Programa de Pós-Graduação em Engenharia de Sistemas

DISSERTAÇÃO

Recife, 03 Janeiro 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

Prof. Dra. Rita de Cássia M. Nascimento.....Engenharia de Sistemas/POLI/UPE

Recife, 03 Janeiro 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".

A todos da Polícia Rodoviária Federal, pelo dados cedidos, em especial ao agente Deivierson, 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.

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, paralizações, 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, utitlizamos o software de georreferenciamento QGIS. 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 PFR, 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 QGIS georeferencing software. Compared to usual navigation approaches, the proposed prediction model represents a breakthrough in terms of mobility and management of tranportation and vehicle trafic , 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	*Adaptative Linear Neuron*
MADALINE	*Many Adaline*
API	*Application Programming Interface*
BG	*Big Data*
DM	*Data Mining*
TDM	*Text Data Mining*
KDD	*Knowledge Discovery Databases*
CRISP-DM	*Cross Insdustry Standard Process for Data Mining*
PRF	*Polícia Rodoviária Federal*
BPRv	*Batalhão de Polícia Rodoviária (estadual)*
NB	*Naïve Bayes*
PB	*PetaByte*
EB	*ExaByte*
ZB	*ZettaByte*
YB	*YottaByte*

Lista de Figuras

Lista de Tabelas

Sumário

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 Nobert 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 á problemática humana e, ao mesmo tempo, com rapidez.

A inteligência artificial é uma área 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.

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.

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 (3).

A RMR é a 5ª 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 (4). 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 (5). 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.

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.

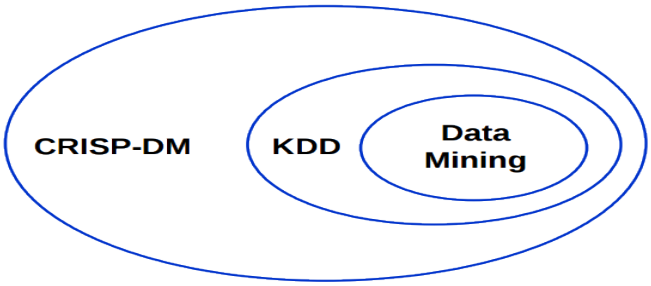
Tabela 2.1: Mineração de dados – contexto de aplicação (1)

Dimensão	**Domínio da aplicação**	**Tipo de Problema**	**Aspecto técnico**	**Ferramentas e Técnicas**
Exemplo	Modelo de resposta	Descrição e sumarização	Dados faltantes	Clementine
—	Predição agitada	Segmentação	*Outlies*	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 (6), 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 (7). 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



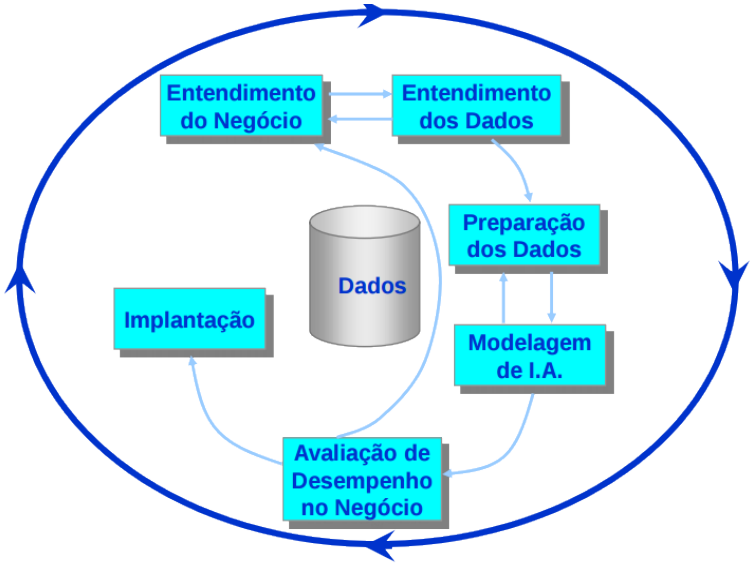
Fonte: Neurotech – 2012

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:

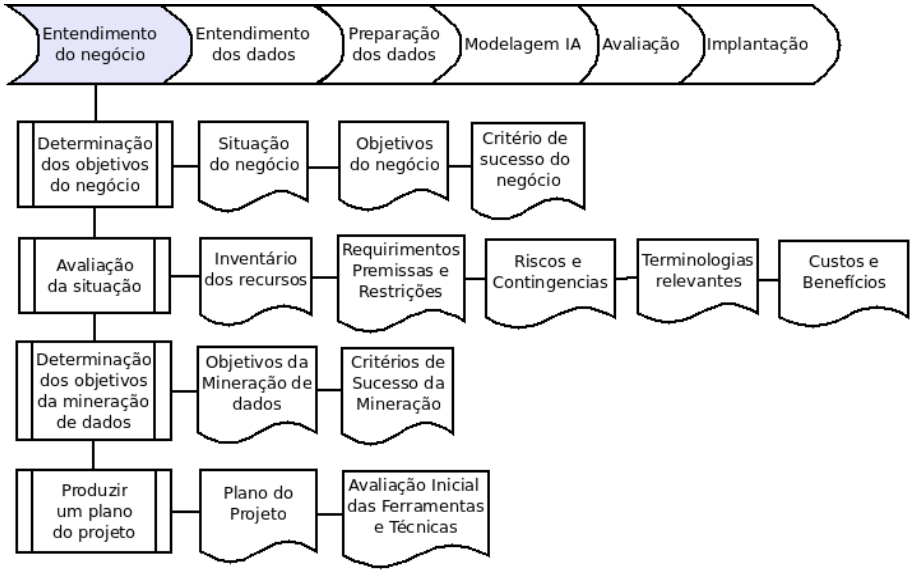
Figura 2.2: O padrão CRISP-DM (1)



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" (CHAPMAN; KERBER; WIRTH et al, 2000, 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.

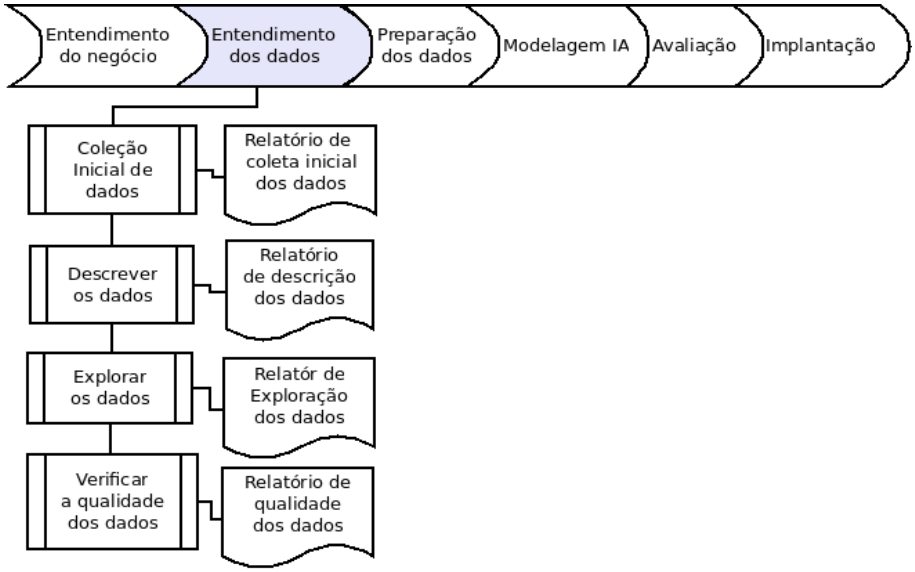
Figura 2.3: Entendimento do negócio



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.

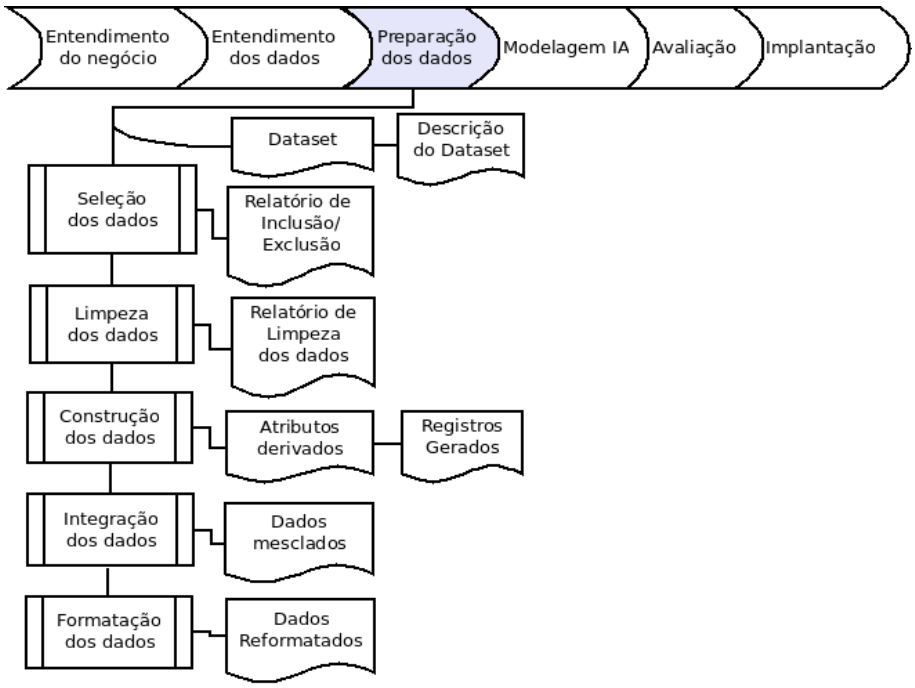
Figura 2.4: Entendimento dos dados



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 (8), vide equação da Entropia: $H_x = -\sum_{\forall x \in X} P(x) log_2 P(x)$ Onde H_x é a medida de entropia, x um atributo do conjunto de variáveis X de variáveis. 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 (9): $H_{Y|X} = \sum_x P(x) H(Y|X = x) = -\sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y|x) log_2 P(y|x)$.

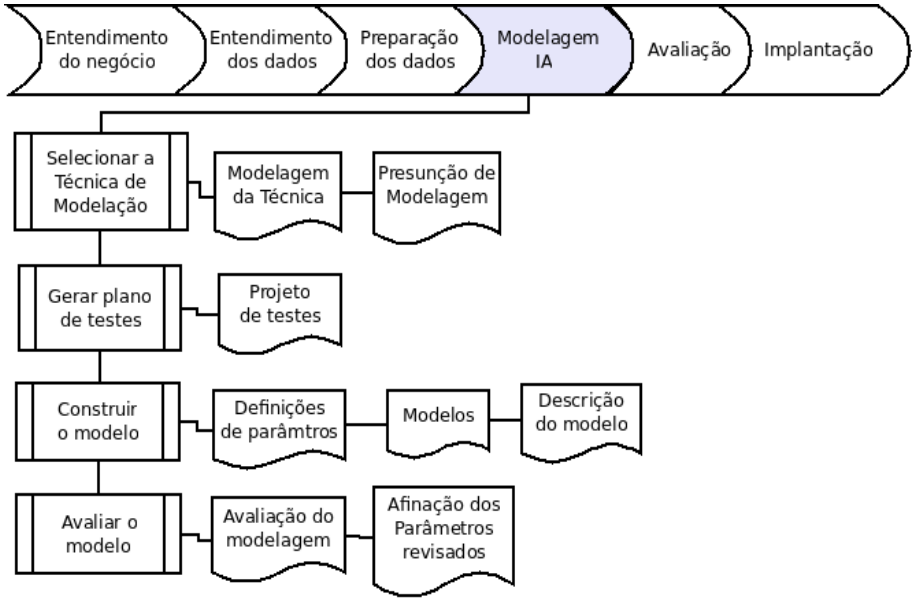
Figura 2.5: Preparação dos dados



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.

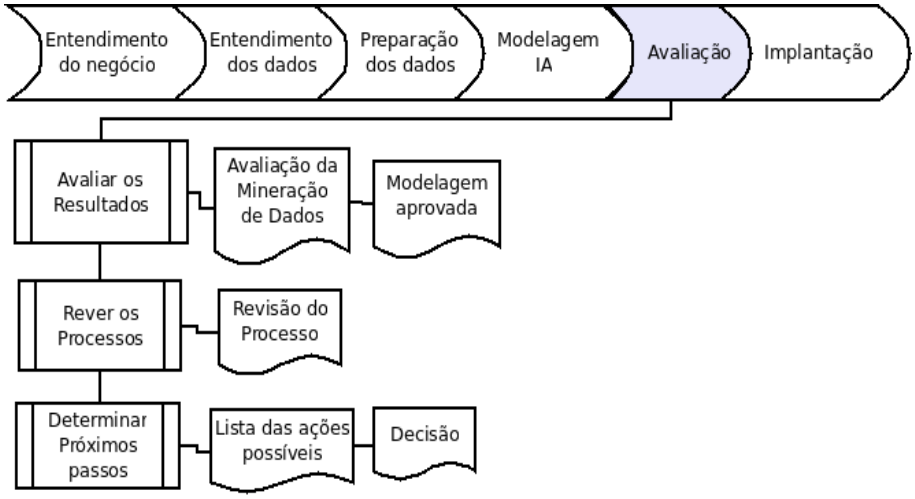
Figura 2.6: Modelagem IA



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.

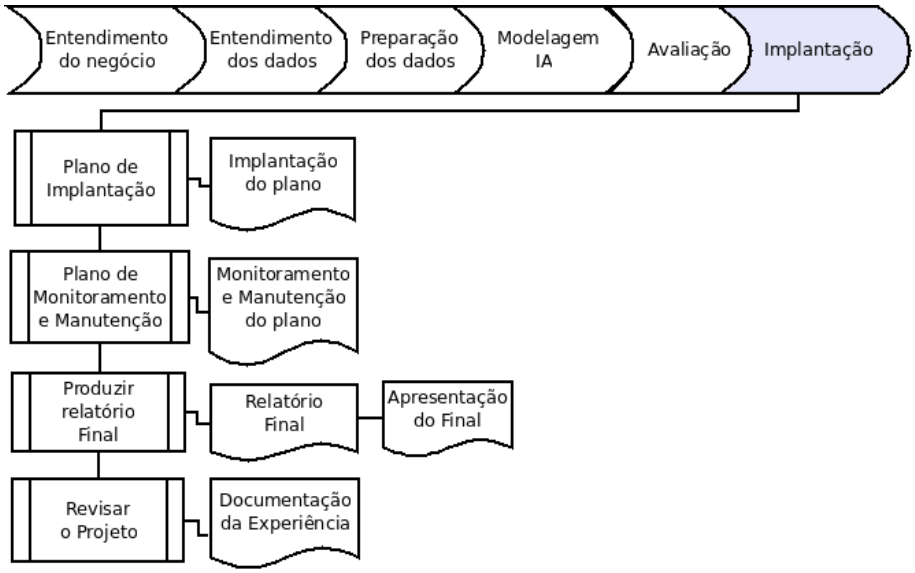
Figura 2.7: Avaliação do modelo



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 especifidades 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 (7).

