



Programa de Pós-Graduação em Engenharia de Sistemas

DISSERTAÇÃO



Recife, 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

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

Banca de qualificação:

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, Janeiro de 2017.

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

Agradecimentos

Agradeço a fulano, ciclano e beltrano de tal,

Resumo

O Transporte de cargas, que atravessa as regiões metropolitanas das grandes cidades brasileiras o fazem principalmente pelas rodovias federais. Essas rodovias estão constantemente congestionadas e têm recebido aumento expressivo de novos veículos a cada ano.

No entorno desses trechos urbanos têm crescido desordenadamente comunidades que demandam por políticas sociais que atendam às suas necessidades. Para reivindicar, dos entes públicos, essas comunidades bloqueiam as rodovias, aumentando a pressão sobre os congestionamentos. Em alguns trechos o traçado das rodovias está próximo a morros e florestas ficando suscetíveis às intempéries climáticas. Essas variáveis impõem constantes paralisações às rodovias, representando atrasos nas entregas, custos adicionais às empresas e prejuízos à competitividade nacional.

Propor novas soluções, que venham minimizar esses constrangimentos é condição 'sine qua non', para mitigar o que vem sendo chamado de "Custo Brasil". Um sistema de suporte a decisão, ao roteamento de cargas, que utilize informações vindas de sítios eletrônicos como google-maps, base de dados históricas da Polícia Rodoviária Federal, das redes sociais e das condições socio-ambientais estão disponíveis em forma de dados na Internet e nos servidores locais dessas instituições, podendo ser processados numa plataforma auto-adaptável.

Assim, esse estudo tem por objetivo desenvolver um sistema de suporte a decisão para a problemática crescente da Logística de Cargas, antecipando prováveis eventos, que possam interferir no fluxo normal das rodovias brasileiras, propondo assim uma alternativa ao traçado de rotas determinísticas e incorporando novas opções de rotas inspiradas na predição de eventos futuros.

Palavras-chave: Mineração de dados, Bases de dados históricas, Redes sociais, Logística de transportes, Roteamento

Abstract

This project we propose to make a predict model receive information of the social networking, data bases, and highways managements, compiling everything is available , to check up routing a self-adapter plataform, in real-time, to drive a vehicle fleet for a destination, safely.

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

Lista de Abreviações e Siglas

API	<i>Application Programming Interface</i>
BG	<i>Big Data</i>
DM	<i>Data Mining</i>
TDM	<i>Text Data Mining</i>
KDD	<i>Knowledge Discovery Databases</i>
PRF	<i>Polícia Rodoviária Federal</i>
BPRv	<i>Batalhão de Polícia Rodoviária (estadual)</i>
TB	<i>TeraByte</i>
PB	<i>PetaByte</i>
EB	<i>ExaByte</i>
ZB	<i>ZettaByte</i>
YB	<i>YottaByte</i>

Lista de Figuras

2.1	O padrão CRISP-DM	15
2.2	Domínio das técnicas aplicadas a mineração de dados	15
2.3	Minerando dados no Big Data	17
2.4	Árvore de decisão	18
2.5	Mapa mental da Mineração em textos	22
3.1	Etapas da gerais da metodologia	24
3.2	Etapas da metodologia	27
A.1	Etapas 1 – Coleta e união das bases históricas de acidentes e interdições . .	32

Lista de Tabelas

2.1	Matriz de Confusão	16
2.2	Matriz modelo de Confusão	16
2.3	Volume de dados no mundo	20
3.1	Variáveis do modelo preditivo	25
A.1	Variáveis originais da base de acidentes	31
A.2	Variáveis originais da base de interdições	32

Sumário

1	Problema, Motivação & Objetivos	11
1.1	Problema	11
1.2	Motivação	12
1.3	Objetivo Geral	12
1.3.1	Objetivos Específicos	12
1.4	Resultados Esperados	13
2	Estado da Arte	14
2.1	Introdução	14
2.2	CRISP-DM	14
2.3	Data Mining	16
2.4	Machine Learning	18
2.4.1	Árvore de Decisão	18
2.4.2	Regressão Logística	18
2.4.3	Redes Neurais	20
2.5	Redes sociais	20
2.5.1	Data Mining - Text Data Mining	22
3	Metodologia	23
3.1	Plano geral da metodologia	23
3.2	Modelo preditivo	24
3.3	Reflexão sobre as tecnologias utilizadas no modelo preditivo	24
3.3.1	As variáveis do modelo preditivo	25
3.4	Extração do conhecimento	26
3.5	Acoplamento com a estrutura dinâmica	27
4	Resumo, Discussão & Trabalhos futuros	28
4.1	Conclusão	28
4.2	Discussão	28
4.3	Trabalhos futuros	29
A	Preprocessamento	30
A.1	Preprocessamento dos dados coleta dos dados da PRF	30
	Referências Bibliográficas	33
	Referências Bibliográficas	33

