



Programa de Pós-Graduação em Engenharia de Sistemas

DISSERTAÇÃO



Recife, 29 Maio de 2017.



Universidade de Pernambuco (UPE)
Escola Politécnica de Pernambuco (POLI)
Instituto de Ciências Biológicas (ICB)

**MODELO PREDITIVO PARA SUGESTÃO DE
ROTEAMENTO RODOVIÁRIO DE CARGAS
CONSIDERANDO DADOS HISTÓRICOS,
SÓCIO-AMBIENTAIS E DE REDES SOCIAIS**

Mestrando: Othon Luiz Teixeira de Oliveira
Orientador: Prof. Dr. Fernando Buarque de Lima Neto

Dissertação de Mestrado apresentada ao
Programa de Pós-Graduação em Engenharia
de Sistemas
Área de concentração: **Cibernética.**

Banca de examinadora:

Prof. Dr. Emerson Alexandre de Oliveira Lima
Prof^a. Dr^a. Maria Lencastre Pinheiro de Menezes e Cruz

Recife, 29 Maio de 2017.

*“Quem escolheu a busca
não pode recusar a travessia”
(Guimarães Rosa)*

*“Ao meu pai que o destino levou pelas mãos,
mas deixou-me as razões e a motivação dessa pesquisa.
Homem forte, alegre, músico e militar dedicado,
economista e professor apaixonado.
Ensinou-me desde a curiosidade em descobrir como tudo
funciona até o gosto pelas ciências exatas”*

Agradecimentos

Ao meu orientador Prof. Dr. Fernando Buarque, sábio e altivo, que sempre soube guiar-me pelos caminhos “não lineares” da pesquisa.

À minha mãe, referência de dedicação e perseverança. Ensinou-me quase tudo que sei, principalmente o gosto pela leitura.

Aos meus filhos Luiz Fellipe e Rafael Luiz, experiência enriquecedora, motivação para fazer melhor e razão para seguir sempre em frente.

À minha amada “Dulcinéa” (Anna Paula), referência de amor e dedicação, interlocutora perspicaz, sempre pronta a ouvir e dialogar. Teve muita paciência com seu cavaleiro errante “Dom Quixote”, que descobriu, em meio às estradas tortuosas, que o fracasso não é um lugar, e sim um caminho.

A todos da Polícia Rodoviária Federal, pelo dados cedidos, em especial ao agente Deiverson, sempre pronto a esclarecer minhas dúvidas.

A todos os professores da UPE, em especial à coordenadora Prof. Dra. Maria de Lourdes, que transformaram esta universidade em referência nacional e o PPGES em referência internacional.

A todos os colegas de mestrado que se transformaram em melhores amigos, em especial: “Mega”, “Rodrigão”, “Felipe San”, Dupleix, “Pastor Charles”, “Fuzzyboy”, “Pedro Malandro” e tantos outros que tornaram o ambiente do PPGES alegre, saudável e fecundo em ideias.

A Júlia, profissional dedicada e divertida, que quando não falava muito era porque algo estava errado.

À UPE pela bolsa de estudos a mim concedida através de seu órgão de fomento PFA.

Aos colegas da disciplina de Mineração de Dados na UFPE, em especial Orlando e Bruno, que se tornaram grandes amigos e interlocutores para todas as horas.

Resumo

As Rodovias federais que atravessam a Região Metropolitana e cidades do interior estão constantemente congestionadas, não apenas pela quantidade de veículos, mas por serem alvo de paralisações das mais diversas matizes, como protestos de trabalhadores, acidentes, danos na via, intempéries naturais e outros fatores de congestionamento. Em situações extremas esses problemas poderiam paralisar até a produção das fábricas no seu entorno, causando grandes prejuízos. Para dirimir alguns destes problemas, esta pesquisa tem por objetivo propor e testar conceitos para uma plataforma autoadaptável que conte tempo um modelo preditivo de comportamento das rodovias federais que atravessam o estado de Pernambuco na região Nordeste do Brasil, de modo que seja possível, antecipar eventos que possam vir a causar constrangimentos, como retenção, redução do fluxo de tráfego (gargálos) e paralisação. A fonte primária de dados desta pesquisa provém da base de dados da Polícia Rodoviária Federal de Pernambuco (PRF/PE) entre 2007 e 2015 tendo considerado veículos, traçado da via e trechos da rodovia relacionados a acidentes. Foram também utilizados dados da rede social Twitter dos últimos anos, tanto da PRF quanto de pessoas que fizeram menção a acontecimentos nas BR's (acidentes, paralisações, etc) no Estado de Pernambuco. Com base nas informações obtidas, foi realizada uma Mineração de Dados utilizando a metodologia CRISP-DM, além de Mineração de Textos para encontrar padrões comportamentais nas rodovias e em seu entorno. As tecnologias empregados para a mineração foram: Árvores de Decisão, Naïve Bayes e Redes Neurais. Os valores da área sob a curva ROC (AUC) obtidos foram acima de 0.8 o que representa um bom grau de confiabilidade. Com os dados do Twitter foram coletados todos os tweets referentes a cada palavra chave, até ao limite imposto pelo aplicativo. As tecnologias utilizadas foram Naïve Bayes, TF-IDF e, para exibir a geolocalização, foi utilizado o software de georreferenciamento Google Maps. Em comparação com abordagens usuais de navegação, o modelo de predição proposto representa um avanço em termos de mobilidade e gestão do transporte, tráfego em rodovias, uma vez que possibilita antecipar eventos e comportamentos, favorecendo a escolha de rotas alternativas e ampliando o espaço temporal de escolha para determinadas rotas.

Palavras-chave: Modelo de Predição, Mineração de dados, CRISP-DM, Controle de tráfego rodoviário.

Abstract

The federal highways that cross at Metropolitan Region of some cities are constantly congested, not only because the number of vehicles, but due to downtime, such as worker protests, accidents, natural events and other types of congestion factors. In extreme situations these problems could paralyze even the production of factories in their surroundings, causing great losses. Thus, this research aim propose and test concepts for a self-adaptive platform that contemplates a predictive model of behavior of the federal highways that cross the state of Pernambuco (Brazil), so that it is possible to anticipate events that may occur in certain stretches of highway that may cause embarrassment, such as traffic reduction and downtime. The primary data source of this research comes from the Federal Highway Police of Pernambuco (PRF/PE) database from 2007 to 2015 onwards, having considered vehicles, track layout and road sections related to accidents. Data from the social network Twitter, of the last years, both from the PRF, and from people who mentioned events in BRs (accidents, stoppages, etc.). Based on the information obtained, a Data Mining was performed using the CRISP-DM methodology to find behavioral patterns on the roads and in its surroundings. The technologies used for Mining were: Decision Trees, Naïve Bayes and Neural Networks. The values of the area under the ROC curve (AUC) obtained were above 0.8 which reflects a good degree of reliability. With Twitter data, all the tweets for each keyword were collected up to the limits of the application. The technologies used were Naïve Bayes and TF-IDF and, to display geolocation, Google Maps georeferencing software was used. Compared to usual navigation approaches, the proposed prediction model represents a breakthrough in terms of mobility and management of transportation and vehicle , since it makes it possible to anticipate events and behaviors, in order to favor the selection of alternative routes.

Keywords: Prediction Model, Data Mining, CRISP-DM, Road Traffic Control

Lista de Abreviações e Siglas

ADALINE	<i>Adaptative Linear Neuron</i>
MADALINE	<i>Many Adaline</i>
API	<i>Application Programming Interface</i>
BG	<i>Big Data</i>
DM	<i>Data Mining</i>
KDD	<i>Knowledge Discovery Databases</i>
TM	<i>Text Mining</i>
TDM	<i>Text Data Mining</i>
KDT	<i>Knowledge Database Text</i>
NB	<i>Naïve Bayes</i>
TF-IDF	<i>Text Frequency - Inverse Document Frequency</i>
CRISP-DM	<i>Cross Industry Standard Process for Data Mining</i>
PRF	<i>Polícia Rodoviária Federal</i>
BPRv	<i>Batalhão de Polícia Rodoviária (estadual)</i>

Lista de Figuras

2.1	Domínio das técnicas aplicadas a mineração de dados	20
2.2	O padrão CRISP-DM (1)	20
2.3	Entendimento do negócio	21
2.4	Entendimento dos dados	22
2.5	Preparação dos dados	23
2.6	Modelagem IA	24
2.7	Avaliação do modelo	25
2.8	Implantação do modelo	26
2.9	Fases da mineração de dados até extração do conhecimento	28
2.10	Mapa mental da Mineração em textos	30
2.11	Grafo de uma rede do Twitter	31
2.12	Descrição da conta Twitter das bibliotecas acadêmicas	34
2.13	Bibliotecas de Universidades e contas no Twitter	35
2.14	Frequência de palavras	35
2.15	Árvore de decisão	40
2.16	Árvore da família TDIDT	41
2.17	Perceptron de McCulloch e Pitts	44
2.18	Rede Neural	46
3.1	Etapas da modelo proposto	52
3.2	Matriz de Mortos 2D - BR 101, Período: 2007 à 2015	54
3.3	Matriz de Gravidade 3D3 ^a -feira - BR 101, Período: 2007 à 2015	55
3.4	Arco cibernético com dados do Twitter	56
3.5	Etapas da metodologia	57
3.6	Rota: Goiana - Recife, a partir da Matriz de Gravidade 3D para 2 ^a -feira .	58
4.1	Etapas da modelo proposto	59
4.2	BR 101: Hora do acidente (1) Concentração em torno da hora (2)	63
4.3	Frequência	63
4.4	BR 104: Hora do acidente (1) Concentração em torno da hora (2)	64
4.5	Frequência	64
4.6	BR 110: Hora do acidente (1) Concentração em torno da hora (2)	65
4.7	Frequência	65
4.8	BR: 116 Hora do acidente (1) Concentração em torno da hora (2)	66
4.9	Frequência	66
4.10	BR 232: Hora do acidente (1) Concentração em torno da hora (2)	67
4.11	Frequência	67
4.12	BR 316: Hora do acidente (1) Concentração em torno da hora (2)	68
4.13	Frequência	68

4.14	BR 407: Hora do acidente (1) Concentração em torno da hora (2)	69
4.15	Frequência	69
4.16	BR 408: Hora do acidente (1) Concentração em torno da hora (2)	70
4.17	Frequência	70
4.18	BR 423: Hora do acidente (1) Concentração em torno da hora (2)	71
4.19	Frequência	71
4.20	BR 424: Hora do acidente (1) Concentração em torno da hora (2)	72
4.21	Frequência	72
4.22	BR 428: Hora do acidente (1) Concentração em torno da hora (2)	73
4.23	Frequência	73
4.24	Tipo de Veículo X Num. Acidentes	74
4.25	Traçado da via X Num. Acidentes	75
4.26	Árvore de Decisão gerada pelo Knime	79
4.27	Km 70, BR 101 (Sul) Pernambuco	81
4.28	Matriz de Mortos: Km 56 – 78, BR 101 (Sul) Pernambuco	81
4.29	Km 64 e 67, BR 104 Pernambuco	82
4.30	Matriz de Mortos: Km 64 e 67, BR 104 - Rota da Sulanca - Pernambuco .	82
4.31	Matriz de Gravidade 3D: Br 116 predição para quinta-feira	83
4.32	Rede Neural utilizada na classificação	86
4.33	Resultado da classificação feita pela Rede Neural – acurácia	87
4.34	Gráfico de frequência de palavras – unigramas	88
4.35	Gráfico tipo Nuvem de Palavras referente aos unigramas	89
4.36	Dendograma de Agrupamento (clustering) do resultado da mineração . .	89
4.37	Gráfico da Agrupamento (clustering) do resultado a mineração	90
A.1	Mapa rodoviário de Pernambuco	96
A.2	Etapa 1 – Coleta e união das bases históricas de acidentes e interdições da Polícia Rodoviária Federal - PE, entre 2007 a 2015	98

Lista de Tabelas

2.1	Mineração de dados – contexto de aplicação (1)	19
2.2	Volume de dados no mundo	31
2.3	Matriz de Confusão	45
2.4	Matriz modelo de Confusão	45
3.1	Variáveis transformadas	53
3.2	Rodovias BRs que atravessam Pernambuco	57
4.1	Detalhe da acurácia para classe Tipo Acidente	75
4.2	Matriz de confusão para a variável Tipo de acidente	75
4.3	Detalhe da acurácia para classe Gravidade	76
4.4	Matriz de confusão para a variável Gravidade	76
4.5	Detalhe da acurácia para classe BR	77
4.6	Matriz de confusão para a variável BRajustada	77
4.7	Linhda da Matriz de confusão correspondente a BR 101	78
4.8	Matriz de confusão para a variável BRajustada	78
4.9	Matriz de confusão para a variável BRajustada	78
4.10	Instâncias classificadas e Erro médio	84
4.11	Detalhe da acurácia para classe BRajusta	84
4.12	Matriz de confusão – BRajustada – Naïve Bayes	85
4.13	Matriz de confusão – BRajustada – Naïve Bayes	85
4.14	Matriz de confusão – BRajustada – Naïve Bayes	85
A.1	Variáveis originais da base de acidentes	97
A.2	Variáveis originais da base de interdições	98

Sumário

1	Introdução	14
1.1	Contextualização	14
1.2	Justificativa	15
1.3	Motivação	16
1.4	Objetivo Geral	17
1.4.1	Objetivos Específicos	17
1.4.2	Organização dessa dissertação	17
2	Revisão da Literatura	18
2.1	Introdução	18
2.2	Mineração de Dados e CRISP-DM	18
2.2.1	Contexto de aplicação do CRISP – DM	19
2.2.2	Ciclo de vida do CRISP-DM	20
2.3	Mineração de dados	27
2.4	Mineração em Dados Textuais	29
2.4.1	Mineração em Textos nas Redes sociais	30
2.5	Aprendizagem de Máquina	37
2.5.1	Tipos de Aprendizagem	37
2.5.2	Algoritmos de Aprendizagem de Máquina	38
2.6	Métricas aplicadas à mineração	45
2.6.1	Classificação para análise preditiva	46
2.6.2	Pesquisas envolvendo inteligência artificial e tráfego em rodovias .	47
3	Contribuição	50
3.1	Modelo Proposto	51
3.2	Reflexão sobre as tecnologias utilizadas no modelo preditivo	52
3.3	Extração do conhecimento - KDD	53
3.4	Arco cibernético com dados do Twitter	55
3.5	Extrapolação para georreferenciamento	56
4	Simulação	59
4.1	Execução do modelo	59
4.2	A construção do Modelo preditivo	61
4.2.1	Aplicação do CRISP-DM	61
4.2.2	Dados encontrados antes da Mineração	62
4.2.3	Dados encontrados após a Mineração	74
4.2.4	Métrica dos classificadores	74
4.2.5	Resultado dos outros classificadores utilizados nesta pesquisa	84
4.3	Acoplamento com a estrutura dinâmica	87

4.3.1	Mineração em texto no Twitter	88
5	Considerações finais	91
5.1	Trabalhos futuros	94
A	Preprocessamento	96
A.0.1	Mapa das de Pernambuco com as rodovias	96
A.1	Coleta e Preprocessamento dos dados da PRF	97
	Referências Bibliográficas	99

*“E se o mundo não corresponde
em todos os aspectos a nossos desejos,
é culpa da ciência ou dos que querem
impor seus desejos ao mundo?”
(Carl Sagan)*

1

Introdução

1.1 Contextualização

O século XXI caracteriza-se como sendo a Era do crescimento exponencial da informação. Esta informação é disponibilizada das mais diversas formas, promovendo uma explosão de fatos e eventos aos quais as pessoas têm acesso. Todavia, justamente por haver uma infinidade de informações, é necessário que elas sejam agrupadas, tratadas, contextualizadas, sob pena de, na ausência de uma organização, as informações se perderem e não oferecerem às pessoas a possibilidade de extrair delas o que é necessário, para uso adequado em sociedade.

Uma questão relevante, que merece destaque é a distinção de Informação e Conhecimento. Xavier e Costa (2) discutem sobre essa questão no âmbito da Ciência da Informação (CI). Inicialmente, esses autores defendem que a Ciência da Informação surge como campo de investigação a partir da explosão de informações no contexto tecnológico, em meados do século XX, e ampliada com o advento da chamada "sociedade da informação".

Le Coadic (3) discute que a Ciência da Informação é um fenômeno sociocognitivo e, por isso mesmo, faz com que a CI necessite de uma abordagem interdisciplinar. Outra abordagem trata a informação como “coisa” (information-as-thing) ou como estrutura (2). Essa abordagem estaria relacionada a um paradigma fisicista, em contraposição ao paradigma sociocognitivo.

Embora alguns pesquisadores utilizem os dois conceitos – informação e conhecimento – neste contexto são colocados como sinônimos, outros recorrem à Epistemologia para diferenciar ambos (2). A primeira perspectiva entende que quanto mais conhecimento, mais informação; e que novos conhecimentos possibilitam a produção de novas informações, numa movimento cíclico. A segunda perspectiva, embora não desconsidere a lógica existente na afirmação acima, entende que o conhecimento é produzido quando as informações são articuladas, em forma de teia, de rede, produzindo algo novo, que não estava contido em cada informação individualmente, mas que se caracteriza pela extração delas.

Essa segunda visão é que acatamos para esta pesquisa. A nossa proposta é de articular as informações produzidas pelas instituições ligadas ao trânsito em Pernambuco

(PRF/PE), as que são produzidas pela rede social Twitter, e, com as tecnologias de IA – Árvores de Decisão, Redes Neurais, dentre outras – promover uma extração, produzindo conhecimentos que serão disponibilizados em um modelo de predição.

As informações podem ser produzidas tanto por seres humanos, quanto por máquinas. Segundo Nobert Wiener (4), a informação tem tanta importância quanto a energia e a matéria. A informação pode ser utilizada para controlar sistemas baseados em comportamento biológico ou mecânico. Esse comportamento, quando controlado por meio de realimentação, tem como alvo atingir um objetivo, um propósito, como compreender, controlar, predizer.

Os dados produzidos pelo ser humano atualmente dobram a cada cinco anos (5). As redes sociais, muito mais do que um ambiente lúdico, se configuram como um espaço onde as pessoas vão buscar informações para a gestão dos seus problemas cotidianos, bem como um lugar de coleta de informações para sistemas inteligentes proporem soluções mais adequadas à problemática humana e, ao mesmo tempo, com rapidez.

Dados governamentais têm sido disponibilizados pelo governo brasileiro desde que este aderiu em 2009 ao movimento mundial para incentivar as autoridades dos países a maior transparência e participação popular conhecido, este movimento foi conhecido como “Open Data” (6). Desde então o Brasil vem se esforçando para disponibilizar informações governamentais para todos os cidadãos. Os dados utilizados nesta pesquisa também podem ser encontrados pelo Sistema BR-Brasil, da Polícia Rodoviária Federal (7) (8). A Polícia Rodoviária Federal regista diariamente boletins de ocorrência, contudo, dados produzidos eletronicamente só estão disponíveis a partir de 2007.

1.2 Justificativa

A inteligência artificial é uma área que pode, através de algoritmos eficientes, propor soluções adaptativas para dar conta das mais diversas necessidades humanas, sobretudo aquelas relacionadas ao contexto social, como logística de transporte, locomoção de pessoas, gestão de tempo, dentre outros. A Mineração de Dados (MD) vai buscar na Inteligência Artificial algoritmos para descoberta de padrões e automatizar tarefas de investigação dos dados. Essa automatização também conhecida como “Machine Learning” aplica-se a quase todos os caminhos na descoberta do conhecimento oferecida pela MD (9).

Uma instância da problemática descrita acima será central nesta pesquisa: os inúmeros acidentes de trânsito que constrange o transporte de mercadorias o fluxo normal de veículos ligeiros e a segurança dos pedestres. Para isso foi necessária a integração de bases de dados heterogêneas disponíveis em computadores de órgãos públicos que contêm informações de qualidade para gerar um modelo preditivo de roteamento logístico de transporte. Para isso foram considerados dados históricos de cada rodovia, com os trechos onde há mais retenções que causam constrangimento nessas vias em determinados períodos do dia, que se repetem em meses e ao longo dos anos, tais como acidentes, protestos, intempéries ambientais. De forma complementar, foram utilizados informações de redes sociais, como o Twitter. A escolha dessa rede social se deu pelo fato de que um dos seus principais objetivos é o de compartilhar informações sucintas e pontuais entre os seus usuários, boa parte delas sobre eventos que influenciam o cotidiano das pessoas.

1.3 Motivação

As rodovias federais que atravessam a região metropolitana do Recife (RMR) e interior do Estado de Pernambuco estão constantemente congestionadas, não apenas pela quantidade de veículos, mas por serem alvo de paralisações das mais diversas matizes, como protestos de trabalhadores, acidentes, danos na via, intempéries naturais e outros tipos de constrangimentos que interferem no fluxo de veículos. Em situações extremas poderiam paralisar até a produção das fábricas no seu entorno (10).

A RMR é a 5^a região mais populosa do Brasil, concentra 3.690.485 habitantes (dados de 2012) em 14 municípios, além da Zona da Mata Norte (ZMN) com 577.191 habitantes e a Zona da Mata Sul (ZMS) com 733.447 habitantes (11). Nessas regiões (RMR, ZMN e ZMS) a frota (automóveis particulares, ônibus, caminhões, motocicletas, tratores e outros veículos) foi contabilizada, em 2015, com mais de 1.270.000 veículos (12). Se considerarmos o interior do estado, essa frota aumenta para mais de 2.700.000 veículos, distribuídos nas regiões do Agreste e Sertão. Algumas cidades se destacam por concentrarem uma frota maior, como Caruaru, no agreste pernambucano, com mais de 150.000 veículos, e Petrolina, no sertão, com quase 130.000.

O que acontece nas grandes cidades do Estado de Pernambuco e no seu entorno é frequente nas grandes cidades brasileiras. Câmeras de monitoramento de trânsito, redes sociais, aplicativos de celular e outros dispositivos, fornecem uma grande quantidade de informações diárias sobre o que acontece nas rodovias e no entorno delas. Essas informações são transformadas em dados eletrônicos, alguns podem ser armazenados em bases de dados históricas, alimentando centros de repositórios conhecidos como *Datacenters* outras devido a grande quantidade, a variedade e a velocidade em que esses dados são produzidas não é possível armazená-las, esse conceito de informação é conhecido como *Big data*¹.

Fora do perímetro urbano as rodovias atravessam outras localidades com problemáticas diversas, tais como pavimento ruim ou ausência de pavimentação, traçados inapropriados e outras intempéries têm causado frequentemente acidentes. A Polícia Rodoviária Federal ou outros órgãos de controle público atendem e registram esses acontecimentos em boletins diários.

A proposição de uma solução para considerar parte dessas informações: dados da PRF, e de outros órgãos de controle de tráfego, requer várias etapas, que engloba algumas técnicas de mineração de dados que foram discutidas nesta pesquisa. Propomos, nesta pesquisa, uma solução para utilização e integração das rotas existentes, definida por critérios cientificamente estudados, que seja materializado em informações para um modelo de predição. Isso poderá ser de suma importância para solucionar a problemática do tráfego em rodovias, fornecendo apoio para que o veículo siga até seu destino de maneira segura e no menor tempo possível.

¹Big Data é um conceito de dados eletrônicos disponíveis na Internet em grande volume, variedade e que precisam de ser acessados com processamento rápido: velocidade. Aos dados que possuem esta característica (volume, variedade e necessidade de velocidade em seu tratamento) chamamo-la de *Big Data* (5) (13)

1.4 Objetivo Geral

Essa pesquisa tem como objetivo principal desenvolver um modelo preditivo de suporte à decisão para a problemática das retenções crescentes nas rodovias pernambucanas. Para isso, propomos uma solução com múltiplas perspectiva através da integração de fontes de dados via diversas tecnologias disponíveis que vão desde a análise dos dados históricos das rodovias à utilização informações de redes sociais e dados governamentais.

1.4.1 Objetivos Específicos

- Caracterizar a problemática de cada rodovia;
- Desenvolver um modelo preditivo dos fenômenos que envolvem as rodovias;
- Desenvolver um ambiente de simulações interativas da estrutura viária e sua dinâmica.
- Propor soluções para melhor experiência dos usuários que utilizam as rodovias federais pernambucanas.

1.4.2 Organização dessa dissertação

Os próximos capítulos dessa dissertação estão organizados da seguinte forma:

No Capítulo 2, a Revisão da Literatura inicia-se com a apresentação do CRISP-DM e todas as suas fases, o contexto da aplicação, ciclo de vida; a Mineração de Dados e suas técnicas empregues nesta pesquisa; mineração de textos e a rede social Twitter; Levantamento de pesquisas envolvendo IA aplicada ao problemas de tráfego em rodovias.

No Capítulo 3, apresentamos nossa Contribuição e a abordagem metodológica.

O Capítulo 4 trata da simulação e execução do modelo proposto, bem como a articulação entre as diversas tecnologias e a discussão dos resultados encontrados.

No Capítulo 5 são apresentadas as considerações finais e propostas de futuras pesquisas.

2

Revisão da Literatura

2.1 Introdução

Neste capítulo, apresenta uma revisão teórica e de pesquisas contemplando três campos, a saber:

- Processo de mineração de dados com respeito à pesquisa em lide;
 - Contextualização do CRISP-DM.
 - Ciclo de vida do CRISP-DM.
- Processo de aplicação da Mineração de Dados e Mineração em textos;
 - Mineração de dados e o Processamento de Linguagem Natural.
 - Mineração de texto em Redes sociais; o Twitter.
- Aprendizagem de máquina em processos de mineração de dados;
 - Naive Bayes
 - Árvore de Decisões.
 - Medição de desempenho e classificação de algoritmos de mineração de dados.

2.2 Mineração de Dados e CRISP-DM

O “CRoss Indrustry Standard Process for Data Mining” – CRISP-DM é um processo para mineração de dados que descreve como especialistas nesse campo aplicam as técnicas de mineração para obter os melhores resultados (1). O CRISP-DM é um processo recursivo, onde cada etapa deve ser revista até quando o modelo apresentar os resultados satisfatórios, preliminarmente definidos. O Analista de Dados ou o Cientista de Dados é o profissional que acompanha e executa o processo.

Esse processo foi concebido, desenvolvido e refinado através de “workshops” entre 1996 e 1999 (1), por três entidades empresariais europeias que formavam um consórcio. Um dos parceiros, a Daimler-Chrysler AG (Alemanha), estava, à época, à frente da maioria das organizações empresariais e comerciais na aplicação de mineração de dados em seus negócios. A SPSS Inc.(EUA), era responsável serviços baseados em mineração de dados desde 1990, tendo lançado o primeiro workbench de mineração de dados comerciais o

Clementine®. E a NCR Systems Engineering Copenhagen (EUA e Dinamarca), com o Teradata®, uma Datawarehouse que estabelecia equipes de consultores especialistas em mineração de dados para atender a seus clientes. Hoje mais de 300 empresas contribuem para o modelo de processo CRISP-DM (14).

2.2.1 Contexto de aplicação do CRISP – DM

A aplicação do CRISP-DM (1) é guiada desde o nível mais genérico até o nível mais especializado, sendo normalmente explicado em quatro dimensões:

- O domínio da aplicação – a área específica onde o projeto de mineração de dados acontece;
- O tipo de problema – descreve as classes específicas do objetivo do projeto de mineração de dados;
- Os aspectos técnicos – cobrem as questões específicas acerca dos desafios usualmente encontrados durante o processo de mineração de dados;
- As ferramentas e técnicas – dimensão específica que cada ferramenta/técnica de mineração de dados é aplicada durante o projeto.

A tabela 2.1 sumariza e exemplifica essas dimensões no contexto de aplicação do CRISP-DM.

Tabela 2.1: Mineração de dados – contexto de aplicação (1)

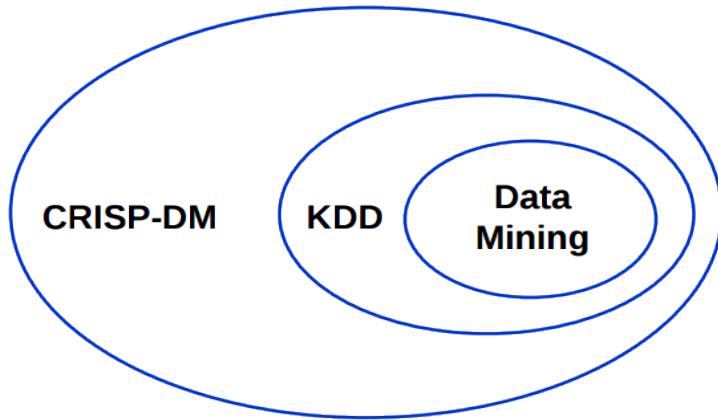
Dimensão	Domínio da aplicação	Tipo de Problema	Aspecto técnico	Ferramentas e Técnicas
Exemplo	Modelo de resposta	Descrição e summarização	Dados faltantes	Clementine
—	Predição agitada	Segmentação	<i>Outlies</i>	MineSet
—	—	Descrição do conceito	—	Árvore de decisão
—	—	Classificação	—	—
—	—	Predição	—	—
—	—	Análise de dependências	—	—

Fonte: CRISP-DM – 1.0 (adapatado)

A tabela 2.1 descreveu que um projeto de Mineração de Dados a ser desenvolvido deverá estar contextualizado nestas quatro dimensões. Mais adiante o ciclo de vida do CRISP-DM (figura 2.2) descreve as fases onde estas dimensões da tabela 2.1 se encaixam. A aplicação das técnicas de mineração de dados identifica padrões ocultos nos dados, inacessíveis pelas técnicas tradicionais, como por exemplo, consultas em banco de dados, técnicas estatísticas, dentre outras. Além disso, possibilita analisar um grande número de variáveis simultaneamente, o que não acontece com o cérebro humano (15). A análise desse processo permite extrair novos conhecimentos a partir dos dados, que é tratado na literatura como KDD – Knowledge Discovery Database (16). Fayyad (16) destaca a natureza interdisciplinar do KDD que contempla a intersecção de campos de pesquisa tais como Aprendizagem de Máquina (Machine Learning), Reconhecimento de Padrões, I.A., Estatística, Computação de Alto Desempenho e outros, propõe que o objetivo principal é

extraí um conhecimento de alto nível a partir de dados de baixo nível num contexto de grandes bases de dados. O CRISP-DM, por sua vez, engloba todos esses elementos como pode ser visto na figura 2.1:

Figura 2.1: Domínio das técnicas aplicadas a mineração de dados



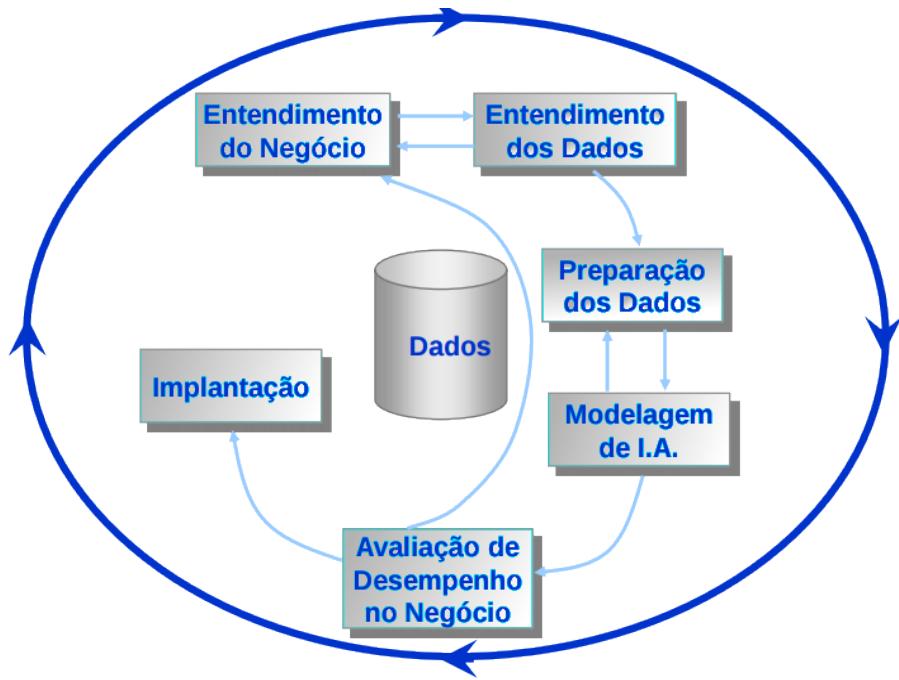
Fonte: Neurotech – 2012

2.2.2 Ciclo de vida do CRISP-DM

O modelo de processo CRISP-DM provê seis fases para um projeto de mineração de dados, sendo assim determina-se um ciclo de vida compreendido para cada uma dessas fases:

A figura 2.2 ilustra as fases do ciclo:

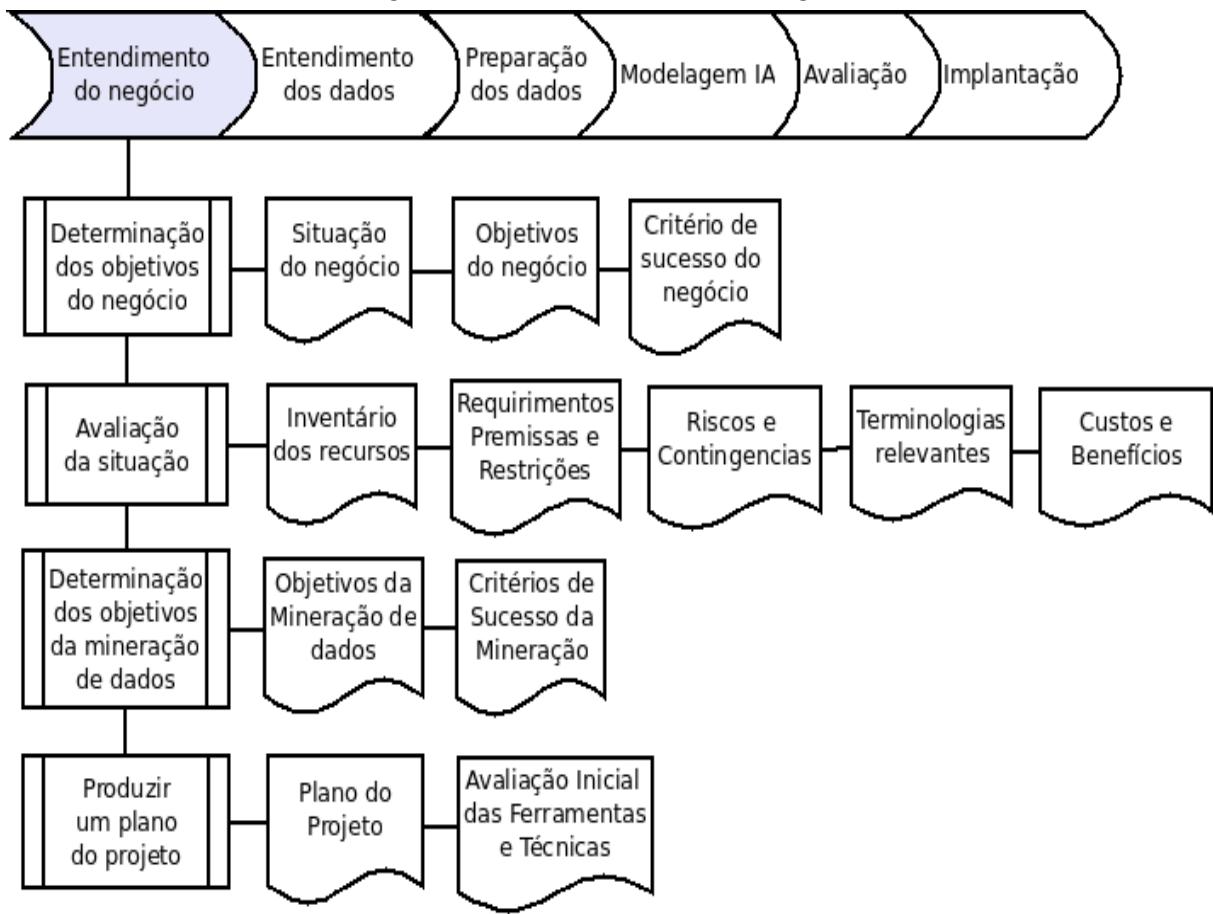
Figura 2.2: O padrão CRISP-DM (1)



Fonte: CRISP-DM 1.0

A primeira fase, conhecida como **Entendimento do negócio**, ou “fase de entendimento dos objetivos e dos requerimentos sob a perspectiva do negócio” (1) é uma fase crucial da mineração, um especialista (ou muitos) deve ser consultado. O analista de dados e o analista do negócio traçam os objetivos da mineração sob a perspectiva do cliente. Questionamentos incorretos ou negligência nesta fase podem acarretar esforços excessivos no processo como um todo a experiência de um profissional da área é condição “sine qua non” nessa fase. Portanto avaliar o negócio, avaliar a situação sob o ponto de vista dos riscos de não conclusão do processo, determinar os objetivos e traçar um plano para execução. Essas etapas são delineadas nas figuras que se seguem. A figura 2.3 é sobre o Entendimento do negócio já descrito anteriormente.

