Inv. Min.	Conf. Min.	Lift Médio	N. Itens Médio
“sexo=Masculino”	0,71	53	0,059	0,12226	0,082749	0,8	1,229	5,151
“sexo=Feminino”	0,22	55	0,00021	0,000636	0,000942	0,8	3,651	5,945
“sexo=Não Informado”	0,06	51	0,01	0,02174	0,179688	0,8	15,5	5,039

Fonte: Elaborado pelo autor.

Para o conjunto de regras com o consequente “sexo=Masculino” foram geradas 53 regras. Para obter esta quantidade de regras o valor de suporte foi ajustado para 0,059. Os seis itens distintos com maior quantidade entre as múltiplas regras geradas são: “tipo_envolvido=Condutor” (53 regras); “estado_fisico=Ileso” (36 regras); “uso_solo=Urbano” (26 regras); “tracado_via=Reta” (20 regras); “tipo_pista=Simples” (20 regras) e “condicao_metereologica=Céu Claro” (18 regras).

Através do valor de confiança inversa, observou-se que as regras para sexo masculino representam uma grande parte das instâncias de sexo masculino o que permite ter uma insuspeição maior referente as regras geradas. Ao relacionar os valores observados nas regras com os gráficos produzidos na análise exploratória, percebeu-se que os valores associados ao sexo masculino descrevem uma grande parte da base de dados, formando um perfil padrão das pessoas envolvidas em acidentes.

Para o segundo conjunto de regras, com o consequente “sexo=Feminino”, gerou-se 55 regras. Para obter esta quantidade de regras, o valor de suporte mínimo foi ajustado para 0,00021. Referente ao antecedente, os valores levantados para discussão foram: “tipo_envolvido=Passageiro” (55 regras); “estado_fisico=Lesões Leves” (29 regras);

“faixa_etaria=(50,60]” (18 regras); “tipo_veiculo=Motocicleta” (17 regras); “faixa_etaria=(40,50]” (16 regras); “uf=SC” e “tipo_veiculo=Automóvel” onde houve empate entre os dois últimos (15 regras).

Dado o baixo valor de suporte e confiança inversa, as regras produzidas podem ser consideradas de baixa qualidade. Foi observado que um dos valores é associado a um estado, logo, levantou-se a hipótese que possivelmente o perfil gerado pelos valores descreva um comportamento regional. Em comparação com as regras de sexo masculino os conjuntos de valores são extremamente diferentes, onde o perfil feminino parece estar associado as lesões, que possivelmente poderia ser explicado por uma predominância no uso veículos menos seguros (motocicletas, e apesar de não ter sido citado, observou-se a ocorrência de motonetas em múltiplas regras). De qualquer forma, apesar das regras para o sexo feminino terem sido geradas, com base nos valores observados nas métricas, conclui-se que as regras são de pouca relevância, representando apenas uma porção minúscula dos acidentados com sexo feminino.

Para gerar o último conjunto de regras, com o consequente “sexo=Não Informado”, gerou-se 51 regras. Para gerar esta quantidade de regras, o valor de suporte mínimo foi ajustado para 0,0115. Os valores levantados para a discussão foram: “faixa_etaria=Não Informado” (49 regras); “estado_fisico=Não Informado” (47 regras); “tracado_via=Reta” (19 regras); “tipo_envolvido=Condutor” (18 regras); “condicao_metereologica=Céu Claro” (15 regras); “tipo_veiculo=Automóvel” e “Uso_Solo=Rural”, onde houve empate entre os dois últimos (10 regras).

Foi notado que mesmo tendo uma quantidade de acidentados com sexo não informado menor do que o feminino, o suporte utilizado foi significativamente maior para as regras de não informado. Isso pode indicar que a qualidade das regras geradas para os valores não informados foi significativamente melhor. Através da confiança inversa, este conjunto de regras pode ser considerado o que mais abrange o valor dentro dos três gerados. Observou-se que valores ausentes em outros atributos são um forte indicador de sexo ausente. Os demais valores são comuns na base de dados, logo, este pode ser o motivo por observar estes valores no conjunto de regras.

4.2.2 Análise de Perfil por Estado Físico

Foi identificado que este atributo possui cinco valores possíveis para um acidentado: “Ileso”; “Lesões Leves”; “Lesões Graves”; “Óbito” e “Não Informado. Com isso, após o

preparo do *data frame* para criação das regras deste atributo, foram criados cinco conjuntos de regras de associação para a análise de cada valor. As métricas do conjunto de regras foram disponibilizados através da tabela 7.

Tabela 7 - Métricas para as regras de associação para o atributo estado físico

RHS	Suporte (RHS)	N. de Regras	Suporte Min.	Suporte Max.	Conf. Inv. Min.	Conf. Min.	Lift Médio	N. Itens Médio
“estado_fisico=Ileso”	0,463	46	0,02	0,05646	0,043197	0,8	1,776	6,022
“estado_fisico=Lesões Leves”	0,348	52	0,00145	0,002395	0,004167	0,8	2,33	5,42
“estado_fisico=Lesões Graves”	0,103	42	0,000036	0,00004892	0,00035	0,8	8,105	5
“estado_fisico=Óbito”	0,032	50	0,00002	0,00005080	0,000625	0,8	26,52	5
“estado_fisico=Não Informado”	0,054	46	0,01	0,02174	0,203704	0,8	15,75	5,087

Fonte: Elaborado pelo autor.

Inicialmente, foram geradas 46 regras para o consequente “estado_fisico=Ileso”. Para gerar as 46 regras, foi necessário ajustar o valor de suporte mínimo para 0,02. Os itens antecedentes destacados são: “tipo_veiculo=Automóvel” (42 regras); “tipo_envolvido=Condutor” (41 regras); “sexo=Masculino” (36 regras); “uso_solo=Rural” (26 regras); “tipo_acidente=Colisão traseira” (24 regras); “tracado_via=Reta” (17 regras).

Algo interessante é que os valores que foram observados se assemelham a aqueles para o sexo masculino. O suporte necessário em conjunto com a confiança inversa foi menor. É valido comentar que a colisão traseira estar associada a acidentes com pessoas ilesas faz sentido, uma vez que intuitivamente aparenta ser menos perigoso para os envolvidos. Outro fator que possivelmente explicaria os casos ilesos é o tipo de veículo observado, já que um automóvel é mais protegido do que motocicletas, motonetas ou nenhum veículo.

Para o consequente “estado_fisico=Lesões Leves” foram geradas 52 regras, utilizando um suporte mínimo de 0,00145. Os itens antecedentes levantados para discussão foram: “tipo_veiculo=Motocicleta” (45 regras); “tipo_pista=Dupla” (18 regras); “tracado_via=Reta” (17 regras); “sexo=Feminino” (17 regras); “sexo=Masculino” (15 regras); “tipo_acidente=Tombamento” (15 regras).

Da mesma forma que os casos com sexo masculino são maioria em regras de ilesos e vice-versa, o mesmo foi observado para o sexo feminino com lesões leves. As lesões parecem estar associadas a veículos que podem ser considerados menos seguros (motocicletas), onde o tombamento ou colisões causam mais danos físicos do que em comparação com um automóvel, e através das regras geradas uma hipótese que surge é que nestes veículos há

maior ocorrência de acidentados com sexo feminino. Apesar dos baixos valores de suporte e confiança inversa observados, os resultados gerados a partir deste conjunto de regras são interessantes.

Foram geradas apenas 42 regras para o consequente “estado_fisico=Lesões Graves”. O suporte mínimo utilizado foi 0,000036. Os valores destacados são: “km=(510,520]” (21 regras); “tipo_veiculo=Ônibus” (16 regras); “municipio=PARAISO DO TOCANTINS” (15 regras); “sexo=Feminino” (15 regras); “municipio=TORRES” (15 regras); “uf=TO” (12 regras).

Com os valores das métricas obtidos, a qualidade das regras pode ser considerada extremamente baixa. Os valores parecem indicar um padrão regional, gerado a partir de poucas ocorrências, ou uma única ocorrência com muitos envolvidos. Seria necessário inspecionar os acidentes na região para comprovar essa hipótese, que foge do escopo deste trabalho. Desejava-se encontrar atributos e valores que possivelmente poderiam ser agravantes na ocorrência de acidentes, mas com este conjunto de regras os resultados são inconclusivos.

Para o consequente “estado_fisico=Óbito” foram geradas 50 regras. Para gerar esta quantidade de regras o suporte mínimo foi ajustado para 0,00002. Os itens antecedentes destacados foram: “tipo_veiculo=Outros” (39 regras); “tipo_envolvido=Pedestre” (36 regras); “fase_dia=Plena Noite” (12 regras); “sexo=Masculino” (11 regras); “br=316” (10 regras); “uso_solo=Urbano” (9 regras).

Novamente, foi observado através das métricas que a qualidade das regras geradas é insatisfatória. Entretanto, observando os valores, pode ser que estas regras sejam válidas para um grupo de envolvidos com uma quantidade muito pequena de acidentes registrados, que no caso seriam os pedestres. Notou-se que o tipo de veículo outros não é definido no dicionário de dados, pode ser que ele seja um indicador para a ausência de veículo (no caso de pedestres). Caso isso seja verdade, então faria sentido pedestres sofrerem acidentes mais graves, já que não há proteção nenhuma contra atropelamentos. Outros valores observados como plena noite em conjunto com solo urbano contribuem para essa hipótese, entretanto, o baixo suporte e confiança inversa causa desconfiança.

O último conjunto de regras geradas tiveram o consequente “estado_fisico=Não Informado”. Utilizando o valor de suporte mínimo 0,011 foram geradas 46 regras. Os itens antecedentes levantados para discussão foram: “sexo=Não Informado” (46 regras); “faixa_etaria=Não Informado” (42 regras); “tracado_via=Reta” (20 regras);

“condicao_metereologica=Céu Claro” (16 regras); “tipo_pista=Simples” (11 regras); “tipo_veiculo=Automóvel” (9 regras).

