UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA - DCT
CURSO DE BACHARELADO EM CIÊNCIA E TECNOLOGIA

DANIEL DA SILVA SANTOS

MINERAÇÃO DE DADOS: UMA ABORDAGEM SOBRE OS DADOS DAS RODOVIAS FEDERAIS DO RIO GRANDE DO NORTE

CARAÚBAS
2019

DANIEL DA SILVA SANTOS

MINERAÇÃO DE DADOS: UMA ABORDAGEM SOBRE OS DADOS DAS RODOVIAS FEDERAIS DO RIO GRANDE DO NORTE

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Ciência e Tecnologia.

Orientador: Wedson Carlos Gomes de Oliveira, Prof. Me.

Coorientador: Sérgio Lins Pessoa, Prof. Me.

CARAÚBAS

2019

Ficha catalográfica elaborada pelo Sistema de Bibliotecas
da Universidade Federal Rural do Semi-Árido, com os dados fornecidos pelo(a) autor(a)

S237m SANTOS, Daniel da Silva.
Mineração de dados: uma abordagem sobre os dados das rodovias federais do Rio Grande do Norte / Daniel da Silva Santos. - 2019.
71 f. : il.

Orientador: Wedson Carlos Gomes de Oliveira. Coorientador: Sérgio Lins Pessoa.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Ciência e Tecnologia, 2019.

1. Mineração de Dados. 2. Acidentes. 3. Descoberta de Conhecimento em Bases de Dados. 4. Rodovias Federais. 5. Rio Grande do Norte. I. OLIVEIRA, Wedson Carlos Gomes de, orient. . PESSOA, Sérgio Lins, co-orient. III. Título.II

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

DANIEL DA SILVA SANTOS

MINERAÇÃO DE DADOS: UMA ABORDAGEM SOBRE OS DADOS DAS RODOVIAS FEDERAIS DO RIO GRANDE DO NORTE

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Ciência e Tecnologia.

Defendida em: 14 de Agosto de 2019

BANCA EXAMINADORA

[redacted]
Wedson Carlos Gomes de Oliveira, Prof. Me. (UFERSA)
Presidente

[redacted]
Oscar Bayardo Ramos Lovon, Prof. Dr. (UFERSA)
Membro Examinador

[redacted]
Mauricio Zuluaga Martinez, Prof. Dr. (UFERSA)
Membro Examinador

AGRADECIMENTOS

Primeiramente a Deus, por até aqui ter me guiado por meio de sua luz, pelas bênçãos que já me proporcionou na vida, pelas amadas pessoas ao meu redor e por todas as forças que me concedeu nos árduos momentos.

A minha família, em especial aos meus amados pais e irmãos José Edilson, Maria Lúcia, Danila Santos e Isaque Edson, por construírem os melhores traços do meu caráter, por me oferecerem a melhor educação que poderiam e por todo o amor, apoio e afeto familiar que me proporcionaram e proporcionam.

A minha namorada, Sabrina Sales, por ter me ajudado diretamente na construção desse trabalho e mesmo indiretamente, por meio do apoio nos mais difíceis momentos e por todo carinho, amor e companheirismo.

Ao meu orientador, o professor Wedson Carlos, pela sua valiosa e sábia orientação que me conduziu a efetivação deste presente trabalho, além de todo o conhecimento, disponibilidade e suporte que me ofereceu.

Ao meu coorientador, o professor Sérgio Lins, que concedeu valiosas contribuições para o desenvolvimento desta pesquisa, sobretudo no inicio, e por atender, sempre com gentiliza, minhas solicitações de auxílio.

A todo o corpo docente e aos demais profissionais da universidade que contribuíram para a minha formação acadêmica/profissional.

Aos meus amigos e colegas, Anderson Picasso, Lucas Amorim, Vanessa Lopes, Mateus Praxedes, Matheus Thomas, Souza Neto, Ellysson Morais, Israel Jonathan e tantos outros não citados, que desde início partilharam comigo dos desafios enfrentados durante a graduação.

Enfim, agradeço a todos que de alguma maneira contribuíram para a realização deste trabalho e/ou da minha formação acadêmica. Meus sinceros agradecimentos!

É uma disciplina que promove, com visão integrada, o gerenciamento e o compartilhamento de todo o ativo de informação possuído pela empresa. Esta informação pode estar em um banco de dados, documentos, procedimentos, bem como em pessoas, através de suas experiências e habilidades.

Gartner Group

RESUMO

O presente trabalho visa aplicar métodos de extração automática de informações (Mineração de Dados) ao registro de acidentes ocorridos em rodovias federais no estado do Rio Grande do Norte. Inúmeras ocorrências de acidentes em vias federais são registradas todos os anos em todo país pela Polícia Rodoviária Federal, essas informações são coletadas com o propósito de suscitar estatísticas e assim esclarecer questões sobre esse quadro. Essas estatísticas são muito bem elaboradas e atualizadas periodicamente por órgãos responsáveis, todavia, percebe-se que investigações em nível municipal e estadual muitas vezes se caracterizam pelo superficialismo. Dessa forma, a Mineração de Dados foi cogitada como uma ferramenta para realização de uma análise minuciosa sobre esse panorama no estado do Rio Grande do Norte e bons resultados advieram dessa metodologia: foram obtidas regras de associações e visualizações que conduziram a inferências sobre o quadro histórico e circunstâncias em que as infelizes casualidades ocorreram, bem como sobre o perfil das pessoas envolvidas nesses incidentes. Certamente, isso pode contribuir para uma melhor percepção a cerca dessa problemática e assim corroborar a favor da segurança das pessoas que trafegam nas rodovias desse estado.

Palavras-chave: Mineração de Dados. Acidentes. Descoberta de Conhecimento em Bases de Dados. Rodovias Federais. Rio Grande do Norte.

ABSTRACT

This work aims to apply methods of automatic information extraction (Data Mining) to the registration of accidents occurred in federal highways in the state of Rio Grande do Norte. Several road accidents are recorded every year throughout the country by the Federal Highway Police, this information is collected in order to raise statistics and thus clarify questions about this situation. These statistics are very well elaborated and periodically updated by responsible agencies, however, it is clear that investigations at the municipal and state levels are often superficially characterized. In this way, Data Mining was considered as a tool to conduct a thorough analysis of this panorama in the state of Rio Grande do Norte and good results came from this methodology: association and visualization rules were obtained and they led to inferences about the historical background and circumstances in which the unfortunate casualties occurred, as well as the profile of the people involved in these incidents. Certainly, this may contribute to a better perception about this problem and thus corroborate in favor of the safety of the people who traffic on the highways of this state.

Keywords: Data Mining. Accidents. Knowledge Discovery in Databases. Federal highways. Rio Grande do Norte.

LISTA DE FIGURAS

Figura 1 - Rodovias do Rio Grande do Norte 18
Figura 2 - Etapas do processo de KDD 23
Figura 3 - Mineração de dados 27
Figura 4 - Ambiente de desenvolvimento (*Jupyter Notebook*) 40
Figura 5 - Arquivos armazenados 41
Figura 6 - Importação das funções 41
Figura 7 - Seleção de dados e atributos 43
Figura 8 - Integração dos dados 44
Figura 9 - Valores ausentes 45
Figura 10 - Valores com ruídos 46
Figura 11 - Dados inconsistentes 47
Figura 12 - Valores incompatíveis e redundantes 47
Figura 13 - Discretização de dados 49
Figura 14 - Nova estrutura dos dados 50

LISTA DE GRÁFICOS

Gráfico 1 - Acidentes ocorridos em cada ano 52
Gráfico 2 - Classificação dos acidentes ocorridos por ano 53
Gráfico 3 - Mapa de calor dos acidentes ocorridos em cada mês e ano 54
Gráfico 4 - Parcela de acidentes acontecidos em cada mês 55
Gráfico 5 - Acidentes distribuídos entre os dias da semana 56
Gráfico 6 - Municípios e rodovias com mais acidentes registrados 57
Gráfico 7 - Tipos de acidentes com vítimas feridas e fatais 58
Gráfico 8 - Principais causas de acidentes com vitímas feridas e fatais 59
Gráfico 9 - Idades das pessoas envolvidas em acidentes 59
Gráfico 10 - Sexo das pessoas envolvidas nas ocorrências registradas 60
Gráfico 11 - Tipo de pistas e classificação dos acidentes 61

LISTA DE QUADROS

Quadro 1 - Variáveis do registro de acidentes (2007 - 2016) 34
Quadro 2 - Novas variáveis inseridas a partir de 2017 35
Quadro 3 - Variáveis Eliminadas 42
Quadro 4 - Interpretação das regras de associação obtidas 63

LISTA DE TABELAS

Tabela 1 - Exemplo de Transações de Itens 30
Tabela 2 - Acidentes em relação ao horário e o dia da semana 56
Tabela 3 - Localidades, tipos de pista e condições climáticas presentes nos registros 61
Tabela 4 - Regras de associação obtidas 62

LISTA DE SIGLAS

AM	Aprendizagem de Máquina
CPF	Cadastro de Pessoas Físicas
DETRAN	Departamento Estadual de Trânsito
DNIT	Departamento Nacional de Infraestrutura de Transportes
IBGE	Instituto Brasileiro de Geografia e Estatística
KDD	*Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bases de Dados)
MD	Mineração de Dados
OMS	Organização Mundial da Saúde
PIB	Produto Interno Bruto
PRF	Polícia Rodoviária Federal
RA	Regras de Associação
RN	Rio Grande do Norte
SQL	*Structured Query Language* (Linguagem de Consulta Estruturada)

SUMÁRIO

1. INTRODUÇÃO 14
2. REFERENCIAL TEÓRICO 16
2.1 Acidentes Viários 16
2.1.1 Acidentes de Trânsito 16
2.1.2 Rodovias Federais do RN 17
2.2 Dados 19
2.2.1 Dados, Informação e Conhecimento 19
2.2.2 Conjuntos de Dados 19
2.2.3 Classificação de Atributos 20
2.2.4 Dados Estruturados e Não Estruturados 21
2.3 Descoberta de Conhecimento em Bases de Dados 22
2.3.1 Definição 22
2.3.2 Etapas do Processo de Descoberta de Conhecimentos em Bancos de Dados 23
2.3.3 Seleção 23
2.3.4 Pré-Processamento 24
2.3.5 Transformação 26
2.3.6 Mineração de Dados 26
2.3.7 Regras de Associação e o Algoritmo *Apriori* 28
2.3.8 Interpretação 32
3. METODOLOGIA 33
3.1 Tipo de Pesquisa 33
3.2 Dados Utilizados 33
3.3 Tecnologias e Ferramentas Utilizadas 36
3.3.1 A Linguagem de Programação *Python* 36
3.3.2 *Numpy* 37
3.3.3 Pandas 38
3.3.4 *Matplolib e Seaborn* 38
3.3.5 *Mlxtend* 39
3.3.6 *Jupyter Notebook* 39
3.4 Aplicação do Processo de KDD 40

3.4.1 Seleção 41
3.4.2 Pré-Processamento 43
3.4.3 Exploração preliminar dos dados 48
3.4.4 Transformação 49
3.4.5 Mineração 50
4. RESULTADOS E DISCUSSÕES 52
4.1 Resultados da Exploração Preliminar 52
4.2 Resultados da Mineração de Dados 61
CONCLUSÃO 65
TRABALHOS FUTUROS 66
REFERÊNCIAS 67

1. INTRODUÇÃO

Em uma sociedade cada dia mais conectada à rede mundial de computadores, a quantidade de dados produzidos e armazenados em massa, devido a essa conexão, aumenta a uma taxa extraordinária. Muitas empresas e organizações modernas exploram e se beneficiam dessa imensa e ascendente disponibilidade de dados. Instituições privadas coletam informações sobre seus clientes a fim de oferecer um serviço ou produto diferencial e maximizar lucros; órgãos do governo os utilizam para planejar e executar políticas públicas racionalmente.

De fato, os dados se tornaram, para a atual sociedade, um fator imprescindível para a administração inteligente, seja pública ou privada. Todavia, de nada seria proveitoso, obter esses recursos, se não fosse possível processá-los, em decorrência, muitas técnicas foram e ainda são desenvolvidas para atender, com máximo nível de eficiência, essa necessidade. Essas técnicas aproveitam o poder computacional das máquinas modernas e permitem extrair informações de conjuntos enormes de dados para solucionar problemas que, em outrora, poderiam ser considerados impossíveis de se resolver, como realizar predições precisas ou descrições automáticas muito complexas. Muitos autores consideram o estudo desses métodos como um novo âmbito do conhecimento, denominado Mineração de Dados (WANG; FU, 2005).

Com efeito, a análise automática de dados atualmente é também utilizada para resolver problemas extrínsecos à esfera administrativa ou econômica, como questões associadas ao processamento de linguagem natural e de imagens, à ciência, à segurança. O presente trabalho pretende, então, aplicar métodos da Mineração de Dados a um estarrecedor problema de âmbito social: os acidentes ocorridos em rodovias federais, especificamente os ocorridos no estado do Rio Grande do Norte.

Inúmeras ocorrências de acidentes em rodovias federais são registradas todos os anos em todo país pela Polícia Rodoviária Federal, essas informações são coletadas com o propósito de suscitar estatísticas e assim esclarecer questões sobre esse quadro. Essas estatísticas são muito bem elaboradas e atualizadas periodicamente por órgãos responsáveis, todavia, percebe-se que investigações em nível municipal e estadual muitas vezes se caracterizam pelo superficialismo. Neste contexto, este trabalho tem como objetivo geral a utilização de técnicas e métodos de mineração de dados, visando extrair padrões relevantes dos repositórios de dados da Polícia Rodoviária Federal (PRF). Espera-se como resultado dessa análise uma descrição minuciosa da atual situação dos acidentes em rodovias do RN,

bem como, encontrar padrões históricos, variáveis relacionadas e as principais circunstâncias que levaram aos acidentes. Além disso, este trabalho se propõe a responder as seguintes questões de pesquisa (QP):

QP 1: Qual o panorama histórico e atual de acidentes em rodovias potiguares?
QP 2: Quais circunstancias que levaram a ocorrência dos acidentes?
QP 3: Qual o perfil das pessoas envolvidas nesses acidentes?

Os objetivos específicos constituem-se em:

- Estudo da problemática abordada (acidentes de trânsito), além das tecnologias e técnicas inerentes a mineração de dados e análise de dados no geral;
- Escolha dos métodos apropriados de mineração de dados, bem como *frameworks* e *softwares* de mineração;
- Análise e discussões dos padrões obtidos.

O presente documento encontra-se organizado da seguinte forma: no capítulo 2 (Referencial Teórico), os principais conceitos e considerações correlacionados ao trabalho serão tratados; no capítulo 3 (Metodologia), os procedimentos e tecnologias utilizadas para que os resultados fossem alcançados serão contextualizados; o capítulo 4 (Resultados e Discussões) foi reservado às interpretações dos resultados obtidos. Por fim, ao final do texto as conclusões a respeito do projeto executado serão discutidas.

2. REFERENCIAL TEÓRICO

Esta seção apresenta os principais temas relacionados a presente pesquisa. Serão apresentados a seguir, os fundamentos teóricos, bem como considerações importantes acerca desses conteúdos.

2.1 Acidentes Viários

2.1.1 Acidentes de Trânsito

De acordo com a Organização Mundial da Saúde (OMS) (2018), cerca de 1,35 milhões de pessoas perdem a vida todos os anos em decorrência de acidentes de trânsito, esse índice indica que incidentes dessa natureza matam mais indivíduos do que doenças como a AIDS/HIV. Além de prejuízos sociais obviamente, incidentes desse tipo, também tem por consequências percas econômicas: ainda segundo esse órgão, esse problema custa cerca de 3% do produto interno bruto (PIB) da maioria dos países do mundo, somado a isso, aproximadamente 93% dessas fatalidades ocorrem justamente em países de frágeis condições economias. Isso posto, é razoável a tipificação desses acontecimentos como um estarrecedor problema de ordem social, econômica e de saúde pública.

Por muito tempo ocorrências desse tipo, entretanto, foram consideradas meros acasos inevitáveis e dessa forma não poderiam ser analisadas racionalmente:

> Na década de 1960 e no início da de 1970, muitos países altamente motorizados começaram a alcançar significativas reduções no número de mortes por meio de abordagens cientificas orientadas à obtenção de resultados. Essa resposta foi estimulada por campanhas propostas por ativistas como Ralph Nader, nos Estados Unidos da América, e que ganharam fundamentação teórica com cientistas, como William Haddon Jr (OMS, 2012, p. 3).

Para Gold (1998), vários fatores corroboram para que um acidente venha a acontecer de fato, no entanto, esses podem ser reunidos em quatro diferentes grupos de maneira geral:

Fatores humanos: referem-se ao estado e comportamento das pessoas envolvidas, como desatenção, nervosismo, embriaguez, etc.

Fatores associados ao veículo: são relativos a possíveis problemas ou condições inadequadas presentes nos veículos utilizados na locomoção, como precariedades nos pneus ou mau funcionamento dos freios.

Fatores vinculados ao ambiente: são aqueles atrelados as localidades e as circunstâncias das vias, incluindo fatores climáticos. Pistas sob más condições, falta de iluminação e chuvas torrenciais podem ser apontados como fatores desse tipo.

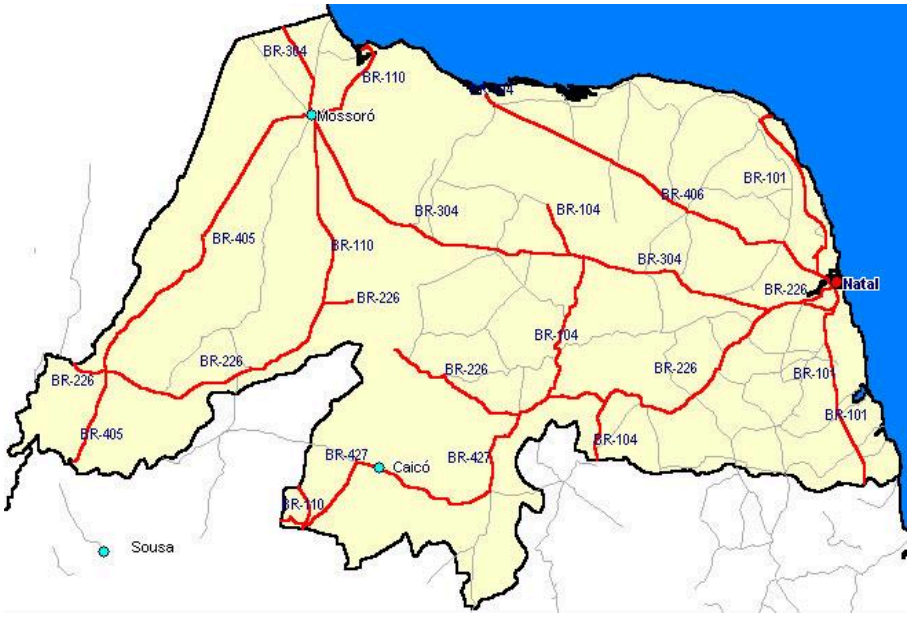
Fatores sociais/institucionais: praticamente correspondem as conjunturas de fiscalizações e as disposições estabelecidas pelo código de trânsito vigente.

No Brasil, o quadro de incidentes não é distinto em comparação com os demais países em desenvolvimento. Segundo o renomado portal de notícias G1 em 2018, entre 2007 e 2017, somente em rodovias federais, mais 83 mil pessoas vieram a falecer no país vítimas de acidentes e em média 23 mortes são registradas por dia nessas vias, por outro lado, a mesma estatística também garante que o país progressivamente vem reduzindo a quantidade dos lamentáveis acontecimentos. As fontes dessa pesquisa remetem ao registro de acidentes disponibilizados pelo principal órgão fiscalizador das rodovias federais brasileiras: a Polícia Rodoviária Federal (PRF). Neste trabalho esses mesmos dados também foram utilizados, todavia, priorizou-se, especificamente, estudar os acidentes ocorridos no estado do Rio Grande do Norte.