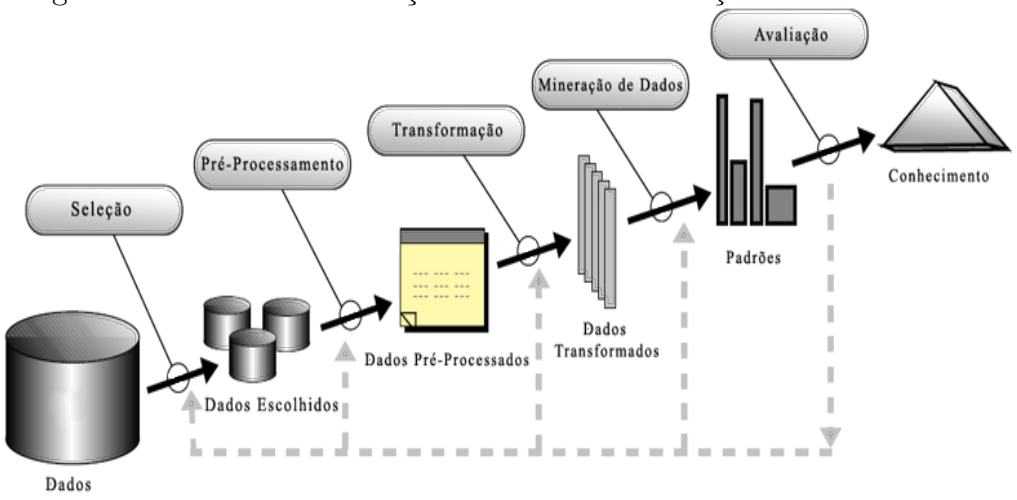
A mineração se distingue das técnicas estatísticas pelo fato de que não trabalha com dados hipotéticos, mas se apoia nos próprios dados para extrair os padrões (CASTANHEIRA, 2008).

FAYYAD (1996), 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 (1996), pode conduzir à descoberta de dados sem significado e padrões inválidos.

Existem vários tipos de dados e informações nesses repositórios que podem ser minerados, contudo esses dados, inicialmente são selecionados e agrupados, a seguir passam por uma fase de preprocessamento, que consiste em tratá-los de forma a prepará-los para a mineração. Essa fase é de fundamental importância na estruturação dos dados, uma vez que em grandes volumes de dados, também conhecido "Datawarehouse", 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 (10), 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



A origem dos dados, os "inputs" estão representados na figura onde se lê "Data" este está repleto de *missing data* e/ou dados inconsistentes, conhecidos como dados não estru-

turados. 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 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 Botnets [1] ou "Swarm Intelligence".]

No balão "Preprocessing" 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 (11). 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 utlizadas. Em "Transformation" 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 "Data Mining", 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) e "Deep Learning". Algumas técnicas de mineração de dados são fortemente influenciadas pelas informações na entrada (input), como as Árvores de decisão (9). As Redes Neurais, dependendo da quantidade de variáveis de entrada, paderão ter milhares de neurônios na camada intermediária, o que inviabilizaria essa metaheurística [2].

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 (12), essa técnica de extração de conhecimento chama-se *Knowledge Discovery Databases* (KDD). O KDD utiliza métodos de Aprendizagem de Máquina para efetuar essa extração.

[1] Botnet é citado no sentido da coleta de informações

[2] Metaheurística são heurísticas aplicadas em problemas onde os custos computacionais não são tratáveis em tempo polionomial, 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 Tipos de Aprendizagem

Quando se fala em algoritmos de IA, adentramos no campo de Aprendizagens e Máquinas. É o princípio da aprensizagem 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 (8). 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 (13) 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 (RUSSEL; NORVIG, 2004), 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 (13).

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

As técnicas algorítmicas apresentadas nas seções subsequentes são parte da grande família de algorítimos 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 (10), 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 predições (10). A seção seguinte analisa as tarefas preditivas.

2.4.1 Aprendizagem de Máquina

Aprendizagem de Máquina ou "Machine Learning" são métodos para analisar dados de forma automatizada e interativa. Segundo Shalev-Shwartz & Ben-David (14) em seu livro "Understandig Machine Learning: From Theory to Algorithms" o termo Aprendizagem de Máquina refere-se à detecção automatizada de Padrões de dados.

Para Nilsson (15) 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 sentimos que 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" (15).

2.4.2 Naïve Bayes

Naïve Bayes é uma classe de algoritmos baseado no teorema da probabilidade condicional de Bayes, serve para rotular classes de variáveis independentes. Em mineração de dados variáveis independentes explicam a variável dependente para fazer predi-

ção. Este classificador tem sido muito empregado para classificar documentos e detectar spam em mensagens (**??**). A probabilidade condicional pode ser explicada por um vetor $x = (x_1, x2, ..., xn)$ que se representa n características (variáveis independentes) que se atribui a estas instâncias de probabilidades $p(C_k|x_i, ..., x_[n])$ para cada K possível ter vindo da classe $C_[k]$. Aplicando o teorema de Bayes da probabilidade condicional temos:

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

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:

$$p(posteriori) = \frac{p(priori) * verossimilhana}{evidncia} \tag{2.2}$$

2.4.3 Aprendizagem Bayesiana

Baseado no teorema de Bayes, dado um conjunto de variáveis aleatórias $\omega = x_1, x_2, .., x_n$ a variável aleatória H (hipótese) denota o tipo de ω, com valores possíveis para $h_1, h2, ..., h_n$. A medida que são inspecionadas as variáveis, são revelados os dados $D_1, D_2, ..., D_n$, 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 deconhecida de X, temos:

$$P(h_i|d) = \sum_i P(X|d, h_i)P(h_i|d) = \sum_i P(X|h_i)P(h_i|d) \tag{2.3}$$

onde cada hipótese h_i determina uma distribuição de probabilidades sobre a variável X (8).

2.4.4 Onde e o que aprender

Historicamente os tipos de aprendizagem computacional estão relacionados em "o que" há para ser aprendido (15). Primeiramente para escolher o que aprender definiremos de "onde" ou sobre quais dados se aprender. Fornecemos um conjunto de treinamento para depois testar o conhecimento aprendido em um conjunto de teste.

2.5 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 predizer 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 $them$ class $= C_1$ ou if L $them$ C_1
- Uma regra de indução seria: if L $them$ R que por sua vez produz novas regras

$age(X, \text{“}youth\text{”})$ AND $income(X, \text{“}high\text{”}) \rightarrow classe(X, \text{“}A\text{”})$

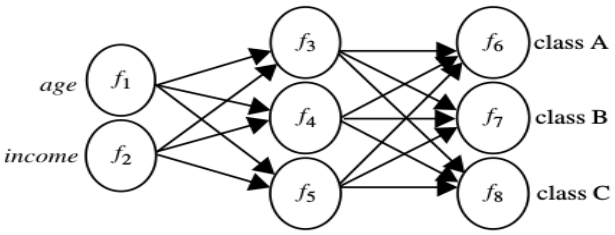
$age(X, \text{“}youth\text{”})$ AND $income(X, \text{“}low\text{”}) \rightarrow classe(X, \text{“}B\text{”})$

$age(X, \text{“}middle - aged\text{”}) \rightarrow classe(X, \text{“}B\text{”})$

$age(X, \text{“}senior\text{”}) \rightarrow classe(X, \text{“}B\text{”})$

A figura a seguir representa uma rede neural com as mesmas características da árvore de decisão anterior:

Figura 2.10: 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 do é fortemente influenciáveis pelas variáveis escolhidas, (9) isso pode representar o desafio maior para implementar esta técnica.

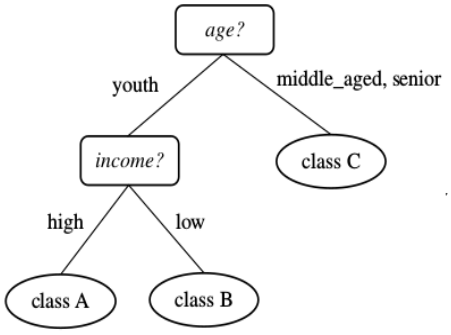
Outro problema que pode ser encontrado em algoritmos de aprendizagem é o “overfitting” ou superadaptação aos modelos. Segundo RUSSEL E NORVIG (2004) [3] o “overfitting” ocorre quando o número atributos é grande.

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

2.6 Árvore de Decisão

No âmbito da inteligência artificial, quando se trata de algoritmos de aprendizagem, as árvores de decisão têm aparecido como uma das ferramentas que possibilitam resultados satisfatórios e de fácil interpretação. Esse parece ser um dos motivos pelos quais tem sido frequente o uso desse algoritmo em pesquisas em IA. Particularmente, nessa pequisa, as árvores de decisão possibilitaram grandes avanços na proposição do modelo de predição, conforme pode ser observado no capítulo dedicado aos resultados. Uma árvore de decisão tem como entrada um conjunto de atributos e como saída, uma decisão. Os atributos podem, ainda, ser de dupla natureza: discretos ou contínuos. O mesmo acontece com as saídas, cujo resultado pode ser uma função de valores discretos - aprendizagem de classificação - ou de valores contínuos - aprendizagem de regressão (RUSSEL; NORVIG, 2004). A decisão gerada aparece em função de uma sequência de testes executados, estando cada um deles relacionados a um nó na árvore. As ramificações que decorrem dos testes são o resultado encontrado a partir da sua realização. Han e Kamber (16) definem indução por árvore de decisão como a aprendizagem de árvore de decisão a partir de classes rotuladas nas tuplas de treinamento. A estrutura da árvore de decisão é semelhante a um fluxograma, onde cada nó interno (não-folha) indica um teste de atributo, cada ramo representa o resultado de um teste e cada nó da folha possui um rótulo de classe. O nó de nível mais superior é chamado de nó-raiz. O exemplo a seguir ilustra uma árvore de decisão simples, onde se vê o nó-raiz e suas ramificações. Tomamos para esse exemplo os dados dessa pesquisa, a serem discutidos de forma mais detalhada posteriormente.

Figura 2.11: Árvore de decisão



Fonte: Han, J. and Kamber, M.

Para Ian e Frank (17), as árvores de decisão 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. As árvores de decisão podem ser classificadas em dois tipos: árvores de regressão (regression trees), que são utilizadas para estimar atributos numéricos, e árvores de classificação (classification trees), usadas para análise de variáveis categóricas. Uma árvore de decisão obedece á regra básica "se-então", partindo do nós (SE) para as folhas (ENTÃO). O conhecimento é representado em cada nó, que apresenta pelo menos duas saídas ou ramificações possíveis, podendo 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 consideradas diz respeito a como propor as regras de forma adequada e relevante para a geração da árvore. É função do algoritmo identificar o melhor

atributo, que será responsável por criar o nó de decisão. As ligações entre os nós representam valores possíveis do teste do nó superior e as folhas indicam a classe (categoria) a qual o registro pertence (CAMILO, SILVA, 2009).

2.6.1 Tipos de Árvores de decisão

O algorítimo *C4.5* é considerado um exemplo clássico de método de indução de árvores de decisão. O *C4.5* (18) foi inspirado no algoritmo *ID3* (19), que produz árvores de decisão a partir de uma abordagem recursiva de particionamento de um conjunto de dados, utilizando conceitos e medidas da Teoria da Informação (**??**).

As árvores de decisão têm uma característica peculiar, a saída do modelo de predição (o output), com regras se – então é claramente perceptivel por analistas humanos. Essa qualidade é utilizada para interpretar os resultados.

2.7 Regressão

2.7.1 Tipos de regressão aplicadas a mineração de dados

incluir e explicar ao menos 2 tipos

2.7.2 Regressão Linear

falta fazer

2.8 Redes Neurais

2.8.1 Introdução

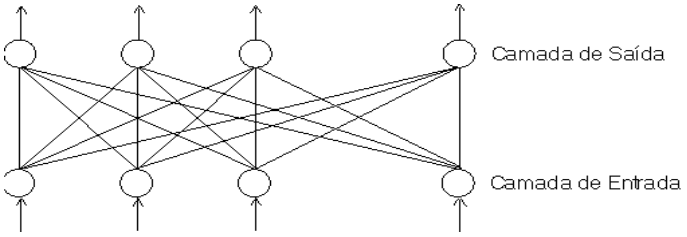
O cérebro humano possui cerca de 10 bilhões de neurônios, que são responsáveis pelo funcionamento do organismo. Esses neurônios se conectam entre si, através de sinapses, formando uma Rede Neural capaz de armazenar e processar grande quantidade de informações.

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 (20) propuseram um modelo simplificado de funcionamento do cérebro humano e, a partir dai 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 Mavin Minsky. Todavia, o primeiro neurocomputador que obteve sucesso surgiu entre 1957 e 1958, o Mark I Perceptron, criado por Rosenblatt, Wightman e colaboradores (20). 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.

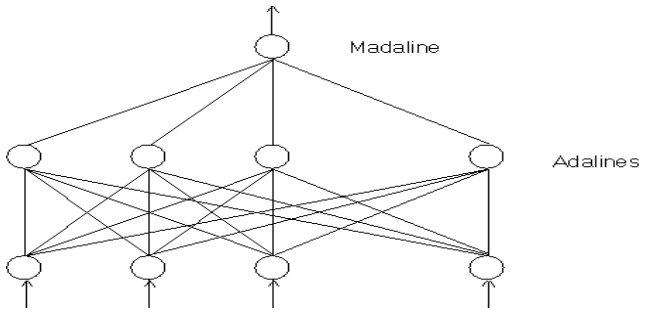
Figura 2.12: Percepton de Rosenblatt



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 abaixo:

Figura 2.13: 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.

2.8.2 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) (21), 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.

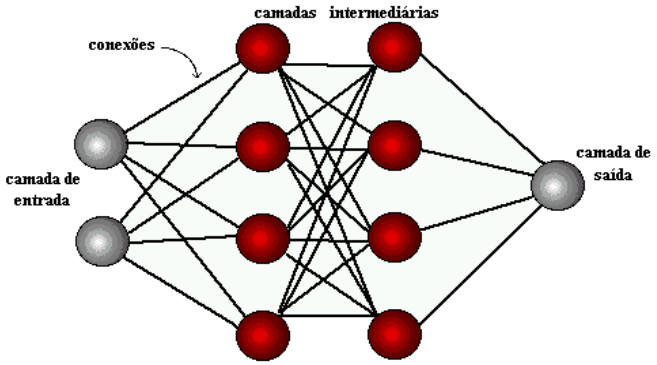
Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento, enquanto que o cérebro de um mamífero pode ter muitos bilhões de neurônios. Uma Rede Neural Artificial (RNA) é um sistema que de neurônios que estabelecem conexões sinápticas, que possuem neurônios de entrada, 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 (22).

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 2.14: Um arranjo de Perceptrons em camadas



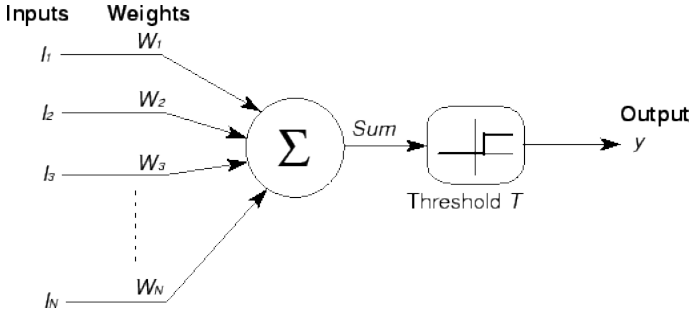
Fonte: http://www.din.uem.br/ia/neurais/neural. Acessado em: 01/10/2016

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 (1943)

Figura 2.15: Perceptron de McCulloch e Pitts



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

Esses canais são os *inputs*, $I_1, I_2, I_3, ..., I_n$, 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:

- A função calculada y é uma função **discriminativa** (classificação) com $y = 0$ e $y = 1$.
  - $net = x_1w_1 + x_2w_2 + ... + x_iw_i + x_dw_d = \sum x_i.w_i = W.X = W^TX$
  - $onde\ w = \{+1, -1\}$

  - $saida = y = f(net)$
  - $f(net) = \begin{cases} 1, & se\ net \geq \mu \\ 0, & se\ net < \mu \end{cases}$
- Fronteira de Decisão:
  - Determina o ponto que separa os dados que vêm de -∞ e +∞
  - Argumento de $f(net)$ é igual a zero: $\sum w_i x_i - \mu \Rightarrow w.x = 0$

O modelo McCulloch e Pitts (1943) 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 "passear" 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.

