



Programa de Pós-Graduação em Engenharia de Sistemas

DISSERTAÇÃO



Recife, 20 Abril de 2017.



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

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

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

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

Banca de examinadora:

Prof. Dr. Carmelo José A. Bastos Filho.....Engenharia de Sistemas/POLI/UPE

Recife, 20 Abril 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 das 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”, “Fuzzuboy”, “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, essa pesquisa teve por objetivo propor e testar conceitos para uma plataforma autoadaptável que contemple 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 causar constrangimentos, como retenção, redução do fluxo de tráfego (gargálos) e paralisação. A fonte primária de dados dessa 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 que reflete um bom grau de confiabilidade. Com os dados do Twitter foram coletados todos os tweets referentes a cada palavra chave, até o limite imposto pelo aplicativo. As tecnologias utilizadas foram Naïve Bayes, TF-IDF e, para exibir a geolocalização, utilizamos 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 the Metropolitan Region of some cities are constantly congested, not only by 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 aimed to 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 source of this research data 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 XX years, both from the PRF, and from people who mentioned events in BRs (accidents, stoppages, etc.) were also used. 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 limit imposed by the application. The technologies used were Naïve Bayes and TF-IDF and, to display geolocation, we used Google Maps georeferencing software. Compared to usual navigation approaches, the proposed prediction model represents a breakthrough in terms of mobility and management of transportation and vehicle traffic , since it makes it possible to anticipate events and behaviors, in order to favor the choice of alternative routes and increasing the time space of choice for certain routes.

Keywords: Data Mining, Data Bases, Social Network, Logistic, Routing

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>
TDM	<i>Text Data Mining</i>
KDD	<i>Knowledge Discovery Databases</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>
NB	<i>Naïve Bayes</i>
PB	<i>PetaByte</i>
EB	<i>ExaByte</i>
ZB	<i>ZettaByte</i>
YB	<i>YottaByte</i>

Lista de Figuras

2.1	Domínio das técnicas aplicadas a mineração de dados	19
2.2	O padrão CRISP-DM (1)	19
2.3	Entendimento do negócio	20
2.4	Entendimento dos dados	20
2.5	Preparação dos dados	21
2.6	Modelagem IA	22
2.7	Avaliação do modelo	22
2.8	Implantação do modelo	23
2.9	Fases da mineração de dados até extração do conhecimento	24
2.10	Mapa mental da Mineração em textos	27
2.11	Grafo de uma rede do Twitter	28
2.12	Descrição da conta Twitter das bibliotecas acadêmicas	32
2.13	Bibliotecas de Universidades e contas no Twitter	32
2.14	Frequência de palavras	33
2.15	Árvore de decisão	38
2.16	Árvore da família TDIDT	39
2.17	Perceptron de Rosenblatt	42
2.18	Rede ADALINE e MADALINE	42
2.19	Um arranjo de Perceptrons em camadas	43
2.20	Perceptron de McCulloch e Pitts	44
2.21	Perceptron e Adaline	45
2.22	Perceptron Multicamadas	46
2.23	Perceptron com Realimentação	46
2.24	Rede Kohone	47
2.25	Rede Neural	50
3.1	Etapas da modelo proposto	53
3.2	Matriz de Mortos 2D	55
3.3	Matriz de Gravidade 3D	56
3.4	56
3.5	O arco cibernético com o Twitter	57
3.6	Etapas da metodologia	57
4.1	Etapas da modelo proposto	58
4.2	Hora do acidente (1) — Concentração em torno da hora (2)	61
4.3	Frequência	61
4.4	Hora do acidente (1) — Concentração em torno da hora (2)	62
4.5	Frequência	62
4.6	Hora do acidente (1) — Concentração em torno da hora (2)	63

4.7	Frequência	63
4.8	Hora do acidente (1) — Concentração em torno da hora (2)	64
4.9	Frequência	64
4.10	Hora do acidente (1) — Concentração em torno da hora (2)	65
4.11	Frequência	65
4.12	Hora do acidente (1) — Concentração em torno da hora (2)	66
4.13	Frequência	66
4.14	Hora do acidente (1) — Concentração em torno da hora (2)	67
4.15	Frequência	67
4.16	Hora do acidente (1) — Concentração em torno da hora (2)	68
4.17	Frequência	68
4.18	Hora do acidente (1) — Concentração em torno da hora (2)	69
4.19	Frequência	69
4.20	Hora do acidente (1) — Concentração em torno da hora (2)	69
4.21	Frequência	70
4.22	Hora do acidente (1) — Concentração em torno da hora (2)	70
4.23	Frequência	71
4.24	Tipo de Veículo X Num. Acidentes	71
4.25	Traçado da via X Num. Acidentes	72
4.26	Árvore de Decisão gerada pelo Knime	75
4.27	Km 70, BR 101 (Sul) Pernambuco	77
4.28	Resultado da classificação feita pelo Naïve Bayes – acurácia	78
4.29	Resultado da classificação feita pela Rede Neural – acurácia	78
4.30	Gráfico de frequência de palavras – unigramas	80
4.31	Nuvem de palavras da Mineração em textos	80
4.32	Dendograma de Clusterização do resultado da mineração	81
4.33	Gráfico da clusterização do resultado da mineração	81
A.1	Etapas 1 – Coleta e união das bases históricas de acidentes e interdições . .	87
A.2	88

Lista de Tabelas

2.1	Mineração de dados – contexto de aplicação (1)	18
2.2	Volume de dados no mundo	27
2.3	Matriz de Confusão	48
2.4	Matriz modelo de Confusão	49
3.1	Variáveis transformadas	54
4.1	Detalhe da acurácia para classe Tipo Acidente	73
4.2	Matriz de confusão para a variável Tipo de acidente	73
4.3	Detalhe da acurácia para classe Gravidade	74
4.4	Matriz de confusão para a variável Gravidade	74
4.5	Detalhe da acurácia para classe BR	74
4.6	Matriz de confusão para a variável BRajustada	75
A.1	Variáveis originais da base de acidentes	86
A.2	Variáveis originais da base de interdições	87

Sumário

1	Introdução	14
1.1	Justificativa do problema	14
1.2	Motivação	15
1.3	Objetivo Geral	16
1.3.1	Objetivos Específicos	16
2	Revisão da Literatura	17
2.1	Introdução	17
2.2	Mineração de Dados e CRISP-DM	17
2.2.1	Contexto de aplicação do CRISP – DM	18
2.2.2	Ciclo de vida do CRISP-DM	19
2.3	Mineração de dados	24
2.4	Mineração de Textos	26
2.4.1	Mineração de Dados/Textos em Redes sociais	26
2.5	Aprendizagem de Máquina	35
2.5.1	Tipos de Aprendizagem	35
2.5.2	Algoritmos de Aprendizagem de Máquina	36
2.6	Medida de desempenho e qualidade aplicadas à mineração	48
2.6.1	Classificação e Regressão para análise preditivas	50
3	Contribuição	51
3.1	Modelo Proposto	52
3.2	Reflexão sobre as tecnologias utilizadas no modelo preditivo	53
3.3	Extração do conhecimento - KDD	54
3.4	Arco cibernético com dados do Twitter	56
3.5	Extrapolação para georreferenciamento	57
4	Simulação	58
4.1	Execução do modelo	58
4.2	A construção do Modelo preditivo	59
4.2.1	Aplicação do CRISP-DM	59
4.2.2	Dados encontrados antes da Mineração	61
4.2.3	Dados encontrados após a Mineração	72
4.2.4	Métrica dos classificadores	72
4.3	Acoplamento com a estrutura dinâmica	79
4.3.1	Mineração em texto no Twitter	79
5	Considerações finais	82
5.1	Trabalhos futuros	84

A Preprocessamento	85
A.1 Coleta e Preprocessamento dos dados da PRF	85
A.2 Preprocessamento dos dados do Twitter	88
Referências Bibliográficas	89

A dissertação

*“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 Justificativa do problema

O século XXI caracteriza-se como sendo a Era do crescimento exponencial da informação. Essas informações são produzidas tanto por seres humanos, quanto por máquinas. Segundo Norbert Wiener (2), a informação tem tanta importância quanto a energia e a matéria. Essa 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. 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” (3). 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. A Polícia Rodoviária Federal regista diariamente boletins de ocorrência, contudo, dados produzidos eletronicamente só estão disponíveis a partir de 2007.

A inteligência artificial é uma área que vai buscar essas informações e, com algoritmos eficientes, propor soluções inteligentes 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 a Inteligência Artificial algoritmos para descoberta de padrões e automatizar tarefas na investigação dos dados, essa automatização também conhecida como “Machine Learning” aplica-se a quase todos os caminhos na descoberta do conhecimento oferecida pela MD.

Uma instância do problemática descrita acima será tratada nesta pesquisa: o tráfego de veículos, transporte de mercadorias e locomoção nas rodovias. Para isso será necessária a integração de bases de dados heterogêneas disponíveis em computadores de órgãos públicos que contenham informações de qualidade para gerar um modelo preditivo de roteamento logístico de transporte. Para isso serão 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, serão 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.

Esta dissertação está organizada da seguinte forma:

No Capítulo 1 a Introdução ...

No Capítulo 2 é a Revisão da Literatura ...

No Capítulo 3 está nossa Contribuição ...

No Capítulo 4 a Simulação e execução do modelo proposto ...

No Capítulo 5 nossas Considerações Finais ...

Os anexos trazem os dados originais e

1.2 Motivação

As rodovias federais que atravessam a região metropolitana 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 (4).

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 (5). 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 (6). 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 frequentemente visto nas grandes cidades brasileiras. Por outro lado, câmeras de monitoramento de trânsito, redes sociais, aplicativos de celular e outros dispositivos, fornecem informações diárias sobre o que acontece nessas rodovias e no entorno delas, atualizando e alimentando bases de dados históricas, em repositórios espalhados pelos centros de monitoramento de trânsito, isso é conhecido como *Big data*.

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 absorver parte dessas informações requer várias etapas, que engloba algumas técnicas de mineração de dados. Propomos, nessa pesquisa, uma solução peculiar para utilização das rotas existentes, definida por critérios cientificamente estudados, que seja materializado num modelo de predição. Isso poderá ser de suma importância para solucionar a problemática do tráfego em rodovias, fornecendo toda informação que se faz necessária para que o veículo até seu destino de maneira segura e no menor tempo possível.

1.3 Objetivo Geral

Essa pesquisa teve 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 multidisciplinar através da integração de diversas tecnologias disponíveis, que vão desde a análise dos dados históricos das rodovias à utilização informações de redes sociais e dados governamentais.

1.3.1 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 com a dinâmica.
- Propor soluções para melhor experiência dos usuários que utilizam as rodovias pernambucanas.

2

Revisão da Literatura

2.1 Introdução

Nesse capítulo, foi realizada uma revisão teórica e de pesquisas contemplando três campos, a saber:

- O primeiro sobre o processo de aplicação da Mineração de Dados e Mineração em textos;
- O segundo relacionado às tecnologias mineração de dados com respeito à pesquisa em lide;
- Finalmente o último campo de pesquisa relacionado às tecnologias de mapeamento através de sistemas de posicionamento global aplicados ao sistema rodoviário.

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.

2.2.1 Contexto de aplicação do CRISP – DM

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

- O domínio da aplicação – a área específica que 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 como os 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 abaixo sumariza e exemplifica essas dimensões no contexto de aplicação do CRISP-DM.

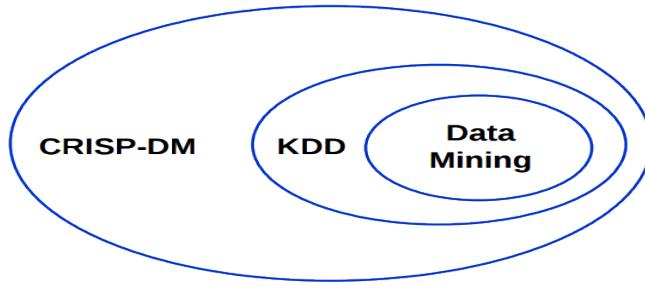
Tabela 2.1: Mineração de dados – contexto de aplicação (1)

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

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 (7), bem como, com outras técnicas. A análise desse processo permite extrair novos conhecimentos a partir dos dados, que é tratado na literatura como KDD – Knowledge Discovery Database (8). Fayyad 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 extrair 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 a seguir:

Figura 2.1: Domínio das técnicas aplicadas a mineração de dados



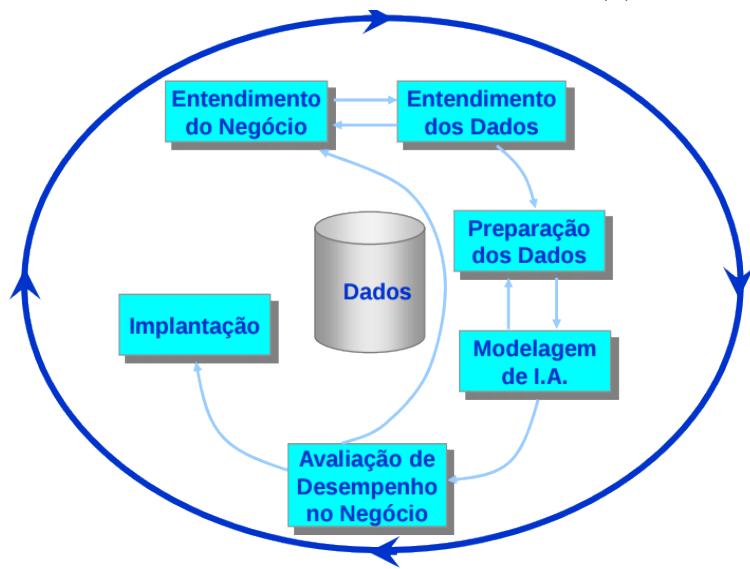
Fonte: Neurotech – 2012

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 a seguir ilustra as fases do ciclo:

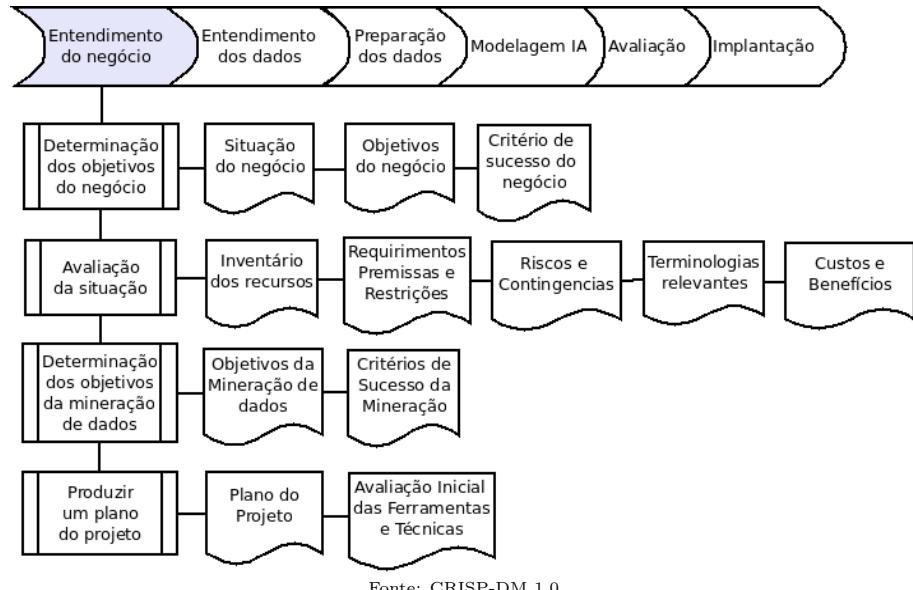
Figura 2.2: O padrão CRISP-DM (1)



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” (9) p.10 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.

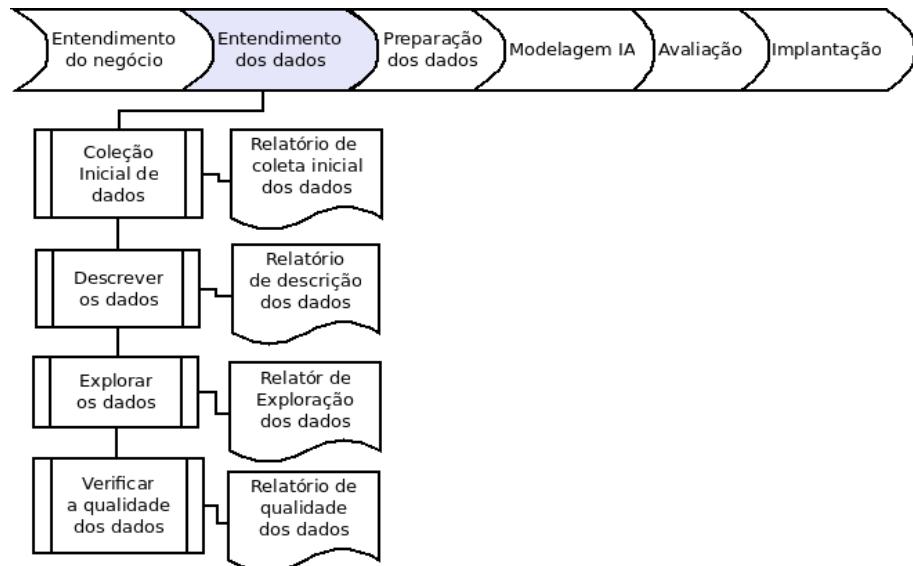
Figura 2.3: Entendimento do negócio



Fonte: CRISP-DM 1.0

Em seguida, o analista de dados passa à segunda fase, **Entendimento dos dados**. 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.