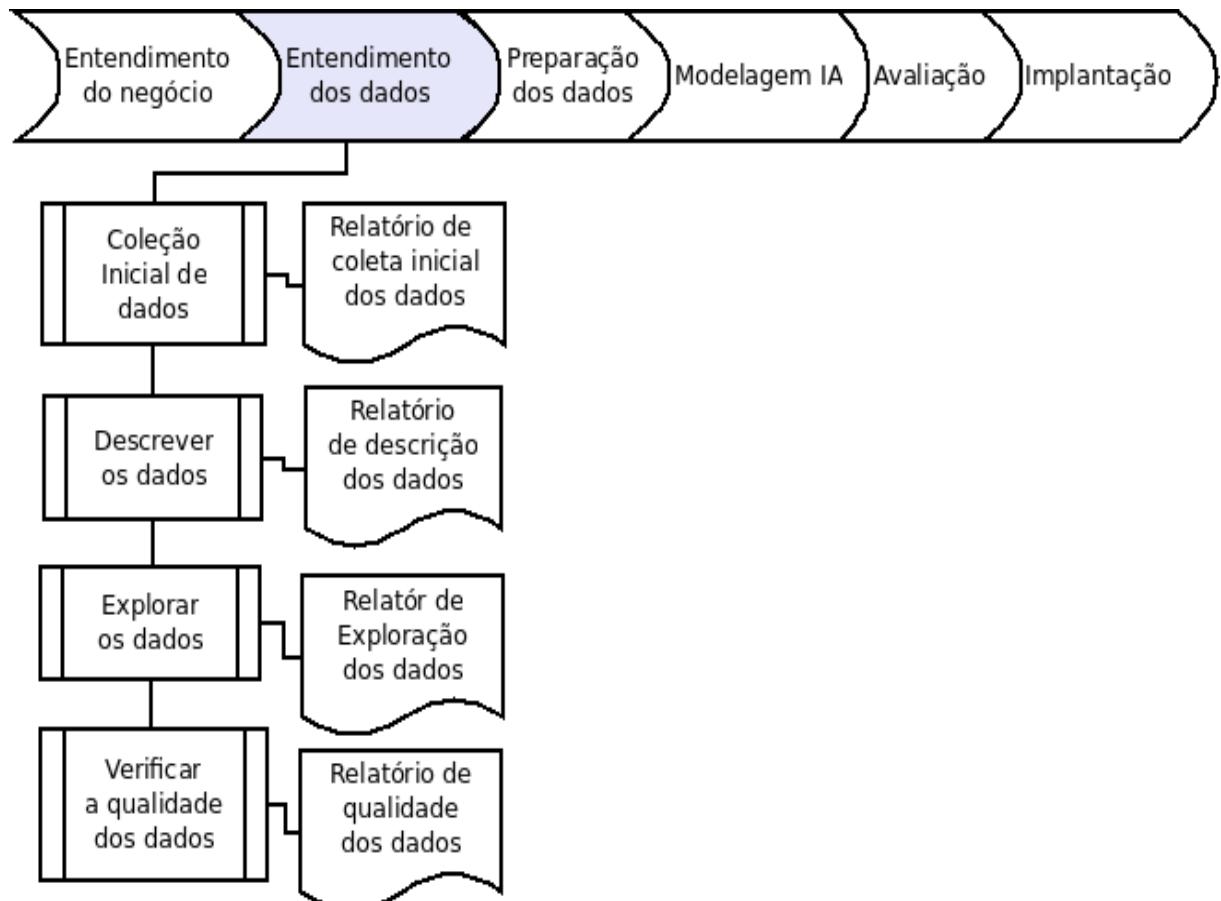
Figura 2.3: Entendimento do negócio



Fonte: CRISP-DM 1.0 (adapatado)

Em seguida, o analista de dados passa à segunda fase, **Entendimento dos dados**, na figura 2.4. Essa fase caracteriza-se pelo exame acurado dos dados, procurando identificar sua qualidade. Dados ausentes – “missing data” – são comuns em bases de dados não estruturadas, configurando-se como um problema a ser considerado, pois seu tratamento pode consumir muito tempo do analista de dados, estima-se cerca de 80% do tempo total (17).

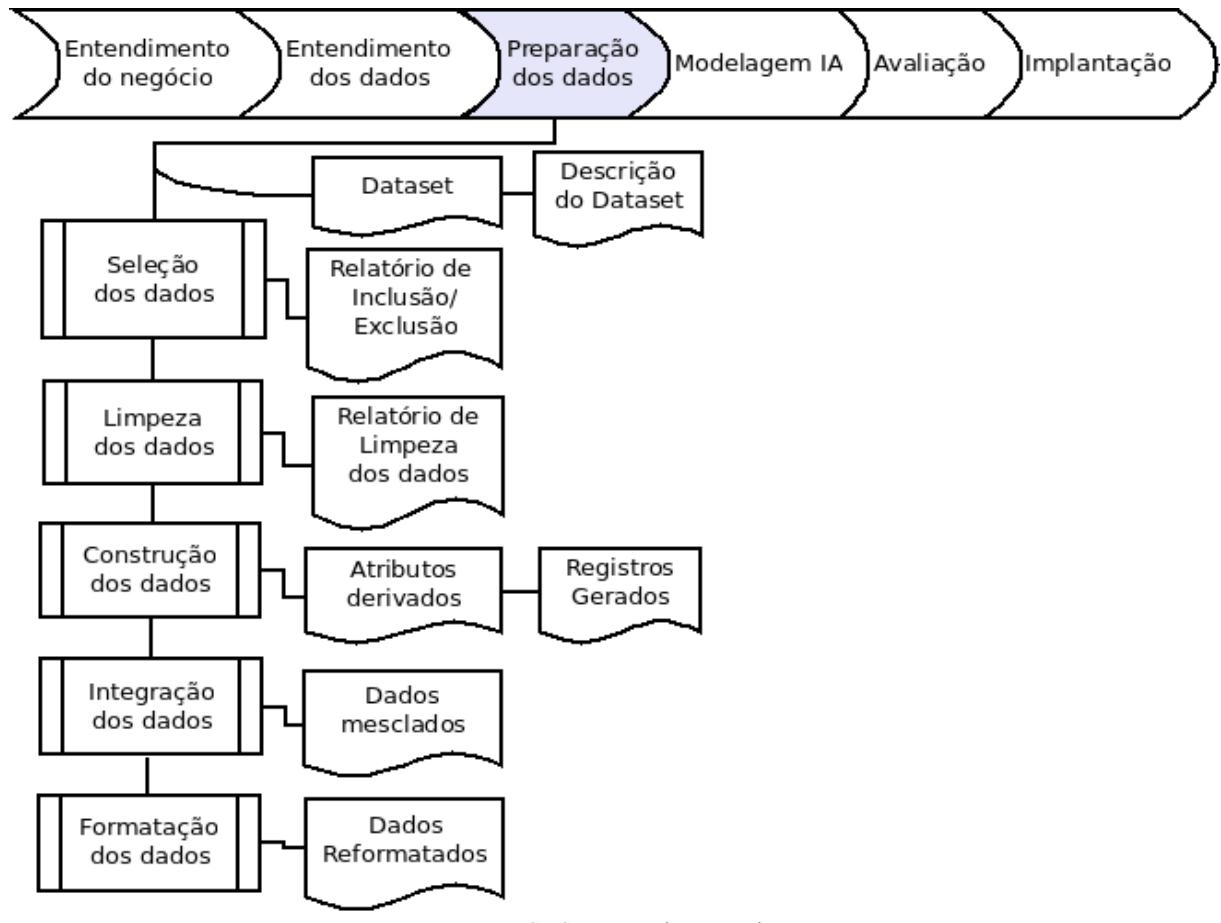
Figura 2.4: Entendimento dos dados



Fonte: CRISP-DM 1.0 (adapatado)

A terceira fase, **Preparação dos dados**, figura 2.5, diz respeito à construção final do conjunto de dados. Preparar os dados significa, criar e selecionar atributos, criar tabelas ou planilhas e registros dos dados. Para selecionar quais dados serão mais relevantes, calcula-se, por exemplo, o coeficiente de correlação linear entre os atributos (variáveis), quando as variáveis são numéricas. Outra forma de qualificar os dados é calculando a quantidade de informação que cada atributo possui. A máxima entropia de cada atributo pode fornecer informações sobre a qualidade da variável quando esta estabelece ganho de informação (18), vide equação da Entropia ¹: $H_x = -\sum_{x \in X} P(x) \log_2 P(x)$ Onde H_x é a medida de entropia, x um atributo do conjunto de variáveis X de variáveis.

Figura 2.5: Preparação dos dados

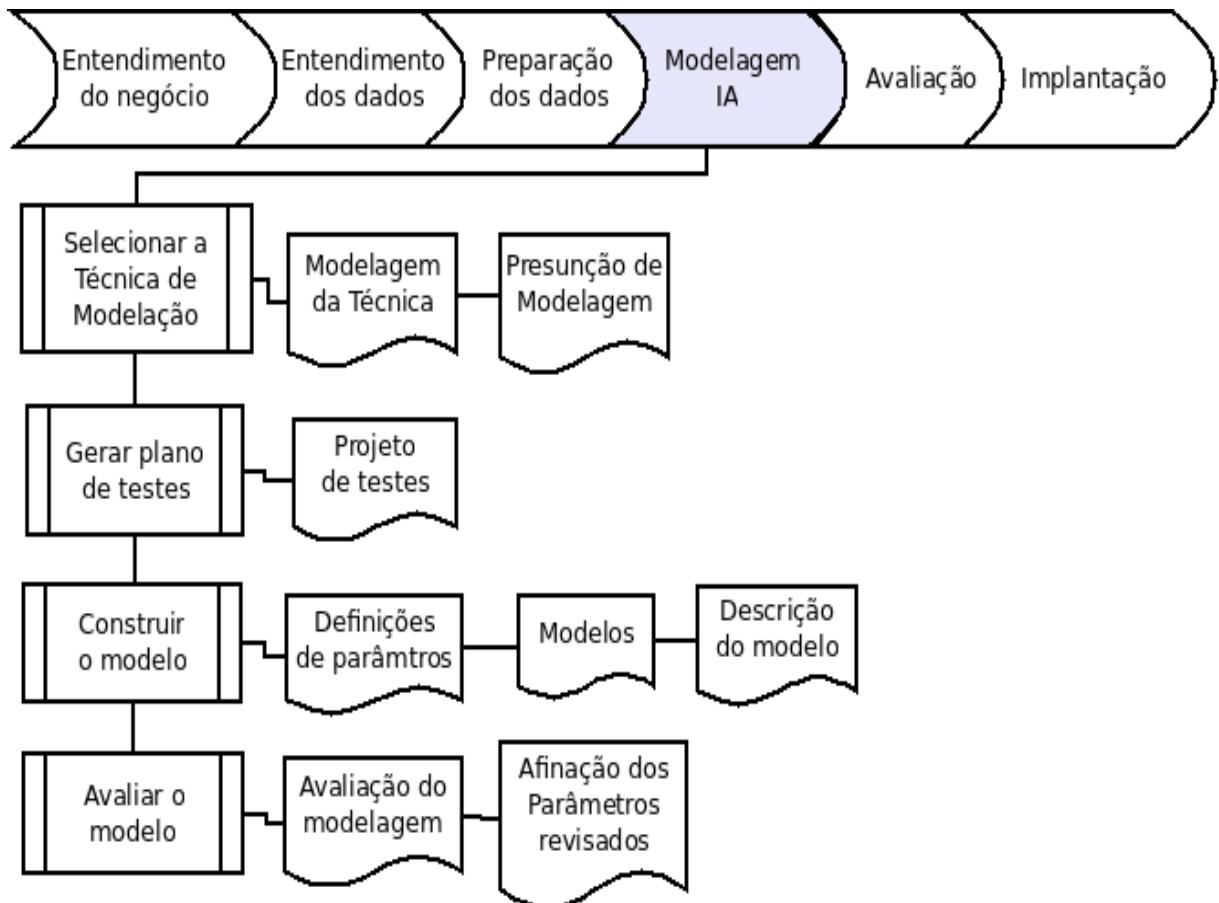


Fonte: CRISP-DM 1.0 (adaptado)

¹O conceito de entropia será discutido na seção 2.6.2, referente a Árvores de Decisão

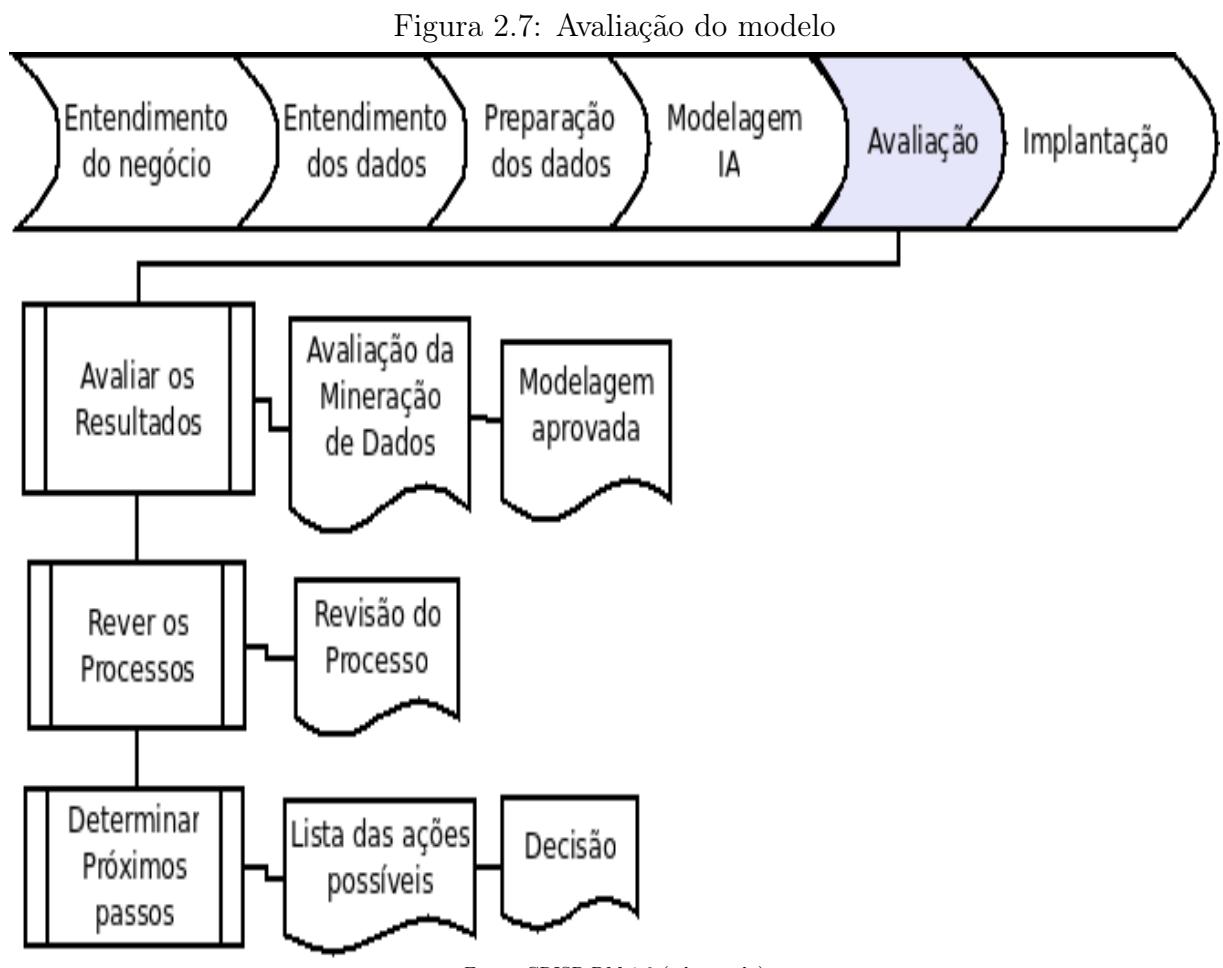
Na quarta fase, **Modelagem de I.A.**, figura 2.6, a tecnologia deve ser escolhida de forma criteriosa, baseada também na experiência do analista de dados. Em sistemas de suporte à decisão, uma tecnologia inadequada pode levar a decisões imprecisas. É comum retornar às fases anteriores para adequar a técnica aos dados. Um modelo de regressão logística para problemas binários, redes neurais para problemas de classificação, e assim por diante.

Figura 2.6: Modelagem IA



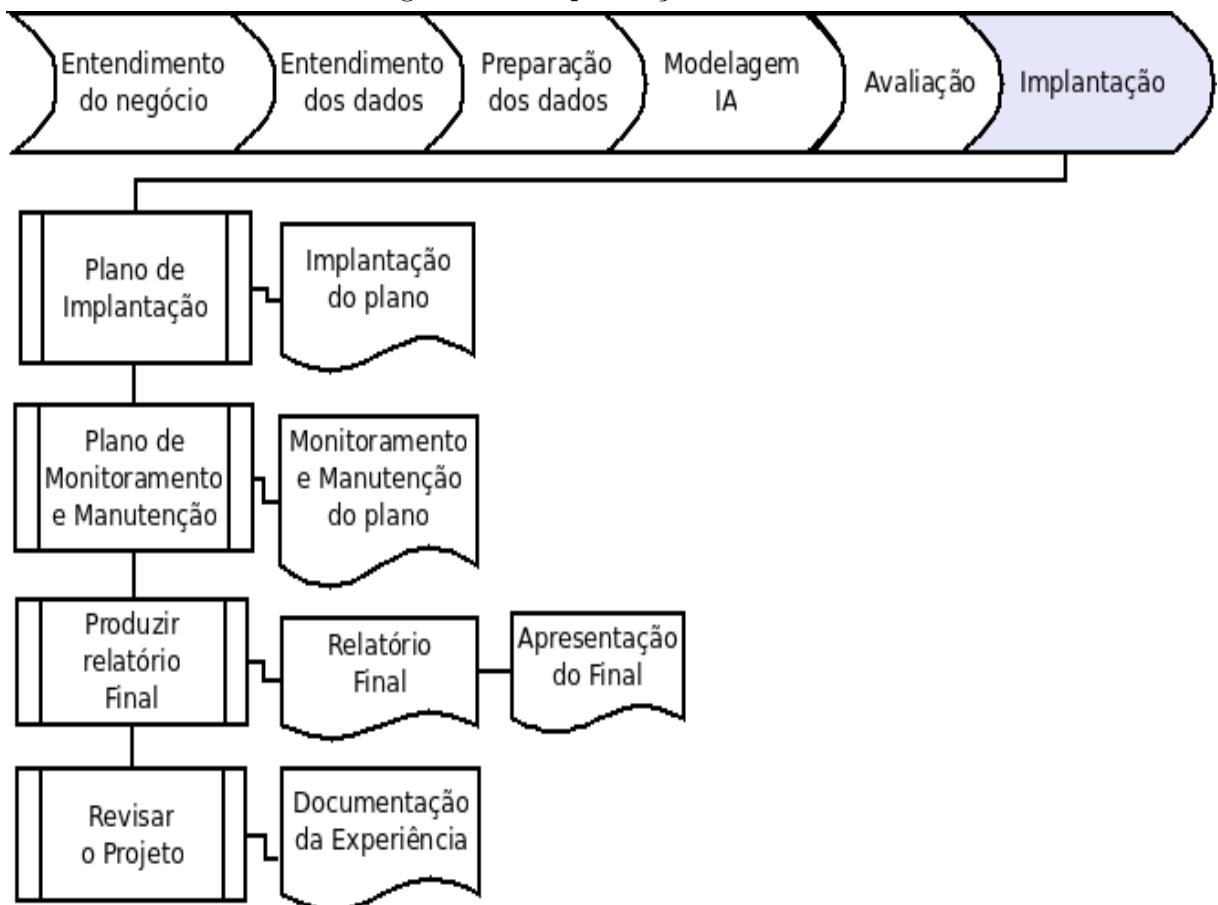
Fonte: CRISP-DM 1.0 (adaptado)

Na fase cinco, **Avaliação de desempenho**, figura 2.7, um ou muitos modelos devem ter sido construídos e testados, de forma que seja possível atingir uma alta qualidade do ponto de vista da análise dos dados, ou seja, que o modelo proposto esteja de adequado aos objetivos do negócio. Para tal é preciso que antes do desenvolvimento final do modelo, os passos executados até então sejam avaliados e revistos.



A sexta e última fase caracteriza-se pela conclusão do modelo, figura 2.8. No entanto, a criação do modelo não é o fim do processo. O conhecimento adquirido precisa ser incrementado, organizado e apresentado de maneira que o cliente possa usá-lo. É importante ressaltar que este ciclo poderá ser retomado até que o modelo esteja adequado às necessidades e especificidades do cliente.

Figura 2.8: Implantação do modelo



Fonte: CRISP-DM 1.0 (adapatado)

2.3 Mineração de dados

No processo de extração do conhecimento, também conhecido como *Knowledge Discovery Databases* (KDD), um dos importantes passos a ser considerado é a mineração de dados, que se caracteriza pela aplicação de algoritmos específicos para descoberta de padrões e/ou comportamentos em grandes bases de dados, também conhecido como repositórios de dados (16).

A mineração se distingue das técnicas estatísticas pelo fato de que não trabalha com dados hipotéticos, mas se apoia nos próprios dados para extrair os padrões (19).

FAYYAD (16), destaca que é necessário distinguir claramente KDD e mineração de dados. Enquanto que a descoberta de conhecimento em bases de dados (KDD) é um processo dividido em etapas bem distintas (Figura 2.9), a Mineração de Dados (MD) é um passo no interior desse processo. Todavia, esse passo é de considerável relevância para que se possa extrair conhecimento adequadamente. A aplicação “cega” dos métodos de mineração de dados, ainda segundo Fayyad (16), pode conduzir à descoberta de dados sem significado e a padrões inválidos.

Nos repositórios dos dados abertos do governo federal (6) existem vários tipos de dados e informações, nesses repositórios que podem ser minerados, contudo esses dados, inicialmente são selecionados e agrupados, a seguir passam por uma fase de preprocessamento, que consiste em tratá-los de forma a prepará-los para a mineração. Essa fase é de fundamental importância na estruturação dos dados, uma vez que em grandes volumes de dados, também conhecido “Datawarehouse”² (20), podem existir inconsistências, faltas (missing data) ou duplicidade e erros de informações.

As técnicas de mineração de dados trabalham com dados estruturados, preenchidos em sua totalidade sem *missing data*, para poder extrair informações relevantes. Existem várias maneiras de se contornar os dados ausentes, como o preenchimento dos dados através de técnicas de inteligência artificial, da média dos valores; em dados numéricos ou com a moda; quando os dados forem categóricos³. Os dados são apresentados como uma tabela onde a primeira coluna ficam as variáveis (qualitativas ou quantitativas), em seguida vêm os dados que seguem segundo a primeira linha a correspondência de que tipo eles são.

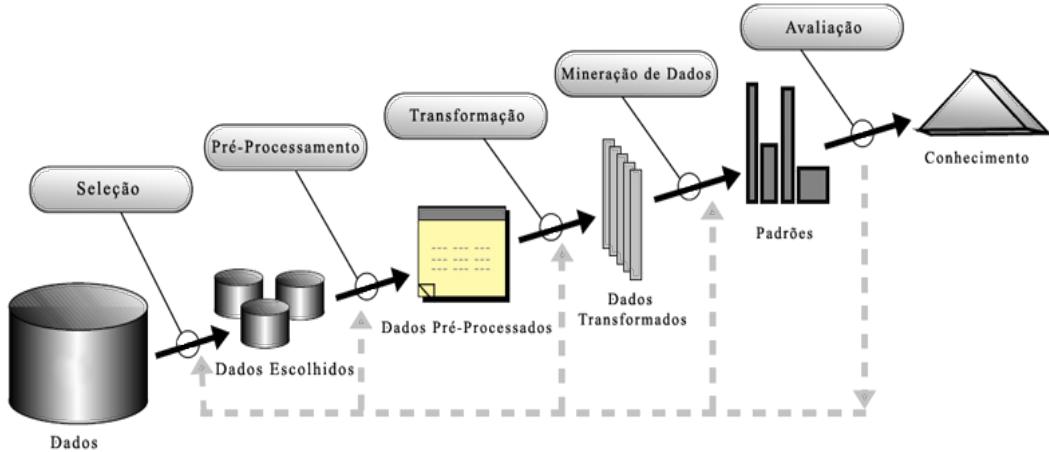
Normalmente são criadas variáveis a partir dos dados numéricos para ampliar o conhecimento sobre esses dados. Para cada tipo de dado (numéricos, categóricos) existem técnicas apropriadas para serem aplicadas sobre eles, algumas mais sensíveis às problemáticas elencadas anteriormente e outras mais robustas (21), que por sua vez estão associadas a classes de problemas que a mineração trata, a tabela 2.1 (página 19) delineou o domínio.

O caminho desde a coleta dos dados até sua mineração e extração de conhecimento é longo, a figura 2.9 tem a ilustração desse caminho:

²Em Datawarehouse os dados também são estruturados, mas de uma maneira específica, que permitem aplicar técnicas conhecidas como Business Intelligence

³Dados categóricos são dados textuais que identificam uma categoria de dados

Figura 2.9: Fases da mineração de dados até extração do conhecimento



Na origem dos dados, os “inputs” estão representados na figura onde se lê “Dados”. Estes geralmente estão repletos de *missing data* e/ou dados inconsistentes, conhecidos como dados não estruturados. O balão onde se lê “Seleção” representa a coleta das informações ou a seleção dos dados em grandes repositórios de dados como Datawarehouse ou *Big Data*.⁴

Em nossa pesquisa esses dados são provenientes das mais diversas fontes, tais como, redes sociais, câmaras de trânsito, informações de satélites meteorológicos e outras fontes.

Armazenar dados provenientes de redes sociais nessa etapa pode ser um grande problema, devido à sua extensão, porém os dados relevantes podem ser armazenados em “Dados Escolhidos” com tecnologia apropriada, utilizando-se técnicas de “Map” e “Reduce” ou mineração de dados em textos *Text Mining* para criar *cluster* de informações e ler os fluxos de dados (stream data). Algumas técnicas de IA podem ser aplicadas nessa etapa como, “Data Mininng Swarm Robotics” através de Botnets⁵ ou “Swarm Intelligence”.

No balão “Pré-Processamento” os dados não-estruturados são tratados, por exemplo, retirando os *missing data*. Para estruturar as informações é preciso utilizar técnicas linguísticas, uma vez que existe lógica entre eles (22). Esses dados normalmente são coletados por técnicas de Mineração de Textos, também conhecidas como Mineração de Dados em Textos, técnicas de IA como “Machine Learning” têm sido muito utilizadas. Em “Transformação” os dados foram estruturados, podendo ser armazenados em Bancos de Dados, conhecidos como Datawarehouse, por exemplo o Hive (23).⁶.

O processo de Mineração dos dados começa no balão “Mineração de Dados”, onde são aplicadas as técnicas de IA conhecidas como classificadores, para extração de padrões, tais como: “Decision Tree” (Árvore de decisão), “Artificial Neural Network” (Redes neurais artificiais), “Logistic Regression” (Regressão Logística), “Naïve Bayes” e “Deep Learning”, dentre outros. Algumas técnicas de mineração de dados são fortemente influenciadas pelas informações na entrada (input), como as Árvores de decisão (24). As Redes Neurais, dependendo da quantidade de variáveis de entrada, puderão ter milhares de neurônios na camada intermediária, o que inviabilizaria essa metaheurística⁷.

⁴A definição de Big data se encontra na página 16

⁵Botnet é citado no sentido da coleta de informações

⁶Hive é um de Datawarehouse que armazena dados não estruturados: *Not Only SQL ou também Not structured SQL* NoSQL. Ferramentas NoSQL servem para trabalhar com o *Big Data*, outros exemplos: HBase, MongoDB e Pig

⁷Metaheurística são heurísticas aplicadas a problemas em que os custos computacionais não são tratá-

Todas essas etapas descritas na figura 2.9 são recorrentes, como indicam as setas pontilhadas que retornam aos passos anteriores. Utilizar técnicas de mineração de dados, extrair dados e conhecimentos, com isso pode-se predizer os resultados futuros na saída do modelo, quando determinados dados ocorrem na entrada (25); essa técnica é o *Knowledge Discovery Databases*. O KDD utiliza métodos de Aprendizagem de Máquina para efetuar essa extração.

2.4 Mineração em Dados Textuais

A mineração em dados textuais ou Mineração em Textos, tal como acontece com a mineração de dados vai buscar os dados em arquivos tipo textos digitalizados, contudo esses arquivos são transformados em documentos antes de serem analisados pelos algoritmos de IA.

De acordo com Hotho, Nürnberg e Paass (26) a expressão *Text Mining* ou descoberta de conhecimento em textos foi referendada em 1995. Entretanto o interesse por extrair conhecimento oriundo de textos remonta à década de 60' (27). Hotho, Nürnberg e Paass (26) discutem, ainda, que é frequente a confusão de termos. Esses mesmos autores (26) na tentativa de definir *Text Mining* afirmam que é preciso considerar a perspectiva específica da área, definindo três possibilidades:

1. A perspectiva sugere que *Text Mining* corresponde à extração de fatos do texto, ou seja, extração de informação;
2. A abordagem assume que mineração em texto se configura como a aplicação de algoritmos e métodos do campo de *Machine Learning*, cujo objetivo seria encontrar padrões usuais;
3. A perspectiva prevê que a mineração em textos segue o modelo de processo de descoberta de conhecimento. É frequente na literatura, a cerca da mineração em texto, entendê-la como uma série de passos para extração de informação bem como o uso de mineração de dados ou processos estatísticos.

Embora as pesquisas no campo da mineração em textos sejam relativamente recentes, os estudos envolvendo Processamento de Linguagem Natural (Natural Language Processing – NLP) datam da década de 40' (28).

As primeiras aplicações computacionais relacionadas à linguagem natural aparecem em torno de 1946 (28), decorrente da expertise de Alan Turing para quebra de códigos inimigos durante a segunda guerra mundial.

Ainda segundo Liddy (28) outro importante avanço é identificado no final da década de 50' quando Noam Chomsky introduziu a ideia de Gramática Gerativa. Neste mesmo período começaram a surgir pesquisas no campo do reconhecimento da fala.

Desde então, essa área de conhecimento tem experimentado grandes avanços, particularmente, nos tempos atuais, com a Mineração em Textos nas Redes Sociais.

A Mineração em textos é inspirada em técnicas de *Machine Learning* (22). Todavia, analisar textos é basicamente entender o seu significado, baseado em regras de associação lógica. O mapa mental na Figura 2.10 mostra um modelo de análise de texto feito por seres humanos.

veis em tempo polinomial, devido às explosões combinatórias geradas pelo grande número de tentativas. Metaheurísticas bioinspiradas metaforizam o comportamento de animais sociais, tais como formigas, pássaros, peixes e outros

Figura 2.10: Mapa mental da Mineração em textos



Minerar dados em texto é um processo que deve ser dividido em várias etapas (29). Minerar em redes sociais exige ao menos uma etapa: a escolha de uma rede social, cada uma delas com sua tecnologia própria.

2.4.1 Mineração em Textos nas Redes sociais

As redes sociais têm assumido, nos dias atuais, um papel essencial na vida de seus usuários. Não apenas como espaço de descontração, mas, sobretudo, como lugar de troca de informações que permitem, dentre outras coisas, tomar conhecimento acerca dos acontecimentos, sejam eles locais ou globais, que influenciarão sua vida. De modo particular, nos grandes centros urbanos, as redes sociais têm servido de fonte de conhecimento acerca de segurança pública, mobilidade urbana e acontecimentos de toda sorte, que possam fazer com que, por exemplo, uma pessoa resolva seguir um ou outro caminho para chegar a um determinado lugar, quer seja ele próximo ou distante de onde se encontre (30) (31) (32).

Além da troca de informações momentâneas, as redes sociais permitem uma atualização praticamente em tempo real, a partir da utilização de seus usuários e de instituições que também dela fazem uso (por exemplo, a Polícia Rodoviária Federal), de modo que possibilita que decisões sejam tomadas e reorientadas, em virtude da alimentação das informações nas redes.

No que diz respeito às escolhas relacionadas ao trânsito, sejam relativas as áreas urbanas, bem como a centenas de quilômetros adiante, pelo interior de um estado, cada vez mais as pessoas não tomam decisão sem antes consultar aplicativos e redes sociais tais como o waze, twitter, facebook, etc.

Aplicativos para celulares até dispõem de GPS, Google Maps e outras fontes que lhe orientem sobre melhores rotas, que levem com maior rapidez e segurança ao seu destino.

Se pensarmos no transporte de cargas, a principal função das redes sociais não é de

caráter lúdico, mas, sim, como uma ferramenta essencial para que não haja qualquer contratempo que possa causar prejuízo à empresa ou empresas envolvidas. Afinal de contas, no que tange ao transporte de mercadorias, sempre há pelo menos duas empresas relacionadas: a de produção do bem e a de transporte do mesmo ao seu destino.

O que discutimos até agora é amplamente sabido por aqueles que analisam o uso das redes sociais na atualidade. O que pretendemos, então, é trazer uma contribuição de natureza científica a essa compreensão e à utilização de forma cada vez mais eficaz dessas ferramentas, a partir do uso da IA, da mineração de dados e dos métodos de extração e produção de conhecimentos (KDD).

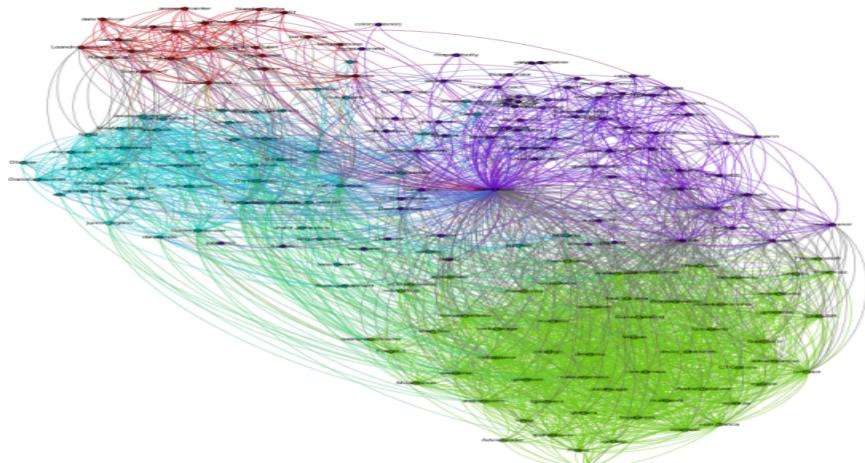
O crescimento dos dados eletrônicos, nos últimos anos, até 2020, pode ser acompanhado na tabela 2.2. Esta demonstra o volume de dados e informações produzidas no mundo entre 2000 à 2020. "Em 2010 empresas e usuários armazenaram mais de 13 exabytes de novos dados"(5).

Tabela 2.2: Volume de dados no mundo

Ano	Qtd	Unidade	Múltiplo
2000	800	terabytes – TB	10^{12}
2006	160	petabytes – PB	10^{15}
2009	500	exabytes – EB	10^{18}
2012	2,7	zettabytes – ZB	10^{21}
2020	35	yottabytes – YB	10^{24}

Uma parte desses dados são as redes sociais. A rede social escolhida para esta pesquisa foi o Twitter por apresentar características como API aberta de fácil conexão para obtenção dos dados. Uma rede social é sobretudo uma de conexão entre pessoas, contudo, sob o ponto de vista tecnológico uma conexão entre nós e arestas, os nós simbolizam as pessoas e as arestas a ligação entre elas, essa “arquitetura social” é conhecido como grafo. A figura a 2.11 foi gerada pelo software Gephi (33) representa um grafo da rede social Twitter.

Figura 2.11: Grafo de uma rede do Twitter



Fonte: <https://github.com/gephi/gephi/wiki/Official-papers>. Acessado em: 10/03/2017

O Twitter

O Twitter caracteriza-se como um microblog onde os usuários escrevem em um espaço delimitado (cerca de 140 caracteres) sobre os mais diversos assuntos. Tais usuários conectam o aplicativo por meio de uma multiplicidade de dispositivos: computadores, tablets e celulares, formando uma grande rede social mundial. Essa rede possui duas diferentes APIs, responsáveis pela captura dos dados: Rest API e Streaming API. O Twitter funciona com o padrão de arquivos JSON e os dados são capturados nesse formato (34). A cada dia centenas de terabytes são inseridos nos seus bancos de dados, conhecidos como Hadoop data warehouse, tornado impossível capturar todas as informações produzidas em um largo espaço de tempo, portanto ou se analisa o Streaming de dados ou a API.

A ideia inicial do Twitter, segundo seus fundadores, era de que essa rede se comportasse como um “SMS da Internet” (30). As informações são enviadas aos usuários, conhecidas como twittes, em tempo real e também enviadas aos usuários seguidores que tenham assinatura para recebê-las: os seguidores. A conexão entre os usuários da rede social se deve à relação entre os seguidores e os seguidos. O comportamento do seguidor para retweetar os usuários seguidos serve como principal mecanismo para compartilhar informações nessas redes.

Nas pesquisas que envolvem as redes sociais, analisar o conteúdo utilizando ferramentas de mineração de textos é um procedimento frequente e tem apontado resultados surpreendentes sobre o comportamento social e suporte à tomada de decisão (35). É comum, ainda, aplicar mineração de textos em bibliotecas e outras instituições. Isso implica em rastrear tópicos, extrair informações, agrupar, categorizar (36).

Em recente artigo, Sandhu (37) indicou a importância do aprendizado sobre mineração de dados e ferramentas de Big Data para as bibliotecas acadêmicas, de forma a melhorar a eficiência da biblioteca e dos serviços de informação. Similarmente, Zhang e Gu (38) alegaram que minerar conhecimento sobre os clientes é importante para as bibliotecas acadêmicas. Na mesma linha de investigação, Sarker et al. (30) destacam que a abordagem da mineração de textos para os dados das mídias sociais tem sido utilizada em muitos campos, como negócios, ciência da saúde, dentre outros.

Estudos atuais têm mostrado o papel da análise sentimental e mineração de opinião nas redes sociais, em particular no Twitter, como forma de investigar padrões de comportamento (39) (40).