Analisando as métricas, torna-se evidente que os resultados são promissores com altos valores de suporte, confiança inversa e contagem, quando comparado com os anteriores. Novamente, foi traçado um paralelo entre as análises de sexo não informado com estado físico não informado. Notou-se que novamente outros atributos ausentes indicam a ausência do valor próprio atributo. Também foi observado a repetição de 3 de 4 pares de atributo-valor na análise de não informado para sexo e estado físico.

4.2.3 Análise de Acidentes Por Tipo de Envolvido

Este atributo possui seis valores possíveis para um acidentado: “Condutor”; “Passageiro”; “Pedestre”; “Testemunha”; “Cavaleiro” e “Não Informado”. Gerou-se quatro conjuntos de regras de associação para a análise de cada valor. O motivo de utilizar apenas quatro dos seis valores é que os valores de “Não Informado” e “Cavaleiro” possuíam uma quantidade muito pequena de instâncias. As métricas para o conjunto de regras geradas foram disponibilizadas na tabela 8.

Tabela 8 - Métricas para as regras de associação para o atributo tipo envolvido

RHS	Suporte (RHS)	N. de Regras	Suporte Min.	Suporte Max.	Conf. Inv. Min.	Conf. Min.	Lift Médio	N. Itens Médio
“tipo_envolvido= Condutor”	0,698	54	0,07	0,12226	0,10029	0,8	1,264	5,019
“tipo_envolvido= Passageiro”	0,26	51	0,007	0,02049	0,02692	0,8	3,212	5,235
“tipo_envolvido= Pedestre”	0,022	49	0,0021	0,004237	0,09545	0,8	41,19	5,204
“tipo_envolvido= Testemunha”	0,197	52	0,00022	0,0004666	0,00112	0,8	41,69	5,692

Fonte: Elaborado pelo autor.

Iniciando com as regras com consequente “tipo_envolvido=Condutor”, gerou se 54 regras utilizando um suporte mínimo de 0,07. Para a discussão, se destacaram os seguintes itens antecedentes: “sexo=Masculino” (51 Regras); “estado_fisico=Ileso” (34 Regras); “tracado_via=Reta” (25 Regras); “fase_dia=Pleno dia” (23 Regras); “condicao_metereologica=Céu Claro” (21 Regras); “uso_solo=Rural” (18 Regras).

Foram obtidos valores altos para as métricas suporte mínimo e confiança inversa mínima. Mais uma vez os itens antecedentes destacados reforçaram os paralelos entre sexo

masculino, estado físico ilesos e tipo envolvido condutor. O valor de confiança inversa mínima similar para estas três análises gerou a hipótese que não há extrema importância na direcionalidade das regras. Aparentemente, há uma forte associação entre estes atributos para este conjunto de valores. Como o suporte dos itens antecedentes comentados foram observados como sendo os maiores dentro de cada atributo, acredita-se que estes valores estão associados pois são os casos mais comuns.

Para as regras com consequente “tipo_envolvido=Passageiro”, gerou-se 51 regras. Para gerar esta quantidade de regras o valor de suporte mínimo foi ajustado para 0,007. Os itens de antecedentes selecionados para discussão foram: “sexo=Feminino” (46 Regras); “uso_solo=Urbano” (40 Regras); “tipo_pista=Simples” (27 Regras); “estado_fisico=Lesões Leves” (24 Regras); “tipo_veiculo=Automóvel” (14 Regras); por último um empate entre “faixa_etaria=(10,20]” e “tipo_acidente=Colisão Frontal” (10 Regras).

Apesar dos valores obtidos para suporte e confiança inversa mínima não serem os melhores dentro dos conjuntos de regras geradas, os resultados são interessantes. Nota-se a ocorrência do sexo feminino na grande maioria das regras, e desta vez com um suporte muito maior. Acredita-se que isso ocorre devido à direcionalidade das regras, onde o sexo feminino geralmente indica que o indivíduo é passageiro, mas a recíproca ocorre em quantidade menor. Para verificar essa direcionalidade, seria necessário contabilizar os sexos dos envolvidos do tipo envolvido passageiro. A ocorrência da faixa etária de 10 à 20 anos como passageiro faz sentido, além da ocorrência de lesões leves em conjunto colisão frontal é interessante já que são valores menos comuns na base de dados.

Em seguida, com o consequente ajustado para “tipo_envolvido=Pedestre”, gerou-se 49 regras. O suporte mínimo utilizado foi 0,0021. Para a discussão, os seguintes itens antecedentes foram selecionados: “tipo_acidente=Atropelamento de Pedestre” (45 Regras); “estado_fisico=Lesões Graves” (33 Regras); “tracado_via=Reta” (25 Regras); “sexo=Masculino” (24 Regras); “causa_acidente=Falta de Atenção do Pedestre” (16 Regras); Com um empate entre “uso_solo=Rural” e “estado_fisico=Óbito” (14 Regras).

Apesar do suporte mínimo utilizado, a confiança inversa mínima pode ser considerada extremamente alta dado o baixo suporte para o RHS. Há um perfil para acidentes que envolvem pedestres, onde o tipo de acidente na maioria das vezes é considerado como atropelamento do mesmo. Novamente ocorreu uma diferença significativa na métrica de confiança inversa devido à direcionalidade das regras, onde aparentemente quando um pedestre é acidentado seu estado físico é associado a lesões graves, mas a recíproca ocorre em quantidade expressivamente menor. Note também que estes acidentes possuem uma gravidade

(de estado físico) que não é normalmente vista entre os outros, já que um pedestre não é protegido por nenhum veículo.

Por último, as regras com consequente “tipo_envolvido=Testemunha”, foram geradas 52 regras utilizando um suporte mínimo de 0,00022. Para a discussão, se destacaram os seguintes itens antecedentes: “estado_fisico=Não Informado” (42 Regras); “uf=MG” (33 Regras); “tipo_acidente=Colisão Frontal” (24 Regras); “sexo=Não Informado” (23 Regras); “fase_dia=Pleno Dia” (23 Regras); “tipo_veiculo=Motocicleta” (23 Regras).

Neste caso as métricas observadas são baixas, algo que pode ser explicado por falta de interesse em dados sobre testemunhas, ou talvez por não ter nenhum padrão de escolha de testemunhas. Algo que foi observado é a ocorrência de valores não informados no antecedente, que provavelmente se deve pela falta de interesse em informações sobre a testemunha, já que deseja se obter informações sobre o acidente. O dicionário de dados não explica o que configura uma testemunha, logo a análise se torna complicada e é meramente especulativa. Acredita-se que os demais valores são para complementar as regras.

5 CONSIDERAÇÕES FINAIS

Referente a documentação da base de dados, especificamente o dicionário de dados, há pouca informação disponível, o que torna trabalhoso a etapa pré-processamento. Outra etapa prejudicada pela falta de informação foi interpretação dos resultados gerados através da mineração de dados, onde é interessante esclarecer certos valores (Como “Outros” observado em tipo veículo, que foi associado a pedestres), já que permitiria realizar deduções mais claras através dos dados. Para leigos, também seria importante destacar como os elementos (acidentes, veículos e pessoas) se relacionam através do dicionário de dados.

Foram observados padrões recorrentes entre indivíduos, onde para regras produzidas com diferentes pares atributo-valor no consequente, valores específicos de certos atributos apareciam se repetir de forma não aleatória no antecedente. Destacando os principais resultados encontrados: Foi notado que indivíduos masculinos, condutores com estado físico ilesos, apareciam representar uma grande parte da base de dados; Identificou-se que indivíduos passageiros apareciam seguir uma tendência de serem do sexo feminino, além de sofrerem lesões (geralmente leves); Percebeu-se que na maioria dos casos sabe-se que há um valor ausente para uma coluna, os valores ausentes ocorrem para vários atributos do indivíduo, como estado físico, sexo, e faixa etária e também que os valores ausentes estão associados a testemunhas, provavelmente pois as características destes indivíduos não eram de interesse; Por último, foi percebido que a gravidade das lesões de um acidentado aparenta, na maioria dos casos, estar ligada ao tipo de veículo – provavelmente, sendo ligada ao tipo de proteção que a estrutura do veículo fornece ao acidentado (no caso de pedestres, nenhuma).

Galvão (2009) destaca seis regras geradas, todas com o consequente sendo a ocorrência de assistência médica, que não é relevante para este estudo. Costa, Bernardini e Filho (2014) apresentam resultados semelhantes aos encontrados para indivíduos ilesos, destacando a associação de indivíduos em automóveis, com faixa etária adulta, colisão traseira, em traçado de retas, ao estado físico ilesos utilizando 0,1 como supmin e 0,9 como confmin. Reis, Silva e Maia (2015) destacam que para acidentes sem vítimas, há um padrão de acidentes como pleno dia, sendo a causa de acidente a não observância da distância mínima de segurança. Para acidentes considerados mais graves (vítimas feridas e vítimas fatais), um padrão regional foi para a cidade de João Monlevade, além de condições meteorológicas ruins e vias com traçado de curva. Outros conjuntos de regras foram destacados e discutidos, mas o estudo destaca situações específicas, que não foram analisados neste trabalho.

6 TRABALHOS FUTUROS

Para futuros trabalhos, há possibilidade de integrar, disponibilizar e traduzir os dados de todos os anos da Polícia Rodoviária Federal, através de um repositório único, como um pacote para a linguagem R. Tornou-se evidente que grande parte do trabalho poderia ser auxiliado por uma ferramenta (construída com Dash¹ ou Shiny²), que realiza as tarefas básicas de limpeza e visualização de dados da Polícia Rodoviária Federal, já que a estrutura dos dados aparentemente não se altera com o passar dos anos.

A etapa de mineração de dados também pode ser melhorada, utilizando outras técnicas como algoritmos de classificação ou agrupamento, para gerar ainda mais conhecimento através do conjunto de dados. Melhorando a etapa de mineração, é possível criar um modelo que possa ser implementado para auxílio durante a tomada de decisões que buscam soluções para reduzir a quantidade e gravidade de acidentes.