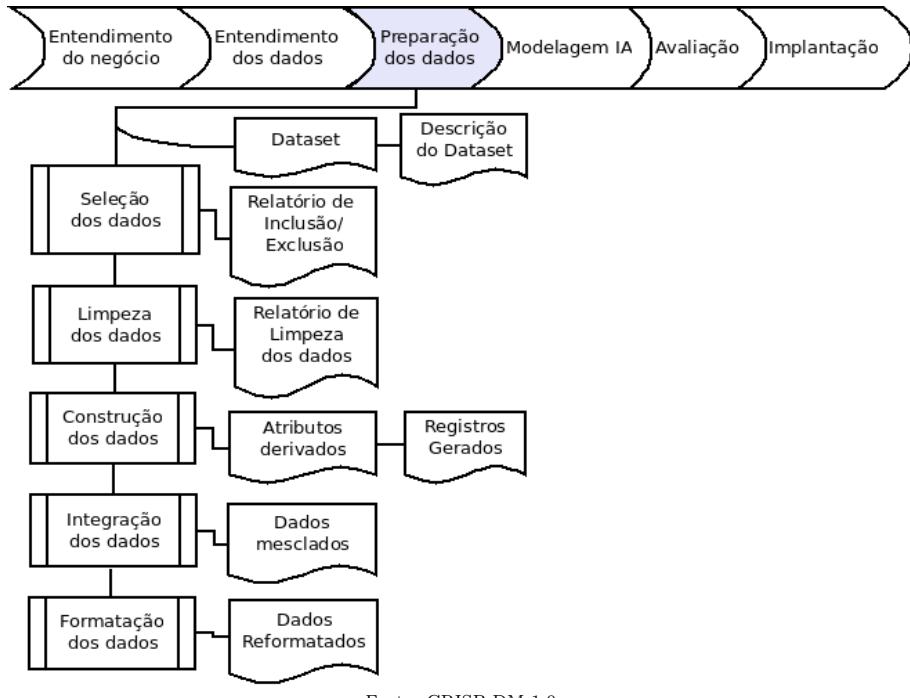
Figura 2.4: Entendimento dos dados



Fonte: CRISP-DM 1.0

A terceira fase, **Preparação dos dados**, 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 (10), 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

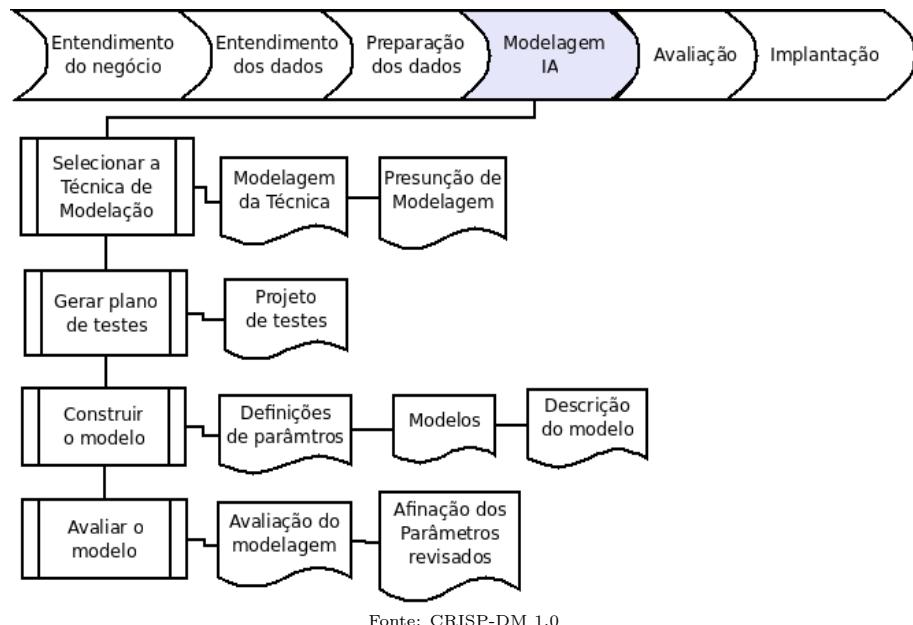


Fonte: CRISP-DM 1.0

Na quarta fase, **Modelagem de I.A.**, a tecnologia deve ser escolhida de forma criteriosa, baseada 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.

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

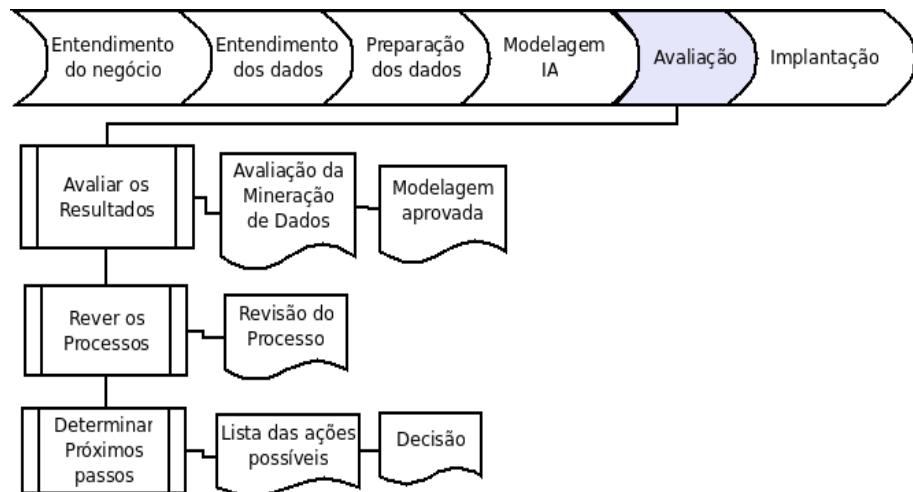
Figura 2.6: Modelagem IA



Fonte: CRISP-DM 1.0

Na fase cinco, **Avaliação de desempenho**, 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.

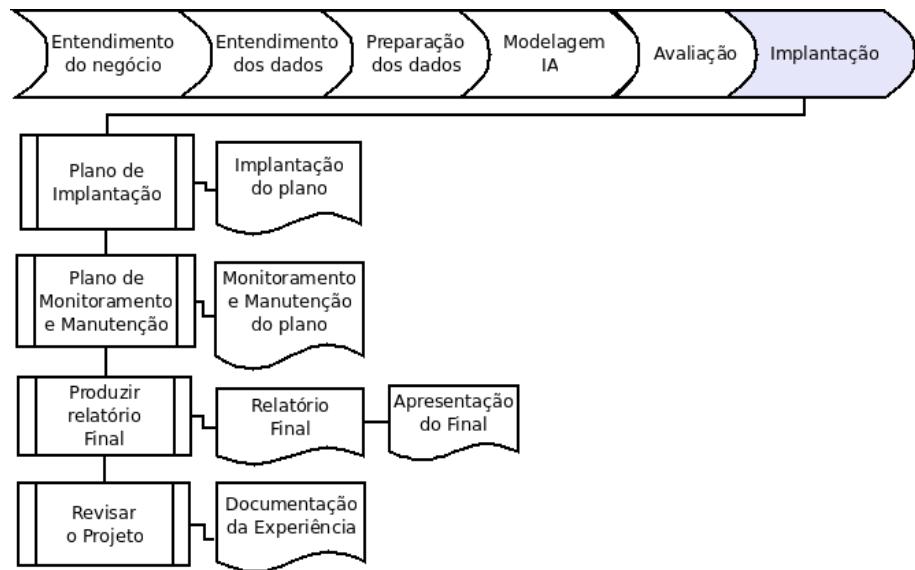
Figura 2.7: Avaliação do modelo



Fonte: CRISP-DM 1.0

A sexta e última fase caracteriza-se pela conclusão do modelo. 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

2.3 Mineração de dados

No processo de extração do conhecimento (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 (8).

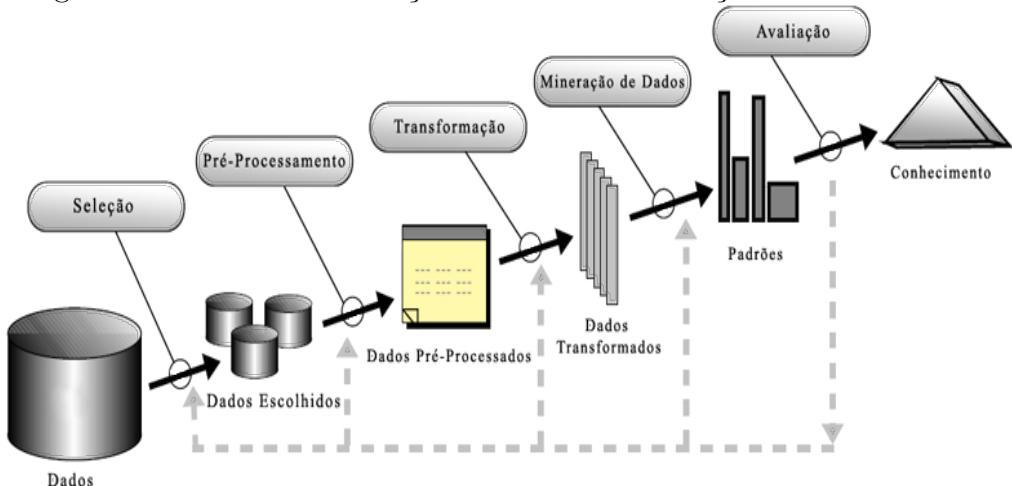
A mineração se distingue das técnicas estatísticas pelo fato de que não trabalha com dados hipotéticos, mas se apoia nos próprios dados para extrair os padrões (11).

FAYYAD (8), destaca que é necessário distinguir claramente KDD e mineração de dados. Enquanto que é um processo, a mineração 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 (8), pode conduzir à descoberta de dados sem significado e padrões inválidos.

Existem vários tipos de dados e informações nesses repositórios que podem ser minerais, 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”, podem existir inconsistências, faltas (*missing data*) ou duplicidade e erros de informações.

Nesse sentido, 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; quando dados numéricos ou com a moda; quando os dados forem categóricos. Para cada tipo de dados existem técnicas apropriadas para serem aplicadas sobre eles, algumas mais sensíveis às problemáticas elencadas anteriormente e outras mais robustas (12), que por sua vez estão associadas a classes de problemas que a mineração trata, a tabela 2.1 delineou o domínio. Isso será tratado na seção Aprendizagem de Máquina (Machine Learning). O caminho da extração dos dados até sua mineração e extração de conhecimento é longo. Na figura a seguir temos a ilustração desse caminho:

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”. Observa-se que este está repleto de *missing data* e/ou dados inconsistentes, conhecidos como dados não estruturados. O balão onde se lê “Selection” representa a coleta das

informações ou a seleção dos dados no *Big Data*. 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 “Target Data” 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 Botnets² 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 (13). 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 em estruturados, podendo ser armazenados em Bancos de Dados, conhecidos como Datawarehouse, por exemplo o Hive.

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 (14). As Redes Neurais, dependendo da quantidade de variáveis de entrada, puderão ter milhares de neurônios na camada intermediária, o que inviabilizaria essa metaheurística³.

Todas essas etapas descritas na figura são recorrentes, como indicam as setas pontilhadas que retornam aos passos anteriores. Utilizar técnicas de mineração de dados, além de extrair dados, extrai conhecimento, com isso pode-se predizer os resultados futuros na saída do modelo, quando determinados dados ocorrem na entrada (15), essa técnica de extração de conhecimento chama-se *Knowledge Discovery Databases* (KDD). O KDD utiliza métodos de Aprendizagem de Máquina para efetuar essa extração.

²Botnet é citado no sentido da coleta de informações

³Metaheurística são heurísticas aplicadas em problemas onde os custos computacionais não são tratáveis 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

2.4 Mineração de Textos

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

De acordo com Hotho, Nürnberg e Paass (16) a expressão *Data Mining* ou descoberta de conhecimento em textos foi referenda em 1995. Entretanto o interesse por extrair conhecimento oriundo de textos remonta à década de 60' (17). Hotho, Nürnberg e Paass (16) discutem, ainda, que é frequente a confusão de termos. Esses mesmos autores (16) na tentativa de definir *Text Mining* afirmam que é preciso considerar a perspectiva específica da área, definindo três possibilidades:

- A primeira perspectiva sugere que *Text Mining* corresponde à extração de fatos do texto, ou seja, extração de informação;
- Uma segunda 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;
- A terceira e última 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' (18).

As primeiras aplicações computacionais relacionadas à linguagem natural aparecem em torno de 1946 (18), decorrente da expertise de Alan Turing para quebra de códigos inimigos durante a segunda guerra mundial.

Ainda segundo Liddy (18) 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 em Redes Sociais.

Minerar dados em texto é um processo que deve ser dividido em várias etapas (19). Minerar em redes sociais exige ao menos mais uma etapa: a escolha de uma rede social, cada uma delas com sua tecnologia própria.

A Mineração em textos é inspirada em técnicas de *Machine Learning* (13). Todavia, analisar textos é basicamente entender o seu significado, baseado em regras de associação lógica. O mapa mental a seguir mostra um modelo de análise de texto feito por seres humanos.

2.4.1 Mineração de Dados/Textos em Redes sociais

Introdução ao estudo das 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 permitam, 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

Figura 2.10: Mapa mental da Mineração em textos



Fonte: o autor

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.

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 essas escolhas relativas as áreas urbanas, bem como a centenas de quilômetros adiante, pelo interior de um estado, por exemplo, cada vez mais as pessoas não tomam qualquer decisão sem antes consultar aplicativos e redes sociais tais como o waze, twitter, facebook, ou até mesmo dispõem, em seus aparelhos celulares, 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, como tanto já referimos nesse trabalho, 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).

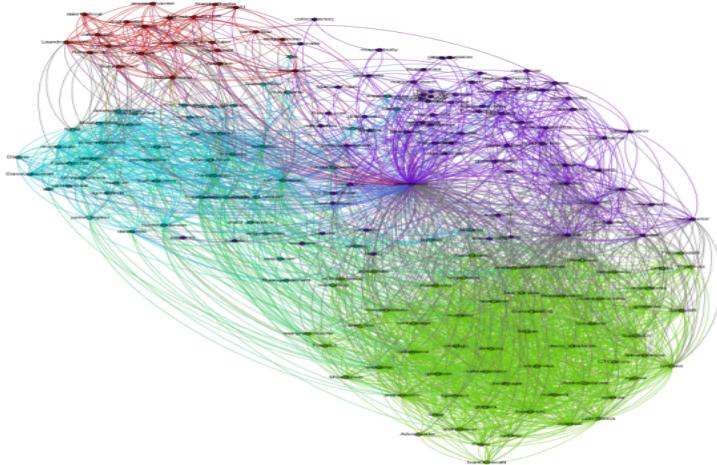
"Em 2010 empresas e usuários armazenaram mais de 13 exabytes de novos dados"(20).

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 seguir foi gerada pelo software Gephi (21) 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 e o Facebook estão entre as mais populares ferramentas de mídias sociais do público em geral. Uma das características-chave dessas ferramentas é que elas habilitam uma comunicação em dois sentidos e interação entre usuários. A natureza do diálogo tipicamente envolve um tópico específico, muitas vezes relacionados a acontecimentos que têm influência direta na vida das pessoas ou que chamam sua atenção, como eventos de cunho político, catástrofes naturais, acidentes com vítimas graves, atentados, dentre outros.

O Twitter, rede social que interessa a essa pesquisa, 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 (22). 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 (23). É comum, ainda, aplicar mineração de textos em bibliotecas e outras instituições. Isso implica em rastrear tópicos, extrair informações, agrupar, categorizar (24).

Em recente artigo, Sandhu (25) 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 (26) alegaram que minerar conhecimento sobre os clientes é importante para as bibliotecas acadêmicas. Na mesma linha de investigação, Sarker et al. (27) 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 (28) (29).

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 (30).

Um relatório publicado em 2013 revelou que o Twitter aparecia 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 (31).

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.

No contexto da Bibliometria e da Cientometria, alguns estudos destacam a utilização da contagem de citações no twitter como o objetivo de avaliar qual o impacto que uma publicação alcança no público leitor, bem como em centros e instituições de pesquisa (33), (34). Nos estudos em questão, pesquisadores examinaram a relação entre características dos autores e grupo de autores e o impacto da produtividade deles. Foi medido o número de publicações produzidas e o número de citações recebidas.

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* (35), em decorrência do número de seguidores, da quantidade de menções feitas, de “retweets”, por exemplo.

Cha, Haddadi e Benevenuto (36) 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.

Na pesquisa conduzida por Suh, Hong e Chi (37), esses pesquisadores encontraram uma correlação positiva entre a existência de uma URL de um “tweet” e a probabilidade de que aconteça um “retweet”. Os autores optaram por uma abordagem indutiva, utilizando-a da seguinte maneira: coletaram 1200 “tweets” recentes, a partir da conta de uma Universidade, considerando todos os membros da Association of American Universities (AAU). A coleta e processamento dos dados deu-se com a utilização do Twitter API com o twitter4j (biblioteca Java), tendo sido adicionados a um código java desenvolvido pelo autor da pesquisa. Tomou-se uma amostra do percentual da Academia.