2.8.3 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.

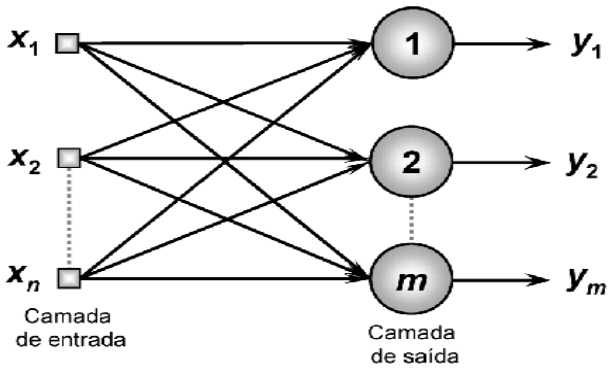
Uma vez que existem distintas possibilidades de aplicação e desenvolvimento de uma RNA, existem, igualmente, diversas maneiras de classificá-las (23). Trataremos de algumas dessas nesse tópico.

(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 Percepton e Adaline

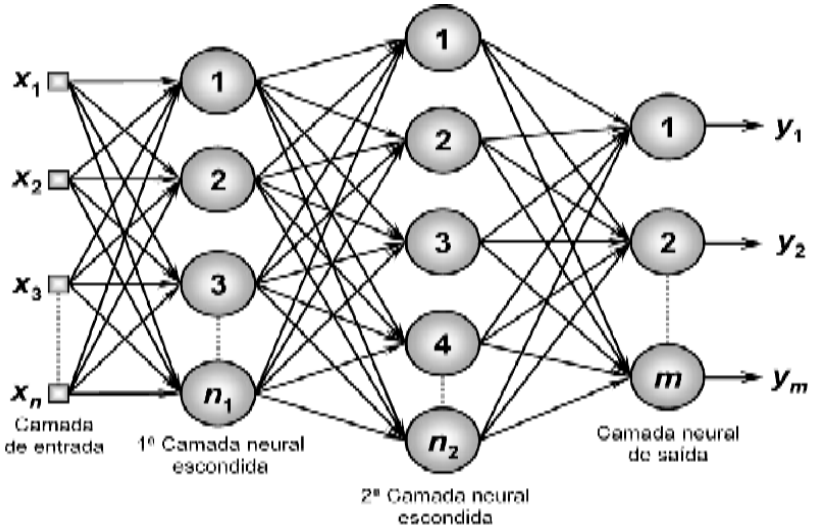
Figura 2.16: Perceptron e Adaline



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)

Figura 2.17: Perceptron Multicamadas



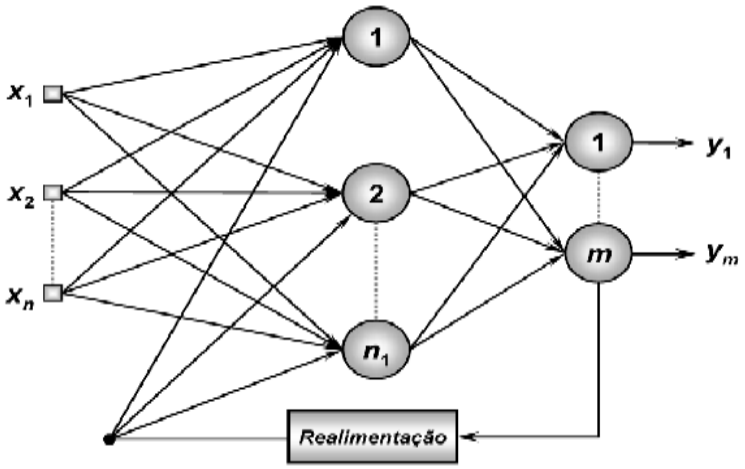
Fonte: http://www.dkriesel.com. Acessado em: 10/10/2016

Alguns autores ainda fazem referência às redes recorrentes ou realimentadas (iii), como a rede de Hopfield e a Perceptron multicamadas (24). 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 Kohone.

(iii) Exemplo de rede recorrente ou realimentada: "backpropagation"

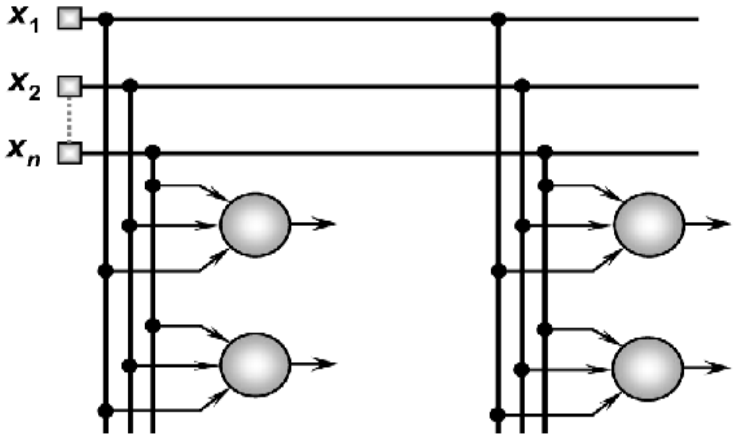
(iv) Exemplo de redes de estrutura reticulada

Figura 2.18: Perceptron com Realimentação



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

Figura 2.19: Rede Kohone



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.

2.8.4 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 contato, por exemplos, por analogias, por exploração, por descoberta, sobretudo

por uma mistura entre os últimos três (25).

- 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 (25). 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 (25). A descrição dessa regra foi apresentada pelo seu propositor, 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.

2.9 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 (26), 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 (27), 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 (28) 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 2.2: Matriz de Confusão

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

Fonte: (29)

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

Tabela 2.3: Matriz modelo de Confusão

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

Fonte: (29)

De acordo com as probabilidades condicionais temos:

$$P(X,Y) = P(X|Y).P(Y) = P(Y|X).P(X) \tag{2.4}$$

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 sobrescrita 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,$\bar{X}$)).

2.10 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 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 decisões 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 empresa 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"(30).

Tabela 2.4: 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}

2.10.1 O Twitter

A ideia inicial do Twitter, segundo seus fundadores, era que essa rede se comportasse como um "SMS da Internet" (31). As informações são enviadas aos usuários, conhecidas como twittes, em tempo real e também enviadas aos usuários seguidores que tenham assinado para recebê-las.

Twitter, 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 a sua atenção, como eventos de cunho politico, catástrofes naturais, acidentes com vítimas graves, atentados, dentre outros. O Twitter se caracteriza 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. Os dados são capturados nesse formato. (A FONTE – BIG SOCIAL DATA: PRINCÍPIOS SOBRE COLETA, TRATAMENTO E ANÁLISE DE DADOS SOCIAIS). A cada dia centenas de terabytes são inseridos no Hadoop data warehouse. 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 assinado para recebê-las. 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 seguido serve como principal mecanismo para espalhar informações nessas redes.

Analisar o conteúdo e minerar textos é um procedimento frequentemente utilizado em pesquisas envolvendo redes sociais, para realização de análise de texto dos usuários em geral e como suporte para tomada de decisão (Abrahams, Fan, Wang, Zhang & Jiao, 2015). É comum, ainda, aplicar mineração de textos em bibliotecas e outras instituições. Isso implica em rastrear tópicos, extrair informações, agrupar, categorizar (Fan, Wallace, Rich e Zhang, 2006). Em recente artigo, Sandhu (2015) 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 os serviços de informação. Similarmente Zhang e Gu (2011) alegaram que minerar conhecimento sobre os clientes é importante para as bibliotecas acadêmicas. A abordagem da mineração de textos para os dados das mídias sociais tem sido utilizada em muitos campos, como negócios, ciência da saúde (Sarker et al., 2015) ciência política. O uso da abordagem de mineração de textos está a ser aplicada na literatura sobre biblioteca e Ciência da Informação. Zang e Gu (2011) analisaram a aplicação da mineração de textos nos dados das redes sociais.

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 Twitter as a Corpus for Sentiment Analysis and Opinion Mining Alexander, 2010 Sentiment Analysis on Twitter, Akshi Kumar and Teeja Mary Sebastian, 2012)

(FONTE: Artigo "Library & Information Science Research) Outro estudo (FONTE ??) 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. (Duggan & Smith, 2013). Um relatório publicado em 2014 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 (Another Pew). 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 (Mitchell & Page, 2013). 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) Heneycutt e Hering (2009) identificaram 11 categorias de conteúdo propagados nos tweets: sobre destinatários, anúncio/propaganda, encorajamento "exhort" (que traduzimos aqui como "auto-ajuda"), informações para outros, informações para si próprio, meta comentários, uso de mídia, opinião, outras experiências, auto experiência, e solicitação de informações. Naaman, Boase e Lai (2010), por sua vez, classificaram os conteúdos 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 (i.e. Adkins & Budd, 2006, Cronin & Overflet, 1994; Cunningham & Dillon, 1997; Lee, 2003). 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 (Haslam et al., 2008; Hinnant et al. 2012; Stivilia et al., 2011). 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 acetiação de wrbsites (Brin & Page, 1998), em decorrência do número de seguidores, da quantidade de menções feitas, de retweets, por exemplo.

(FONTE: SE É POSITIVO OU NEGATIVO)

Cha, Haddadi e Benevenuto (2010) avaliaram a influência dos usuários no Twitter, 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). Em outro estudo, André et al. (2012) utilizaram a abordagem "crowdsourcing" para identificar que tipo de tweets chamam atenção e são os preferidos dos seguidores, como também, os que são mais rejeitados. Eles encontraram uma preferência dos usuários em fazer perguntas aos seguidores, bem como em partilhar informações e autopromoção (que acreditamos que diz respeito a algo como falar de si próprio, de seus sentimentos, percepções e ideias). Os usuários, por outro lado, demonstram não ter preferência por tweets categorizados como: manutenção de presença, conversações e atualizações do "status" corrente do usuário.

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. Do ponto de vista metodológico .....

Na pesquisa conduzida por Suh, Hong e Chi (2010), eles 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.

(REESCREVER)

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. Consequentemente métodos não paramétricos foram usados para examinar a relação entre um tweet as características da conta. O texte Krusal-Walls apresentou uma diferença significativa entre o conteúdo da categoria do tweet para o númro de favoritos $X = 15.11, df = 8, p < 0,057$. O teste Sperman, 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 retweetar 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 que também se interessou por conteúdos de natureza mais acadêmica, conduzido por (FONTE - ANO),

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 "compartilhar informações e conhecimentos, incrementar serviços e promoções, interação com estudantes usuários das bibliotecas, a um custo mínimo" (Chue & Due, 2013, p. 72). Tal prática têm promovido considerável mudança na interação com usuários e relacionamento com os clientes (Del Bosque, Leif e Skaf, 2012; Cavanagh, 2015), tendo sido utilizado frequentemente como alternativa para estabelecer uma conexão personalizada com os seus usuários (Boaten & Quan, 2014). 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. (FONTE: Abraham et al., 2015)

(Stieglitz & Dang-Xuan, 2013)

O corpus de dados desta pesquisa foi obtido a partir da "timeline" de dez bibliotecas acadêmicas (i.e. todos Twittes 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, e a seleção restringiu-se às universidades de língua inglesa e a apenas uma biblioteca, para o caso de a universidade ter mais que uma (no artigo original é a tabela 1). 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 prepocessing" – o grupo de dados recuperado foi tratado, para reduzir os "ruídos", seguindo uma abordagem consistente com outros estudos de mineração em textos, tal como Ralston, O'Neil, Wigmore e Harrison (2014) e Yoon, Elhadad e Bakken (2013). O processo contempla 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 analisar os históricos de tweets da bibliotecas escolheu-se a análise de conteúdo. A frequência dos tweets, retweets e distribuição dos tweets e retweets foi identificado e contabilizado. Em seguida, os tweets marcados com o PamTaT, uma ferramenta "text mining" desenvolvida por Pamplin Collage do Instituto Politécnico de Negócios de Virgínia da Universidade Estadual de Virgínia (Poderia colocar essa observação numa nota de rodapé). O PamTaT é baseado na interface do Microsoft Excel para Python nltk – "natural language processing framework" (Bird, Loper & Klien, 2009) e permite a análise de grande volume de textos pelos usuários finais, não necessitando de conhecimento de programação da linguagem Python. O PamTaT para serve para determinar a frequência de palavras simples (unigrams), de duas palavras (bigrams) e sequência de três palavras (trigrams) que aparecem no texto fonte. Com isso permite desenvolver uma matriz de frequência de termos-tweet, mostrando como sequências de palavras simples e múltiplas palavras (n-grams) são usadas pela biblioteca acadêmica selecionada. Também foi utilizado o Harvard General Inquirer (Stone et al. 1966) 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 identificou duas centenas de palavras incluindo, por exemplo: como positivo, negativo, 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. Hurtwitz (2002) forneceu uma lista abrangente de categorias de palavras reconhecidas pela Harvard General Inquirer e apresentou uma lista completa de palavras específicas que pertenciam a cada categoria de palavras.

Imagem tabela 1 do twitter

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 biblioteca da Cambrige University, respectivamente. A biblioteca da Universidade da Califórnia San Diego tem a conta do Twitter mais antiga (maio de 2008), mas apresenta um pequeno número de tweets, comparativamente à biblioteca da Stanford University, que começou no Twitter com uma conta em abril de 2012, mas possui o segundo maior número de tweets. A análise estatística da tabela segue comparando esses dados, e por ser visto no artigo aqui referendado. A tabela 2, por sua vez, apresenta as menções, hashtags, e retweets das bibliotecas.

A análise do conteúdo dos Tweets foi desenvolvida da seguinte maneira: tomando a frequência de unigramas (palavras únicas) observou-se (fig. 4), 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 anúncio para abertura do espaço para os estudantes, exposições, abertura da casa (biblioteca), etc. O segundo termo mais frequente foi "research" (pesquisa), que foi usado 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 incluidos em tweets realcionados 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".