Outro estudo aponta que em 2013 um número superior a 70% dos indivíduos adultos que faziam uso da internet estavam conectados a redes sociais. Cerca de 20% utilizavam o twitter, sendo que aproximadamente 46% conectando-o diariamente e algo em torno de 29% mais de uma vez ao dia (31).

Um relatório publicado em 2013 revelou que o Twitter estava posicionado entre as três maiores mídias sociais, em termos de adesão e utilização, perdendo apenas para o facebook e o youtube. Esse relatório revelou, ainda, que naquele ano, nos EUA 8% dos adultos que tinham entre 18 e 29 anos de idade utilizavam o twitter como principal mídia social. Em outras idades, esse percentual subia para 45% no mesmo país. As notícias são o principal interesse dos usuários dessa rede (41).

Em 2014, os dados revelavam que, em um dia típico, sem qualquer evento extraordinário, essa rede era conectada por cerca de 230 milhões de usuários, responsáveis pela produção de aproximadamente 500 milhões de “tweets” (postagem tipo microblog) (Twitter, 2014).

Naaman, Boase e Lai (32), classificaram os conteúdos propagados nos tweets em oito categorias: Informações compartilhadas, auto promoção, opinião/queixas, declarações e

pensamentos aleatórios, “eu agora” (me now), perguntas aos seguidores, manutenção de presença e piadas.

Na Web, por sua vez, a quantidade de citações (i.e. links de URL) e estrutura dos links são utilizadas pelos motores de busca (ex: o Google) com o objetivo de identificar a relevância e aceitação de *websites* (42), em decorrência do número de seguidores, da quantidade de menções feitas, de “retweets”, por exemplo.

Cha, Haddadi e Benevenuto (43) avaliaram a influência dos usuários no Twitter na rede, como um todo, analisando o número de “retweets”, menções e seguidores. Esses pesquisadores identificaram uma correlação positiva entre o número de seguidores e o número de “retweets” pelo top 10 (do Twitter) e o primeiro percentil dos mais conectados, com base no grau do link (i.e. número de seguidores).

Modalidades e ferramentas de análise do Twitter

Nesse tópico abordaremos, em maior detalhe, as escolhas metodológicas e ferramentas analíticas utilizadas em estudos que levam em conta dados do twitter, apresentando alguns dessas pesquisas.

Em artigo conduzido por Chu & Du (44), os autores justificaram que as mídias sociais têm sido utilizadas cada vez mais para promoção das bibliotecas, com o objetivo de incrementar a relação com os clientes, permitindo “facilitar informações e compartilhar conhecimentos, incrementar serviços e promoções, interação com estudantes usuários das bibliotecas, a um custo mínimo” (p. 72) (livre tradução do autor)⁸. Tal prática têm promovido considerável mudança na interação com usuários e relacionamento com os clientes (45), tendo sido utilizada frequentemente como alternativa para estabelecer uma conexão personalizada com os seus usuários (46).

A pesquisa em questão interessou-se em investigar quantas vezes a biblioteca acadêmica usa o Twitter; tipo de conteúdo compartilhado pela biblioteca acadêmica no twitter; temas associados com os “tweets” da biblioteca acadêmica (47). Nesta pesquisa foi obtido a partir da “timeline” de dez bibliotecas acadêmicas (i.e. todos “tweets” desde a adesão à plataforma), através de um serviço de arquivamento (twimemachine.com), em dezembro de 2014. Foram selecionadas as 10 maiores bibliotecas ranqueadas pelo Shanghai Ranking, tendo a seleção se restringido às universidades de língua inglesa e a apenas uma biblioteca por instituição, para o caso de a universidade ter mais de uma.

Na etapa de preprocessamento – “dataset preprocessing” – o grupo de dados recuperado foi tratado, para reduzir os “ruídos”, seguindo uma abordagem consistente com outros estudos de mineração em textos, tal como o de Ralston, O’Neil, Wigmore e Harrison (48) e de Yoon, Elhadad e Bakken (49). O processo contemplou a aplicação de certo número de filtros. Por exemplo, foram removidas as “stopwords”, pontuação e numeração, todos os nomes de usuários seguido por um símbolo “”, “hashtags” após o símbolo “#” e “hyperlinks” após o “http”. Também foi removido a abreviação do Twitter tal como “RT” (retweetes), e “MT” (tweet modificado) e palavras tal como “via”. O nome do usuário do Twitter para cada biblioteca acadêmica também foi excluído.

Para análise do conjunto de dados, utilizou-se a mineração em textos e para investigar os históricos de “tweets” das bibliotecas escolheu-se a análise de conteúdo. A frequência dos “tweets”, “retweets” e sua distribuição foi identificada e contabilizada. Em seguida, os

⁸Facilitate information and knowledge sharing, service enhancement and promotion, interaction with student library users, at minimal costs.

“tweets” marcados com o PamTaT ⁹.

Quando se tem um *corpus linguístico* ou, em outras palavras, um documento que está sendo analisado, baseado nas informações linguísticas que ele contém, uma ferramenta estatística frequentemente utilizada em mineração de textos é a TF-IDF. Esta sigla é a abreviação, em língua inglesa, de *Term Frequency*, isto é, *frequência do termo*, e *Inverse Document Frequency* (50), que significa frequência inversa do documento. Em termos gerais, o TF considera que o peso de uma palavra em um documento está relacionado à frequência com que essa palavra aparece no mesmo. Em nossa pesquisa, a palavra “acidente”, quando temos o trígrama *ocorrência de acidente*, as palavras *ocorrência* e *acidente* têm um importante peso relacionado a frequência, enquanto que o *de*, sem estar acompanhado dos outros dois, não tem significado relevante para os dados minerados, pois o mesmo deverá aparecer inúmeras vezes, sem estar relacionado ao que é investigado. Nesse caso, o IDF aponta exatamente para a *frequência inversa*, ou seja a palavra *acidente* é provável que apareça no texto em menor frequência que *ocorrência*, portanto a ordem inversa deverá ser: “acidente” < “ocorrência” < “de”, como pretendíamos. A equação 2.1 demonstra como se encontra o *idf*.

$$idf(term) = \ln\left(\frac{n_{documents}}{n_{documents \ containing \ term}}\right) \quad (2.1)$$

A técnica serve, para determinar a frequência de palavras simples (unigrams), de duas palavras (bigrams) e sequência de três palavras (trigrams) que aparecem no texto fonte. Com isso, permite desenvolver uma matriz de frequência de termos-tweet, mostrando como sequências de palavras simples e múltiplas palavras (n-grams) são usadas pela biblioteca acadêmica selecionada.

Figura 2.12: Descrição da conta Twitter das bibliotecas acadêmicas

Library name	Tweets	Following	Followers	Favorites	Twitter account(s) created
Harvard University	2177	1006	15,500	144	Jul-09
Stanford University	3771	1130	1787	304	Apr-12
MIT	2284	585	10,700	135	Mar-09
Cambridge University	3114	148	9164	4	Mar-09
Columbia University	1315	148	2507	45	Aug-09
University of Oxford	2830	156	25,900	38	Apr-10
Yale University	1991	116	4690	50	Apr-09
University of California San Diego	1874	756	988	440	May-08
University of Washington	2009	231	2697	143	Apr-09
Johns Hopkins University	5900	551	449	1007	Aug-09
Total	27265	4827	74,432	2310	

Fonte: AL-DAIHANI, S. M. and ABRAHAMS, A. – 2016

Observa-se que foi incluído o número de “tweets”; a quantidade de usuários seguidos pela biblioteca; a quantidade de seguidores e o número de “tweets” favoritos da biblioteca. A Universidade Johns Hopkins, como se pode observar na tabela, tem o maior número de “tweets”, seguida pela biblioteca da Stanford University e pela biblioteca da Cambridge

⁹Ferramenta “text mining” desenvolvida por Pamplin Collage do Instituto Politécnico de Negócios de Virgínia da Universidade Estadual de Virgínia (Bird, Loper & Klien, 2009)

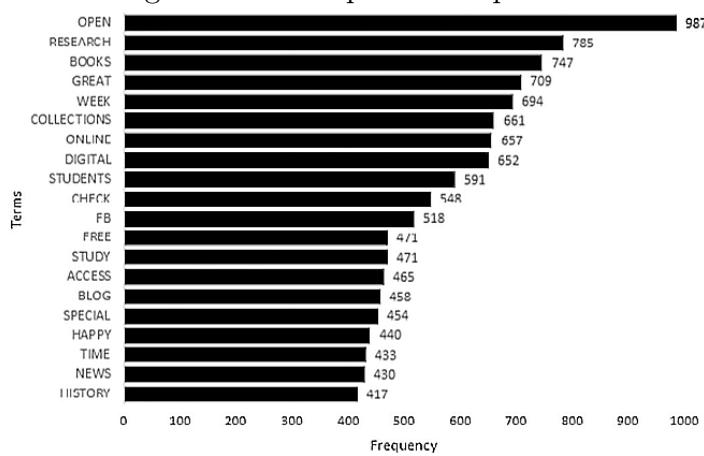
Figura 2.13: Bibliotecas de Universidades e contas no Twitter

Nome Biblioteca	Tweets	Menções	Hashtags	Retweets
Cambridge University	3076	1403	440	475
Columbia University	1305	923	632	232
Harvard University	2154	950	1674	99
Johns Hopkins University	3190	3900	373	822
MIT	2190	1283	701	414
University of Oxford	2795	2544	1109	901
Stanford University	3177	3994	1042	1960
University of Washington	2001	1057	377	452
University of California San Diego	1857	1144	1219	491
Yale University	1962	650	58	128
Total	23,707	17,848	7625	5974

University, respectivamente. A biblioteca da Universidade da Califórnia San Diego tem a conta do Twitter mais antiga (maio de 2008), mas apresenta um pequeno número de “tweets”, comparativamente à biblioteca da Stanford University, que começou no Twitter com uma conta em abril de 2012, mas possui o segundo maior número de “tweets”.

A análise do conteúdo dos Tweets foi desenvolvida da seguinte maneira: tomando a frequência de unigramas (palavras únicas), observou-se que a palavra mais frequente foi “open”, utilizada em uma variedade de contextos pelos “tweets” da biblioteca. Por exemplo: foi usada em um anúncio sobre a mudança do horário de funcionamento, bem como em um anúncio para abertura do espaço para os estudantes, exposições, abertura da casa (biblioteca), etc.

Figura 2.14: Frequência de palavras



Fonte: AL-DAIHANI, S. M. and ABRAHAMS, A. – 2016

O segundo termo mais frequente foi “research”, que foi utilizado também em diferentes contextos, relacionados frequentemente a apoio, a investigação, por exemplo: workshop research, ferramentas de pesquisa e software, abertura ao acesso para pesquisar, dados e laboratório de pesquisas, guia de pesquisa e ajuda e campos de pesquisa. Outros termos tais como “livros”, “coleções” (acervos) e “on-line” foram utilizados no contexto dos “tweets” sobre os recursos da biblioteca. Tais termos foram incluídos em “tweets” relacionados a “e-books”, “textbooks”, “livros raros”, “solicitando e renovando livros”, “comentários de livros”, “novos livros”, “livros de coleções especiais”, “livros recomendados”, “política de circulação de livros”.

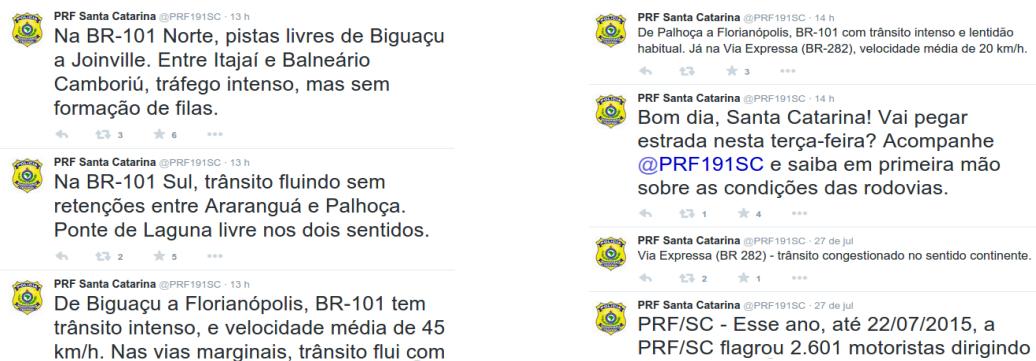
Na distribuição dos bigramas (sequência de duas palavras) no conjunto de “tweets”,

observa-se que o mais frequente foi “special collections”. O segundo mais popular bigrama foi “open access”, utilizado em diferentes contextos, tal como política de acesso, publicidade, recursos, treinamentos e workshops, dicas e orientação, serviços, eventos e notícias. O resultado mostrou a ênfase colocada na iniciativa de promover “open access” com as instituições acadêmicas.

O terceiro maior bigrama foi “reading room”, relativo à atividade de suporte aos estudantes com as instituições acadêmicas. Essas salas são um dos mais importantes espaços da biblioteca, usadas para leitura e estudo. Os “tweets” eram, em sua maioria, relacionados a notícias de abertura e fechamento das salas de leitura: “reading rooms”.

Dentre os mais importantes trigramas (sequência de três palavras), destacou-se o trígrama “save the date”. Essa expressão era utilizada para requerer especial atenção dos seguidores para os eventos importantes que estavam para acontecer. Este trígrama é seguido, como o segundo mais frequente, por “pleased to announce”, outra expressão usada para enfatizar a importância de eventos especiais. O terceiro mais usado foi “open access week” (seguido muito próximo por “open access policy”), que novamente destaca os esforços na iniciativa de espaço aberto (open access).

No caso particular da pesquisa em lide, o interesse voltou-se para os “tweets” da Polícia Rodoviária Federal do estado de Pernambuco, bem como de seus seguidores e de usuários que fizeram menção a eventos que afetam diretamente o tráfego nas principais rodovias do estado. A seguir, a título de exemplo, pode-se verificar uma sequência de “tweets” da Polícia Rodoviária Federal do estado de Santa Catarina:



Fonte: <https://twitter.com/PRF191SC> Acessado em: 10/10/2016

A Polícia Rodoviária Federal disponibilizou às 13h através do canal @PRF191SC, informações relevantes sobre o trânsito naquela localidade, num espaço temporal variado. Por exemplo: entre Itajaí e Balneário Camboriú foi informado que o trânsito estava intenso, sugerindo que a frota de caminhões deva ter uma rota alternativa, caso a situação persista por muito tempo. No primeiro twitte da segunda coluna, é informado em Via Expressa (BR 282) que o trânsito está lento com velocidade de 20km/h (praticamente congestionado).

Outra rede social conhecida pelos condutores de veículos é o Waze. O Waze é um aplicativo de navegação para o trânsito, e funciona em aparelhos celulares e tablets. Os utilizadores desse aplicativo são conhecidos como wazers e compartilham informações sobre o trânsito, em tempo real. Todavia, as informações somente estão disponíveis no momento em que são postadas pelos utilizadores, por um período de tempo pequeno. Caso não haja usuários trafegando pelas vias, ou caso os mesmos não tenham disponibilidade para compartilhar informações, não há o que se ver. Outro problema identificado em relação ao waze é que caso não haja conexão à Internet, não há como acessar os dados dos ‘wazers’, para navegação.

Além dos dados que chegam ao *Big Data* através das redes sociais, as grandes cidades têm disponíveis câmeras de monitoramento do trânsito nos semáforos ou próximas a eles. Algumas com cobertura por canais de televisão, bem como câmaras de segurança próximas às rodovias, coletando informações em tempo real. Os dados desses dispositivos são gravados, sendo conhecidos como *stream* de dados. Esses *streams* podem ser disponibilizados na Internet, em sítios eletrônicos especialmente construídos para isso, como o <http://vejoao vivo.com.br> (acessado em 10/10/2016) dentre outros.

Os dados disponibilizados pelos diversos meios de comunicação não estão em formato que possam ser utilizados imediatamente, precisando antes serem processados. Tais dados não processados são conhecidos como “dados frios”. O processo de tratar as informações, retirando-lhes o “lixo” e transformando dados “frios” em dados “quentes”, é um processo que tem um custo temporal elevado, devido ao volume dos dados.

2.5 Aprendizagem de Máquina

Historicamente, a aprendizagem computacional está relacionada com “o que” há para ser aprendido (51). Para escolher o que aprender é necessário definir de “onde” ou sobre quais dados aprender. Deve ser fornecido um conjunto de treinamento, para, em seguida, testar o conhecimento aprendido em um “conjunto de teste”.

Aprendizagem de Máquina ou “Machine Learning” são métodos para analisar dados de forma automatizada e interativa. Segundo Shalev-Shwartz & Ben-David (9), o termo Aprendizagem de Máquina refere-se à detecção automatizada de Padrões de dados.

Para Nilsson (51), o aprendizado ocorre quando uma máquina modifica sua estrutura interna, programa ou dados (baseados nos inputs ou em uma resposta para informação externa) de tal maneira que melhora o desempenho futuro. Por exemplo, quando uma máquina de reconhecimento da fala melhora após “ouvir” várias amostras de fala humanas e que nós percebemos que está pronta, neste caso podemos dizer que a máquina aprendeu.

Sistemas que executam tarefas de inteligência artificial, tais como Reconhecimento de Padrões, Diagnóstico, Controle de Robôs, Predição e outros, precisam ser modificados para executarem “Machine Learning” (51).

2.5.1 Tipos de Aprendizagem

Quando se fala em algoritmos de IA, adentramos no campo de Aprendizagens e Máquinas. É o princípio da aprendizagem que faz com que o algoritmo estabeleça a decisão adequada para o problema proposto. No campo da aprendizagem de máquina, é possível apontar três tipos de aprendizagem: a aprendizagem supervisionada, aprendizagem não-supervisionada e aprendizagem por reforço (18). As duas primeiras serão aqui descritas de maneira sucinta, e consideradas mais adiante, uma vez que interessam particularmente a essa pesquisa, sobretudo quando da utilização de redes neurais e árvores de decisão.

A aprendizagem supervisionada (52) se caracteriza pelo acesso ao conjunto de exemplos de treinamento pelo algoritmo de aprendizagem, também conhecido como induutor, de modo que haja especificação da saída desejada. No caso da aprendizagem não-supervisionada (18), os valores de entrada são estabelecidos, mas não são definidos os valores de saída. O induutor terá o papel de estabelecer aproximações, propondo agrupamentos (clusters) em função de determinadas categorias como, por exemplo, similaridade (52).

Para “aprender” sobre uma determinada função f definimos uma amostra em um conjunto de treinamento $X = x_1, x_2, \dots, x_n$.

As técnicas algorítmicas apresentadas nas seções subsequentes são parte da grande família de algoritmos que compõem o aprendizado de máquina aplicado a mineração de dados.

A descoberta de conhecimento através da aplicação das técnicas de mineração de dados podem ser agrupadas de acordo com suas funcionalidades (21), essas funcionalidades têm como característica principal a maneira como são descobertos os padrões no dados, elas podem estar em uma das duas categorias: tarefas descritivas ou tarefas preditivas. As tarefas mineração descritivas preocupam-se nas características dos dados no conjunto de dados; o “data set”. As tarefas de mineração preditivas induzem regras nos dados correntes para produzirem previsões (21). A seção seguinte analisa as tarefas preditivas.

2.5.2 Algoritmos de Aprendizagem de Máquina

Os algoritmos de Aprendizagem de Máquina selecionados para esta pesquisa contemplaram robustez, devido a grande quantidade de dados (cerca de 55 000 registros) a qualidade no tratamentos desses dados como e tolerância a falhas.

Naïve Bayes

Dentre os algoritmos de machine learning que destamos nessa dissertação está o Naïve Bayes. Essa classe de algoritmos é baseado no teorema da probabilidade condicional de Bayes (53), que serve para rotular classes de variáveis independentes. O classificador Naïve Bayes é um modelo probabilístico relativamente simples, todavia, muito potente. Produz estimativas de probabilidade, em vez previsões.

Um classificador Naïve Bayes produz probabilidade como output (54). A probabilidade que é produzida pode ser utilizada para distinguir a que grupo classificado, uma amostra não classificada pertence, por exemplo, dependendo da mais alta probabilidade obtida entre um grupo de probabilidades.

Resumidamente, um classificador Naïve Bayes sugere que a presença ou ausência de um atributo particular de uma determinada classe não está relacionado à presença ou ausência de qualquer outro atributo. Ainda que esses atributos dependam entre si, ou da existência de outro atributo, esse tipo de classificador considera todas as propriedades como contributos independentes da probabilidade final (54).

Em mineração de dados variáveis independentes explicam a variável dependente para fazer previsão. Este classificador tem sido muito empregado para classificar documentos e detectar spam em mensagens. A probabilidade condicional pode ser explicada por um vetor $x = (x_1, x_2, \dots, x_n)$ que se representa n características (variáveis independentes) que se atribui a estas instâncias de probabilidades $p(C_k|x_1, \dots, x_n)$ para cada K possível ter vindo da classe C_k . Aplicando o teorema de Bayes da probabilidade condicional temos:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} \quad (2.2)$$

Em outras palavras, a medida que se conhece os resultados das probabilidades pode-se predizer os novos resultados porque o conjunto de testes torna-se menor. A probabilidade condicional também pode ser entendida como:

$$p(posteriori) = \frac{p(priori) * verossimilhana}{evidencia} \quad (2.3)$$

Aprendizagem Bayesiana

Baseado no teorema de Bayes, dado um conjunto de variáveis aleatórias $\omega = x_1, x_2, \dots, x_n$ a variável aleatória H (hipótese) denota o tipo de ω , com valores possíveis para h_1, h_2, \dots, h_n . A medida que são inspecionadas as variáveis, são revelados os dados D_1, D_2, \dots, D_n , onde D_i é uma variável aleatória com valores possíveis para cada variável do conjunto ω de variáveis. Sendo D a representação dos dados do espaço de variáveis para uma predição sobre a parte desconhecida de X , temos:

$$P(h_i|d) = \sum_i P(X|d, h_i)P(h_i|d) = \sum_i P(X|h_i)P(h_i|d) \quad (2.4)$$

onde cada hipótese h_i determina uma distribuição de probabilidades sobre a variável X (18).

Um aspecto relevante e positivo que deve ser mencionado acerca do classificador Naïve Bayes é que este requer apenas um pequeno grupo de dados de treinamento para estimar os parâmetros que são necessários para que seja feita a classificação. Outro aspecto que merece destaque é que ele se mostra eficiente para aprendizagem supervisionada, trabalhando de forma rápida com dados complexos (54).

Árvores de Decisão

No âmbito da inteligência artificial, quando se trata de algoritmos de aprendizagem, uma classe de algoritmos que tem se revelado potente para problemas das mais diversas naturezas é a árvore de decisão. Além do universo das pesquisas no campo da informática engenharia e ciências da computação, as árvores de decisão têm sido utilizadas, sobretudo, em pesquisas relacionadas à Medicina, à Economia, nos mais diversos sistemas de suporte à decisão, como diagnósticos de doenças, investigação de fraudes, dentre outros (55).

A escolha desse algoritmo está relacionada, em larga medida, a uma relação que cotidianamente chamamos de “custo benefício”. Uma árvore de decisão é gerada de maneira relativamente simples e os resultados produzidos são, em sua maioria - e a depender da área específica - de grande poder de abrangência e de fácil interpretação. Todavia, para que se faça opção por essa ferramenta, o pesquisador precisa ter clareza sobre a que classe de problemas ela atende, bem como, de que maneira pode ser gerada e realimentada, e de que forma seus resultados devem ser adequadamente interpretados.

Na pesquisa apresentada nessa dissertação, particularmente, as árvores de decisão possibilitaram grandes avanços na proposição do modelo de predição, conforme pode ser observado no capítulo dedicado aos resultados.

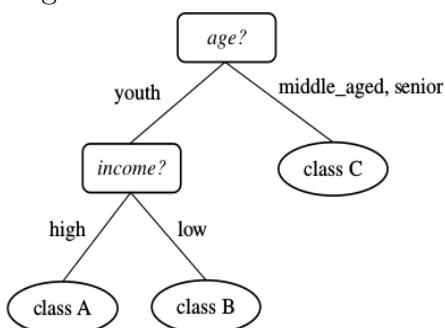
Conforme o nome sugere, o que se espera de uma árvore de decisão é que ao final do processo resulte a melhor decisão. Todavia, para que a decisão satisfatória apareça, é necessário que o pesquisador faça as escolhas adequadas, de modo a poder tirar o melhor proveito dessa ferramenta. Para tal, é preciso compreender a natureza desse algoritmo e os processos a ele relacionados. A seguir serão apresentados os principais elementos que precisam ser conhecidos pelo pesquisador que deseja se aventurar por esse campo.

Breve histórico e conceituação

De maneira sintética, uma árvore de decisão tem como entrada um conjunto de atributos e como saída uma decisão. Os atributos podem, ainda, ser classificados de duas maneiras: discretos ou contínuos. Em virtude dos atributos de entrada, tem-se, como resultado uma função de valores discretos – aprendizagem de classificação – ou de valores contínuos – aprendizagem de regressão (18).

A decisão gerada aparece em função de uma sequência de testes executados, estando cada um deles relacionados a um nó da árvore. As ramificações que decorrem dos testes são o resultado encontrado a partir da realização do experimento. O exemplo a seguir ilustra uma árvore de decisão simples, onde se vê seu nó-raiz e suas ramificações.

Figura 2.15: Árvore de decisão



Fonte: Han, J. and Kamber, M.

Uma árvore de decisão obedece à regra básica “se-então”, de maneira que, parte-se do nó (se) até as folhas (então). O conhecimento é representado em cada nó, que apresenta pelo menos duas saídas ou ramificações possíveis, que pode, ou não, converter-se em um novo nó, relacionado a um novo nível.

Embora seja um algoritmo simples e de fácil interpretação, uma das mais importantes questões a ser considerada é como propor as regras de forma adequada e relevante para a geração da árvore. É necessário identificar o melhor atributo, que será responsável por criar o nó de decisão. As ligações entre os nós representam os valores possíveis do teste do nó superior e as folhas indicam à classe (categoria) a qual o registro pertence (55).

A origem das árvores de decisão, como algoritmo no campo da inteligência artificial data da segunda metade do século XX. A literatura aponta que as árvores de decisão foram propostas por Ross Quinlan, pesquisador australiano, no final da década de 70 e início dos anos 80, sendo o ano de 1983 aquele em que foi apresentado o primeiro algoritmo para geração de árvores de decisão: o ID3 (Iterative Dichotomiser), utilizado até hoje e considerado um dos mais importantes.

Todavia, HSSINA; MERBOUHA; MOHAMMED (56), sugerem que autores no campo da estatística descrevem que seu surgimento na deu década de 60, com Sonquist e Morgan, que utilizaram árvores de decisão para predição e explicação, com o algoritmo AID (Automatic Interation Detection), sem tomar conhecimento das pesquisas de Quinlan. A partir desse modelo, houve uma expansão para problemas de classificação e discriminação, cuja abordagem teria culminado no CART (Classification and Regression Tree), método desenvolvido por Breiman e seus colaboradores.

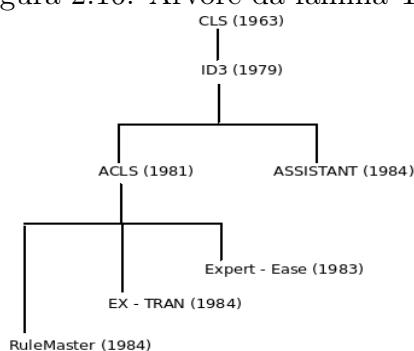
Quinlan (57) discute que desde que a inteligência artificial começou a se desenvolver como campo de teorização e investigação, em meados dos anos 50, as máquinas de apren-

dizagem (machine learning) ocuparam um lugar de particular interesse dos pesquisadores, sobretudo pela na compreensão e modelização de comportamentos inteligentes.

Tal interesse, ainda segundo esse autor, instigou a busca pelo desenvolvimento de sistemas competentes baseados no conhecimento (knowledge-based) e tomou vulto nas pesquisas em inteligência artificial. Quinlan (57) avança, pontuando o interesse de muitos pesquisadores, dos quais ele próprio, no que ele chamou de “microcosmo de máquinas de aprendizagem e de uma família de sistemas de aprendizagem que têm sido utilizadas para construir sistemas baseados em conhecimentos de um tipo simples” (p.82) (livre tradução do autor). ¹⁰

O primeiro grupo de árvores de decisão, ainda conforme Quinlan (57) era responsável por tarefas de classificação, desenvolvendo uma decisão das raízes às folhas, sendo conhecido com *Top – Down*. ¹¹ O modelo proposto comportaria – como o é até hoje – inúmeras análises e reanálises, em todos os estágios e durante todo o processo. Assim, esse primeiro modelo faria parte da família de algoritmos do tipo Top-Down Induction of Decision Tree – TDIDT, representada por Quinlan, em forma de árvore, na figura a seguir (p. 84).

Figura 2.16: Árvore da família TDIDT



Fonte: J. R. Quinlan.

Para Quinlan (57), o pai da família TDIDT é o Sistema de Aprendizagem de Conceito (Concept Learning System – CLS) de Hunt, proposto em 1963, que constrói uma árvore de decisão que visa diminuir o custo de classificar um objeto. Para cada etapa o CLS explora o conjunto de decisões possíveis, seleciona uma ação que minimize o custo, para então mover a um nível abaixo na árvore. O ID3 configura-se, então, como um entre uma série de programas desenvolvidos em função desse Sistema (CLS).

O ACLS (57), por sua vez, seria uma generalização do ID3. Enquanto que o CLS e o ID3 têm propriedades que possibilitam descrever objetos apenas com valores de um grupo específico, o ACLS permite propriedades com valores não restritos, podendo ser aplicado para tarefas das mais complexas, tal como o reconhecimento de imagens.

O “Assistant” (ver figura anterior), por sua vez, generaliza atributos do ACLS, permitindo atributos valores contínuos. E, embora não produza uma árvore de decisão iterativa, como acontece com o ID3, tem o poder de escolher um conjunto de treinamento para os objetos disponíveis. Essa classe de algoritmos tem sido bastante utilizada no campo médico, por exemplo.

Ainda na figura anterior, os três sistemas à esquerda são derivações comerciais do

¹⁰This paper focusses on one microcosm of machine learning and on a family of learning systems that have been used to build knowledge-based systems of a simple kind.

¹¹Em uma árvore de decisão, embora seu desenvolvimento se dê das raízes às folhas, a sua representação começa pelo nó-raiz, na parte superior, descendo em direção às folhas.

ACLS, que não estão relacionados a avanços teóricos consideráveis, mas incorporam inovações simples e de sucesso na geração e utilização de árvores de decisão.

Conforme mencionado, a tarefa das árvores de decisão desse tipo é a de classificação. Seu produto será, pois, uma classe. Para tal, (57) descreve que a estratégia subjacente a essa árvore é não-incremental, ou seja, um grupo de casos relevantes é apresentado, os exemplos são dados, mas não há uma ordem específica de apresentação dos mesmos.

Ainda do ponto de vista histórico, no final da década de 80 e início da década de 90 (58) é desenvolvido o algoritmo C4.5, uma evolução significativa do ID3, que consegue lidar tanto com atributos categóricos, quanto contínuos. É também capaz de lidar com valores desconhecidos, representados por “?” e sendo tratados de forma especial no processo.

No W.E.K.A. (Waikato Environment Knowledge Analysis) é disponibilizada sua implementação, passando a ser chamado J48, que é a implementação na linguagem Java do algoritmo utilizado na pesquisa contemplada nessa dissertação.

O C4.5 é capaz de analisar a medida de ganho, introduzindo um conceito fundamental para o avanço desse algoritmo: a “poda”, que é realizada utilizando medidas estatísticas para identificar e, posteriormente, excluir ramos. Tal processo permite o recorte de ramos que não apresentam contribuição significativa, melhorando o desempenho do algoritmo, que se tornou um dos mais utilizados na literatura que contempla árvores de decisão.

São identificadas dois momentos em que são realizadas as podas. O primeiro é o de pré-poda, efetivado no treinamento e que se caracteriza pela interrupção do processo de divisão do nó “em função da avaliação de um conjunto de medidas, transformando o nó em folha rotulada com a classe majoritária” (59). A pós-poda, por sua vez, é executada findo o processo de construção da árvore, e é aplicada recursivamente, na direção de baixo para cima.

Segundo Quinlan (58), os dados de entrada do C4.5 são caracterizados por uma coleção de casos de treinamento, cada uma com um tupla de valores para um grupo de atributos (variáveis independentes) e uma classe de atributos (variáveis dependentes). Um atributo, por sua vez, pode ser contínuo ou discreto.

Para Ian e Frank (60), as árvores de decisão geradas a partir do C4.5 podem ser representadas por uma abordagem “dividir para conquistar” para resolução de problemas de aprendizagem, a partir de um conjunto de instâncias independentes. Os nós em uma árvore de decisão “testam” um atributo específico, comparando seu valor com uma constante. No entanto, algumas árvores podem comparar dois atributos com outros ou utilizarem uma função para tal.

O último algoritmo de árvores de decisão que pretendemos contemplar nessa breve revisão histórica é o CART – Classification and Regression Trees (61). Esse algoritmo produz tanto árvores de classificação (para o caso de atributos discretos) quanto de regressão (para atributos contínuos).

O CART é conhecido, sobretudo, pela técnica de partição recursiva binária, tendo em vista que cada nó é dividido em dois outros nós, que podem ser divididos, cada um, em mais dois nós, sucessiva e recursivamente. É estabelecido um conjunto de regras que dão suporte à divisão de cada nó, até a decisão de que a árvore está completa.

Diferentemente do C4.5, o CART não realiza pré-poda. A poda acontece ao final – pós-poda - da árvore gerada, em seu tamanho máximo, e por meio da relação custo-complexidade (61), obtendo, muitas vezes, subárvores, que são analisadas e, via de regra, a melhor delas é escolhida. Ainda no que diz respeito ao CART, o critério de classificação utilizado é o índice de Gini. Esse índice tem como base o cálculo da entropia, e é utilizado frequentemente como parâmetro de pesquisa no campo sócio-econômico.

A entropia é um conceito, utilizado na química e na física, para medir a quantidade de desorganização da matéria. WIENER e SHANON (62) lançaram mão desse conceito para analisar a desorganização da informação. Quando há alta entropia, pode-se dizer que a informação está com nível considerável de desorganização ou de medida de incerteza.

No caso das árvores de decisão, a entropia é citada por RUSSEL & NORVIG (18) relacionada ao ganho de informação. Quando um atributo é identificado como aquele que está relacionado a um maior ganho de informação (ou maior redução de entropia), ele é escolhido como o atributo teste para o nó. Tal atributo teria, então, a função de diminuir a aleatoriedade ou impureza nas partições (59). A seguir a equação que representa o cálculo padrão da entropia (já mencionada na seção 2.2.2)

$$H_x = - \sum_{\forall x \in X} P(x) \log_2 P(x) \quad (2.5)$$

H_x é a medida de entropia, x um atributo do conjunto de variáveis X de variáveis.

A entropia condicional, formalizada na equação seguinte, é a entropia restante dos atributos de Y no valor y quando o atributo X é dado como x (24):

$$H_{Y|X} = \sum_x P(x) H(Y|X=x) = - \sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y|x) \log_2 P(y|x) \quad (2.6)$$

Com essa abordagem é possível reduzir o número de testes necessários para que uma árvore seja produzida.

Redes Neurais

O cérebro humano possui cerca de 10 bilhões de neurônios, que são responsáveis pelo funcionamento do organismo. Esses neurônios se conectam entre si, através de sinapses, formando uma Rede Neural capaz de armazenar e processar grande quantidade de informações (18).

De forma semelhante às Redes Neurais naturais, foram desenvolvidas as Redes Neurais Artificiais, que recebem esse nome por se caracterizarem como um sistema cujo funcionamento é semelhante à arquitetura das redes neurais humanas.

No início da década de 40, precisamente em 1943, McCulloch e Pitts (63) propuseram um modelo simplificado de funcionamento do cérebro humano e, a partir daí sugeriram a construção de uma máquina que fosse inspirada nesse funcionamento.

Definições e aprendizado em Redes Neurais Artificiais

São várias as definições que podem ser encontradas sobre o que vem a ser uma Rede Neural Artificial (RNA) (19), em função da complexidade de tal Rede. Do ponto de vista computacional, uma RNA configura-se como uma técnica para solucionar problemas de Inteligência Artificial (IA) que estabelece um modelo matemático baseado em funções de um modelo neural biológico simplificado, com capacidade de estabelecer conexões sinápticas, bem como, de aprendizado, generalização, associação e abstração.

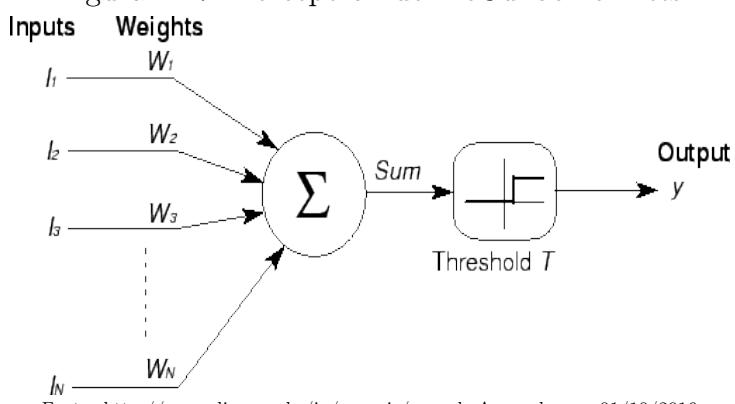
Os neurônios, nessas redes são conhecidos como Perceptrons. O arranjo em camadas desses perceptrons é chamado *Multilayer Perceptron*, que é responsável pela resolução de problemas mais complexos que não seriam passíveis de resolução pelo modelo de neurônio básico.