2.1.2 Rodovias Federais do RN

O estado do Rio Grande do Norte (RN), segundo o Instituto Brasileiro de Geografia e Estatística (IBGE) (2018), conta com uma população de aproximadamente 3,5 milhões de habitantes, uma frota de quase 1,3 milhões de veículos e com 52,8 mil km² de área territorial. Sobre essa extensão, de acordo com o Departamento Nacional de Infraestrutura de Transportes (DNIT) (2019), há trechos de nove diferentes rodovias federais, oito das quais podem ser observadas na figura 1, a BR 437 não está representada nesse mapa. As subsequentes descrições para cada uma dessas rodovias também estão de acordo com esse departamento.

Figura 1 - Rodovias do Rio Grande do Norte



Fonte: DNIT (2019)

BR 101: segunda maior rodovia federal do Brasil, com uma extensão total de aproximadamente 4,8 mil km, essa rodovia atravessa doze estados distintos, desde o RN até o Rio Grande do Sul.

BR 104: com cerca de 670 km, essa via conecta quatro diferentes estados da região nordeste incluindo o estado Potiguar: Rio Grande do Norte / Paraíba / Pernambuco / Alagoas.

BR 110: possui uma extensão de aproximadamente 1 mil km, essa rodovia passa por cinco estados diferentes: Rio Grande do Norte / Paraíba / Pernambuco / Alagoas / Bahia.

BR 226: contém 1,7 mil km de extensão total, também atravessa cinco unidades federativas: Rio Grande do Norte / Ceará / Piauí / Maranhão / Tocantins.

BR 304: Com um comprimento de 420 km, essa via liga apenas o RN ao estado do Ceará.

BR 405: Conecta o RN ao estado da Paraíba, seu comprimento total é pouco maior do que 250 km.

BR 406: Rodovia presente apenas no estado Potiguar e liga a capital (Natal) ao município de Macau, o tamanho dessa via é entorno de 175 km.

BR 427: Também liga o RN ao estado Paraibano e seu comprimento é próximo de 213 km.

BR 437: Rodovia que adentra o estado de menor comprimento (80 km), também liga o Rio Grande do Norte ao Ceará. Trata-se de uma via sem pavimento algum até a data do presente trabalho.

2.2 Dados

2.2.1 Dados, Informação e Conhecimento

Os principais objetos de estudo deste trabalho são, primordialmente, os dados. É fundamental que isso esteja esclarecido, uma vez que, toda a teoria apresentada nesta obra e procedimentos realizados, diretamente ou não, têm como finalidade estudar, manipular ou processar esses elementos. Torna-se imprescindível, dessa maneira, a conceituação desse objeto, mais do que isso, é importante explorar a distinção entre os termos “dados”, “informação” e “conhecimento”, os quais, apesar de intimamente relacionados, são distintos e, por isso, podem gerar ambiguidades.

Nesse sentido, dados são registros considerados insumos brutos, por consequência, incapazes de sortir conhecimentos de forma individual, mas que, através de um tratamento adequado, podem vir a se tornarem informação (FEDELI; POLLONI; PERES, 2011). Em contrapartida, os autores compreendem a informação como um produto resultante da exploração e interpretação de um conjunto de dados, um resultado descritivo realmente com valor elucidativo intrínseco. Já o conhecimento é, segundo Costa *et al.*, (2014) compreendido como o reconhecimento de padrões ou comportamentos, a partir das informações, que possibilite a execução de análises de cunho preditivo, tais como decisões.

2.2.2 Conjuntos de Dados

Alguns conjuntos de dados podem ser representados, de maneira formal, como uma matriz matemática com “n” quantidade de linhas e “d” número de colunas. Desta maneira, as linhas representam os registros do conjunto, também referidos como exemplos, instâncias ou objetos, já as colunas, abstraem as variáveis, também denominadas de atributos, campos ou propriedades (FACELI *et al.*, 2011). A quantidade de atributos em uma matriz de dados (“d”) define a dimensão dos dados (ZAKI; JR, 2014). Quando há apenas uma única variável, considera-se que o conjunto é formado por dados univariados que, em contrapartida, é

formado por dados bivariados caso haja dois atributos, ou ainda multivariados, no caso de múltiplos (DEVORE, 2006).

2.2.3 Classificação de Atributos

A ciência clássica que estuda conjuntos de dados é a estatística. Esse ramo da matemática classifica os atributos dos dados de acordo com o tipo e nível de mensuração - também conhecido como escala – de seus valores (LARSON; FARBER, 2015). Em relação ao tipo, muitas bibliografias consideram duas classificações: Dados qualitativos ou quantitativos. Por sua vez, são especificados, segundo os autores acima mencionados, ao nível de mensuração como: nominais ou ordinais para dados qualitativos e intervalar ou racional para quantitativos.

O tipo qualitativo, também chamado simbólico, é atribuído a campos que podem ser considerados categóricos, ou seja, conforme apontam Bussab e Morettin (2012), os valores podem ser associados às categorias, como o sexo de uma pessoa, os dias da semana ou as cores de um objeto. Essas categorias são representadas normalmente por conjuntos de caracteres, os quais podem incluir algarismos numéricos. Um exemplo disso, são os valores do Cadastro de Pessoas Físicas (CPF), que, são quase totalmente constituídos por símbolos numéricos. Todavia, não se pode aplicar operações aritméticas a dados simbólicos (FACELI *et al.,*2011). Dessa forma, percebe-se que não há sentido na atitude de somar, subtrair, multiplicar ou dividir os dígitos presentes em dois valores do CPF.

Já os atributos tipificados como quantitativos também denominados numéricos, refletem quantias, de modo que, seus valores podem ser manipulados por operações aritméticas, (FACELI *et al.,* 2011). A idade de uma pessoa, a distância entre dois locais e o saldo de uma conta bancária são variáveis consideradas desse tipo. Ademais, dados numéricos podem ser discretos ou contínuos. Sendo assim, a primeira subclassificação refere-se às quantias que podem assumir infinitos valores, como a medida de uma temperatura e a segunda, apenas a valores dentro de um subconjunto numerável, como a idade de uma pessoa, que assume apenas valores inteiros (CRESPO, 2009).

O nível de mensuração, ou escala, é a propriedade que determina as cabíveis operações matemáticas que podem ser realizadas sobre os dados de um atributo (LARSON; FARBER, 2015). Como já discutido, certos valores não podem ser processados por meio de operações aritméticas, os dados qualitativos possuem essa característica, ou seja, o nível de mensuração dos dados dessa natureza é, de maneira geral, considerado baixo, todavia, é regular a

classificação desses como valores pertencentes à escala nominal ou ordinal. (LARSON; FARBER, 2015). A primeira, é o menor nível de mensuração, corresponde a dados categorizados por rótulos ou nomes relacionados minimamente e, por isso, segundo Pinheiro *et al.,* (2009) a única operação que pode ser executada sobre tais é uma comparação de igualdade ou desigualdade. Por outro lado, os autores destacam que algumas categorias, mesmo que qualitativas, podem ser organizadas entre si, de modo que uma ordem pode ser definida. Por exemplo, é totalmente admissível ordenar os nomes dos meses. Dados com esse comportamento pertencem à escala ordinal.

Dados numéricos, também possuem níveis de mensuração distintos (ZAKI; JR, 2014). A escala intervalar caracteriza variáveis que consertem o cálculo da diferença entre seus valores, porém, não o cálculo da razão, pois, isso ocorre quando o valor zero, inserido no intervalo de valores que esse conjunto assume, não é o usual da matemática (FACELI *et al.,* 2011). Para exemplificação, as horas são registros que se comportam dessa maneira. Por fim, o nível de mensuração racional é a propriedade atribuída a grupos numéricos nos quais a razão entre seus valores possui sentido matemático e o valor zero, de fato, o significado usual de nulo (LARSON; FARBER, 2011), a quantidade de acidentes ocorridos em uma rodovia é naturalmente o conjunto pertencente a essa escala. Desta maneira, conforme apontam Faceli, *et al.,* (2011), o nível de mensuração é um importante fator que indica a quantidade de informação atrelada aos dados.

2.2.4 Dados Estruturados e Não Estruturados

Devido ao desenvolvimento tecnológico e, sobretudo, ao surgimento da rede mundial de computadores, a quantidade de dados digitais vem aumentando de maneira exponencial (GALDINO, 2016). Estima-se que o número de registros em banco de dados no mundo inteiro duplica a cada 20 meses (WITTEN; FRANK, 2005). Esses são dados tipicamente estruturados, que, de acordo com Cavique (2014), representam apenas uma pequena parcela do montante de dados produzidos, de maneira contínua e acelerada, atualmente.

Nessa perspectiva, dados estruturados são aqueles que podem ser organizados conforme o modelo formal de conjuntos de dados, isto é, como uma estrutura matricial fixa (CIELEN; MEYSMAN; ALI, 2016). Devido a essa conveniente particularidade, dados dessa natureza são frequentemente armazenados em bancos de dados relacionais, os quais utilizam estruturas semelhantes a tabelas como suporte de persistência, e uma linguagem que segue o

padrão SQL (*Structured Query Language*[1]) para manipulação e extração desses dados quando necessário (COSTA, 2011).

A maior parcela de dados, atualmente, não se encontra mais em bases de dados relacionais, mas em arquivos, textos, áudios, vídeos e imagens, que são os denominados dados não estruturados. Esses representam cerca de 90% da quantidade total de dados disponíveis (TESSAROLO; MAGALHÃES, 2013). Nos últimos anos, esses dados cativaram o interesse, sobretudo, das empresas, e obtiveram notoriedade como abundantes fontes de valiosas informações. Devido a isso, HURWITZ *et al*., (2013) enfatizam que várias tecnologias, ferramentas e abordagens capazes de armazenar, extrair, e processar dados desse tipo foram desenvolvidas recentemente.

2.3 Descoberta de Conhecimento em Bases de Dados

2.3.1 Definição

A clássica definição para Descoberta de Conhecimento em Bases de Dados ou *Knowledge Discovery in Databases* (KDD), referida em muitas literaturas, foi dada por Fayyad, Piatetsky-Shapiro e Smyth como "um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e, em última instância, compreensíveis em dados" (1996, p.649, tradução nossa[2]). Ademais, segundo Foster *et.al* (2013), as técnicas provenientes desse método se aplicam a grandes volumes de dados, quantidades que representam um desafio no tocante à obtenção de conhecimentos, quando convencionais técnicas de exploração e análise são utilizadas. Dessa maneira, KDD pode ser compreendido como um procedimento sistemático e eficiente que busca extrair padrões e informações dotadas de valor, não evidentes, de grandes conjuntos de dados (COSTA *et al.,* 2019).

Ademais, o KDD é um método de análise cuja execução pode ser decomposta em várias etapas, e se caracteriza como um processo interativo e iterativo, uma vez que essas etapas não são definitivas, mas interdependentes e que podem ser retroalimentadas (PRASS, 2016). Além disso, as técnicas em cada etapa desse procedimento se aplicam a dados estruturados. Conforme Feldman e Sanger (2006), vários métodos análogos também foram

[1] Na Língua Portuguesa: Linguagem de Consulta Estruturada.

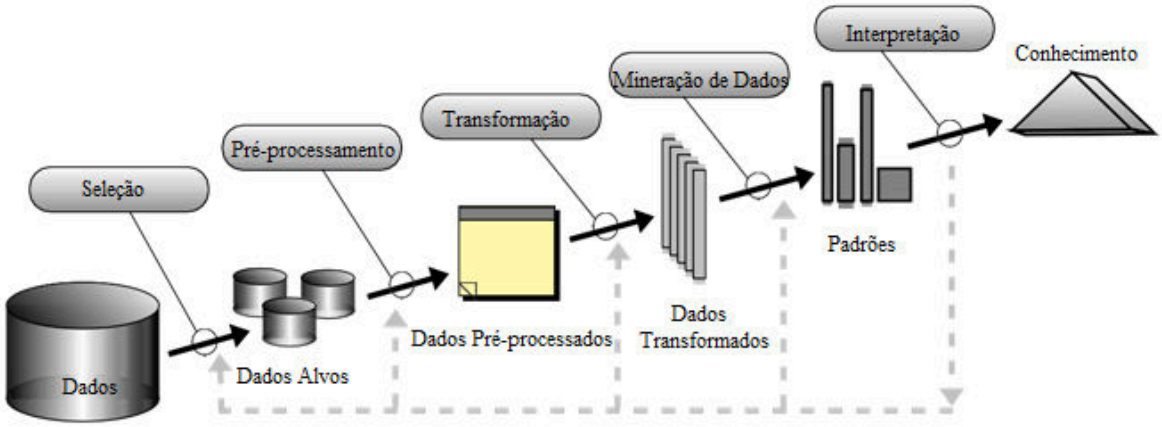
[2] No original: "Is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p.649).

desenvolvidos para atender a necessidade do processamento de dados não estruturados. Neste trabalho, entretanto, somente as primeiras serão abordadas.

2.3.2 Etapas do Processo de Descoberta de Conhecimentos em Bancos de Dados

As cinco etapas mais reconhecidas em meio acadêmico que subdividem o processo de KDD também foram sugeridas por Fayyad, Piatetsky-Shapiro e Smyth (1996), estas são: Seleção, Pré-processamento, Transformação, Data Mining e Interpretação ou Avaliação. Ainda segundo esses, em cada uma dessas fases o nível de conhecimentos requeridos é alto e o número de decisões a serem tomadas é significativo, o que torna o processo, como um todo, complexo (1996). A ordem de operação dessas etapas e os relacionamentos entre essas estão esquematizadas na Figura 2. As seguintes seções tratarão minuciosamente de cada uma dessas fases.

Figura 2 - Etapas do processo de KDD



Fonte: FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996).

2.3.3 Seleção

A seleção, após o estudo dos dados, é a primeira fase do processo de KDD. Neste nível, um subconjunto de interesse (*target data*) é extraído a partir da massa original de dados, quando se realiza uma filtragem dos registros e atributos que são julgados relevantes para a análise (ZHANG, S.; ZHANG, C.; XINDONGWU, 2004). Para isso, é necessário, não apenas

que a fonte de dados seja conhecida, como também que se saiba qual tipo de exploração será feita, bem como, quais dados serão importantes para que os resultados desejáveis sejam obtidos. O processo de seleção é particularmente significativo, pois mesmo poderosas técnicas de extração de conhecimentos provenientes do KDD podem ter seu desempenho comprometido perante grandes quantidades de objetos ou variáveis (FACELI *et al.,* 2011).

2.3.4 Pré-Processamento

O pré-processamento é uma etapa essencial do processo de extração de conhecimentos a qual requer maior esforço e tempo dentre todas as demais, mais de 50% do tempo gasto em todo o processo é investido nessa fase (OLSON; DELEN, 2008). Larose (2005) também destaca a necessidade desta etapa, dado que, os dados presentes em bases de armazenamento muitas vezes apresentam ruídos, redundâncias e inconsistências ou ainda podem ser dados obsoletos ou incompletos. Além disso, de acordo com o autor, todos esses fatores podem comprometer o procedimento de análise e gerar resultados imprecisos. Esse estágio é, segundo Pyle (1999) responsável por realizar a limpeza dos dados a serem analisados, essa limpeza deve ser feita conforme as características das técnicas que serão utilizadas durante o processo de extração de conhecimentos.

Dentre os problemas frequentemente deparáveis ao se analisar dados, situa-se a ausência de valores em atributos de interesse (*missing data*) ou a presença de dados nulos, que não agregam qualquer informação à análise feita (LAROSE, 2005). A ausência de valores em atributos se dá devido a várias situações, tais como, erros na coleta ou na entrada de dados, desconhecimento do valor da variável durante o registro, ou até devido ao caráter optativo que essa possa possuir em relação ao preenchimento (FACELI *et al.,* 2011). No tocante às abordagens a serem tomadas no tratamento de dados faltosos, Han *et. al.,* (2012) lista várias técnicas que podem ser adotadas para esse propósito, dentre as quais está a simples eliminação dos atributos incompletos ou preenchimento desses valores através de métodos estatísticos, como média, moda e mediana, manualmente de acordo com os valores mais prováveis ou ainda através de uma constante global que representa os dados ausentes. A técnica que deverá ser utilizada depende, obviamente, de inúmeros fatores como a quantidade de registros, de campos, a natureza dos dados, entre outros.

Em bases de dados multivariados é possível que haja também objetos com propriedades incoerentes, isto é, registros cuja relação dos próprios atributos é logicamente conflitante, dados com esse problema são denominados inconsistentes (SÁ, 2015). Como

exemplo, em um conjunto de dados que reúna características de pessoas, certamente haveria inconsistência no registro de um indivíduo recém-nascido com 1,6 m de altura. Em análises complexas, instâncias com esse tipo de problema não são reconhecíveis facilmente, um refinado domínio sobre o conjunto de dados utilizado pode ser requerido para a identificação desses, ou ainda métodos matemáticos especiais (SÁ, 2015). Em contrapartida, segundo Côrtes, Porcaro e Lifschitz, (2002), uma correção manual desses dados pode ser eficaz em alguns casos, para isso deve-se recorrer a conhecimentos extrínsecos aos dados.

Outra situação recorrente que se deve lidar na fase de pré-processamento, se refere à presença de valores discrepantes, compreendidos como registros que assumem valores destoantes em relação aos demais, chamados de *outliers* (PRASS, 2016). Frequentemente, esses valores representam erros aleatórios, nesse caso, diz-se que são dados com ruídos (CÔRTES; PORCARO; LIFSCHITZ, 2002). De acordo com Olson e Delen, "os *outliers* podem ser causados por vários motivos, como erros humanos ou erros técnicos, ou podem ocorrer naturalmente em um conjunto de dados devido a eventos extremos" (2008, p. 13, tradução nossa[3]). Logo, algumas técnicas que podem ser utilizadas no tratamento e identificação de dados ruidosos e *outliers* são: *binning,* agrupamento ou regressão. A primeira se refere à suavização de dados de acordo com os valores ao seu entorno por meio da média ou mediana estatística; a segunda consiste em agrupar dados semelhantes na tentativa de encontrar dados atípicos; por último, a regressão equivale a determinar uma linha de tendência de valores numéricos, a qual pode ser utilizada tanto na identificação como na substituição de dados com ruídos (HAN; KAMBER; PEI, 2012). Outrossim, segundo Pyle (1999), *outliers* podem ser tratados como valores ausentes.

A ocorrência de registros e de variáveis redundantes é outra situação que requer atenção e tratamento nessa fase. A redundância, nesse caso, não significa apenas valores repetidos, mas registros e atributos que refletem a mesma informação (OLSON; DELEN, 2008). A idade de uma pessoa e a data de nascimento podem ser citados como de atributos tipicamente redundantes, nesse sentido. Redundâncias em dados não beneficiam o processo de KDD, ao invés disso, podem perturbar os resultados (WANG; FU, 2005). Para Pyle (1999), a solução ideal para esse problema é simplesmente a remoção do atributo ou objeto que não agrega novas informações. É importante salientar, no entanto, que dados redundantes podem ter sido submetidos à circunstâncias distintas, e por isso, mesmo contendo informações

[3] No Original: "Outliers may be caused by many reasons, such as human errors or technical errors, or may naturally occur in a data set due to extreme events." (OLSON e DELEN, 2008, p. 13).

essencialmente iguais, dois atributos podem ter nível de qualidade diferentes em relação a consistência e ausência de valores (YE, 2003).

De acordo com Pyle (1999), a ocorrência de valores nulos, inconsistentes e redundantes pode estar associada à uma má integração de dados. Esta tarefa também pode ser realizada no pré-processamento, na qual ocorre a junção de dados oriundos de várias fontes distintas em único conjunto (SOCZEK; ORLOVSKI, 2004). É importante verificar, as particularidades de cada dessas fontes nesse processo, uma vez que isso pode acarretar justamente nos problemas já citados (CAMILO; SILVA, 2009).