Procederam com o preprocessamento dos dados, de modo a preparar a análise. Nessa etapa, os “retweets” foram retirados das amostras. Também foram removidos da amostra os “tweets” com conteúdo pouco significativo para a pesquisa como, por exemplo, breves agradecimentos (e.g. “welcome”), pequenos comentários ou “gírias” de algum “tweet” (e.g. “lol”), encorajamentos pontuais (e.g. “keep going”) e conversas pessoais, sem relação com o conteúdo dos “tweets”. Com isto, a amostra foi reduzida de 1200 para 752 “tweets” que, em seguida, foram distribuídos em nove categorias. Os resultados apontaram que, dos 752 “tweets” analisados, 271 apresentavam pelo menos um “retweet”, e 131 receberam um “favorito”. Em média, “tweets” recebidos 0,67 (desvio padrão “SD” = 1,4) “retweets” e 0,23 (SD=0,6) favoritos. Adicionalmente, em média, “tweets” incluídos 0,47 (SD=0,51) URLs, 0,61 (SD=0,91) menções de usuários e 0,04 (SD=0,2) media de entidades. Em média o tamanho dos “tweets” foi 107 (SD=29,81) caracteres.

Na média, foram enviados, nas seis contas de Twitter das bibliotecas analisadas, cerca de 1817 (SD=1126) “tweets”, seguido 1062 (SD=641) usuários, estes seguidos por 2006 (SD=788) usuários, e estes 1503 (SD=450) dias (duração). O teste Shapiro-Wilk mostrou que nove característica de “tweets” normalmente distribuídas. Métodos não paramétricos foram usados para examinar a relação entre um tweet as características da conta. O teste Krusal-Walls apresentou uma diferença significativa entre o conteúdo da categoria do tweet para o número de favoritos $X = 15.11, df = 8, p < 0,057$. O teste Spearman, por sua vez, demonstrou haver uma correlação negativa entre o número de “retweets” e o número de menções a usuários, sugerindo que “tweets” com conexões pessoais podem ter pouco valor de ‘uso geral’, de maneira que os usuários demonstram certa relutância em retweetear conteúdos advindos desses seguidores. O teste Spearman encontrou, ainda, uma pequena correlação positiva entre o número de favoritos e o número de usuários seguidos, bem como uma baixa correlação negativa entre o número de favoritos e tempo de uso da conta (desde o cadastro).

Em outro artigo, conduzido por Chu & Du (38), 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)⁴. Tal prática têm

⁴Facilitate information and knowledge sharing, service enhancement and promotion, interaction with

promovido considerável mudança na interação com usuários e relacionamento com os clientes (39), tendo sido utilizada frequentemente como alternativa para estabelecer uma conexão personalizada com os seus usuários (40).

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 (41). 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.

As informações relevantes dos “tweets”, utilizadas para a pesquisa eram: data do tweet; número de vezes em que o tweet foi marcado como favorito por outro usuário; número de vezes em que houve um re-tweet, ou seja, em que ele foi passado para frente; data que se “juntou” ao Twitter.

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 (42) e de Yoon, Elhadad e Bakken (43). 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 “tweets” marcados com o PamTaT⁵.

O PamTaT é baseado na interface do Microsoft Excel para Python nltk – “natural language processing framework” e permite a análise de grande volume de textos pelos usuários finais, não necessitando de conhecimento de programação da linguagem Python. A ferramenta serve, ainda, 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.

Foi utilizado o Harvard General Inquirer (17) para análise semântica e sentimental dos “tweets”. Essa ferramenta de análise de textos permite ao usuário final repostar a frequência de categorias de palavras diferentes usadas no texto fonte. Aplicações reportadas no “General Inquirer” para diferentes textos-fontes identificaram duas centenas de palavras incluindo, dentre elas, palavras que davam o sentido de algo positivo, negativo, ou ainda relacionadas a vontade (prazer), relacionadas a dor, relacionadas a localização, relacionadas a hora (tempo), relacionadas à Academia, relacionadas a exagero (overstatement) ou subavaliação (understatement), e assim por diante.

Hurwitz forneceu uma lista abrangente de categorias de palavras reconhecidas pela student library users, at minimal costs.

⁵Ferramenta “text mining” desenvolvida por Pamplin Collage do Instituto Politécnico de Negócios de Virgínia da Universidade Estadual de Virgínia (Bird, Loper & Klien, 2009)

Harvard General Inquirer e apresentou uma lista completa de palavras específicas que pertenciam a cada categoria de palavras.

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

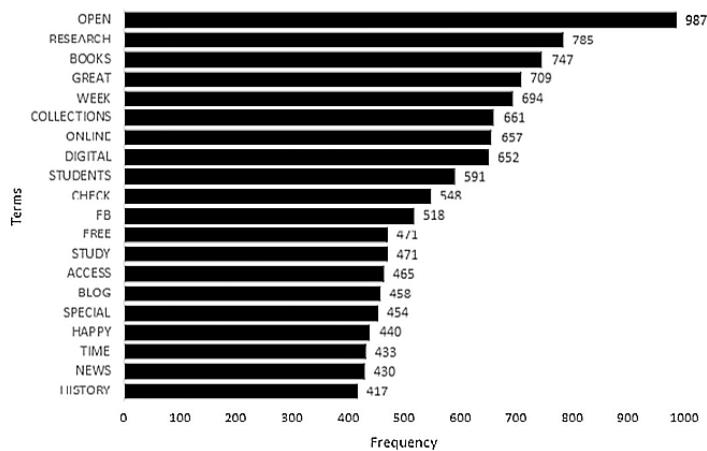
Observa-se que foi incluído o número de “tweets”; a quantidade de usuários seguidos pela biblioteca; a quantidade de seguidores e o número de “tweets” favoritos da biblioteca. A Universidade Johns Hopkins, como se pode observer na tabela, tem o maior número de “tweets”, seguida pela biblioteca da Stanford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

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

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:

The image shows five tweets from the official account of the Federal Highway Police of Santa Catarina (@PRF191SC). The tweets are as follows:

- PRF Santa Catarina @PRF191SC · 13 h Na BR-101 Norte, pistas livres de Biguaçu a Joinville. Entre Itajaí e Balneário Camboriú, tráfego intenso, mas sem formação de filas.
- PRF Santa Catarina @PRF191SC · 13 h Na BR-101 Sul, trânsito fluindo sem retenções entre Araranguá e Palhoça. Ponte de Laguna livre nos dois sentidos.
- PRF Santa Catarina @PRF191SC · 13 h De Biguaçu a Florianópolis, BR-101 tem trânsito intenso, e velocidade média de 45 km/h. Nas vias marginais, trânsito flui com
- PRF Santa Catarina @PRF191SC · 14 h De Palhoça a Florianópolis, BR-101 com trânsito intenso e lentidão habitual. Já na Via Expressa (BR-282), velocidade média de 20 km/h.
- PRF Santa Catarina @PRF191SC · 14 h Bom dia, Santa Catarina! Vai pegar estrada nesta terça-feira? Acompanhe @PRF191SC e saiba em primeira mão sobre as condições das rodovias.

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 (44). 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 (45), o termo Aprendizagem de Máquina refere-se à detecção automatizada de Padrões de dados.

Para Nilsson (44), 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” (44).

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 (10). 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 (46) se caracteriza pelo acesso ao conjunto de exemplos de treinamento pelo algoritmo de aprendizagem, também conhecido como indutor, de modo que haja especificação da saída desejada. No caso da aprendizagem não-supervisionada (10), os valores de entrada são estabelecidos, mas não são definidos os valores de saída. O indutor terá o papel de estabelecer aproximações, propondo agrupamentos (clusters) em função de determinadas categorias como, por exemplo, similaridade (46).

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 (12), essas funcionalidades tem 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 (12). A seção seguinte analisa as tarefas preditivas.

2.5.2 Algoritmos de Aprendizagem de Máquina

Naïve Bayes

Dentre os algoritmos de machine learning que destacaremos nessa dissertação está o Naïve Bayes. Essa classe de algoritmos é baseado no teorema da probabilidade condicional de Bayes (47), 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 predições.

Um classificador Naïve Bayes produz probabilidade como output (48). 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 (48).

Em mineração de dados variáveis independentes explicam a variável dependente para fazer predição. 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 $p(C_k|x_i, \dots, x[n])$ para cada K possível ter vindo da classe $C[k]$. Aplicando o teorema de Bayes da probabilidade condicional temos:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} \quad (2.1)$$

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:

$$p(\text{posteriori}) = \frac{p(\text{priori}) * \text{verossimilhana}}{\text{evidencia}} \quad (2.2)$$

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:

$$P(h_i|d) = \sum_i P(X|d, h_i)P(h_i|d) = \sum_i P(X|h_i)P(h_i|d) \quad (2.3)$$

onde cada hipótese h_i determina uma distribuição de probabilidades sobre a variável X (10).

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 (48).

Árvores de Decisão

Aspectos introdutórios

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, sobremaneira, em pesquisas relacionadas à Medicina, à Economia, nos mais diversos sistemas de suporte à decisão, como diagnósticos de doenças, investigação de fraudes, dentre outros (49).

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.

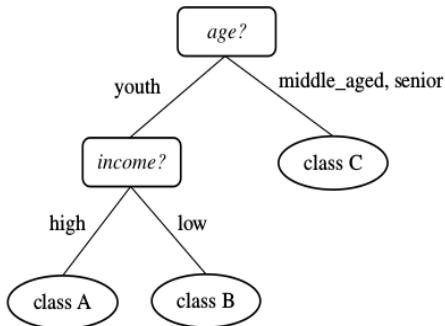
Conforme o nome sugere, o que se espera de uma árvore de decisão é que ao final do processo resulte a melhor decisão. Todavia, para que a decisão satisfatória apareça, é necessário que o pesquisador faça as escolhas adequadas, de modo a poder tirar o melhor proveito dessa ferramenta. Para tal, é preciso 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 (10).

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 (49).

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

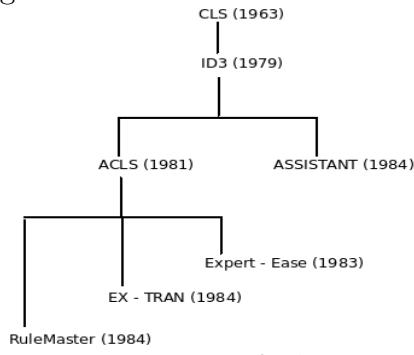
Tal interesse, ainda segundo esse autor, instigou a busca pelo desenvolvimento de sistemas competentes baseados no conhecimento (knowledge-based) e tomou vulto nas pesquisas em inteligência artificial. Quinlan (51) avança, pontuando o interesse de muitos pesquisadores, dos quais ele próprio, no que ele 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). ⁶

O primeiro grupo de árvores de decisão, ainda conforme Quinlan (51) era responsável

⁶This 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.

por tarefas de classificação, desenvolvendo uma decisão das raízes às folhas, sendo conhecido com *Top – Down*.⁷ 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 (51), 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 (51), 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 ACLS, que não estão relacionados a avanços teóricos consideráveis, mas incorporaram 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, (51) 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 (52) 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

⁷Em 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.

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 algoritmos, 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” (53). 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 (52), 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 (54), 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 (55). 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 (55), 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 (56) 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 (10) 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 (53). A seguir a equação que representa o cálculo padrão da entropia (já mencionada na seção 2.2.2)

$$H_x = - \sum_{\forall x \in X} P(x) \log_2 P(x) \quad (2.4)$$

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 (14):

$$H_{Y|X} = \sum_x P(x) H(Y|X=x) = - \sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y|x) \log_2 P(y|x) \quad (2.5)$$

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

Redes Neurais

Introdução

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 (10).

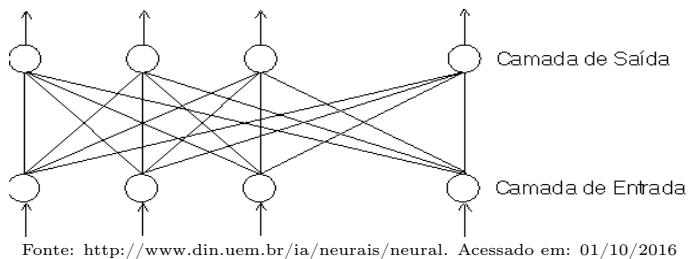
De forma semelhante ao funcionamento das Redes Neurais naturais, foram desenvolvidas as Redes Neurais Artificiais, que recebem esse nome por se caracterizarem como um sistema cujo funcionamento é semelhante à arquitetura das redes neurais humanas.

Nesse contexto, em que se pretende criar modelos computacionais com funcionamento semelhante ao modelo neurológico humano, surge a chamada neurocomputação. No início da década de 40, precisamente em 1943, McCulloch e Pitts (57) propuseram um modelo simplificado de funcionamento do cérebro humano e, a partir daí sugeriram a construção de uma máquina que fosse inspirada nesse funcionamento. A partir da proposição de McCulloch e Pitts, vários trabalhos começaram a ser desenvolvidos, tomando o funcionamento do cérebro humano como modelo.

Em 1949, Hebb explicitou matematicamente as sinapses dos neurônios humanos. Dois anos depois, em 1951, o primeiro neurocomputador, chamado Snark, foi desenvolvido por Marvin Minsky. Todavia, o primeiro neurocomputador que obteve sucesso surgiu entre 1957 e 1958, o Mark I Perceptron, criado por Rosenblatt, Wightman e colaboradores (57). O interesse principal desses pesquisadores era o de desenvolver, com esse neurocomputador, a capacidade de reconhecimento de padrões. Nesse contexto, os estudos na área se aprofundaram de tal forma, que muitos consideram Rosenblatt como o fundador da neurocomputação, tal qual encontramos hoje.

A figura a seguir ilustra, de maneira simplificada, a Rede de Perceptrons, conforme proposta de Rosenblatt.

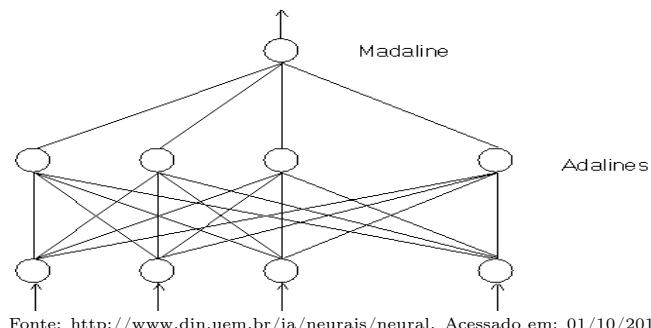
Figura 2.17: Perceptron de Rosenblatt



Fonte: <http://www.din.uem.br/ia/neurais/neural>. Acessado em: 01/10/2016

Dando continuidade e indo mais além dos trabalhos de Rosenblatt e seus colaboradores, Widrow desenvolveu, em conjunto com alguns alunos, o Adaline, um tipo de processamento de redes neurais dotado de uma potente lei de aprendizado, em uso ainda nos dias atuais, que pode ser representado pela figura a seguir:

Figura 2.18: Rede ADALINE e MADALINE



Fonte: <http://www.din.uem.br/ia/neurais/neural>. Acessado em: 01/10/2016

Muitos estudos foram realizados nas décadas seguintes, mas o que marcou esse período foram as elucubrações sobre o desenvolvimento de máquinas tão potentes quanto o cérebro humano, muito mais do que a publicação de pesquisas realmente contundentes na área.

A partir dos anos 80, a pesquisa em neurocomputação deu outro grande salto qualitativo e em 1987 teve lugar, em São Francisco a primeira conferência de redes neurais nos tempos modernos: a IEEE International Conference on Neural Networks, sendo também formada a International Conference on Neural Networks Society (INNS). Em 1989 foi fundado o INNS Journal, e em 1990 o Neural Computation e IEEE Transations on Neural Networks.

Definições e funcionamento de uma Rede Neural Artificial

São várias as definições que podem ser encontradas sobre o que vem a ser uma Rede Neural Artificial (RNA) (11), 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 aprendizado, generalização, associação e abstração.

Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento, enquanto que o cérebro de um mamífero pode ter muitos bilhões de neurônios. Uma Rede Neural Artificial (RNA) é um sistema que de neurônios que estabelecem conexões sinápticas, que possuem neurônios de entrada, que recebem os estímulos provenientes

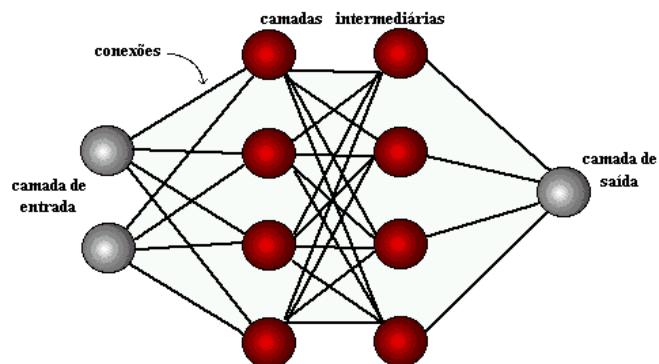
do meio exterior, os neurônios internos ou hidden (neurônios ocultos) e os neurônios de saída, que se comunicam com o mundo externo (58).

Cabe destacar que, de acordo com esse modelo, os neurônios internos têm considerável importância nesse processo, uma vez que são responsáveis pela resolução de problemas linearmente inseparáveis. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede.

Os neurônios, nessas redes são conhecidos como Perceptrons. O arranjo em camadas desses perceptrons é chamado *Multilayer Perceptron*. O *multilayer perceptron* é responsável pela resolução de problemas mais complexos que não seriam passíveis de resolução pelo modelo de neurônio básico. Para aprender os perceptrons tem que estar dispostos em camadas, um único perceptron pode realizar algumas operações do tipo XOR, contudo seria incapaz de aprendê-la.