Outra possibilidade é explorar frameworks alternativos para trabalhar com um volume de dados maior. Utilizar um framework como o Apache Spark pode ser uma solução para processar o conjunto de dados completo da Polícia Rodoviária Federal, desde 2007. Enquanto a análise de dos três anos focados neste estudo levava um tempo considerável para processar em uma única máquina, explorar ferramentas de processamento distribuído possibilita extrair conhecimento de um volume muito maior de dados.

1 O Dash é um framework que permite rapidamente construir aplicações web com Python, R e Julia.
2 O Shiny é um pacote R que permite construir aplicações web com R.

REFERÊNCIAS

- ABULATIF, L. I. **Processo De Integração De Dados: Um Modelo De Gestão Da Informação Para Múltiplas Bases De Dados De Acidentes De Trânsito No Brasil.** Serv. Saúde, Brasília, v. 27, n. 2, e2017160, 2018. Disponível em: <https://doi.org/10.5123/s1679-49742018000200018>.
- AGRAWAL, R.; SRIKANT, R. **Fast Algorithms For Mining Association Rules.** VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases, p. 487-499, 1994.
- ALLAIRE, J. J. **RStudio: Integrated Development Environment for R.** 2012. Disponível em: <https://www.r-project.org/conferences/useR-2011/abstracts/180111-allairejj.pdf>. Visitado em: 26 de fevereiro de 2022.
- AZEVEDO, A. I. R. L.; SANTOS, M. F. **KDD, SEMMA and CRISP-DM: a parallel overview.** Instituto Politécnico do Porto. Instituto Superior de Contabilidade e Administração do Porto. 2008. Disponível em: <https://recipp.ipp.pt/handle/10400.22/136>. Acesso em: 18 de fevereiro de 2021.
- BACHE, S.; WICKHAM, H. **Magrittr: A Forward-Pipe Operator for R.** 2022. Disponível em: <https://magrittr.tidyverse.org>. Visitado em: 03 de março de 2022.
- BRASIL. LEI N° 12.527. 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 21 de abril de 2021.
- BRASIL. Ministério da Infraestrutura. **Rodovias Federais.** Disponível em: <https://antigo.infraestrutura.gov.br/rodovias-brasileiras.html>. 2019a. Acesso em: 22 de janeiro de 2021.
- BRASIL. Ministério da Infraestrutura. **Transportes 2018.** Disponível em: http://canaldoservidor.infraestrutura.gov.br/images/2019/Documentos/Transportes_2018_-web.pdf. 2019b. Acesso em: 22 de janeiro de 2021.
- BRASIL. **Portal Brasileiro De Dados Abertos.** 2021. Disponível em: <https://dados.gov.br/>. Acesso em: 21 de abril de 2021.
- CARVALHO, C. H. R. **Mortes Por Acidentes De Transporte Terrestre No Brasil: Análise Dos Sistemas De Informação Do Ministério Da Saúde.** Instituto de Pesquisa Econômica Aplicada. 2016. Disponível em: <http://repositorio.ipea.gov.br/handle/11058/6869>.

- CHE, D.; SAFRAN, M.; PENG, Z. **From Big Data to Big Data Mining: Challenges, Issues, and Opportunities.** Lecture Notes in Computer Science, p. 1–15. 2013.
- COENEN, F. **Data mining: Past, present and future.** The Knowledge Engineering Review, v. 26(1), p. 25-29. 2011. Disponível em: <https://www.cambridge.org/core/journals/knowledge-engineering-review/article/abs/data-mining-past-present-and-future/EE2E494D98BCE76EBE3FE07897540C43#.Yh6gWMOhaO4.link>.
- COSTA, J. J.; BERNARDINI, F. C.; FILHO, J. V. 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, [S.l.], v. 3, n. 2, p. 139-157, dec. 2014. ISSN 2237-826X. Disponível em: <http://dx.doi.org/10.5380/atoz.v3i2.41346>.
- CRAN. **Wordcloud2 Introduction.** 2018. Disponível em: <https://cran.r-project.org/web/packages/wordcloud2/vignettes/wordcloud.html>. Visitado em: 27 de fevereiro de 2022.
- DEDIĆ N.; STANIER C. **Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery.** Innovations in Enterprise Information Systems Management and Engineering. ERP Future 2016. Lecture Notes in Business Information Processing, v. 285. 2017. Disponível em: https://doi.org/10.1007/978-3-319-58801-8_10.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining To Knowledge Discovery In Databases.** AI Magazine, v. 17, n. 3, p. 37-54, 1996.
- GALVÃO, N. D. **Aplicação Da Mineração De Dados Em Bancos Da Segurança E Saúde Pública Em Acidentes De Transporte.** Tese (Doutorado em Ciências) - Escola Paulista de Medicina, Programa de Pós-graduação em Enfermagem, Universidade Federal de São Paulo. São Paulo, p. 1-120. 2009.
- GAO, F.; KHANDELWAL, A.; LIU, J. **Mining Frequent Itemsets Using Improved Apriori on Spark.** Proceedings of the 3rd International Conference on Information System and Data Mining. p. 87–91. 2019. Disponível em: <https://doi.org/10.1145/3325917.3325925>.
- GODOI, R. S.; GUIMARÃES, M. P. **Mineração De Dados No Auxílio À Redução De Gastos Públicos Em Saúde No Trânsito.** 2014. Disponível em: https://www.academia.edu/39179898/Mineração_de_dados_no_auxílio_à_redução_de_gastos_públicos_em_saúde_no_trânsito.
- GROLEMUND, G.; WICKHAM, H. **Dates and Times Made Easy with lubridate.** Journal of Statistical Software, v. 40(3), p. 1-25. 2011. Disponível em: <https://www.jstatsoft.org/v40/i03/>.

- HAHSLER, M.; GRÜN, B.; HORNIK, K. Arules – A Computational Environment For Mining Association Rules And Frequent Item Sets. *Journal of Statistical Software*, v. 14, n. 15, p. 1-25, 2005.
- HAHSLER, M.; CHELLUBOINA, S. **Visualizing Association Rules: Introduction to the R-extension Package arulesViz.** 2017. Disponível em: <https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf>. Acesso em: 02 de março de 2022.
- HAHSLER, M. **A Probabilistic Comparison Of Commonly Used Interest Measures For Association Rules.** Disponível em: https://michael.hahsler.net/research/association_rules/measures.html. Acesso em: 19 de julho de 2020b.
- HAHSLER, M.; BUCHTA, C.; GRUEN, B.; HORNIK, K. **Package ‘arules’.** 2022. Disponível em: <https://cran.r-project.org/web/packages/arules/arules.pdf>. Visitado em: 02 de março de 2022.
- INMON, W. H. **The Data Warehouse And Data Mining.** *Communications of the ACM*, v. 39(11), p. 49–50. 1996.
- LIU, L.; WEN, J.; ZHENG, Z.; SU, H. **An improved approach for mining association rules in parallel using Spark Streaming.** *International Journal of Circuit Theory and Applications*, v. 49(4), p. 1028–1039. 2021. Disponível em: <https://doi.org/10.1002/cta.2935>.
- MASSAÚ, G. C.; ROSA, R. G. **Acidentes De Trânsito E Direito À Saúde: Prevenção De Vidas E Economia Pública.** *Revista De Direito Sanitário*, v. 17(2), p. 30-47. 2016.
- MESQUITA FILHO, M. **Acidentes De Trânsito: As Consequências Visíveis E Invisíveis À Saúde Da População.** *Revista Espaço Acadêmico*, v. 11(128), p. 148-157. 2012. Disponível em: <http://www.periodicos.uem.br/ojs/index.php/EspacoAcademico/article/view/13630>.
- MIOSLAVSKAYA, N.; TOLSTOY, A. **Big Data, Fast Data and Data Lake Concepts.** *Procedia Computer Science*, v. 88, p. 300-305. 2016. Disponível em: <https://doi.org/10.1016/j.procs.2016.07.439>.
- PEARSON, R. K. **Exploratory Data Analysis Using R.** Boca Raton: CRC Press. 2018.
- PEBESMA, E.; BIVAND, R. S. **Classes and Methods for Spatial Data: the sp Package.** 2005. Disponível em: https://cran.r-project.org/web/packages/sp/vignettes/intro_sp.pdf. Visitado em: 27 de fevereiro de 2022.
- PLOTLY. **Plotly R - Open Source Graphing Library.** 2022. Disponível em: <https://plotly.com/r/>. Visitado em: 27 de fevereiro de 2022.

POLÍCIA RODOVIÁRIA FEDERAL. **Dados Abertos – Acidentes.** 2020. Disponível em: <https://portal.prf.gov.br/dados-abertos-acidentes>. Acesso em: 18 de junho de 2020.

POLÍCIA RODOVIÁRIA FEDERAL. **Acesso a Informação – Institucional.** 2021a. Disponível em: <https://www.gov.br/prf/pt-br/acesso-a-informacao/institucional>. Acesso em: 30 de abril de 2021.

POLÍCIA RODOVIÁRIA FEDERAL. **Anuário 2020.** 2021b. Disponível em: <https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/anuario-2020.html>. Acesso em: 30 de abril de 2021.

REIS, C. V. R.; SILVA, J. T. M.; MAIA, L. C. G. O Uso Da Descoberta De Conhecimento Em Banco De Dados Nos Acidentes Da Br-381. XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB). 2015. Disponível em: <http://www.ufpb.br/evento/index.php/enancib2015/enancib2015/paper/view/3096>.

RIBEIRO, C. J. S.; ALMEIDA, R. F. Dados Abertos Governamentais (OPEN Government Data): Instrumento Para Exercício De Cidadania Pela Sociedade. XII Encontro Nacional de Pesquisa em Ciência da Informação (XII ENANCIB). 2011.