2.3.5 Transformação

No estágio de transformação, o principal objetivo é preparar definitivamente os dados para a próxima etapa, a qual se encarrega, de fato, da extração de informações (SILVA, 2004). Para isso, deve ser realizada a redução de objetos, categorias e de atributos presentes na coleção de dados, bem como a conversão entre tipos (simbólicos para numéricos ou vice-versa). Isso se torna necessário, uma vez que, muitas das técnicas utilizadas na obtenção de informações trabalham exclusivamente com um desses tipos de dados. Ademais, essas se tornam impraticáveis quando a dimensão do conjunto e o número de exemplos são muito grandes (CAMILO; SILVA, 2009).

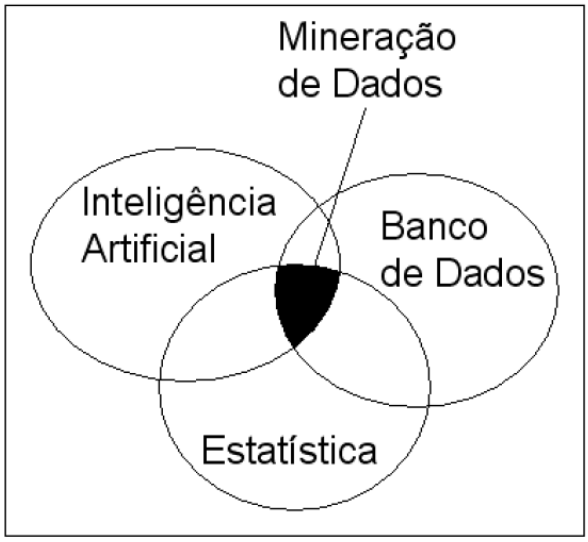
Muitas vezes, modificações nos atributos também são necessárias nesta fase, dependendo da técnica a ser utilizada, como agregações, normalizações e generalizações, há casos, inclusive, que novas variáveis devam ser construídas (CÔRTES; PORCARO; LIFSCHITZ, 2002). A agregação é uma operação que se refere ao agrupamento de valores, por exemplo, agrupar ocorrências diárias de acidentes em meses, normalização à tarefa de adaptar valores numéricos a uma nova escala como um intervalo de 0 a 1. Já generalização ou discretização significa realizar a conversão de atributos quantitativos em simbólicos através da atribuição de categorias a valores numéricos como, por exemplo, a definição da classe social de uma pessoa de acordo com o salário que essa recebe.

2.3.6 Mineração de Dados

Dentre todas as etapas do KDD, a mais importante é a mineração de dados (MD) ou *data mining* (GOLDSCHMIDT; PASSOS, 2005). Justamente devido a isso, muitos autores tratam o processo de KDD como sinônimo de mineração de dados. Esta é, dessa forma, a fase

na qual informações relevantes, como padrões, associações, mudanças e anomalias são extraídas do grupo de dados de fato. Para Silva (2004), pode-se compreender a mineração de dados como uma aplicação de conhecimentos dentro da interseção de três áreas distintas: a Estatística, a Inteligência Artificial e a Teoria sobre Bancos de Dados (Figura 3). Por sua vez, Thomé define que "*data mining* é a concepção de modelos computacionais capazes de identificar e revelar padrões desconhecidos, mas existentes entre dados pertencentes a uma ou mais bases de dados" (2002, p. 13). Em suma, portanto, a fase de mineração consiste em aplicar técnicas e algoritmos computacionais ao conjunto de dados selecionado, já processado e transformado, com o intuito de obter informações úteis.

Figura 3 - Mineração de dados



Fonte: SILVA (2004).

A Mineração de Dados engloba mecanismos ou métodos que objetivam, essencialmente, realizar previsões ou descrições. A primeira finalidade consiste em, a partir de variáveis conhecidas, estimar propriedades desconhecidas ou que podem apenas ser futuramente determinadas; a segunda trata da identificação de padrões elucidativos e compreensíveis para humanos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Para Goldschmidt e Passos (2005), as principais tarefas da MD utilizadas para prever valores e descrever padrões são:

Descoberta de Associações: Busca-se identificar quais elementos ocorrem simultaneamente e com frequência em uma base de dados.

Classificação: A partir dos padrões presentes nos dados conhecidos, tenta-se modelar uma função capaz de atribuir categorias a novos registros.

Regressão: Semelhante à classificação, porém se aplica a dados numéricos. Os valores desconhecidos são estimados também por meio de um modelo, e, nesse caso, uma função matemática como a regressão linear pode ser utilizada.

Sumarização: Consiste na identificação e descrição das principais características dos objetos presentes em um conjunto de dados.

Clusterização: Equivale a agrupar objetos com características semelhantes em subgrupos (*clusters*) de maneira automática.

Para executar essas tarefas, a MD faz uso, como já discutido, da teoria e de algoritmos provenientes da inteligência artificial, sobretudo, da área conhecida como Aprendizagem de Máquina (AM) ou *Machine Learning* (SILVA, 2005; NILSSON, 1996). Para Ratner (2011) *apud* Samuel (1959), AM pode ser compreendida como o campo de estudo dentro da computação que busca produzir algoritmos capazes de desenvolver um comportamento próprio diretamente através dos dados, não havendo a necessidade de programá-los explicitamente. Algoritmos desse tipo vêm sendo utilizados em diversas áreas do conhecimento e obtendo ótimos resultados na resolução de distintos problemas (FOSTER; FAWCETT, 2013).

De acordo com Nelli (2015), a AM abrange vários métodos com propriedades específicas e distintas, a escolha de algum depende da natureza dos dados e do modelo que se deseja construir, fatores que definem o tipo de problema de aprendizagem, que segundo o mesmo autor, são dois: aprendizagem supervisionada e não supervisionada. No primeiro tipo de aprendizagem, os dados possuem uma característica de particular interesse (*target*), a qual, frequentemente, se deseja monitorar o comportamento de acordo com as demais. Desta maneira, conforme ressalta Madhavan (2015), um algoritmo de AM pode então ser utilizado para aprender como se dá esse comportamento e então conseguir atribuir um valor a essa característica em novos objetos. Em contrapartida, na aprendizagem não supervisionada não há uma variável de particular interesse, mas é desejável muitas vezes uma descrição dos dados ou um agrupamento entre semelhantes. Em ambos os casos o próprio algoritmo AM encarrega da inferência do resultado (VERMEULEN, 2018).

2.3.7 Regras de Associação e o Algoritmo *Apriori*

Na metodologia desse trabalho utilizou-se a técnica de MD conhecida como Regras de Associação (RA). No entanto, antes de explorar esse ponto, para que ambiguidades sejam evitadas, é importante distinguir minuciosamente os já citados termos: tarefa, técnica e algoritmo. Assim sendo, as tarefas são as funções que a mineração de dados pode realizar, as técnicas dizem respeito às maneiras de como essa função poderá ser feita, já o algoritmo representa uma ferramenta que poderá ser utilizada. Uma tarefa na mineração de dados pode ser realizada por meio de várias técnicas que por sua vez utilizam algoritmos específicos (GOLDSCHMIDT; PASSOS, 2005). Uma vez feito esse esclarecimento, nota-se com facilidade que a técnica de Regras de Associação e a tarefa de Descoberta de Associações são conceitos distintos, apesar de relacionados.

A técnica de Regras de Associação (RA) possibilita encontrar inter-relações entre categorias de uma base de dados de maneira automática, ou seja, padrões entre ocorrências simultâneas de elementos (TAN; STEINBACH; KUMAR, 2005). Uma clássica aplicação dessa técnica é a resolução do problema da "cesta de mercado", cujo objetivo é identificar quais produtos de um comércio são frequentemente vendidos em conjunto, e quais desses, quando comprados, influenciam na aquisição de um novo grupo (YE, 2003, p. 27). Já uma aplicação mais recente se refere ao seu uso na análise de navegação de um usuário pela internet para fins promocionais (RUD, 2001). Essas utilizações, evidentemente, são de interesse comercial, entretanto, segundo Tan, Steinbach e Kumar (2005) a aplicabilidade das RA's e da própria tarefa de associação não se restringe a esse âmbito:

> Além da cesta de mercado, a análise de associação também é aplicável a outros domínios de aplicação, como bioinformática, diagnóstico médico, *web mining* e análise de dados científicos. Na análise dos dados da ciência da Terra, por exemplo, os padrões de associação podem revelar conexões interessantes entre os processos oceânicos, terrestres e atmosféricos. (2005, p.328, tradução nossa[4])

No tocante à teoria matemática sobre a extração de regras de associação, pressupõe-se que os dados que serão explorados estão organizados em uma estrutura tabular do tipo: $T \times P$, em que T é um conjunto matemático de n transações $T = \{ t_1, t_2, t_3, \dots, t_n\}$ e P um conjunto de m itens $P = \{ p_1, p_2, p_3, \dots, p_m\}$. Transações se referem a registros que, de alguma maneira, constroem uma relação de itens. Estas podem representar, por exemplo, uma compra ou um usuário da internet, por outro lado, os itens seriam os produtos ou os *websites*

[4] No Original: "Besides market basket, association analysis is also applicable to other application domains such as bioinformatics, medical diagnosis, web mining and scientific data analysis. In the analysis of Earth science data, for example, the association patterns may reveal interesting connections among the ocean, land, and atmospheric processes" (TAN; STEINBACH; KUMAR, 2005, p.328).

acessados (ZAKI e JR, 2014). Um item p pode ou não estar presente em uma transação t, o que indica uma ocorrência. Desse modo, essa informação deve estar no conjunto de dados por meio de valores binários como {Sim, Não} ou {0,1}. A Tabela 1 é um exemplo de conjunto de dados com a estrutura conveniente para extração de RA's.

Tabela 1 - Exemplo de Transações de Itens

Transação	**A**	**B**	**C**	**D**
1	Sim	Sim	Sim	Sim
2	Não	Não	Sim	Não
3	Não	Sim	Não	Sim
4	Sim	Não	Não	Não
5	Sim	Sim	Não	Sim

Quando uma transação contém um item ou um subconjunto de itens (*itemset*), diz-se que essa dá suporte a esse item ou a esse subconjunto. Formalmente, um *k-itemset* é um subconjunto I com k elementos, tal que $I \subset P$, o suporte de I – ou de um item p, se k=1 - é definido como a razão entre quantidade de transações que contém esse subconjunto de itens, pela quantidade total de transações (YE, 2003). No exemplo da tabela, o suporte do 2-*itemset* {A,B} é igual a 0,4, pois no total de cinco transações, em duas houveram as ocorrências do item A e também do item B.

Uma regra de associação, enfim, pode ser compreendida como uma implicação frequentemente representada da seguinte maneira: $X \rightarrow Y;\ X \cap Y = \emptyset$, onde X é denominado termo antecedente e Y representa o termo consequente, ou seja, Y é uma provável consequência de X e ambos são k-*itemsets* sem elementos em comum (AGRAWAL; SRIKANT, 1994.). No processo de extração dessas regras, duas métricas são particularmente importantes: o próprio suporte dessa e sua confiança. A primeira é definida pela equação (1) e representa a percentagem de casos que a regra se demonstrou válida. Na busca por um RA, frequentemente, um suporte mínimo é estabelecido.

$$suporte(X \rightarrow Y) = suporte(X \cup Y) \qquad \textbf{(1)}$$

Já a métrica de confiança, está relacionada com a probabilidade de a consequência ocorrer entre as transações que contém o k-i*temset* antecedente, formalmente, a confiança pode ser definida pela expressão (2) (AGRAWAL; IMIELINSKI; SWAMI, 1993).

$$confiança(X \rightarrow Y) = \frac{suporte(X \cup Y)}{suporte(X)} \quad \textbf{(2)}$$

Para minerar regras de associação em grandes bases de dados, um algoritmo eficiente se torna necessário. Nessa lógica, dentre os desenvolvidos para atender esse propósito, o clássico algoritmo *Apriori*, indubitavelmente, é o mais simples e importante (RAO; GUPTA, 2012). Proposto por R. Agrawal e R. Srikant, esse algoritmo é executado em duas etapas, a primeira consiste na busca de *itemsets* mais frequentes, enquanto a segunda, na elaboração das regras (LAROSE, 2005).

Para determinar os conjuntos de itens mais frequentes, o algoritmo necessita que um suporte mínimo seja especificado primeiramente, então um processo iterativo é inicializado. Na primeira iteração, o suporte de cada 1-*itemset* é calculado e comparado com o fornecido, os subconjuntos cujo suporte seja maior ou igual ao especificado serão selecionados e combinados em 2-i*temsets*. Na próxima iteração, o suporte desses também é calculado e os que atendem o critério já citado, de maneira análoga, são combinados em 3-*itemsets* e, dessa maneira, o procedimento se repete até que a máxima combinação frequente seja encontrada (LAROSE, 2005). Percebe-se que não são determinadas todas as combinações de k-*itemsets* nesse processo, (certamente uma tarefa inviável), isso porque o *Apriori* foi desenvolvido a partir do princípio de que a combinação de itens frequentes possui um suporte máximo igual ao desses individualmente. Esse princípio é denominado "Propriedade *Apriori*", a essência desse algoritmo (HAN; KAMBER; PEI, 2012).

Na fase da elaboração das regras, além dos conjuntos de itens frequentes, é requerida uma confiança mínima que servirá de parâmetro na seleção das regras (GOLDSCHMIDT; PASSOS, 2005). Com esses requisitos, para cada k-*itemset* tal que $k >= 2$, as possíveis combinações para o termo antecedente X e para o consequente Y são determinadas e então é calculada a confiança da regra $X \rightarrow Y$, caso seja maior ou igual ao parâmetro fornecido, a regra é selecionada como válida (AGRAWAL; SRIKANT, 1994).

Além das clássicas métricas de suporte e confiança, vários trabalhos propuseram outros parâmetros a serem utilizados na seleção e na avaliação de regras associativas, como o *lift*, e a convicção. O *lift* é um indicador de interesse para uma regra, pode ser calculada por

meio da equação (3). Considera-se que quanto maior o suporte do termo consequente menos interessante é uma norma de associação. Também pode ser utilizada na avaliação do relacionamento entre os fatores de uma regra, em particular, um valor próximo de 1 indica uma independência estatística entre X e Y (BRIN *et al.*, 1997).

$$lift(X \rightarrow Y) = \frac{confiança(X \cup Y)}{suporte(Y)} \quad \textbf{(3)}$$

A convicção, por sua vez, é uma métrica que avalia a dependência entre o antecedente e o consequente, um alto valor de convicção significa que Y é muito dependente de X, e que esse ocorre raramente sem ocasionar aquele e, da mesma forma que o *lift,* um valor entorno do unitário indica uma frágil relação entre os fatores (BRIN *et al.*, 1997). A expressão (4) permite o cálculo da convicção.

$$convicção(X \rightarrow Y) = \frac{1 - suporte(Y)}{1 - confiança(X \rightarrow Y)} \quad \textbf{(4)}$$

2.3.8 Interpretação

Na fase de interpretação, finalmente, as informações obtidas na fase de mineração são analisadas para que o objetivo final do processo de KDD seja alcançado: o conhecimento (CARVALHO; TSUNODA, 2018). Entretanto, esse estágio não significa o fim do procedimento necessariamente, uma vez que, segundo Prass (2016), dependendo da qualidade dos resultados, será necessário retornar a alguma fase anterior para que equívocos sejam corrigidos e finalmente os resultados obtidos sejam realmente valiosos.

Essas fases do processo de KDD, bem como todos os conceitos em torno desse procedimento apresentados nesta seção, foram fundamentais no tocante a elaboração e desenvolvimento da metodologia adotada neste trabalho, bem como a interpretação dos resultados posteriormente auferidos. Já na próxima seção deste documento (Metodologia), será perceptível a relevância de toda essa teoria então abordada.

3. METODOLOGIA

Este capítulo aborda os dados coletados, as principais ferramentas utilizadas e as tecnologias necessárias para a realização da pesquisa, também mostra a metodologia adotada e o processo usado para a execução do processo de KDD.

3.1 Tipo de Pesquisa

Por ter como objetivo principal analisar os acidentes ocorridos nas rodovias federais do Rio Grande do Norte através do processo de descoberta e conhecimentos em bases de dados, a presente pesquisa pode ser considerada de abordagem quantitativa. Visto que a metodologia adotada em sua efetivação, bem como o comportamento adotado diante do objeto de estudo e dos resultados, estão em conformidade com particularidades desse tipo de investigação.

> Diferentemente da pesquisa qualitativa, os resultados da pesquisa quantitativa podem ser quantificados. Como as amostras geralmente são grandes e consideradas representativas da população, os resultados são tomados como se constituíssem um retrato real de toda a população alvo da pesquisa. A pesquisa quantitativa se centra na objetividade. Influenciada pelo positivismo, considera que a realidade só pode ser compreendida com base na análise de dados brutos, recolhidos com o auxílio de instrumentos padronizados e neutros. A pesquisa quantitativa recorre à linguagem matemática para descrever as causas de um fenômeno, as relações entre variáveis, etc (Fonseca, 2002, p. 20).

3.2 Dados Utilizados

A fonte de dados elementar explorada neste trabalho foi o registro de acidentes disponibilizado, abertamente, pela Polícia Rodoviária Federal (PRF) em seu portal na *internet*: www.prf.gov.br.[5] Segundo a própria página, os dados ali publicados estão isentos de qualquer restrição de licenças, patentes ou mecanismos de controle. Dessa forma, tratam-se de dados abertos e, devido a isso, são convenientes para pesquisas. Nesse portal são encontradas

[5] Informações retiradas do portal da PRF - Disponivel em: <https://www.prf.gov.br/portal/dados-abertos/>. Acesso em: 14 Julho 2019.

bases de dados sobre acidentes ocorridos em rodovias e sobre infrações de trânsito, conquanto, apenas a primeira foi utilizada na metodologia desse trabalho. Ademais, os dados sobre acidentes se encontram agrupados de duas maneiras:

- **Acidentes agrupados por ocorrência:** cada instância representa a ocorrência de único acidente, dados sobre as circunstâncias do acontecimento estão presentes, mas não há campos sobre os dados pessoais dos envolvidos, apenas quantificações quanto ao número de feridos, mortos, etc.
- **Acidentes agrupados por pessoas:** várias linhas presentes na base podem representar o mesmo acidente, pois os registros se referem a cada pessoa envolvida no acidente individualmente. Dados pessoais, além dos associados às circunstâncias e aos veículos, como o sexo e idades estão presentes nesse conjunto.

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

Organizada cronologicamente desde 2007 até o ano atual (2019), a fonte de dados escolhida conta com uma quantidade de atributos consideradamente ampla, em grande parte, do tipo qualitativo. Convenientemente, as autoridades que fornecem esses dados também disponibilizam, na mesma página inclusive, dicionários que incluem informações sobre todos esses atributos, ou seja, metadados. Ao estudá-los, verificou-se que, a partir de 2017 houve uma descontinuidade na metodologia utilizada para a coleta desses dados. Antes desse ano, segundo esses arquivos, o registro de acidentes era realizado por meio do sistema BR-Brasil que foi descontinuado e, desde então, um novo sistema foi utilizado para registro das ocorrências. Por conseguinte, os conjuntos de dados a partir de 2017 apresentam alguns atributos a mais em relação aos registros anteriores.

O quadro 1, apresenta e resume todas as variáveis presentes na base de dados escolhida nos registros anteriores a 2017, outrossim, o quadro 2 inclui as novas propriedades inseridas a partir desse ano.

Quadro 1 - Variáveis do registro de acidentes (2007 - 2016).