São três as camadas usualmente identificadas em uma rede de perceptrons:

- Camada de entrada: nessa camada apresenta-se os padrões à rede;
- Camadas Intermediárias (ocultas): aqui é realizada a maior parte do processamento por conexões ponderadas. São consideradas como as camadas extratoras de características;
- Camada de saída: responsável pela conclusão e apresentação do resultado.

As unidades de processamento são conectadas por canais de comunicação, associados a determinado peso, como mostra a Figura 2.17, proposta por McCulloch e Pitts (64). Esses canais são os *inputs*, $I_1, I_2, I_3, \dots, I_n$, e cada um tem um peso associado, que serão calibrados de acordo com a aproximação do resultado esperado pela rede neural, produzido na saída (fase de “forward”). Essa aproximação é conhecida como erro ou erro padrão. Esse erro será propagado de volta à entrada, retroalimentando a rede neural (fase de “backward”), caso o modelo de rede de aproximação seja o “backpropagation”. Dessa forma a rede neural se aproxima cada vez mais do resultado que foi previamente estimado na fase de treinamento.

Figura 2.17: Perceptron de McCulloch e Pitts



Uma rede neural passa por um processo aprendizado, resultante de uma fase de treinamento, estabelecido a partir de casos reais, que a faz adquirir, a partir de então, a sistemática que é necessária para executar o processo desejado satisfatoriamente. Uma vez “treinada” os pesos estão calibrados para solucionar a classe de problemas para o qual foi desenhada.

O treinamento supervisionado tem sido o mais comumente utilizado em redes neurais (65). Em linhas gerais, como o nome indica, no treinamento supervisionado há um agente externo que indica explicitamente à rede um comportamento bom ou ruim, com base no padrão de entrada. Os valores iniciais dos pesos são aleatórios, e o ajuste se dá a partir do algoritmo de aprendizado, na próxima interação ou ciclo seguinte. São apresentados sinais de entrada e de saída à rede, e os ajustes vão sendo feitos paulatinamente. O treinamento pode levar um período considerável de tempo, em função dos ajustes que vão sendo realizados. O treinamento é concluído quando a rede neural atinge um determinado patamar de desempenho, tendo alcançado a precisão estatística esperada. Não havendo mais necessidade de treinamento, “congela-se” os pesos, para sua aplicação.

O aprendizado não supervisionado, por sua vez, não depende de um agente externo, pois funciona de forma autorregulatória, apresentando mecanismos que analisam as regularidades ou tendências dos padrões de entrada, possuindo a capacidade de se adaptar automaticamente às necessidades da rede.

2.6 Métricas aplicadas à mineração

Quando são desenvolvidos sistemas de predição e análise de diagnóstico, avalia-se o desempenho e a qualidade dos resultados encontrados. Um método gráfico eficiente para detecção e avaliação da qualidade de sinais, conhecido como *Receiver Operating Characteristic* – ROC, ou curva ROC (66), foi criado e desenvolvido na década de 50 do século passado, para avaliar a qualidade da transmissão de sinais em um canal com ruído. Recentemente a curva ROC tem sido adotada em Mineração de dados e Aprendizagem de Máquina (67), em sistemas de suporte à decisão na medicina, para analisar a qualidade da detecção de um determinado teste bioquímico, na psicologia para detecção de estímulos (68) em pacientes, e na radiologia para classificação de imagens.

Essas métricas são amplamente utilizadas na classificação binária de resultados contínuos. Para isso ser construído utiliza-se a Matriz de Contingência que classifica as probabilidades como: verdadeiro positivo, falso positivo, falso negativo e verdadeiro negativo, respectivamente *True Positive* – TP , *False Positive* – FP , *False Negative* – FN e *True Negative* – TN , também conhecida como matriz de confusão, descrita na tabela a seguir:

Tabela 2.3: Matriz de Confusão

	Preditivo	
Real	TP FN	Positive – POS
Real	FP TN	Negative – NEG
—	PP PN	—

Fonte: (69)

A matriz da Tabela 2.3 sintetiza a matriz da Tabela 2.4, portanto as duas tabelas são equivalentes.

Tabela 2.4: Matriz modelo de Confusão

	\bar{Y}	
X	$P(X,Y) P(\bar{X},\bar{Y})$	Positive – POS
\bar{X}	$P(\bar{X},Y) P(\bar{X},\bar{Y})$	Negative – NEG
—	$P(Y) P(\bar{Y})$	—

Fonte: (69)

De acordo com as probabilidades condicionais temos:

$$P(X,Y) = P(X|Y).P(Y) = P(Y|X).P(X) \quad (2.7)$$

Então, a taxa de verdadeiros positivos será $P(Y|X)$ e a probabilidade de falsos alarmes ou taxa de falsos positivos será $P(Y,\bar{Y})$, a barra sobreposta em \bar{X} (ou \bar{Y}) representa negação.

A curva ROC será construída cruzando-se a taxa dos verdadeiros positivos ($tpr = P(Y|X)$) com a taxa dos falsos positivos ($fpr = P(Y,\bar{X})$).

2.6.1 Classificação para análise preditiva

Classificação é um processo para encontrar um modelo que descreve e distingue classes de dados. Esse modelo tem como base de análise um conjunto de treinamento (i.e. objetos de dados para os quais serão encontrados rótulos que os classifiquem).

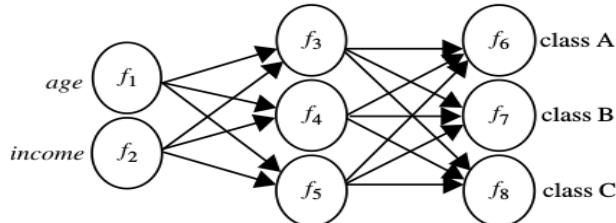
Esse modelo é usado para prever quais rótulos de classes terão os objetos desconhecidos. O modelo pode ser representado por regras de classificação do tipo “IF - THEN”, por árvores de decisão, redes neurais e outros. Regras de classificação se distinguem de regras de indução da seguinte forma:

- Uma regra de classificação poderia ser: $if L \text{ then } \text{class} = C_1$ ou $if L \text{ then } \text{class} = C_1$
- Uma regra de indução seria: $if L \text{ then } R$ que por sua vez produz novas regras

$age(X, "youth") \text{ AND } income(X, "high") \rightarrow classe(X, "A")$
$age(X, "youth") \text{ AND } income(X, "low") \rightarrow classe(X, "B")$
$age(X, "middle-aged") \rightarrow classe(X, "B")$
$age(X, "senior") \rightarrow classe(X, "B")$

A Figura 2.18 representa uma rede neural com as mesmas características da árvore de decisão anterior:

Figura 2.18: Rede Neural



Fonte: Han, J. and Kamber, M.

As árvores de decisão produzem regras de indução, são algoritmos rápidos, contudo dados impuros podem comprometer o desempenho desse algoritmo. A fase de extração dos dados é fortemente influenciável pelas variáveis escolhidas, (24) isso pode representar o desafio maior para implementar esta técnica.

Outro problema que pode ser encontrado em algoritmos de aprendizagem é o “overfitting” ou superadaptação aos modelos. Segundo RUSSEL E NORVIG (18)¹² o “overfitting” ocorre quando o número atributos é grande.

¹²Foi observado que redes neurais muito grandes generalizam bem, desde que os pesos sejam mantidos pequenos. Essa restrição mantém os valores de ativação na região linear da função sigmóide $g(x)$ onde x é próximo de zero. Por sua vez isso faz com que a rede se comporte como uma função linear, com um número muito menor de parâmetros.

2.6.2 Pesquisas envolvendo inteligência artificial e tráfego em rodovias

Nesta sessão apresentaremos algumas pesquisas realizadas no Brasil e em outras partes do mundo, que contemplam elementos de interesse para nossa pesquisa: tráfego em rodovias e o uso de técnicas de inteligência artificial, como árvores de decisão e redes neurais, dentre outras.

Para tal, tomamos como referência a última década de produção, priorizando artigos a partir de 2010, embora tenham sido identificadas pesquisas desde meados dos anos 90, já contemplando tráfego urbano ou em rodovias e inteligência artificial.

No Brasil, sobretudo no eixo sul-sudeste, algumas pesquisas têm sido conduzidas nesse campo. Corrêa (70) desenvolveu um estudo exploratório, em que buscou identificar pesquisas no Brasil que tratasse da utilização de redes neurais aplicadas ao setor de transporte, comparando com aquelas produzidas sob a mesma temática, em países desenvolvidos. Essa autora identifica a produção de pesquisas nesse cenário desde os anos 90, corroborando a análise que fizemos sobre pesquisas que foram produzidas nesse campo, com auxílio da ferramenta de busca google acadêmico.

Identificamos campos de interesse que podemos resumir da seguinte forma: pesquisas interessadas em utilizar técnicas de IA para identificar comportamento de tráfego ou causas de acidentes graves, notadamente os acidentes fatais, podendo, ou não preocupar-se com a predição de acidentes futuros, mas sem propor algum tipo de solução prática ou desenvolvimento de ferramentas auxiliares do tráfego. Outro grupo de pesquisas que se interessa pela utilização da IA para proposição de soluções práticas, aplicativos, ferramentas, para lidar com os constrangimentos de tráfego, sem preocupação específica com, por exemplo, causas de acidentes fatais. Outro grupo que mobiliza técnicas de IA e que além da identificação de problemas relativos ao tráfego em vias urbanas e rodovias, preocupa-se com a modelização e predição de rotas alternativas, nas quais essa pesquisa se insere, embora seja importante destacar que há aspectos dos outros dois grupos que também nos interessam, como, por exemplo, as causas de acidentes, sobretudo que culminam com óbito.

A seguir, destacamos alguns estudos, mesmo oriundos de contextos tão diversos, apontam para interesses semelhantes. Optamos por destacar pesquisas relacionadas às mais diversas regiões, no sentido de destacar o quanto esse campo é vasto e de interesse mundial.

Em uma pesquisa conduzida nos Estados Unidos, com dados obtidos a partir do Sistema de Relatório de Análise de Fatalidades (FARS), no Alabama, Shanthi e Ramani (71) buscaram analisar em que medida os algoritmos de classificação de mineração de dados são eficazes para previsão dos fatores que influenciam acidentes de trânsito, levando em conta a gravidade da lesão decorrente do acidente. Nessa pesquisa foi comparado o desempenho de algoritmos como C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes e Árvore de randomização, para modelar a gravidade da lesão ocorrida no acidente. A acurácia foi avaliada com base na precisão e valores de recall e os resultados apontaram para a eficácia desses algoritmos, com destaque para as árvores de randomização.

Boa parte das pesquisas desenvolvidas no Brasil têm focado na identificação de problemas relacionados a congestionamento de vias urbanas ou rodovias no interior do estado, propondo, a partir da utilização de técnicas de IA, como redes neurais, árvores de decisão, lógica fuzzy, identificar o comportamento dos veículos nas vias, como a pesquisa de Ferreira (72), que buscou utilizar combinação de técnicas de IA para previsão do comportamento do tráfego na cidade de São Paulo.

Pertencente a outro campo de interesse - aquele voltado para a proposição de soluções práticas - Braga et al (73) utilizaram o algoritmo de IA Anytime e executaram os testes na Plataforma Raspberry, com o objetivo de estabelecer um planejamento de rotas nas malhas viárias das cidades de Manaus e do Rio de Janeiro. A plataforma utilizada na pesquisa mostrou-se eficaz para a proposição de um protótipo de planejamento de rota, para uso veicular e abre caminho para o desenvolvimento de um aplicativo para smartphones.

Outra pesquisa, também em língua portuguesa, conduzida em Portugal (74), propôs o desenvolvimento de soluções inteligentes, a partir do uso de data mining e técnicas de simulação, para estabelecer previsões das condições de tráfego em tempo real, de modo a conter o aumento dos congestionamentos em rodovias e região metropolitana. Essa recente pesquisa aproxima-se, em certo sentido, do nosso interesse, uma vez que tem como foco uma abordagem de previsão, tal qual propomos nessa pesquisa. No caso da pesquisa desenvolvida por Reis, os dados históricos utilizados foram capturados a partir dos sensores localizados nas estradas, diferentemente de nossa proposta, que contempla dados históricos da Polícia Rodoviária Federal de Pernambuco (mineração de dados) e da rede social Twitter (mineração em textos).

A preocupação com acidentes graves, muitos deles implicando em mortes, também se revela em pesquisas em países desenvolvidos ou países em que é sabido haver problemas ligados ao tráfego nas capitais e nas importantes cidades do interior, como é o caso de alguns países do Oriente Médio. Essas pesquisas estariam relacionadas ao primeiro campo que destacamos acima: as que buscam identificar e alertar as autoridades sobre o impacto de acidentes fatais relacionados ao tráfego.

Uma pesquisa conduzida nos Estados Unidos por Chong, Abraham, Paprzycki (75), destaca o grande impacto social que esse tipo de fatalidade promove. Esses pesquisadores utilizaram as redes neurais e árvores de decisão para modelizar os acidentes graves e fatais, identificando três causas fundamentais que promovem acidentes dessa natureza: a não utilização de cinto de segurança, a condição de luminosidade na rodovia (que em nossa pesquisa está incluída na variável condições de visibilidade) e consumo de álcool pelo condutor.

Pesquisa semelhante foi desenvolvida por Akgüngör e Dogan (76), que utilizaram redes neurais e algoritmos genéticos para propor um modelo para estimar o número de acidentes (A), com vítimas fatais (F) e prejuízos (I), em Ankara (Turquia), utilizando dados de um espaço temporal de 20 anos (1986 a 2006), de modo a oferecer um retrato fiel dos impactos sociais causados pelas condições precárias do tráfego nas grandes cidades turcas.

Kashani, Shariat-Mohaymany, Ranjbari (77), em pesquisa desenvolvida no Irã, utilizaram a árvore de regressão e classificação (CART) para analisar os acidentes fatais de trânsito que acontecem no Irã. Segundo os dados da pesquisa, setenta por cento dos acidentes com vítimas fatais no Irã acontecem nas áreas rurais. A pesquisa identificou que as principais causas de morte relacionadas ao tráfego são: a ausência do uso de cinto de segurança, as ultrapassagens indevidas e o consumo de álcool por parte do condutor. Tomando como referência a nossa pesquisa, destacamos que as causas consideradas como mais relevantes dizem respeito ao condutor do veículo, mais do que às condições da via.

Em uma pesquisa conduzida na Jordânia, Jadaan, Al-Fayyad e Gammoh (78), tiveram como objetivo propor um modelo de previsão de futuros acidentes de tráfego, utilizando-se de Redes Neurais. A pesquisa considerou o espaço temporal entre 1990 e 2011, e tomou como dados de entrada das Redes Neurais: número de veículos registrados, população, comprimento de estradas pavimentadas, produto interno bruto. Foram propostos quatro modelos alternativos, e um deles (4o.) mostrou-se o mais adequado, com coeficiente de de-

terminação ($R^2 = 0,992$), a partir do qual propuseram o modelo de predição, considerando os dados atuais e o número previsto de acidentes.

Beshah e Hill (79) desenvolveram uma pesquisa na Etiópia, utilizando algoritmos de IA, para identificar as principais causas de acidentes graves, principalmente aqueles com mortes, que vitimizam aproximadamente 1 milhão e 200 mil pessoas, além de deixar cerca de 50 milhões de feridos. Os pesquisadores utilizaram árvores de decisão, Naïve Bayes e outros classificadores, para identificar as principais causas dos acidentes. Consideraram algumas variáveis que também interessaram particularmente à nossa pesquisa, como condições de visibilidade da via, sentido da via, gravidade do acidente. Também conforme o interesse de nossa pesquisa, o estudo em questão interessou-se pelo estabelecimento de um modelo preditivo, a partir das tecnologias de IA utilizadas.

Interessados pela mesma temática, Olutayo e Eludire (80) desenvolveram pesquisa na Nigéria, utilizando técnicas de Árvores de Decisão e Redes Neurais para analisar o tráfego em rodovias de tráfego intenso, com o objetivo de, segundo eles, reduzir a carnificina nas grandes rodovias do país. Coletaram dados históricos da instituição de segurança rodoviária daquele país, referentes a 40 km de rodovia, entre as cidades de Ibadan e Lagos, classificando-os como contínuos ou como categóricos. Para os dados contínuos foram utilizadas a técnica de Redes Neurais Artificiais, e os dados categóricos foram tratados com Árvores de Decisão. Os resultados apontaram que as principais causas de acidentes fatais são ruptura de pneu, perda de controle e alta velocidade. Assim como outros estudos, as causas relacionadas aos condutores aparecem como aquelas mais fortemente associadas a acidentes fatais. Os autores propuseram, ainda, estabelecer uma comparação entre RNA e AD, como técnicas de IA associadas à resolução do problema de pesquisa, e identificaram que as RNA um relatório de erro mais baixo, bem como maior número de instâncias corretamente classificadas e taxa de precisão mais elevada.

Também com o objetivo de comparar os resultado obtidos utilizando Redes Neurais e Árvores de Decisão, pesquisa conduzida na Tailândia (81) propôs uma técnica para identificar níveis de congestionamento em rodovias, com base nos dados de sensores de velocidade de alta acurácia; e utilizaram GPS, webcam e uma pesquisa de opinião, essa última para detector percepções humanas/dos condutores sobre os níveis de congestionamento. Essa pesquisa identificou o J48 como o algoritmo de aprendizagem que mostrou o melhor resultado, com acurácia de 91,29%. Um aspecto que diferenciou essa pesquisa das demais, foi o fato de que considerou as percepções humanas para sua análise. Em nossa pesquisa, os dados obtidos a partir do Twitter e a mineração de textos e análise sentimental, permitiu-nos, igualmente, contemplar a dimensão humana, para além da utilização de dados históricos.

Outros estudos poderiam ainda ser considerados, contemplando ainda a comparação entre algoritmos para gestão segurança rodoviária no sul da Itália (82) análise de acidentes em áreas rurais da Espanha (83); utilização da Plataforma GIS para o desenvolvimento de um sistema de reconhecimento automático de acidentes de trânsito na rodovia Konya-Afyonkarahisar, na Turquia (84), mas acreditamos ter feito o levantamento dos principais pesquisas desenvolvidas nesse campo, na atualidade, o que corrobora com o que discutimos sobre a importância de desenvolvemos um estudo dessa natureza, sobretudo contemplando aspectos ainda não encontrados na literatura, como a integração das redes sociais a essas pesquisas, propondo um trabalho de mineração de texto e análise sentimental, em articulação com os dados produzidos pelos algoritmos de IA.

3

Contribuição

A contribuição dessa pesquisa é de cunho metodológico-prático. Do ponto de vista metodológico, pela aplicação do processo CRISP-DM, usado para construir o modelo preditivo e classificativo, em termos práticos pela integração entre mineração de dados históricos e mineração em textos; pela utilização de algoritmos de classificação e de previsão.

No que diz respeito à mineração de dados, as técnicas de classificação utilizadas foram Árvores de Decisão e Naïve Bayes. Quanto às técnicas de previsão, utilizou-se Redes Neurais (regressão logística, como caso particular). Para mineração de textos foi utilizado o Text frequency - Inverse Document Frequency (TF-IDF).

Ainda do ponto de vista prático, propomos um modelo que integre previsão à API de mapas de posicionamento global, fornecendo informação suficiente para um utilizador tomar uma decisão acerca dos dias e horários que ofereçam menor risco de acidentes ou qualquer evento que implique na retenção do fluxo de veículos.

As soluções disponíveis que existem, tais como; Google Maps, Waze e outras ferramentas dessa natureza somente exibem, informações momentâneas, produzidas e compartilhadas pelos utilizadores dos aplicativos, ou por informações provindas de GPS. Contudo, não analisam dados históricos dessas rodovias, nem fazem previsões sobre o seu comportamento.

Outra contribuição dessa pesquisa é a proposição de um arco cibernético construído com a API de redes sociais. Os “feeds” de notícias das redes sociais como o Twitter permitem analisar o contexto das rodovias com defasagem temporal muito pequena. Os utilizadores dessas redes sociais contribuem com informações relevantes como, por exemplo, o anúncio de uma paralisação que ocorrerá daqui a alguns dias. A PRF de Pernambuco é outro contribuidor permanente. Com seu canal no Twitter: @PRF191PE, fornece diariamente informação das rodovias, além de dados estatísticos relativos a acidentes e retenção de tráfego, como aqueles decorrentes de protestos, que, quando feitos dentro da lei devem ser informados à PRF, com dia e hora marcados antecipadamente.

A monitoração de redes sociais foi feita a partir de Mineração em Textos - Knowledge Database Text (KDT) ou simplesmente Text Mining (TM), tendo, a princípio, sido executada no Twitter, no canal @PRF191PE e pela busca em outros canais, com as pa-

lavras: “protesto”, “planejar paralização”. Posteriormente pesquisou-se em: @PRF191PB, @PRF191ALAGOAS, @PRF191RN, @PRF191CEARA, @PRFBAHIA, 191SAOPAULO, @PRF191MG, @PRFRJ191, @PRF191PR, @PRFGOIAS.

Para ter-se acesso aos tweets da PRF, os usuários do canal, inclusive a própria PRF, postam comentários diretamente na API do Twitter ou por um navegador. Para essa pesquisa as palavras-chave tais como: protestos, acidentes, paralisação, foram de suma importância.

Uma vez capturados, esses tweets foram tratados por Mineração em Textos e analisados instantaneamente por algoritmos de I.A. A técnica utilizada para análise dos textos foi a análise de palavras ou termos frequentes (TF-IDF). O algoritmo de classificação escolhido foi Naïve Bayes, por ser um classificador rápido e eficiente, e por ter sido utilizado na primeira fase dessa pesquisa, servindo como comparativo à Árvore de Decisão. O algoritmo de agrupamento (clustering) foi o K-means.

3.1 Modelo Proposto

A metodologia utilizada nessa pesquisa contempla um plano em três etapas, cada uma dividida em fases atinentes. A primeira etapa da nossa metodologia contempla todas as fases do CRISP-DM, onde está o modelo classificativo, o preditivo e a descoberta de conhecimento sobre o comportamento das rodovias estudadas. A descoberta de conhecimento sobre esse comportamento tem a ver com o “modus operandi” dos seus utilizadores. A priori, especulou-se sobre possíveis erros de traçados e outros que pudessem ser identificados pelos algoritmos de mineração empregados no processo.

Os algoritmos escolhidos contemplaram algumas características especiais, tais como, robustez, tolerância à faltas de dados (missing data), aprendizagem e facilidade de interpretação dos dados processados.

No quesito tolerância à faltas e facilidade de interpretação dos dados, a Árvore de Decisão e o Naïve Bayes se destacaram por não necessitar de qualquer requisito extra interpretar os resultados.

No quesito robustez, tolerância à faltas e capacidade de aprendizagem relativamente alta, as redes neurais artificiais (RNA), com a topologia Perceptron multicamadas com retroalimentação “backpropagation”, se destacaram por terem adequada capacidade de generalização e especificidade em modelos de previsão.

A extração do modelo preditivo deverá ocorrer quando este se integrar a uma estrutura dinâmica, composta pelas redes sociais e mapas vetoriais, dado um espaço temporal pré-determinado por um agente: o utilizador.

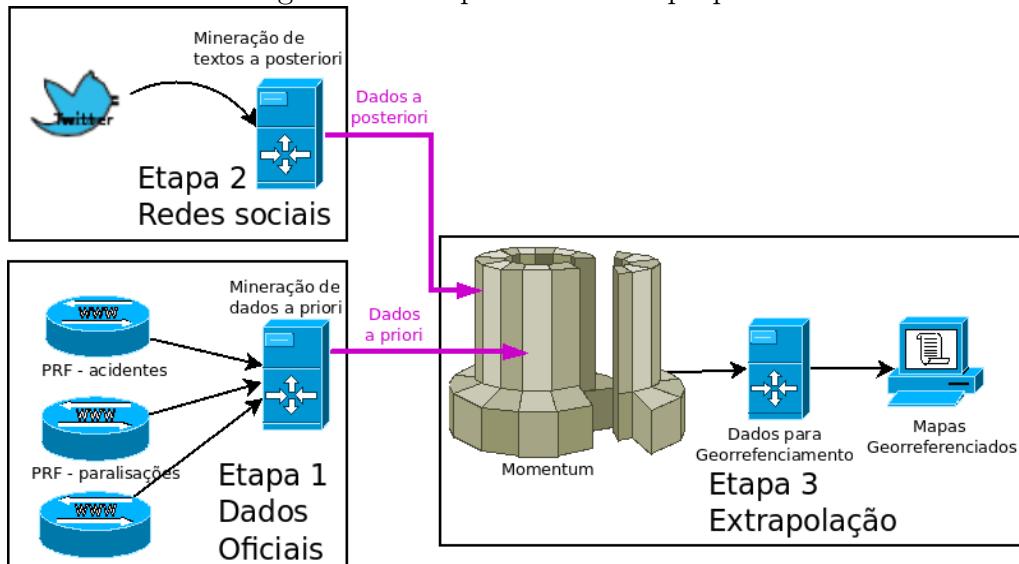
Através de API’s, os mapas vetoriais permitem a geolocalização dos pontos classificados ou pontos onde haverá grande número de retenções, conhecidos como gargalo.

Para a integração às redes sociais, foi escolhida a API do Twitter. Esta “interface” é simples de ser configurada. A quantidade de informações produzidas pelos utilizadores gera poucos dados em cada postagem, mas é eficaz. O utilizador tem que ser sucinto ao publicar suas postagens em um espaço de 140 caracteres. Isso facilita a forma como os dados são extraídos pela quantidade diminuta, bem como a quantidade de conexões à Internet. Contudo, esta rede social tem uma crescente quantidade de postagens no formato imagens, dificultando a mineração em textos. Esse aspecto foi particularmente relevante para essa pesquisa, quando foi constatado que a PRF também aumentou consideravelmente o número de tweets no formato de imagem.

A alternativa escolhida para tratar essa dificuldade foi buscar conexões dos outros utilizadores do canal da PRF, uma vez que as redes sociais, do ponto de vista tecnológico, se caracterizam como grafos e subgrafos interconectados, permitindo descobrir outras sub-redes que possivelmente conterão as informações desejadas.

A figura a seguir, Figura 3.1, ilustra essa metodologia descrita.

Figura 3.1: Etapas da modelo proposto



Fonte: autor

3.2 Reflexão sobre as tecnologias utilizadas no modelo preditivo

Na fase de transformação de dados, da primeira etapa, onde são relacionadas as variáveis, o acoplamento entre as bases heterogêneas foi conseguida utilizando regras de indução da lógica proposicional (18). Nesta pesquisa, bases heterogêneas foram integralizadas em um único grande conjunto de dados, o “data set”. As variáveis desse “data set” foram consideradas variáveis independentes, e preservadas aquelas com maior relevância ou as que continham a maior quantidade de conhecimento embutido, descoberto pelo cálculo da entropia e correlação linear. Foram também construídas novas variáveis, nas bases onde não havia correspondência, respeitando a lógica do negócio.

A tabela 3.1 descreve as variáveis transformadas na base de dados de acidentes da PRF.

Tabela 3.1: Variáveis transformadas

KM	Numeração do quilômetro arredondada
BR	Nome da BR
Condição Pista	Condição da pista: seca, molhado, ...
Restrição de Visibilidade	Restrição de visibilidade: inexistente, neblina, .., outros
Tipo Acidente	Tipo de Acidente: atropelamento, colisão lateral,..
Causa Acidente	A possível causa do acidente: Falta de atenção, ...
Traçado Via	Tipo de traçado da via: reta, curva, cruzamento, ...
Tipo veículo	Tipo de veículo envolvido no acidente
Dia da Semana	dia em que ocorreu o acidente
Hora	que ocorreu o acidente/ocorrência no formato hh/mm/ss
Qtd Mortos	Quantidade de mortos envolvidos
Gravidade	Quantidade de acidentes graves
Período	turno do dia em que se deu a ocorrência

3.3 Extração do conhecimento - KDD

O processo de descoberta do conhecimento iniciou-se com a coleta das bases de dados de acidentes da PRF. Optamos por coletar os dados dessa base diretamente na fonte, ou seja dos servidores da PRF. Tais dados nos foram cedidos após alguns procedimentos burocráticos de praxe (ver anexos). Essa escolha foi motivada para tentar mitigar o problema da qualidade dos dados. No artigo “Uma análise da qualidade dos dados relativos aos boletins de ocorrências das rodovias federais para o processo de Mineração de Dados”, COSTA, BERNARDINI, LIMA et al (85) destacam a não padronização e não aceitação dos dados pela comunidade internacional. EAVES, D. (86) sugere que os dados sejam disponibilizados da maneira como foram coletados.

A PRF tem ao menos duas bases ¹ de dados referentes às ocorrências nas rodovias BRs. A base de acidentes rodoviários e a base de intervenções que guarda as ocorrências que paralisaram as rodovias, tais como: protestos ou paralisações dessa natureza.

Para traçar um painel da diversidade das rodovias pernambucanas, foi efetuado a priori uma classificação através do algoritmo Árvore de Decisão (81) e comparado com o classificador Naïve Bayes (83). Mediú-se a acurácia dos classificadores, comparando-se uma técnica algorítmica com a outra. A variável “BRajustada” mostrou ser a melhor variável para exprimir o nó raiz da Árvore de Decisão. Isso porque classificou os dados com o maior número de verdadeiros positivos e menor número de falsos positivos, muitos iguais a zero, obtendo curvas ROC com índices acima de 0.90. O Naïve Bayes também obteve índices de acurácia próximos a este, quando utilizado com o mesmo procedimento.

Foram construídas matrizes dos resultados da classificação e predição, chamadas “Matriz de Mortos” e “Matriz de Gravidade”. As matrizes de mortos refletem a quantidade de óbitos ocorridos em cada hora do dia, em determinado quilômetro da rodovia, contemplando, assim, duas dimensões (ver Figura 3.2). Foi estabelecida uma terceira dimensão para detalhar os dias da semana, uma vez que a utilização da rodovia tem características diferentes nos dias de semana e fins de semana.

¹Somente mencionamos bases de dados que interessaram à essa pesquisa.

Figura 3.2: Matriz de Mortos 2D - BR 101, Período: 2007 à 2015

		MatrizMortos2D																							
		Km (0 - 213)																							
		HORA (0 - 23)																							
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	0
0	0	0	1	0	0	0	17	0	0	0	5	0	0	0	0	0	0	1	2	0	1	1	1	0	
0	2	0	11	0	16	1	3	2	0	0	0	0	0	7	0	1	0	0	0	0	0	4	0	0	
0	0	1	1	0	0	0	1	0	0	0	1	0	0	2	0	0	0	0	0	0	3	0	0	0	
0	0	0	0	4	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	
0	0	1	0	0	1	0	0	1	1	0	2	0	0	0	0	0	0	1	1	0	10	0	0	5	
0	1	6	4	3	5	0	0	4	0	1	0	6	0	0	0	0	0	0	0	0	0	0	2	0	0
4	1	17	2	0	11	0	0	5	0	0	2	1	0	6	0	0	0	0	2	0	8	4	2	10	
0	8	13	5	3	15	1	9	2	2	0	0	4	0	1	0	4	0	0	6	0	2	3	0	11	
0	13	7	5	2	11	0	31	5	0	4	2	0	0	3	1	0	0	1	4	0	6	2	0	0	
0	3	17	10	2	4	0	12	1	2	0	0	0	5	0	1	0	1	0	2	0	19	0	0	4	
0	1	0	1	0	1	0	28	24	2	0	0	3	8	9	1	0	0	0	6	2	17	1	0	0	
4	2	3	0	2	8	0	13	19	4	0	0	2	0	1	0	0	0	0	0	0	10	5	0	4	
3	3	2	4	1	4	2	23	1	0	1	0	0	0	0	2	1	0	0	0	0	7	18	1	8	
2	2	0	0	6	2	0	15	5	0	0	0	8	20	0	10	0	5	0	0	10	0	0	0	2	
0	0	2	4	10	3	0	8	31	7	0	0	2	0	0	0	0	0	0	0	5	0	1	16		
0	3	4	1	2	0	11	10	19	1	0	6	3	0	2	0	0	0	0	16	12	10	1	6		
0	1	3	6	8	1	0	13	33	0	0	0	3	0	1	0	1	0	6	1	0	19	6	3	2	
8	0	10	0	2	7	0	12	5	9	0	3	0	1	1	0	0	3	0	0	0	11	0	0	4	
4	32	0	0	0	10	0	20	21	5	0	0	2	0	9	0	0	0	1	4	2	22	0	0	0	
0	3	2	0	0	0	2	2	22	0	3	4	0	0	6	0	0	0	1	0	8	17	7	0	1	
1	0	3	0	0	0	8	0	13	2	3	4	12	5	6	6	0	0	2	0	0	24	3	0	3	
0	0	0	0	0	0	10	0	19	0	4	0	0	0	7	7	0	0	0	1	0	2	0	0	1	
0	0	7	0	1	4	2	1	11	2	0	0	0	0	9	0	0	0	0	0	0	1	3	0	6	
1	0	10	0	0	0	2	2	12	0	0	2	0	0	0	0	0	0	0	2	0	0	1	1	0	

Fonte: o Autor

O exemplo acima apresenta as ocorrências na BR 101 (período 2007 à 2015), onde as linhas representam as 24 horas do dia (0-23 horas) e as colunas representam cada km da rodovia, no caso dessa matriz 214 km (ou 214 colunas - omitida as restantes), início no km-0 até o fim no km-213.

As matrizes de mortos foram propostas com dois objetivos principais o de demarcar os pontos críticos em cada quilômetro, devido a esse tipo de sinistro e, o de estimar o tempo de paralização da rodovia devido à gravidade deste tipo de acidente. Todavia, nem sempre um acidente com óbito implica em maior tempo de paralisação. Por exemplo, uma colisão que envolva muitos veículos, ainda que não produza óbitos, pode criar um maior constrangimento na rodovia do que, por exemplo, um atropelamento que tenha projetado a vítima para fora da pista. Em virtude disso, foram criadas as matrizes de gravidade, que contemplaram a maioria das variáveis transformadas, descritas anteriormente, resultando numa equação que relacionou os fatores mais relevantes nos quesitos: importância da variável medida pela correlação linear entre elas e o valor da entropia associada à variável e que estejam relacionados a uma ocorrência na rodovia, não somente ao sinistro da representado pela Matriz de Mortos.

A equação que gera a Matriz de Gravidade, contém algumas das variáveis descritas na Tabela 3.1 e foi proposta da seguinte forma:

$$ProbAcid_{101} = (RestVisibi + CondPista + TracadoVia) * Erro_{101} + Gravidade \quad (3.1)$$

Onde:

- RestVisibi - Restrição de Visibilidade;
- CondPista - Condição da Pista;
- TraçadoVia - Traçado da Via;
- Erro_{101} - Taxa de Erro do classificador Árvore de Decisão para a BR 101

A figura 3.3 seguir representa a Matriz de Gravidade 3D da BR 101 para um terça-feira. As linhas contemplam as 24 horas (0-23 horas); as colunas referem-se aos quilômetros da rodovia (0-213 km); e a terceira dimensão (que não aparece nessa imagem), contempla os dias da semana (1 - segunda-feira, 2 - terça-feira, 3 - quarta-feira,..., 7 - domingo).

Figura 3.3: Matriz de Gravidade 3D3^afeira - BR 101, Período: 2007 à 2015

MatrizGravidade3D2															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	0	0	0	0	0	0	4,578	0	0	0,699	0	0	0	0
1	1,518	0	0	1,88	0	0	0	0	0	2,06	0	0	0	0	0
2	0	0	0	0	0	0	0	2,518	0	0	0	0	0	0	0
3	0	0	0	0	2,518	0	0	2,518	0	2,524	0	0	0	0	0
4	0	0	0	1,524	0	0	0	0	3,06	0	0	0	0	0	0
5	0	0	0	0	0	0	1,416	0	0	0	5,12	0	0	2,06	0
6	0	0	2,524	0,258	0	5,12	0	0	0	0	0	2,312	0	0	5,12
7	0	0	3,06	3,501	1,983	3,06	2,518	3,06	4,584	0	0	1,518	3,06	0	0
8	0	2,524	1,159	0,983	2,06	0	1,518	2,423	3,06	0	0	0	0	1,416	0
9	0	0	0	1,524	0	3,06	0	0	0	0	0	3,578	0	0	0
10	0	0	3,578	4,382	1,524	1,524	0	6,12	4,947	0	0	0	0	9,18	0
11	0	0	3,06	0	2,678	0	2,06	3,06	7,156	0	0	0	3,941	2,06	0
12	0	0	0	0	0	0	0	4,85	2,524	1,423	0	2,06	0	0	0
13	0	0	0	0	0	0	0	0	0	2,06	0	0	0	0	0
14	0	0	0,152	0	0,888	3,06	0	0	7,456	2,518	0	2,06	0	2,06	0
15	0	0	2,524	0	2,06	0	3,06	0	4,483	2,06	0	2,524	2,518	0	0
16	0	0	0	0,983	1,524	0	0	3,06	4,94	0	1,524	0	0	2,06	0
17	0	0	0	0	3,06	0	0	0	0	1,8	0	2,524	0	2,423	0
18	0	0	0	0	0	0	2,06	3,06	0	1,778	0	0	3,06	1,518	0
19	0	0	0	0	0	0	0	0	2,311	0	1,518	4,12	0	0	6,12
20	0	0	0	0	0	2,423	0	0	2,06	0	0	0	0	0	3,06
21	0	0	0	1,524	0	0	0	0	0	2,06	0	1,518	0	0	3,06
22	0	0	3,06	0,983	2,518	0	0	0	0	3,06	0	0	0	0	0
23	0	0	3,06	0,876	2,06	0	3,06	0	0	1,518	0	0	0	0	0

Km (0 - 213)

Fonte: o Autor

Ainda sobre as matrizes - Mortos e Gravidade - a primeira contempla o número absoluto de óbitos que ocorreu em cada local da rodovia em todo o período contemplado na base de dados da PRF. A matriz de gravidade, por sua vez, apresenta cada uma das variáveis como uma probabilidade de ocorrência (ver equação 3.1). A variável “gravidade” é binária, sendo o valor 1 para ocorrência de óbito em determinado local e 0 para ausência de óbito. O “Erro” descrito na equação configura-se como um ajuste aos termos que aparecem entre parênteses, e foi utilizado com base na técnica ensemble, que afirma que o produto de dois classificadores resulta em um valor mais preciso.