ROBU, V.; DOS SANTOS, V. D. Mining Frequent Patterns in Data Using Apriori and Eclat: A Comparison of the Algorithm Performance and Association Rule Generation. 6th International Conference on Systems and Informatics (ICSAI). p. 1478-1481. 2019. Disponível em: <https://ieeexplore.ieee.org/document/9010367>.

RSTUDIO. Leaflet for R – Introduction. 2022a. Disponível em: <https://rstudio.github.io/leaflet/>. Visitado em: 27 de fevereiro 2022.

RSTUDIO. DT: An R interface to the DataTables library. 2022b. Disponível em: <https://rstudio.github.io/DT/>. Visitado em: 27/02/2022.

SAS INSTITUTE. Introduction to SEMMA. 2017. Disponível em: <https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jn8bbjjm1a2.htm&docsetVersion=14.3&locale=en>. Acesso em: 19 de fevereiro de 2021.

SINGH, P.; SINGH, S.; MISHRA, P. K., GARG, R. RDD-Eclat: Approaches to Parallelize Eclat Algorithm on Spark RDD Framework. Lecture Notes on Data Engineering and Communications Technologies. p. 775-768. 2020. Disponível em: http://dx.doi.org/10.1007/978-3-030-37051-0_85.

SMYTH, P.; PREGIBON, D.; FALOUTSOS, P. Data-Driven Evolution of Data Mining Algorithms. 2002. Communications of the ACM, 45, 33-37.

- SOWMYA, R.; SUNEETHA, K. R. **Data Mining with Big Data.** 11th International Conference on Intelligent Systems and Control (ISCO). p. 246-250. 2017.
- THE R FOUNDATION. **What is R?** 2022. Disponível em: <https://www.r-project.org/about.html>. Visitado em: 02 de março de 2022.
- THURAISINGHAM, B. **A Primer For Understanding And Applying Data Mining.** IT Professional, v. 2(1), p. 28–31. 2000.
- TUFFÉRY, S. **Data Mining and Statistics for Decision Making.** Wiley. 2011. DOI 10.1002/9780470979174. Disponível em: <http://dx.doi.org/10.1002/9780470979174>.
- WANG, L. **Data Mining, Machine Learning and Big Data Analytics.** International Transaction of Electrical and Computer Engineers System. v. 4, n. 2, p. 55-61. 2017, Disponível em: <http://pubs.sciepub.com/iteces/4/2/2>.
- WICKHAM, H.; FRANÇOIS, R.; HENRY, L.; MÜLLER K. **Dplyr: A Grammar of Data Manipulation.** 2022. Disponível em: <https://dplyr.tidyverse.org>. Acesso em: 27 de fevereiro de 2022.
- WICKHAM H. **Ggplot2: Elegant Graphics for Data Analysis.** Springer-Verlag New York. de 2016. Disponível em: <https://ggplot2.tidyverse.org/>. Acesso em: 27 de fevereiro de 2022.
- WIRTH, R.; HIPP, J. **CRISP-DM: Towards a Standard Process Model for Data Mining.** Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, p. 29-39. 2000.
- XIE, Y. **knitr: A General-Purpose Package For Dynamic Report Generation in R.** 2022. Disponível em: <https://yihui.org/knitr/>. Acesso em: 02 de março de 2022.
- ZAHARIA, M.; GHODSI, A.; XIN, R. ARMBRUST, M. **Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics.** 11th Conference on Innovative Data Systems Research, CIDR 2021. Disponível em: http://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf.

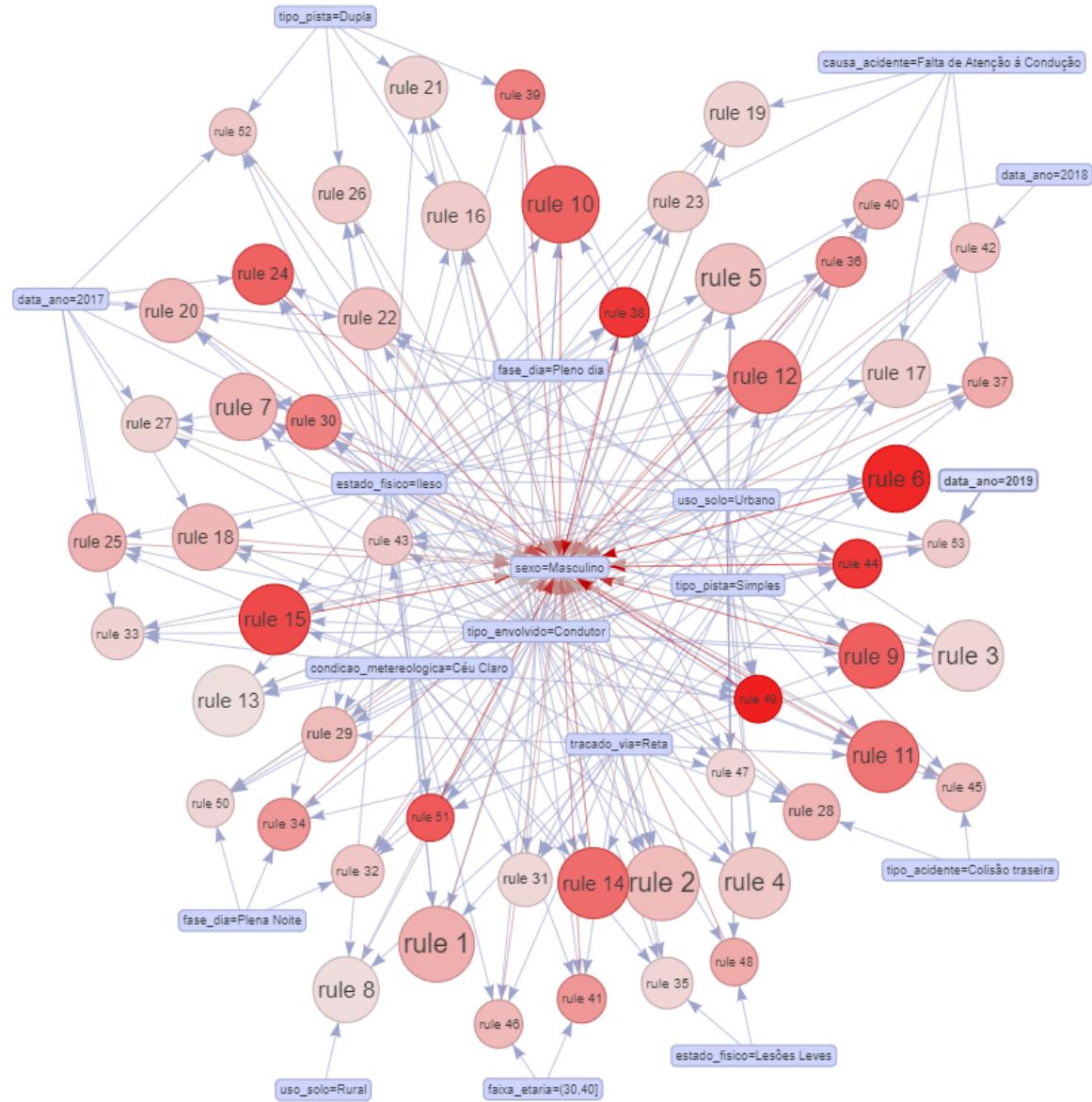
APÊNDICE A – GRAFOS DE REGRAS DE ASSOCIAÇÃO EXTRAÍDAS

Conforme citado no trabalho, a exploração das regras de associação foi realizada com meios visuais. O motivo por realizar a análise desta forma foi com o objetivo de encontrar padrões que se repetiam entre múltiplas regras.

Para a interpretação das regras de associação, utilizou-se da visualização de regra associação por grafo. Estes grafos são úteis para conseguir analisar um conjunto de regras pequenas. A interatividade dos grafos permite analisar um conjunto maior de itens. Os grafos representam regras através de vértices e utilizam arestas para associar estas regras a itens. As métricas de uma regra são identificadas pelo tamanho do vértice (suporte) e cor (lift) (HAHSLER & CHELLUBOINA, 2017).