Variável	Descrição	Exemplos
id	Identificação única para cada acidente	173746,182345
data_inversa	Data da ocorrência	20/01/2007, 25/10/2015
dia_semana	Dia da semana da ocorrência	Segunda-feira, Sábado.

horario	Horário da ocorrência	16:42:00, 21:15:00
uf	Unidade da Federação.	RJ, SP, DF, RN.
br	Representa o identificador da BR do acidente.	101, 304.
km	Identificação do quilômetro onde ocorreu o acidente na rodovia.	623.2, 305.5.
municipio	Nome do município de ocorrência do acidente	Fortaleza, Recife, Natal.
causa_acidente	Causa presumível do acidente.	Falta de atenção, Velocidade incompatível.
tipo_acidente	Identificação do tipo de acidente.	Colisão frontal, Saída de pista.
classificação_acidente	Classificação quanto à gravidade do acidente.	Sem Vítimas, Com Vítimas Feridas.
fase_dia	Fase do dia no momento do acidente.	Amanhecer, Pleno dia.
sentido_via	Sentido da via considerando o ponto de colisão.	Crescente e decrescente.
condição_meteorologica	Condição meteorológica no momento do acidente.	Céu claro, chuva.
tipo_pista	Tipo da pista considerando a quantidade de faixas.	Dupla, simples ou múltipla.
tracado_via	Descrição do traçado da via.	Reta, Curva.
uso_solo	Descrição sobre as características do local do acidente.	Urbano ou rural.
id_veiculo	Identificando única de cada veículo envolvido nos acidentes registrados.	328381, 424452.
tipo_veiculo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro.	Automóvel, Caminhão.
marca	Descrição da marca e modelo do veículo.	GM/Celta, Ford/KA
ano_fabricacao_veiculo	Ano de fabricação do veículo.	2005, 2001
pesid	Identificação única para cada pessoa envolvida.	173715, 175351
tipo_envolvido	Tipo de envolvido no acidente conforme sua participação no evento.	Condutor, Passageiro.
estado_fisico	Condição do envolvido conforme a gravidade das lesões.	Morto, Ferido.
idade	Idade do envolvido.	31,40
sexo	Sexo do envolvido.	Masculino ou Feminino.
nacionalidade	Indica a nacionalidade do envolvido.	Brasil, Argentina.
naturalidade	Indica a naturalidade do envolvido.	Natal, Fortaleza.

Fonte: Polícia Rodoviária Federal (2019).

Quadro 2 - Novas variáveis inseridas a partir de 2017.

Variável	Descrição	Exemplos
causa_principal	Identifica se a causa do acidente foi identificada como principal pelo policial.	Sim ou Não
ordem_tipo_acidente	Valor numérico que identifica a sequência dos eventos sucessivos que ocorreram no acidente.	1, 2
Ilesos	Valor binário que identifica se o envolvido foi classificado como ileso.	Sim ou Não
feridos_leves	Valor binário que identifica se o envolvido foi classificado como ferido leve.	Sim ou Não
feridos_graves	Valor binário que identifica se o envolvido foi classificado como ferido grave.	Sim ou Não
Mortos	Valor binário que identifica se o envolvido foi classificado como morto.	Sim ou Não
Latitude	Latitude do local do acidente em formato geodésico decimal.	-27,8101, -6,4622
Longitude	Longitude do local do acidente em formato geodésico decimal.	-49,20166969, 36,1899

Regional	Setor Regional da PRF que registrou o ocorrido	SR-GO, SR-CE
Uop	Unidade Operacional da Polícia que registrou o ocorrido	UOP03/RS, UOP01/BA
delegacia	Delegacia em que o acidente foi registrado	Colisão frontal, Saída de pista.

Fonte: Polícia Rodoviária Federal (2019).

No total, registros de doze anos diferentes foram selecionados e salvos em um computador pessoal para a realização da pesquisa: todos referentes aos anos entre 2007 e 2018. Além disso, por se tratar de um conjunto ainda incompleto, optou-se por não utilizar os dados registrados no ano de 2019. Todos esses dados obtidos foram armazenados em doze arquivos distintos com a extensão ".CSV" (*Comma-separated values*), formato de arquivo semelhante a uma planilha.

3.3 Tecnologias e Ferramentas Utilizadas

Atualmente, várias ferramentas e tecnologias para processamento e mineração de dados, bem como para se trabalhar com aprendizagem de máquina estão disponíveis, muitas das quais são completamente livres e gratuitas, inclusive. Trata-se de *softwares* especializados como o *Weka* e o *Rapidminer* ou de linguagens de programação como *R* e *Python* com suas extensas bibliotecas de funções. Percebe-se que as ferramentas do segundo tipo são cada vez mais preferíveis pelos profissionais e estudiosos desse âmbito, uma vez que são mais flexíveis e possibilitam, com maior facilidade, a implementação dos resultados no desenvolvimento de aplicações e sistemas. A linguagem de programação *Python,* especificamente, vem ganhando notoriedade, segundo o *Kaggle*, o maior site do mundo que promove competições de análise dados e aprendizagem de máquina, em uma pesquisa realizada em outubro de 2018, cerca de 65% de seus usuários afirmaram utilizar *Python* regularmente em seus trabalhos. Essa popularidade se deve a vários prós que serão discutidos na próxima subseção. Ademais, essas vantagens também justificaram a adoção dessa linguagem como a principal ferramenta utilizada neste trabalho.

3.3.1 A Linguagem de Programação *Python*

A linguagem Python foi desenvolvida em 1991 pelo matemático e programador holandês Guido Van Rossum, segundo a própria documentação oficial, é uma linguagem de programação interpretada (não compilada), portável, de alto nível, multiparadigma e de

propósito geral. Devido à eficiência e precisão no tratamento de tipos numéricos, essa linguagem progressivamente conquistou notoriedade quanto ao desenvolvimento de aplicações científicas e de análises de dados (BOSCHETTI; MASSARON, 2015). Atualmente, devido a extensa quantidade de poderosas bibliotecas de funções (pacotes), *Python* pode ser considerada uma plataforma completa e eficiente para processamento, extração de informações e exploração de grandes conjuntos de dados (MASSARON. MUELLER, 2015). No tocante, especificamente, às vantagens que levaram a escolha dessa linguagem, podem ser citadas:

- **Simplicidade**: característica que torna o tempo necessário para o aprendizado dessa linguagem minimamente curto e sua utilização prazerosa;
- **Código aberto**: *Python* é mantida pela *Python Software Foundation* (PSF), uma comunidade sem fins lucrativos que licencia a linguagem como uma tecnologia de código aberto e gratuita;
- **Disponibilidade de pacotes**: pacotes (bibliotecas e extensões) como *Numpy, Matplotlib*, Pandas e vários outros estendem as funcionalidades da linguagem e tornam tarefas comumente requeridas na análise e mineração de dados, demasiadamente simples de serem feitas;
- **Gerenciador de pacotes**: *Python* conta com poderoso gerenciador de pacotes o PyPI, muitas vezes apenas um único comando é necessário para que uma biblioteca seja corretamente instalada;
- **Popularidade**: devido à popularidade dessa linguagem, a quantidade de livros, manuais e vídeos disponíveis sobre essa é muita ampla, isso, evidentemente, facilita sua utilização.

Uma distribuição mundialmente conhecida que inclui o *Python* e praticamente todos os pacotes necessários para se trabalhar com análise de dados, mineração e aprendizagem de máquina é o “Anaconda”. Segundo sua página oficial na internet, a distribuição conta com mais de 1500 pacotes de qualidade garantida que servem a esse propósito, alguns dos quais já são pré-instalados e outros podem ser obtidos conforme a necessidade. Na execução desse trabalho, muitas dessas extensões foram utilizadas, as quais serão brevemente abordadas nas seguintes subseções.

3.3.2 *Numpy*

Scipy é uma biblioteca para *Python* que inclui uma coleção de ferramentas para programação de aplicações científicas. Dentre essas a mais importante é a *Numpy* (*Numerical Python*), módulo que implementa uma gama de funções e operações matemáticas essenciais como trigonométricas, estatísticas e vetoriais e que oferece uma estrutura de dados com um desempenho de mais alto nível: o *ndarray* (matriz multidimensional), que, segundo a documentação, suporta não apenas tipos numéricos, mas também simbólicos como cadeias de caracteres (*strings*) e valores booleanos. *Numpy,* mais do que importante, vem se tornando essencial, uma vez que serve de base para praticamente todos os outros módulos do *Python* que necessitam realizar cálculos matemáticos eficientemente. Pandas e *Matplotlib*, que ainda serão apresentadas, são exemplos de bibliotecas dependentes do *Numpy.*

3.3.3 Pandas

Pandas é um pacote completo e muito poderoso para manipulação e exploração de dados. Construída a partir do *Numpy,* essa biblioteca oferece uma coleção de funções ainda mais ampla, que tornam o processamento de grandes conjuntos de dados uma tarefa muito mais simples e rápida de ser executada, além disso, conta com estruturas de dados ainda mais robustas, ricas em funcionalidades e que suportam tipos de dados muito mais complexos do que os tradicionais *ndarrays*: os *DataFrames* e as *Series*.

O Pandas foi, seguramente, a ferramenta mais utilizada durante todo o processo de KDD desenvolvido neste trabalho, a economia de tempo promovida devido sua utilização, sobretudo na etapa de pré-processamento, realmente a legitimou como uma tecnologia de qualidade impar e justificou sua escolha.

3.3.4 *Matplolib e Seaborn*

Ao se trabalhar com dados, uma tarefa frequentemente necessária é a construção de figuras e gráficos, haja vista que esses recursos visuais refletem, de maneira muito mais eficiente, as informações contidas nesses. Dentre várias ferramentas disponíveis para construção de visualizações na plataforma *Python*, optou-se pelas bibliotecas *Matplotlib* e *Seaborn.* A primeira concerne, praticamente, em um ecossistema gráfico completo. A extensa quantidade de funções ofertadas por essa biblioteca, muito semelhantes às do *software* MATLAB, permite a construção de quase que todo tipo de visualização como histogramas, gráficos de setores, diagramas de dispersão dentre vários outras. Todavia, a quantidade de

código necessária para a construção de plotagens complexas pode se tornar extensa ao se utilizar apenas a *Matplolib,* e isso justificou a utilização da *Seaborn* em casos específicos, uma biblioteca construída a partir da primeira, que permite desenvolver gráficos estatísticos complexos de uma maneira muito mais simples.

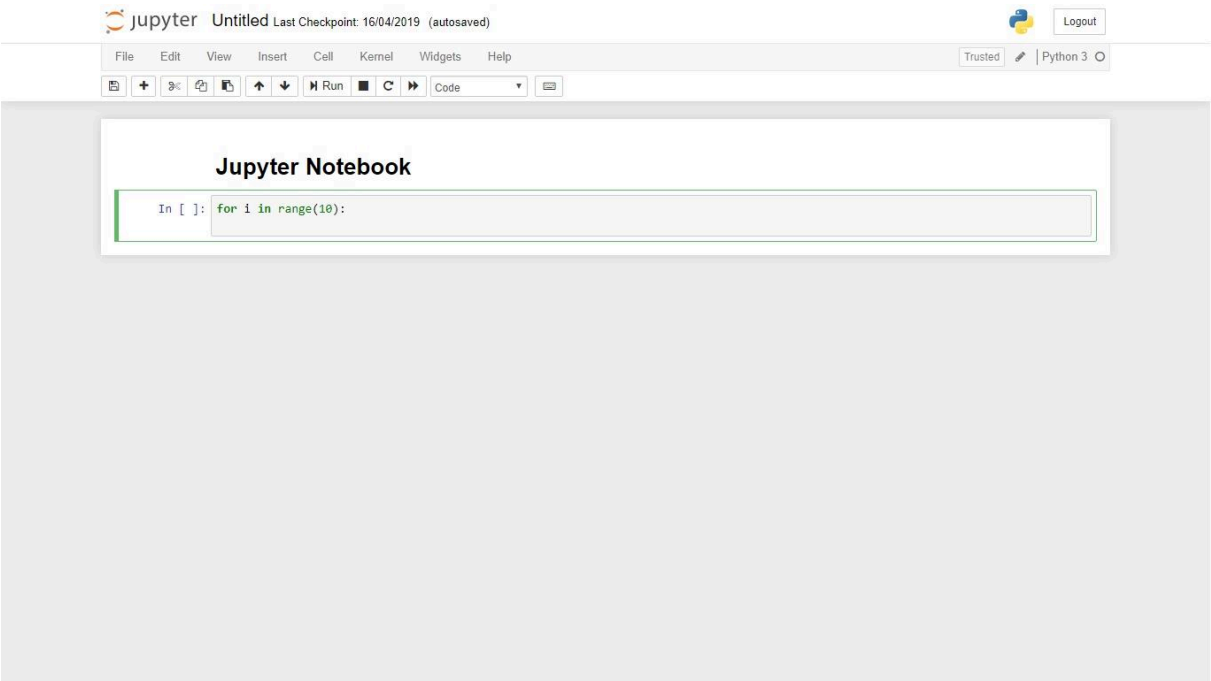
3.3.5 *MLxtend*

A plataforma *Python,* atualmente, talvez contenha as mais completas e poderosas ferramentas gratuitas para se trabalhar com aprendizagem de máquina e inteligência artificial. Dentre as tais, destaca-se as bibliotecas *Scikit-Learn* e o *TensorFlow*, amplamente utilizadas em projetos pessoais, científicos e coorporativos. Todavia, essas não suportam totalmente algoritmos de associação como o *Apriori* e, devido a isso, na presente pesquisa foi feito o uso de uma alternativa menos popular, mas de qualidade análoga: a *MLxtend* (*machine learning extensions*). Segundo a página oficial dessa biblioteca na *internet*, a *MLxtend* implementa uma variedade de utilitários que, em verdade, complementa o pacote *Scikit-Learn*, para a finalidade dessa pesquisa, no entanto, somente os recursos oferecidos por essa se demonstraram suficientes, uma vez que, a implementação do algoritmo *Apriori* se mostrou tão eficiente e completa, quanto a disponível em conceituados *softwares* como o *Weka.*

3.3.6 *Jupyter Notebook*

O *Jupyter Notebook* não se refere a mais uma das bibliotecas disponíveis para a linguagem *Python*, mas a um ambiente de desenvolvimento amigável e rico em recursos que torna todo o processo de análise e mineração de dados muito mais confortável e produtivo. Com uma interface limpa e intuitiva, essa ferramenta organiza, da melhor maneira, blocos de códigos e gráficos desenvolvidos como um documento e ainda possibilita a inserção de comentários, imagens, vídeos e até mesmo de equações matemáticas para uma melhor descrição sobre esses. Tudo isso torna os resultados obtidos e os procedimentos utilizados muito mais apresentáveis. A Figura 4 demonstra a *interface* do *Jupyter Notebook*. Vale salientar ainda que esse ambiente é compatível com outras linguagens de programação além do *Python*, como a linguagem *R*, por exemplo.

Figura 4 - Ambiente de desenvolvimento (*Jupyter Notebook*)



3.4 Aplicação do Processo de KDD

Antes mesmo da preparação dos dados, certas configurações foram necessárias. A primeira dessas se refere à reunião dos arquivos que contém os dados escolhidos em uma única pasta digital denominada “acidentes”, conforme a figura 5. Ao total, cerca de 1,25 GB (*Giga Bytes*) de dados brutos foram coletados, divididos em 12 arquivos distintos conforme o ano em que os registros foram coletados. Com os dados reunidos, um *notebook* (Arquivo do *Jupyter Notebook*) foi criado, todos os comandos, tabelas, comentários e gráficos utilizados foram registrados e executados nesse arquivo. A figura 6 demonstra o primeiro comando executado: a importação dos pacotes instalados, necessário para que as funções, provenientes desses, possam ser executadas.

Figura 5 - Arquivos armazenados

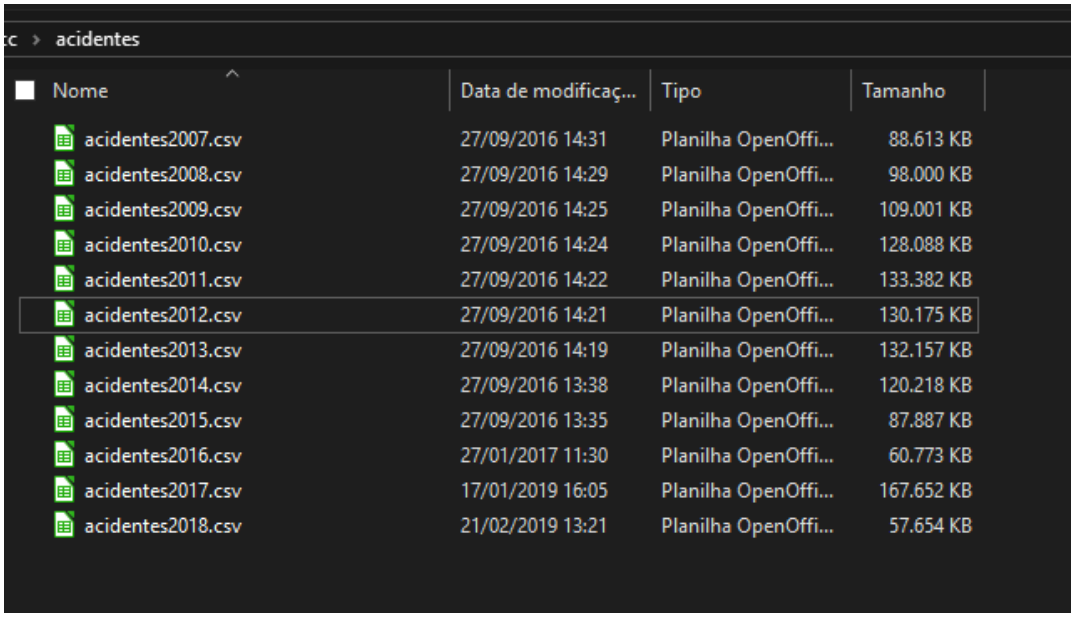


Figura 6 – Importação das funções

```
# importações
import pandas as pd, seaborn as sns,matplotlib.pyplot as plt, mlxtend.frequent_patterns, mlxtend.preprocessing, os,warnings
```

Percebe-se que o *Numpy* não foi importado explicitamente e isso seria desnecessário, pois o Pandas e o *Matplotlib* já utilizam internamente essa biblioteca. Em contrapartida, os módulos denominados *"os" e "warnings"* foram inseridos. Essas funções se referem utilitários nativos e pouco relevantes que foram utilizados somente para ler os arquivos de dados e para controlar avisos de erros. Com o ambiente devidamente configurado e os dados convenientemente armazenados, iniciou-se, efetivamente, o processo de descoberta de conhecimentos nessa base de dados. As próximas subseções tratarão das etapas desse processo, desde a seleção de dados e atributos até a mineração. Já a etapa de análise será discutida na seção em que os resultados serão apresentados.

3.4.1 Seleção

Nesse primeiro estágio, a primeira tarefa realizada foi a seleção dos registros de acidentes que ocorreram apenas no estado do Rio Grande do Norte, visto que, esse trabalho tem por interesse estudar apenas essa amostra. Ademais, após uma verificação minuciosa, apenas uma fração dos atributos presentes no conjunto de dados foi considerada relevante ou

adequada. Sendo assim, muitas variáveis, por motivos distintos, não foram utilizadas após essa etapa. O quadro 3 resume quais desses campos foram eliminados e as justificativas dessa ação para cada caso.