A figura a seguir apresenta o arranjo dos perceptrons em camadas, conforme discutido anteriormente.

Figura 2.19: Um arranjo de Perceptrons em camadas



Fonte: <http://www.din.uem.br/ia/neurais/neural>. Acessado em: 01/10/2016

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

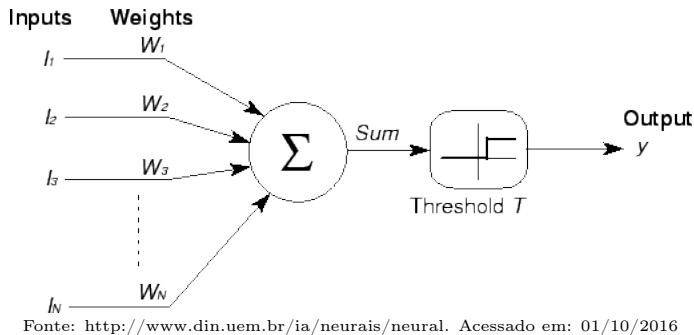
- Camada de entrada: nessa camada apresenta-se os padrões à rede;
- Camadas Intermediárias (ocultas): aqui é realizada a maior parte do processamento por conexões ponderadas. São consideradas como as camadas extratoras de características;
- Camada de saída: responsável pela conclusão e apresentação do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

O comportamento inteligente de uma Rede Neural Artificial vem das múltiplas interações existentes entre as unidades de processamento dessa rede. As unidades de processamento são conectadas por canais de comunicação, associados a determinado peso, como mostra a figura a seguir, proposta por McCulloch e Pitts (59)

Esse 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; fase de treinamento. Uma vez que o erro tornar-se infinitamente pequeno dizemos que a rede neural “não aprende mais” há uma degradação do sistema, o “overfitting”.

Da figura acima podemos extrair duas coisas:

Figura 2.20: Perceptron de McCulloch e Pitts



- A função calculada y é uma função **discriminativa** (classificação) com $y = 0$ e $y = 1$.
 - $net = x_1w_1 + x_2w_2 + \dots + x_iw_i + x_dw_d = \sum x_i.w_i = W.X = W^T X$
 - onde $w = \{+1, -1\}$
 - saída $y = f(net)$
 - $f(net) = \begin{cases} 1, & \text{se } net \geq \mu \\ 0, & \text{se } net < \mu \end{cases}$
- Fronteira de Decisão:
 - Determina o ponto que separa os dados que vêm de $-\infty$ e $+\infty$
 - Argumento de $f(net)$ é igual a zero: $\sum w_i x_i - \mu \Rightarrow w.x = 0$

O modelo McCulloch e Pitts (59) leva em conta cinco hipóteses fundamentais, a saber:

1. A atividade de um neurônio é binária. Isso quer dizer que os neurônios respondem a valores **verdadeiro** ou **falso** ou 0 ou 1;
2. As RNA são formadas por linhas direcionadas, que são inspiradas em sinapses, e que ligam os neurônios. Tais linhas podem ser positivas (excitatórias) ou negativas (inibitórias);
3. Os neurônios, numa RNA, têm um limiar fixo, nomeado como L . Isso posto, o processo só é disparado se a entrada for igual ou maior que esse limiar;
4. Uma única sinapse inibitória evita, por completo, o disparo do neurônio, ainda que venham, ao mesmo tempo, várias sinapses excitatórias;
5. A quinta e última hipótese propõe que cada sinal leva determinada unidade de tempo para “passar” de um neurônio a outro.

Uma rede neural passa por um processo 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. Isso faz com que as RNA tenham uma característica diferente da computação programada, que exige um conjunto de regras pré-fixadas e algoritmos. A Rede Neural, por sua vez, extrai regras básicas a partir de dados reais, ou seja, aprendem através de exemplos. Uma vez “treinada” os pesos estão calibrados para solucionar a classe de problemas para o qual foi desenhada. Essa rede neural portanto pode ser considerado um aproximador de funções, uma vez dada uma série de “inputs” ela poderá produzir um “output” baseada nas funções de internas.

Aplicações e Tipos de Redes Neurais

São várias as aplicações das redes neurais. Elas podem ser utilizadas para reconhecimento e classificação de padrões; processamento de sinais e de imagens; identificação

e controle de sistemas; predição, dentre outras funções. No caso específico desse estudo, nosso interesse está centrado, fundamentalmente, na predição.

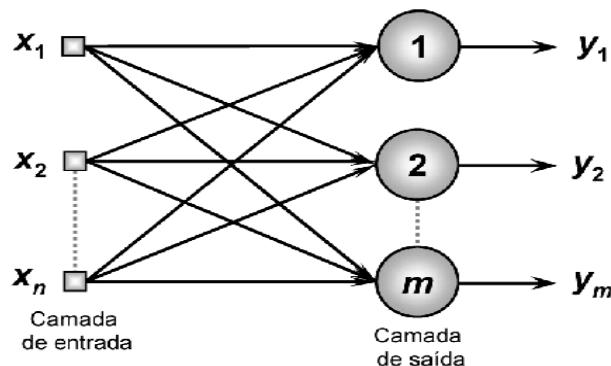
Para o desenvolvimento de uma rede neural algumas importantes fases precisam ser consideradas. Em primeiro lugar, é necessário um estudo detalhado do problema, para que possam ser feitas as escolhas adequadas à sua resolução. Em seguida, passa-se à fase de desenvolvimento do modelo neural, a partir de neurônios biológicos, e das estruturas e conexões sinápticas. A etapa seguinte implica na escolha de um algoritmo de aprendizado de regras, com ajuste de pesos ou forças de conexões intermodais, e de um conjunto de treinamento. Passa-se, então, à fase de treinamento, propriamente dita, aos testes e, por fim, utilização da rede neural.

Uma vez que existem distintas possibilidades de aplicação e desenvolvimento de uma RNA, existem, igualmente, diversas maneiras de classificá-las (60). Trataremos de algumas dessas nesse tópico.

- (a) Quanto à sua arquitetura: estática, dinâmica ou fuzzy; de única camada/camadas simples ou de múltiplas camadas. Exemplos de RNA de uma (i) ou múltiplas (ii) camadas:

- (i) São exemplos de redes Perceptron e Adaline

Figura 2.21: Perceptron e Adaline



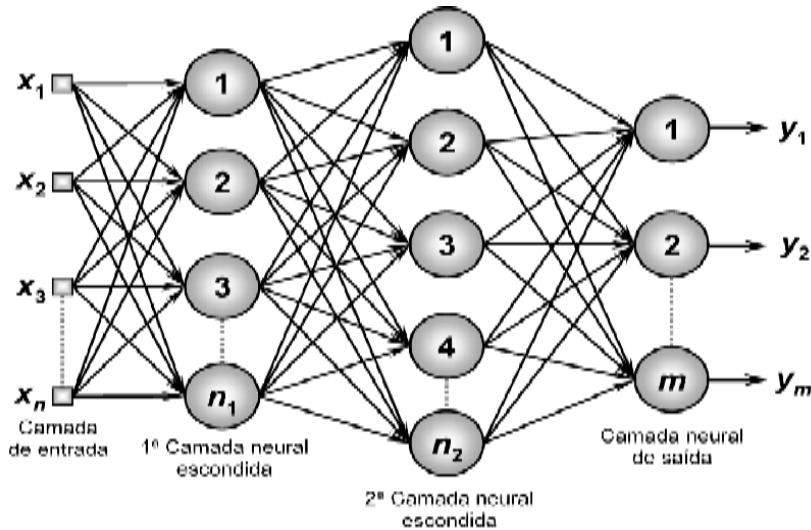
Fonte: <http://www.dkriesel.com>. Acessado em: 10/10/2016

- (ii) São exemplo dessas redes as de Perceptron multicamadas (PMC/MLP) e redes de base radial (RBF)

Alguns autores ainda fazem referência às redes recorrentes ou realimentadas (iii), como a rede de Hopfield e a Perceptron multicamadas (61). Tais redes, segundo esses autores, são ideais para processamento dinâmico, como previsão de séries temporais, controle de processos, etc. Referem também a existência das redes com estrutura reticulada (iv), como a rede de Kohonen.

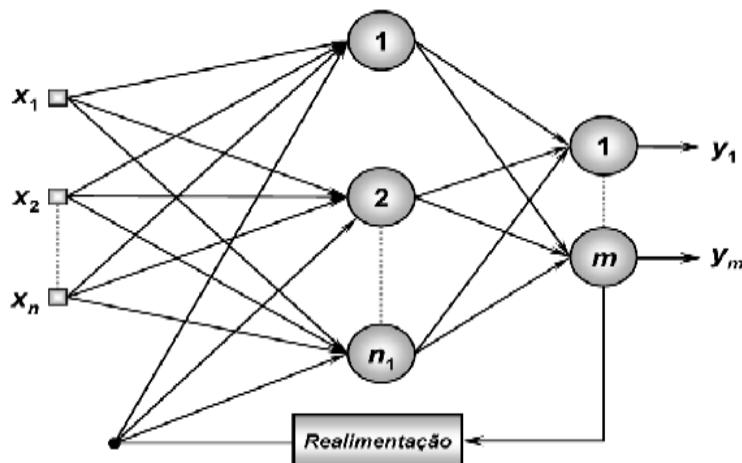
- (iii) Exemplo de rede recorrente ou realimentada: “backpropagation”
 - (iv) Exemplo de redes de estrutura reticulada

Figura 2.22: Perceptron Multicamadas



Fonte: <http://www.dkriesel.com>. Acessado em: 10/10/2016

Figura 2.23: Perceptron com Realimentação



Fonte: <http://www.dkriesel.com>. Acessado em: 10/10/2016

- (b) Quanto às conexões, os tipos de RNA são: no sentido de ida; no sentido de ida e volta; lateralmente conectadas; topologicamente ordenadas; híbridas.
- (c) Quanto à aplicação: reconhecimento de padrões e classificação; processamento de imagem e visão; identificação de sistema e controle; processamento de sinais.

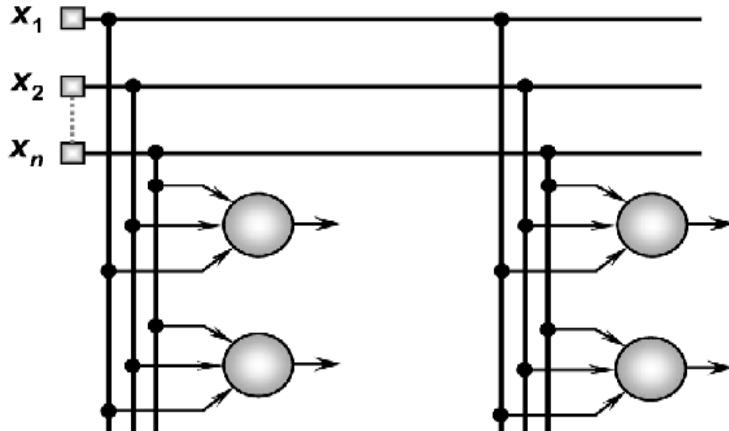
Aprendizado em Redes Neurais

O aprendizado caracteriza-se pela capacidade que a RNA tem de resolver uma determinada classe de problemas para os quais foi destinado seu desenvolvimento. Nesse processo, é proposto um algoritmo de aprendizado, que se caracteriza como um conjunto de regras bem delineadas que permitirão a resolução do problema.

O aprendizado das RNA resulta da fase de treinamento, através de um processo iterativo de ajuste dos pesos. O conhecimento é armazenado nas sinapses, ou seja, nos pesos que são atribuídos às conexões que existem entre os neurônios da rede.

- Por independência de quem aprende: o aprendizado pode ser por memorização, por

Figura 2.24: Rede Kohone



Fonte: <http://www.dkriesel.com>. Acessado em: 10/10/2016

contato, por exemplos, por analogias, por exploração, por descoberta, sobretudo por uma mistura entre os últimos três (62).

- Por retroação do mundo: quando há uma realimentação específica vinda do mundo exterior, podendo ser supervisionado ou não-supervisionado.

O treinamento supervisionado tem sido o mais comumente utilizado em redes neurais (62). 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.

O aprendizado de uma rede se dá em um determinado tempo e a partir de determinado padrão. A velocidade de aprendizado depende de variáveis que precisam ser consideradas como, por exemplo, a complexidade da rede proposta, a quantidade de camadas que ela possui, a arquitetura adotada, o algoritmo que foi utilizado, a precisão esperada. É preciso que se esteja atento a todos esses elementos, uma vez que dependendo de como essas variáveis forem consideradas, o treinamento pode se estender por um período bastante longo e, nem sempre apresentando um resultado satisfatório.

Quanto aos algoritmos de aprendizagem, são muitos os que podem ser utilizados, mas boa parte deles são variações do princípio de Hebb, que é a regra mais utilizada (62). A descrição dessa regra foi apresentada pelo seu proposito, Donald Hebb, em 1949. A ideia básica dessa regra é a de que quando um neurônio recebe um entrada a partir de outro neurônio, isso significa que ambos estão ativos e que os pesos entre os neurônios precisam ser excitados.

Uma segunda lei utilizada é a de Hopfield. Ela se inspira no princípio de Hebb, mas

acrescenta a ideia de definição da magnitude da excitação ou inibição. Uma terceira regra, que é a mais comumente utilizada hoje em dia, é a regra Delta de Widrow. Essa regra propõe a alteração dos pesos sinápticos, minimizando o erro quadrático da rede, reduzindo a diferença entre o valor de saída desejado e o atual valor de saída da unidade de processamento. Assim, o erro da saída é transformado pela derivação da função de transferência e utilizado pra regular os pesos de entrada da camada prévia da rede, realizando assim um processo de retropropagação dos erros. Para a utilização desse tipo de regra deve-se observar que o conjunto dos dados de entrada esteja organizado de forma aleatória.

Há ainda a lei de aprendizado de Teuvo Kohonen, que foi inspirada em sistemas biológicos, em que há uma competição entre os elementos para aprender, ou atualizar e ajustar seus pesos. A unidade de processamento mais apta será aquela que possuir o melhor sinal de saída e terá a capacidade de inibir os ajustes sinápticos de seus concorrentes e excitar seus vizinhos, de maneira que apenas essa unidade e seus vizinhos poderão realizar o ajuste dos pesos.

2.6 Medida de desempenho e qualidade 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 (63), 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 (64), 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 (65) 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: (66)

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

De acordo com as probabilidades condicionais temos:

$$P(X, Y) = P(X|Y).P(Y) = P(Y|X).P(X) \quad (2.6)$$

Tabela 2.4: Matriz modelo de Confusão

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

Fonte: (66)

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 sobreescrita 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 e Regressão para análise preditivas

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).

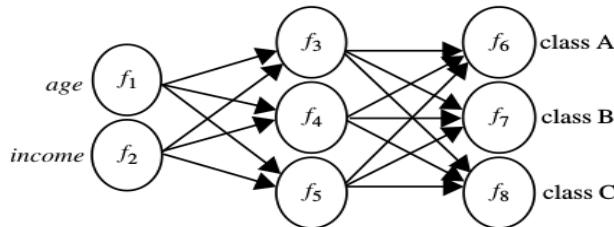
Esse modelo é usado para prever quais rótulos de classes terão os objetos desconhecidos. O modelo pode ser representado por regras de classificação do tipo “IF - THEN”, por árvores de decisão, redes neurais e outros. Regras de classificação se distinguem de regras de indução da seguinte forma:

- Uma regra de classificação poderia ser: $if L \text{ then } \text{class} = C_1$ ou $if L \text{ then } C_1$
- 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 a seguir representa uma rede neural com as mesmas características da árvore de decisão anterior:

Figura 2.25: 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, (14) 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 (10)⁸ o “overfitting” ocorre quando o número atributos é grande.

⁸Foi 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 sigmoidal $g(x)$ 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.

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; pela integração entre mineração de dados históricos e mineração de textos; pela utilização de algoritmos de classificação e de predição.

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

Do ponto de vista prático, pela proposição de um modelo que integre predição 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, até o momento dessa pesquisa, 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), tendo, a princípio, sido executada no Twitter, no canal @PRF191PE e pela busca em outros canais, com as palavras: “protesto”, “planejar paralisação”, por exemplo.

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 (missing data), taxa de 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 para entender e interpretar os resultados.

No quesito robustez, tolerância à faltas e taxa 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 predição.

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

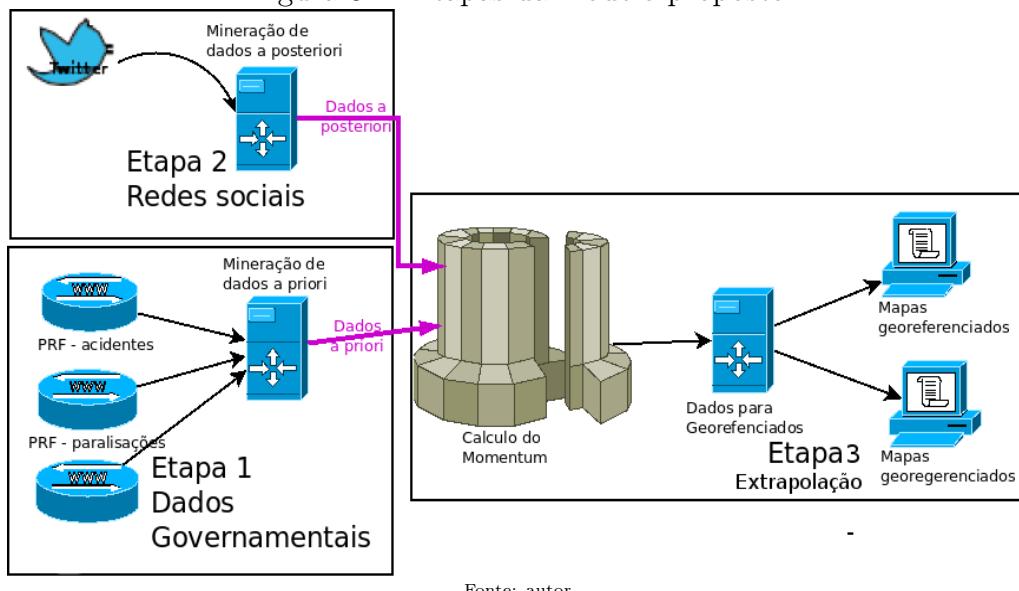
Através de API's, os mapas vetoriais permitem a geolocalização dos pontos classificados ou pontos onde haverá grande número de retenções, conhecidos como gargalo.