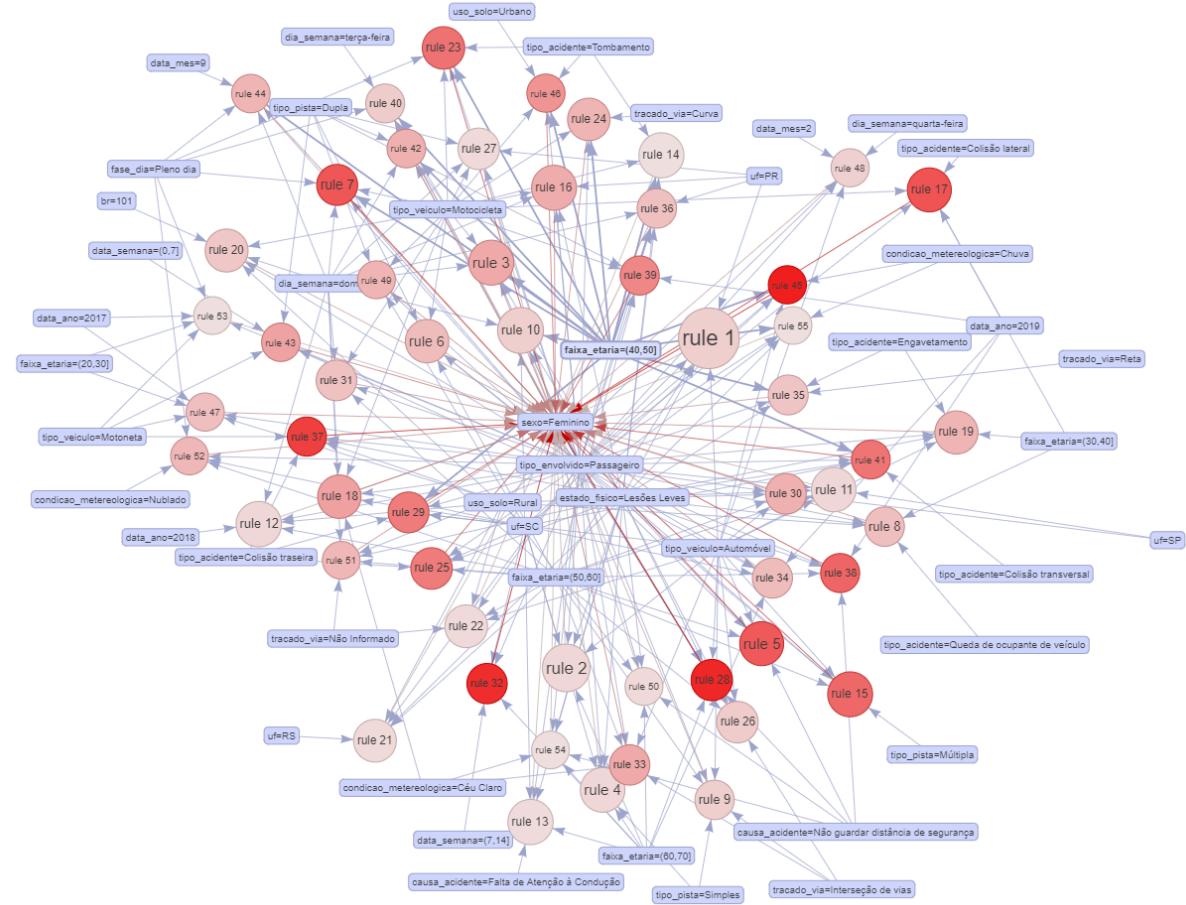
A.1 GRAFOS PARA O ATRIBUTO SEXO

Figura 14 - Grafo para o conjunto de regras extraídas para "sexo=Masculino"



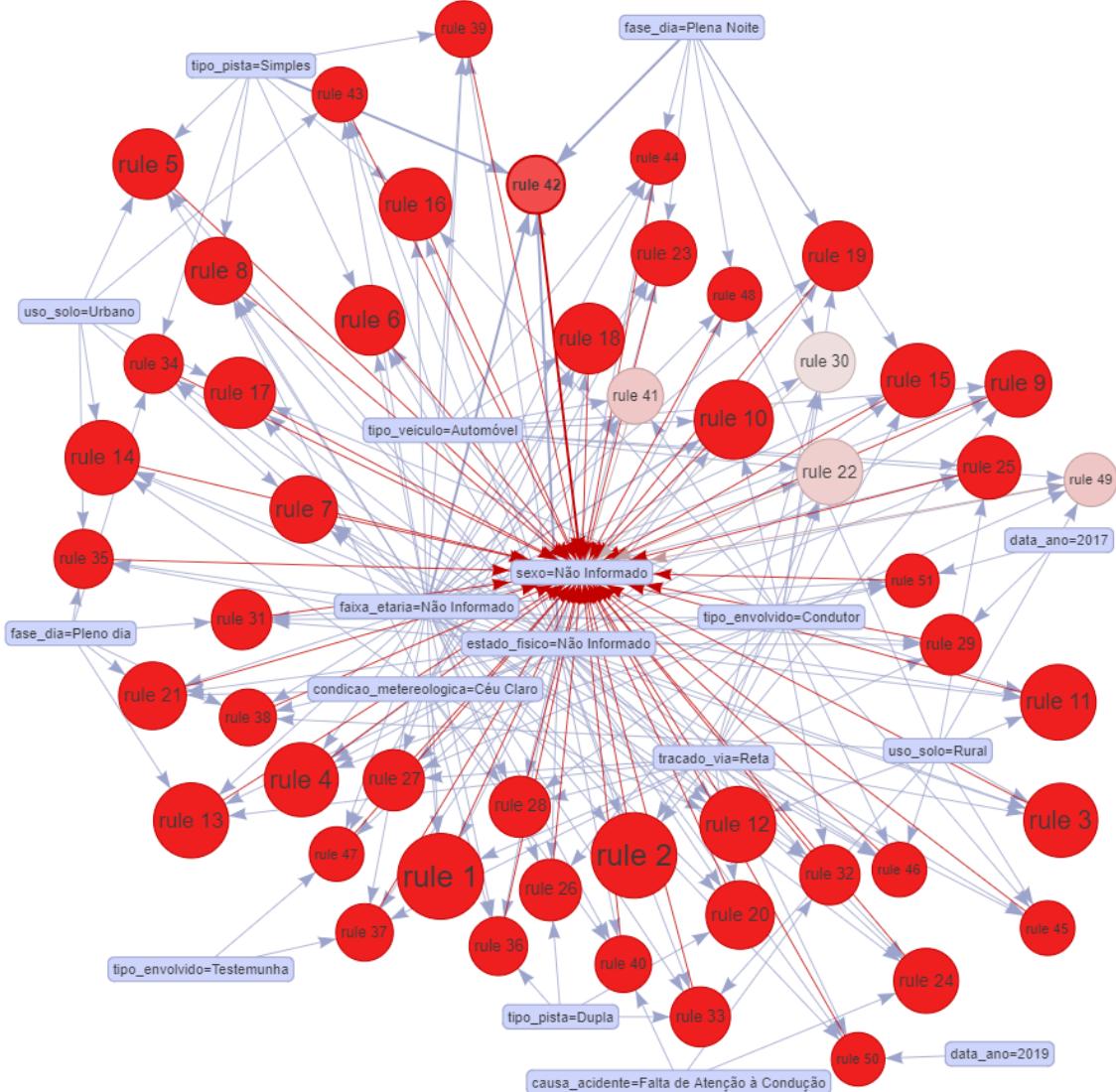
Fonte: Elaborado pelo autor.

Figura 15 - Grafo para o conjunto de regras extraídas para "sexo=Feminino"



Fonte: Elaborado pelo autor.

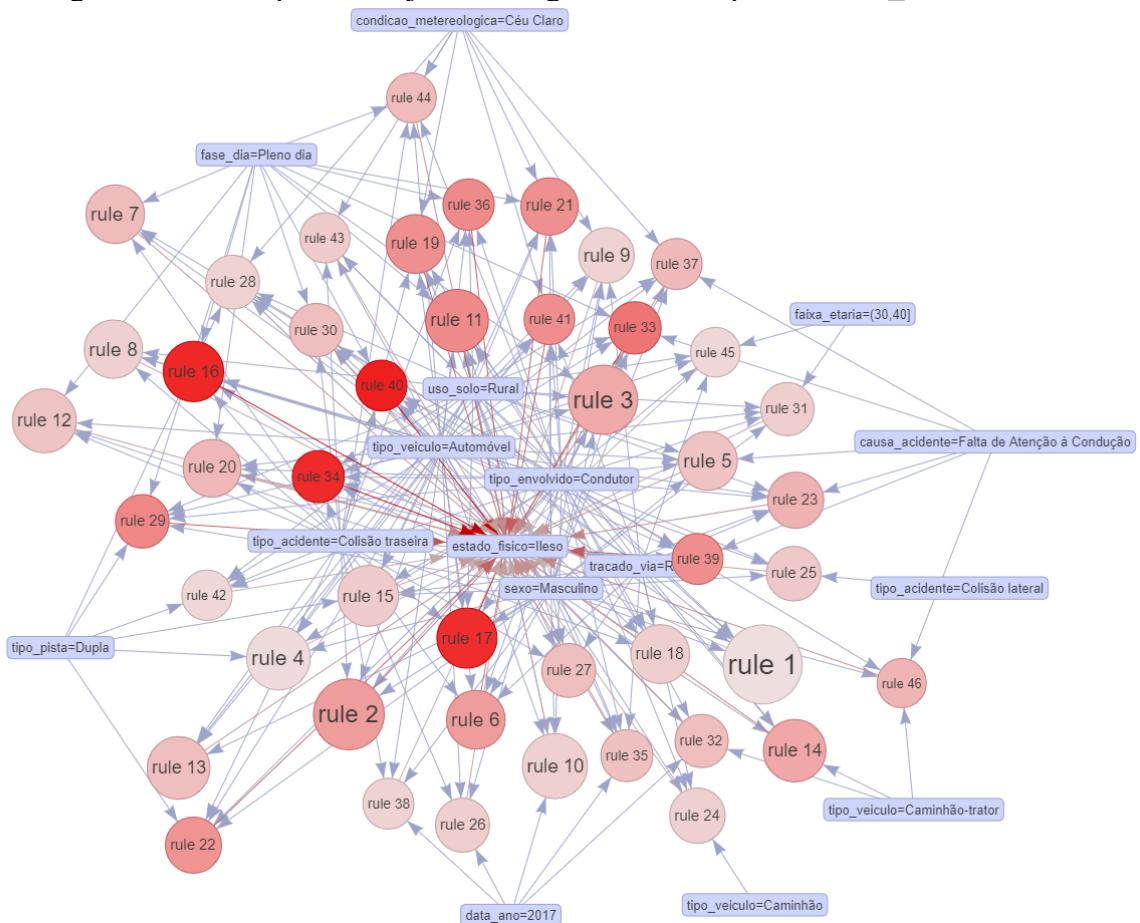
Figura 16 - Grafo para o conjunto de regras extraídas para "sexo=Não Informado"



Fonte: Elaborado pelo autor.