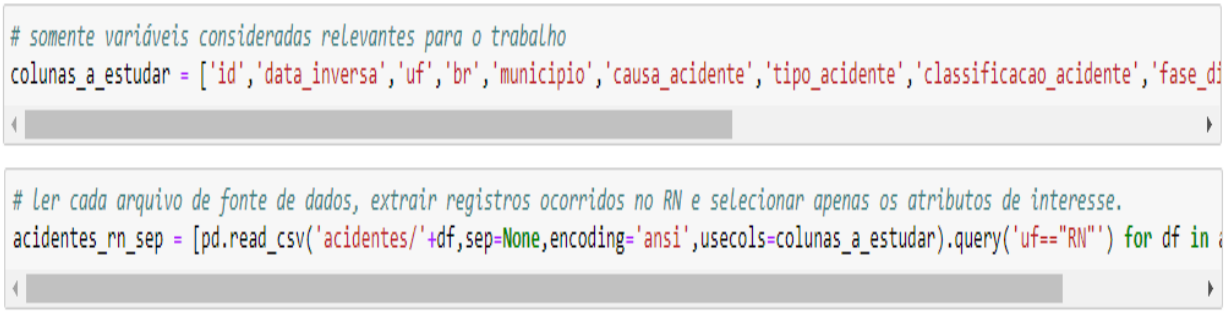
Cabe salientar que, um parâmetro relevante, porém não determinante, na escolha das variáveis diz respeito ao seu tipo: campos qualitativos são preferíveis, uma vez que a técnica de mineração utilizada (regras de associação) é aplicável apenas aos dados dessa natureza, além disso, priorizaram-se as propriedades associados às circunstâncias que levaram ao acidente bem como as características dos indivíduos envolvidos.

Quadro 3 - Variáveis Eliminadas

Variáveis	Justificativa
dia_semana	Consideradas redundantes em relação a outras variáveis, as quais foram preferíveis por conveniência ou por possuírem mais informações agregadas.
horario	
estado_fisico	
Mortos	
feridos_graves	
feridos_leves	
Ilesos	
km	Apesar ser de interesse a localização das ocorrências em relação aos estados e municípios, uma posição mais específica não é necessária para os fins dessa análise.
Latitude	
Longitude	
sentido_via	As informações intrínsecas a esses campos não foram consideradas relevantes.
id_veiculo	
Marca	
ano_fabricacao_veiculo	
tipo_veiculo	
Pesid	
Nacionalidade	
Naturalidade	
Delegacia	
Uop	
Regional	
tracado_via	Possuem valores incompatíveis entre as bases de dados, devido à mudança no sistema de registros de acidentes, foi preferido elimina-los ao se ter que lidar com várias inconsistências.
causa_principal	
ordem_tipo_acidente	

A figura 7 demonstra os comandos utilizados nesse processo inicial. Percebe-se que ao invés da remoção de cada um dos atributos discutidos, foi feita uma seleção de fato, na qual apenas as variáveis julgadas relevantes foram carregadas a partir da base de dados (terceiro comando presente na figura). É importante esclarecer ainda que mesmo essas variáveis pré-selecionadas não foram, em totalidade, mantidas durante todo o procedimento de preparação.

Figura 7 - Seleção de dados e atributos



3.4.2 Pré-Processamento

A etapa de pré-processamento de dados, como já esperado, foi a que mais demandou tempo e esforços dentre todas as demais. Mesmo na fonte de dados utilizada, de qualidade incontestável, muitos problemas foram encontrados, como dados com ruídos, redundâncias, má formatações e inconsistências, por consequência, a principal meta nesta fase, foi, justamente, o tratamento dessas perturbações.

A primeira atividade realizada nesse estágio não consistiu já em uma correção, mas na integração dos dados que estavam armazenados em diferentes arquivos, especificamente, esses foram compilados em um *DataFrame*[6] único. Isso permitiu que manipulações fossem facilmente realizadas. A figura 8 exibe os códigos desenvolvidos para integrar os vários registros e também demonstra as primeiras instâncias presentes no *DataFrame* resultante.

[6] Principal estrutura de dados fornecida pelo Pandas, semelhante à uma tabela virtual.

Figura 8 - Integração dos dados

```
# integração de dados
acidentes_rn = pd.concat(acidentes_rn,ignore_index=True)
```

	id	data_inversa	uf	br	municipio	causa_acidente	tipo_acidente	classificacao_acidente	fase_dia	condicao_metereologica	tipo_pista	uso_s
0	173746.0	01/01/2007	RN	101.0	EXTREMOZ	Falta de atenção	Saída de Pista	Sem Vítimas	Plena noite	Ceu Claro	Simples	Ru
1	173937.0	01/01/2007	RN	304.0	MACAIBA	Outras	Colisão com objeto móvel	Com Vítimas Feridas	Plena noite	Ceu Claro	Simples	Ru
2	173943.0	01/01/2007	RN	101.0	SAO GONCALO DO AMARANTE	Não guardar distância de segurança	Colisão traseira	Sem Vítimas	Anoitecer	Ceu Claro	Dupla	Urb
3	173943.0	01/01/2007	RN	101.0	SAO GONCALO DO AMARANTE	Não guardar distância de segurança	Colisão traseira	Sem Vítimas	Anoitecer	Ceu Claro	Dupla	Urb
4	173965.0	01/01/2007	RN	101.0	PARNAMIRIM	Ingestão de álcool	Colisão traseira	Sem Vítimas	Pleno dia	Ceu Claro	Dupla	Ru

Após a fusão dos dados, foram utilizadas funções do pacote Pandas para encontrar possíveis valores ausentes entre os registros. Inicialmente, apenas três campos foram identificados com esse problema: a coluna “br”, “idade” e “sexo”; posteriormente, percebeu-se que vários outros atributos continham valores que não refletiam qualquer informação, portanto, também nulos, dados desse tipo foram encontrados nas colunas “fase_dia”, “condicao_metereologica”, “classificacao_acidente”, dentre outras. Na figura 9 é possível observar os campos que foram automaticamente identificados como problemáticos devido à ausência de dados, já os demais, com o problema análogo, foram reconhecidos por meio de uma inspeção manual dos valores assumidos, na mesma figura inclusive, demonstra-se como essa inspeção foi realizada para o atributo “fase_dia”.

Figura 9 - Valores ausentes

```
# verificar valores ausentes
acidentes_rn.isnull().any()
```

```
id                          False
data_inversa                False
uf                          False
br                           True
municipio                   False
causa_acidente              False
tipo_acidente               False
classificacao_acidente      False
fase_dia                    False
condicao_metereologica      False
tipo_pista                  False
uso_solo                    False
tipo_envolvido              False
idade                        True
sexo                         True
dtype: bool
```

```
# exibir as fases do dias
acidentes_rn['fase_dia'].sort_values().unique()
```

```
array(['(Null)', 'Amanhecer', 'Anoitecer', 'Plena Noite', 'Pleno Dia'],
      dtype=object)
```

Ainda sobre a figura 9, percebe-se que a etiqueta “True” próximo ao campo indica a ausência de valores nesse, já em relação ao atributo “fase_dia”, no primeiro quadro não há indício de registros faltosos, todavia, ao se observar os valores assumidos por esse campo no segundo quadro, nota-se a categoria ‘(Null)’, a qual, evidentemente, não se refere a qualquer informação relevante. Vale salientar que essa situação ocorre, pois no último caso todos os registros realmente estão preenchidos, mas a natureza de algumas instâncias é a mesma referente a registrados não presentes.

No tocante ao tratamento de valores faltosos, seguiram-se nesse trabalho os procedimentos propostos por Han et. al., (2012), tais como eliminação e preenchimento, entretanto, foi preferível, em toda a etapa de pré-processamento manter todos os registros. Dessa forma, os dados ausentes foram corrigidos por meio de preenchimento, alguns por meio de técnicas matemáticas, como foi o caso do campo ‘br’, no qual se inferiu em qual rodovia o acidente ocorreu por meio da moda estatística calculada para cada município; outros campos foram preenchidos por meio de constantes que representassem objetos desconhecidos como a categoria “Ignorado”; já em poucos casos, foi possível inserir o valor em falta por meio de uma consulta à base de dados original, como foi a situação da variável “fase_dia”, uma vez que no registro, antes da etapa de seleção, havia o campo com a hora exata da ocorrência.

Em relação à presença de dados com ruídos, poucos registros com esse problema foram encontrados, uma vez que os atributos trabalhados eram predominantemente do tipo qualitativo, por conseguinte, menos susceptíveis a essas anomalias. De fato, deparou-se com

valores destoantes significativos apenas na coluna que representava a idade dos indivíduos envolvidos nos acidentes (atributo quantitativo discreto). Como se observa na figura 10, pessoas com supostamente 128, 906 e até 1916 anos foram catalogadas, nitidamente, esses *outliers* são oriundos de erros e por isso foram tratados.

Ainda na mesma imagem, percebe-se a ocorrência de idades com valor "-1", essas instâncias não refletem erros, mas registros desconhecidos. Na condição dessa variável, os dados discrepantes foram somente ignorados, idades acima de 110 anos tiveram valores substituídos por "-1" que representaria desconhecimento a respeito. Essa atitude se justificou, pois era ínfima a quantidade de casos com esse impasse.

Figura 10 - Valores com ruídos

```
# observar valores de idades
acidentes_rn['idade'].unique()
```

```
array([  40,   23,   -1,   32,   24,   28,   42,   33,   26,   21,   55,
         20,   71,   49,   41,   46,   58,   63,   51,   15,   17,   44,
         25,   35,   45,   34,   68,   31,   27,   66,   47,   14,   19,
         48,   50,   43,   22,   29,   64,   30,   38,   36,   52,   37,
         53,   70,   54,   59,   65,   39,   56,   57,    5,   60,   61,
         18,    0,   13,   69,   62,   67,   16,   77,   76,   10,   11,
         12,   75,    9,   80,   81,    2,   90,   86,    6,    3,   74,
         78,   72,    7,    4,   73,   79,   88,    1,    8,   84,   82,
         87,   91,   83,  109,   94,   85,  125,   89,  111,  127,  106,
         95,  128,  906, 1916,  117,  939], dtype=int64)
```

Sobre a ocorrência de inconsistências, alguns atributos, de fato, foram considerados conflitantes. O quadro do atributo "br" foi o mais intrigante, de acordo com os dados coletados (Figura 11), houveram acidentes que ocorreram em rodovias federais que não se situam no estado do RN (BR 343) ou que se quer existem (BR 268 e BR 501). Isso foi confirmado ao se realizar uma pesquisa no mapa rodoviário disponibilizado pelo DNIT. Incoerências dessa natureza foram solucionadas de maneira análoga aos demais problemas encontrados, especificamente, na situação da variável "br", utilizou-se, mais uma vez, a moda estatística das rodovias para cada município como recurso de inferência, visto que Côrtes, Porcaro e Lifschitz, (2002) garantem a validade desse método também para esse caso.

Durante a realização da limpeza dos dados, foi perceptível, que a identificação de problemas relacionados a inconsistências representa uma tarefa mais desafiadora, devido, sobretudo, à recorrente necessidade de se coletar informações extrínsecas aos dados em análise.

Figura 11 - Dados inconsistentes

```
# observar resultado
acidentes_rn['br'].unique()
```

```
array(['101', '304', '226', '110', '406', '405', '427', '343', '501',
       '268', '104'], dtype=object)
```

Dados em falta, contendo ruídos ou inconsistentes instituem problemas ordinariamente deparáveis em qualquer fonte de dados, de tal maneira que a etapa de pré-processamento praticamente consiste no tratamento desses. Contudo, os impasses mais frequentemente encontrados no registro de acidentes da PRF se deram por causa da incompatibilidade entre valores e atributos nas bases de dados de anos distintos.

Como já discutido, após 2017 o órgão responsável pelo conjunto de dados em estudo alterou o sistema de registros utilizado até então, isso refletiu em novos atributos e novas categorias ou em alterações presentes nos registros de 2017 e 2018 e, consequentemente, em problemas de assimetria e redundâncias em relação aos conjuntos de anos anteriores.

Como exemplo, na figura 12 é possível verificar as possíveis causas que levaram aos acidentes, compiladas a partir de todos os anos em estudo, nota-se uma quantidade ampla de valores atribuídos. Entretanto, nos registros até 2016, constam-se apenas onze causas distintas, já no conjunto de 2018 vinte e quatro diferentes foram consideradas, ou seja, muitas categorias demonstradas não estão presentes na maior parcela de dados.

Figura 12 - Valores incompatíveis e redundantes

```
# exibir as causas de acidentes compiladas a partir dos dados
acidentes_rn['causa_acidente'].sort_values().unique()
```

```
array(['Agressão Externa', 'Animais Na Pista',
       'Avarias E/Ou Desgaste Excessivo No Pneu',
       'Carga Excessiva E/Ou Mal Acondicionada', 'Condutor Dormindo',
       'Defeito Mecânico Em Veículo', 'Defeito Mecânico No Veículo',
       'Defeito Na Via',
       'Deficiência Ou Não Acionamento Do Sistema De Iluminação/Sinalização Do Veículo',
       'Desobediência À Sinalização',
       'Desobediência Às Normas De Trânsito Pelo Condutor',
       'Desobediência Às Normas De Trânsito Pelo Pedestre', 'Dormindo',
       'Falta De Atenção', 'Falta De Atenção Do Pedestre',
       'Falta De Atenção À Condução', 'Fenômenos Da Natureza',
       'Ingestão De Substâncias Psicoativas', 'Ingestão De Álcool',
       'Ingestão De Álcool E/Ou Substâncias Psicoativas Pelo Pedestre',
       'Mal Súbito', 'Não Guardar Distância De Segurança',
       'Objeto Estático Sobre O Leito Carroçável', 'Outras',
       'Pista Escorregadia', 'Restrição De Visibilidade',
       'Sinalização Da Via Insuficiente Ou Inadequada',
       'Ultrapassagem Indevida', 'Velocidade Incompatível'], dtype=object)
```

Uma vez que a causa dos acidentes foi considerada um parâmetro muito relevante para o estudo, a simples remoção do atributo não foi uma opção cogitada, desse modo outros métodos foram aplicados na resolução desse quadro. Em primeiro momento, foram percebidas muitas categorias que, essencialmente, são iguais (redundantes), como “Defeito Mecânico Em Veículo” e “Defeito Mecânico No Veículo” ou “Dormindo” e “Condutor Dormindo”, para esses casos, simplesmente alguns valores redundantes foram renomeados; já em certas circunstâncias generalizações foram feitas, como exemplo, “Desobediência À Sinalização”, “Desobediência Às Normas de Trânsito Pelo Condutor” e “Desobediência Às Normas de Trânsito Pelo Pedestre” foram todas categorias consideradas como “Desobediência Às Normas de Trânsito”; em casos em que a incompatibilidade de valores não pôde ser solucionada, genericamente se atribuiu a categoria “Outras” que representaria causas não consideradas especificamente. Apesar de apenas o atributo “causa_acidente” aqui ser exemplificado, muitos outros também apresentaram problemas similares como “tipo_acidente”, “sexo” e “uso_solo”, as soluções adotadas também foram análogas para esses.

Além da resolução de erros e empecilhos, nessa etapa do trabalho também foram feitas algumas alterações nos dados, como correções de títulos de categorias e padronização de valores. Algumas variáveis foram também removidas já outras criadas, por exemplo, criou-se a uma variável que continha o mês do ocorrido a partir da data completa, já o campo que indicava o estado onde o acidente ocorreu foi eliminado, pois se tornou redundante. Já em relação ao número de instâncias, preocupou-se em manter todos os registros, haja vista que na próxima etapa realizada, uma exploração íntegra e quantitativa sobre os dados foi realizada.

3.4.3 Exploração preliminar dos dados

Essa fase do desenvolvimento do trabalho não faz parte da estrutura clássica do processo de KDD apresentada, isso, de maneira alguma, comprometeu esse procedimento dinâmico, mas o complementou, uma vez que foi considerado necessário realizar uma visualização das principais relações presentes nos dados. Dessa forma, o principal objetivo dessa etapa foi desenvolver recursos visuais como gráficos e tabelas com o intuito de explorar padrões, mesmo que superficiais, entre os atributos, a mineração posteriormente realizada possibilitou a identificação de informações mais profundas sobre esses, todavia, panoramas

gerais também foram considerados de interesse e isso justificou a realização desse estágio. Nesse nível, explorações quantitativas foram realizadas, por exemplo, o número de acidentes ocorridos em cada ano, esse foi o motivo pelo qual todos os registros foram mantidos na etapa de limpeza dos dados.

3.4.4 Transformação

Como proposto por várias bibliografias, destinou-se esse estágio de transformação a preparar definitivamente os dados para a aplicação da técnica de mineração. Incluem-se nas atividades realizadas nessa etapa: a eliminação de atributos e registros já não mais necessários, discretização do atributo numérico “idade” e a alteração na estrutura dos dados. Sobre os dados eliminados, após a exploração preliminar, não foram mais consideradas relevantes ou necessárias o ano em que o acidente ocorreu e nem a identificação dessa ocorrência, além disso, todos os registros com constantes que indicavam desconhecimento sobre algum atributo foram removidos. Essa limpeza permitiu que o algoritmo de extração de padrões fosse executado de maneira mais eficiente, além de evitar que a qualidade dos resultados fosse comprometida.

A variável “idade” continha valores numéricos, tipo de dados não conveniente para a técnica de mineração que seria realizada. Dessa maneira, uma discretização desses dados foi necessária. As idades foram então agrupadas em faixas etárias. A figura 13 demonstra os comandos necessários para a realização dessa tarefa e também os grupos de idades utilizados, esses grupos foram determinados ao fracionar o intervalo de idades assumidas aproximadamente de maneira igual em 18 e 17 anos.

Figura 13 - Discretização de dados

```
acidentes_rn.loc[:,'idade'] = pd.cut(acidentes_rn['idade'].astype('int'),6,labels=[
                                                    'criança ou adolescente de até 17 anos',
                                                    'jovem adulto entre 18 e 35 anos',
                                                    'adulto entre 35 e 53 anos',
                                                    'pessoa de idade avançada entre 53 a 70 anos',
                                                    'idoso entre 70 e 88 anos',
                                                    'idoso de idade avançada com mais de 88 anos'])
```

O último passo da preparação foi enfim alterar a estrutura dos dados para uma forma semelhante à tabela 1, ou seja, como um registro de transações, cujos valores são binários,

indicando a ocorrência ou não de uma categoria. Isso foi feito utilizando as ferramentas do pacote *MLxtend* conforme se pode observar na figura 14.

Figura 14 - Nova estrutura dos dados

```
# alterar estrutura dos dados
te = mlxtend.preprocessing.TransactionEncoder()

transformacao = te.fit(acidentes_rn.values).transform(acidentes_rn.values)

acidentes_rn_transformado = pd.DataFrame(transformacao,columns=te.columns_)

acidentes_rn_transformado.head()
```

	Abril	Acari	Acu	Agosto	Almino Afonso	Amanhecer	Angicos	Animais Na Pista	Anoitecer	Antonio Martins	...	Upanema	Urbano	Velocidade Incompatível	Vento	adulto entre 35 e 53 anos	criança ou adolescente de até 17 anos
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	True	False	False	True	False
4	False	False	False	False	False	False	False	False	False	False	...	False	True	False	False	False	False

3.4.5 Mineração

Enfim, a mineração dos dados foi feita com o intuito de se extrair padrões frequentes presentes na base de dados (associações), isso permitiu uma análise descritiva do panorama de acidentes ocorridos no RN. Dentre várias técnicas que poderiam ser utilizadas para esse fim, optou-se pelas regras de associação. Vários motivos levaram a essa decisão, dentre os quais se podem citar:

- **Simplicidade:** Todos os conceitos matemáticos associados à técnica de regras de associação são extremamente simples, portanto, compreensíveis facilmente;
- **Fácil interpretação:** Os resultados podem ser representados por regras do tipo "causa" e "consequência", de maneira que a interpretação se torna quase que imediata;
- **Conveniência:** Essa técnica é aplicável apenas a variáveis qualitativas, tipo de atributos, convenientemente, majoritários na base de dada estudada;
- **Aplicação em trabalhos anteriores:** Alguns trabalhos anteriores como o de Reis (2014) obtiveram bons resultados a partir da aplicação dessa técnica em registros de acidentes.