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 postagem 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 vencer 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 ilustra (um overview) essa metodologia descrita graficamente.

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 criadas novas variáveis, a proximidade entre as bases heterogêneas foi conseguida utilizando regras de indução da lógica proposicional (10). 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 a seguir 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 (67) destacam a não padronização e não aceitação dos dados pela comunidade internacional. EAVES, D. (68) 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, feitos pelas pessoas que vivem no entorno das rodovias.

Para traçar um painel da diversidade das rodovias pernambucanas, foi efetuado a priori uma classificação através do algoritmo Árvore de Decisão e comparado com o classificador Naïve Bayes. 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 imagem a seguir). 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 3.3: Matriz de Gravidade 3D

Figura 3.4:

MatrizGravidade3D2																
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	1488	0	2	0	0	0	0	0	0
1	0	0	0	0	106	0	0	0	488	1091	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	1091	0	0	0
3	0	0	0	0	0	0	0	1001	0	0	488	82	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	1082	0	0	0	0	0
5	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	15	0	1096	0	0	1976	0	0	0	0	0	1488	0
7	0	1005	0	0	0	0	11	0	0	0	0	0	0	0	0	0
8	0	1488	0	0.03	0	1488	0	1488	11	0	0	488	0	0	0	0
9	0	1.57	1091	1082	1173	0	0	0	0	0	0	0	0	3652	0	1488
10	0	0	0	0	91	1011	0	0	1.49	0	0	0	1002	0	0	0
11	0	0	1488	22	0	0	0	1488	2976	488	0	0	0	91	0	0
12	0	0	0	0	0	0	0	1488	0	488	0	0	0	0	0	0
13	0	0	0	0	0	0	0	1082	1488	0	0	0	0	0	1488	0
14	0	82	0	1015	1	1976	488	1488	1488	0	0	0	1011	0	0	0
15	0	0	1082	5	488	0	0	0	1488	0	0	0	0	0	0	0
16	0	0	0	0	0	0	11	3978	0	0	0	0	0	0	0	0
17	0	0	1015	0	91	1488	0	0	1488	0	488	1976	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	91	1015	0	0	0	1976	0	15	0	0	0	0	0	0	0
20	0	0	1015	0	0	0	0	0	0	1	0	0	1488	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1016	0
22	0	0	0	488	0	1015	488	0	0	1	0	0	0	488	0	0
23	0	0	0	0	15	0	0	1488	0	0	0	0	0	82	0	0

Km (0 - 213)

Fonte: autor

3.4 Arco cibernetico com dados do Twitter

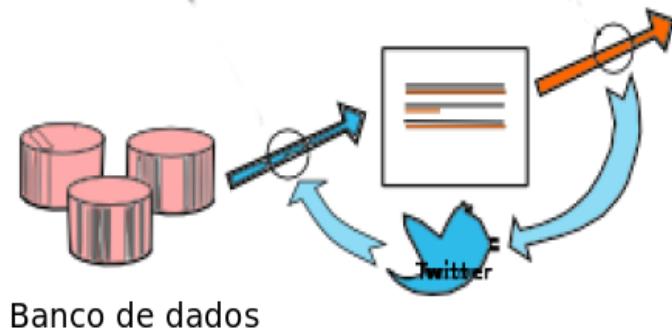
Os dados do Twitter permitem uma busca imediata por novas informações que poderão ser confrontadas com o modelo preditivo, aumentando o nível de confiança deste. Com isso, as informações construirão um Arco cibernetico. Segundo Wiener (1948), novas informações permitem realimentação aos sistemas, com 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 sociais ², para estes casos a decisão de qual ação a ser tomada sempre estará “nas mãos” do agente, do observador ou do utilizador.

As informações das redes sociais que comporão o arco cibernetico não deverão retroalimentar o modelo de predição já construído, pois o fluxo decisório já foi tomado pelo observador. Os dados a posteriori não servem para um modelo de predição, uma vez que enviesam o sistema preditivo.

²O sistema de predição é baseado em cálculos probabilísticos

Figura 3.5: O arco cibernético com o Twitter

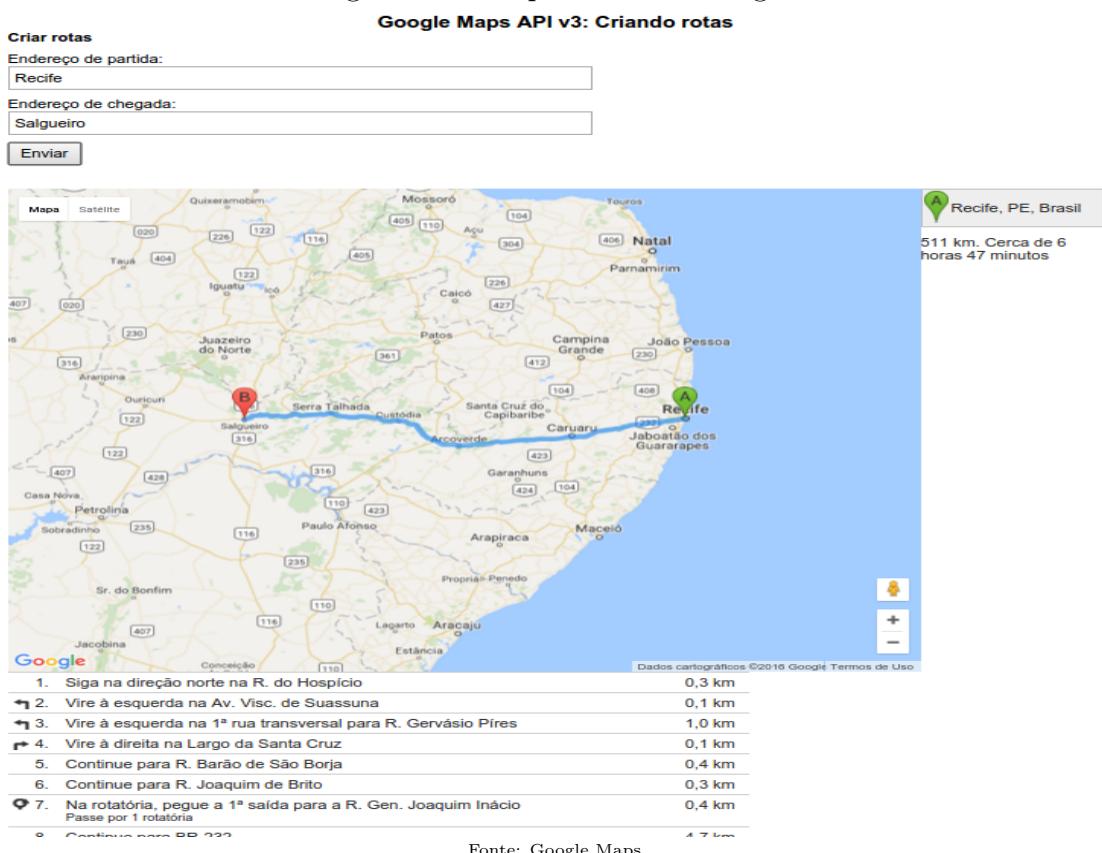


Fonte: autor

3.5 Extrapolação para georreferenciamento

Os pontos críticos das matrizes geradas (de morte ou de gravidade) constituem-se como marcadores no mapa, que deverão ser levados em conta pelos utilizadores quando estiverem passando pelo local referenciado, conforme imagem a seguir.

Figura 3.6: Etapas da metodologia



O capítulo a seguir contemplará os resultados encontrados após a execução de todas as fases aqui apresentadas.

4

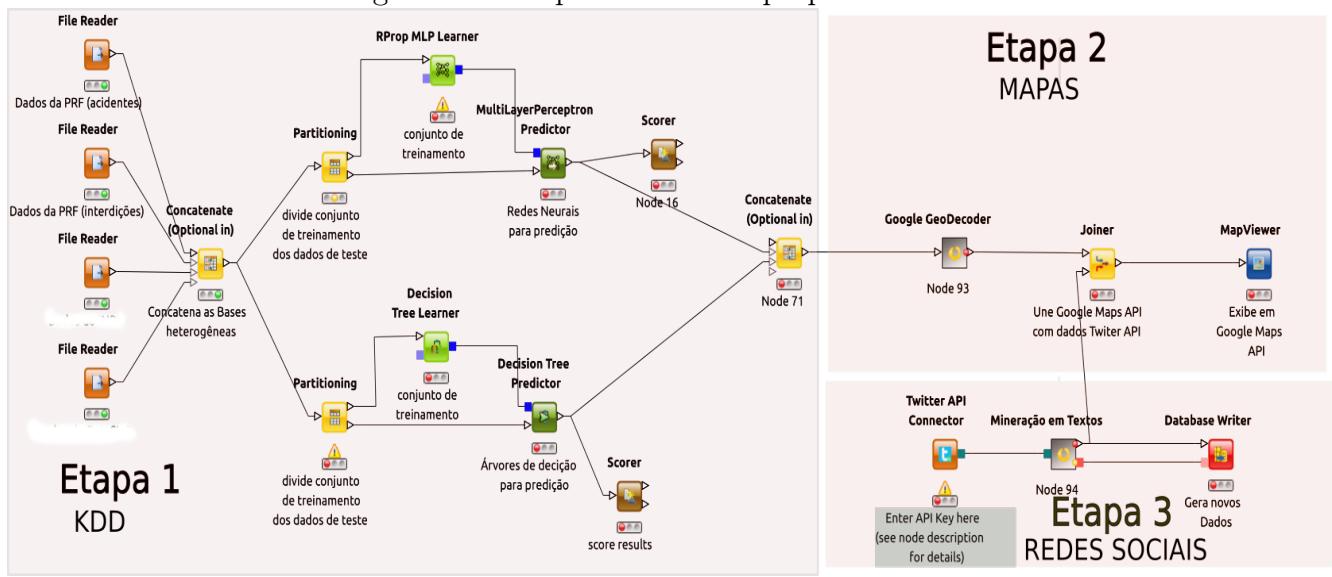
Simulação

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.

4.1 Execução do modelo

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

Figura 4.1: Etapas da modelo proposto



Fonte: autor

A **etapa 1** contempla a fase da coleta das bases históricas, 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.

2. 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.
3. A conclusão dessa etapa ocorre com a Mineração dos dados e a extração de conhecimento. Os "outputs" dessa etapa consistem nos pontos críticos localizados nas rodovias, que permitem traçar uma rota. As localizações geográficas, indicadas pelo Km, são agrupadas formando "clusters" de dados exibidos em mapas vetoriais. A partir da definição dos pontos críticos foram geradas matrizes – de mortos e de gravidade – que serviram para localizar tais pontos no mapa e traçar rotas. As principais ferramentas de I.A. utilizadas nessa etapa foram Árvores de Decisão e Redes Neurais, com a finalidade classificar e predizer os pontos críticos em cada quilômetro das rodovias.

A **segunda** etapa contempla:

1. Representação da malha viária em mapas de bases vetoriais;
2. Um ambiente de simulação interativa que utiliza uma plataforma baseada na API do Google Maps.
3. Quando houver informações vindas do twitter sobre ocorrências na rodovia, deverá ser feita a geolocalização.

A **terceira** e última etapa consiste em um módulo com as seguintes características:

1. Um módulo dinâmico onde são capturados "feeds" da rede social Twitter. Essa técnica faz um arco cibernético mantendo o utilizador atualizado com as informações recentes.

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 (67) destacam a não padronização e não aceitação dos dados pela comunidade internacional. EAVES, D. (68) 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 (69), R (70) e Weka (71), com objetivo de estabelecer uma comparação entre eles, cuja intenção era produzir um classificador mais robusto. Nessa direção, a técnica Ensamble de classificadores (72) 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 “rparty” implementa esse algoritmo.

Para escolha das variáveis de input 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 input foi a entropia, que é um elemento considerado importante pela literatura (10). 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.

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

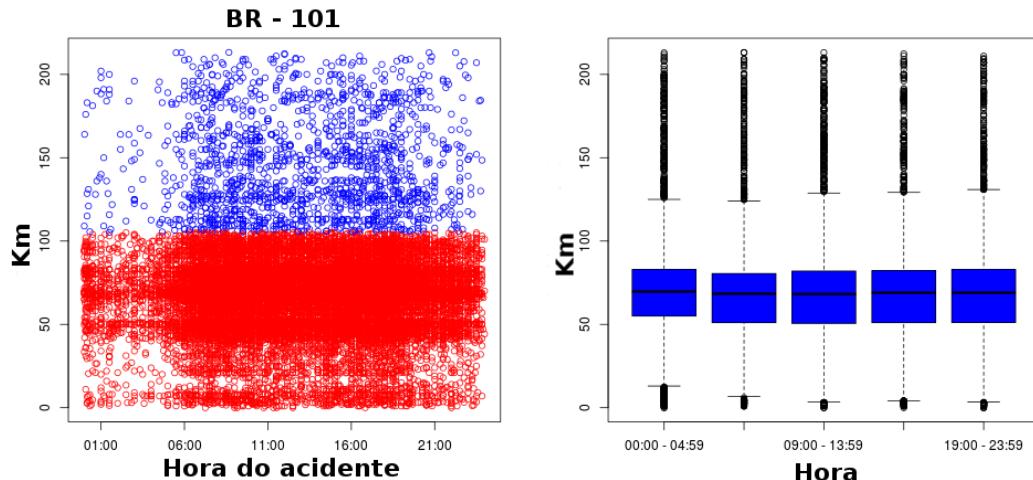
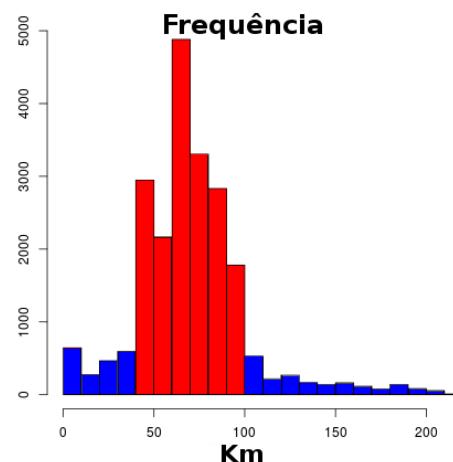


Figura 4.3: Frequência



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 4.4: Hora do acidente (1) — Concentração em torno da hora (2)

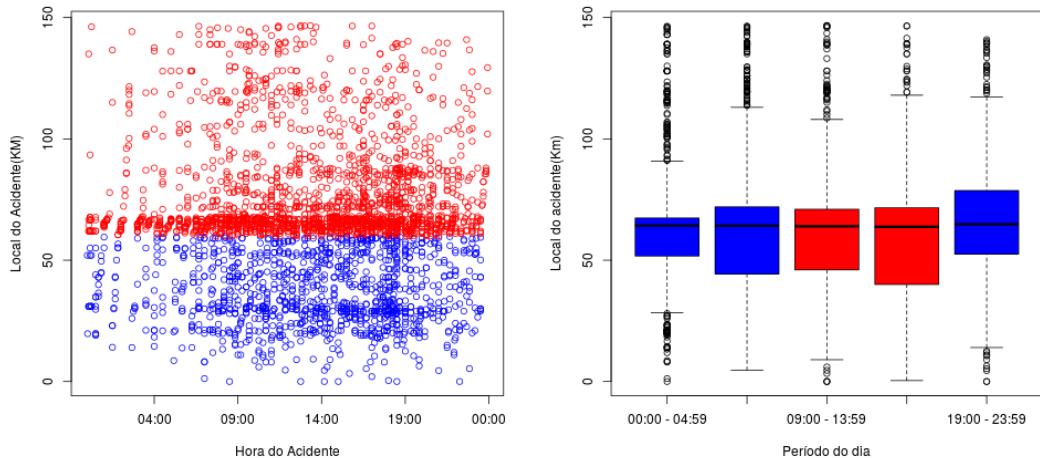
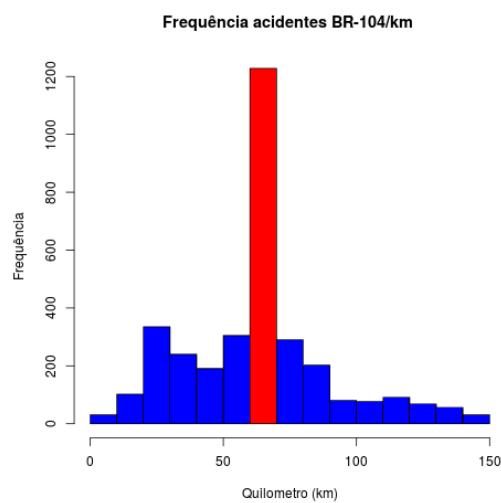


Figura 4.5: Frequência



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) corresponde à frequência do local onde ocorreram esses acidentes. Percebe-se que em torno do Km 60 concentra-se o maior número de ocorrências. O terceiro gráfico (boxplot) apresenta as ocorrências em torno da mediana dessa localidade (Km). No gráfico 1 são identificados padrões, por exemplo no Km 60 ocorrem acidentes que se estendem das 04h às 23h.