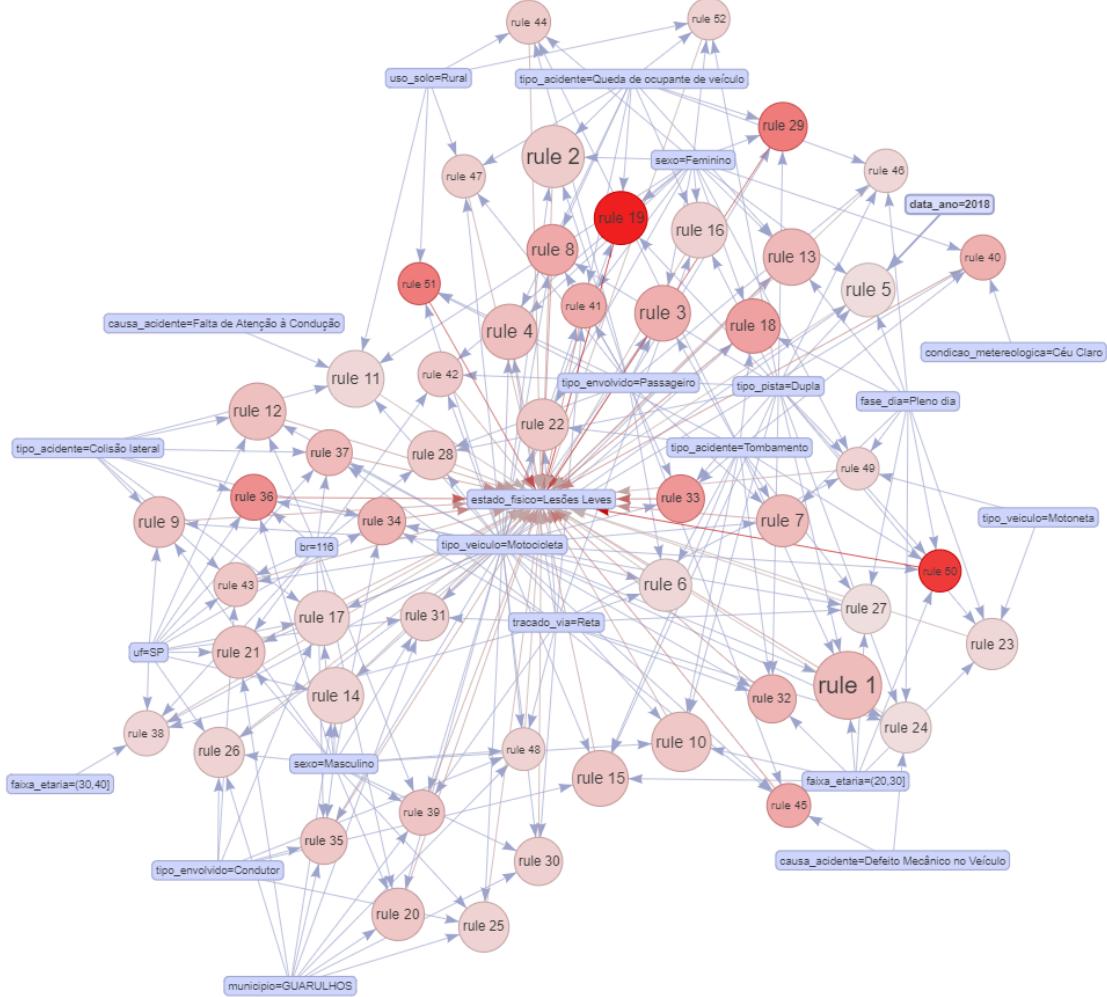
A.2 GRAFOS PARA O ATRIBUTO ESTADO FÍSICO

Figura 17 - Grafo para o conjunto de regras extraídas para "estado_fisico=Ileso"



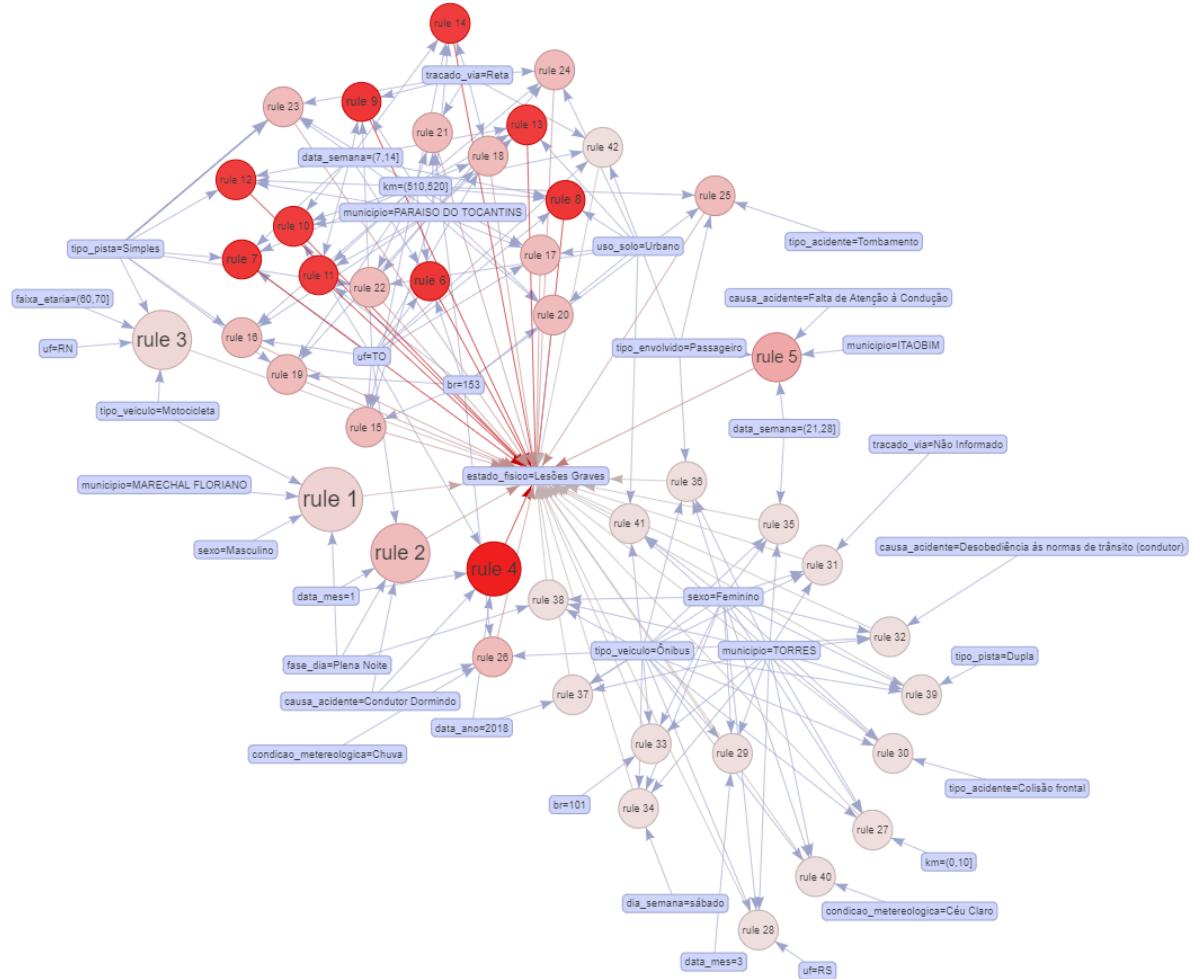
Fonte: Elaborado pelo autor.

Figura 18 - Grafo para o conjunto de regras extraídas para "estado_físico=Lesões Leves"



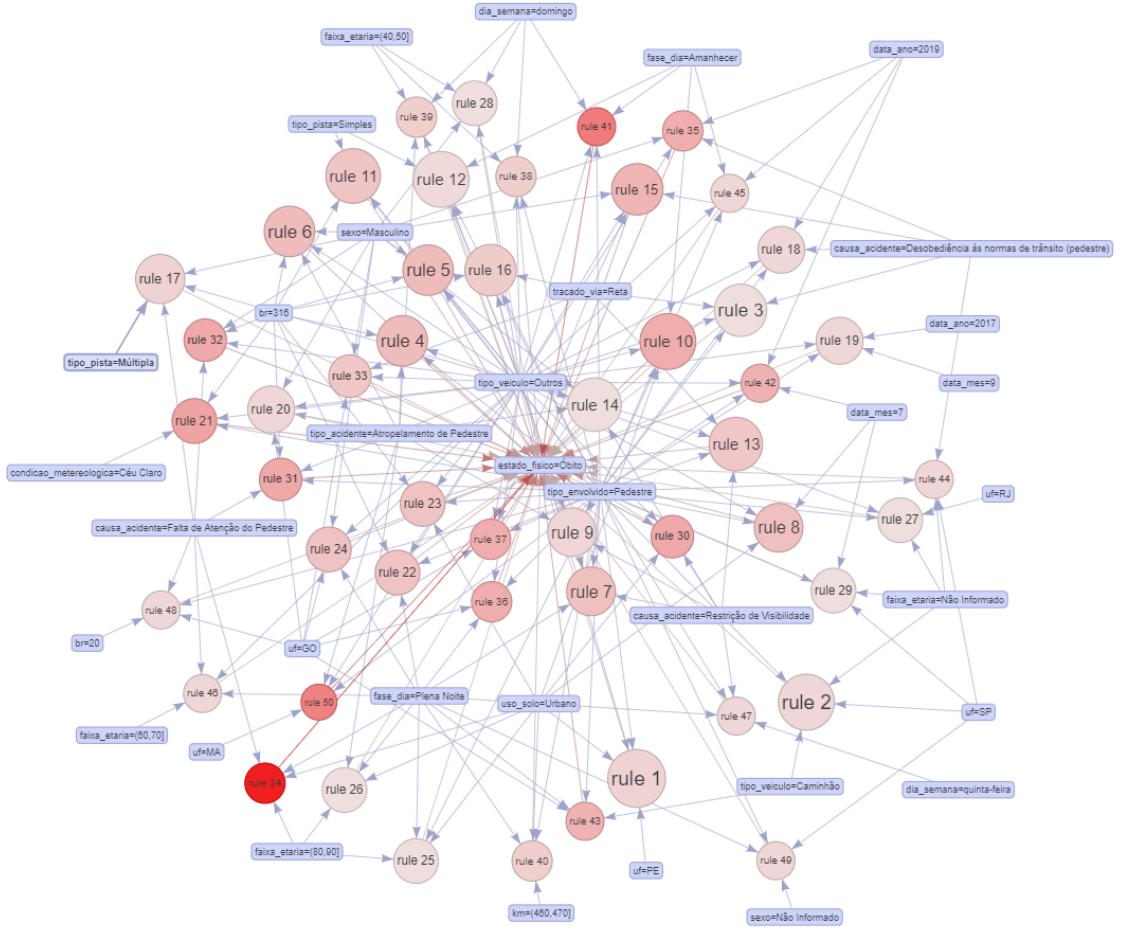
Fonte: Elaborado pelo autor.