Estabelecida a técnica que seria utilizada, foram estudados os algoritmos que serviriam ao propósito, a opção natural foi pelo *Apriori*, uma vez que se demonstrou o mais importante, frequentemente referido em bibliografias e manuais como a melhor opção em situações

análogas a deste trabalho. Esse algoritmo então foi selecionado e aplicado ao conjunto de dados transformado e, conforme requerido por esse, foram definidos os valores que servirão de parâmetros para a avaliação das regras. Nesta mineração, o suporte mínimo requerido foi de 0,01 (1%) e a confiança 0,80 (80%), em verdade, esses parâmetros foram ajustados inúmeras vezes, essa combinação final, entretanto resultou em regras subjetivamente mais interessantes. Demais métricas como *lift* e convicção foram analisadas manualmente quando as associações mais relevantes foram selecionadas a partir do montante gerado. Por conveniência, essas serão apresentadas na próxima seção.

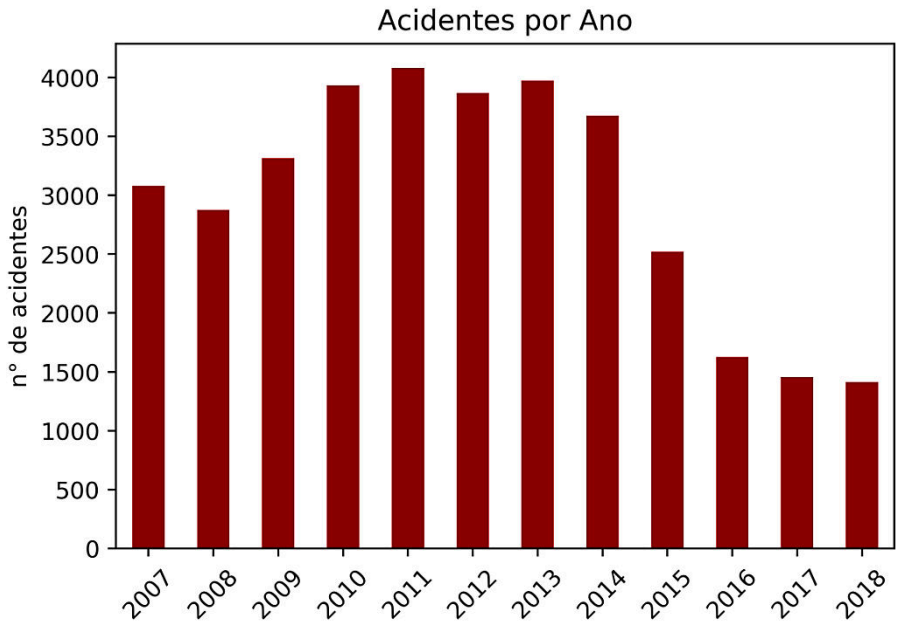
4. RESULTADOS E DISCUSSÕES

Nesta seção serão apresentados e analisados os resultados obtidos ao final do trabalho desenvolvido. Primeiramente, serão discutidos os gráficos e tabelas criados na etapa de exploração preliminar, na qual nenhuma técnica de aprendizagem e de mineração foi utilizada. Posteriormente, uma análise será feita sobre as regras de associação obtidas a partir da aplicação do algoritmo *Apriori.*

4.1 Resultados da Exploração Preliminar

Na base de dados da PRF, constataram-se um total de 78.356 registros referentes aos acidentes e as pessoas envolvidas nessas ocorrências dentro do estado do Rio Grande do Norte. Dessa quantia, o número de instâncias que de fato representaria a quantidade de distintos acidentes acontecidos em rodovias federais entre 2007 e 2018 foi 35.822. A distribuição desses acontecimentos em relação a esse período pode ser observada no primeiro gráfico dessa seção (gráfico 1).

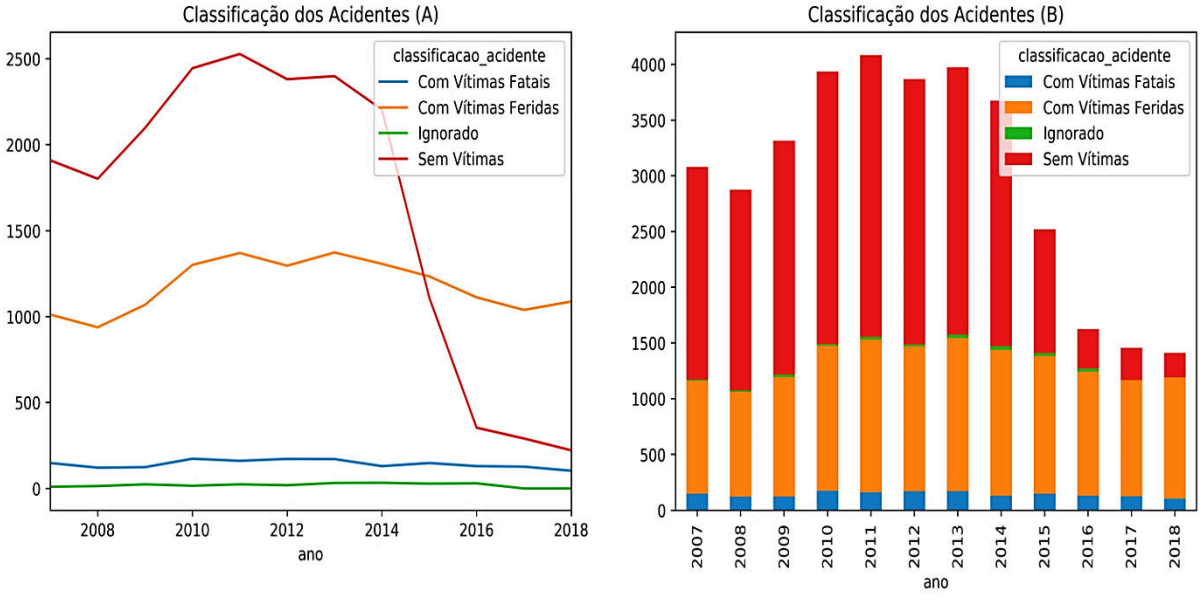
Gráfico 1 - Acidentes ocorridos em cada ano



Indubitavelmente, a informação que se sobreleva ao interpretar esse primeiro gráfico é a acentuada redução de acidentes registrados em rodovias federais nos últimos anos. Enquanto entre 2010 e 2014 a quantidade ocorrências era superior a 3.500 em cada ano, a partir de 2016

essa quantia se torna próxima de 1500, sendo, de fato, uma redução muito expressiva. É provável que a minimização desse quadro se deu pelas medidas preventivas e de conscientização que a própria PRF vem implantando, de maneira mais ativa, nos últimos anos: em 2016 foi iniciada a **Operação Rodovida** sob a responsabilidade desse órgão, ação que consiste em fiscalizações intensivas em períodos do ano, horários e pontos de rodovias considerados criticamente perigosos, além da promoção de campanhas de conscientização. Atrelado a isso, pode-se apontar a manutenção e criação de deliberações legislativas mais rígidas nesse período como outro fator que corroborou para redução de incidentes, como a revisão da Lei Seca (n° 11.705) e da lei n° 13.281 (coibição do uso do telefone celular durante condução de veículos).

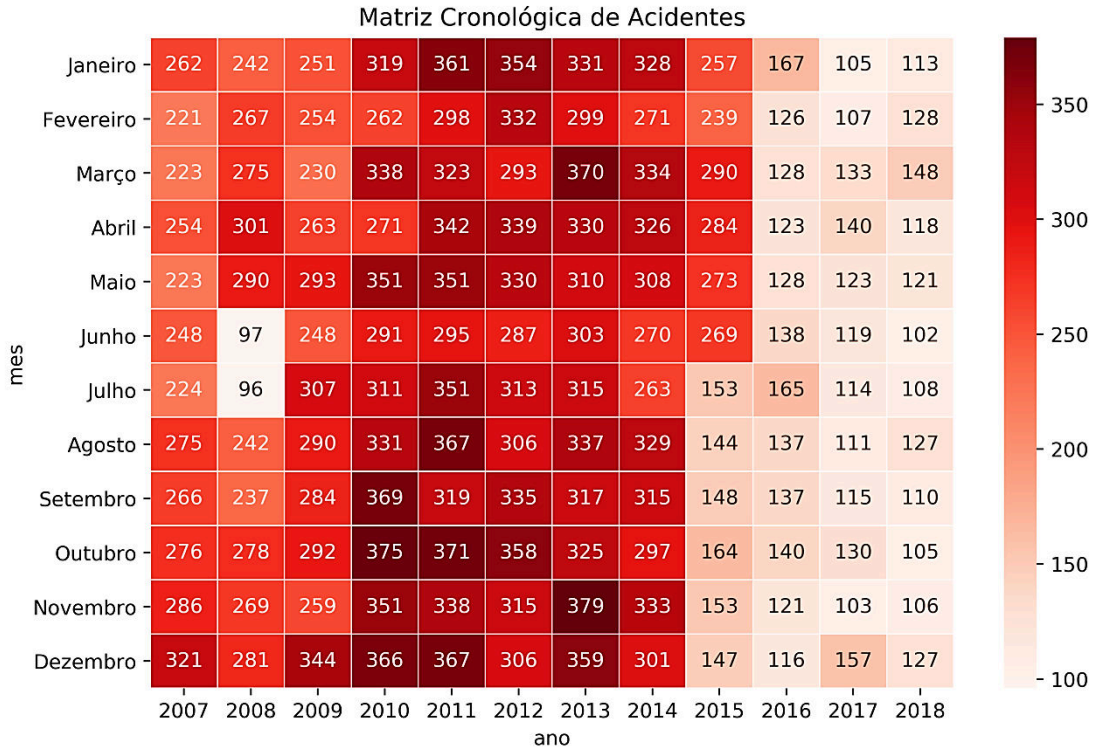
Gráfico 2 - Classificação dos acidentes ocorridos por ano



Ao se analisar o gráfico 2, por sua vez, percebe-se que mesmo a quantidade total de ocorrências terem diminuído, o número acidentes com vítimas feridas ou fatais persiste praticamente de maneira constante durante os anos. Em relação aos incidentes com vítimas feridas especificamente, nos dois últimos anos a quantia é maior do que em 2007 e 2008. A conclusão mais sensata a respeito dessa situação é de que houve uma redução significativa apenas dos acidentes de menor gravidade, os quais não causam ferimentos aos envolvidos. Cabe salientar ainda que esses dados não representam, realmente, todos os acidentes ocorridos em rodovias nesses anos, mas apenas os registrados formalmente pelas autoridades. Nesse

sentido, a quantidade total de incidentes provavelmente foi maior, mas se admite nessa pesquisa que somente a amostra estudada reflete satisfatoriamente o panorama geral.

Gráfico 3 - Mapa de calor dos acidentes ocorridos em cada mês e ano

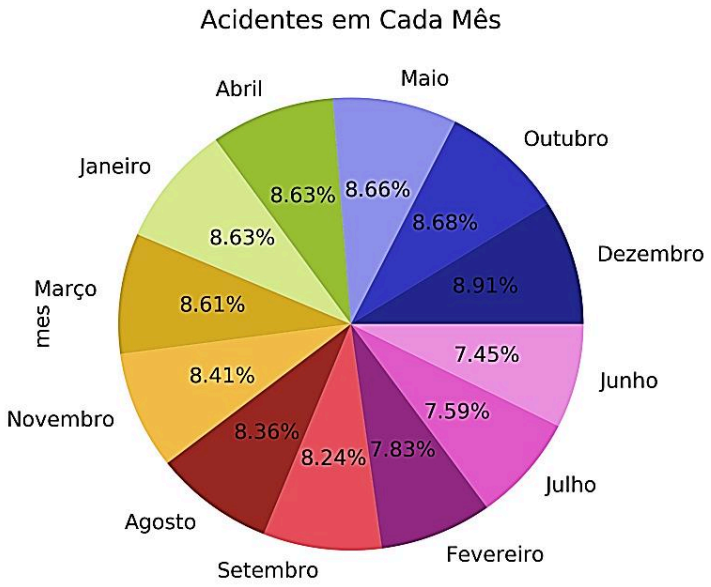


Para uma observação um pouco mais minuciosa a respeito da distribuição cronológica das ocorrências em estudo, o gráfico 3 foi construído. Trata-se de um mapa de calor, plotagem em que regiões mais escuras representam valores mais expressivos, maiores quantidades de acidentes nesse caso. A partir dessa visualização, é possível chegar a mesma conclusão obtida ao se analisar o primeiro gráfico, o número de acidentes registrados foi amenizado. Mais do que isso, entretanto, é possível observar o comportamento dos dados no decorrer de cada ano. Dessa forma, percebe-se que nos últimos 12 anos, o mês no qual foi registrado a maior quantidade de acidentes foi Novembro de 2013, de maneira geral, nota-se que os incidentes tendem a se concentrarem no mês de Dezembro, provavelmente, devido às festividades que ocorrem nesse mês, como o *Réveillon* e o Natal.

O gráfico 4 torna mais nítida a distribuição das ocorrências entre os meses de um ano. De fato, uma parcela um pouco maior desses acidentes ocorreram no fim do ano e como já discutido isso pode estar relacionado com celebrações que acarretam em maior tráfego nas rodovias, entretanto, o mês de Junho e Julho, época em que ocorrem as populares festas

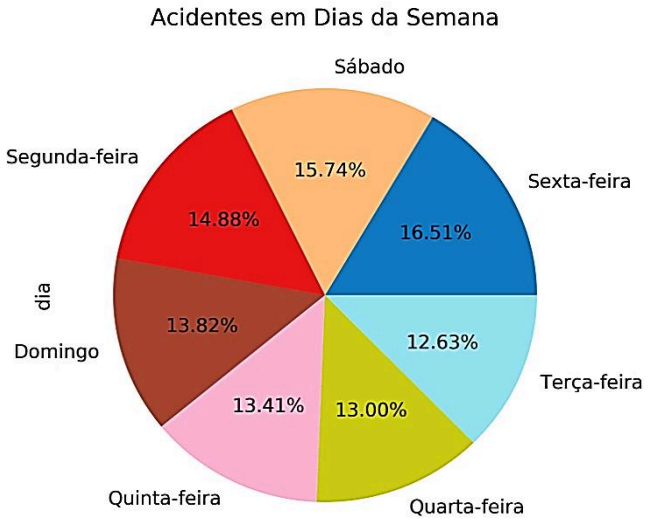
juninas na região, apresentaram, inusitadamente, os menores índices. Já em outros meses tipicamente festivos como Fevereiro e Março (período em que ocorre o Carnaval) também não apresentaram realces em relação à distribuição discutida. Diante disso, o que se pode inferir é que a disposição dos infelizes acontecimentos, em relação aos meses, pode estar associada a outros fatores além da ocorrência de eventos.

Gráfico 4 - Parcela de acidentes acontecidos em cada mês



Uma vez estudado o comportamento dos registros em relação aos anos e meses, também foi de interesse verificar como se arranjam no decorrer dos dias da semana. A partir do gráfico 5, essa análise pode ser feita e o que se observa, naturalmente, é que no Sábado e, especialmente, na Sexta-feira é mais provável que se ocorra acidentes nas rodovias do estado. Essa observação se justifica uma vez que no fim da semana muitas pessoas tendem a visitarem outros locais e municípios do RN ou de estados adjacentes, sobretudo, aquelas que trabalham e/ou residem em localidades distantes do seu grupo familiar.

Gráfico 5 - Acidentes distribuídos entre os dias da semana



Uma complementação para o gráfico anterior se refere à tabela 2, na qual estão presentes as quantidades de acidentes registrados em cada dia da semana subdivididos ainda entre as fases do dia. Por meio dessa, de maneira previsível, observa-se que a maioria das instâncias é relativa a incidentes que ocorreram em pleno dia, todavia, nesse intervalo, a Segunda-feira se demonstra o dia mais propício a esses acontecimentos. Já durante o período noturno, o Domingo e o Sábado podem ser considerados os principais dias em relação à probabilidade de ocorrência de um incidente em rodovias federais.

Tabela 2 - Acidentes em relação ao horário e o dia da semana

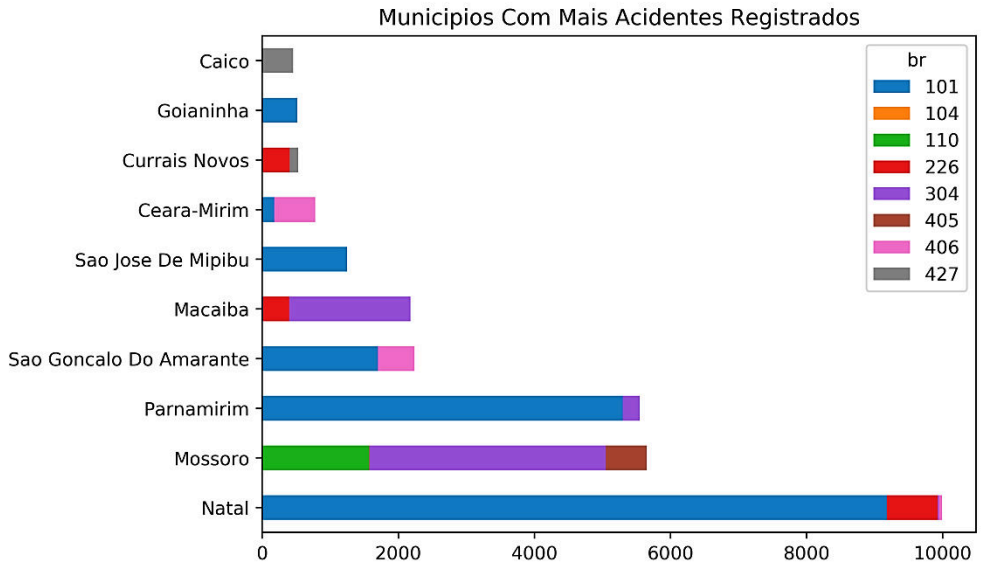
Dia da Semana	**Fase do Dia**				**Total**
	Amanhecer	**Anoitecer**	**Plena Noite**	**Pleno Dia**	
Domingo	252	353	2.135	2.212	4.952
Segunda-feira	191	295	1.333	3.512	5.331
Terça-feira	138	279	1.225	2.884	4.526
Quarta-feira	93	265	1.303	2.995	4.656
Quinta-feira	131	308	1.379	2.986	4.804
Sexta-feira	166	360	1.926	3.463	5.915
Sábado	233	280	2.118	3.007	5.638

No tocante às principais localidades em que os acidentes rodoviários ocorreram, a próxima visualização construída (gráfico 6) exibe os dez municípios do Rio Grande do Norte com maiores quantias de casos registrados, bem como as rodovias em que esses aconteceram. Nitidamente, a capital do estado se demonstra o município mais perigoso em relação às casualidades dessa natureza.

Com uma população de aproximadamente 800 mil habitantes e uma frota em entorno de 400 mil veículos segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), é natural que Natal detenha essa classificação em virtude de possuir um tráfego mais ativo. Ademais, Mossoró e Parnamirim também são municípios populosos e importantes polos econômicos para o estado assim como a capital, o que justifica os altos índices apresentados também nesses locais.