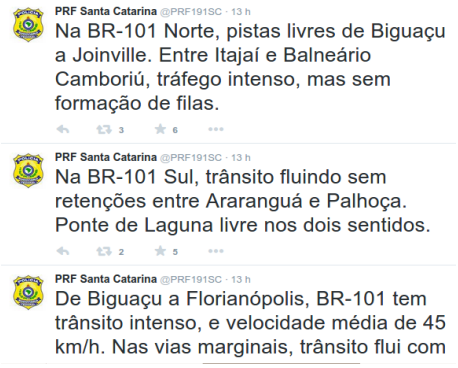
Imagem tabela 2 do twitter

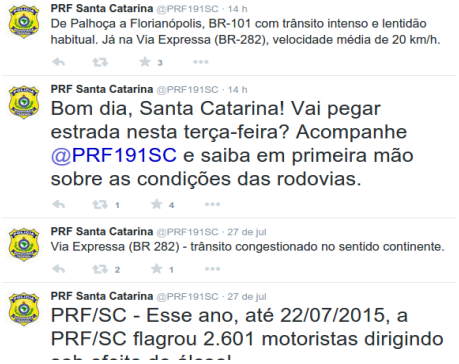
Imagem tabela 3 do twitter

A fig-5, por sua vez, mostra a distribuição dos bigramas (sequência de duas palavras) no conjunto de tweets. Observa-se que o mais frequente bigrama foi "special collections". O segundo mais popular bigrama é "open access", que foi usado em diferentes contextos tal como política de acesso, publicidade, recursos, trenamentos e workshops, dicas e orientação, serviços, eventos e notícias. O resultado mostrou a enfase colocada na iniciativa de promover "open access" com as instituições acadêmicas. O terceiro maior bigrama foi "reading room", relative à atividade de suporte aos estudantes com as instituições acadêmicas. As salas são um dos mais importantes espaços da biblioteca, usadas para leitura e estudo. Os tweets são tipicamente relados para noticias da abertura e fechamento das salas de letura "reading rooms". A fig-6 mostra os mais importantes trigramas (sequência de três palavras), dos quais destacou-se o trigrama "save the date". Essa expressão é utilizada para requerer especial atenção dos seguidores para os eventos importantes que estavam para acontecer. Este trigrama é seguido, como o Segundo que mais aparece, 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)

Imagem tabela 4 do twitter

A seguir pode-se verificar uma sequência de twittes da Polícia Rodoviária Federal de Santa Catarina:





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 o transito está intenso. Isso sugere 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). Essa informação sugere que deve ser pensada uma rota alternativa, caso o congestionamento persista por muito tempo.

Outra rede social conhecida pelos condutores de veículos é o Waze. O Waze é um aplicativo de navegação para o trânsito, funciona em aparelhos celulares e tablets. Os utilizadores desse aplicativo são conhecidos como wazers e compartilham informações sobre o trânsito, em tempo real. Toda via, 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 em postar informações, não há o que se compartilhar. Outro problema levantado com o 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 próximos 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://vejoaovivo.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.

2.10.2 Data Mining - Text Data Mining

Minerar dados em texto nas redes sociais não é uma tarefa atômica, devendo ser divida em várias etapas, com processos específicos em cada uma delas, como descrito anteriormente. Extrair conhecimento dos dados não processados não faz sentido, tratá-los apenas "per si" exige muito trabalho de IA, como Mineração de dados em textos. A Mineração em textos é inspirada em técnicas de "Machine Learning" (11). Contudo analisar textos é basicamente entender o significado do texto, baseado em regras de associação lógica. O mapa mental a seguir mostra um modelo de análise de texto feito por seres humanos.

Figura 2.20: Mapa mental da Mineração em textos

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; do ponto de vista prático pela proposição de um modelo que integre predição à API de mapas de posicionamento global, fornecendo informação suficiente a um gestor para decidir quando e por onde enviar uma carga por determinada rodovia que apresente retenções crescentes de logística de cargas.

As soluções disponíveis que existem tais como; Google Maps, Waze e outros dessa natureza somente exibem informações momentâneas, produzidas e compartilhadas pelos utilizadores dos aplicativos ou por informações provindas de GPS, contudo não analisam dados históricos dessas rodovias nem fazem predições 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 muita informação relevante como por exemplo o anúncio de uma paralisação que ocorrerá daqui a uma semana, a PRF de Pernambuco é outro contribuidor permanente; com seu canal no Twitter: @PRF191PE fornece diariamente informação das rodovias além de dados estatísticos.

A monitoração de redes sociais é feita por Mineração de dados em textos conhecida como ou Mineração em Textos ou "Text Mining"(TM). A TM a princípio foi executa somente no Twitter no canal PRF191PE. Agentes da PRF disseram que os protestos quando feitos dentro da lei devem ser informados à PRF, com dia e hora marcados antecipadamente. Para ter-se acesso aos twittes postados da PRF os usuários do canal, inclusive a própria PRF, postam comentários diretamente na API do Twitter ou por um navegador. Para nossa pesquisa as palavras chaves tais como: protestos, acidentes, paralisação, são de suma importância.

Uma vez capturadas esses tweets são tratadas por Mineração em Textos e analisados instantaneamente por algoritmos de I.A. A técnica utilizada para analise dos textos é o Processamento de Linguagem Natural (PLN). O algoritmo escolhido foi Naive Bayes por ser um classificador rápido e eficiente e por ter sido utilizado na primeira fase desta pesquisa, lá serviu como comparativo à Árvore de Decisão.

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 completa o ciclo todo do processo 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 esses comportamento tem a ver com o "modus operandi" dos utilizadores. A priori especulou-se sobre possíveis erros de traçados e outros que possam 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 falas e facilidade de interpretação dos dados a Árvore de Decisão e o Naïve Bayes se destacaram por não necessitar de nenhum requisito extra para entender e interpretar os resultados [referencia].

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. As redes neurais têm capacidade de generalização e especificidade em modelos de previsão[referencia].

A extrapolação do modelo preditivo ocorre quando este se integra o modelo preditivo à uma estrutura dinâmica a serem exibidas em mapas vetoriais, dado um espaço temporal pré-determinado por um agente; o utilizador.

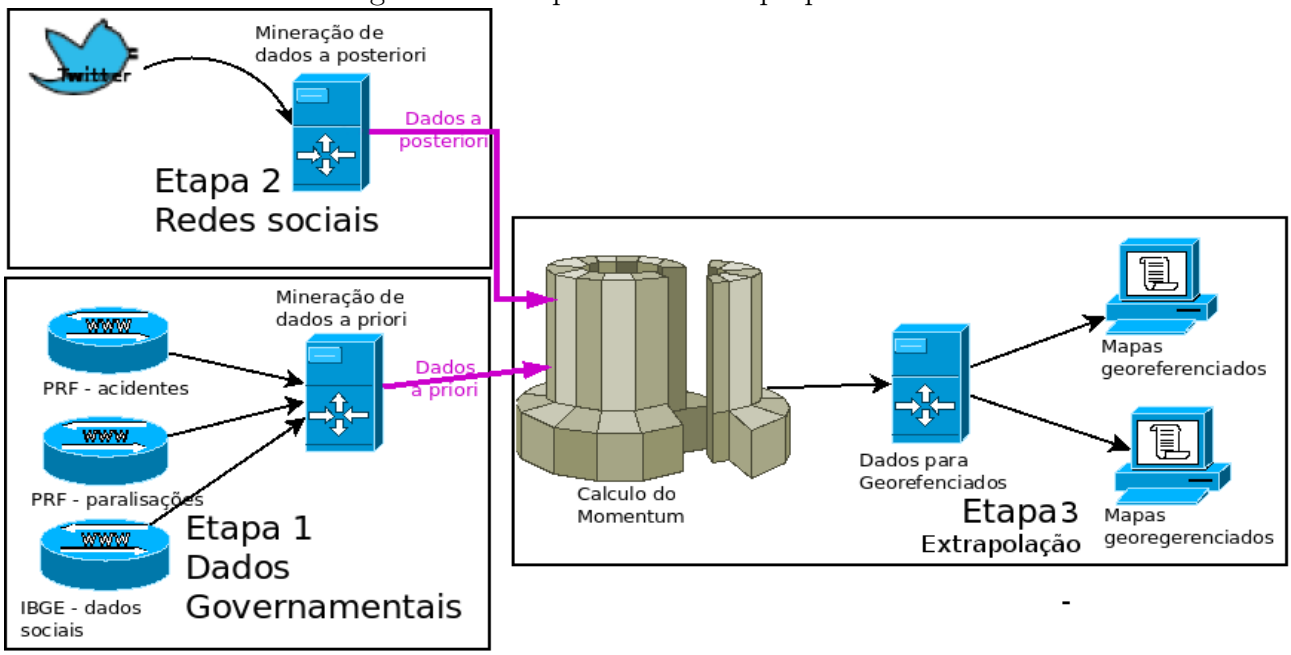
Através de API's os mapas vetoriais permitem a geolocalização dos pontos classificados ou os pontos onde haverá grande número de retenções, conhecido no meio da logística de cargas como **gargalo**.

A API do Google-Maps é o "front end", foi escolhida por permitir maior pela portabilidade e simplicidade para integração da estrutura dinâmica com a preditiva.

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 geram poucos dados, porém são eficazes; 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 conexões à Internet, contudo esta rede social tem uma crescente quantidade de postagens no formato imagens, isso dificulta a mineração em textos. A API do Twitter tem a finalidade de complementar o modelo preditivo, contudo acrescenta um padrão dinâmico aos mapas vetoriais às redes sociais. Esta "interface" é responsável por fornecer informações instantâneas à terceira etapa do modelo proposto, servindo de "busca local" das informações mais recentes das redes sociais, relativas à trechos das rodovias; os "feeds" da "timeline" do Twitter, os twetts fornecem dados que serão minerados e interpretados fornecendo informações instantâneas ao modelo proposto.

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

As técnicas preditivas tradicionais que contemplam análise de grandes massas de dados como base homogêneas. Não existe uma técnica de mineração que generalize os mais diversos ambientes preditivos, mas sim um "pool" dessas técnicas onde uma complementa outra como identificamos nos artigos de ZENG, HUANG, PEI *et al* (2016), POSSAS, BAV e CARVALHO (1998), e BESHAS e HILL (2010).

Na fase de transformação de dados, da primeira etapa, onde são criadas novas variáveis, a proximidade entre as bases heterogêneas foi conseguido utilizando de regras de indução da lógica proposicional (8). Nesta pesquisa, bases heterogêneas foram integralizadas num única grande conjunto de dados o "data set". As variáveis desse "data set" são consideradas variáveis independentes, foram preservadas as com maior relevância ou as que continham a maior quantidade de conhecimento embutido. construídas novas, nas bases onde não haviam correspondência, respeitando a lógica do negócio.

A tabela a seguir descreve as variáveis originais na base de dados de acidentes da PRF

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. Esses dados nos foram cedidos após alguns procedimento 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 (2012) destacam a não padronização e não aceitação dos dados pela comunidade internacional. EAVES, D. (2009) sugere que os dados sejam disponibilizados na 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 dessas 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. Mediu-se a acurácia dos classificadores comparando-se uma técnica algorítmica com a outra. A variável "BRajustada" mostrou ser a variável melhor para exprimir o nó rais da Árvore de Decisão porque classificou os dados com o menos número de falsos positivos e com o maior número de verdadeiros positivos, obtendo curvas ROC com índices acima de 0.90. O Naïve Bayes também obteve índices de acurácia próximos a este.

As técnicas como Redes Neurais Artificias (MLP) (32), Árvores de decisão (CART) (16), fornecem visão generalizada dos fatores preponderantes, levantando padrões ocultos nos dados. Esta fase é conhecida como Aprendizagem de Máquina (acrônimo de Machine Learning)

a Redes Neurais Artificias do tipo *Multi Layer Perceptron* – (MLP) têm capacidade de receber várias entradas ao mesmo tempo e distribuí-las de maneira organizada, além são simples de implementar e trazem resultados satisfatórios em grandes bases de dados.

b Árvores de decisão do tipo *Classification and Regression Tree* – (CART) foi empregue por Pakgohar et al no artigo *The role of human factor in incident and severity of road crashes based on the CART and LR regression a data mining approach* para classificar acidentes com nível de acurácia próximo aos 80%

c Regressão logística tipo *Multinomial Logistic Regression* – (MLR) fornece a possibilidade de aprofundamento em vários níveis de busca sendo a mais apropriada, já que Regressão logística tradicional não permite aprofundamento desse tipo no espaço de busca.

3.4 Reflexão sobre as tecnologias utilizadas no modelo preditivo a posteriori

(Artigo "Library & Information Science Research") Tecnologia Para encaminhar as pesquisas, os autores usaram uma abordagem indutiva. Os mais recentes 1200 tweets foram coletados da conta (Tweet de uma Universidade) de todos os membros da Association of American Universities (AAU). Os dados foram coletados e processados usando

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

o Twitter API com o twitter4j (biblioteca Java) e adicionados a um código java desenvolvido pelo autor. ...Foi reduzido a uma amostra do percentual da academia.... Os dados coletados foram preprocessados para preparar a analise. Retweets foram removidos das amostras. Em seguida tweets com pouco conteúdo foram removidos, tal como breves agradecimentos (e.g. "welcome") breves comentários de algum tweet (e.g. "lol") pequenos encorajamentos (e.g. "keep going") e conversas pessoais foram removidas das amostras. Isto reduziu a amostra de 1200 para 752 tweets. Estes tweets foram em seguida categorizados em nove categorias... Dos 752 tweets analisados, 271 tinham ao 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, as seis contas de Twitter das bibliotecas analisadas enviaram 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. Consequentemente métodos não paramétricos foram usados para examinar a relação entre um tweet as características da conta. O teste Kruskal-Walls revelou uma diferença estatística significativa entre o conteúdo da categoria do tweet para o número de favoritos ($X^2 = 15.11$, df =8, $p < 0{,}057$). O teste Spearman mostrou uma correlação negativa entre o número de retweets e o número de menções a usuários. Isso sugere que tweets com conexões pessoais podem ter pouco valor de 'uso geral' e esses usuários podem estar relutantes em retweetar o conteúdo desses seguidores. O teste Spearman encontrou uma pequena correlação positiva entre o número de favoritos e o número de usuários seguidos, e encontrou uma baixa correlação negativa entre o número de favoritos e tempo de uso da conta (desde o cadastro).

(Artigo "A Text Mining Analysis of Academic Libraries Tweets") Análise de textos Em mídias sociais, analisar o conteúdo e minerar textos é frequentemente usado para analise texto de usuários geral e suporte para tomada de decisão (Abrahams, Fan, Wang, Zhang & Jiao, 2015). Aplicar mineração de textos em bibliotecas e outras instituições inclui extrair informações, rastrear tópicos, sumarização, categorização, agrupamento, ligar conceitos, visualização de informação e perguntas de questionários (Fan, Wallace, Rich e Zhang, 2006). De acordo com Zhang e Gu (2011) as bibliotecas acadêmicas poderiam utilizar mineração de textos para beneficiar os aplicativos no objetivo de suporte a tomada de decisões. No recente artigo, Sandhu (2015) indicou a importância do aprendizado sobre mineração de dados e ferramentas de Big Data para as bibliotecas acadêmicas para melhorar a eficiência da biblioteca e os serviços de informação. Similarmente Zhang e Gu (2011) alegaram que minerar conhecimento sobre os clientes é importante para as bibliotecas acadêmicas. A abordagem de mineração de textos para os dados das mídias sociais tem sido utilizados em muitos outros campos, como negócios (business) (Abraham et al., 2015), ciência da saúde (Sarker et al., 2015) ciência política (Stieglitz & Dang-Xuan, 2013). O uso da abordagem de mineração de textos está lentamente começando a ser aplicada na literatura sobre biblioteca e Ciência da Informação... Papatheodorou, Kapidakis, Sfakakis e Vassiliou (2003) estudaram o uso das técnicas de mineração para analisar o uso das comunidades digitais. Zang e Gu escreveu um artigo sobre aplicação de "text mining" nos dados das vidas sociais. Shulman et al. (2015) analisou a rede de Twitter em duas bibliotecas acadêmicas os esforços para identificar a influência dos "players" e o recrutamento deles para fins de disseminar informações.

Conjunto de Dados O conjunto de dados desta pesquisa foram coletados completa-

mente da "timeline" de 10 bibliotecas acadêmicas (i.e. todos Twittes desde a adesão à plataforma) através de um serviço de arquivamento (twimemachine.com) em dezembro de 2014. «« Interessante essa ferramenta, instalei e pode ser usada para fazer mineração em textos »» As bibliotecas escolhidas foram as 10 maiores ranqueadas pelo Shanghai Ranking. A seleção foi restringidas para universidades que "falavam inglês" e a uma biblioteca se a universidade tinha mais que uma (no artigo original é a tabela 1). Adicionalmente para o conteúdo do tweet, as seguintes informações foram recuperadas: Data do tweet – "tweet date", Número de vezes marcados como favorito – "number of times marked as a "favorite" by other users", Número de vezes retweetado "number of times "retweeted" e, Data que se "juntou" (entrou) ao Twitter.