3.4 Arco cibernetico com dados do Twitter

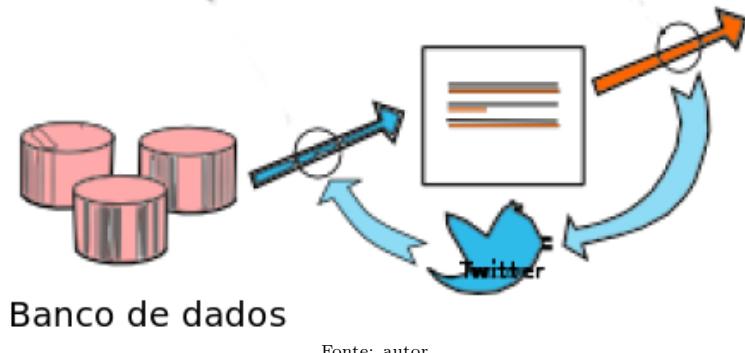
Os dados do Twitter permitem uma busca imediata por novas informações que poderão ser confrontadas com o modelo preditivo, aumentando o nível de confiança deste. Com isso, as informações construirão um arco cibernetico para reforçar ou não inferências anteriores(vide Figura 3.5). Segundo Wiener (1948), novas informações permitem realimentação aos sistemas, dotando-os de maior potencial adaptativo, como pode ser visto nos exemplos de tweets identificados nessa pesquisa: no trecho da Br 101, na altura do km 5, no Município de Goiana, um utilizador do twitter publicou que a comunidade que mora no entorno dessa localidade fará um protesto nas primeiras horas do dia seguinte, devido a um atropelamento ocorrido na BR; ou a PRF publicou que o km 80 da Br 232, na altura do Município de Gravatá, será interditada no dia seguinte, por 2h, para remoção/explosão de rochas.

No entanto, quando as informações provenientes do modelo de predição entrarem em

conflito com as informações provenientes das redes sociais², para estes casos a decisão de qual ação será tomada estará “nas mãos” do agente, do observador ou do utilizador.

As informações das redes sociais que comporão o arco cibرنético não deverão retroalimentar o modelo de predição já construído, pois o fluxo decisório já foi tomado pelo observador. Os dados a posteriori geralmente não servem para um modelo de predição, uma vez que podem enviesar o sistema preditivo, mas a depender de sua qualidade podem ser considerados.

Figura 3.4: Arco cibرنético com dados do Twitter



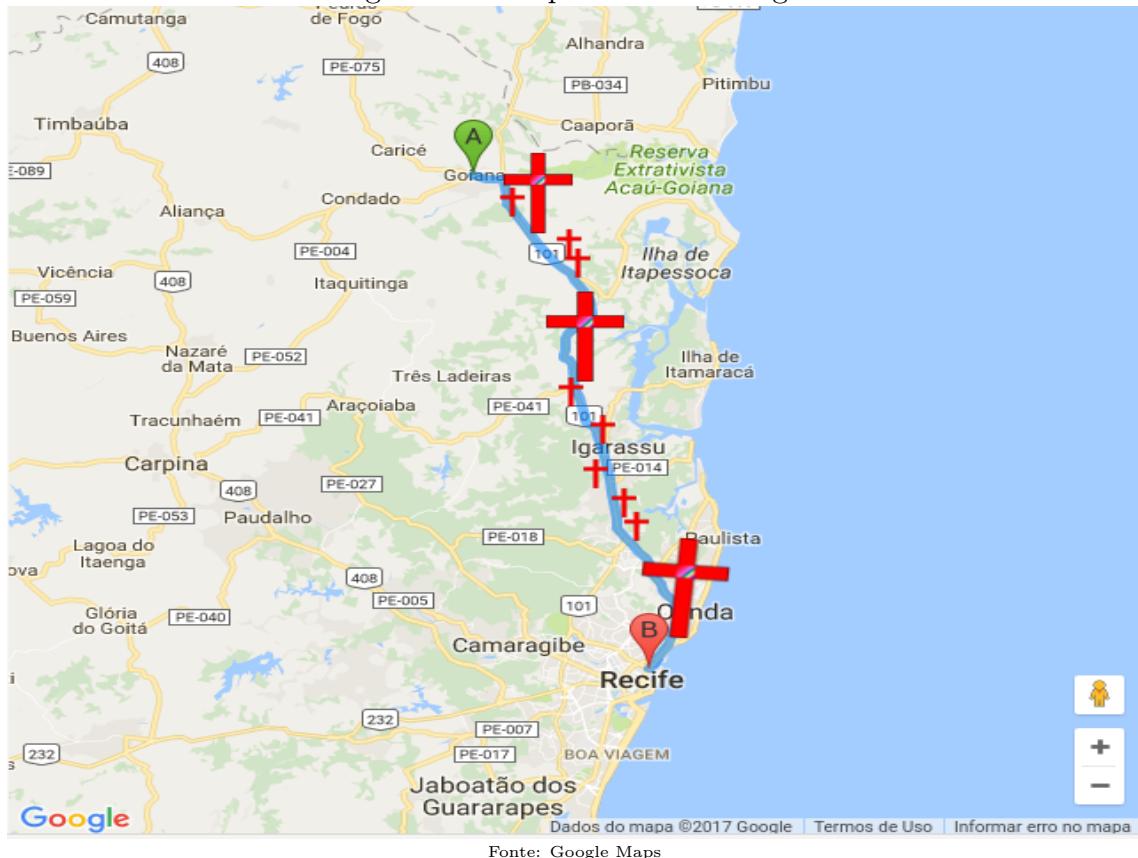
Fonte: autor

3.5 Extrapolação para georreferenciamento

As matrizes apresentadas anteriormente (Figura 3.2 e Figura 3.3) representam pontos críticos, aqueles relacionados com a ocorrência de acidentes graves. Quando a entrada é 0 (zero) nas matrizes, significa que não houve ocorrência naquele local da BR, portanto, quanto maior o valor do ponto crítico, maior gravidade ou maior probabilidade de ocorrência de acidentes na rodovia. Estes pontos podem ser extrapolados para um mapa, a fim de auxiliar os condutores que estejam usando essa tecnologia, quando trafegarem por um determinado trecho da rodovia. Tais pontos são conhecidos nessa pesquisa como extração para georreferenciamento, conforme a Figura 3.5.

²O sistema de predição é baseado em cálculos probabilísticos

Figura 3.5: Etapas da metodologia



Os locais nos Mapas em que aparecem pontos críticos (representados pelas cruzes: cruzes grandes - grande concentração de acidentes graves; cruzes pequenas - demais acidentes) são utilizados para traçar o roteiro de uma viagem, por exemplo, partindo do ponto A para o ponto B, indicado na rodovia. A seguir, Tabela 3.2, uma relação de todas as BRs que cruzam o Estado de Pernambuco que contém esses pontos. O georreferenciamento para esta pesquisa se dará pelo indicador do Km em cada BR.

Tabela 3.2: Rodovias BRs que atravessam Pernambuco

Rodovia	Início	Município	Fim	Município
BR101	km 0.0	Goiana	231.2	Xexéu
BR104	km 0.0	Taquaritinga do Norte	148.0	Quipapá
BR110	km 95.4	Ibimirim	149.1	Ibimirim
BR116	km 0.0	Salgueiro	92.2	Belém do São Francisco
BR232	km 4.0	Recife	553.3	Terra Nova
BR316	km 0.0	Serrolândia	369.5	Petrolândia
BR407	km 0.0	Afrânio	130.1	Petrolina
BR408	km 20.3	Timbaúba	105.7	Jaboatão do Guararapes
BR423	km 18.3	São Caetano	196.2	Itaíba
BR424	km 0.0	Arcoverde	133.9	Correntes
BR428	km 0.0	Belém do S. Francisco	193.4	Petrolina

Para exemplificar a utilização dessas matrizes, foi escolhido o roteamento entre dois pontos, partindo da cidade de Goiana (km 0 da BR 101) à cidade do Recife (km 60 da BR 101). Este trecho é conhecido por ter um tráfego intenso, com velocidade baixa, devido ao número de veículos que trafegam em determinados horários, bem como, pela inúmera quantidade de acidentes graves, limitadores de velocidade (lombadas eletrônicas e físicas) e elevada quantidade de buracos. Estipularemos portanto, 20km/h em média, com fluxo de trânsito descontinuo. Normalmente, o fluxo contínuo ocorre quando não estão presentes os limitadores elencados anteriormente.

Figura 3.6: Rota: Goiana - Recife, a partir da Matriz de Gravidade 3D para 2^a-feira

		km 0 a 20																				
		km 21 a 40																				
		km 41 a 60																				
1 ^a Hora		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	SOMA-1
5		0	0	0	0	0	0	1,416	0	0	0	5,12	0	0	2,06	0	0	0	2,06	0	0	0
6		0	0	2,524	0,258	0	5,12	0	0	0	0	0	2,312	0	0	5,12	0	0	0	0	3,06	0
7		0	0	3,06	3,501	1,983	3,06	2,518	3,06	4,584	0	0	1,518	3,06	0	0	1,518	0	0	2,06	0	0
2 ^a Hora		21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	SOMA-2
6		0	0	1,423	0	0	0	0	0	0	1,518	0	0	0	0	0	0	0	1,983	0	2,518	7,442
7		0	0	0	0	0	0	0	0	0	0	1,524	0	0	0	0	0	0	0	4,12	0	5,644
8		2,518	0	5,12	3,94	0	0	2,524	0	0	0	4,12	0	3,06	1,416	0	2,518	6,001	0	3,097	3,06	37,374
3 ^a Hora		41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	SOMA-3
7		3,06	4,584	4,12	1,431	2,06	0	5,12	2,518	2,06	14,764	2,06	8,644	5,578	6,638	10,698	6,096	10,698	4,85	14,53	10,216	119,725
8		2,423	5,584	7,156	4,12	8,18	9,18	5,036	8,18	7,007	25,462	9,18	2,431	8,18	3,043	0	8,698	23,752	10,434	12,572	15,276	175,894
9		11,698	3,06	0	2,06	3,06	3,06	5,12	0	5,638	20,791	10,887	1,796	2,518	2,06	3,06	6,644	7,434	5,578	16,156	9,627	120,247

Fonte: autor

A Figura 3.6 pode ser analisada da seguinte forma: Três veículos partiram em uma segunda-feira de Goiana com destino ao Recife em horários diferentes. O Veículo-X partiu às 5h da manhã, Veículo-y às 6h e o Veículo-z partiu às 7h. Os primeiros, segundos e terceiros trechos de 20 km (km 0 - 20; km 21 a 40; km 41 a 60) foram obtidos a partir da Matriz de Gravidade 3D, calculada para uma segunda-feira. Os valores de preenchimento dessa matriz são fornecido pela aplicação da Equação 3.1. Na primeira linha do primeiro trecho (km 0 a 20) o Veículo-x obteve somatório 10.656, o Veículo-y 18,394 e o terceiro, Veículo-z, 29,922.

No cômputo geral, para cada veículo, em cada trecho de uma hora, foram obtidos os valores abaixo:

- Veículo-x: $10,656 + 7,442 + 119,725 = 137,823$
- Veículo-y: $18,394 + 5,644 + 175,894 = 199,932$
- Veículo-z: $29,922 + 37,374 + 120,247 = 187.543$

Segundo a aplicação da Equação 3.1, o Veículo-x chegará primeiro ao Recife, seguido pelo Veículo-z e, por último, o Veículo-y. Portanto, no exemplo acima, um condutor que parte de Goiana às 5h da manhã (2^a-feira) poderá chegar ao Recife em torno de três horas (velocidade média: 20km/h). Todavia, se por algum motivo esse condutor não conseguir partir às 5h da manhã, deverá fazê-lo às 7h, e não às 6h.

4

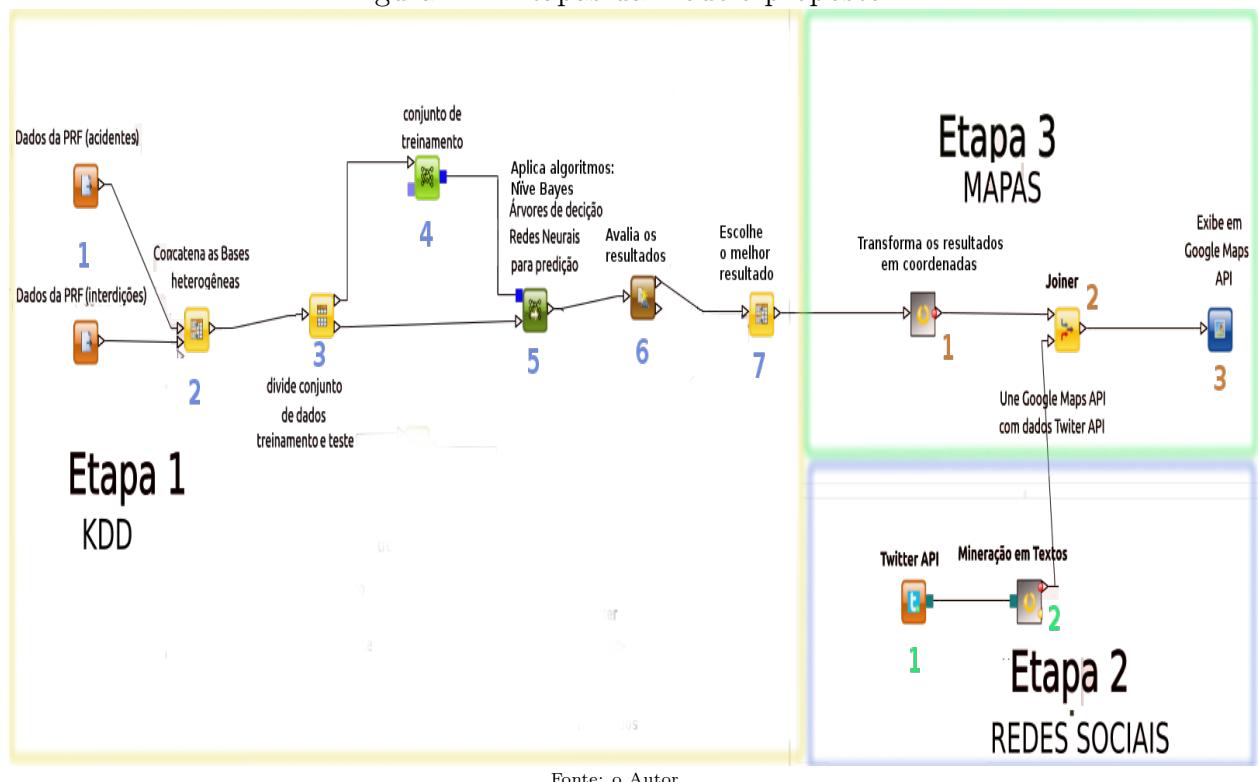
Simulação

4.1 Execução do modelo

Esse capítulo apresenta os resultados encontrados em cada etapa, no processo de desenvolvimento do modelo de predição de ocorrências nas principais BRs do estado de Pernambuco, para definição de melhores rotas e horários para os usuários dessas vias.

A figura a seguir resume as três etapas contempladas na proposição do modelo.

Figura 4.1: Etapas da modelo proposto



A Etapa 1 - KDD contempla a fase da coleta das bases de dados históricos, preparação dos dados, construção das variáveis do modelo preditivo, descoberta dos pontos críticos das rodovias (que serão utilizados na etapa de georreferenciamento);

1. O modelo preditivo integra as bases de dados da Polícia Rodoviária Federal – PRF. Algumas dessas informações também estão disponíveis em base de dados abertas, como sugere o Portal da Transparência, nos servidores da PRF, além de outras informações complementares;
2. Em seguida as bases heterogêneas são concatenados de acordo com a lógica do negócio que se propôs, como por exemplo algumas variáveis podem ser construídas a fim de permitir a concatenação dos dados;
3. Uma vez concatenados e limpos, dados tratamento, retirados os “missing data”, os dados serão divididos em conjunto de treinamento e de testes;
4. Uma parte do conjunto de dados, o conjunto de treinamento serve para treinar os diversos algoritmos de IA já previamente escolhidos;
5. Nesta fase os algoritmos de IA são utilizados a fim de produzirem resultados esperados pelas técnicas escolhidas. Pode haver recorrência nesta fase caso os algoritmos não tenham produzidos os resultados esperados. A conclusão dessa etapa ocorre com a Mineração dos dados e a extração de conhecimento.
6. A fase de avaliação dos resultados ocorre pela análise das diversas métricas definidas como: Instâncias corretamente classificadas, Áreas sob a curva ROC (AUC) e Matriz de Confusão.
7. A escolha do melhor resultado se dá pelo maior valor da área encontrada sob a curva ROC concomitantemente com o maior número de instâncias classificadas corretamente e o menor número de instâncias classificadas incorretamente com o menor erro médio absoluto.

A Etapa 2 - Redes Sociais contempla:

1. Um módulo dinâmico onde são capturados “feeds” da rede social Twitter. Essa técnica faz um arco cibرنético mantendo o utilizador atualizado com as informações recentes.
2. Após a captação dos dados do Twitter é feita a Mineração nos textos para localizar informações que permitam antever alguma paralização futura nas rodovias.

A Etapa 3 - Mapas é última etapa, consiste em um módulo com as seguintes características:

1. Conversão dos resultados encontrados na Etapa 1, localizações geográficas dos pontos críticos da malha viária, indicadas pelo Km, são agrupadas formando ”clusters“ de dados exibidos em mapas vetoriais;
2. Um ambiente de simulação interativa que utiliza uma plataforma baseada na API do Google Maps.
3. Exibe informações vindas do twitter sobre ocorrências na rodovia, encontrando sua geolocalização a ser transformado em marcos “milestone” para representação sobre mapas de bases vetoriais.

4.2 A construção do Modelo preditivo

O modelo preditivo foi construído utilizando bases de dados históricos da PRF (de acidentes e de paralisações, por exemplo, protestos) entre Janeiro de 2007 a Dezembro 2015. Cada ano correspondia a uma base de dados independentes, tendo sido integradas, formando uma base única com aproximadamente 85 mil registros.

4.2.1 Aplicação do CRISP-DM

O CRISP-DM nesta pesquisa ajudou a guiar as escolhas nos momentos em que os resultados pareciam não fazer sentido. Todavia, por ser um processo recursivo, o retorno aos fundamentos dessa metodologia prevê que haja ajustes necessários, a fim de se atingir os objetivos da proposta.

A proposta metodológica delineada para esta pesquisa contemplou todas as fases do KDD, conforme descrito a seguir.

Fases da Mineração ao KDD

Seleção: Nesta etapa foram coletadas as informações provenientes das bases de dados da Policia Rodoviária Federal de Pernambuco entre 2007 e 2015. Segundo informações da própria PRF/PE, apenas a partir de 2007 esses dados passaram a ser armazenados eletronicamente. A PRF/PE dispõe em banco de dados relacionais alguns desses dados na Internet, contudo no artigo “Uma análise da qualidade dos dados relativos aos boletins de ocorrências das rodovias federais para o processo de Mineração de Dados”, COSTA, BERNARDINI, LIMA (85) destacam a não padronização e não aceitação dos dados pela comunidade internacional. EAVES, D. (86) sugere que os dados sejam disponibilizados na maneira como foram coletados.

A primeira base de dados coletada diretamente dos servidores da PRF continha relatório de acidentes e a segunda a de interdições. A partir dos dados capturados na base da PRF utilizamos como variáveis de entrada:

- Condição da Pista: Seca, Com buracos, Molhada, Em obras, Com material granulado, Oleosa, Enlameada, Com gelo, Outras
- Restrição de visibilidade: Inexistente, Veículo Estacionado, Poeira/Fumaça/neblina, Vegetação, Ofuscamento, Cartazes/faixas, Placas
- Traçado da via: Reta, Curva, Cruzamento, Defeito
- Tipo de veículo: Automóvel, Caminhonete, Motocicletas, Caminhão, Caminhão-trator, Bicicleta, Ônibus, Motoneta, Micro-ônibus, Trator de rodas, Carroça, Caminhão-Tanque, Semi-Reboque, Utilitário, Ciclomotor, Charrete, Carro-de-mão, Quadriciclo, Trator misto, Reboque, Trator de esteiras, Não informado, Não se aplica, Não identificado

Preprocessamento: Nesta fase foram retiradas as variáveis que continham inconsistência e “missing data”, como, por exemplo, informações acerca de latitude e longitude. Cabe destacar que a base de dados, como um todo, apresentava sérias inconsistências, uma vez que, por exemplo, um mesmo acidente, quando envolvia dois ou mais veículos, era lançado na base duas ou mais vezes, em função da quantidade de veículos envolvidos.

Foram eliminadas variáveis em duplicidade (i.e. as variáveis Mês, Ano que apareciam separadamente, já haviam sido contempladas na variável Data.).

Transformação: Foram criadas as variáveis “Tipo de paralisação”, contemplando acidentes sem mortos e com, no máximo, dois veículos envolvidos; “Dias da semana” (domingo, segunda-feira,...sábado); “Ajuste de horas” (i.e. 17h58, 17h59, 18h, 18h01, 18h02, arredondadas para 18h); “Ajuste de Km” (seguiu a mesma lógica do ajuste de horas).

Mineração de dados: O algoritmo escolhido para essa fase da pesquisa foi Árvore de Decisão, que possibilita uma interpretação imediata e de fácil compreensão. Como ferramentas, foram escolhidas o Knime (87), R (88) e Weka (89), com objetivo de estabelecer uma comparação entre eles, cuja intenção era produzir um classificador mais robusto. Nessa direção, a técnica Ensamble de classificadores (90) estabelece que a combinação de um ou mais classificadores iguais, ou mais de um classificador diferente, aumenta a precisão.

Tanto na ferramenta Knime como Weka o algoritmo de árvore de decisão é chamado de J48, uma vez que se trata da implementação Java do algoritmo C4.5 no R. A biblioteca “rpart” implementa esse algoritmo.

Para escolha das variáveis de entrada foi calculada a correlação linear entre todas as variáveis. Entre as variáveis BR e Delegacia (variável que agrupa municípios) obteve-se correlação linear de 0,653. Entre Tipo de Acidente e Traçado via a correlação foi baixa, apenas 0,14. Variáveis com correlação linear abaixo disso foram descartadas.

Outra métrica para a escolha de variáveis de entrada foi a entropia, que é um elemento considerado importante pela literatura (18). Nesse caso, variáveis que seriam desconsideradas em virtude da baixa correlação linear, foram reconsideradas por conta da entropia.

Interpretação/Avaliação: Produção de árvores de decisão a partir do estabelecimento de diferentes nós-raízes, definidos em virtude da correlação linear e/ou da entropia.

4.2.2 Dados encontrados antes da Mineração

Os dados revelaram que a grande maioria dos acidentes ocorre com pista seca, sem restrição de visibilidade. A cor vermelha remete aos trechos BR 101 em que ocorrem mais acidentes. A cor azul diz respeito à menor frequência de acidentes.

O gráfico 4.2(1), 4.2(2) e 4.3 contêm dados da BR 101, uma das mais importantes para o nordeste brasileiro, uma vez que atravessa a maioria dos estados dessa região, nas localidades mais densamente povoadas. Em virtude disso, seu tráfego intenso. O gráfico 4.2(1) representa os acidentes que ocorreram a cada hora (abscissa) em cada Km (ordenada) nos últimos nove anos. O gráfico 4.2(2) corresponde à frequência do local onde ocorreram esses acidentes. É possível perceber que há determinados locais na rodovia onde se concentram os acidentes. O terceiro gráfico, tipo ‘boxplot’, exibe a concentração das ocorrências em torno da mediana dessa localidade (Km). Especulou-se, a priori, que a variável “traçado da rodovia” ou que as condições climáticas poderiam ser de grande influência na ocorrência de acidentes, contudo mais adiante descobrimos outros condicionantes que influenciam mais fortemente esses acontecimentos. É possível perceber no gráfico 4.2(1) e no 4.3 que especialmente em determinados locais (Km) – por exemplo na BR 101, entre os Km 40 e 100 – os acidentes ocorrem desde as 05h da manhã até as 23h.

Figura 4.2: BR 101: Hora do acidente (1) Concentração em torno da hora (2)

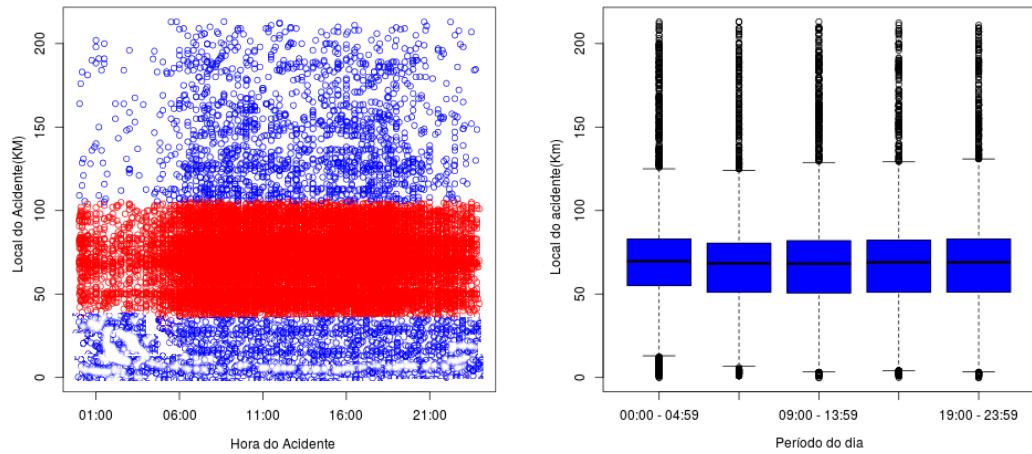
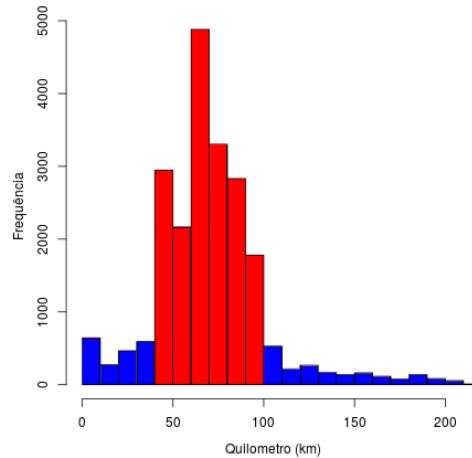


Figura 4.3: Frequência



O gráfico 4.4(1), 4.4(2) e 4.5 apresentam dados da BR 104, que atravessa seis municípios de Pernambuco, dentre eles Caruaru, que é responsável por uma das maiores frotas de veículos do interior e por onde passam cerca de 50 mil veículos por dia. A gráfico 4.4(1) representa, nos últimos nove anos, os acidentes que ocorreram a cada hora (abcissa) em cada Km (ordenada). O gráfico 4.4(2) (boxplot) apresenta as ocorrências em torno da mediana dessa localidade (Km). O terceiro gráfico (4.5) corresponde à frequência do local onde ocorreram esses acidentes. Percebe-se que em torno do Km 60 concentra-se o maior número de ocorrências. No gráfico 1 são identificados padrões, por exemplo no Km 60 ocorrem acidentes que se estendem das 04h às 23h.

Figura 4.4: BR 104: Hora do acidente (1) Concentração em torno da hora (2)

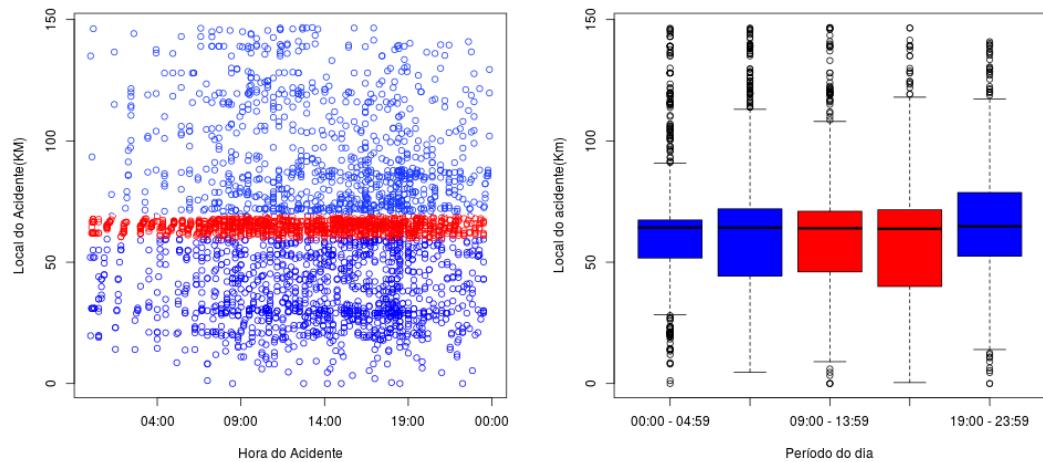
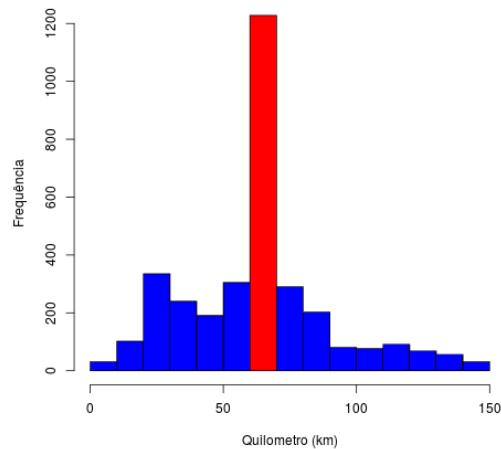


Figura 4.5: Frequência



O gráfico da Figura 4.6 (1) e 4.6 (2) representa a BR 110. Esta é a menor rodovia federal de Pernambuco. Esta rodovia inicia-se e termina em Ibimirim. Esta rodovia também pode ser considerada como a que possui o menor número de sinistro, consequentemente a mais segura. O local com maior frequência de acidentes se encontra no entorno do quilômetro 110, seguido pelos quilômetros 130 a 150.

Figura 4.6: BR 110: Hora do acidente (1) Concentração em torno da hora (2)

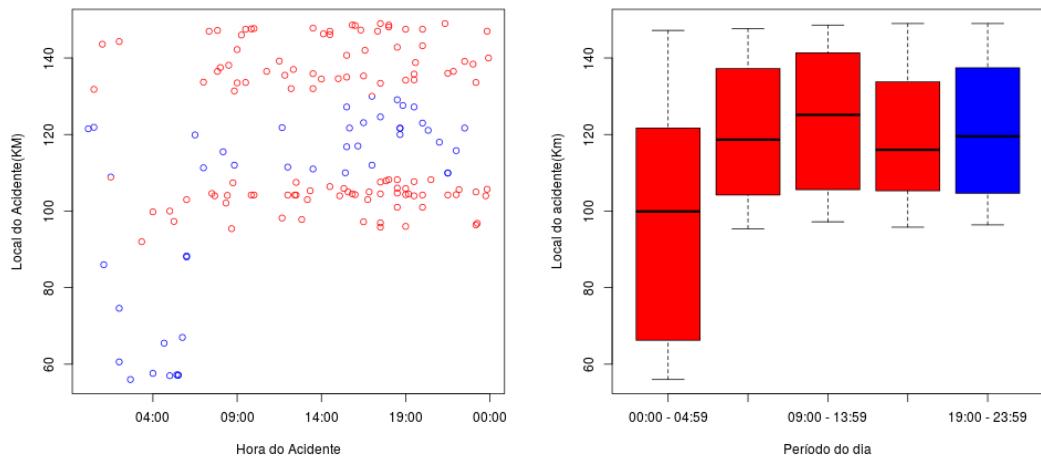
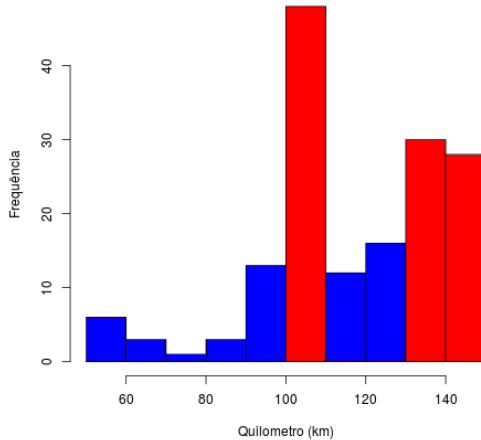


Figura 4.7: Frequência



O gráfico 4.8(1), 4.8(2) e 4.9 apresentam dados da BR 116, que percorre os estados do Brasil que vão desde o Rio Grande do Sul até o Ceará. O gráfico 4.8(1) mostra os acidentes que ocorreram a cada hora (abcissa) em cada Km (ordenada), também nos últimos nove anos. O gráfico 4.8(2) corresponde à frequência do local onde esses acidentes aconteceram. Há determinados locais na rodovia em que a maioria dos acidentes se concentram. O gráfico tipo ‘boxplot’ exibe a concentração das ocorrências em torno da mediana dessa localidade (Km). Inicialmente acreditou-se que a variável “traçado da rodovia” ou que as condições climáticas poderiam ter grande influência na causa dos acidentes. Entretanto, mais adiante foram descobertos outros condicionantes dessas ocorrências. É possível perceber no gráfico 4.9, que os acidentes em torno do Km 30 ocorrem com mais frequência a partir da 04h da manhã, estendendo-se até próximo às 22h.

Figura 4.8: BR: 116 Hora do acidente (1) Concentração em torno da hora (2)

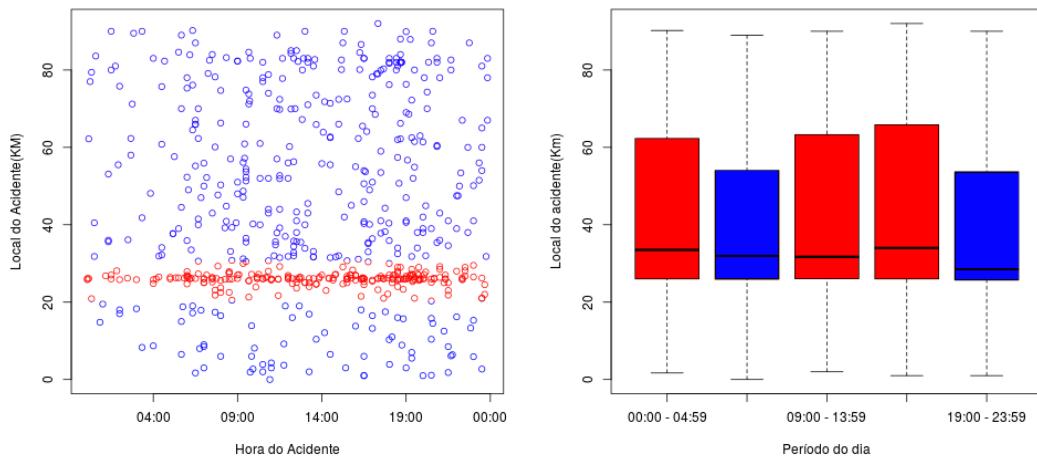
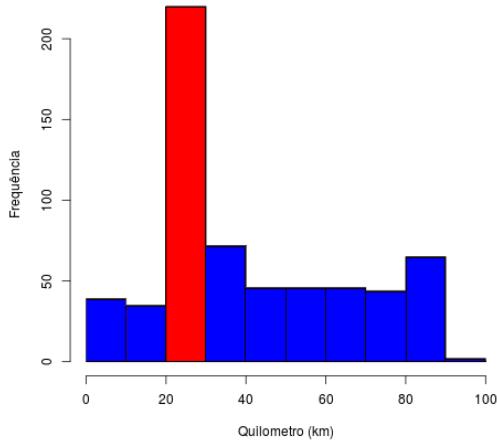


Figura 4.9: Frequência



O gráfico 4.10(1), 4.10(2) e 4.11 trazem dados da BR 232, cuja importância se destaca pelo fato de atravessar todo o estado de Pernambuco, de leste a oeste. O primeiro gráfico apresenta os acidentes dos últimos nove anos, em cada hora (abcissa) e Km (ordenada). O gráfico 4.10(2) corresponde ao local onde aconteceram esses acidentes. É possível perceber no gráfico 4.11, que nessa BR há um número maior de acidentes nos Km 0, 90, 110, 260, 410 e 500, desde a 00h até as 23h. O terceiro gráfico ('boxplot') exibe a concentração das ocorrências em torno da mediana dessa localidade (Km). Especulou-se a priori que a variável “traçado da rodovia” ou que as condições climáticas eram as principais responsáveis pelo grande número de acidentes.