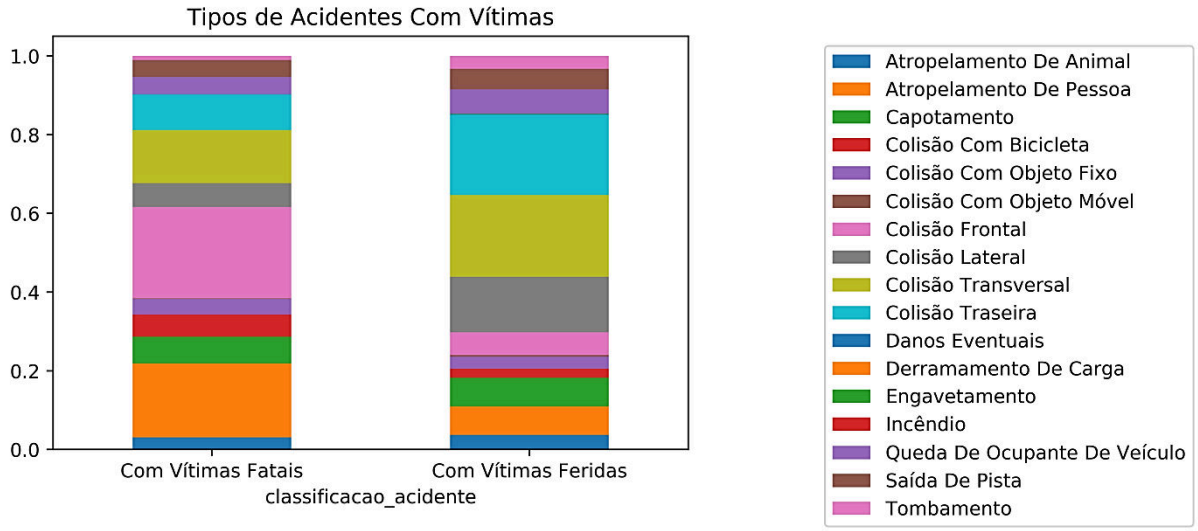
Ainda sobre o mesmo gráfico, nota-se que a quantidade de causalidades registradas na BR 101 chega a ser destoante em relação às demais. Sendo a segunda maior rodovia federal e uma das linhas de transportes mais importantes do Brasil, essa rodovia conta com trechos de tráfego intensos e de condições perigosos inclusive dentro do RN. Segundo o Departamento Nacional de Infraestrutura de Transportes (DNIT), os pontos mais perigosos dessa via dentro do estado situam-se entre os km 96.4 ao km 104.6 na área urbana de Natal e Parnamirim. Dessa forma, compreende-se que, de fato, essa rodovia se trata da mais perigosa do estado, dado que, na pesquisa realizada mais 53% dos acidentes catalogados ocorreram apenas nessa rodovia.

Gráfico 6 - Municípios e rodovias com mais acidentes registrados



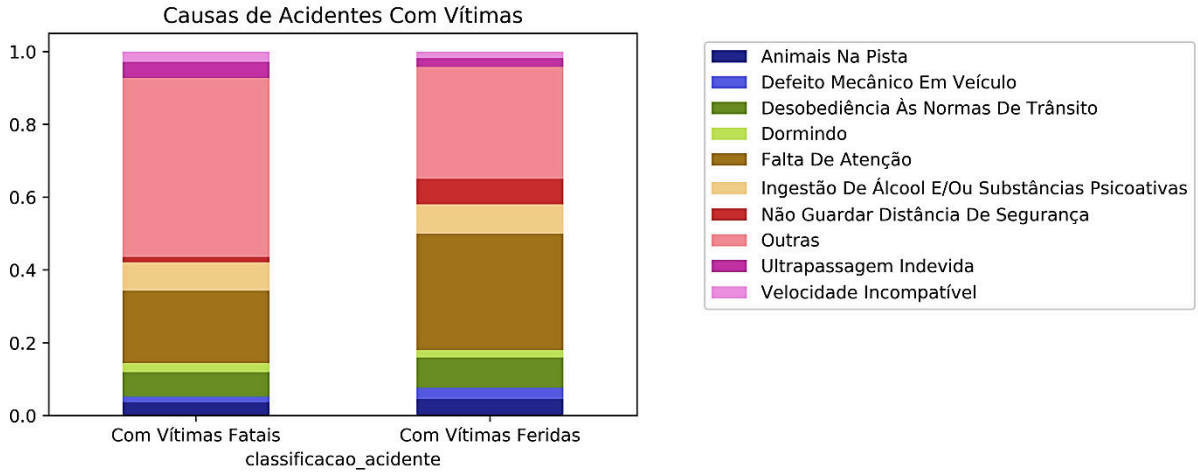
Uma vez constatado que o número de ocorrências com vítimas não teve uma redução significativa como o número total de acidentes, conforme já discutido, houve a preocupação, nessa pesquisa, sobre quais tipos e quais as principais causas levam aos casos mais perigosos, nos quais as vítimas sofrem ferimentos ou são levadas a óbitos. Para investigar especificamente esses tipos de causalidades, o gráfico 7 foi desenvolvido, cujos valores estão representados em escala normalizada. Segundo esse, os incidentes com vítimas feridas são frequentemente resultantes de três diferentes tipos de colisões: traseira, transversal e lateral, sendo a primeira a mais corriqueira. Em contrapartida, colisões frontais e atropelamentos podem ser considerados os tipos de acidentes mais perigosos e, ordinariamente, levam às fatalidades.

Gráfico 7 - Tipos de acidentes com vítimas feridas e fatais



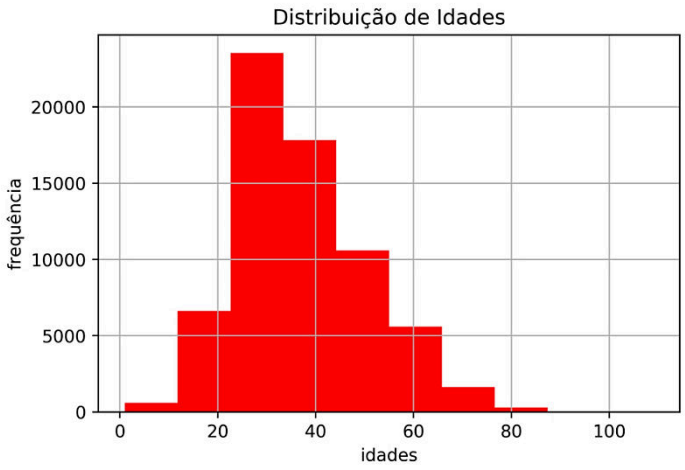
Já em relação às causas das ocorrências com maior nível de periculosidade, conforme o gráfico 8, nota-se que a falta de atenção é a principal circunstância acarretadora de casos em que as vítimas sofreram ferimentos. Relativo aos registros em que houve mortes, os dados desconhecem as mais importantes causas, entretanto, a desatenção também pode ser considerada uma condição que propicia perecimentos em rodovias potiguares. Vale salientar ainda que a ingestão de álcool e de substâncias psicoativas também pode ser apontada como um fator significativo que ocasiona acidentes tanto com vítimas feridas quanto fatais.

Gráfico 8 - Principais causas de acidentes com vitímas feridas e fatais



Analisadas as circunstâncias que levam as indesejáveis causalidades, foram investigados, também, os perfis das pessoas envolvidas nesses acontecimentos. Em relação às idades dessas em particular, conforme se pode notar por meio do histograma a seguir (gráfico 9), os indivíduos vítimas de acidentes, frequentemente, pertencem a faixa etária de 24 a 32 anos, isso é, jovens representam o grupo de cidadãos do estado que estão mais susceptíveis a invalidez por ferimentos e a fatalidades em decorrências de incidentes em rodovias. Além de prejuízos sociais evidentemente, esse quadro também representa percas econômicas ao estado, haja vista que as pessoas com essas idades representam uma significativa parcela da população economicamente ativa, de maneira recíproca, pode-se compreender que, justamente, devido a esse fato, os resultados podem ser justificados, visto que, pessoas que trabalham tendem a periodicamente trafegar entre diferentes municípios e regiões do estado.

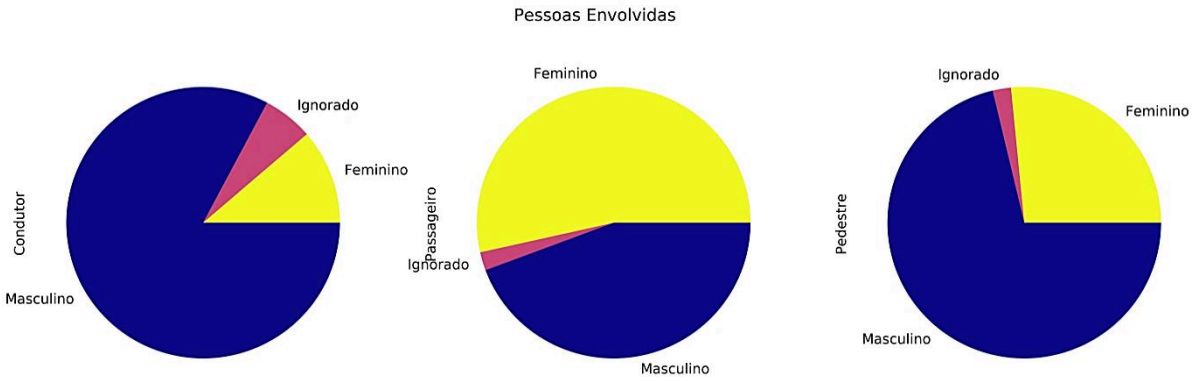
Gráfico 9 - Idades das pessoas envolvidas em acidentes



No tocante ao sexo das vítimas por sua vez, o próximo gráfico (10) indica a percentagem dos gêneros agrupados quanto ao envolvimento nos acidentes, respectivamente, para os condutores, passageiros e pedestres da esquerda para direita. A informação que se sobressai é a de que enquanto os condutores dos veículos são majoritariamente homens, as mulheres representam a maior parcela dos passageiros acidentados. Por sua vez, os pedestres em sua maioria também foram registrados como pessoas do sexo masculino.

Os primeiros resultados demonstram-se válidos, de fato, o número de homens habilitados é ainda significantemente maior na região segundo o Departamento Estadual de Trânsito (DETRAN), dessa forma, indivíduos do sexo oposto tendem a se locomovem como passageiros, em contrapartida, a última situação é deveras inusitada, certamente, um equilíbrio entre pessoas de sexos distintos seria previsível em relação aos pedestres, o que não se concretiza segundo os dados.

Gráfico 10 - Sexo das pessoas envolvidas nas ocorrências registradas



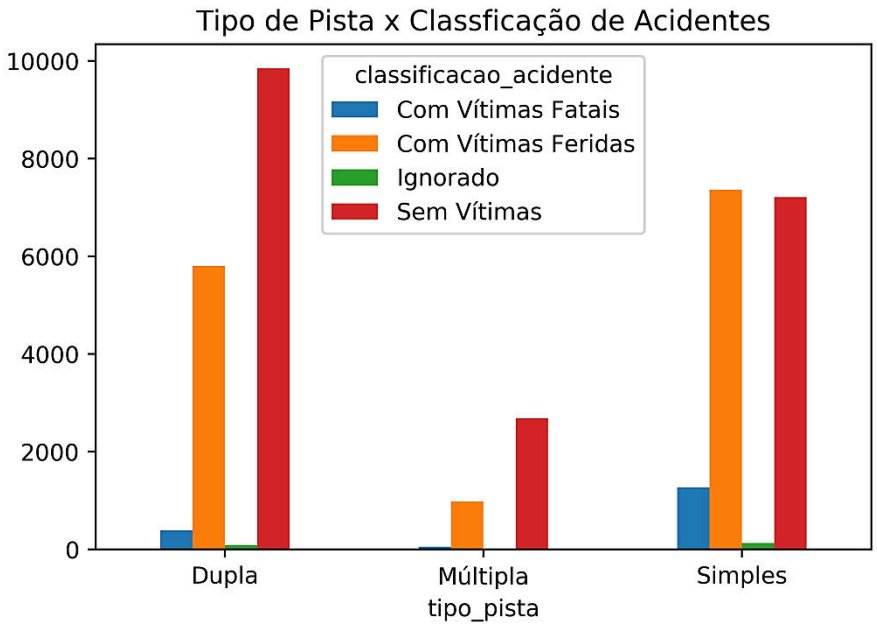
A última tabela desenvolvida na seção de exploração preliminar (tabela 3) relaciona as demais variáveis estudadas na base de dados. Essas refletem as condições das vias e do tempo sob as quais os incidentes foram registrados. Como se percebe, a maioria dos fatídicos eventos aconteceram dentro do perímetro urbano em boas condições climáticas e, curiosamente, em pistas duplicadas (teoricamente mais seguras).

O gráfico a seguir (gráfico 11), foi construído com intuito de elucidar essa questão, e o que se nota a partir desse é que realmente pistas duplas, apesar dos maiores índices registrados, são menos perigosas, uma vez que o número de acidentes com vítimas fatais e feridas é muito menor em relação a pistas simples por exemplo. Isso ocorre pois em trechos duplicados, há uma separação físicas entre as vias e o sentido do tráfego é separado reduzindo assim colisões, sobretudo frontais, tipo de acidente de mais alto risco como já discutido.

Tabela 3 - Localidades, tipos de pista e condições climáticas presentes nos registros

	Condição Meteorológica						
Local	**Tipo de Pista**	**Chuva**	**Céu Claro**	**Ignorado**	**Nublado**	**Sol**	**Vento**
Rural	**Dupla**	264	1.184	25	208	222	11
	Múltipla	12	34	-	12	5	-
	Simples	550	6.023	201	750	1.040	81
Urbano	**Dupla**	1.047	9.191	169	1.522	2.144	139
	Múltipla	288	2.283	51	334	698	12
	Simples	311	5.171	89	640	1057	54

Gráfico 11 - Tipo de pistas e classificação dos acidentes



4.2 Resultados da Mineração de Dados

Após a realização da MD, uma análise criteriosa foi feita a respeito das regras de associação obtidas. Dentre o total encontrado pelo algoritmo, algumas regras foram consideradas relevantes e nessa subseção serão apresentadas e analisadas. A tabela 4 resume as principais associações mineradas a partir da base de dados estudada.

Tabela 4 - Regras de associação obtidas

#	Antecedentes	Consequências	Confiança	*Lift*	Convicção
1	BR 101, Natal, Sol, Falta de atenção, colisão traseira	Pleno Dia, Urbano	99,0%	2,11	51,55
2	BR 110, Sem vítimas, colisão traseira, Urbano	Mossoró	98,4%	6,00	53,71
3	Rural, Atropelamento de animal, BR 304, sem vítimas	Simples, Animais na pista	97,4%	19,25	37,16
4	Simples, Condutor, Ingestão De Álcool E/Ou Substâncias Psicoativas, adulto entre 35 e 53 anos	Masculino	96,8%	1,16	5,18
5	Não guardar distância de segurança, chuva, sem vítimas, Urbano	Colisão traseira	95,4%	2,30	12,64
6	Ultrapassagem indevida, colisão frontal	Simples	94,5%	2,26	10,52
7	Queda de ocupante de veículo, urbano	Com vítimas feridas	92,8%	2,51	8,70
8	Simples, pleno dia, saída de pista	Rural	90,9%	3,48	8,15
9	BR 101, São Gonçalo do Amarante, com vítimas feridas, urbano	Dupla	88,5%	1,91	4,65
10	Jovem adulto entre 18 e 35 anos, passageiro, Dupla	Com vítimas feridas	86,4%	2,26	4,56
11	BR 427, com vítimas feridas	Simples, Céu Claro	83,9%	2,76	4,32
12	Plena noite, Mossoró, dupla	Céu Claro, Urbano	82,3%	1,67	2,86
13	Desobediência às normas de trânsito, BR 304, Urbano	Mossoró	80,5%	4,90	4,29
14	BR 101, Atropelamento de pessoa	Com vítimas feridas, urbano	80,1%	3,14	3,75
15	Rural, Sábado	Simples	80,0%	1,88	2,88

Uma regra dessa natureza pode ser interpretada como uma relação em que as consequências provavelmente ocorrem se os antecedentes também acontecerem. Logo, a última associação presente na tabela acima, por exemplo, pode ser compreendida como: os acidentes ocorridos em áreas rurais no dia de sábado provavelmente acontecem em tipos de pistas simples. A probabilidade de que a consequência venha a se concretizar sob essas condições é dada pela confiança numérica, no caso da décima quinta regra, nota-se que 80% dos casos ocorreram conforme a regra. Outrossim, o interesse dessas relações, como já discutido em seções anteriores, pode ser quantificada pela métrica de *lift* e a dependências entre os fatores, pela convicção. Um alto valor de *lift* significa que o suporte da regra é muito maior do que o esperado, mesmo que as consequências ocorram com pouca frequência no conjunto de dados. Já um valor expressivo de convicção representa uma relação muito forte entre os antecedentes e as consequências, em particular, caso o primeiro nunca ocorra sem acarretar o segundo, o valor da convicção tende ao infinito. Ainda sobre o mesmo exemplo, a

valor de *lift* e de convicção, como se ver, vale 1,88 e 2,88 respectivamente, valores baixos, porém coerentes.

O seguinte quadro reúne as interpretações feitas a partir das regras presentes na tabela 4, os números à esquerda (coluna '#') identificam cada regra demonstrada.

Quadro 4 - Interpretação das regras de associação obtidas

#	Interpretação
1	Na rodovia 101 dentro do município de Natal e sob boas condições climáticas, 99% dos incidentes classificados como colisões traseiras provocadas por desatenção ocorreram em pleno dia e dentro do perímetro urbano. Essa regra se demonstrou praticamente duas vezes mais interessante do que o esperado (*lift*= 2,11) e em raros casos, sob tais condições, os acidentes ocorreram em outro horário ou área, de acordo com a convicção muito alta (51,55) apresentada.
2	Na BR 110, dentro de uma área urbana, Mossoró se demonstra o município mais provável no qual uma colisão traseira de baixo risco possa ocorrer. 98, 4% desses casos ocorreram em tal local. Devido ao alto valor de interesse e de convicção essa regra se demonstra bastante válida.
3	Incidentes acontecidos em zonas rurais na BR 304 envolvendo atropelamento de animais que não resultaram em ferimentos ou óbito das vítimas, naturalmente, aconteceram devido à presença desses na pista, entretanto, mais do que isso, majoritariamente ocorreram em pistas simples. Os altos valores das métricas garantem a qualidade dessa regra, bem como, a forte relação entre os fatores, o que de certa forma é intrigante, pois é senso comum haver feridos nesses tipos de acidentes.
4	Nas tristes eventualidades registradas em pistas simples cuja principal causa foi atribuída ao consumo de bebidas alcoólicas ou outras drogas análogas pelos envolvidos, os condutores entre 35 e 53 anos eram, tipicamente, homens. O valor de *lift* para essa regra é próximo de 1, o que demonstra que a essa não é uma boa associação, pois o sexo masculino é uma categoria muito frequente, todavia, o valor de convicção é significativo e demonstra uma relação entre as consequências e os antecedentes. Dessa forma, homens, provavelmente, representam a maioria das pessoas com essa faixa etária e que utilizam esse tipo de substâncias.
5	Não aguardar a distância de segurança, principalmente em solo urbano e em tempo de chuva, acarreta frequentemente em colisões traseiras, entretanto, não há vítimas feridas nesses casos. As métricas para essa regra a favorecem como uma boa associação.
6	Acidentes cuja causa deve-se a ultrapassagens indevidas e classificados como colisões frontais também ocorrem com frequência em trechos de pista simples. A partir das métricas associadas a essa relação, percebe-se que de fato acidentes dessa natureza costumam acontecerem nesses tipos de vias.
7	Quedas de ocupantes de veículos, principalmente em zonas urbanas, seguramente a partir das métricas, levam a acidentes cujas vítimas sofrem ferimentos.
8	Em trechos de vias simples durante o dia, cerca de 90% dos incidentes causados devido à saída da pista foram registrados em áreas rurais. Essa regra se trata de uma associação interessante, subjetivamente e quantitativamente, a qual demonstra que fora do perímetro urbano é mais provável que acidentes dessa forma aconteçam, isso pode estar relacionado a fatores já discutidos como a presença de animais em pistas ou talvez nessas se encontrem precariedades com mais facilidade.
9	Na cidade de São Gonçalo do Amarante na BR 101, curiosamente, acidentes de alto risco são mais frequentes em trechos de rodovias duplicadas.
10	Passageiros jovens entre 18 e 35 anos frequentemente representam as pessoas envolvidas em acidentes com vítimas feridas acontecidos em pistas duplicadas.