Preprocessamento do conjunto de dados – "dataset preprocessing" O "dataset" recuperado foi "limpo" para reduzir os ruídos seguindo uma abordagem consistente com outros estudos de mineração (em textos) tal como Ralston, O'Neil, Wigmore e Harrison (2014) e Yoon, Elhadad e Bakken (2013). O passo-a-passo incluiu a aplicação de um 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.

Analise do Conjunto de dados – "Dataset" Mineração em textos e técnicas de análise de conteúdo foram usadas para analisar os históricos de tweets da bibliotecas. A frequência dos tweets, retweets e distribuição dos tweets e retweets o tempo todo foi identificado e "plotado". Então os tweets foram limpos, abertos e marcados com o PamTaT, uma ferramenta "text mining" desenvolvida por Pamplin Collage do Instituto Politécnico de Negócios da Virgínia da(ligado à) Universidade Estadual de Virgínia. PamTaT é baseado interface do Microsof Excel para Python nltk – "natural language processing framework" (Bird, Loper & Klien, 2009). PamTaT permite que grande volumes de textos sejam analisado pelos usuários finais sem requer conhecimento de programação da linguagem Python. Em particular foi usado PamTaT para determinar a frequência de palavras simples (unigramas), de duas palavras (bigramas) e sequencias de três palavras (trigramas) que aparecem no texto fonte. Isso permite-nos 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. Adicionalmente "correu-se" com o Harvard General Inquirer (Stone et al. 1966) para análise semântica e sentimental dos tweets. "Harvard General Inquirer" é uma ferramenta de análise de textos de propósito geral, esta permite ao usuário final repostar com que frequentemente categoria de palavras diferentes usadas no texto fonte. Aplicações reportadas no "General Inquirer" para diferentes textos fontes identificou (categoria) duas centenas de palavras incluindo, por exemplo: como positivo, negativo, relacionadas a vontade (prazer), relacionadas a dor, relacionadas a localização, relacionadas a hora (tempo), relacionadas a academia, relacionadas a exagero (overstatement) ou subavaliação (understatement) e assim por diante. Hurtwitz (2002) forneceu uma lista abrangente de categorias de palavras reconhecidas pela Harvard General Inquirer e proveu uma lista completa de palavras específicas que pertencem a cada categoria de palavras. No estudo desse artigo foi usou-se o Harvard General Inquirer para determinar o número de palavras em cada categoria de palavras usadas em cada biblioteca.

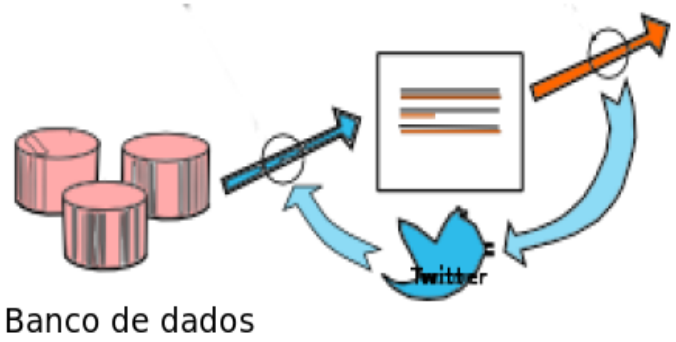
3.5 Arco cibernético com dados do Twitter

Os dados do Twitter, permite uma busca imediata por novas informações que poderão ser confrontadas com o modelo preditivo aumentando o nível de confiança deste, com isso a informação construirá um Arco cibernético, que segundo Wiener (1948) a informação permite realimentação aos sistemas, com controle mais eficaz, por exemplo: no trecho da Br 101, na altura do km 5, no Município de Goiana alguém publicou que a comunidade que mora no entorno dessa localidade fará um protesto daqui a dois dias devido ao acidente ocorrido ou a PRF publicou que o km 80 da Br 232, na altura do Município de Gravatá será interditada amanhã, por 2h, para remoção/explosão de rochas. Essas informações, por serem a posteriori às predições, podem aumentar o nível de confiança sugerida pelo modelo preditivo e controle por parte do utilizador e dentro de um universo temporal mais restrito servir de comprovação do das predições.

No entanto pode ocorrer o contrário quando as informações provenientes do modelo de predição entrar 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, o observador ou o utilizador.

As informações das redes sociais, armazenadas em um banco de dados, poderão servir futuramente para novas predições. Essas informações que comporão o arco cibernético não deverão retroalimentar o modelo de predição já construído, pois o fluxo decisório já foi tomado pelo observador, sendo que dados a posteriori não servem para um modelo de predição, enviesa o sistema preditivo. Uma nova fase de Mineração de Dados, desta vez mineração de dados em textos com modelo de predição. Dessa forma compõe-se um novo arco cibernético, mais genérico à proposição inicial descrita.

Figura 3.2: O arco cibernético com o Twitter

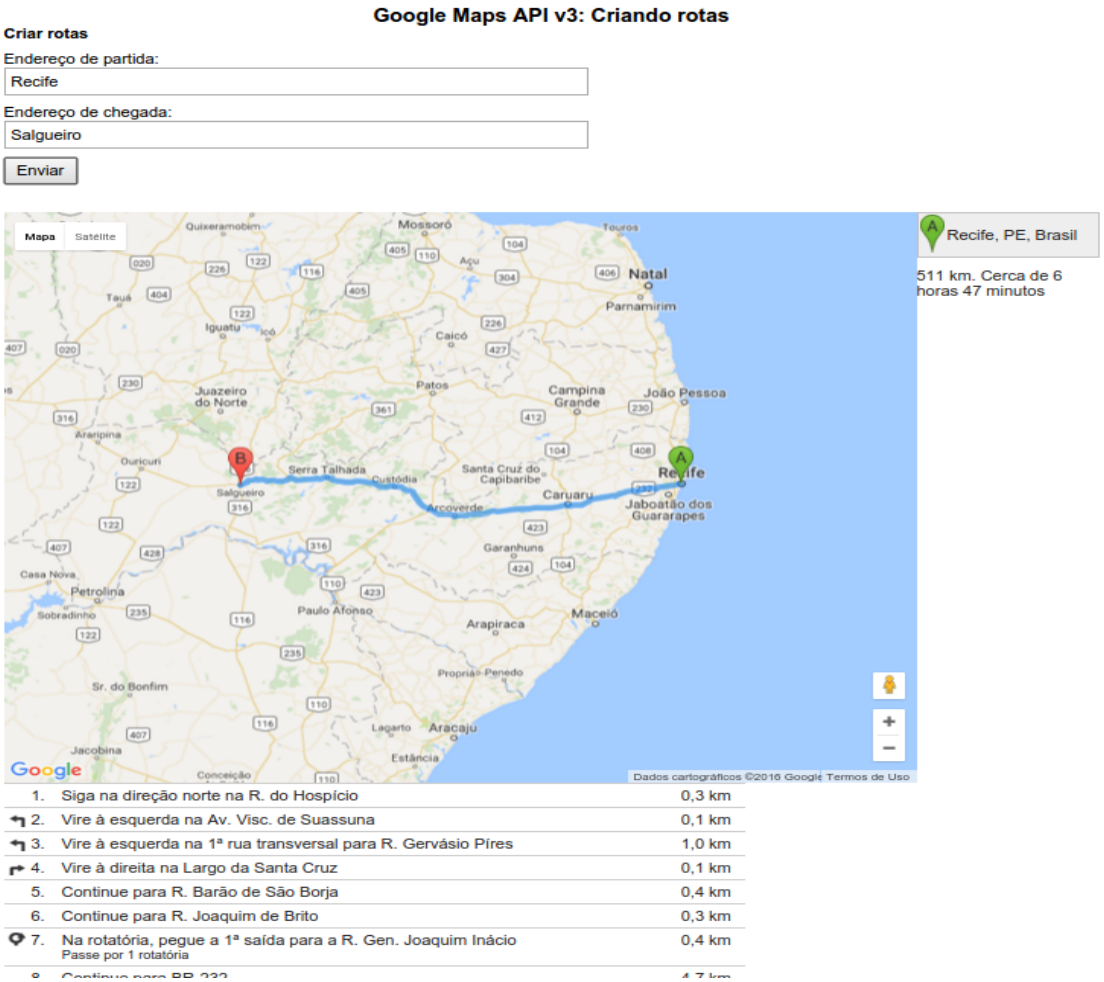


Fonte: autor

[2] O sistema de predição é baseado em cálculos probabilísticos

3.6 Extrapolação para georreferenciamento

Figura 3.3: Etapas da metodologia

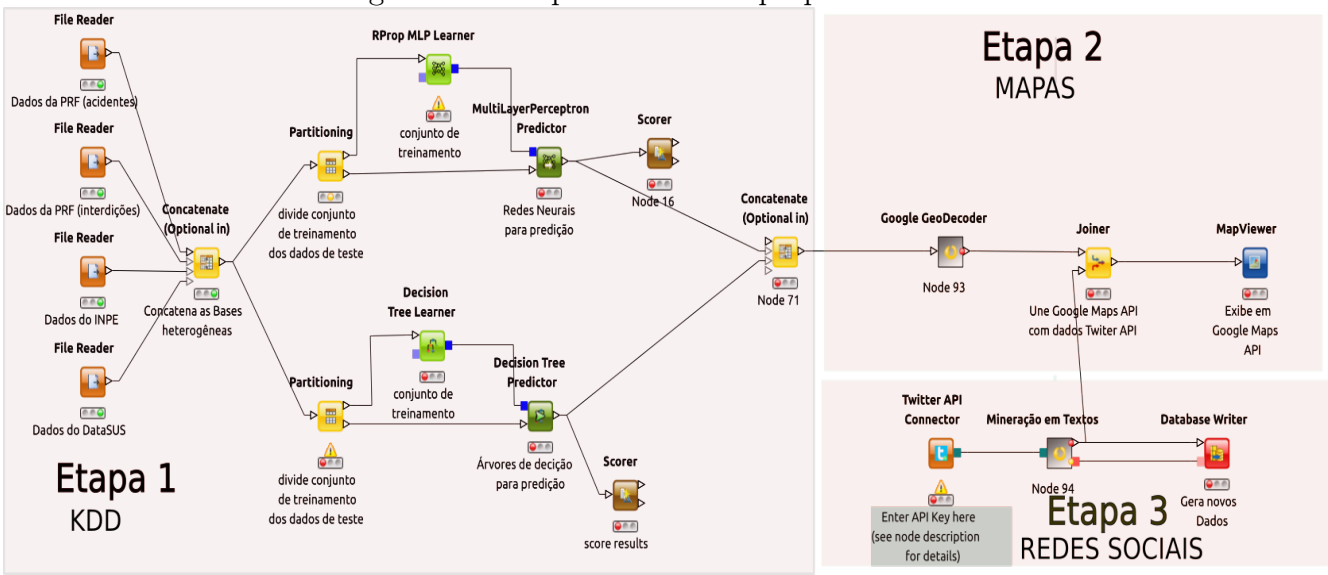


4 Simulação

4.1 O modelo proposto

A figura a seguir ilustra essa metodologia descrita graficamente, onde as três etapas.

Figura 4.1: Etapas da modelo proposto



Fonte: autor

A **etapa 1** contempla a fases da coleta das bases de dados históricas, preparação dos dados e construção das variáveis do modelo preditivo.

1. O modelo preditivo integra bases de dados, tais como: Polícia Rodoviária Federal – PRF, Batalhão de Polícia de Transito – BPRv.
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 para complementar o sistema estão disponíveis na Internet sendo atualizadas pela PRF através de uma API aberta, esta pode ser configurável para se ligar ao nosso sistema.

3. A conclusão dessa etapa ocorre com a Mineração dos dados e a extração de conhecimento. Os ”outputs“ dessa etapa consiste em transformar os dados provenientes da mineração em coordenadas geográficas. As coordenadas geográficas são agrupadas a priori formando ”cluster“ de dados exibidos em mapas vetoriais.

A **segunda** etapa da metodologia contempla:

1. Representação da malha viária em mapas de bases vetoriais;
2. Um ambiente de simulação interativa que utiliza uma plataforma baseada em API como QGIS e Google Maps.

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

1. Um módulo dinâmico onde são capturados “feeds” de redes sociais, por exemplo pelo Twitter. Essa técnica faz um arco cibernético mantendo o sistema atualizado de informações.

4.2 A construção do Modelo preditivo

O modelo preditivo foi construído utilizando bases de dados históricas da PRF (de acidentes e de paralisações ex: protestos) entre Janeiro de 2007 a Dezembro 2015. As bases de dados do Batalhão de Polícia de Rodoviária estadual – BPRv vieram entre Janeiro/2010 a Julho/2016, cortes em ambas as bases foram feitos para adequar as datas até 2015. Essas bases de dados são integradas gerando um único e complexo modelo preditivo acoplado a estrutura dinâmica.

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 algum, contudo 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 ideia metodológica proposta para esta pesquisa também contemplou todas as fases do KDD conforme descrito a seguir.

4.2.2 Aplicação das fases da Mineração ao KDD

Seleção: Nesta etapa foram coletadas as informações provenientes das bases de dados da Policia Rodoviária Federal (PRF) de 2007 a 2015, uma vez que nosso interesse era o de analisar os últimos dez anos, no entanto, como a base de dados só dispunha de informações a partir de 2007, foram considerados os nove anos disponíveis. A PRF 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 (33) destacam a não padronização e não aceitação dos dados pela comunidade internacional EAVES, D. (34) 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 burados, 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, Caminhoneta, Motocicletas, Caminhão, Caminhão-trator, Bicicleta, Caminhonete, 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, sendo consideradas por conterem inconsistência e "missing data", como, por exemplo, informações acerca de latitude e longitude. Cabe destacar que a base, 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 aviam 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 a pesquisa foi Árvore de decisão que possibilita uma interpretação imediata e de fácil compreensão. Como ferramentas, foram escolhidas o Knime (35) e R (36) e o Weka (37), com objetivo de estabelecer uma comparação entre ambos, cuja intenção era produzir um classificador mais preciso. Nessa direção, a técnica Ensamble de classificadores (38) estabelece que a combinação de um ou mais classificadores iguais, ou mais de um classificador diferente, aumenta a precisão. Tanto na ferramenta Knime com Weka o algoritmo 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 calculado a correlação linear entre todas as variáveis, entre as variáveis BR e Delegacia (variável que agrega municípios) obteve correlação linear de 0,653, já entre Tipo de Acidente e Traçado via a correlação foi baixa, apenas 0,14, variáveis com correlação linear abaixo disso foram descartadas.

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

4.2.3 Dados encontrados antes da Mineração

A grande maioria dos acidentes ocorre com pista seca, sem restrição de visibilidade. A cor vermelha é referente a BR 101, a cor azul à BR 232, seguido das outras menos significativas.