Figura 19 - Grafo para o conjunto de regras extraídas para "estado_físico=Lesões Graves"



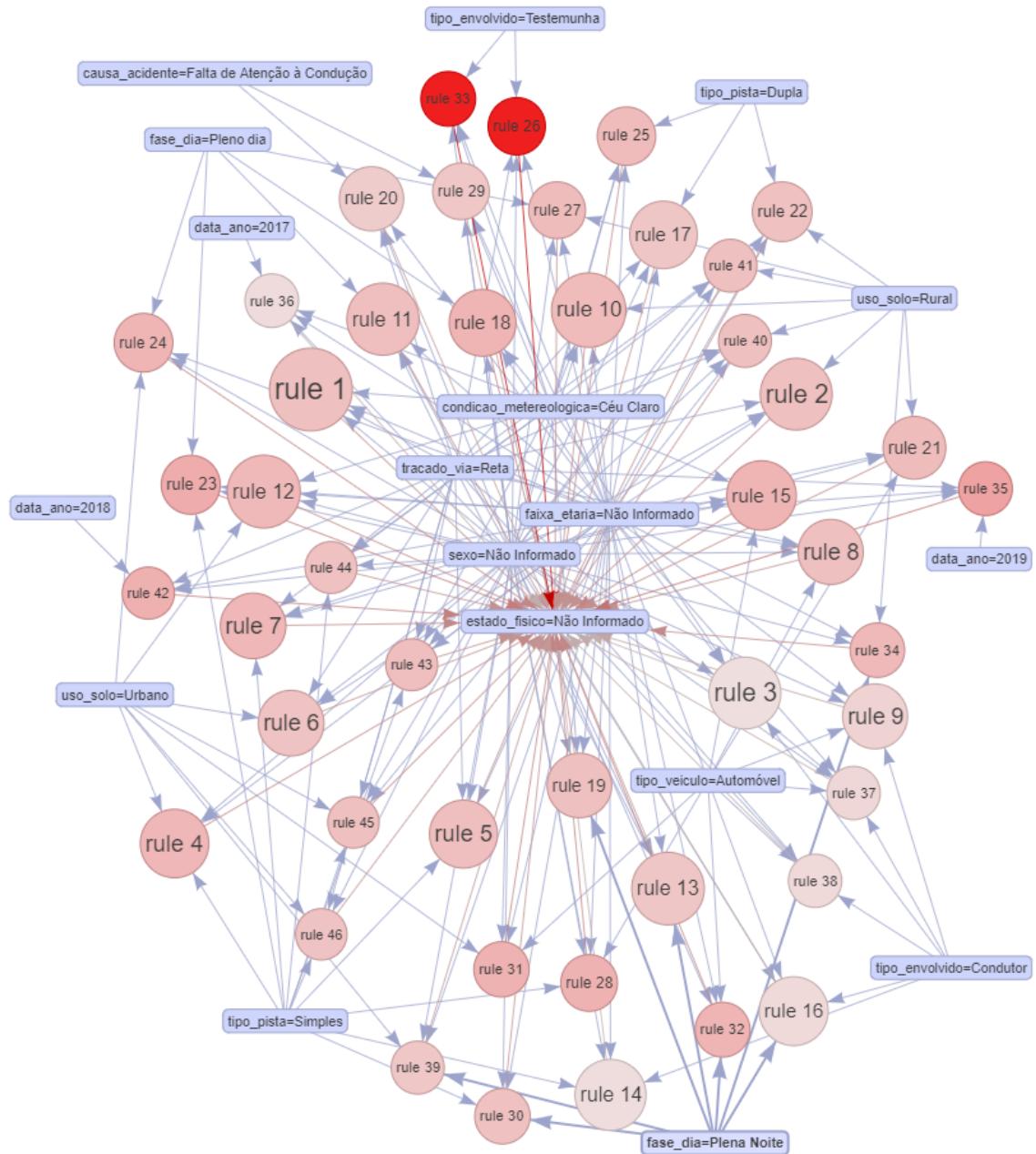
Fonte: Elaborado pelo autor.

Figura 20 - Grafo para o conjunto de regras extraídas para "estado_físico=Óbito"



Fonte: Elaborado pelo autor.

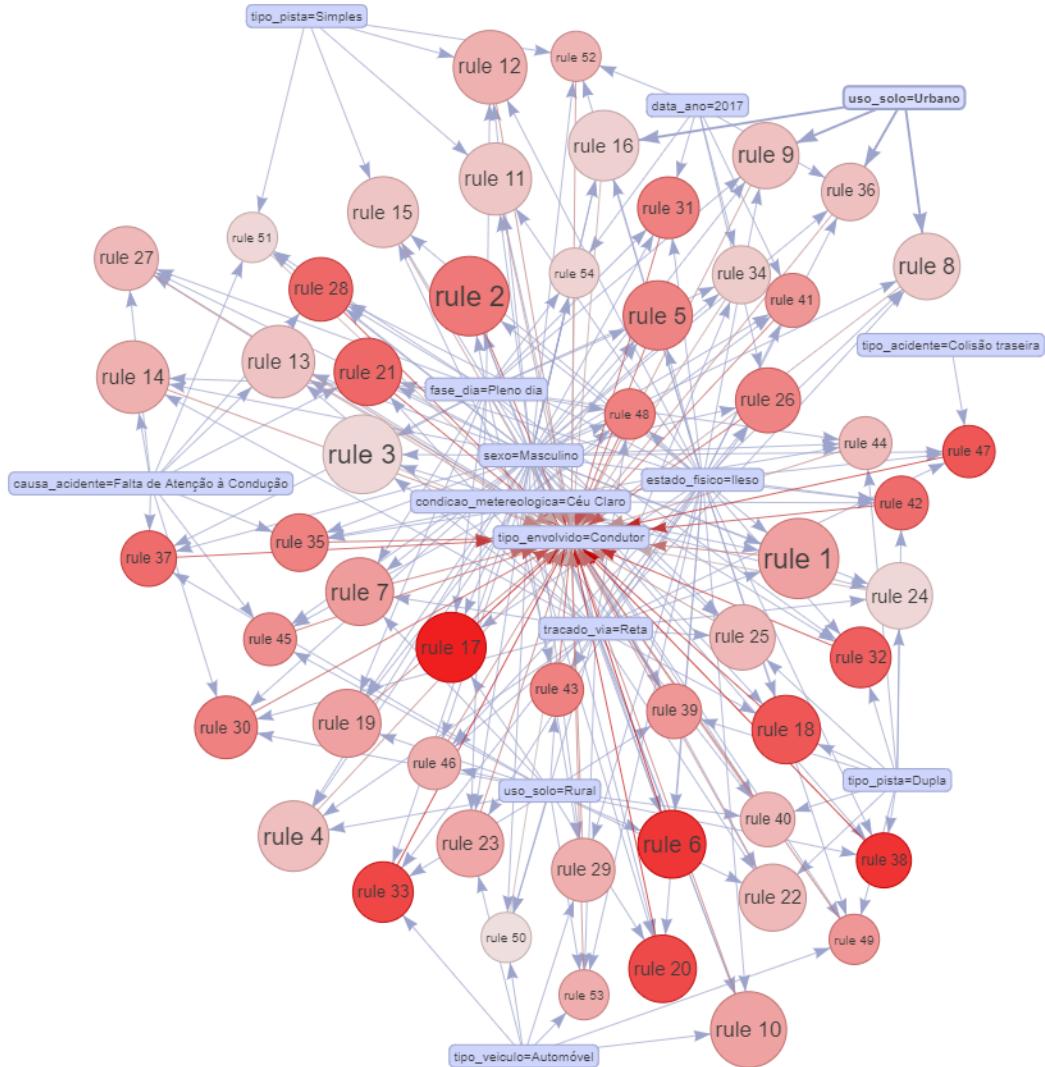
Figura 21 - Grafo para o conjunto de regras extraídas para "estado_físico=Não Informado"



Fonte: Elaborado pelo autor.