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

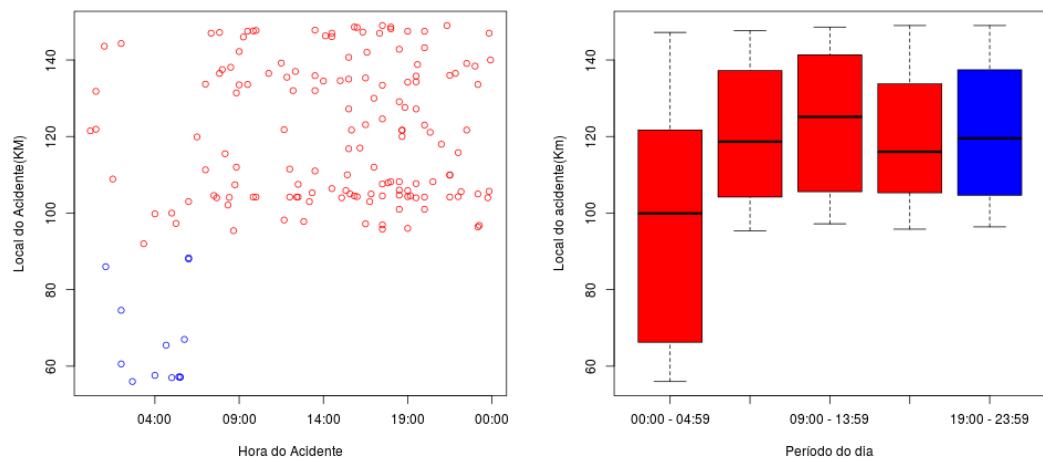


Figura 4.7: Frequência

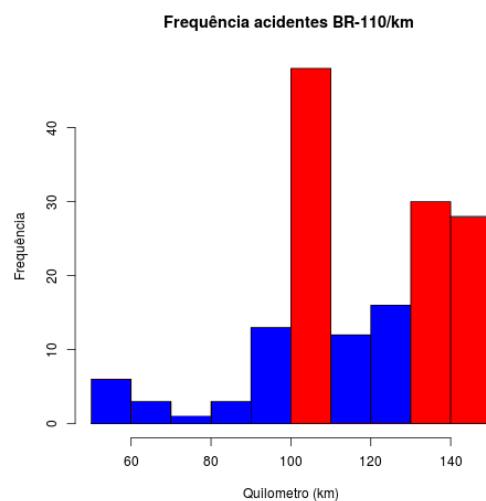


Figura 4.8: Hora do acidente (1) — Concentração em torno da hora (2)

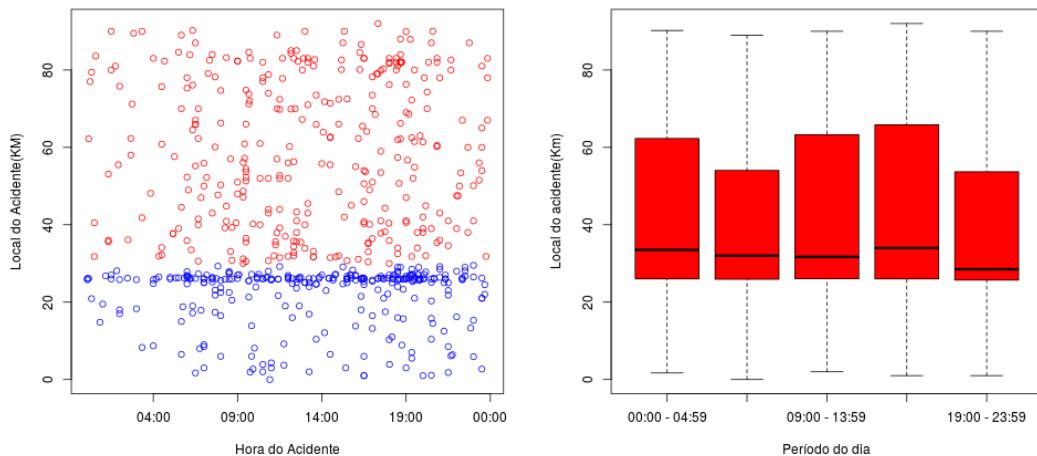
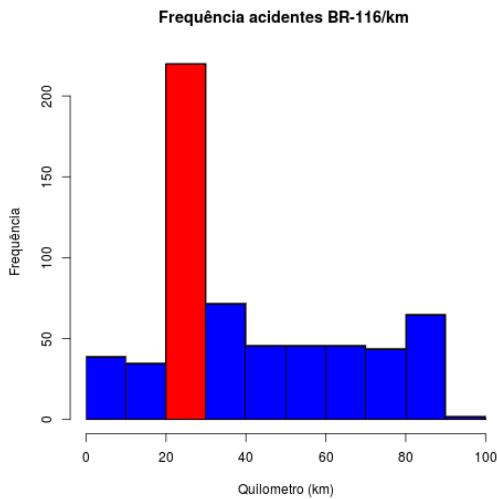


Figura 4.9: Frequência



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 4.10: Hora do acidente (1) — Concentração em torno da hora (2)

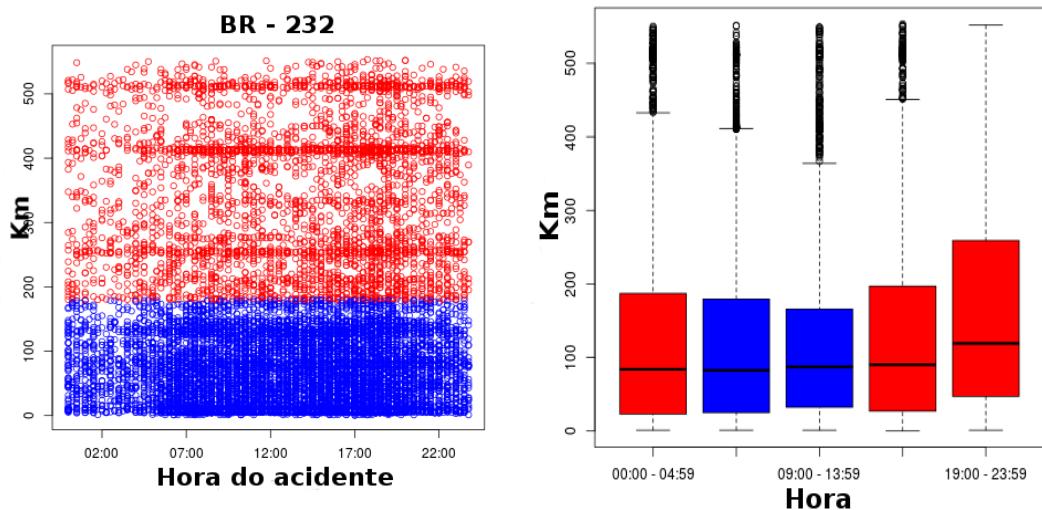
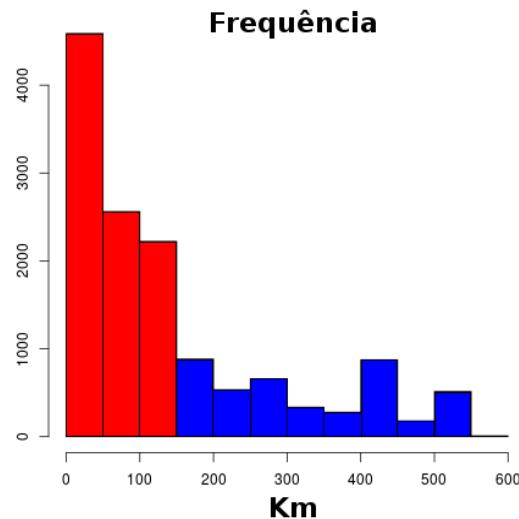


Figura 4.11: Frequência



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. Posteriormente foram identificadas outras causas que influenciam mais fortemente essas ocorrências.

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

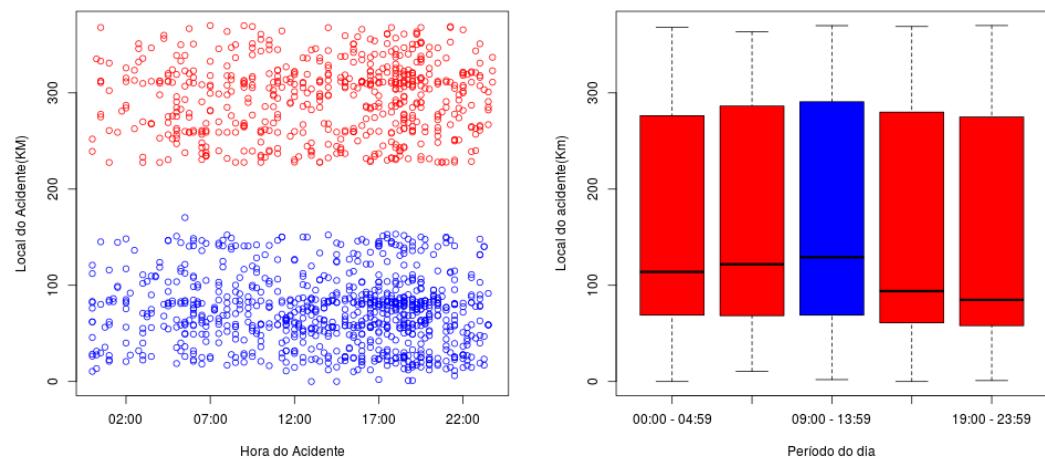


Figura 4.13: Frequência

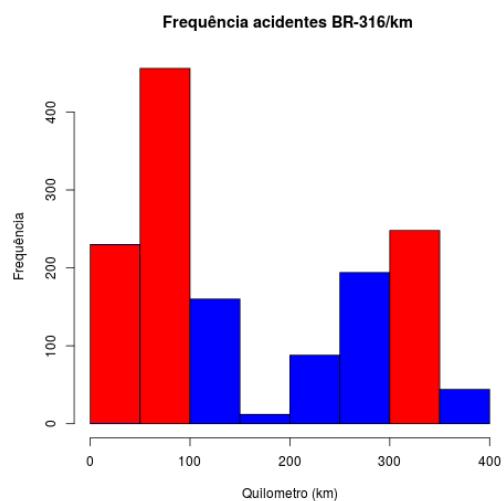


Figura 4.14: Hora do acidente (1) — Concentração em torno da hora (2)

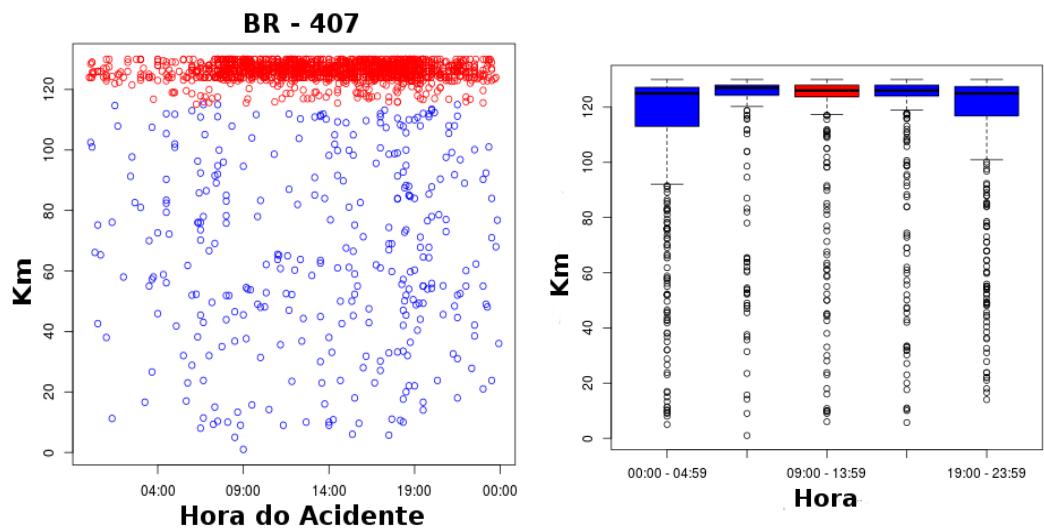
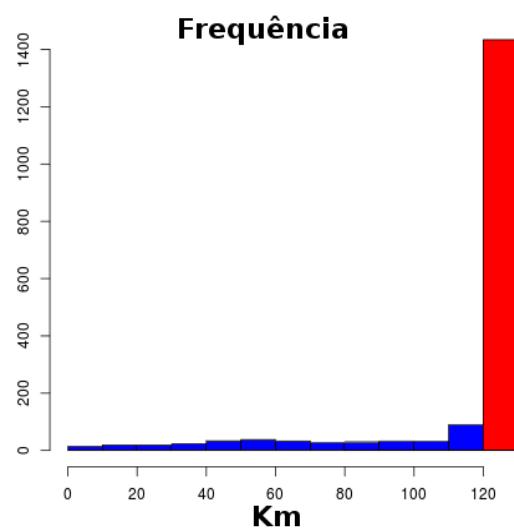


Figura 4.15: Frequência



Na BR 407 os acidentes se concentram na altura do Km 130.

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

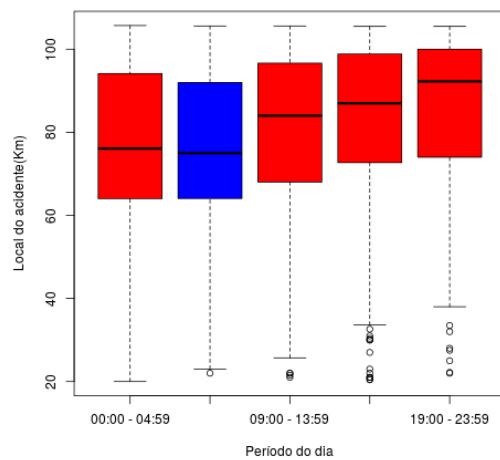


Figura 4.17: Frequência

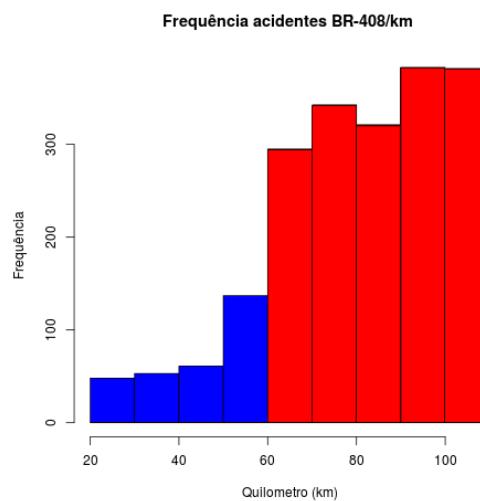


Figura 4.18: Hora do acidente (1) — Concentração em torno da hora (2)

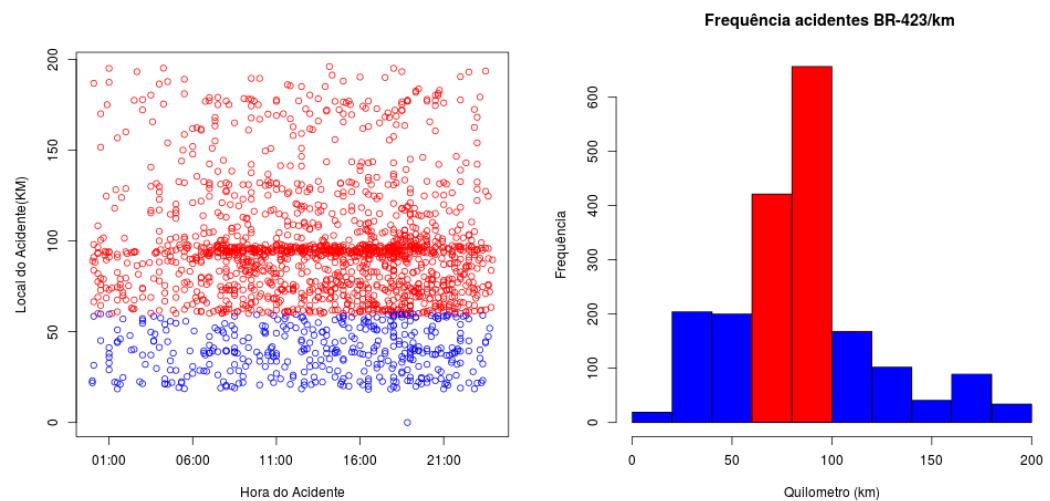


Figura 4.19: Frequência

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

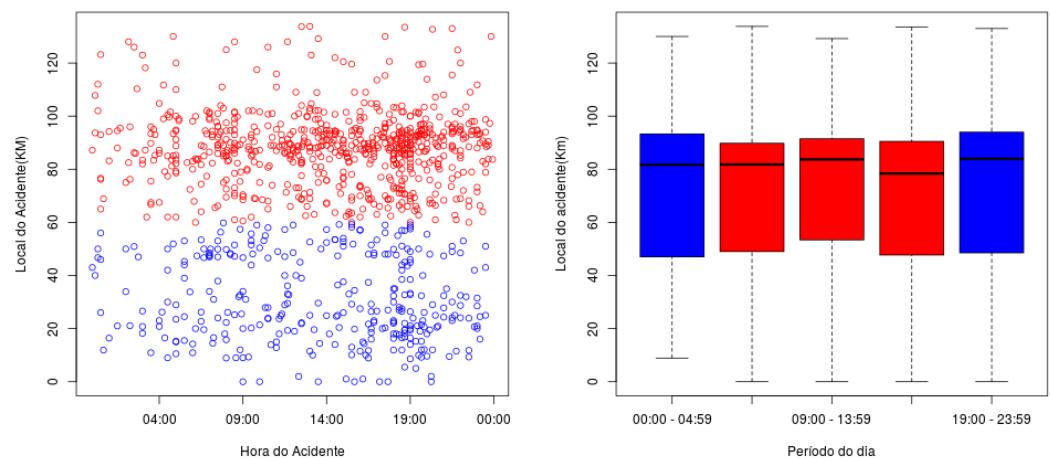


Figura 4.21: Frequência

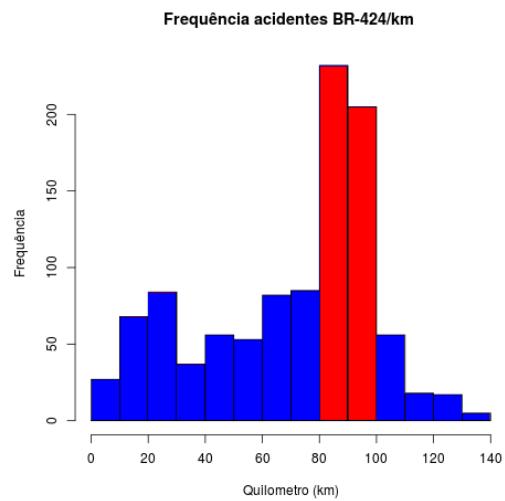


Figura 4.22: Hora do acidente (1) — Concentração em torno da hora (2)