Figura 4.10: BR 232: Hora do acidente (1) Concentração em torno da hora (2)

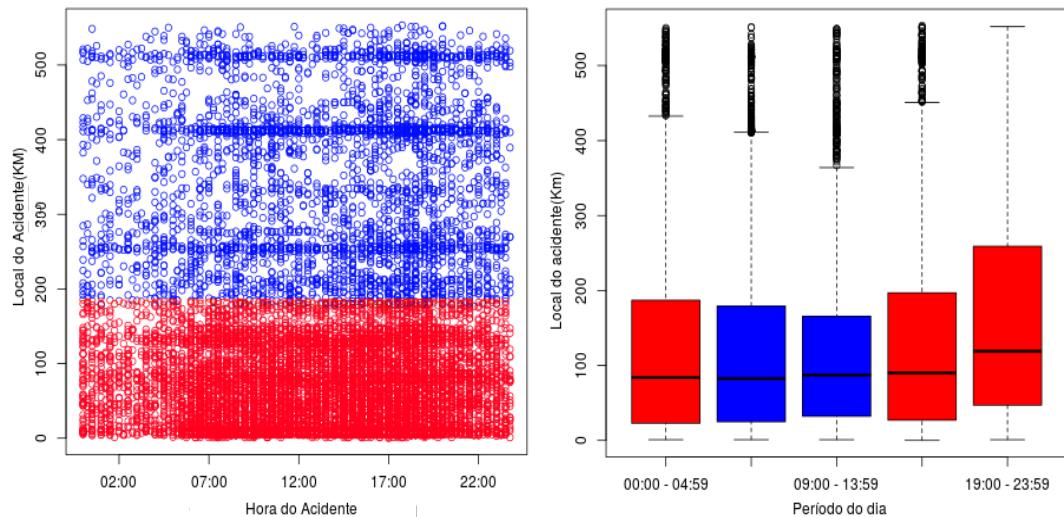
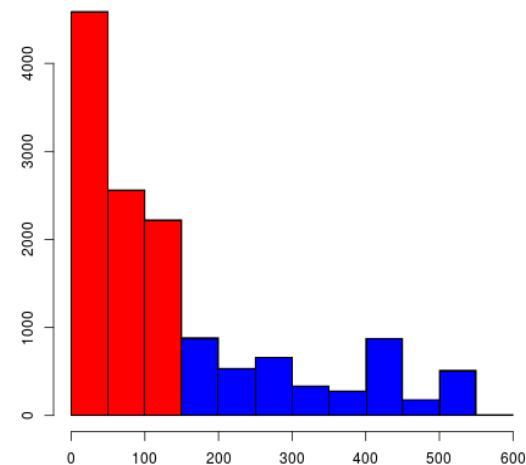


Figura 4.11: Frequência



O gráfico 4.12(1), 4.12(2) e 4.13 trazem dados da BR 316. Esta BR é uma das menores do Estado de Pernambuco, de oeste a leste. O primeiro gráfico apresenta uma área em branco no centro, os acidentes se concentram nas extremidades, no entorno do km 90 e 300 (da abscissa) e acontecem com maior frequência em torno das 17h00 (ordenada).

Figura 4.12: BR 316: Hora do acidente (1) Concentração em torno da hora (2)

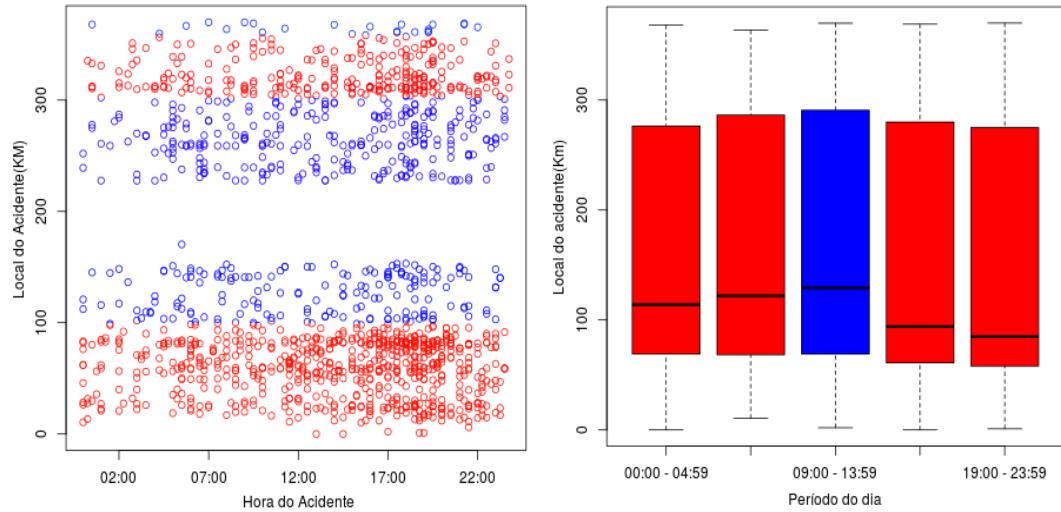
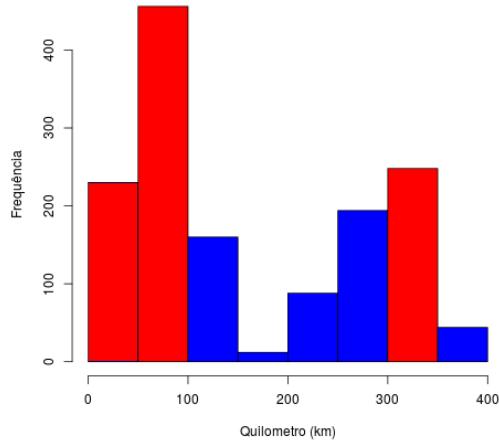


Figura 4.13: Frequência



Na BR 407 os acidentes se concentram na altura do Km 130, representados nos gráficos 4.14 (1) e 4.14 (2). Esta BR situa-se no extremo oeste do Estado, ligando a cidade de Afrânio a Petrolina. A peculiaridade desse trecho (km 130) é verificada na descida da ponte que atravessa o rio São Francisco.

Figura 4.14: BR 407: Hora do acidente (1) Concentração em torno da hora (2)

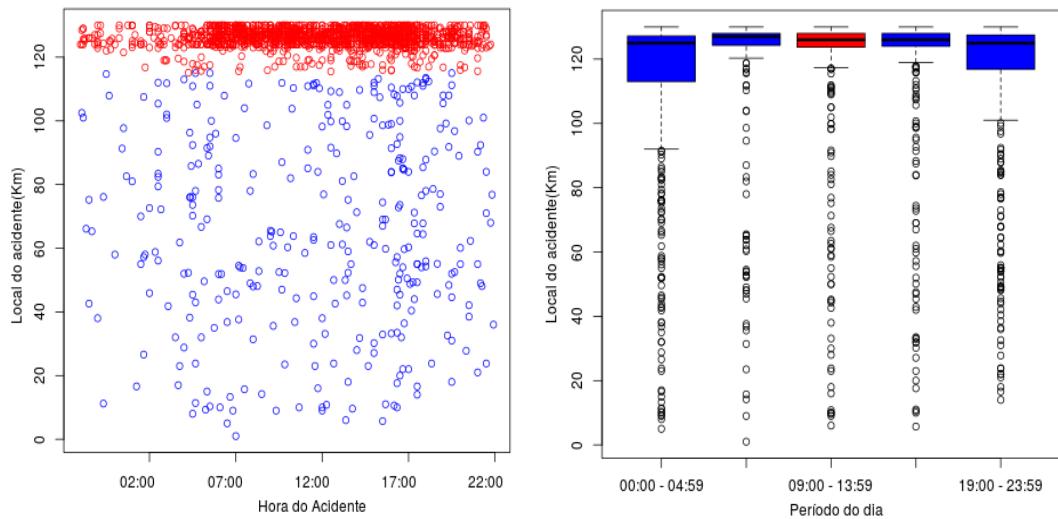
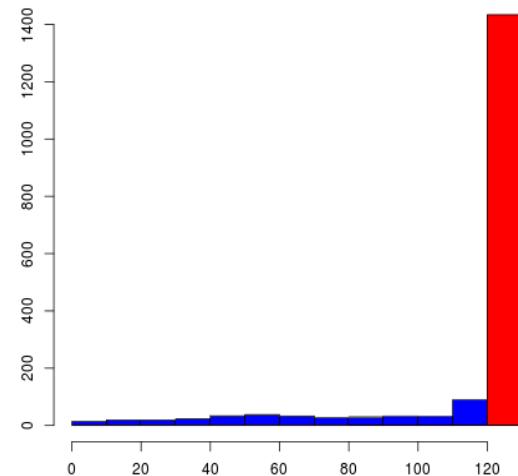


Figura 4.15: Frequência



O gráfico 4.16 (1), 4.16 (2) e 4.17 representam a BR 408. Em Pernambuco esta BR liga a cidade de Timbaúba a Jaboatão dos Guararapes. Esta BR tem importância a medida integra o setor industrial de autopeças no norte de Pernambuco proporcionado pela fábrica da Fiat (FCA) ligando-se ao polo industrial do sul do Estado.

Figura 4.16: BR 408: Hora do acidente (1) Concentração em torno da hora (2)

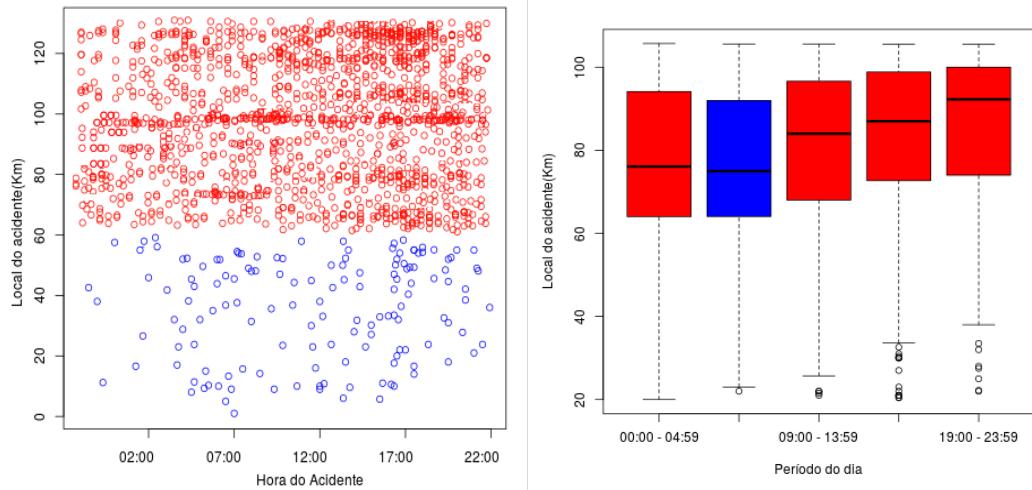
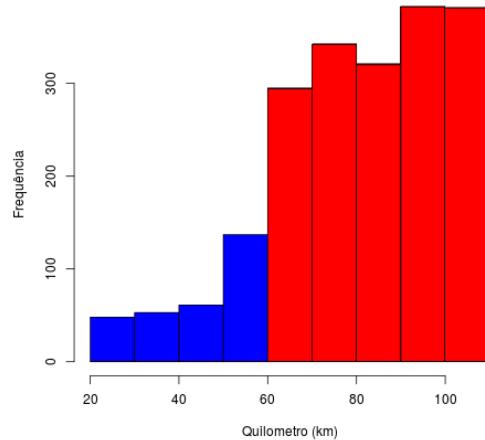


Figura 4.17: Frequência



A BR 423 4.18 (1), 4.18 (2) e 4.19, esta BR é a mais estratégica rodovia do Nordeste por ser o elo entre a BR 232 (a mais extensa BR do Estado de Pernambuco) à principal hidroelétrica da região; Paulo Afonso. Esta rodovia inicia-se no município de São Caetano indo até a fronteira com o estado de Alagoas no município de Itaíba, com 196,2 km é a 4^a mais extensão do Estado.

Figura 4.18: BR 423: Hora do acidente (1) Concentração em torno da hora (2)

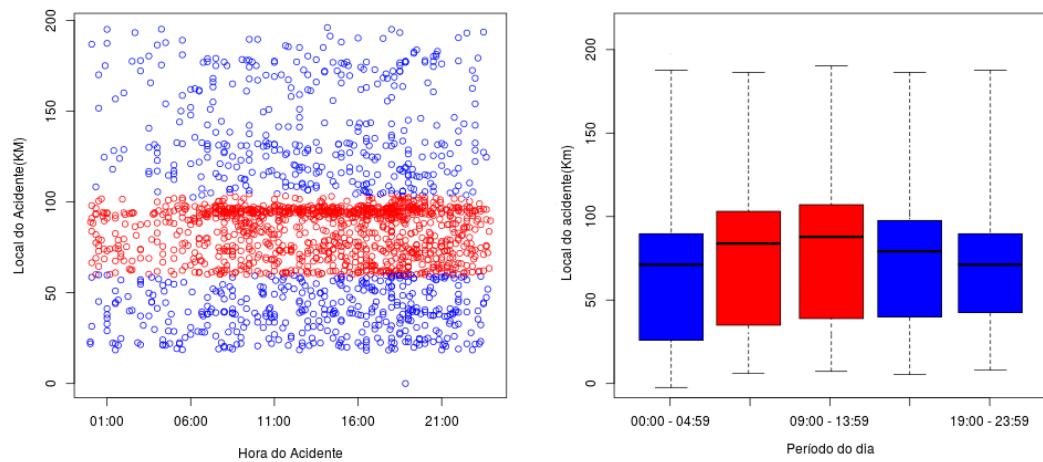
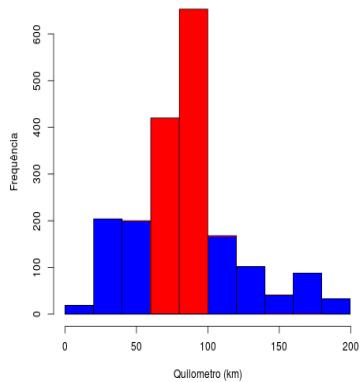


Figura 4.19: Frequência



A rodovia BR 424 inicia em Arcoverde e termina em Correntes. Com extensão de 133,9 km tem características semelhantes à BR 423 (vide gráficos 4.20 (1 e 2) e 4.21), contudo esta rodovia tem importância secundária interligando o importante município do Agreste, Garanhuns, à BR 424.

Figura 4.20: BR 424: Hora do acidente (1) Concentração em torno da hora (2)

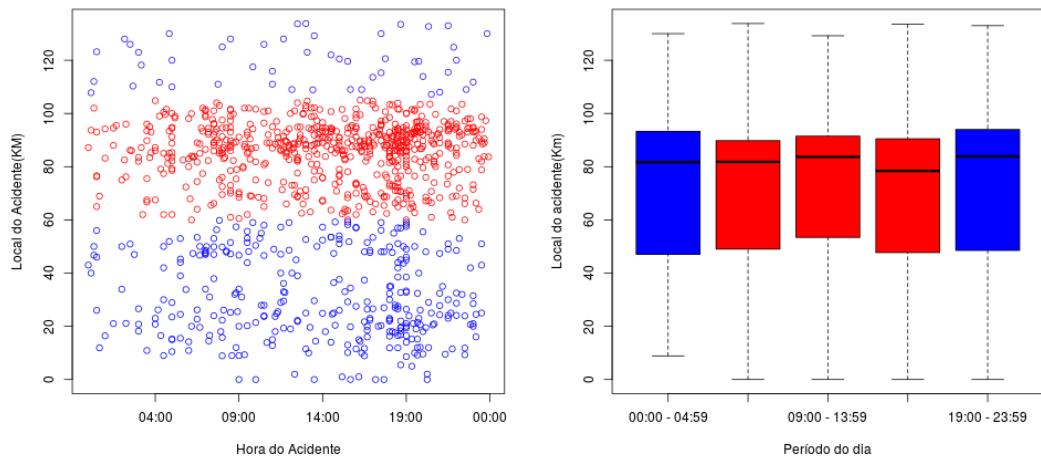
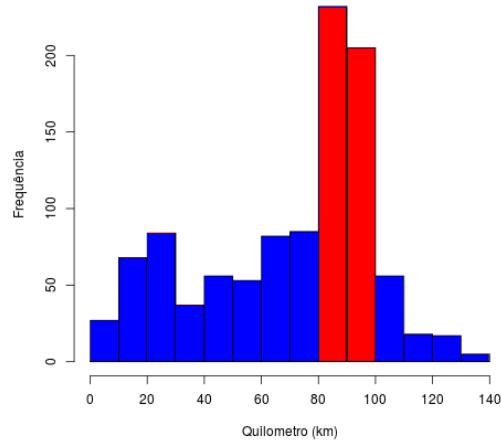


Figura 4.21: Frequência



A BR 428 é a 11^a rodovia federal estudada nesta pesquisa, ela conecta os municípios de Belém do São Francisco ao município de Petrolina, o maior do Sertão pernambucano. Esta rodovia contorna o Rio São Francisco pela margem norte. Nesta rodovia o ponto crítico situa-se no entorno do km 180 sendo que, cerca das 19h00 o horário que ocorrem o maior número de acidentes (vide gfáfico 4.22).

Figura 4.22: BR 428: Hora do acidente (1) Concentração em torno da hora (2)

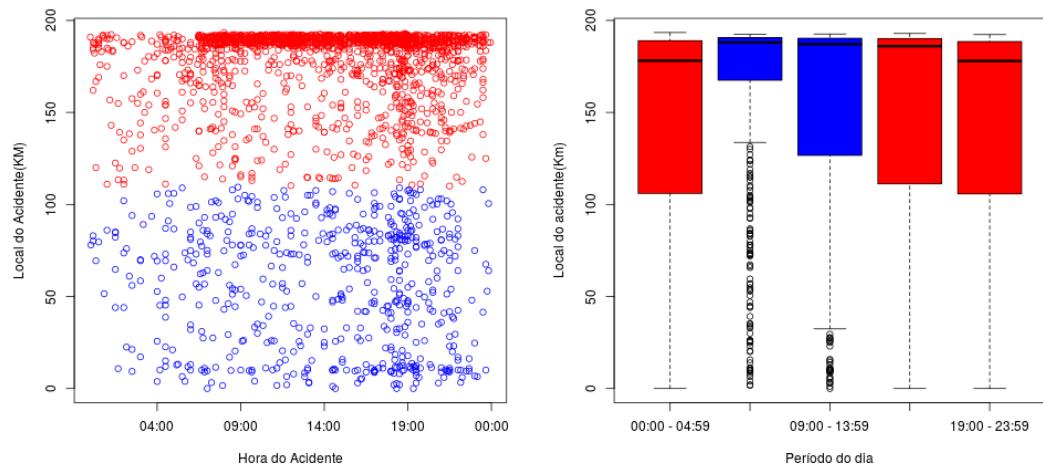
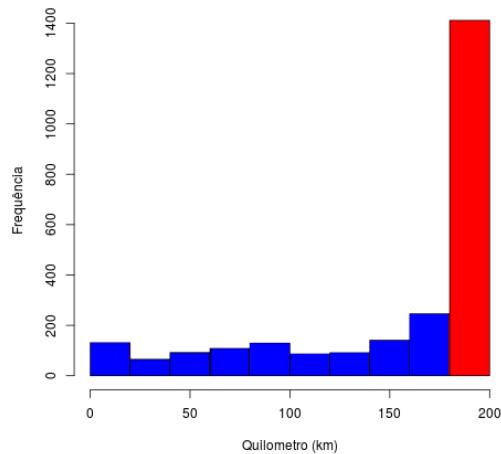


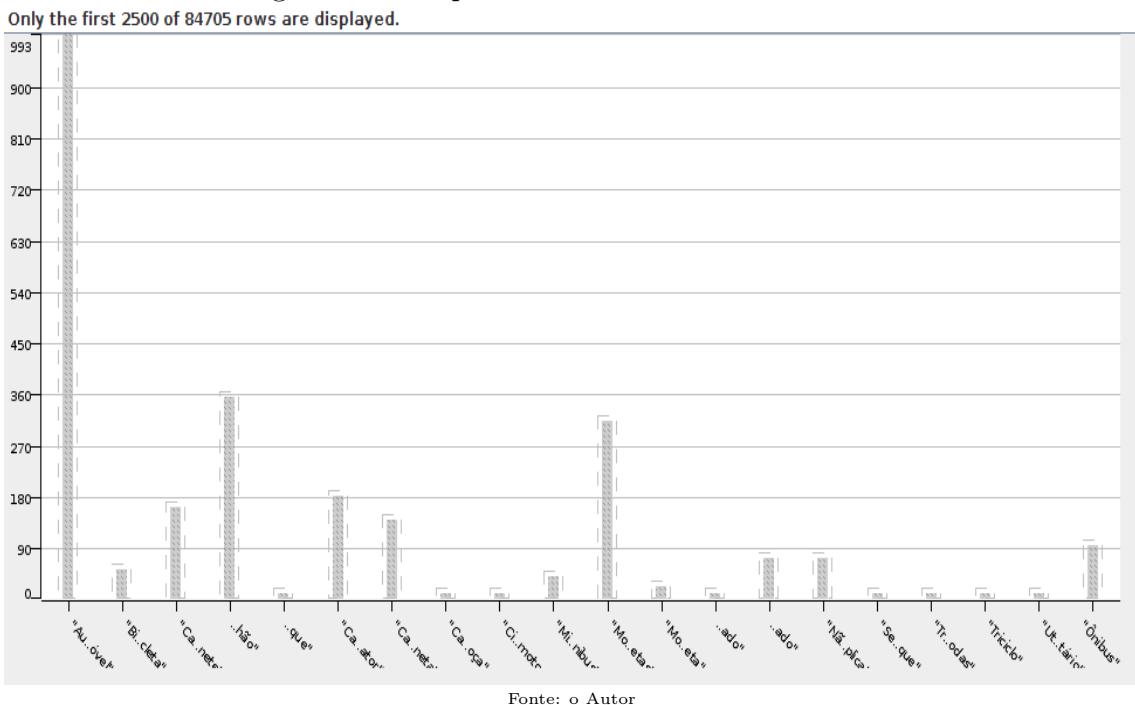
Figura 4.23: Frequência



O número elevado de acidentes que tem por causa o automóvel de passeio que trafega por via retilínea, provavelmente condutores comuns, não profissionalizados. O Caminhão é o segundo veículo que mais se envolve em acidentes, seguido das motonetas (motocicletas com potência limitada).

Os dados do gráfico 4.25 demonstram que o traçado da via (em linha reta) não parece influenciar a causa elevada no número de acidentes, pois, a maioria dos acidentes ocorre em pista retilínea sugerindo assim que o condutor é o principal responsável pelas ocorrências nas BRs desse tipo de sinistro, com isso nosso foco se direciona para analisar e tentar antever o comportamento do condutor, perante às condições externa, para além das condições da rodovia.

Figura 4.24: Tipo de Veículo X Num. Acidentes



4.2.3 Dados encontrados após a Mineração

Os resultados dos classificadores serão demonstrados a seguir.

As variáveis “Tipo de Acidente”, “Gravidade” e “BRajustada” foram escolhidas pelas características de ganho de informação, dado pelo cálculo da entropia. A variável “BRajustada” significa que essa variável foi transformada de dado numérico para categórico. A literatura (18) aconselha que os nós da raiz dos classificadores, em especial Árvores de Decisão, sejam aqueles que apresentam maior entropia, como a variável “Tipo de Acidente”. A seguir apresenta-se a métrica para avaliar um classificador, também conhecida como acurácia.

- TP: True Positive;
- FP: False Positive;
- Prec.: Precision = $TP / (TP + FP)$;
- Recall = $TP / (TP + FN)$;
- F-Me: F-measure ou f-score = $2 * Precision * Recall / (Precision + Recall)$;
- AUC: Area Under Curve (Roc);

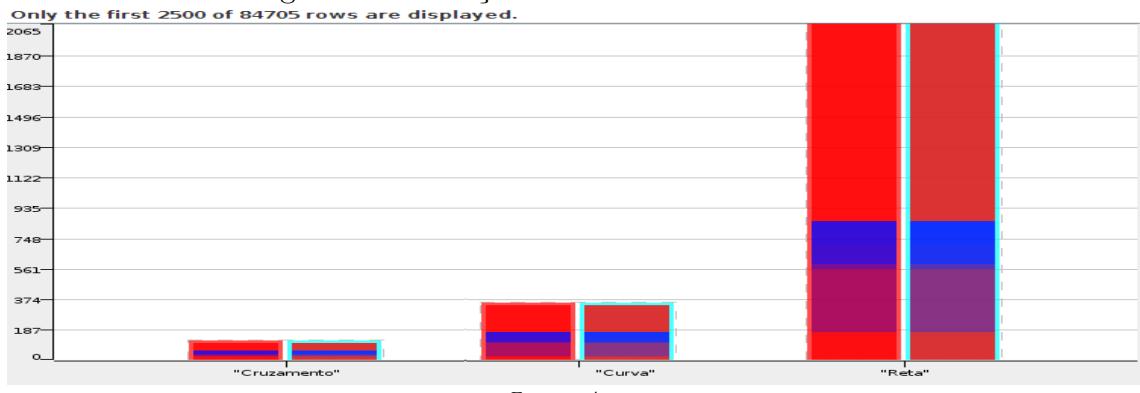
4.2.4 Métrica dos classificadores

- (i) Variável: Tipo de Acidente (Entropia: 3.0686)

Descrição	Valores	Percentual
Instâncias Corretamente Classificadas	7987	47.6324%
Instâncias Incorretamente Classificadas	8781	52.3676%
Erro médio absoluto	0.0786	—
Erro médio quadrático	0.2083	—

Fonte: o Autor

Figura 4.25: Traçado da via X Num. Acidentes



Fonte: o Autor

Tabela 4.1: Detalhe da acurácia para classe Tipo Acidente

TP	FP	Prec.	Recall	F-Me.	AUC	Classe
0.337	0.059	0.372	0.337	0.354	0.738	Colisão transversal
0.026	0.012	0.066	0.026	0.038	0.684	Colisão com objeto fixo
0.925	0.003	0.920	0.925	0.923	0.980	Atropelamento de pessoa
0.463	0.157	0.448	0.463	0.455	0.731	Colisão lateral
0.682	0.259	0.545	0.682	0.606	0.773	Colisão traseira
0.485	0.024	0.409	0.485	0.443	0.893	Queda de Moto/bicicleta
0.322	0.002	0.528	0.322	0.400	0.744	Colisão com bicicleta
0.122	0.026	0.229	0.122	0.159	0.786	Capotamento
0.890	0.014	0.655	0.890	0.755	0.954	Atropelamento de animal
0.048	0.007	0.243	0.048	0.081	0.729	Colisão frontal
0.440	0.089	0.366	0.440	0.399	0.792	Saída de Pista
0.000	0.000	0.000	0.000	0.000	0.658	Colisão c/ objeto móvel
0.096	0.006	0.292	0.096	0.144	0.774	Tombamento
0.000	0.000	0.000	0.000	0.000	0.616	Derramamento de Carga
0.041	0.000	0.400	0.041	0.074	0.627	Danos Eventuais
0.000	0.000	0.000	0.000	0.000	0.733	Incêndio

Fonte: o Autor

Tabela 4.2: Matriz de confusão para a variável Tipo de acidente

a	b	c	d	e	f	g	h	Classificadas
527	7	2	385	483	46	2	24	Colisão transversal
16	14	0	69	154	15	0	47	Colisão com objeto fixo
8	0	483	16	14	0	0	0	Atropelamento de pessoa
336	30	8	1674	1217	102	8	48	Colisão lateral
250	51	9	835	3573	105	11	59	Colisão traseira
44	4	1	74	120	266	2	0	Queda de Moto/bicicleta
8	0	0	22	38	3	38	1	Colisão com bicicleta
28	34	5	85	236	1	2	120	Capotamento
—	—	—	—	—	—	—	—	—

Fonte: o Autor

Os valores restantes foram omitidos por não representarem uma amostra adequada,

pois a acurácia foi consideravelmente baixa, por exemplo o classificado não acerta na maioria das vezes qual a classe deve ser escolhida para todas os atributos. As variáveis de classe são as mesmas da tabela anterior.

(ii) Variável: Gravidade (Entropia: 0,9997)

Descrição	Valores	Percentual
Instâncias Corretamente Classificadas	12110	72.2209%
Instâncias Incorretamente Classificadas	4658	27.7791%
Erro médio absoluto	0.3816	—
Erro médio quadrático	0.4368	—

Fonte: o Autor

Tabela 4.3: Detalhe da acurácia para classe Gravidade

TP	FP	Prec.	Recall	F-Me.	AUC	Classe
0.907	0.608	0.727	0.907	0.807	0.721	S
0.392	0.093	0.703	0.392	0.504	0.721	N

Fonte: o Autor

Tabela 4.4: Matriz de confusão para a variável Gravidade

a	b	Classificadas
9747	996	a = S
3662	2363	b = N

Fonte: o Autor

(iii) Variável: BRajustada (Entropia: 2,4128)

Descrição	Valores	Percentual
Instâncias Corretamente Classificadas	13507	80.5522%
Instâncias Incorretamente Classificadas	3261	19.4478%
Erro médio absoluto	0.0469	—
Erro médio quadrático	0.1656	—

Fonte: o Autor

Tabela 4.5: Detalhe da acurácia para classe BR

TP	FP	Prec.	Recall	F-Me.	AUC	Classe
0.902	0.178	0.812	0.902	0.854	0.917	BR101
0.873	0.003	0.957	0.873	0.913	0.992	BR104
0.213	0.001	0.357	0.213	0.267	0.816	BR110
0.457	0.003	0.669	0.457	0.543	0.961	BR116
0.760	0.068	0.787	0.760	0.774	0.919	BR232
0.893	0.006	0.800	0.893	0.844	0.985	BR316
0.951	0.007	0.857	0.951	0.901	0.995	BR428
0.761	0.012	0.693	0.761	0.725	0.974	BR423
0.461	0.006	0.599	0.461	0.521	0.957	BR424
0.814	0.001	0.961	0.814	0.881	0.999	BR407
0.158	0.010	0.460	0.158	0.235	0.781	BR408

Fonte: o Autor

Tabela 4.6: Matriz de confusão para a variável BRajustada

a	b	c	d	e	f	g	h	Classificadas
6960	0	0	625	0	0	0	0	BR101
0	1071	0	156	0	0	0	0	BR104
0	0	0	625	0	0	26	11	BR110
0	0	85	0	90	11	0	0	BR116
970	9	0	3185	1	0	1	0	BR232
0	0	27	11	377	7	0	0	BR316
0	0	0	0	0	95	0	0	BR407
643	0	0	66	0	0	0	0	BR408
0	39	0	0	0	0	449	92	BR423
0	0	0	625	0	0	172	154	BR424
0	0	15	0	3	675	0	0	BR428

Fonte: o Autor

A área sob a curva ROC, AUC (Area Under Curve) mede a relação de verdadeiros positivos contra os falsos positivos. Quanto maior a área da curva tanto melhor será o classificador. Portanto, um número de verdadeiros positivos acima de 80% (vide coluna AUC da Tabela 4.5), o número de falsos positivos próximo a 0% (vide Tabela 4.5 coluna FP (Falsos Positivos)) traduzem uma área sob a curva ROC (AUC) que dá maior confiabilidade aos testes.

A variável “BRajustada” não teve o maior coeficiente de entropia encontrado, contudo esta variável apresentou índices de classificação das instâncias corretas acima dos 90%, somente em dois casos ficou abaixo desse patamar que foram para a BR 110 (área = 0,816) e para a BR 408 (área = 0,781) e, o menor índice de classificação incorreta dentre os três classificadores utilizados, podemos constatar isso da seguinte forma, na primeira linha da Matriz de Confusão demonstrada na Tabela 4.6 temos:

Tabela 4.7: Linhda da Matriz de confusão correspondente a BR 101

a	b	c	d	e	f	g	h	Classificadores
6960	0	0	625	0	0	0	0	BR101

Fonte: o Autor

A Tabela 4.7 foi retirada da Tabela 4.6 correspondente à linha da BR 101. O classificador Árvore de Decisão encontrou 6960 (classe a) de instâncias corretamente e 625 (classe d) de instâncias erradas, as entradas restantes estão em zero, significa que o classificador não "errou"em outras classes. A área sob a curva ROC correspondente a BR 101 foi de 0,917 ou 91,7%.

A segunda linha da Matriz de Confusão da Tabela 4.6 está demonstrada na Tabela 4.8.

Tabela 4.8: Matriz de confusão para a variável BRajustada

a	b	c	d	e	f	g	h	Classificadores
0	1071	0	156	0	0	0	0	BR104

Fonte: o Autor

Como se pode constatar na Tabela 4.8 (BR 104) o classificador encontrou 1071 instâncias corretas (classe b) e 156 erradas (classe d), com área sob a curva ROC de 0,992 ou 99,2%.

A 4^a linha (de baixo para cima) da Tabela 4.5 correspondente à BR 408 é demonstrada na Tabela 4.9

Tabela 4.9: Matriz de confusão para a variável BRajustada

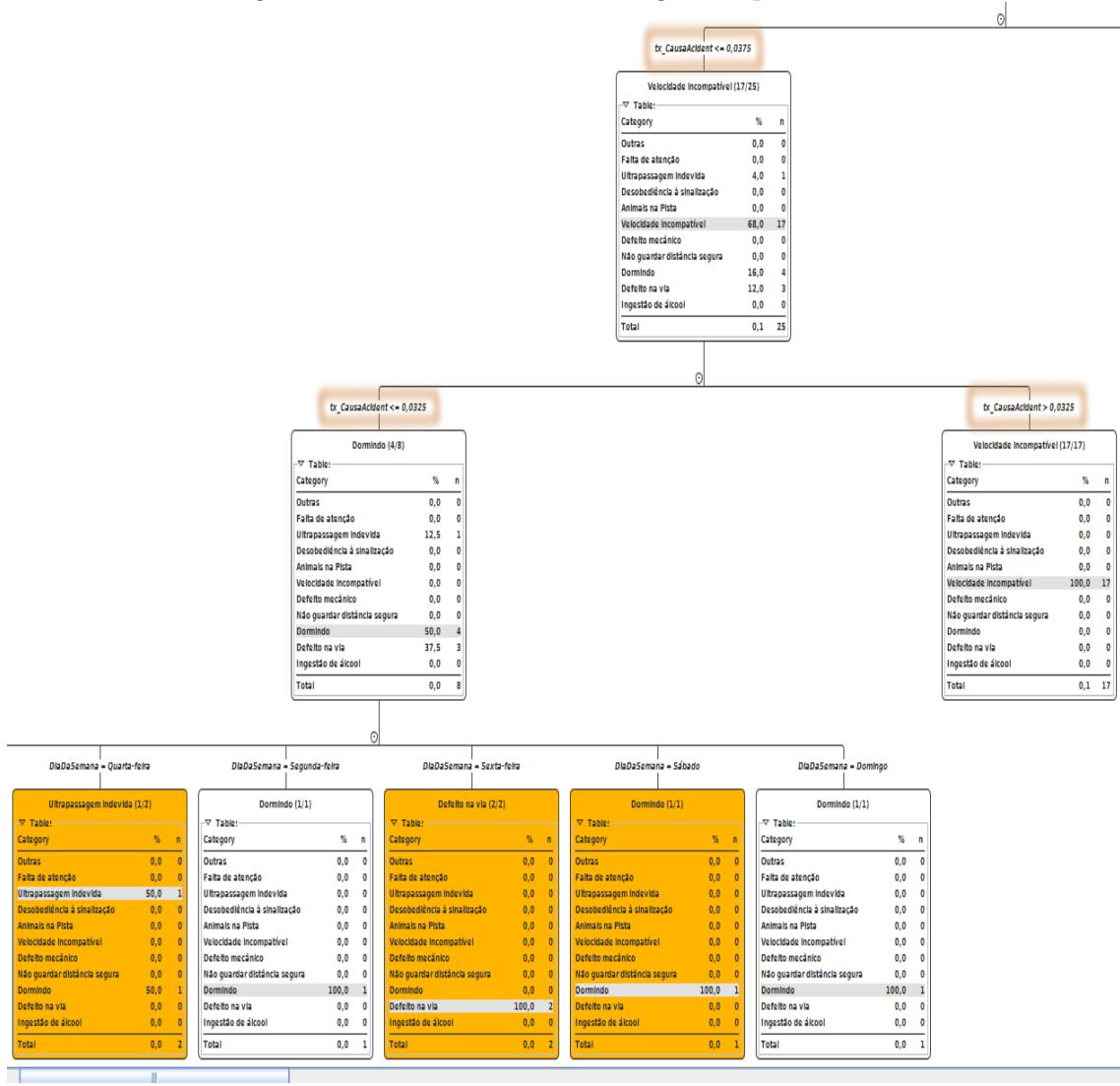
a	b	c	d	e	f	g	h	Classificadores
643	0	0	66	0	0	0	0	BR408

Fonte: o Autor

O classificador Árvore de Decisão encontrou a menor área sob a curva ROC para essa BR, (vide última linha da Tabela 4.5) na coluna AUC (AUC = 0,781) e o número de falsos positivos (FP) foi 0,10 (10%), acreditamos que devido ao baixo número ocorrências (acidentes e paralisações) dessa rodovia o classificador obteve este desempenho, contudo considerado razoável (66).

Com os resultados encontrados e demonstrados acima pelo algoritmo Árvore de Decisão, a variável BRajustada foi escolhida para explicar o comportamento das rodovias. Foram construídas Árvores de Decisão com as ferramentas R-cran, Weka e Knime. O Weka construiu uma árvore ininteligível, o R-cran será discutida a seguir e uma parte da Árvore de Decisão que o Knime construiu é demonstrada na Figura 4.26.

Figura 4.26: Árvore de Decisão gerada pelo Knime



Fonte: o Autor

A árvore construída pelo Knime para a mesma variável “Causa do Acidente” => velocidade incompatível. Devido a grande quantidade de nós folha foi escolhida parte da Árvore onde foram classificadas os acidentes em dias-da-semana, com índices que chegaram a 100%. O valor “p-value” abaixo de 0,05 é destacado na Figura 4.6, também para exemplificar, o nó folha classificou como causas dos acidentes: nas quartas-feiras: “ultrapassagem indevida”; nas sextas-feiras: “defeito na via”; e no sábado: “dormindo ao volante”.