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

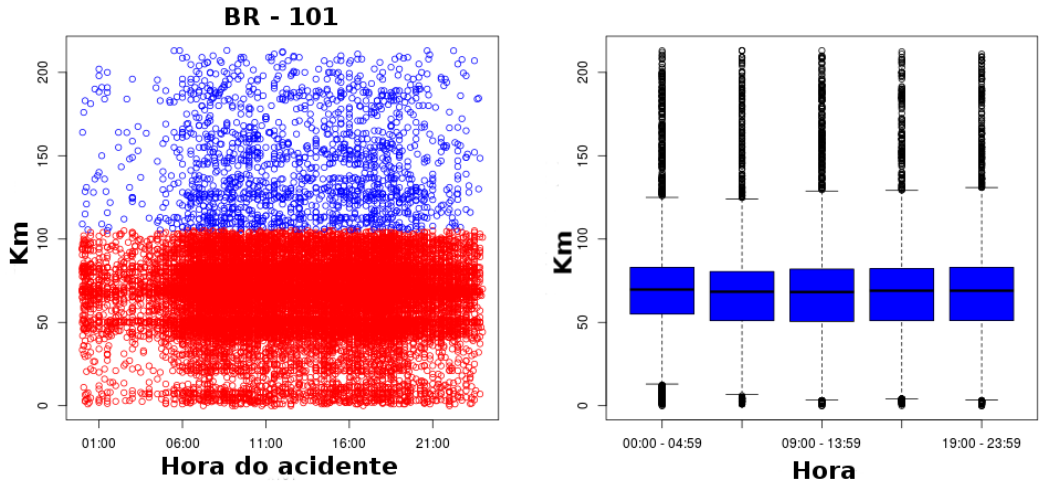
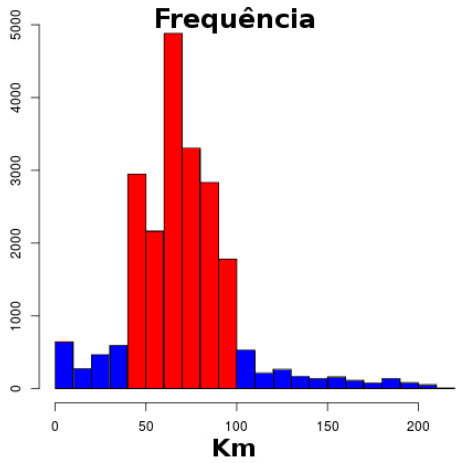


Figura 4.3: Frequência



O gráfico 4.2(1), 4.2(2) e 4.3 contém dados da BR 101 uma das mais importantes para o nordeste brasileiro pelo seu tráfego intenso. A gráfico 4.2(1) representa os acidentes que ocorreram a cada hora (abcissa) 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 ter influência na causa dos acidentes, contudo mais adiante descobrimos outros condicionantes que influenciam mais fortemente essas ocorrências. É

possível perceber no gráfico i, que alguns padrões especialmente em determinados locais (Km), por exemplo na BR 101 entre os Km 40 e 100 ocorrem acidentes a partir da 05h da manhã até as 23h.

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

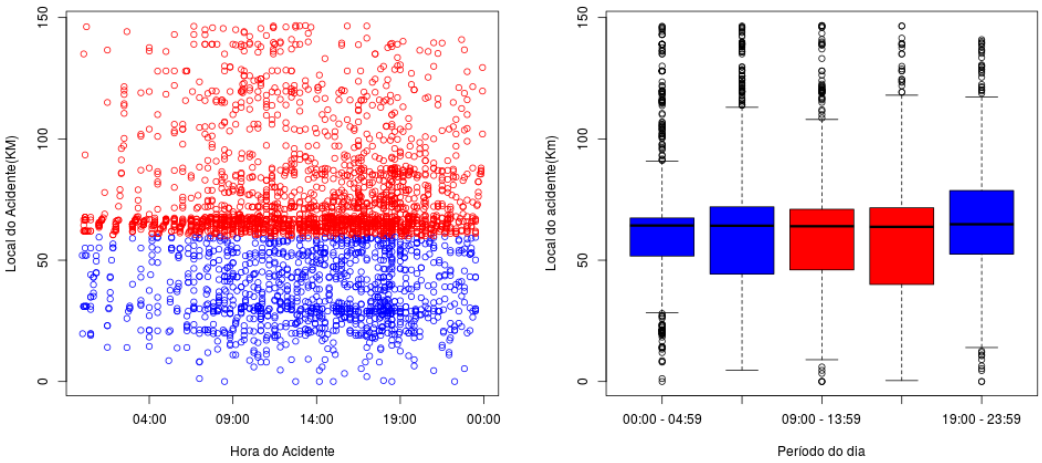
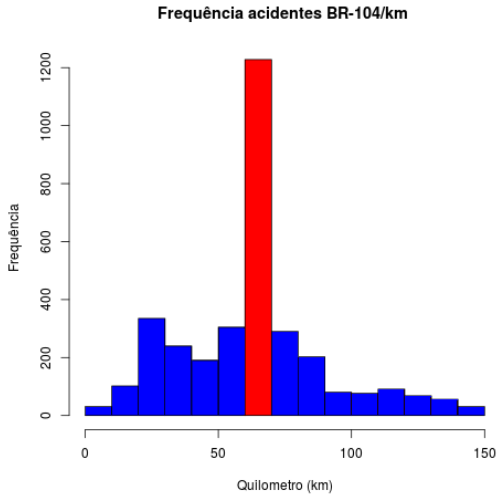


Figura 4.5: Frequência



O gráfico 4.4(1), 4.4(2) e 4.5 contém dados da BR 104 que atravessa seia municípios de Pernambuco, detre ele Caruaru que apresenta uma das maiores frotas de veículos do interior. A gráfico 4.4(1) representa os acidentes que ocorreram a cada hora (abcissa) em cada Km (ordenada) nos últimos nove anos. O gráfico 4.4(2) corresponde à frequência do local onde ocorreram esses acidentes. É possível perceber que há determinados locais, em torno do Km 60 na rodovia, onde se concentram os acidentes. O terceiro gráfico, tipo 'boxplot', exibe a concentração das ocorrências em torno da mediana dessa localidade (Km). É possível perceber no gráfico i, que alguns padrões especialmente em determinados locais (Km), por exemplo no Km 60 ocorrem acidentes a partir da 04h da manhã até as 23h. Destaca-se que esta rodovia atravessa o trecho urbano de Caruaru por onde passa por dia mais de 50.000 veículos.

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

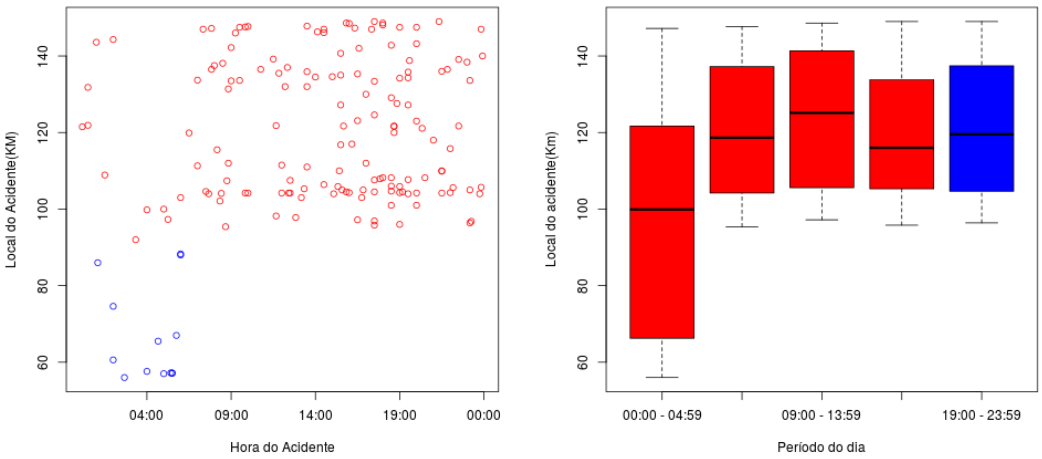


Figura 4.7: Frequência

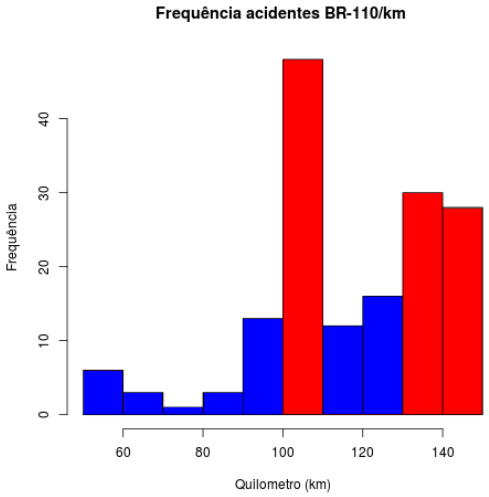
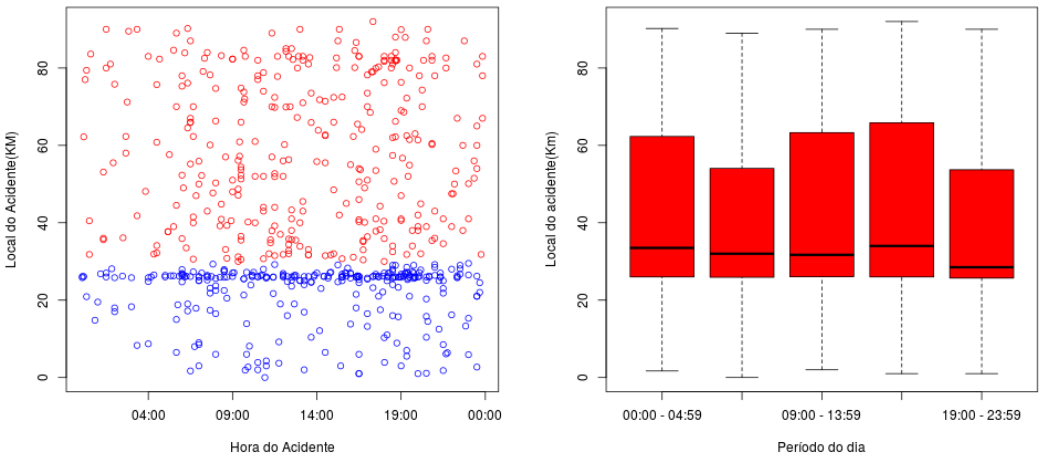
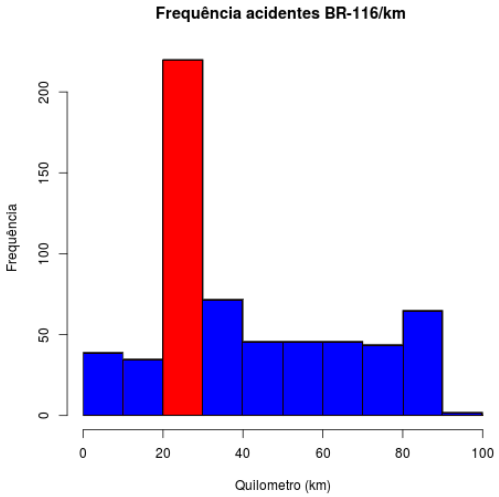


Figura 4.8: Hora do acidente (1) — Concentração em torno da hora (2)



O gráfico 4.8(1), 4.8(2) e 4.9 contém dados da BR 116 uma das mais importantes

Figura 4.9: Frequência



para o nordeste brasileiro e do paíspelo seu tráfego intenso, atravessa o país inteiro, desde o Rio Grande do Sul até o Ceará. A gráfico 4.8(1) representa os acidentes que ocorreram a cada hora (abcissa) em cada Km (ordenada) nos últimos nove anos. O gráfico 4.8(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 ter influência na causa dos acidentes, contudo mais adiante descobrimos outros condicionantes que influenciam mais fortemente essas ocorrências. É possível perceber no gráfico 4.9, que alguns padrões especialmente em determinados locais (Km), por exemplo, há um padrão nos acidentes nos Km 30 onde ocorrem acidentes a partir da 04h da manhã até as 22h.

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

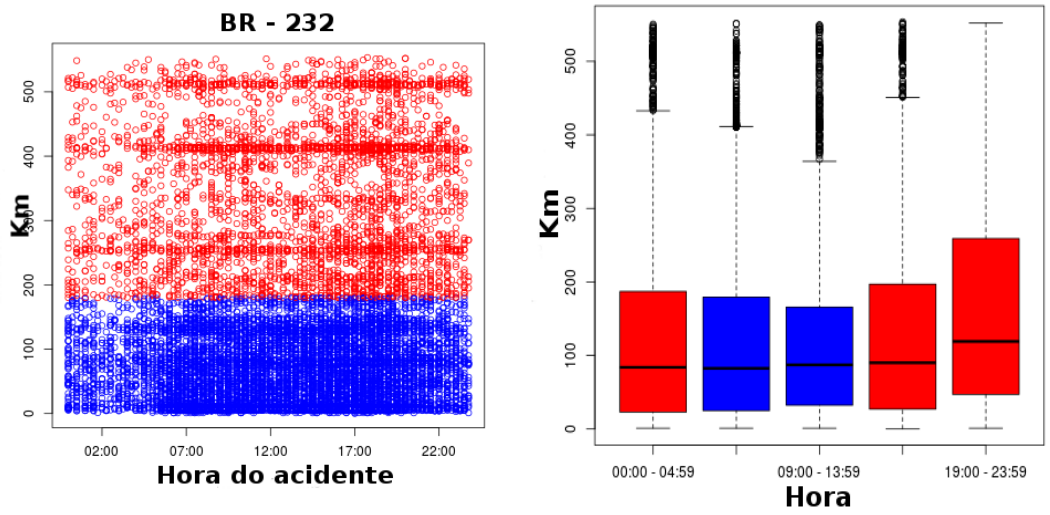
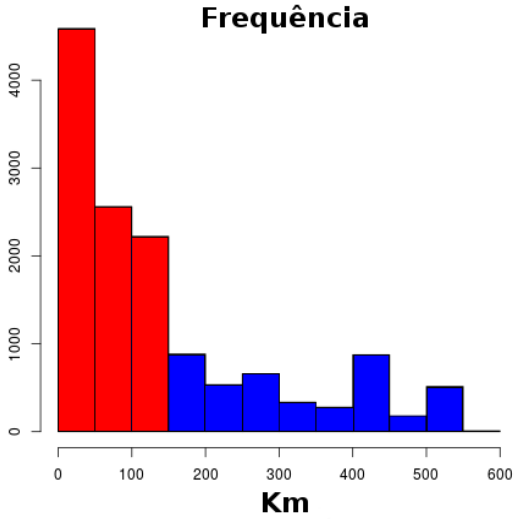


Figura 4.11: Frequência



O gráfico 4.10(1), 4.10(2) e 4.11 contém dados da BR 101 uma das mais importantes para o nordeste brasileiro pelo seu tráfego intenso. A gráfico 4.10(1) representa os acidentes que ocorreram a cada hora (abcissa) em cada Km (ordenada) nos últimos nove anos. O gráfico 4.10(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 ter influência na causa dos acidentes, contudo mais adiante descobrimos outros condicionantes que influenciam mais fortemente essas ocorrências. É possível perceber no gráfico 4.11, que alguns padrões especialmente em determinados locais (Km), por exemplo, na BR 232 há um padrão nos acidentes nos Km 0, 90, 110, 260, 410 e 500. ocorrem acidentes a partir da 00h da manhã até as 23h.