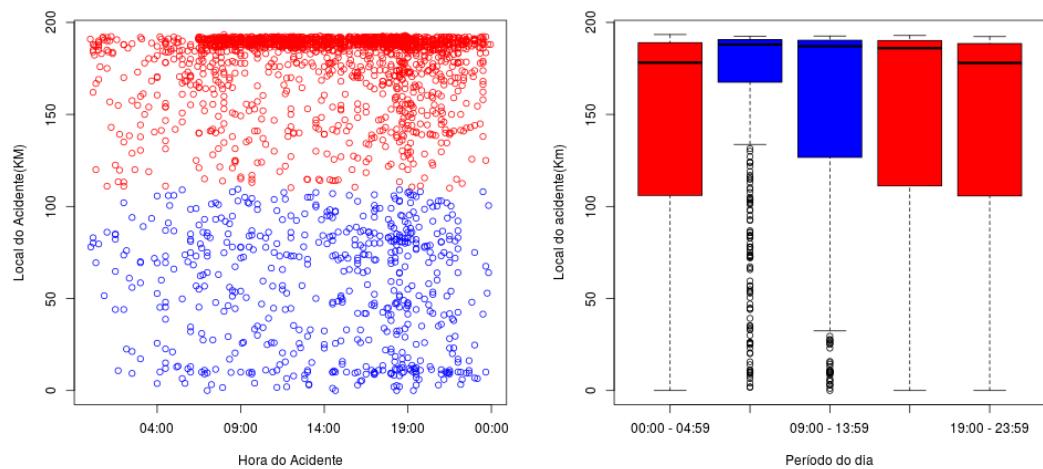


Figura 4.23: Frequência

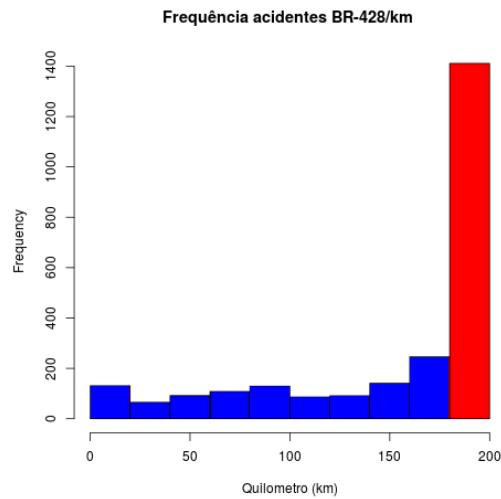
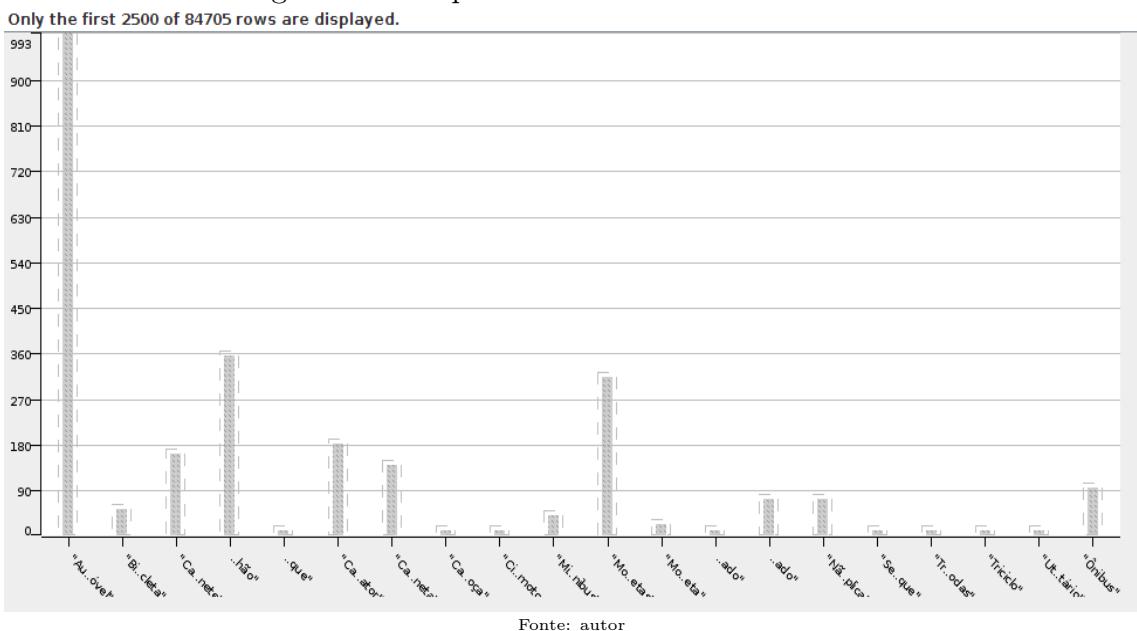


Figura 4.24: Tipo de Veículo X Num. Acidentes

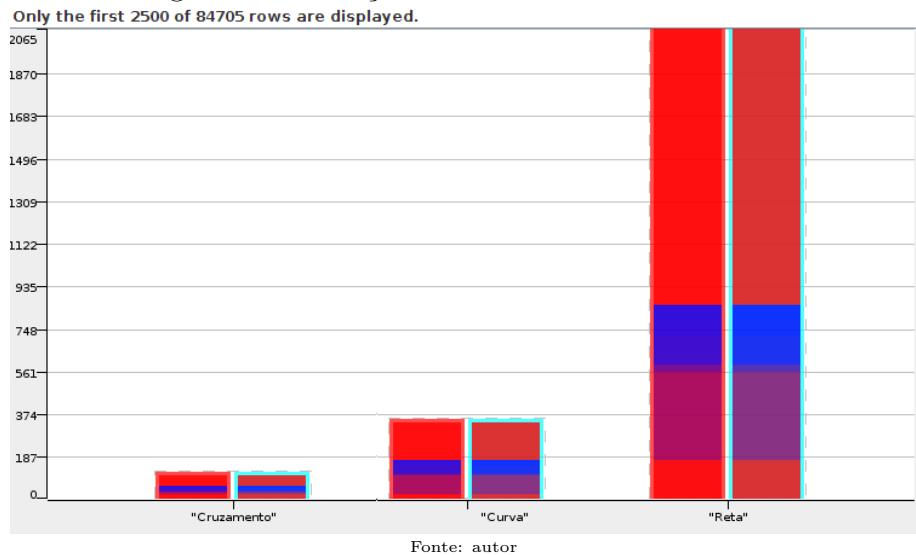


Fonte: autor

O maior número de acidentes ocorre com automóvel de passeio, provavelmente condutores comuns, não profissionalizados. O Caminhão é o segundo veículo que mais se envolve em acidentes, seguido das motonetas.

Os dados revelam que o tipo de traçado da via não tem influência primordial nas estatísticas de acidentes, pois a maioria dos acidentes ocorre em pista retilínea. Os dados sugerem, ainda, que o condutor é o principal responsável pelo maior número de ocorrências nas BRs. Isso direciona essa pesquisa para analisar e antever o comportamento do condutor, além das condições da rodovia.

Figura 4.25: Traçado da via 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 (10) 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.: Precison = $TP / (TP + FP)$;
- Recall = $TP / (TP + FN)$;
- F-Me: F-measure ou f-score = $2 * \text{Precison} * \text{Recall} / (\text{Precision} + \text{Recall})$;
- 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	—

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

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

a	b	c	d	e	f	g	h	Classificadores
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
—	—	—	—	—	—	—	—	—

Os valores restantes foram omitidos por não representarem uma amostra adequada, pois a acurácia foi consideravelmente baixa. 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 quadratico	0.4368	—

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

Tabela 4.4: Matriz de confusão para a variável Gravidade

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

(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	—

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

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% e o número de falsos positivos próximo a 0% traduzem uma área da 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 80% e o menor índice de classificação incorreta dentre os três classificadores utilizados. Esta variável foi a escolhida para encabeçar os algoritmos para os resultados encontrados.

Tabela 4.6: Matriz de confusão para a variável BRajustada

a	b	c	d	e	f	g	h	Classificadores
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

A árvore construída pelo Knime para a mesma variável “Causa do Acidente” => velocidade incompatível é apresentada na próxima figura.

Figura 4.26: Árvore de Decisão gerada pelo Knime



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 curva ROC acima dos 90as 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 para a variável “Delegacia” 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]

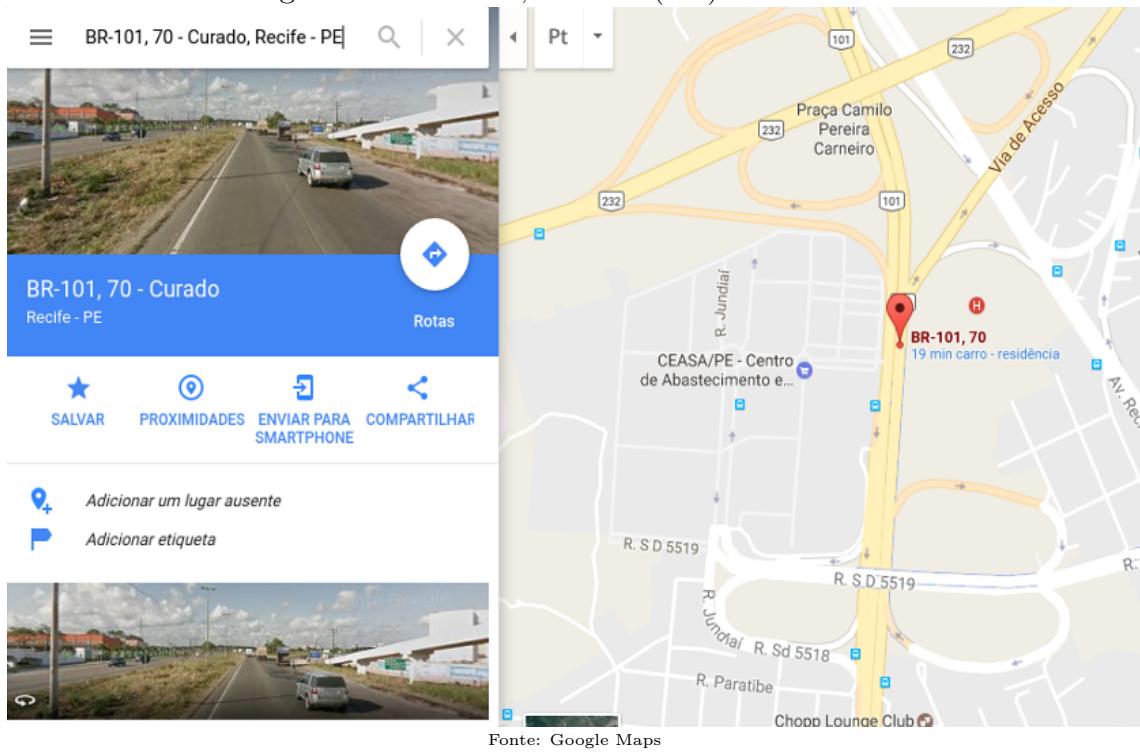
Essa variedade de causas explica que o condutor dessa região não respeita a sinalização, os limites de velocidade, dentre outras regras de trânsito. Pode-se dizer que é um condutor indisciplinado, pois todas as causas de acidentes elencadas foram encontrados. Caso se considere um raio de 50 Km no entorno 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.

A figura a seguir, obtida a partir da API do Google Maps, demonstra o local aproximado (Km 70, Br 101 – sul) destacado na Matriz de Mortos. Ao ser consultada a Árvore de Decisão (acima) para explicar as causas do alto índice de óbitos, constatou-se que eram, em sua maioria, mortes por atropelamento. A imagem do Google Maps 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



Para 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

Os dados gerados pelos outros classificadores – Naïve Bayes e Redes Neurais – permitiram confirmar que os resultados encontrados pelas Árvores de Decisão são os mais adequados para o modelo proposto.

A seguir exemplificamos alguns dos resultados desses classificadores:

Figura 4.28: Resultado da classificação feita pelo Naïve Bayes – acurácia

NaiveBayesCausaAccident								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
431	180	443	431	437	254	680	414	Outras
765	436	466	765	579	311	713	530	Falta de atenção
25	5	196	25	45	56	729	126	Velocidade incompatível
108	15	155	108	128	112	785	98	Ultrapassagem indevid
208	25	277	208	237	210	794	203	Desobediência à sinaliz
109	8	343	109	165	177	791	200	Defeito mecânico
742	38	455	742	564	559	938	695	Animais na Pista
77	13	450	77	132	148	839	354	Não guardar distância
82	5	277	82	127	140	838	144	Dormindo
144	8	409	144	213	227	841	237	Defeito na via
180	32	235	180	204	168	803	180	Ingestão de álcool