11	Na rodovia federal 427 acidentes com vítimas feridas são constantemente registrados em trajetos simples e sob boas condições climáticas, cerca de 80% dos casos aconteceram conforme essa regra. Em verdade, segundo o DNIT (2019) não há trechos duplicados na BR 427, certamente isso justifica o fato de que, mesmo em condições meteorológicas favoráveis, incidentes de alto risco costumam ocorrer nessa via.
12	Em Mossoró durante a noite, acidentes em trechos de pista dupla predominantemente (82,3% das vezes) ocorrem em condições de céu claro e dentro do perímetro urbano do município. Essa associação se demonstra interessante subjetivamente, uma vez que, o horário parece influir significativamente sobre as ocorrências nessa cidade.
13	Casualidades cuja principal causa se deva a desobediência as normas de trânsito em zona urbana na BR 304, trivialmente, são também registradas no município de Mossoró. As métricas intrínsecas a essa regra são significativas, por isso, é razoável apontar esse município como um local em que a negligência é um forte fator que propicia os acontecimentos de incidentes.
14	Na rodovia federal 101, atropelamentos de pessoas seguramente acarretarão em vítimas feridas e ocorrerão em centros urbanos. Em 80% dos casos essa regra se demonstrou válida e as demais métricas também garantem a qualidade dessa associação.
15	Em zonas rurais no sábado, 80% dos casos registrados aconteceram em trechos simples. Novamente se percebe o quanto esse tipo de pista favorece a ocorrências de incidentes.

Sobre os acidentes nas rodovias federais do RN, a partir das regras, já analisadas, e mesmo das demais não apresentadas nessa seção, em síntese, pode-se afirmar que esses são favorecidos, de maneira geral, não por circunstâncias insopitáveis, mas por fatores e causas que podem ser evitados. Várias das associações encontradas apontaram a desatenção durante a condução, o descaso as normas de trânsito e o uso de substâncias indevidas como recorrentes causas de incidentes. A presença de animais nas rodovias, sobretudo nas zonas rurais da BR 304, também se mostrou um quadro estarrecedor no que tange aos antecedentes que levam as casualidades. Os maiores centros urbanos do RN, Natal e Mossoró principalmente, polarizam os números de casos registrados, uma observação razoável, uma vez que, se tratam dos municípios mais populosos e importantes economicamente para o estado. Por fim, trechos de pistas simples se demonstraram importantes cenários acarretadores de acidentes rodoviários, inclusive dos mais agravantes, lamentavelmente, ainda no estado a quantidade de pontos em que rodovias são duplicadas é ínfimo (DNIT, 2019), uma condição que, certamente, compromete a segurança das pessoas que trafegam nessas vias.

CONCLUSÃO

Ao reaver os objetivos inicialmente propostos e os resultados finalmente obtidos, percebe-se que o presente trabalho se demonstrou válido. Todo o procedimento de Descoberta de Conhecimento em Bases de Dados foi empregado com ponderação ao conjunto de dados estudado e bons consequentes derivaram dessa metodologia. De maneira geral, tanto os resultados obtidos na exploração preliminar quanto na mineração de fato, foram considerados satisfatórios e coerentes de modo que possibilitaram a resolução das questões de pesquisa inicialmente levantadas.

Dentre as constatações alcançadas a respeito do quadro histórico e atual dos acidentes (QP 1), se situa a significativa redução dos casos em rodovias do estado, as ações que veem sendo tomadas, de maneira mais ativa por órgãos como a Polícia Rodoviária, foram apontadas como um importante fator que corroborou para isso, entretanto, a minimização desse quadro, aparentemente, afetou apenas incidentes pouco agravantes, visto que aqueles de maior risco não tiveram uma acentuada diminuição registrada. Dessa forma, enquanto em outrora a quantidade total de acidentes rodoviários era mais significativa, atualmente esse número é menos expressivo, entretanto, aqueles que causam mortes e ferimentos persistem praticamente em mesma proporção.

Ademais, relativo às circunstâncias das casualidades (QP 2), percebe-se que as causas desses acontecimentos muitas vezes estão atreladas aos estados e comportamentos indevidos das pessoas envolvidas, um ressalto pode ser atribuído a Mossoró, município no qual a negligência as normas de trânsito aparenta ser uma recorrente causa de acidentes, todavia, fatores como o tipo de pista e a presença de animais nas rodovias (sobretudo na BR 304) também se demonstraram importantes condições que propiciam a efetivação dos infelizes eventos.

Em relação ao perfil dos indivíduos vítimas dos acidentes em estudo (QP 3), pôde-se notar que em sua maioria são homens de idade ativa economicamente, esses também representam os motoristas e pilotos que conduzem sob efeito de drogas como o álcool, em contrapartida, os passageiros estudados em sua maioria eram do sexo feminino. Dessa forma, ao final desse trabalho, a partir dessas observações, consideram-se resolvidos todos os levantamentos inicialmente estabelecidos e a importância dessa pesquisa, justificada, visto que as informações aqui expostas por meio das regras e visualizações podem complementar as estatísticas utilizadas por órgãos responsáveis pela segurança nas rodovias do estado ou

suscitar novas pesquisas de cunho científico a respeito do mesmo tema e dessa forma contribuir, mesmo que minimamente, na atenuação do problema estudado.

TRABALHOS FUTUROS

Em relação aos possíveis trabalhos futuros relacionados ao tema abordado, podem-se citar os seguintes:

- Aplicações de técnicas de caráter preditivo ou de outros algoritmos ao mesmo banco de dados estudado.
- Estudo análogo sobre rodovias específicas presentes no estado, inclusive não federais.
- Pesquisas que envolvam outras bases de dados correlatas como registro de infrações que a própria Polícia Rodoviária Federal também disponibiliza.
- Desenvolvimento de *Softwares* educativos, visto que muitos acidentes foram ocasionados devido a imprudências, ou de aplicativos que auxiliem condutores, pois a desatenção também foi identificada como uma importante causa de incidentes.

REFERÊNCIAS

AGRAWAL, R; IMIELINSKI, T; SWAMI, A. **Mining Association Rules between Sets of Items in Large Databases.** San Jose: IBM Almaden Research Center, 1993.

AGRAWAL, R.; SRIKANT, R., **Fast Algorithms for Mining Association Rules.** San Jose: IBM Almaden Research Center, 1994.

ANACONDA D. **Anaconda**, 2019. Disponivel em: <https://www.anaconda.com/distribution/>. Acesso em: 14 Julho 2019.

BOSCHETTI, A.; MASSARON, L. **Python Data Science Essentials:** Become an efficient data science practitioner by thoroughly understanding the key concepts of Python. 1ª. ed. [S.l.]: Packt Publishing, 2015.

BRASIL, **Lei Seca.** Lei nº 11.705, 19 de Junho de 2008.

_____., **Lei nº 13.28,** 4 de Maio de 2016.

BRIN, S., *et al.* Dynamic Itemset Counting and Implication Rules for Market Basket Data. **SIGMOD Record,** Volume 6, Number 2: New York, June 1997, pp. 255 - 264.

BUSSAB, W. D. O.; MORETTIN, P. A. **Estatística Básica**. 7ª. ed. São Paulo: Saraiva, 2012.

CAMILO, C. O.; SILVA, J.C. **Mineração de Dados:** Conceitos, Tarefas, Métodos e Ferramentas. Goiânia: Instituto de Informática Universidade Federal de Goiás, 2009.

CAVIQUE, L. Big data e data science. **Boletim da APDIO**, p. 11–14, 2014.

CIELEN, D.; MEYSMAN, A. D. B.; ALI, M. **Introducing Data Science:** Big data, machine learning, and more, using Python tools. 1ª. ed. Nova York: Manning Publications, 2016.

CONTENTS. **Matplotlib.org**, 2019. Disponivel em: <https://matplotlib.org/contents.html>. Acesso em: 14 Julho 2019.

CÔRTES, S. C.; PORCARO, R. M; LIFSCHITZ, S. **Mineração de Dados –** Funcionalidades, Técnicas e Abordagens**.** p. 35, Rio de Janeiro: PUC-RioInf. MCC, 2002.

COSTA, C. N., *et al.,* Descoberta de Conhecimento em Bases de Dados. **Revista Eletrônica: Faculdade Santos Dumont. 2019, Disponível em:** <https://fsd.edu.br/revistaeletronica/arquivos/2Edicao/artigo9.pdf> Acesso em: 27 jun. 2019.

COSTA, E. R. **Bancos de dados relacionais.** 2011. 64 f. TCC (Graduação) - Curso de Tecnólogo em Processamento de Dados, Faculdade de Tecnologia de São Paulo, São Paulo, 2011.

COSTA, J. A. F. *et al.* **Uma introdução à Mineração de Dados:** Conceitos e Aplicações. Mossoró: EdUFERSA, 2014.

CRESPO, A. A. **Estatística Fácil**. 19ª. ed. São Paulo: Saraiva, 2009.

DADOS abertos. **Portal da Polícia Rodoviária Federal**, 2019. Disponivel em: <https://www.prf.gov.br/portal/dados-abertos/>. Acesso em: 14 Julho 2019.

DEVORE, J. L. **Probabilidade e estatística:** para engenharia e ciências. 6ª. ed. São Paulo: Cengage Learning, 2006.

CONDIÇÕES DAS RODOVIAS – DNIT (2017). Disponível em: <http://servicos.dnit.gov.br/condicoes/condicoesdrf.asp?BR=101&Estado=Rio+Grande+do+Norte&drf=14>Acesso em: 19 de Julho de 2019.

DNIT (2019). Disponível em: < http://servicos.dnit.gov.br/vgeo/> Acesso em: 19 de Julho de 2019.

DOCUMENTATION. **Python.org**, 2019. Disponivel em: <https://docs.python.org/3/>. Acesso em: 13 Julho 2019.

IBGE – CIDADES E ESTADOS DO RIO GRANDE DO NORTE (2019). Disponível em: <https://cidades.ibge.gov.br/brasil/rn/natal/panorama> Acesso em: 19 de Julho de 2019.

FACELI, K. *et al.* **Inteligência Artificial:** Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 2011.

FAYYAD, U; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine,** Califórnia, v. 17, n. 3, 199, p. 37-54.

_____., Knowledge Discovery and Data Mining: Towards a Unifying Framework. **KDD-96 Proceedings.** California, 1996 p. 82-88.

_____., The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM,** New York, v. 39, n. 11, 1996, p. 27–34.

FEDELI, R. D.; POLLONI, E. G. F.; PERES, F. E. **Introdução à Ciência da Computação**. 2ª. ed. São Paulo: Cengage Learning Editores, 2011.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook:** Advanced Approaches in Analyzing Unstructured Data. 1ª. ed. Nova York: Cambridge University Press, 2006.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002. Apostila.
GIL, A. C. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas, 2007.

FOSTER, P.; FAWCETT, T. **Data Science for Business**. 1ª. ed. [S.l.]: O'Reilly Media, Inc., 2013.

GALDINO, N. Big Data: Ferramentas e Aplicabilidade. In: **Anais do XIII SEGeT – Simpósio de Excelência de Gestão e Tecnologia.** AEDB, 2016.

GLOBAL STATUS REPORT ON ROAD SAFETY 2018. Geneva: World Health Organization; 2018. Licence: CC BYNC-SA 3.0 IGO.

GOLD, P. A. **Segurança de trânsito:** Aplicações de Engenharia para reduzir acidentes. EUA: BID, 1998.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático. 1ª. ed. Rio de Janeiro: Elsevier, 2005.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining Concepts and Techniques**. 3ª. ed. San Francisco: Elsevier, 2012.

HURWITZ, J. *et al.* **Big Data For Dummies**. 1ª. ed. New Jersey: John Wiley & Sons, 2013.

KAGGLE ML & DS SURVEY. **Kaggle**, 2019. Disponivel em: <https://www.kaggle.com/kaggle/kaggle-survey-2018#multipleChoiceResponses.csv>. Acesso em: 13 Julho 2019.

LAROSE, D. T. **Discovering Knowledge in Data An Introduction to Data Mining**. 1ª. ed. New Jersey: John Wiley & Sons, 2005.

LARSON, R.; FARBER, B. **Estatística aplicada**. 6ª. ed. São Paulo: Pearson Education do Brasil, 2015.

MADHAVAN, S. **Mastering Python for Data Science:** Explore the world of data science through Python and learn how to make sense of data. 1ª. ed. [S.l.]: Packt Publishing, 2015.

MASSARON, L.; MUELLER, J. P. **Python for Data Science For Dummies®**. 1ª. ed. Nova Jersey: John Wiley & Sons, 2015.

NELLI, F. **Python Data Analytics:** Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language. 1ª. ed. Nova York: Apress Media LLC, 2015.

NILSSON, N. J. **Introduction to machine learning:** An early draft of a proposed textbook. 1ª. ed. Palo Alto: [s.n.], 1996.

NUMPY. **numpy.org**, 2019. Disponivel em: <http://www.numpy.org/>. Acesso em: 13 Julho 2019.

OLSON, D. L.; DELEN, D. **Advanced Data Mining Techniques**. 1ª. ed. Nova York: Springer-Verlag, 2008.

OPAS/OMS. (2018) Disponível em: <https://www.paho.org/bra/index.php?option=com_content&view=article&id=5147:acidentes-de-transito-folha-informativa&Itemid=779> Acesso em: 21 de Julho de 2019.

PINHEIRO, J. I. D. *et al.* **Estatística Básica:** A Arte de Trabalhar Com Dados. 1ª. ed. Rio de Janeiro: Elsevier, 2009.

PORTAL G1/DF (2018). Disponível em: <https://g1.globo.com/df/distrito-federal/noticia/2018/12/06/acidentes-em-rodovia-federais-mataram-mais-de-83-mil-pessoas-no-brasil-em-10-anos.ghtml> Acesso em: 21 de Julho de 2019.

PRASS, F. S. **KDD –** uma Visão Geral Do Processo. 2016. Disponível em: <https://docplayer.com.br/7949839-Kdd-uma-visal-geral-do-processo.html> Acesso em: 25 jun. 2019.

PYLE, D. **Data Preparation for Data Mining.** San Francisco: Morgan Kaufmann Publishers, Inc. 1999.

PYTHON. **Python Data Analysis Library**, 2019. Disponivel em: <https://pandas.pydata.org/>. Acesso em: 14 Julho 2019.

RAO, S.; GUPTA, P., Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm**. International Journal of Computer Science And Technology** v. 3, n. 1, p. 5, 2012.

RATNER, B. **Statistical and Machine-Learning Techniques for Better Predictive Modeling**. 2ª. ed. [S.l.]: CRC Press, 2011.

REIS, C. V. R. **O uso da descoberta de conhecimento em Banco de Dados nos acidentes da BR-381.** (Mestrado em Sistemas de Informação e Gestão do Conhecimento). Fundação Mineira de Educação e Cultura, Minas Gerais, 2014.

RUD, O. P. **Data Mining Cookbook:** Modeling Data for Marketing, Risk, and Customer Relationship Management. 1ª. ed. Nova York: John Wiley & Sons, 2001.

SÁ, E. M. **Detecção e monitoramento de dados inconsistentes através do método axiomático**, 2015. Disponível em: <https://www.academia.edu/37111819/Detec%C3%A7%C3%A3o_e_monitoramento_de_dados_inconsistentes_atrav%C3%A9s_do_m%C3%A9todo_axiom%C3%A1tico> Acesso em: 28 jun. 2019

SAÚDE, O. M. **Relatório mundial sobre prevenção de lesões causadas pelo trânsito:** resumo. Brasília: Biblioteca da Opas, 2012. 73 p.

SEABORN: statistical data visualization. **seaborn**, 2019. Disponivel em: <https://seaborn.pydata.org/index.html>. Acesso em: 14 Julho 2019.

SILVA, M. P. S., **Mineração de Dados:** Conceitos, Aplicações e Experimentos com Weka. Livro da Escola Regional de Informática Rio de Janeiro – Espírito Santo. Porto Alegre: Sociedade Brasileira de Computação, 2004, v. 1, p. 1-20.

SOCZEK, F. C.; ORLOVSKI, R. Mineração de Dados: Conceitos e aplicação de algoritmos em uma Base de Dados na área da saúde. **Semana Acadêmica,** vol.1, 2004.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. 1ª. ed. [S.l.]: Pearson, 2005.

TESSAROLO, P. H; MAGALHÃES, W. B. **A era do big data no conteúdo digital:** os dados estruturados e não estruturados. Paraná: Unipar, 2013. Disponível em:

<http://web.unipar.br/~seinpar/2015/_include/artigos/Pedro_Henrique_Tessarolo.pdf> Acesso em: 26 jun. 2019

THOMÉ, A. C. G. **Redes neurais:** uma ferramenta para KDD e Data Mining. s.l.: [Universidade Federal do Rio de Janeiro], s.d. (Apostila), 2002. Disponível em: <http://equipe.nce.ufrj.br/thome/grad/nn/mat_didatico/apostila_kdd_mbi.pdf>. Acesso em: 28 jun. 2019.

VERMEULEN, A. F. **Practical Data Science:** A Guide to Building the Technology Stack for Turning Data. 1ª. ed. [S.l.]: Apress, 2018.

WANG, L.; FU, X. **Data Mining with Computational Intelligence**. 1ª. ed. Berlim: Springer-Verlag Berlin, 2005.

WITTEN, I. H.; FRANK, E. **Data Mining:** Practical machine learning tools and techniques. 2ª. ed. San Francisco: Elsevier, 2005.

YE, N. **The handbook of Data Mining**. 1ª. ed. [S.l.]: Lawrence Erlbaum Associates, 2003.
ZAKI, M. J.; JR, W. M. **DATA MINING AND ANALYSIS**. Nova York: Cambridge University Press, 2014.

ZHANG, S.; ZHANG, C.; XINDONGWU. **Knowledge Discovery in Multiple Databases**. 1ª. ed. [S.l.]: Springer-Verlag, 2004.

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal Rural do Semi-Árido - UFERSA
CAMPUS DE CARAÚBAS

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Às 16h30min., do dia 14 de agosto do ano de dois mil e dezenove, na sala 10 do bloco central de aulas I da Universidade Federal Rural do Semi-árido – UFERSA Campus de Caraúbas-RN, sob a presidência do professor Me. Wedson Carlos Gomes de Oliveira reuniu-se a Banca Examinadora de defesa do Trabalho de Conclusão de Curso (TCC) de autoria do aluno **DANIEL DA SILVA SANTOS**, aluno do curso de Bacharelado em Ciência e Tecnologia desta Universidade, com o título "*MINERAÇÃO DE DADOS: UMA ABORDAGEM SOBRE OS DADOS DAS RODOVIAS FEDERAIS DO RIO GRANDE DO NORTE*". A Banca Examinadora ficou assim constituída: professor Me. Wedson Carlos Gomes de Oliveira, presidente da banca e orientador do TCC; Prof. Dr. Oscar Bayardo Ramos Lovon, e o Prof. Dr. Mauricio Zuluaga Martinez, como membros. Foram registradas as seguintes ocorrências:_____ Concluída a defesa, procedeu-se o julgamento pelos membros da banca examinadora, em reunião fechada, tendo o aluno obtido as seguintes notas: 10 (Dez); 10 (Dez) e 10 (Dez). Apuradas as notas verificou-se que o aluno foi APROVADO com *média geral* 10 (Dez), fazendo jus, portanto, ao título de Bacharel em Ciência e Tecnologia. E para constar, eu, Wedson Carlos Gomes de Oliveira, presidente da banca lavrei a presente ata que, após lida e aprovada pelos membros da banca examinadora, será assinada por todos. Caraúbas, 14 de agosto de 2019.

Assinatura dos membros da Banca Examinadora.

Wedson Carlos Gomes de Oliveira
Prof. Me. Wedson Carlos Gomes de Oliveira
Presidente

Oscar Bayardo Ramos Lovon
Prof. Dr. Oscar Bayardo Ramos Lovon
Membro

Mauricio Zuluaga M.
Prof. Dr. Mauricio Zuluaga Martinez
Membro