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

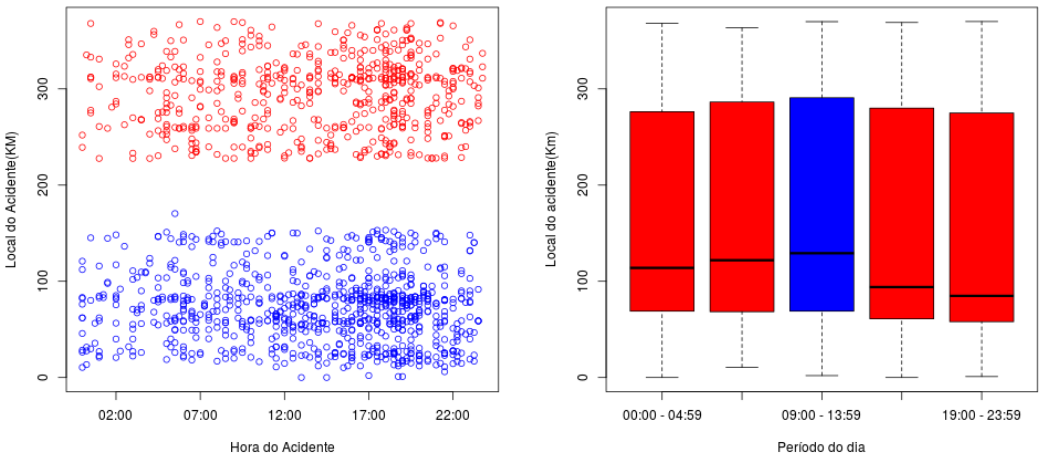


Figura 4.13: Frequência

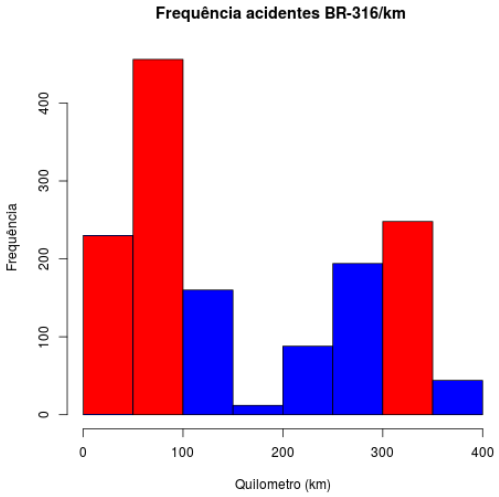


Figura 4.14: Hora do acidente (1) — Concentração em torno da hora (2)

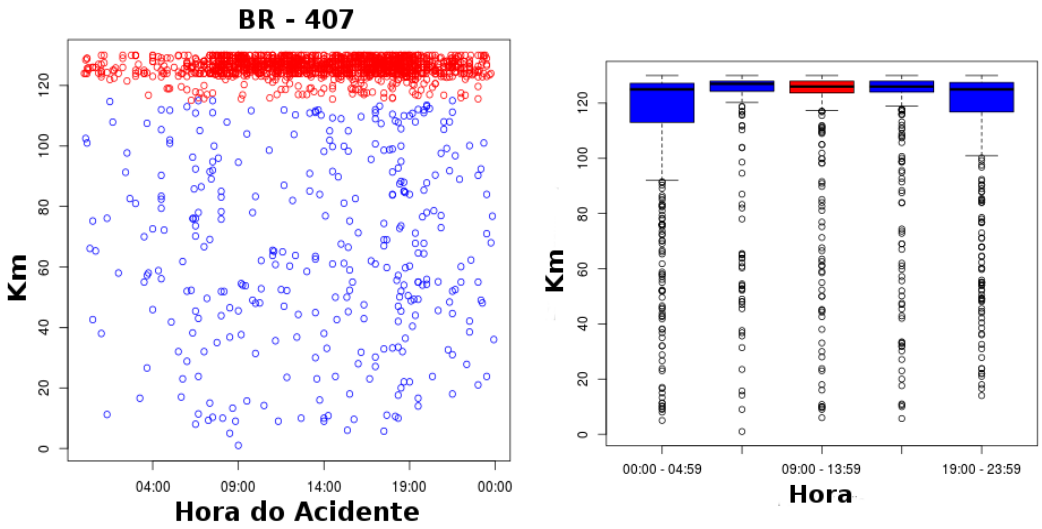
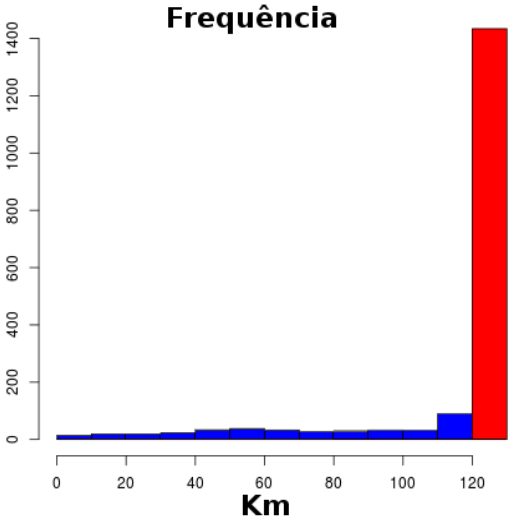


Figura 4.15: Frequência



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)

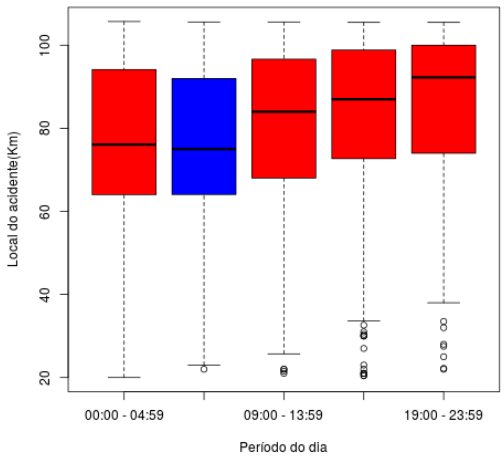


Figura 4.17: Frequência

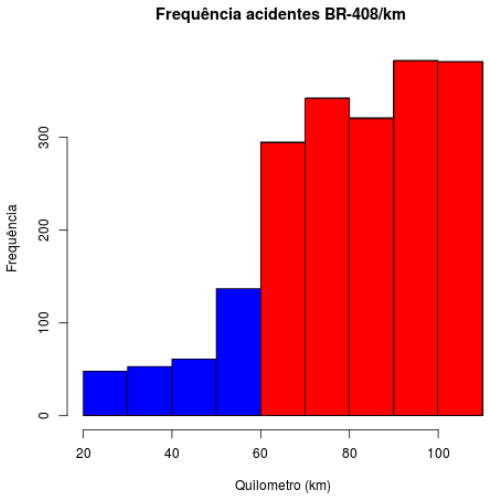


Figura 4.18: Hora do acidente (1) — Concentração em torno da hora (2)

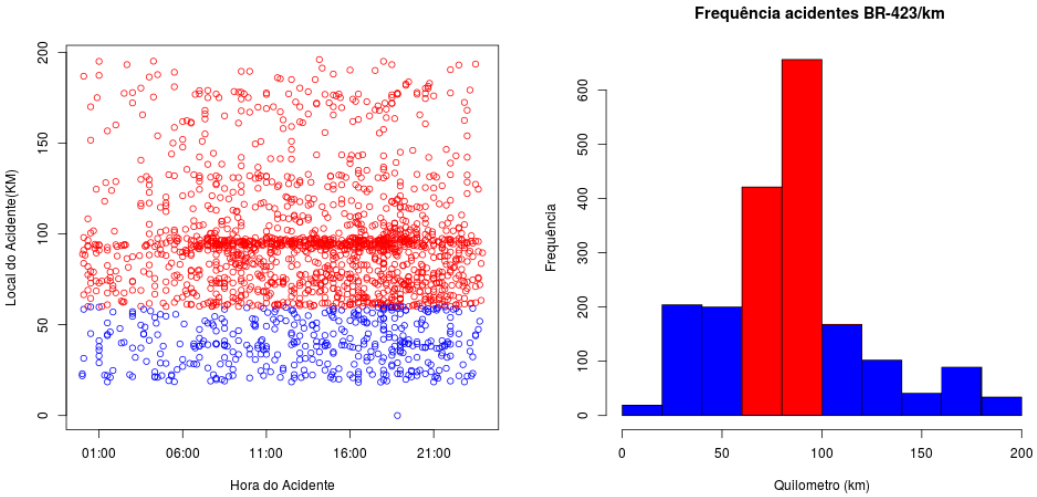


Figura 4.19: Frequência

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

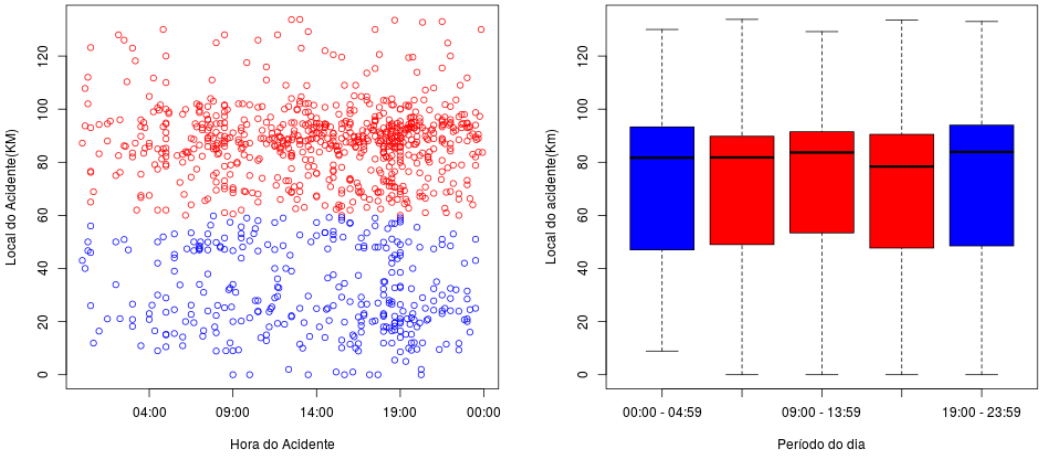


Figura 4.21: Frequência

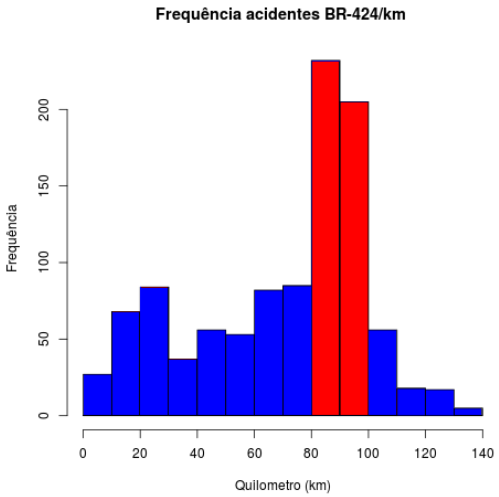


Figura 4.22: Hora do acidente (1) — Concentração em torno da hora (2)

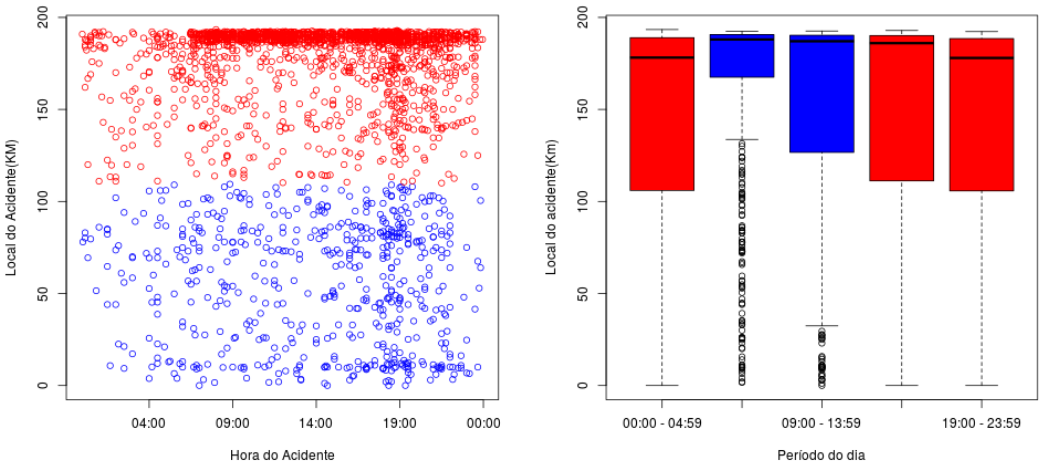


Figura 4.23: Frequência

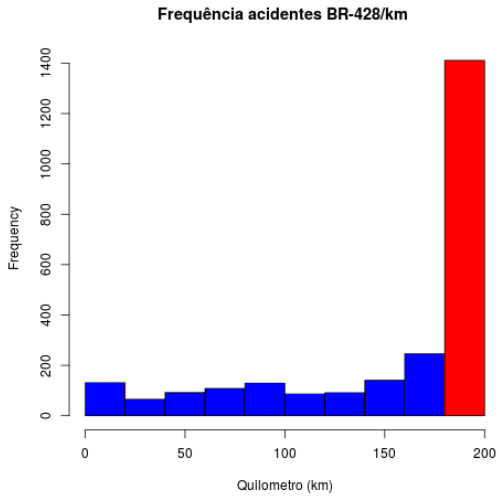
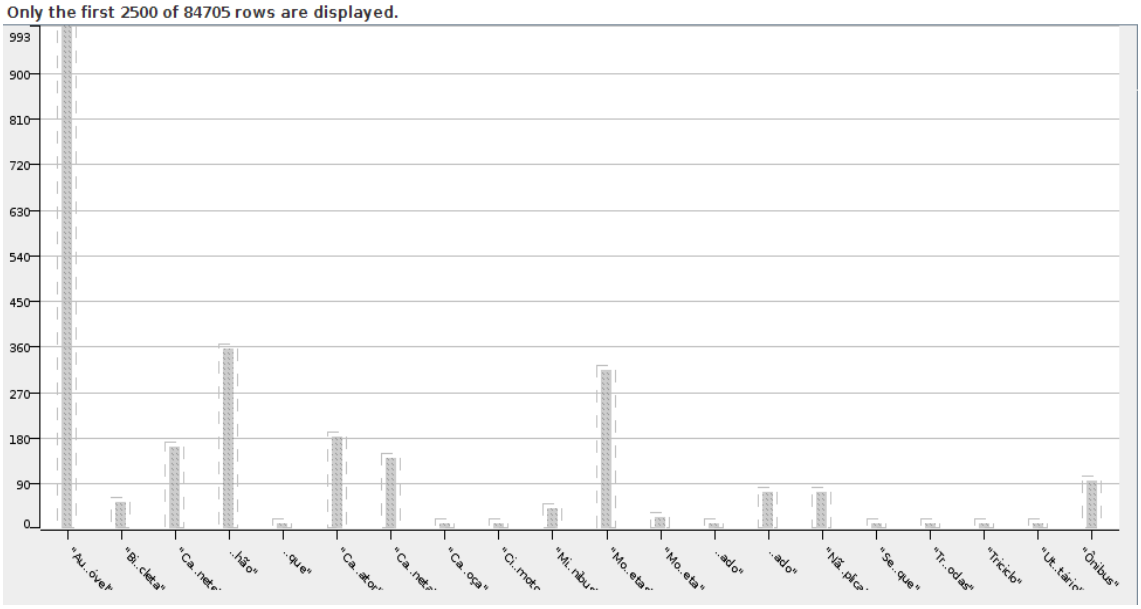


Figura 4.24: Tipo de Veículo X Num. Acidentes

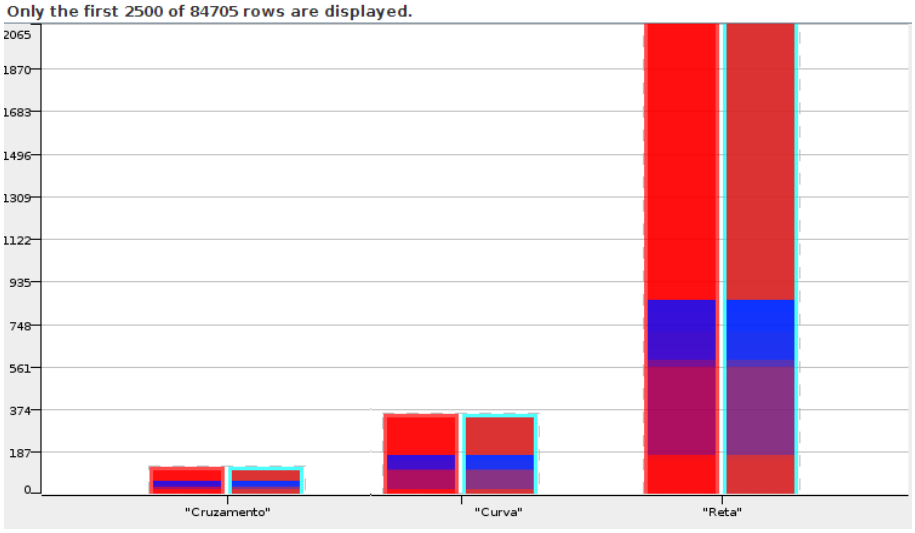


Fonte: autor

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. Esses dados são referentes ao período entre 2007 à 2015.