Fonte: O Autor

Figura 4.29: Resultado da classificação feita pela Rede Neural – acurácia

```
Redes Neurais -- classificador
-----
Threshold      -7.651297135217415
Attrib CondPista    9.154976078003095
Attrib RestrVisibili   -4.4972310838186305
Attrib TipoAcidente   1.3788162272305897
Attrib CausaAccident   -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 TipoAcidente -1.1503782569405412
Attrib CausaAccident -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.

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.

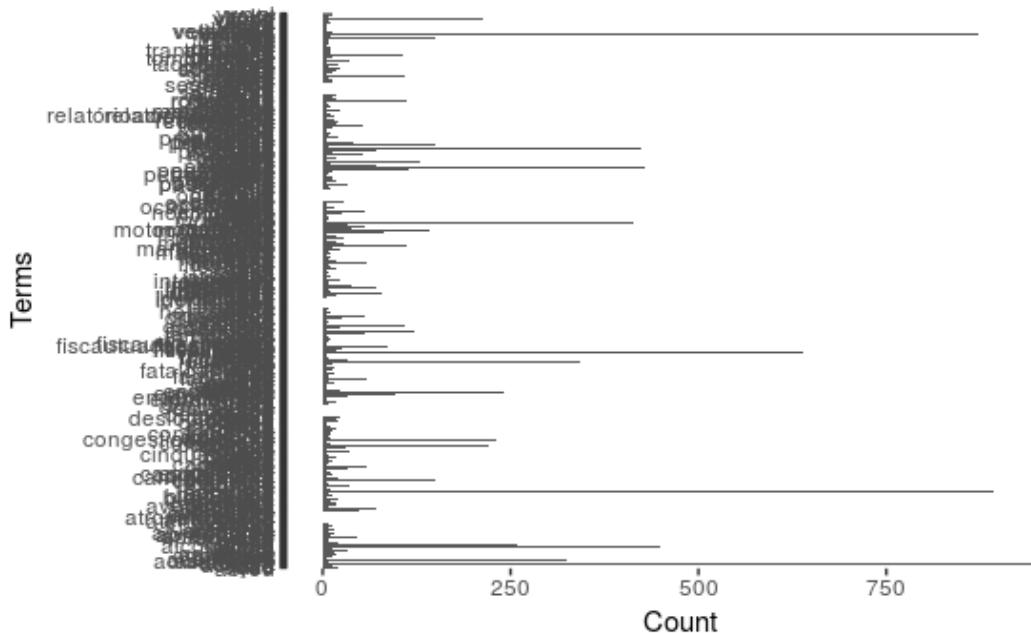
Para localizar novos canais de informação, foram construídos subgrafos contemplando, além dos tweets da PFR, retweets dos seguidores desse canal. Esperava-se, com isso, encontrar novos Hubs na rede. No entanto, a busca em subgrafos é uma tecnologia que precisa ser investigada mais a fundo, dada a sua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

Conceitos como betwenness, 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 que 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 a seguir 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 – e bigramas, termos com duas palavras. Foram encontradas palavras como: fiscalização, colisão, vítima fatal, acidentes, detre outros.

Figura 4.30: Gráfico de frequência de palavras – unigramas



A nuvem de palavras 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 (acid), balanço, dentre outras.

Figura 4.31: Nuvem de palavras da Mineração em textos



Figura 4.32: Dendograma de Clusterização do resultado da mineração

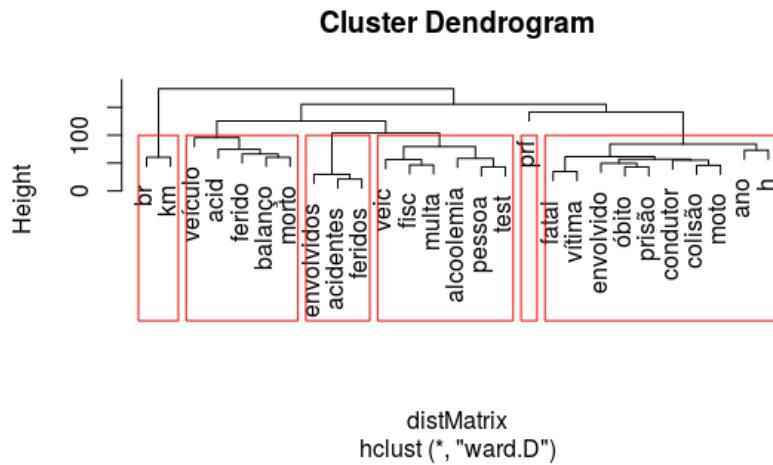
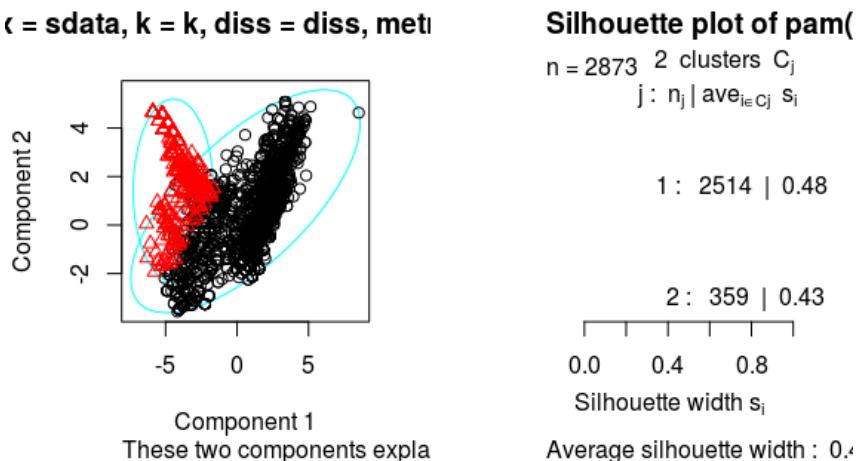


Figura 4.33: Gráfico da clusterização do resultado da mineração



O dendrograma 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.

O próximo gráfico ilustra a utilização de outro algoritmo de agrupamento (clusterização), conhecido com K-means.

Uma análise mais minuciosa dos gráficos produzidos pela mineração de textos pode trazer contribuições de considerável relevância, que permitem refinar o modelo de predição.

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 para os mais diversos atores que fazem uso das rodovias.

Os resultados apontam 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.

Nesse sentido, órgãos federais de controle de estradas podem se beneficiar de pesquisas dessa natureza. Os dados encontrados nessa pesquisa corroboram com os resultados encontrados em estudos semelhantes, quer seja no Brasil ou em outros países, sugerindo que há um padrão de comportamento em rodovias que pode ser analisado, de maneira a facilitar o transporte de cargas e o tráfego de veículos em geral, em seu curso.

A pesquisa também contribuiu para a compreensão das causas de constrangimentos e acidentes 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. Por exemplo, falta de sinalização, passarelas, limitador de velocidade e faixa de pedestre são fortes condicionantes para as situações de atropelamento. A quantidade de mortes por atropelamento entre os quilômetros 66 e 71 da BR 101 - mais de 9.500 mortes em nove anos - corroboram com essa análise.

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 algo-

ritmos. 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.

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), embora tenha sido uma ferramenta útil, há a necessidade de um monitoramento intensivo das redes sociais, para que pudesse 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.

Do ponto de vista prático, a contribuição da pesquisa se dá pela 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.

Em relação a outras pesquisas dessa natureza, algumas delas consideradas no Estado da Arte dessa dissertação, essa pesquisa avança em relação ao que até então foi proposto, 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. Outro avanço em relação às pesquisas identificadas diz respeito à utilização de uma ampla gama de técnicas de I.A. para encontrar soluções, bem como, propor articulações entre essas técnicas.

Alguns aspectos merecem destaque em relação às dificuldades enfrentadas 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, dados relativos à latitude e longitude estavam ausentes ou 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 uma alternativa, uma vez que a localização era fundamental para o georreferenciamento.

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 a questão proposta com respeito 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. Nesse sentido, entendemos que novas pesquisas precisam ser condizidas, para a ampliação do modelo sugerido.

Trabalhos futuros incluem a incorporação desta proposta em modelos formais de decisão, por exemplo, de roteamento rodoviário metropolitano. Através da API do twitter, implementar algoritmos de busca em redes sociais, para encontrar os Hubs difusores de informações.

A API Google Maps, o “front-end” do modelo proposto, em uma futura aplicação poderá ser executada em um aparelho celular do tipo “Smartphone”, com capacidade para executar aplicativos gráficos mais complexos.

A

Preprocessamento

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* (3), 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.(67) 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

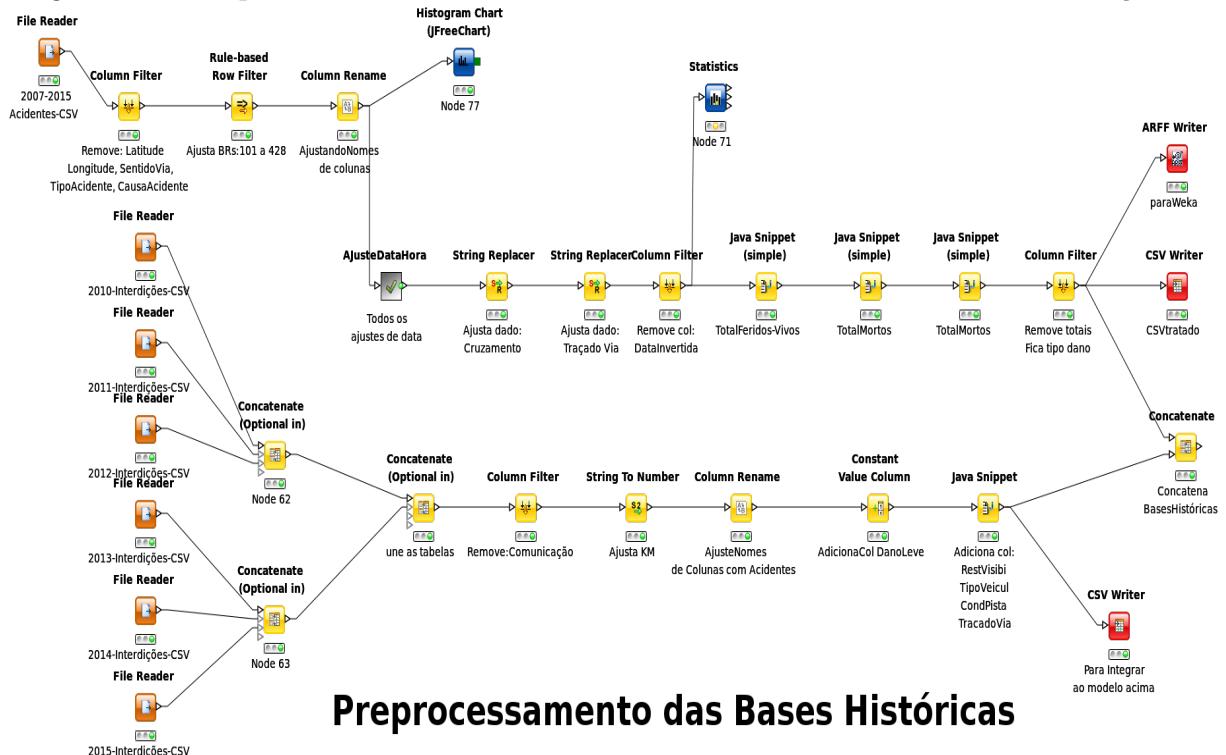
Ano	Ano da ocorrência do acidente
Mês	Mês de ocorrência do acidente
Num	Número do mês do acidente ex: 1 = Janeiro
KM	Numeração do quilômetro
BR	Numeração da Br
Latitude	Latitude da ocorrência
Longitude	Longitude da ocorrência
Condição Pista	Condição da pista: seca, molhado, ...
Restrição de Visibilidade	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 seguinte; 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.1: Etapas 1 – Coleta e união das bases históricas de acidentes e interdições



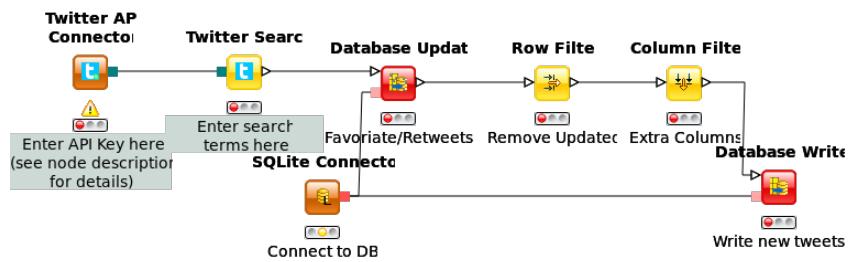


Figura A.2:

A.2 Preprocessamento dos dados do Twitter

Referências Bibliográficas

- 1 WIRTH, R. Crisp-dm 1.0 – step-by-step data mining guide. p. 7–10, 2000.
- 2 SALLES, F. R. *A relevância da cibernetica*. Tese (Dissertação de Mestrado) — Universidade de São Paulo - USP, 2007.
- 3 TRANSPORTES, P. e. A. Governo Federal Ministério dos. 2016. Disponível em: <<http://dados.gov.br/dados-abertos/>>.
- 4 BNDES. Perspectivas do investimento, n. 2, out. 2013. *Perspectivas do Investimento 2014-2017*, p. 2, 2013.
- 5 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>.
- 6 IBGE, I. B. de Geografia e E. Região metropolitana do recife no contexto de pernambuco no censo 2010. 2014. Disponível em: <http://www.cidades.ibge.gov.br/painel/frota.php?codmun=261160&search=pernambuco/recife/infograficos:frota-municipal-de-veiculos/&lang=_ES>.
- 7 POSSAS, B. et al. Data mining: técnicas para exploração de dados. *Universidade Federal de Minas Gerais*, 1998.
- 8 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.
- 9 CHAPMAN, P. et al. Crisp-Dm 1.0. *CRISP-DM Consortium*, p. 76, 2000.
- 10 RUSSEL, S.; NOVING, P. *Inteligência Artificial*. [S.l.]: Elsevier, Rio de Janeiro, 2004. 716–721 p. ISBN 8535211772.
- 11 CASTANHEIRA, L. G. Aplicação de técnicas de Mineração de Dados em Problemas de Classificação de Padrões. n. 5531.
- 12 HAN, J.; KAMBER, M. Data mining: Concepts and techniques. Elsevier, San Francisco, v. 2 edition, p. 15–16, 2006.
- 13 ARANHA CHRISTIAN E PASSOS, E. *A Tecnologia de Mineração de Textos*. 2006. 1–8 p.
- 14 Srivastava, V. K.; SINGH, N. Review of Decision Tree Algorithm: Big Data Analytics. *International Journal of Informative & Futuristic Research*, v. 2, n. 10, p. 3644–3654, 2015.

- 15 AMIN, A. et al. A comparison of two oversampling techniques (smote vs mtdf) for handling class imbalance problem: A case study of customer churn prediction. v. 353, p. 215–225, 2015. Disponível em: <<http://link.springer.com/10.1007/978-3-319-16486-1>>.
- 16 HOTHÓ, 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.
- 17 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.
- 18 LIDDY, E. D. Natural language processing. 2001.
- 19 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.
- 20 JAGADISH H. V., G. J. L. A. P. Y. P. J. M. R. R.; SHAHABI, C. Exploring the inherent technical challenges in realizing the potential of big data. *Comunication of the ACM*, v. 57, n. 7, p. 86–96, July 2014.
- 21 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>>.
- 22 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.
- 23 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.
- 24 FAN, W. et al. Tapping the power of text mining. *Communications of the ACM*, ACM, v. 49, n. 9, p. 76–82, 2006.
- 25 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.
- 26 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.
- 27 SARKER, A. et al. Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics*, Elsevier, v. 54, p. 202–212, 2015.
- 28 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.
- 29 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.
- 30 MADDEN, M. et al. Teens, social media, and privacy. *Pew Research Center*, v. 21, p. 2–86, 2013.

- 31 MITCHELL, A.; HOLCOMB, J.; PAGE, D. News use across social media platforms. *Washington, Pew Research Center*, 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 ADKINS, D.; BUDD, J. Scholarly productivity of us lis faculty. *Library & Information Science Research*, Elsevier, v. 28, n. 3, p. 374–389, 2006.
- 34 CUNNINGHAM, S. J.; DILLON, S. M. Authorship patterns in information systems. *Scientometrics*, Springer, v. 39, n. 1, p. 19, 1997.
- 35 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.
- 36 CHA, M. et al. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, v. 10, n. 10-17, p. 30, 2010.
- 37 SUH, B. et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: IEEE. *Social computing (socialcom), 2010 ieee second international conference on*. [S.l.], 2010. p. 177–184.
- 38 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.
- 39 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.
- 40 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.
- 41 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>>.
- 42 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.
- 43 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.
- 44 NILSSON, N. J. Introduction to Machine Learning. *Machine Learning*, v. 56, n. 2, p. 387–99, 2005. ISSN 10959572. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/21172442>>.
- 45 BEN-DAVID, S.; SHALEV-SHWARTZ, S. *Understanding Machine Learning: From Theory to Algorithms*. [s.n.], 2014. 449 p. ISBN 9781107057135. Disponível em: <<http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf>>.

- 46 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.
- 47 MONTGOMERY, D. C.; RUNGER, G. C.; CALADO, V. *Estatística aplicada e probabilidade para engenheiros*. [S.l.]: Grupo Gen-LTC, 2000.
- 48 POLICARPO, R. A. C. S. A. Semantic classification of nouns. 2015.
- 49 CAMILO, C. O.; SILVA, J. C. da. *Mineração de dados: Conceitos, tarefas, métodos e ferramentas*. [S.l.], 2009. 1–29 p.
- 50 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>>.
- 51 QUINLAN, J. R. Induction of decision trees. *MACH. LEARN*, v. 1, p. 81–106, 1986.
- 52 KOHAVI, R.; QUINLAN, R. Decision tree discovery. In: *IN HANDBOOK OF DATA MINING AND KNOWLEDGE DISCOVERY*. [S.l.]: University Press, 1999. p. 267–276.
- 53 SIMÕES, A. C. D. A. Mineração de Dados baseada em Árvores de Decisão para Análise do Perfil de Contribuintes. *Discovery*, p. 32, 2008.
- 54 WITTEN, I. H.; FRANK, E. Data mining: Practical machine learning tools and techniques. *Elsevier, San Francisco*, v. 2 edition, 2005.
- 55 BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CRC press, 1984.
- 56 PINEDA, V. C. A Entropia Segundo Claude Shannon: o desenvolvimento do conceito fundamental da teoria da informação. *Dissertação Mestrado*, 2006.
- 57 HEATON, J. *Introduction to Neural Networks for Java*. [S.l.: s.n.], 2008. 440 p. ISBN 1604390085.
- 58 TATIBANA, C. Y.; KAETSY, D. Y. Acessado em: 01.out.2016. Disponível em: <<http://www.din.uem.br/ia/neurais/#neural>>.
- 59 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.
- 60 R. ZENG, H. H. P. X. W. S. C. G. M. C. Rule extraction from an optimized neural network for traffic crash frequency modeling.
- 61 KRIESEL, D. *A Brief Introduction to Neural Networks*. [s.n.], 2007. Disponível em: <Disponível em:<http://www.dkriesel.com>>.
- 62 BARRETO, J. M. Jorge M. Barreto. 2002.
- 63 EGAN, J. P. Signal detection theory and roc analysis. New York, USA: Academic Press, 1975.
- 64 ANAESTHETIST, T. Acessado em: 20.jan.2015. Disponível em: <<http://www.anaesthetist.com/mnm/stats/roc/Findex.htm>>.

- 65 R., S. C. Acessado em: 20.jan.2015. Disponível em: <<http://crsouza.com/2009/07/analise-de-poder-discriminativo-atraves-de-curvas-ro>>.
- 66 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.
- 67 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>>.
- 68 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>>.
- 69 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.
- 70 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. Disponível em: <<https://www.R-project.org/>>.
- 71 ZEALAND, U. of W. N. 2017. Disponível em: <<http://www.cs.waikato.ac.nz/>>.
- 72 BERNARDINI, F. C. “combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéticos” - tese de doutorado. Instituto de Ciências e Matemática Computacional/USP, 2006.