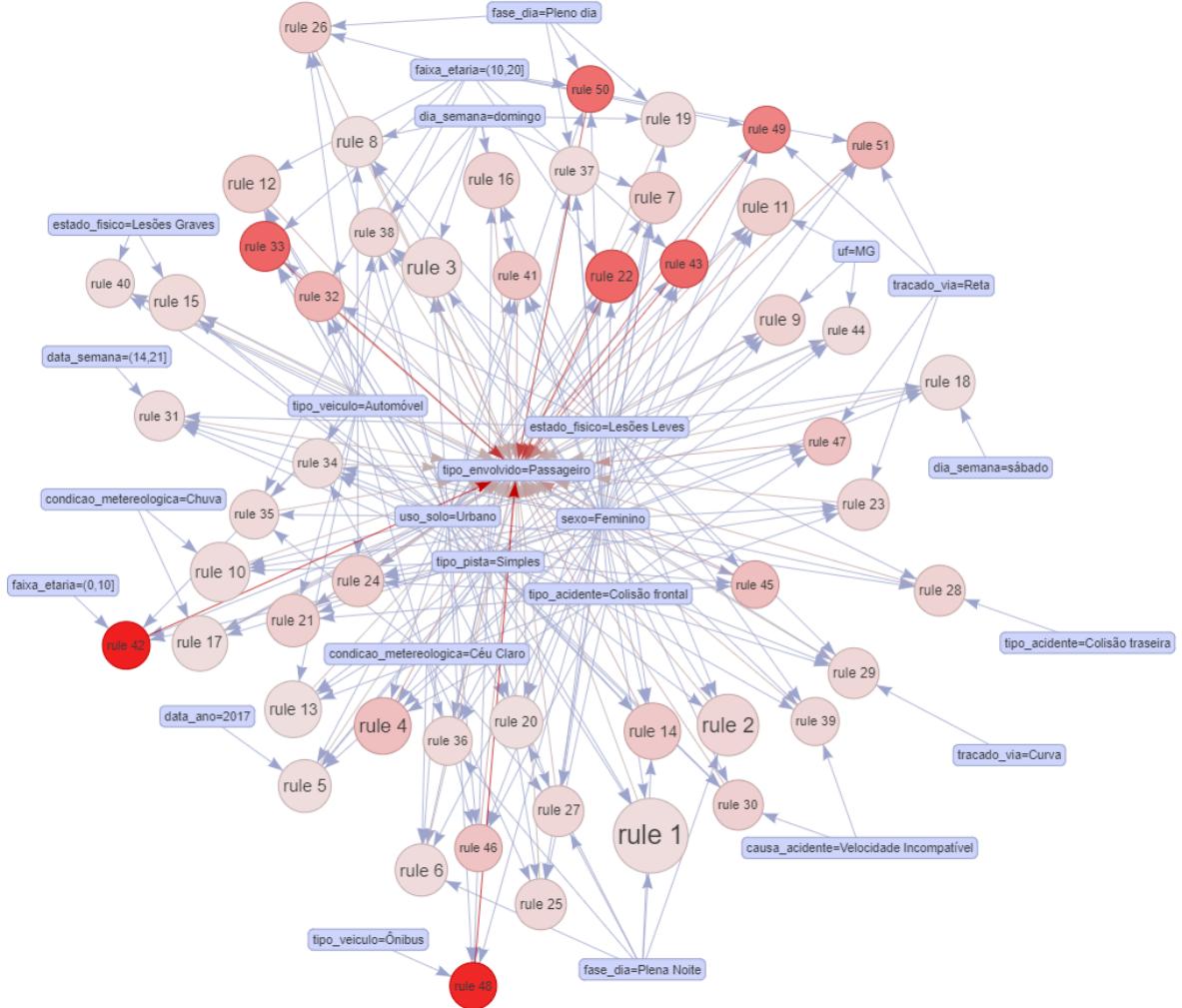
A.3 GRAFOS PARA O ATRIBUTO TIPO ENVOLVIDO

Figura 22 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Condutor"



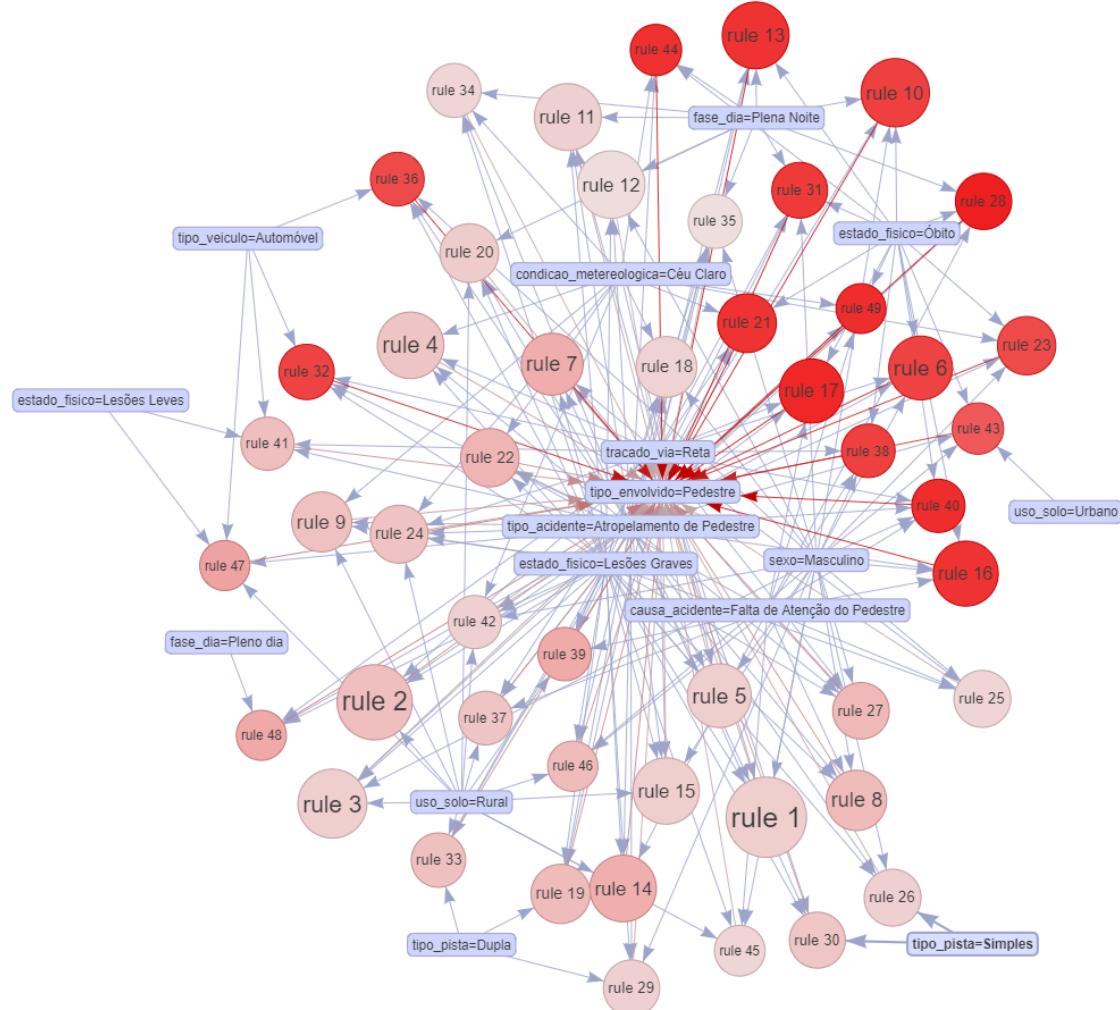
Fonte: Elaborado pelo autor.

Figura 23 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Passageiro"



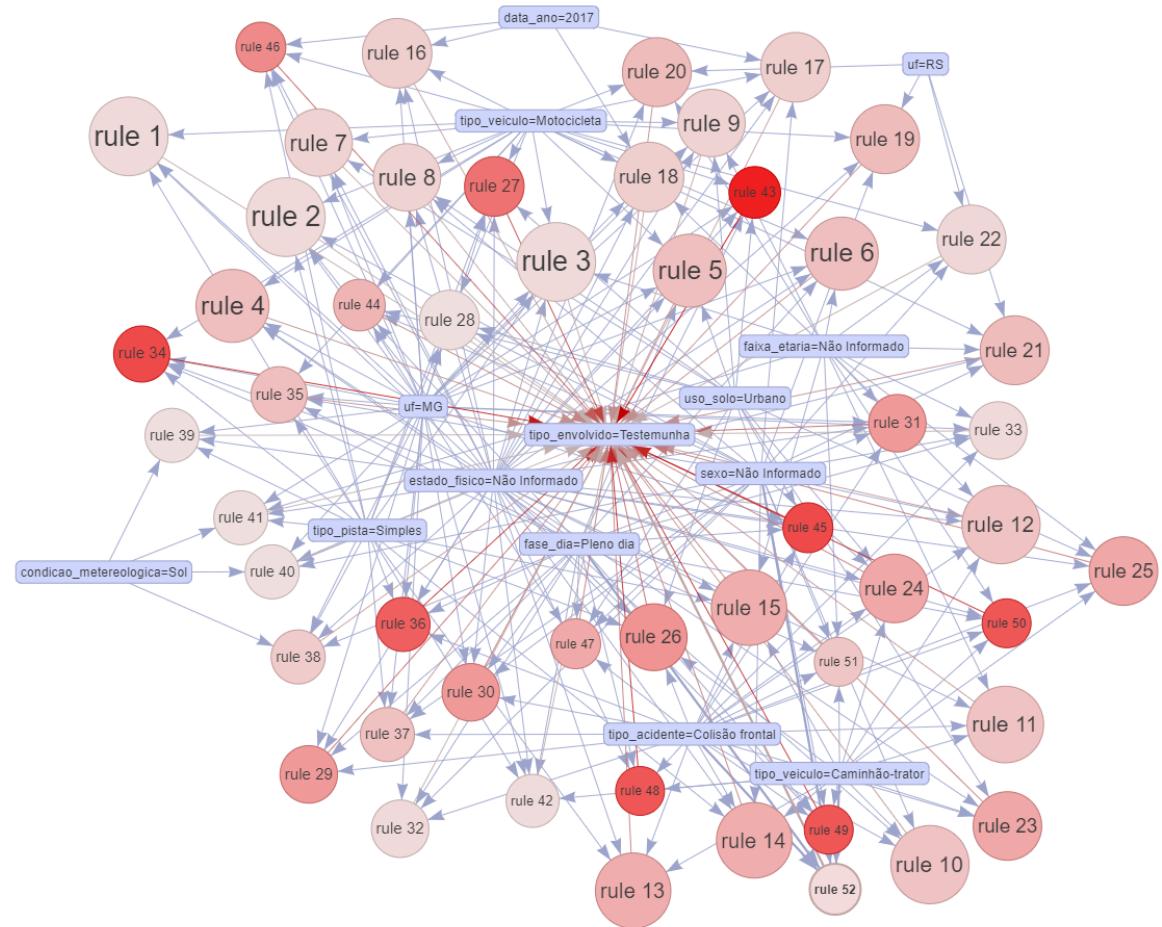
Fonte: Elaborado pelo autor.

Figura 24 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Pedestre"



Fonte: Elaborado pelo autor.

Figura 25 - Grafo para o conjunto de regras extraídas para "tipo_envolvido=Testemunha"



Fonte: Elaborado pelo autor.