O tipo de traçado da via não influência nas estatísticos de acidenes, pois a grande maioria dos acidentes ocorre em linha reta. Esse comportamento do condutor nos faz crer o condutor é o responsável pelo maior número de ocorrências nas BRs. Isso direciona nossa pesquisa para analisar e antever o comportamento do condutor, as condições da rodovia e ambientais nessa base de dados não são fatores relevantes.

Figura 4.25: Traçado da via X Num. Acidentes



Fonte: autor

4.2.4 Dados encontrados após a Mineração

Os resultados da classificação encontrados estão contidos a seguir, dentre os classificadores disponíveis no Weka os que melhor apresentaram resultados foram: o Naive Bayes e a Árvore de Decisão As variáveis "Tipo de Acidente", "Gravidade" e "BRajustada" foram escolhidas pelas características de ganho de informação dado pelo cálculo da entropia. "BRajustada" significa ... A literatura aconselha que os nós da raiz dos classificadores, em especial Árvores de decisão, tenham maior entropia, como a variável "Tipo de Acidente", no entanto o grande número de ramificações que esta variável gerou não foi interessante para o objetivo da pesquisa; explicar o porquê das causas dos acidentes (pontos fortemente destacados nos gráficos com rótulo (1)). A seguir apresenta-se a nomenclatura da acurácia encontrada na classificação.

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

4.2.5 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 quadratico	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 representar uma amostra adequada, pois..... 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 quadratico	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 ou 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 dão maior confiabilidade aos testes.

A variável "BRajustada" não teve o maior coeficiente de entropia encontrado, contudo esta variável apresentou índices de classificação das instâncias correta acima dos 80% e o menor índice de classificação incorreta dentre os dois classificadores utilizados. Esta variável foi a escolhida para encabeçar os algoritmos para os resultados encontrados.

A árvore construída pelo Knime para a mesma variável "Causa do Acidente" => velocidade incompatível está na próxima figura.

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

Figura 4.26: Árvore de Decisão gerada pelo Knime

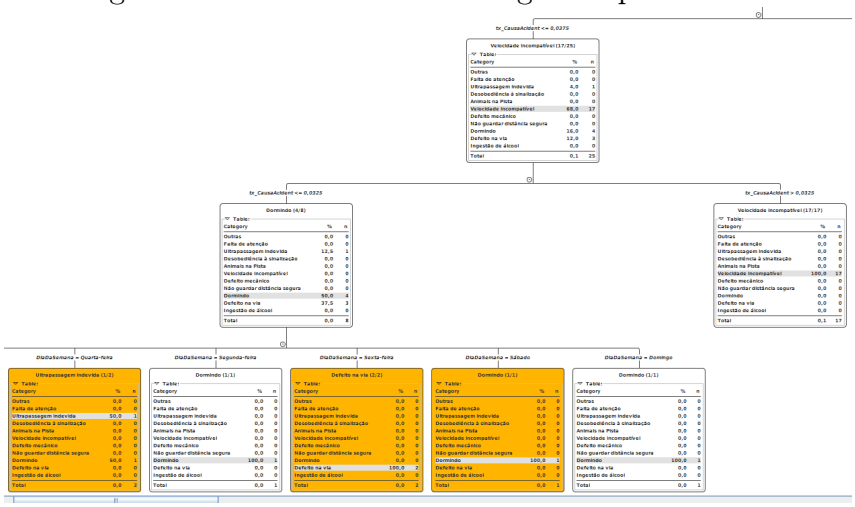


Figura 4.27:

Também para exemplificar, o nó folha classificou, que nas quartas-feiras a causa "ultrapassagem indevida"; sextas-feiras: "defeito na via"; e, caso seja um sábado, a causa é "dormindo". 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, inclusive o classificador Naive Bayes obteve um desempenho semelhante as Arvores de decisão com essa variável, somente na BR 408 e BR110 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 as leis de transito, pode se dizer que e indisciplinado, pois todos os tipos de causa foram encontrados. Caso se considere um raio de 50 Km no entorno da capital Recife pode-se dizer que os motoristas tem 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 fata de atenção foi condição "sine qua non" ocorreram acidentes na região metropolitana do Recife. Para a região no entrono da BR 116 os acidentes com mortes [Gravidade = S] na quinta-feira o curioso foi que quase todos os tipos de veículos se envolveram nesse tipo ocorrências. 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

4.3 Acoplamento com a estrutura dinâmica

As predições feitas na primeira fase têm como "output" coordenadas geográficas do tipo latitude e longitude.

A estrutura dinâmica é composta por duas API's, uma disponibilizada pela Google, através do Google Maps que está atualmente na versão V3 e outra uma API do Twitter. A API do Google Maps proporciona uma "leitura" atualizada em forma de mapa no momento em que a estrutura dinâmica "roda".

A API do Twitter possibilita atualizar o utilizador de informações recentes, pois o modelo preditivo teve análise temporal fixada pelas bases de dados até 2015, portanto o fluxo decisório da predição é baseado até esse período, 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.

5
Conclusão

Após a realização do estudo, os dados sugerem que o modelo de predição ora proposto dá conta das questões dessa pesquisa, ou seja, possibilita a gestão logística relativa a horários de utilização das rodovias. Os resultados do 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 estudos dessa natureza. Os dados encontrados dessa pesquisa corroboram com os resultados encontrados em estudos semelhantes, 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 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 atenção do condutor, que resulta, na maioria das vezes, em colisões traseiras e/ou laterais. Os acidentes acontecem nos horários em que há mais veículos trafegando na via, independentemente da hora do dia. Quando a via exige maior atenção, por condições que lhes são peculiares (um cruzamento, por exemplo), uma pequena restrição para ser amplificada, aumentando consideravelmente a quantidade de acidentes. Outra contribuição da pesquisa a ser destacada é de cunho metodológico-prático. Do ponto de vista metodológico, pela contribuição da aplicação do processo CRISP-DM, usado para construir o modelo preditivo. O algoritmo Árvores de Decisão mostrou-se robusto, quando aplicado a esse tipo de problema. Quanto á mineração de textos em redes sociais, no caso do Twitter, possibilita a identificação de comportamentos do condutor, antes mesmo de utilizar a via. Todavia, embora tenha sido uma ferramenta útil, para que pudesse haver uma influência maior nos resultados, seria necessário ampliar o escopo para outras redes sociais que tenham o perfil de troca de informações pontuais sobre 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 re-

tenções crescentes de logística de cargas. Tal modelo se configura como um avanço em relação às funcionalidades das soluções disponíveis que existem até o momento, tais como: Google Maps, Waze e outros dessa natureza somente exibem informações momentâneas, produzidas e compartilhadas pelos utilizadores dos aplicativos ou por informações provindas de GPS, com a nossa abordagem dados históricos de rodovias podem auxiliar predições sobre seu comportamento. Pode-se destacar alguns aspectos que figuraram como elementos dificultadores na realização da pesquisa. As informações da base de dados da PRF .....

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 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 metropolitanos. A API Google Maps, o "front-end" do sistema, 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* (39), 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.(33) 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

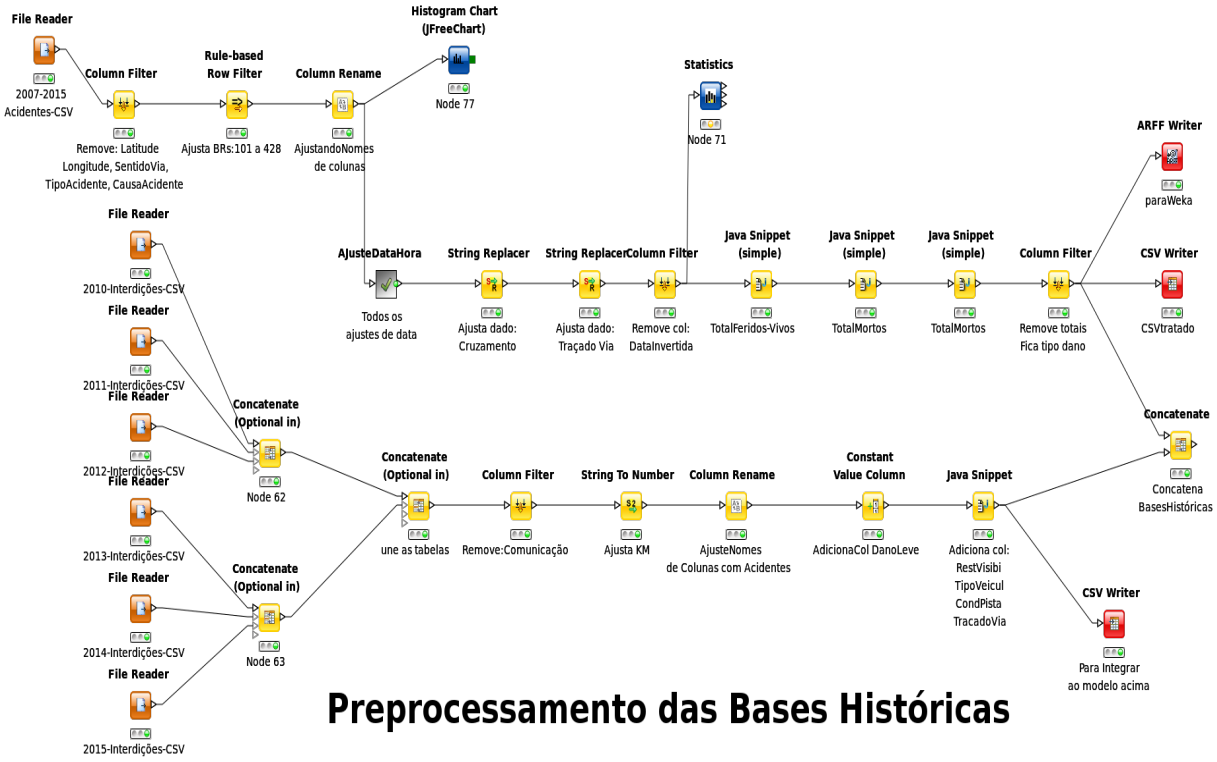
Ano	Ano da ocorrência do acidente
Mês	Mês de ocorrência do acidente
Num	Número do mês do acidente ex: 1 = Janeiro
KM	Numeração do quilômetro
BR	Numeração da Br
Latitude	Latitude da ocorrência
Longitude	Longitude da ocorrência
Condição Pista	Condição da pista: seca, molhado, ...
Restrição de Visibilidade	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



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 cibernética*. Tese (Dissertação de Mestrado) — Universidade de São Paulo - USP, 2007.

3 BNDES. Perspectivas do investimento, n. 2, out. 2013. *Perspectivas do Investimento 2014-2017*, p. 2, 2013.

4 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>.

5 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>.

6 POSSAS, B. et al. Data mining: técnicas para exploração de dados. *Universidade Federal de Minas Gerais*, 1998.

7 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.

8 RUSSEL, S.; NOVING, P. *Integência Artificial*. [S.l.]: Elsevier, Rio de Janeiro, 2004. 716–721 p. ISBN 8535211772.

9 A. 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.

10 HAN, J.; KAMBER, M. Data mining: Concepts and techniques. Elsevier, San Francisco, v. 2 edition, p. 15–16, 2006.

11 ARANHA CHRISTIAN E PASSOS, E. *A Tecnologia de Mineração de Textos*. 2006. 1–8 p.

12 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>.

13 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.

14 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>.

15 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>.

16 HAN, J.; KAMBER, M. Data mining: Concepts and techniques. Elsevier, San Francisco, v. 2 edition, 2006.

17 WITTEN, I. H.; FRANK, E. Data mining: Practical machine learning tools and techniques. *Elsevier, San Francisco*, v. 2 edition, 2005.

18 QUINLAN, J. R. Induction of decision trees. *Expert Systems*, Learning, Machine and Learning, Machine and Publishers, p. 81–106, 2007. ISSN 0885-6125, 1573-0565.

19 QUINLAN, J. R. Discovering rules by induction from large collections of examples. *Expert systems in the micro electronic age. Edinburgh University Press*, In D. Michie, 1979.

20 HEATON, J. *Introduction to Neural Networks for Java*. [S.l.: s.n.], 2008. 440 p. ISBN 1604390085.

21 CASTANHEIRA, L. G. Aplicação de técnicas de Mineração de Dados em Problemas de Classificação de Padrões. n. 5531.

22 TATIBANA, C. Y.; KAETSY, D. Y. Acessado em: 01.out.2016. Disponível em: <http://www.din.uem.br/ia/neurais/#neural>.

23 ZENG Q., H. H. P. X. W. S. C. G. M. Rule extraction from an optimized neural network for traffic crash frequency modeling.

24 KRIESEL, D. *A Brief Introduction to Neural Networks*. [s.n.], 2007. Disponível em: <availableathttp://www.dkriesel.com>.

25 BARRETO, J. M. Jorge M. Barreto. 2002.

26 EGAN, J. P. Signal detection theory and roc analysis. New York, USA: Academic Press, 1975.

27 ANAESTHETIST, T. Acessado em: 20.jan.2015. Disponível em: <http://www.anaesthetist.com/mnm/stats/roc/Findex.htm>.

28 SOUZA, C. R. Acessado em: 20.jan.2015. Disponível em: <http://crsouza.com/2009/07/analise-de-poder-discriminativo-atraves-de-curvas-roc>.

29 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.

30 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.

31 F E. WILLIAMS, B. S. D.; GLASS, N. Twitter. July 2015. Disponível em: <https://pt.wikipedia.org/wiki/Twitter>.

32 WEST, D. Neural network credit scoring models – computers and operations research. p. 1131 – 1152, 2000.

33 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>.

34 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/>.

35 AG, K. 2017. Disponível em: <https://www.knime.org/>.

36 R-CRAN. 2017. Disponível em: <https://www.r-project.org/>.

37 ZEALAND, U. of W. N. 2017. Disponível em: <http://www.cs.waikato.ac.nz/>.

38 BERNARDINI, F. C. "combinação de classificadores simbólicos utlizando medidas de regras de conhecimento e algoritmos genéticos" - tese de doutorado. Instituto de Ciências e Matemática Computacional/USP, 2006.

39 2016. Disponível em: <http://dados.gov.br/dados-abertos/>.