Contudo, os melhores resultados, de acordo com mais alta precisão, segundo a métrica dos classificadores, foi a variável “BRajustada” com nó raiz, com curva ROC acima dos 90% em quase todas as classes. O classificador Naïve Bayes obteve um desempenho semelhante, com essa variável. Somente na BR 408 e BR 110 ficou abaixo, o que confirma os valores encontrados pelo Weka. Os valores das regras encontradas pelo algoritmo Árvore de Decisão com a variável “Delegacia”, (abaixo do nó raiz BRajustada) foram:

(a) “Delegacia” [1101(Região Metropolitana)], [BR 101], [KM: 4], [Traçado da via: Reta], [Gravidade = S (acidente com mortes) = [Causa Acidente: Falta atenção] [Causa Acidente: Velocidade incompatível] [Causa Acidente: Ultrapassagem indevida] [Causa Acidente: Defeito mecânico] [Causa Acidente: Não guardar distância] [Causa Acidente: Dormindo] [Causa Acidente: Ingestão de álcool]

(b) “Delegacia” [1101(Região Metropolitana)], [BR 232], [KM: 17], [Condição pista: Seca], [Tipo Auto: automóvel] = [Causa Acidente: Velocidade incompatível] [Causa Acidente: Ultrapassagem indevida] [Causa Acidente: Desobediência à sinalização] [Causa Acidente: Não guardar distância] [Causa Acidente: Dormindo] [Causa Acidente: Ingestão de álcool]

Essa variedade de causas explica que o condutor dessa região não respeita a sinalização, os limites de velocidade, dentre outras regras de trânsito. Pode-se dizer que é um condutor indisciplinado, pois todas as causas de acidentes elencadas foram encontrados. Caso se considere um raio de 50 Km no entorno da capital Recife, acredita-se que os motoristas têm a mesma característica, pelo tipo de acidente que acomete nessa área. Os valores das regras encontradas pelo algoritmo para a variável “Tipo do Acidente” foram: (a) “Tipo de Acidente” [região metropolitana]: [Atropelamento de pessoa], [pista seca], [período: noite], [Br < 116 (101, 104)] , [Dia da semana: terça-feira]: [Gravidade = N (sem morte)], [Km <= 69] => falta de atenção. [Gravidade = S (com morte)] => outras.

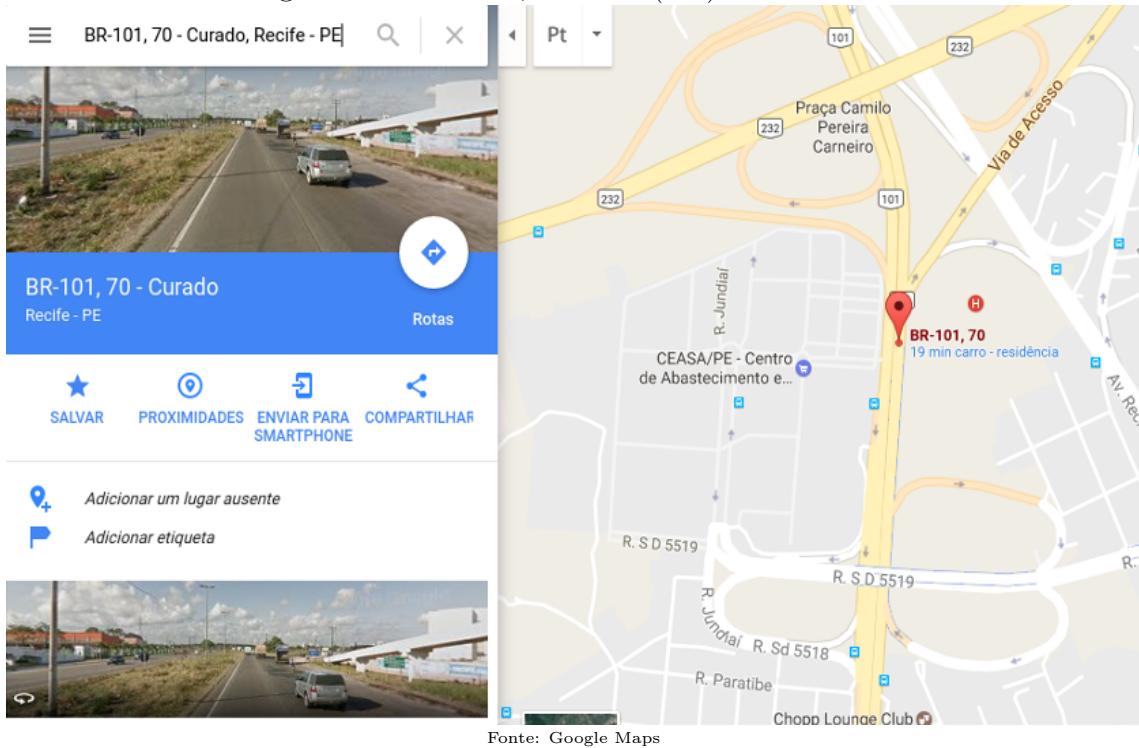
Tipo de Acidente: [Atropelamento de pessoa], [pista seca], [período: noite], [Br < 116 (101, 104)] , [Dia da semana: sexta- feira]: [Gravidade = N (sem morte)], [Km <= 58] => falta de atenção. [Gravidade = S (com morte)] => [Km > 58] [Km <= 67] => falta de atenção.

A falta de atenção foi condição “sine qua non” que determinou os acidentes na região metropolitana do Recife. Os dados revelam, ainda, que em torno do Km 67 encontra-se o maior número de acidentes com morte de todo estado de Pernambuco.

As Figura 4.27 e 4.28 correspondem ao mesmo ponto na BR 101, a Figura 4.27 foi obtida a partir da API do Google Maps, ela demonstra o local (aproximadamente o Km 70, Br 101 – sul) destacado na Matriz de Mortos, Figura 4.28, logo abaixo. Ao ser consultado, este ponto nas Árvores de Decisão, para explicar as causas do alto índice de óbitos, constatou-se que eram, em sua maioria, mortes por atropelamento.

A imagem do Google Maps, Figura 4.27, define que o local é próximo à CEASA, que é principal Centro de Abastecimento de alimentos da região, com um grande fluxo de pessoas – muitas delas das comunidades do entorno – e de veículos que vêm de diversas regiões do país, para comercialização dos produtos em grosso e varejo. Esse exemplo aponta para a ideia de extração das ferramentas utilizadas nessa pesquisa.

Figura 4.27: Km 70, BR 101 (Sul) Pernambuco



Em seguida a Matriz de Mortos e a Árvore de Decisão correspondente a esse trecho. Chamamos a atenção para o km 68 desta rodovia o número de óbitos somente neste local é o dobro dos outros. Provavelmente o local mais perigoso para se atravessar em todo Estado de Pernambuco.

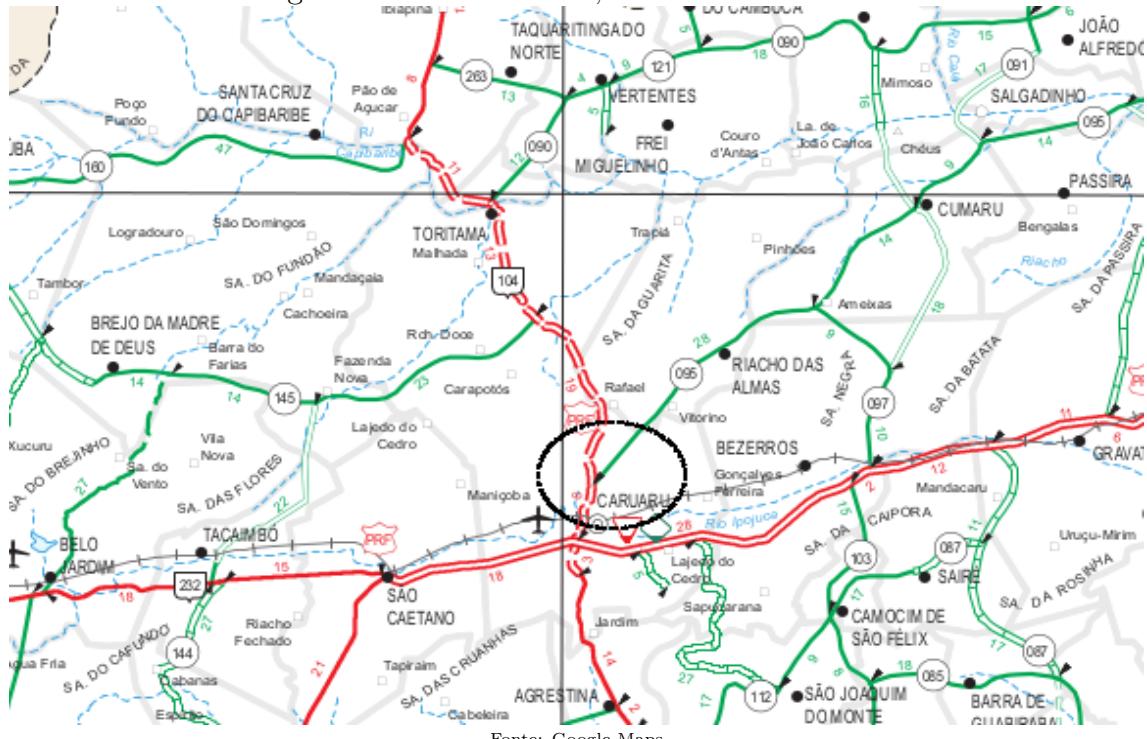
Figura 4.28: Matriz de Mortos: Km 56 – 78, BR 101 (Sul) Pernambuco

		Km (56 - 78)																							
		56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	
Hora (0 - 23)		0	0	10	14	10	1	0	0	3	4	3	9	11	19	17	19	20	4	7	6	8	9	0	8
1	0	1	2	1	0	0	0	0	7	8	1	3	29	11	4	3	7	0	2	1	2	2	2	2	
2	5	3	15	4	0	8	21	2	12	1	0	1	16	20	4	0	4	8	3	0	0	0	0	8	
3	0	11	0	0	0	8	0	4	0	2	1	7	6	16	16	8	0	4	2	0	4	0	0	0	
4	0	8	2	8	0	0	0	0	10	10	4	5	13	9	25	4	1	2	2	2	4	6	1	1	
5	0	11	6	11	0	6	3	8	7	20	34	11	25	4	32	13	11	9	15	0	2	13	10	10	
6	2	17	24	26	31	13	0	23	38	26	44	50	106	30	45	45	14	38	21	39	7	52	22	22	
7	8	49	38	29	39	50	5	63	51	67	106	75	182	116	119	111	64	72	75	54	34	22	64	64	
8	4	25	45	42	18	48	5	41	53	90	114	73	244	109	163	109	36	88	60	24	39	51	50	50	
9	9	34	37	22	24	21	7	48	55	41	87	62	203	114	100	66	67	50	79	29	14	54	51	51	
10	1	40	42	27	28	13	2	37	49	33	52	64	145	66	55	80	25	45	25	10	19	23	37	37	
11	2	70	27	19	8	5	4	48	60	27	56	63	156	77	71	95	26	51	29	38	30	31	76	76	
12	1	29	25	6	14	7	0	15	27	26	73	122	259	64	51	66	26	25	24	12	29	9	25	25	
13	0	10	21	11	19	17	2	36	14	13	62	34	157	65	66	61	25	33	21	16	26	39	56	56	
14	0	24	57	4	13	13	3	45	53	41	53	63	205	59	94	82	24	29	19	19	11	26	36	36	
15	7	33	37	16	14	32	2	20	33	29	84	84	213	115	67	38	46	36	34	25	33	22	44	44	
16	2	16	51	16	21	11	20	43	48	35	43	79	176	63	99	54	21	22	32	17	9	54	18	18	
17	0	23	22	21	9	8	5	25	32	34	67	81	191	158	110	63	31	15	69	13	12	18	35	35	
18	0	42	36	34	10	14	8	44	50	29	120	98	287	133	128	85	32	54	31	19	24	11	27	27	
19	4	12	23	2	5	10	1	16	21	35	69	58	194	120	76	37	26	29	6	27	0	10	12	12	
20	2	12	18	20	20	1	2	9	9	14	23	37	89	24	19	21	21	23	10	12	5	11	13	13	
21	4	16	7	8	3	7	2	18	9	14	16	23	52	42	43	28	8	11	15	2	16	26	5	5	
22	0	8	13	3	1	1	0	6	9	13	10	27	87	22	52	17	10	9	10	5	8	8	12	12	
23	0	8	5	12	3	2	4	2	2	6	13	8	73	11	46	35	9	7	6	7	7	4	1	1	

Fonte: o Autor

Outro trecho, agora na BR 104 (rota da Sulanca) é demonstrado na Figura 4.29.

Figura 4.29: Km 64 e 67, BR 104 Pernambuco



A Matriz de Mortos correspondente a esse trecho é a Figura 4.30, Destacamos os pontos: km 64 e km 67 desta rodovia o número de óbitos nestes locais é grande.

Figura 4.30: Matriz de Mortos: Km 64 e 67, BR 104 - Rota da Sulanca - Pernambuco

		Km (51 - 70)																				
		51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	
)	0	0	0	0	0	0	0	0	4	0	0	1	0	2	2	5	0	1	3	0	0	
)	1	0	0	0	0	0	0	0	0	0	0	6	13	0	5	0	0	8	2	0	0	
)	2	0	0	1	1	0	0	0	2	0	7	0	2	3	1	4	0	0	0	0	0	
)	3	1	0	0	0	1	0	0	0	0	4	2	6	2	5	0	1	0	0	0	0	
)	4	2	1	1	2	0	0	0	1	5	0	0	2	0	2	4	0	6	4	6	0	
)	5	1	0	0	0	0	0	0	1	5	0	2	0	13	4	1	1	3	4	5	0	
)	6	0	0	2	2	0	0	4	0	1	1	8	7	12	17	0	1	12	7	12	0	
)	7	0	2	2	2	6	1	0	0	4	2	0	8	16	7	13	15	29	10	11	0	
)	8	0	6	9	4	0	0	2	0	7	0	7	17	11	13	28	10	41	5	2	0	
)	9	0	3	9	8	2	1	5	0	14	16	6	16	5	38	28	21	32	10	3	0	
)	10	4	19	1	0	10	2	2	0	4	0	5	7	14	32	21	23	21	7	13	0	
)	11	1	4	15	1	2	0	1	13	4	6	13	12	13	24	28	12	23	20	7	0	
)	12	0	2	0	2	1	0	6	2	11	1	4	17	17	27	27	11	23	13	21	0	0
)	13	0	11	3	0	2	0	0	0	2	13	4	26	6	21	22	9	35	7	12	0	
)	14	0	4	4	0	13	0	8	0	9	0	4	13	8	18	16	9	37	14	0	0	
)	15	1	1	0	15	7	6	0	9	7	0	7	21	34	20	18	15	33	13	0	0	
)	16	0	0	9	0	4	0	2	4	4	6	12	19	13	28	20	26	40	22	34	0	
)	17	3	5	5	1	11	3	0	0	0	5	4	18	26	27	20	21	25	16	0	2	
)	18	0	10	6	13	0	1	3	21	12	10	9	13	30	35	10	26	26	18	2	3	
)	19	3	5	3	0	2	1	0	0	15	4	15	23	7	13	25	4	25	13	0	0	
)	20	0	1	0	0	2	0	0	0	8	4	1	20	0	7	11	6	12	9	0	0	
)	21	0	2	0	0	13	0	3	1	4	3	0	5	9	13	8	15	1	7	0	0	
)	22	0	6	1	1	0	0	0	1	0	0	3	10	10	17	10	1	5	8	4	0	
)	23	0	3	0	0	0	0	0	0	4	0	5	0	10	2	3	4	0	8	0	0	
somatório de acidentes / km																						
Σ	16	85	71	52	76	15	36	59	116	86	113	278	250	388	302	246	432	220	119	5		

Fonte: o Autor

A região no entorno da BR 116, os acidentes com mortes [Gravidade = S] ocorrem frequentemente às quinta-feira, envolvendo todos os tipos de veículos. Os valores das regras encontradas pelo algoritmo para a variável “Causa do Acidente” foram: [Ingestão de álcool], [Tipo de auto: não identificado], [Período: Manhã] o tipo de acidente => colisão traseira. [Ingestão de álcool], [Tipo de auto: automóvel], [Traçado da via: Reta], [Condição da pista: molhada], [Dia da semana]: [Segunda-feira] => colisão frontal [Terça-feira] => colisão transversal [Quarta-feira] => colisão transversal [Quinta-feira] => saída de pista [Sexta-feira] => colisão traseira [Sábado]: [BR = 232] => colisão traseira [BR > 232] => colisão frontal

Figura 4.31: Matriz de Gravidade 3D: Br 116 predição para quinta-feira
Km (20 - 28)

	20	21	22	23	24	25	26	27	28
0	0	0	0	0	0	2,697	0	0	0
1	0	0	0	0	0	0	1,697	0	0
2	0	0	0	0	2,697	0	0	0	1,656
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1,256	0	0
7	0	0	0	2,697	0	0	2,697	0	0
8	0	2,697	0	0	0	2,179	0	1,697	0
9	0	0	0	0	0	0	3,394	0	1,697
10	0	0	0	0	2,697	2,697	3,876	0	0
11	0	0	0	0	0	0	1,697	0	0
12	0	2,697	0	0	0	2,179	1,697	0	0
13	0	0	0	0	0	0	0	0	0
14	0	0	0	2,697	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	1,697	0
17	0	0	0	0	0	0	4,876	0	0
18	0	1,697	0	0	0	0	1,697	1,256	0
19	0	0	0	0	0	0	1,179	0	0
20	0	0	0	0	0	0	2,876	0	0
21	0	0	0	0	0	0	1,697	0	0
22	0	0	0	0	0	2,697	1,697	0	0
23	0	0	0	0	0	0	0	0	0

Fonte: o Autor

4.2.5 Resultado dos outros classificadores utilizados nesta pesquisa

Os dados gerados pelos classificadores – Naïve Bayes e Redes Neurais no software Weka permitiram confirmar que os resultados encontrados pelas Árvores de Decisão são os mais adequados para o modelo proposto.

O resultado do algoritmo Naïve Bayes é demonstrado na Tabela 4.10 (Instâncias classificadas corretamente e incorretamente), pela Tabela 4.11 (Acurácia) e pela Tabela 4.12 (Matriz de Confusão).

Tabela 4.10: Instâncias classificadas e Erro médio

Descrição	Valores	Percentual
Instâncias Corretamente Classificadas	9232	73,3339%
Instâncias Incorretamente Classificadas	3357	26,6661%
Erro médio absoluto	0,0588	—
Erro médio quadrático	0,1908	—

Fonte: o Autor

Tabela 4.11: Detalhe da acurácia para classe BRajusta

TP	FP	Prec.	Recall	F-Me.	AUC	Classe
0,947	0,277	0,734	0,947	0,827	0,914	BR101
0,895	0,008	0,893	0,895	0,894	0,997	BR104
0,159	0,000	0,538	0,159	0,246	0,996	BR110
0,833	0,010	0,510	0,833	0,633	0,995	BR116
0,429	0,052	0,748	0,429	0,546	0,834	BR232
0,589	0,016	0,510	0,589	0,546	0,983	BR316
0,772	0,003	0,901	0,772	0,832	0,996	BR407
0,165	0,013	0,284	0,165	0,209	0,890	BR408
0,607	0,013	0,650	0,607	0,628	0,983	BR423
0,377	0,006	0,594	0,377	0,461	0,988	BR424
0,875	0,009	0,849	0,875	0,862	0,994	BR428

Fonte: o Autor

O algoritmo Naïve Bayes encontrou o percentual de instâncias corretamente classificadas um pouco abaixo do percentual da Árvore de Decisão. A área sob a curva ROC (AUC) para a variável BRajustada, o Naïve Bayes (NB) obteve valores próximos aos da Árvore de Decisão (AD) para quase todas as classes, com exceção para a área da classe BR 408:

- para NB foi 0,890
 - para AD foi 0,781
- diferença portanto de 10,9% acima para o NB.

A Tabela 4.12 corresponde à Matriz de Confusão do NB para a variável BRajustada.

Tabela 4.12: Matriz de confusão – BRajustada – Naïve Bayes

a	b	c	d	e	f	g	h	i	j	k	Classificadas
5317	0	0	0	198	0	0	102	0	0	0	a = BR101
0	761	0	0	88	0	0	0	1	0	0	b = BR104
0	0	7	0	12	0	0	0	20	5	0	c = BR110
0	0	0	130	1	24	0	0	0	0	1	d = BR116
1605	69	0	0	1424	159	59	0	0	0	0	e = BR232
0	0	0	94	47	206	0	0	0	0	3	f = BR316
0	0	0	0	1	1	346	0	0	0	100	g = BR407
323	0	0	0	1	0	0	64	0	0	0	h = BR408
0	22	5	0	95	0	0	0	290	66	0	i = BR423
0	0	1	0	36	0	0	0	135	104	0	j = BR424
0	0	0	31	0	14	38	0	0	0	583	k = BR428

Fonte: o Autor

As classes a, b, c,...,k correspondem à coluna “Classificadas” (BR 101, BR 104,...,BR 428). Podemos constatar que há um maior número de entradas diferentes de zero em cada linha dessa tabela, portanto o classificador NB “errou” mais vezes ao tentar encontrar determinadas classes corretas, por exemplo:

Tabela 4.13: Matriz de confusão – BRajustada – Naïve Bayes

a	b	c	d	e	f	g	h	i	j	k	Classificadas
5317	0	0	0	198	0	0	102	0	0	0	a = BR101

Fonte: o Autor

Nas classes “e” e “h” da linha a = BR101, o classificador “errou”, respectivamente em 198 e 102, portanto deveria ter entradas a zero caso tivesse classificado corretamente todas as instâncias.

Tabela 4.14: Matriz de confusão – BRajustada – Naïve Bayes

a	b	c	d	e	f	g	h	i	j	k	Classificadas
1605	69	0	0	1424	159	59	0	0	0	0	e = BR232

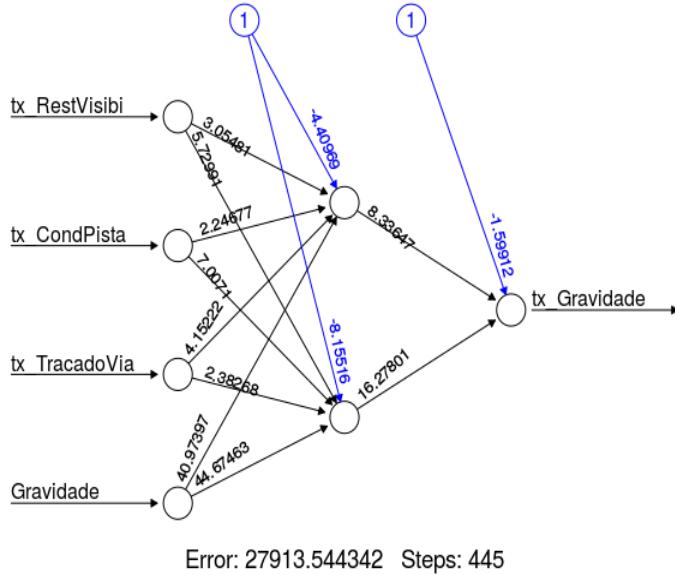
Fonte: o Autor

Nas classes “a”, “b”, “f” e “g” da linha e = BR232, o classificador “errou”, respectivamente, em 1605, 69, 159 e 59. Podemos constatar que na classe “a”, o NB obteve mais erros (1605) do que acertos (1424), portanto ele não foi capaz de encontrar uma quantidade adequada de classes corretas para e = BR232. Concluímos, assim, que o algoritmo Árvore de Decisão supera o Naïve Bayes.

A Rede Neural utilizada na classificação dos pontos críticos das rodovias está descrita na Figura 4.32. As variáveis de entrada são descritas segundo a equação a seguir:

$$ProbAcid = (RestVisibi * CondPista * TracadoVia * TipoAcident * CausaAcident) + Gravidade \quad (4.1)$$

Figura 4.32: Rede Neural utilizada na classificação



Fonte: O Autor

As variáveis da equação 4.1 contemplam as variáveis da equação 3.1 mais as seguintes variáveis:

- RestVisibi - Restrição de Visibilidade;
- CondPista - Condição da Pista;
- TraçadoVia - Traçado da Via;
- TipoAcident - Tipo de Acidente;
- CausaAcident - Causa do Acidente;
- Gravidade - Gravidade do acidente.

A Equação 4.1 foi baseada nos parâmetros escolhidos para a Equação 3.1 que por sua vez foram baseados no entropia e correlação linear forte entre as variáveis. Ao acrescentar mais parâmetros à Equação 4.2 procurou-se chegar mais próximo dos resultados encontrados pelos outros classificadores utilizados nesta pesquisa, contudo os resultados foram irrelevantes. A Figura 4.33 demonstra os resultados da Rede Neural como classificador

A Figura 4.32 é um esboço da rede neural utilizada em um dos experimentos. Algumas variáveis como CausaAciden e TipoAcident não se encontram contempladas neste modelo, contudo nos experimentos elas não alteraram o comportamento do classificador.

A partir desse resultado podemos concluir que a Árvore de Decisão encontrou resultados mais promissores, apesar da taxa de erro médio absoluto (vide Figura 4.33) da Rede Neural ser o menor dos classificadores utilizados nesta pesquisa.

Figura 4.33: Resultado da classificação feita pela Rede Neural – acurácia

```

Redes Neurais -- classificador

Threshold -7.651297135217415
Attrib CondPista 9.154976078003095
Attrib RestrVisibili -4.4972310838186305
Attrib TipoAcident 1.3788162272305897
Attrib CausaAcident -3.3073519252019596
Attrib TracadoVia -0.0796848784283105
Attrib TipoAuto 10.246140682910276
Attrib Hour -3.9835710841832386
Attrib Delegacia 0.0780543602125268
Attrib tx_TracadoVia 2.336680234369794
Sigmoid Node 5
Inputs Weights
Threshold 2.7475234065506204
Attrib CondPista 0.7451431464473918|
Attrib RestrVisibili 6.228899014250266
Attrib TipoAcident -1.1503782569405412
Attrib CausaAcident -0.08546733165157612
Attrib TracadoVia 2.381725131910798
Attrib TipoAuto 18.505265905598662
Attrib Hour -3.0447003266782393
Attrib Delegacia -0.9246810390821438
Attrib tx_TracadoVia -2.448400403741353
Class
Input
Node 0

Time taken to build model: 100.17 seconds

==== Evaluation on training set ====
==== Summary ===

Correlation coefficient 0.9973
Mean absolute error 0.0095
Root mean squared error 0.0197
Relative absolute error 3.7444 %
Root relative squared error 7.4925 %
Total Number of Instances 45320

```

Fonte: O Autor

4.3 Acoplamento com a estrutura dinâmica

As predições feitas na primeira fase têm como “output” georreferenciamento que localiza um ponto no mapa a partir do quilômetro (Km). O georreferenciamento mais preciso seria a latitude e longitude. Todavia, esses dados apresentaram grave inconsistência na base de dados da PFR/PE, tendo sido descartados. Sentimos necessidade de informações sobre a teoria do fluxo de tráfego, que consiste de leis matemáticas, físicas aplicadas ao tráfego de automóvel, tais como o fluxo (veículos/hora), a velocidade (km/hora) e a densidade de tráfego (veículos/km).

A estrutura dinâmica é composta por duas API’s, uma disponibilizada pelo Google, através do Google Maps, que está atualmente na versão V3, e a outra por uma API do Twitter. A API do Google Maps proporciona uma “leitura” atualizada e precisa, de forma que os dados do Km da rodovia podem ser localizados no mapa.

A API do Twitter, por sua vez, possibilita atualizar o utilizador com informações recentes. Contudo, o objetivo desta API é fazer um Arco Cibernético das informações, retroalimentando, com dados recentes, um banco de dados de redes sociais. Isso permite um visualização instantânea do ambiente como um todo.

4.3.1 Mineração em texto no Twitter

A mineração de dados textuais na rede social Twitter demonstrou ser uma ferramenta promissora, uma vez que oferece uma ampla gama de informações, atualizadas em tempo real. Entretanto, para o caso de pesquisas dessa natureza, o monitoramento precisa ser constante, tendo em vista que novas informações são produzidas a todo momento; e outros canais precisam ser monitorados, a fim de ampliar o universo de dados disponíveis, para produzir o efeito esperado no modelo.

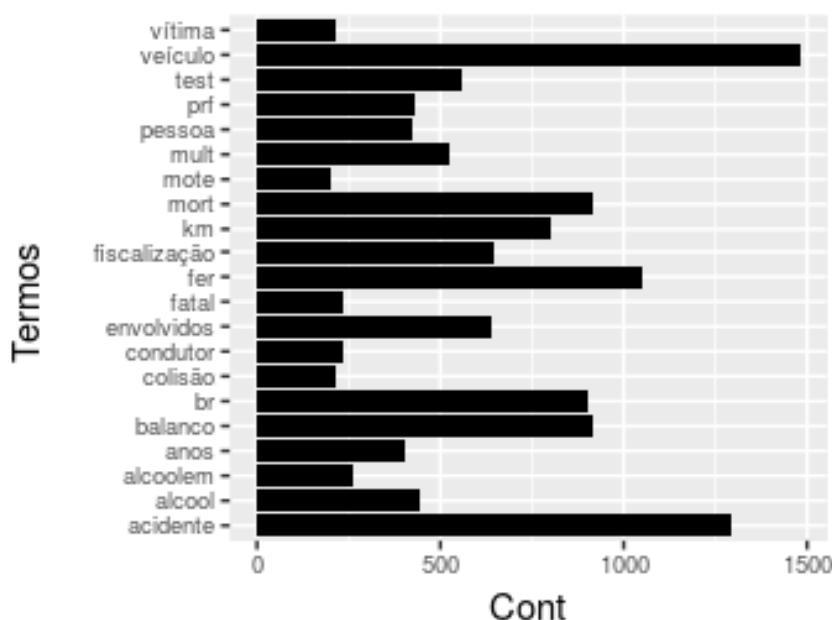
Para localizar novos canais de informação, foram construídos subgrafos contemplando, além dos tweets da PFR, retweets dos seguidores desse canal. Esperava-se, com isso, encontrar novos Hubs na rede. No entanto, a busca em subgrafos é uma tecnologia que precisa ser investigada mais a fundo, dada a sua complexidade.

Conceitos como intermediação, centralidade, peso das arestas, precisam ser adequadamente compreendidos pelo pesquisador, para que a tecnologia seja explorada em todo seu potencial. Isso implica no estudo de novos algoritmos, próprios para mineração em grafos de redes sociais, o que, em princípio, fugia do escopo dessa pesquisa.

Ainda que consideremos a ferramenta não tenha sido adequadamente explorada, por todas as questões propostas acima, a mineração de dados textuais do twitter permitiu inferir que o utilizador dessa rede social faz referência ao que acontece nas rodovias, com especial atenção para fatos que possam ter implicação em seu cotidiano, tanto imediatamente (por exemplo, congestionamento numa via que ele utiliza para ir ao trabalho), quanto num universo temporal mais distante (por exemplo, condição da rodovia que dá acesso à cidade em que ele vai passar um feriado). Esse último aspecto é aquele que interessa particularmente ao modelo proposto nessa pesquisa.

Os dados da Figura 3.34 mostram a busca dos termos frequentes encontrado no documento textual, extraído a partir de 3.200 tweets. O primeiro gráfico apresenta unigramas – termos frequentes que contém uma palavra. Foram encontradas palavras como: fiscalização, colisão, vítima fatal, acidentes, entre outros.

Figura 4.34: Gráfico de frequência de palavras – unigramas



A nuvem de palavras da Figura 4.35 é um gráfico de frequência de palavras em que os termos mais frequentes aparecem em destaque – com letras em tamanho maior – seguidos pelos próximos termos mais frequentes – em tamanho um pouco menor – e assim sucessivamente, chegando a contemplar dezenas ou até centenas de palavras, a depender da escolha do pesquisador. No caso da nuvem de palavras a seguir, aparecem cerca de 50 termos, e destacam-se como mais frequentes (em virtude do tamanho): veículo, BR, acidente, balanço, dentre outras.

Figura 4.35: Gráfico tipo Nuvem de Palavras referente aos unigramas



O dendrograma da Figura 4.36 é um gráfico que agrupa palavras de acordo com o assunto. No dendrograma a seguir foram identificados seis agrupamentos (clusters), como, por exemplo: ferido, balanço e morto; envolvidos, veículo (veic), fiscalização (fisc), multa, alcoolemia, pessoa, test; PRF; fatal, vítima, envolvido, óbito, prisão, condutor, colisão, moto, ano, h (hora). Os agrupamentos, quando analisados adequadamente, permitem identificar, com um número reduzido de palavras, o acontecimento ao qual está sendo feita referência. Por exemplo, no último agrupamento exemplificado acima é possível deduzir que houve uma colisão envolvendo uma moto, que resultou no óbito de um dos envolvidos e na prisão daquele que foi o responsável pela colisão.

Figura 4.36: Dendrograma de Agrupamento (clustering) do resultado da mineração

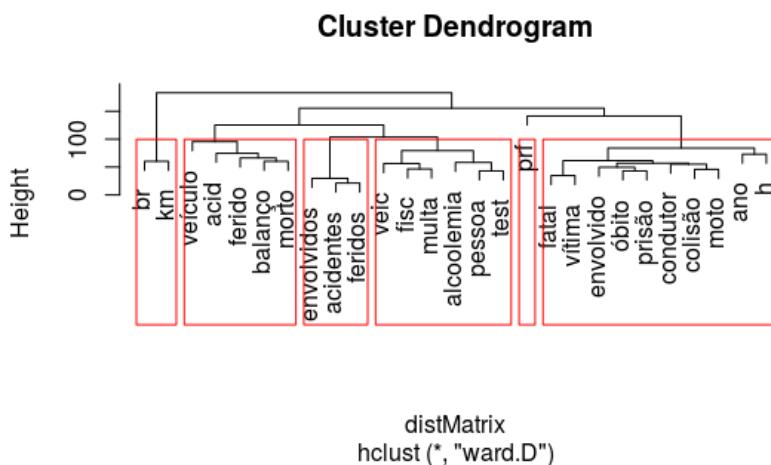
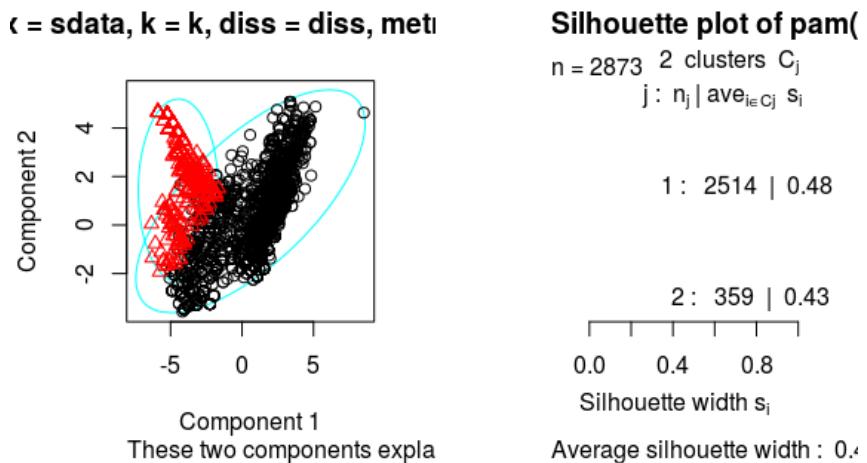


Figura 4.37: Gráfico da Agrupamento (clustering) do resultado a mineração



O Gráfico 4.37 ilustra a utilização de outro algoritmo de agrupamento, conhecido com K-means.

Uma análise mais minuciosa dos gráficos produzidos pela mineração em textos pode trazer contribuições de considerável relevância, que permitem refinar o modelo de predição da Etapa 1, descrita na Figura 4.1 do modelo proposto.

5

Considerações finais

Após a realização dessa pesquisa, os dados sugerem que é possível propor um modelo de predição que possibilite a gestão logística relativa a dias e horários de utilização das rodovias para os mais diversos atores que dela fazem uso das rodovias.

Os resultados encontrados apontam na mesma direção de trabalhos realizados no Brasil e em outros países, que destacam para a eficácia da aplicação da I.A. para analisar questões relativas ao tráfego de veículos e problemas que atingem as rodovias, comprometendo o deslocamento daqueles que a utilizam, tanto para uso privado, quanto relacionados ao contexto profissional e logístico.

Isto posto, órgãos federais de controle podem se beneficiar de pesquisas dessa natureza para compreender o comportamento do tráfego nas rodovias e propor soluções para os problemas que afetam esse cenário diariamente. Os dados encontrados sugerem que há um padrão de comportamento em rodovias que pode ser analisado, de maneira a facilitar o tráfego de veículos em geral, em seu curso.

Outro aspecto que pode impactar positivamente no georreferenciamento é o registro adequado e utilização constante, pelos órgãos de trânsito, das informações de latitude e longitude, que apareceram precariamente nos dados da PRF/PE. Essa medida facilitaria o trabalho de pesquisadores e profissionais que desenvolvem sistemas de posicionamento global. Por exemplo, o KM da BR 101 existe em Pernambuco, no Rio de Janeiro e, possivelmente, em outros locais. Assim, uma das principais informações utilizadas pela PRF para registro de ocorrências apresenta inconsistência.

A pesquisa também contribuiu para a compreensão das causas de constrangimentos e incidentes nas vias. Os dados revelaram que a maioria dos acidentes ocorre em via reta, com pista seca e em boas condições, sugerindo como uma das principais causas de acidente a falta de prudência do condutor, que resulta, na maioria das vezes, em colisões traseiras e/ou laterais, essa última sendo a que mais culmina em óbito.

Numa análise restrita a acidentes com morte por atropelamento, os dados revelaram que próximo aos perímetros urbanos não há uma estrutura segura para o pedestre atravessar a via, ou que o induza a obedecer as regras de trânsito, como por exemplo: atravessar somente na faixa de pedestre, quando estiver em trechos urbanos de rodovias; fazer uso dos semáforos para atravessá-las, quando houver; utilizar as passarelas (quando há) para

cruzar as BRs com fluxo intenso. Outros condicionantes são, por exemplo: falta de sinalização, passarelas, limitador de velocidade e faixa de pedestre. A quantidade de mortes por atropelamento entre os quilômetros 66 e 71 da BR 101 - mais de 9.500 mortes em nove anos - apontam nessa direção.

Esse padrão de comportamento também foi identificado em outras pesquisas discutidas nessa dissertação e sugere que as variáveis do condutor são aquelas mais fortemente relacionadas a acidentes em rodovias, quer seja no perímetro urbano ou no interior. Isso pode indicar que não apenas no Brasil, mas em outros países, inclusive os mais desenvolvidos (como Estados Unidos, Portugal, Espanha e Itália) o motorista não obedece às leis de trânsito, e mesmo com punições rigorosas e multas com altos valores, continuam dirigindo após ingestão de álcool, sem utilizar cinto de segurança, com velocidade acima da permitida, fazendo ultrapassagens indevidas, como revelam as pesquisas.

Nessa dissertação, identificamos que os acidentes acontecem nos horários em que há mais veículos trafegando na via, entre seis e nove horas da manhã, e entre quatro e sete horas da tarde. Quando a via exige maior atenção, por condições que lhes são peculiares (um cruzamento, por exemplo), uma pequena restrição pode ser amplificada, aumentando consideravelmente a quantidade de acidentes.

Outra contribuição da pesquisa a ser destacada é de cunho metodológico-prático. Destaca-se, inicialmente, a articulação entre os resultados envolvendo diferentes algoritmos. A pesquisa utilizou Naïve Bayes, TF-IDF para os dados minerados do Twitter; Árvores de Decisão, Redes Neurais e Naïve Bayes, para o trabalho com os dados da PRF no modelo de classificação.

Observa-se que essa parece ser uma tendência nos trabalhos que envolvem análise de tráfego em rodovias: a utilização de diferentes técnicas de I.A. para explicar o comportamento das rodovias. Todavia, o objetivo de parte dessas pesquisas foi a comparação entre as técnicas, para estabelecer qual a melhor delas. No caso específico dessa pesquisa, o nosso interesse foi o de articular os resultados, não apenas identificando qual o melhor algoritmo, mas propor uma correlação entre eles.

Um exemplo que podemos dar a esse respeito, discutido no capítulo dedicado aos resultados, diz respeito ao fato de que a Matriz de Mortos e a Matriz de Gravidade nos deram informação sobre locais com maior número de acidentes, inclusive com óbitos, e no cruzamento com a árvore de decisão produzida identificamos a causa desses óbitos: morte por atropelamento. Ainda na perspectiva de articulação entre os dados, identificando a partir da API do Google Maps, que tratava-se de um lugar de grande tráfego de veículos e de pessoas, uma vez que era o trecho próximo à CEASA/PE. Assim, reiterando o que dissemos anteriormente, nossa pesquisa propõe, mais do que a comparação, o cruzamento entre os dados, propondo uma análise multidimensional e pluritecnológica, ou seja, considerando mais de uma dimensão e levando em conta mais de uma tecnologia.

Ainda do ponto de vista metodológico, ressalta-se a contribuição da aplicação do processo CRISP-DM, utilizado para construir o modelo de mineração de dados. O algoritmo Árvores de Decisão mostrou-se robusto quando aplicado a esse tipo de problema.

Quanto à mineração de textos em redes sociais (Twitter), foi possível encontrar informações, classificá-las e extrair delas dados relevantes. Entretanto, um contraponto que é importante estabelecer refere-se ao fato de que não conseguimos estabelecer um valor numérico relativo a essas informações, para haver uma extração para a equação do momentum, proposta para o estabelecimento do modelo de predição.

Outro aspecto a ser mencionado é que embora tenha sido uma ferramenta útil, há a necessidade de um monitoramento intensivo das redes sociais, para que possa haver uma

influência maior nos resultados. Seria também necessário ampliar o escopo para outras redes sociais que pudessem informar sobre o comportamento de usuários de rodovias, uma vez que nem sempre o utilizador dessa rede social está disposto a informar o que ele pretende fazer - como usuário de rodovias - por essa rede, cujas informações precisam ser sucintas, dadas as características do aplicativo. Ele pode, por exemplo, não propagar a informação no Twitter, mas fazê-lo no facebook.

Todavia, embora possamos admitir que o estudo das redes sociais precisaria ser mais aprofundado, destacamos que não foram encontrados estudos envolvendo o campo de nosso interesse - comportamento de tráfego em rodovias – e as redes sociais, com vistas à articulação entre mineração de dados e mineração de textos.

Podemos ainda destacar a contribuição da pesquisa com a proposição de um modelo que integre predição à API de mapas de posicionamento global, fornecendo informação suficiente a um gestor para decidir quando enviar, por exemplo, uma frota de caminhões por determinada rodovia que apresente retenções frequentes, ou um usuário da rodovia decidir em qual horário deve viajar. Nesse segundo caso, por exemplo, é comum muitos motoristas optarem por viajar pela manhã bem cedo, quando o sol ainda não está a pique e quando (supostamente) há menos veículos na rua. Nossa pesquisa mostra, por outro lado, que nas primeiras horas da manhã concentra-se um alto número de acidentes, contrariando a preferência que encontramos no senso comum.

Ainda no que diz respeito aos trabalhos dessa natureza, alguns deles considerados no Estado da Arte dessa dissertação, essa pesquisa avança em relação ao que até então foi proposto em outros trabalhos, pelo fato de que além de identificar ocorrências nas rodovias, levando em conta passado e presente, propõe um modelo que contempla o futuro, possibilitando ao usuário escolhas mais assertivas.

Alguns aspectos merecem destaque em relação às dificuldades enfrentada na proposição do modelo de classificação e predição. Em primeiro lugar, as informações da base de dados da PRF apresentavam muitas lacunas (missing values), devido ao tipo de registro feito na ocorrência. Por exemplo, um acidente que envolva dois veículos (ou mais) – um motocicleta e um carro – aparecia, muitas vezes, com dois registros: acidente envolvendo carro e acidente envolvendo moto. Como essa, várias situações de dados duplicados foram identificados, tornando mais complexo o trabalho de preprocessamento dos dados.

Ainda em relação ao registro dos dados pela PRF, relativos à latitude e longitude estavam ausentes ou mal preenchidos, muitas vezes, de forma equivocada, o que comprometeu a localização exata do evento (colisão, atropelamento), fazendo com que fosse necessário desprezar essas variáveis e encontrar outra alternativa, uma vez que a localização era fundamental para o georreferenciamento.

Outra falta percebida em relação aos dados foi a necessidade de informações quanto ao fluxo de tráfego rodoviário. Para suprir nosso modelo proposto de informações semelhantes e para atingir os objetivos já referidos é que foram construídas as Matrizes de Mortos e Gravidade em cada ponto de cada rodovia, para cada cenário.

Foi percebida também certa imprecisão em relação à quilometragem em que se dava a ocorrência, com erros que variavam de alguns metros até quilômetros, sendo necessário o olhar atento do pesquisador e o confronto cuidadoso de informações, de modo que essas falhas fossem dirimidas.

Outra questão que merece destaque é o complexo processo de limpeza de dados da “timeline” do Twitter na etapa de preprocessamento. Atualmente, a própria PRF, para dar mais destaque, tem utilizado imagens para informar as ocorrências nas BRs, em vez de textos, o que tem contribuído para diminuir a quantidade de dados textuais. Com isso

foi necessário procurar outras “timelines” para ampliar a quantidade de informações do Twitter.

5.1 Trabalhos futuros

Essa pesquisa não encerra as questões propostas, relativas ao desenvolvimento de um modelo preditivo. O que foi apresentado, sobretudo, foi a intenção de um modelo que servirá como ponto de partida para o desenvolvimento de uma ferramenta que atenda ao fim proposto de forma eficaz. Assim, entendemos que novas pesquisas precisam ser conduzidas para a ampliação do modelo sugerido.

Trabalhos futuros incluem a incorporação desta proposta em modelos formais de decisão, por exemplo, de roteamento rodoviário. Através da API do twitter, implementar algoritmos de busca em redes sociais, para encontrar os Hubs difusores de informações e minerar os seus textos, de forma a alimentar a equação do momentum.

Particularmente sobre a equação do momentum, a mesma foi concebida a fim de integrar duas situações espaço-temporais: o passado – com as bases históricas – e o futuro – com as informações das redes sociais – a ser extrapolado para um painel de controle (dashboard). Esse painel funcionaria da seguinte maneira: se o peso das informações das redes sociais fosse maior do que o peso das informações dos dados históricos, o sistema se guiaria primeiramente pelas redes sociais. Caso contrário, o sistema se guiaria pelos dados históricos. Por exemplo, na ocasião em que o utilizador traça sua rota para uma viagem futura, o Dashboard informa que se Guiará pelos dados históricos. No entanto, no momento da viagem, o sistema faz nova varredura nas redes sociais e identifica que haverá uma paralisação no trecho escolhido. Nesse momento ele sugere ao usuário postergar a viagem ou utilizar outra rota.

O sistema também poderia incorporar a análise sentimental das postagens dos usuários, de maneira que seja possível antecipar atitudes errôneas dos seus utilizadores. Por exemplo, se o utilizador postou que está bebendo em um bar ou que está exausto e com sono, e que vai “pegar a estrada” em seguida, e sabendo-se que o itinerário do utilizador passará por pontos já detectados como de alto risco, o sistema emitiria um alerta para o condutor redobrar a atenção ou não conduzir o veículo, nem utilizar a via.

Esse mesmo sistema poderia disponibilizar um conjunto de alertas, quando o condutor traçar um roteiro de viagem, e informar sobre locais de alto risco ou alertar para horários críticos, para que o usuário fique mais atento, reduza a velocidade quando necessário, etc.

Todas essas sugestões poderiam culminar no desenvolvimento de um aplicativo para Smartphone. O aplicativo disponibilizaria informações antecipadas de bases históricas e de redes sociais, sobre o itinerário escolhido pelo usuário para uma data futura, diferentemente dos aplicativos disponíveis atualmente, em que o usuário consulta sobre rota a ser utilizada no espaço temporal quase imediato. A partir do momento em que o utilizador traçasse a rota para uma viagem a ser realizada, por exemplo dali a uma semana, quinze dias ou outra data futura, escolhendo origem e destino – semelhante ao que acontece no GPS, Waze, etc. – o aplicativo automaticamente carregaria as informações de bases históricas em forma de alertas, e ao mesmo tempo faria uma varredura nas redes sociais, para identificar a existência de alguma informação nova, não contemplada na base de dados.

Outra implementação do aplicativo poderia ser a emissão de relatórios gerenciais acerca das viagens anteriores do usuário. Por exemplo, indicar no relatório se as informações disponíveis foram bem utilizadas, ou seja, se o condutor fez uso racional das informações, viajando no melhor horário e evitando trechos perigosos em momentos críticos.

Tudo o que foi discutido anteriormente sugere o desenvolvimento de um sistema de suporte à decisão, que dê apoio ao utilizador para a escolha mais segura, eficaz e econômica.

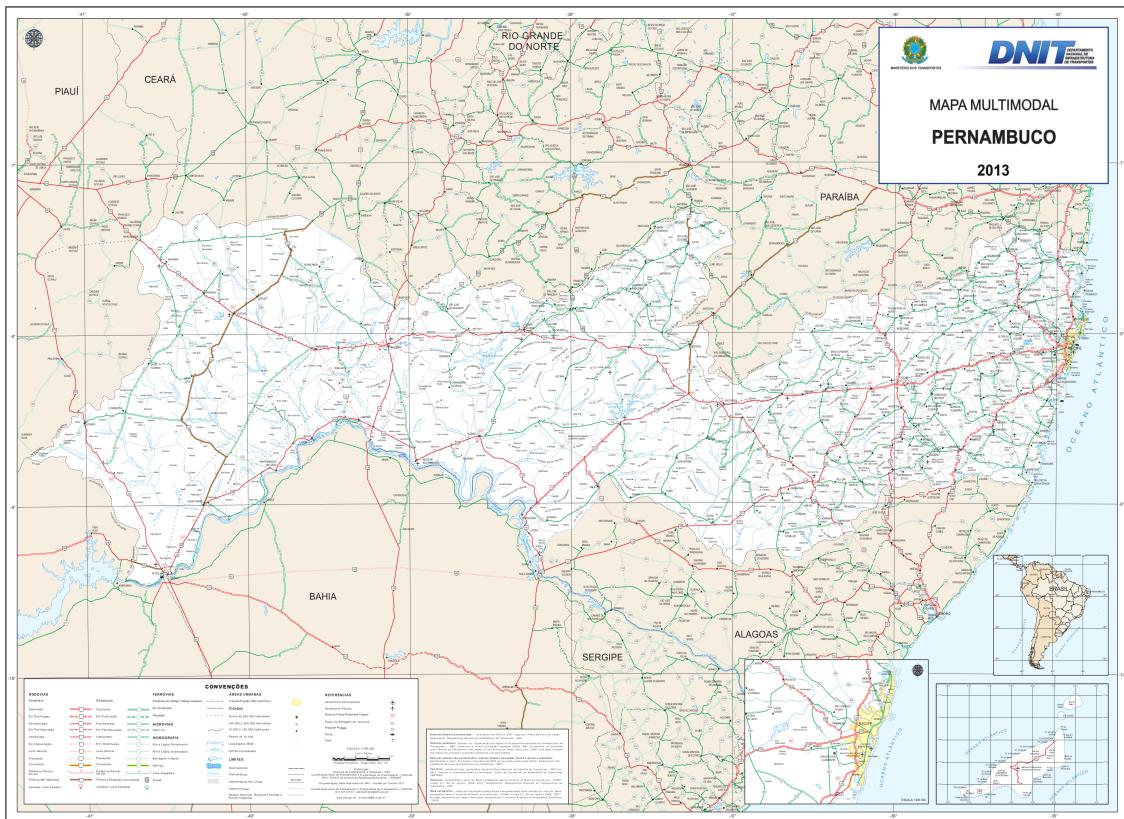
Por fim, no trabalho do pesquisador não cabe um ponto final. Quando uma questão de pesquisa é resolvida, novas questões surgem; questões essas que não teriam sido possíveis se não tivesse sido conduzida a pesquisa proposta. Isso reafirma o quanto espírito do pesquisador deve ser o daquele que faz boas perguntas, não apenas o que o que encontra boas respostas.

A

Preprocessamento

A.0.1 Mapa das de Pernambuco com as rodovias

Figura A.1: Mapa rodoviário de Pernambuco



Fonte: DNIT

A.1 Coleta e Preprocessamento dos dados da PRF

As informações para suprir nosso modelo preditivo estão disponíveis na Internet, em sua maioria são Dados Governamentais Abertos, tais como os dados da PRF, INPE e IBGE. Isto são iniciativas governamentais para fomentar a participação popular, dentro outros motivos, essas informações são também conhecidas como *open data* (6), contudo os dados referentes à PRF e ao BPRv, para esta pesquisa, foram cedidos pelos respectivos órgãos governamentais (ver anexos) já em formato CSV para serem utilizados exclusivamente nesta pesquisa. Isso possibilitou ganho qualitativo nos dados evitando passar pelos transtornos como descreve Costa (2015) quando coletou os dados diretamente da Internet.(85) As bases de dados do INPE e do base de dados do IBGE apresentaram boa qualidade o que justificou serem serem coletados diretamente da Internet.

Tabela A.1: Variáveis originais da base de acidentes

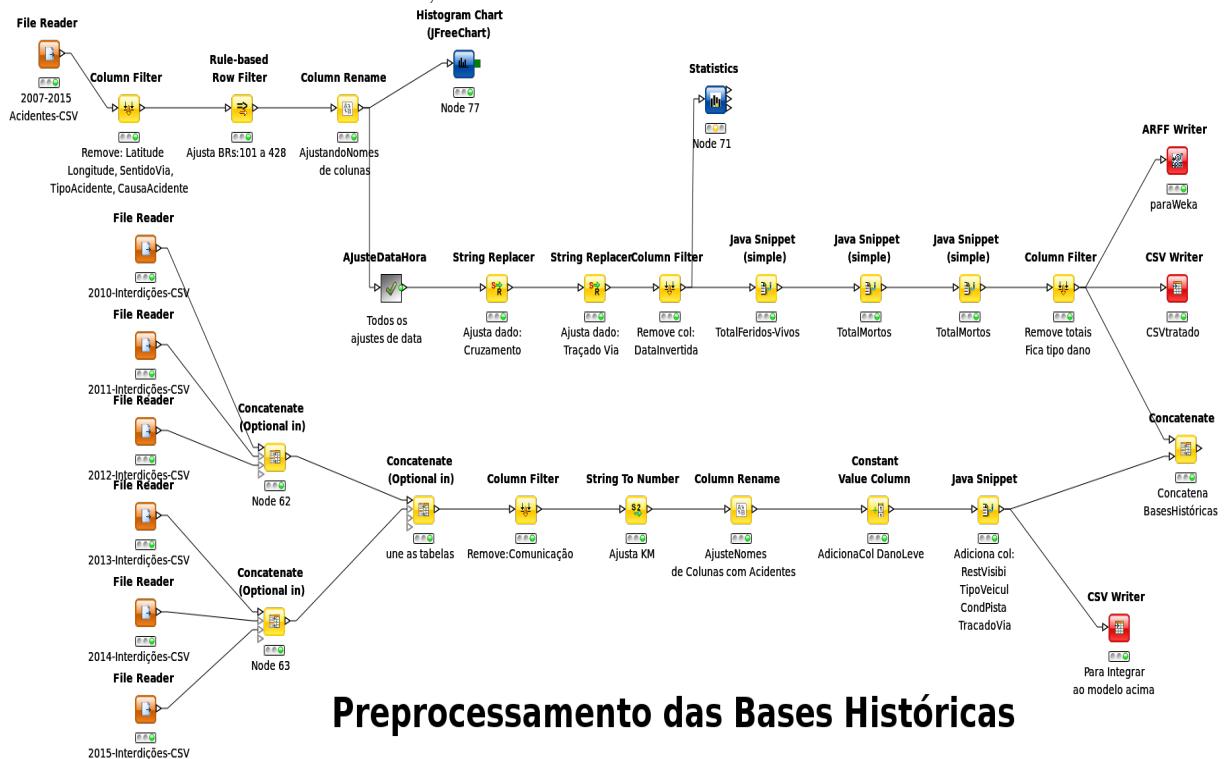
Ano	Ano da ocorrência do acidente
Mês	Mês de ocorrência do acidente
Num	Número do mês do acidente ex: 1 = Janeiro
KM	Numeração do quilômetro
BR	Numeração da Br
Latitude	Latitude da ocorrência
Longitude	Longitude da ocorrência
Condição Pista	Condição da pista: seca, molhado, ...
Restrição de Visibilidade	Restrição de visibilidade: inexistente, neblina, .., outros
Tipo Acidente	Tipo de Acidente: atropelamento, colisão lateral,..
Cauda Acidente	A possível causa do acidente: Falta de atenção, ...
Sentido Via	Sentido da via: crescente, decrescente
Traçado Via	Tipo de traçado da via: reta, curva, cruzamento, ...
Município	Localidade onde ocorreu
Tipo veículo	Tipo de veículo envolvido no acidente
Data Inversa	Data do acidente no formato dd/mm/aa
Horário	Hora que ocorreu o acidente no formato hh/mm/ss
Qtd Feridos Graves	Quantidade de feridos graves envolvidos
Qtd Feridos Leves	Quantidade de feridos leves envolvidos
Qtd Ilesos	Quantidade de ilesos envolvidos
Qtd Mortos	Quantidade de mortos envolvidos
Qtd Pessoas	Quantidade de pessoas envolvidos
Qtd Veículos	Quantidade de veículos envolvidos
Qtd Acidentes Graves	Quantidade de acidentes graves
Qtd Ocorrências	Quantidade de ocorrências

Na Tabela A.2; as variáveis originais da base de dados da PRF com interdições das vias (somente interdições que paralisaram as BRs, não contém acidentes, exemplo: passeatas, protestos)

Tabela A.2: Variáveis originais da base de interdições

Comunicação	Código do agente que comunicou o incidente
Data Hora	Data hora no formato dd/mm/aa mm:ss
BR	Numeração da Br do incidente
KM	Numeração do quilômetro do incidente
Trecho	Local onde ocorreu o incidente

Figura A.2: Etapa 1 – Coleta e união das bases históricas de acidentes e interdições da Polícia Rodoviária Federal - PE, entre 2007 a 2015



Referências Bibliográficas

- 1 CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINAERTZ, T.; SHEARER, C.; and WIRTH, R. Crisp-dm 1.0 – step-by-step data mining guide. *CRISP-DM Consortium*, CRISP-DM consortium, p. 7–10, 2000.
- 2 XAVIER, R. C. M.; COSTA, R. O. da. Relações mútuas entre informação e conhecimento: o mesmo conceito? *Ciência da Informação*, v. 39, n. 2, 2011.
- 3 FLEISHER C. & MARINÉ, R. M. LE COADIC, Yves François. A Ciência da Informação. tradução de Maria Yêda FS de Filgueiras Gomes. Brasília: Briquet de Lemos. *Perspectivas em Ciência da Informação*, v. 1, n. 2, 1996.
- 4 SALLES, F. R. *A relevância da cibernética*. Tese (Dissertação de Mestrado) — Universidade de São Paulo - USP, 2007.
- 5 JAGADISH, J. G. H. V.; SHAHABI, C. *Exploring the inherent technical challenges in realizing the potential of Big Data*. [S.l.: s.n.], 2014. 86–96 p.
- 6 Ministério dos Transportes, Portos e Aviação – Governo Federal. 2016. Disponível em: <<http://dados.gov.br/dados-abertos/>>.
- 7 JUSTICA, P. e Ministério da. 2014. Disponível em: <<http://dados.gov.br/dataset/acidentes-rodovias-federais>>.
- 8 CIDADANIA, P. M. da Justica e. Dados abertos. Acessado em: 20.Out.2015. Disponível em: <<https://www.prf.gov.br/portal/dados-abertos>>.
- 9 BEN-DAVID, S.; SHALEV-SHWARTZ, S. *Understanding Machine Learning: From Theory to Algorithms*. [s.n.]. 449 p. Acessado em: 02.Fev.2016. ISBN 9781107057135. Disponível em: <<http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf>>.
- 10 BNDES. Perspectivas do investimento, n. 2, out. 2013. *Perspectivas do Investimento 2014-2017*, p. 2, 2013.
- 11 BITOUN, J. et al. Região metropolitana do recife no contexto de pernambuco no censo 2010. p. 25, 2012. Disponível em: <http://www.observatoriodasmetropoles.net/download/Texto_BOLETIM\RECIFE_FINAL.pdf>.
- 12 IBGE – Instituto Brasileiro de Geografia e Estatística. Região metropolitana do recife no contexto de pernambuco no censo 2010. 2014. Acessado em: 01.Set.2015. Disponível em: <http://www.cidades.ibge.gov.br/painel/frota.php?codmun=261160&search=pernambuco/recife/infograficos:frota-municipal-de-veiculos/&lang=_ES>.

- 13 DILSIZIAN, L. E. S. E. Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. p. 1.
- 14 CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINAERTZ, T.; SHEARER, C.; and WIRTH, R. Crisp-dm 1.0. *CRISP-DM Consortium*, v. 76, p. 1–2, 2000.
- 15 POSSAS, B.; CARVALHO, M.L.B.; REZENDE, R.S.F.; & MEIRA JR, W. Data mining: técnicas para exploração de dados. *Universidade Federal de Minas Gerais*, 1998.
- 16 FAYYAD, P.; PIATETSKY-SHAPIRO, U. & SMYTH, G. From data mining to knowledge discovery in databases. *Advances in Knowledge Discovery and Data Mining*, v. 17, n. 3, p. 1–36, 1996.
- 17 HAN, J. & KAMBER, M. Data mining: Concepts and techniques. Elsevier, San Francisco, v. 2 edition, 2006.
- 18 RUSSEL, S.; NOVING, P. *Inteligência Artificial*. [S.l.]: Elsevier, Rio de Janeiro, 2004. 716–721 p. ISBN 8535211772.
- 19 CASTANHEIRA, L. G. Aplicação de técnicas de mineração de dados em problemas de classificação de padrões.
- 20 Ralph Kimball. *The Data Warehouse Toolkit*. [S.l.: s.n.], 2013. 1689–1699 p. ISSN 1098-6596. ISBN 9788578110796.
- 21 HAN, J.; KAMBER, M. Data mining: Concepts and techniques. Elsevier, San Francisco, v. 2 edition, p. 15–16, 2006.
- 22 ARANHA, C.; PASSOS, E. *A Tecnologia de Mineração de Textos*. 2006. 1–8 p.
- 23 BURTICA, R. et al. Practical application and evaluation of no-sql databases in cloud computing. In: *2012 IEEE International Systems Conference SysCon 2012*. [S.l.: s.n.], 2012. p. 1–6.
- 24 SRIVASTAVA A.; KATIYAR, V.; SINGH, N. Review of Decision Tree Algorithm: Big Data Analytics. *International Journal of Informative & Futuristic Research*, v. 2, n. 10, p. 3644–3654, 2015.
- 25 AMIN, A. et al. A comparison of two oversampling techniques (smote vs mtddf) for handling class imbalance problem: A case study of customer churn prediction. v. 353, p. 215–225, 2015. Acessado em: 20.Jan.2016. Disponível em: <<http://link.springer.com/10.1007/978-3-319-16486-1>>.
- 26 HOTHO, A.; NÜRNBERGER, A.; PAASS, G. A brief survey of text mining. In: *Ldv Forum*. [S.l.: s.n.], 2005. v. 20, n. 1, p. 19–62.
- 27 STONE, P. et al. The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, Wiley Online Library, v. 8, n. 1, p. 113–116, 1968.
- 28 LIDDY, E. D. Natural language processing. 2001.

- 29 LIMA, A. C. E. S.; De Castro, L. N. Automatic sentiment analysis of twitter messages. *Proceedings of the 2012 4th International Conference on Computational Aspects of Social Networks, CASoN 2012*, n. March 2017, p. 52–57, 2012.
- 30 SARKER, A. et al. Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics*, Elsevier, v. 54, p. 202–212, 2015.
- 31 MADDEN, M. et al. Teens, social media, and privacy. *Pew Research Center*, v. 21, p. 2–86, 2013.
- 32 NAAMAN, M.; BOASE, J.; LAI, C.-H. Is it really about me?: message content in social awareness streams. In: ACM. *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. [S.l.], 2010. p. 189–192.
- 33 BASTIAN, M.; HEYMANN, S.; JACOMY, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. Disponível em: <<http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>>.
- 34 FRANÇA, T. C. et al. Big social data: Princípios sobre coleta, tratamento e análise de dados sociais. *XXIX Simpósio Brasileiro de Banco de Dados-SBBD*, v. 14.
- 35 ABRAHAMS, A. S. et al. An integrated text analytic framework for product defect discovery. *Production and Operations Management*, Wiley Online Library, v. 24, n. 6, p. 975–990, 2015.
- 36 FAN, W. et al. Tapping the power of text mining. *Communications of the ACM*, ACM, v. 49, n. 9, p. 76–82, 2006.
- 37 SANDHU, R.; SOOD, S. K. Scheduling of big data applications on distributed cloud based on qos parameters. *Cluster Computing*, Springer, v. 18, n. 2, p. 817–828, 2015.
- 38 ZHANG, Y.; GU, H. Text mining with application to academic libraries. In: *Computer Science for Environmental Engineering and EcoInformatics*. [S.l.]: Springer, 2011. p. 200–205.
- 39 PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: *LREc*. [S.l.: s.n.], 2010. v. 10, n. 2010.
- 40 KUMAR, A.; SEBASTIAN, T. M. Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, Citeseer, v. 9, n. 3, p. 372–378, 2012.
- 41 MITCHELL, A.; HOLCOMB, J.; PAGE, D. News use across social media platforms. *Washington, Pew Research Center*, 2013.
- 42 BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, Elsevier, v. 30, n. 1, p. 107–117, 1998.
- 43 CHA, M. et al. Measuring user influence in twitter: The million follower fallacy. *Icwsrm*, v. 10, n. 10-17, p. 30, 2010.
- 44 CHU, S. K.; DU, H. S. Social networking tools for academic libraries. *Journal of Librarianship and Information Science*, v. 45, n. 1, p. 64–75, 2012. ISSN 0961-0006.

- 45 BOSQUE, D. D.; LEIF, S. A.; SKARL, S. Libraries atwitter: trends in academic library tweeting. *Reference Services Review*, Emerald Group Publishing Limited, v. 40, n. 2, p. 199–213, 2012.
- 46 BOATENG, F.; LIU, Y. Q. Web 2.0 applications' usage and trends in top us academic libraries. *Library Hi Tech*, Emerald Group Publishing Limited, v. 32, n. 1, p. 120–138, 2014.
- 47 AL-DAIHANI, S. M.; ABRAHAMS, A. A text mining analysis of academic libraries' tweets. *Journal of Academic Librarianship*, Elsevier Inc., v. 42, n. 2, p. 135–143, 2016. ISSN 00991333. Disponível em: <<http://dx.doi.org/10.1016/j.acalib.2015.12.014>>.
- 48 RALSTON, M. R. et al. An exploration of the use of social media by surgical colleges. *International Journal of Surgery*, Elsevier, v. 12, n. 12, p. 1420–1427, 2014.
- 49 YOON, S.; ELHADAD, N.; BAKKEN, S. A practical approach for content mining of tweets. *American journal of preventive medicine*, Elsevier, v. 45, n. 1, p. 122–129, 2013.
- 50 HIEMSTRA, D. A probabilistic justification for using $tf \times idf$ term weighting in information retrieval. *International Journal on Digital Libraries*, Springer, v. 3, n. 2, p. 131–139, 2000.
- 51 NILSSON, N. J. Introduction to Machine Learning. *Machine Learning*, n. 2, p. 387–99. ISSN 10959572. Acessado em: 03.Fev.2016. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/21172442>>.
- 52 MONARD M., B. J. A. “conceitos sobre aprendizado de máquina – sistemas inteligentes fundamentos e aplicações. Manole Ltda, Barueri - SP, v. 1, p. 89–114, 2003.
- 53 MONTGOMERY, D. C.; RUNGER, G. C.; CALADO, V. *Estatística aplicada e probabilidade para engenheiros*. [S.l.]: Grupo Gen-LTC, 2000.
- 54 POLICARPO, R. A. C. S. A. Semantic classification of nouns. 2015.
- 55 CAMILO, C. O.; SILVA, J. C. da. *Mineração de dados: Conceitos, tarefas, métodos e ferramentas*. [S.l.], 2009. 1–29 p.
- 56 HSSINA ABDELKARIM MERBOUHA, H. E. B.; ERRITALI, M. *A comparative study of decision tree ID3 and C4.5*. 2014. Disponível em: <<http://dx.doi.org/10.14569/SpecialIssue.2014.040203>>.
- 57 QUINLAN, J. R. Induction of decision trees. *MACH. LEARN*, v. 1, p. 81–106, 1986.
- 58 KOHAVI, R.; QUINLAN, R. Decision tree discovery. In: *IN HANDBOOK OF DATA MINING AND KNOWLEDGE DISCOVERY*. [S.l.]: University Press, 1999. p. 267–276.
- 59 SIMÕES, A. C. D. A. Mineração de Dados baseada em Árvore de Decisão para Análise do Perfil de Contribuintes. *Discovery*, p. 32, 2008.
- 60 WITTEN, I. H.; FRANK, E. Data mining: Practical machine learning tools and techniques. *Elsevier, San Francisco*, v. 2 edition, 2005.
- 61 BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CRC press, 1984.

- 62 PINEDA, V. C. A Entropia Segundo Claude Shannon: o desenvolvimento do conceito fundamental da teoria da informação. *Dissertação Mestrado*, 2006.
- 63 HEATON, J. *Introduction to Neural Networks for Java*. [S.l.: s.n.], 2008. 440 p. ISBN 1604390085.
- 64 MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.
- 65 BARRETO, J. M. Jorge M. Barreto. 2002.
- 66 EGAN, J. P. Signal detection theory and roc analysis. New York, USA: Academic Press, 1975.
- 67 ANAESTHETIST, T. September 2001. Acessado em: 20.Jan.2015. Disponível em: <<http://www.anaesthetist.com/mnm/stats/roc/Findex.htm>>.
- 68 SOUZA, C. R. Análise de poder discriminativo atraés de curvas roc. July 2009. Acessado em 01.Set.2016. Disponível em: <<http://crsouza.com/2009/07/analise-de-poder-discriminativo-atraves-de-curvas-roc>>.
- 69 BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, v. 30, n. 7, p. 1145–1159, 1997. ISSN 00313203.
- 70 CORRÊA, F. et al. Aplicação de redes neurais artificiais no setor de transportes no brasil. Universidade Federal de São Carlos, 2008.
- 71 SHANTHI, S.; RAMANI, R. G. Feature relevance analysis and classification of road traffic accident data through data mining techniques. In: *Proceedings of the World Congress on Engineering and Computer Science*. [S.l.: s.n.], 2012. v. 1, p. 24–26.
- 72 FERREIRA, R. P. et al. Combinação de técnicas da inteligência artificial para previsão do comportamento do tráfego veicular urbano na cidade de são paulo. Universidade Nove de Julho, 2011.
- 73 BRAGA, M. d. L. et al. Planejamento de rotas com algoritmos anytime em redes veiculares na plataforma raspberry pi. *IV Simpósio Brasileiro de Engenharia de Sistemas Computacionais*, p. 1–6, 2014.
- 74 REIS, J. M. M. dos. Previsão em tempo real de condições de tráfego em redes veiculares. 2016.
- 75 CHONG, M. M.; ABRAHAM, A.; PAPRZYCKI, M. Traffic accident analysis using decision trees and neural networks. *arXiv preprint cs/0405050*, 2004.
- 76 AKGÜNGÖR, A. P.; DOĞAN, E.
- 77 KASHANI, A. T.; SHARIAT-MOHAYMANY, A.; RANJBARI, A. A data mining approach to identify key factors of traffic injury severity. *PROMET-Traffic&Transportation*, Fakultet prometnih znanosti Sveučilišta u Zagrebu, v. 23, n. 1, p. 11–17, 2011.

- 78 JADAAN, K. S.; AL-FAYYAD, M.; GAMMOH, H. F. Prediction of road traffic accidents in jordan using artificial neural network (ann). *Journal of Traffic and Logistics Engineering*, v. 2, n. 2, 2014.
- 79 BESHAH, T.; HILL, S. Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in ethiopia. In: *AAAI Spring Symposium: Artificial Intelligence for Development*. [S.l.: s.n.], 2010.
- 80 OLUTAYO, V.; ELUDIRE, A. Traffic accident analysis using decision trees and neural networks. *International Journal of Information Technology and Computer Science (IJITCS)*, v. 6, n. 2, p. 22, 2014.
- 81 THIANNIWET, T.; PHOSAARD, S.; PATTARA-ATIKOM, W. Classification of road traffic congestion levels from vehicle's moving patterns: A comparison between artificial neural network and decision tree algorithm. In: *Electronic Engineering and Computing Technology*. [S.l.]: Springer, 2010. p. 261–271.
- 82 LUCA, M. D. A comparison between prediction power of artificial neural networks and multivariate analysis in road safety management. *Transport*, Taylor & Francis, p. 1–7, 2015.
- 83 OÑA, J. de; MUJALLI, R. O.; CALVO, F. J. Analysis of traffic accident injury severity on spanish rural highways using bayesian networks. *Accident Analysis & Prevention*, Elsevier, v. 43, n. 1, p. 402–411, 2011.
- 84 DURDURAN, S. S. A decision making system to automatic recognize of traffic accidents on the basis of a gis platform. *Expert Systems with Applications*, Elsevier, v. 37, n. 12, p. 7729–7736, 2010.
- 85 COSTA, J. D. J.; BERNARDINI, F. C.; FILHO, J. V. A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. *AtoZ: novas práticas em informação e conhecimento*, v. 2, n. 2014, p. 1–26, 2015. Disponível em: <<http://ojs.c3sl.ufpr.br/ojs/index.php/atoz/rt/printerFriendly/41346/25356>>.
- 86 EAVES, D. The three laws of open government data. Acessado em: 24.Out.2016. Disponível em: <<http://eaves.ca/2009/09/30/three-law-of-open-government-data/>>.
- 87 BERTHOLD, M. R. et al. KNIME: The Konstanz Information Miner. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. [S.l.]: Springer, 2007. ISBN 978-3-540-78239-1. ISSN 1431-8814.
- 88 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Acessado em: 02.Fev.2016. Disponível em: <<https://www.R-project.org/>>.
- 89 ZEALAND, U. of W. N. Weka - university of waikato. Acessado em: 10.Fev.2016. Disponível em: <<http://www.cs.waikato.ac.nz/>>.
- 90 BERNARDINI, F. C. “combinação de classificadores simbólicos utlizando medidas de regras de conhecimento e algoritmos genéticos” - tese de doutorado. Instituto de Ciências e Matemática Computacional/USP, 2006.