O Projeto

*“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

Problema, Motivação & Objetivos

1.1 Problema

A partir do início do século XXI o mundo digital, especialmente a Internet, conheceu sua primeira grande crise (1). As empresas ligadas a esse mundo, conhecidas como PontoCom, para sobreviverem, adaptaram-se à Internet abrindo suas estruturas, desde então houve um *boom* nesse segmento. As informações geradas e disponibilizadas à Internet, nos mais recentes anos, representam 90% de tudo o que já foi criado nos anos anteriores pela humanidade ou desde que nossa civilização aprendeu a guardar informação.

Para armazená-los, seriam necessários milhões de computadores; se fosse possível dispô-los num único *DataCenter*, esses ocupariam uma área do tamanho do estado de São Paulo.

Os dados produzidos pelo ser humano atualmente dobram a cada 5 anos; astrônomos atualizam suas descobertas numa base de dados disponíveis para outros utilizarem; as ciências biológicas agora têm tradição em depositar seus avanços científicos em repositórios públicos (2); redes sociais estão focadas na Web: Facebook, LinkedIn, Tweeter e outras sobrevivem coletando informações, vendendo espaço publicitário e repassando-as às empresas de telemarketing, empresas de comércio eletrônico como Amazon.com, Submarino.com.br, Americanas.com, MagazineLuiza.com.br, utilizam essas informações para vender mais e melhor. Por outro lado, artigos científicos dos mais variados assuntos e das mais variadas áreas alimentam todos os dias, com milhões de informações, os *Data Centers*.

Uma instância do problema descrito anteriormente será tratado nesta pesquisa, para isso 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 cargas rodoviárias, considerando 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. Associadas ao problema em lixeiras as redes sociais são um arcabouço de informações, os utilizadores dessas redes fornecem uma grande quantidade de dados que podem ser filtrados para dentro de uma aplicação, através de

técnicas adequadas.

1.2 Motivação

As rodovias federais que atravessam a Região Metropolitana do Recife (RMR) estão constantemente congestionadas, não apenas pela quantidade de veículos, mas por serem alvo de paralisações das mais diversas matizes, como protestos de trabalhadores, acidentes, buracos, intempéries naturais e outros tipos de paralisações. 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) (4) em 14 municípios, além da RMR será considerada para a pesquisa a Zona da Mata Norte (ZMN) com 577.191 habitantes e a Zona da Mata Sul (ZMS) com 733.447 habitantes. Nessas regiões (RMR, ZMN e ZMS) a frota (automóveis particulares, ônibus, caminhões, motocicletas, tratores e outros veículos) foi contabilizada, em 2014, com 635.686 veículos (5).

O que acontece na região metropolitana do Recife, é frequentemente visto no entorno das principais 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 o *Big data*.

A proposição de uma solução para resolver esse problemática requer várias etapas, para além da proposição de algumas técnicas de mineração dos dados. Propomos, nesse mestrado, uma solução peculiar, ao enviar essa frota de caminhões por diversas rotas, escolhidas por critérios cientificamente estudados. Isso poderá ser de suma importância para solucionar a problemática do transporte de cargas, que poderá advir com a plena produção da FCA, e que não se aplica apenas à região metropolitana do Recife, mas a toda e qualquer fábrica do país. Permitirá fornecer toda informação que se faz necessária para acompanhar os caminhões na transposição dos obstáculos que possam surgir ao transitar por Pernambuco, conduzindo-os até seu destino de maneira segura e no menor tempo possível.

1.3 Objetivo Geral

Esse mestrado tem como objetivo principal desenvolver um modelo preditivo suporte a decisão para a problemática das retenções crescentes de transporte de cargas rodoviárias nas BRs brasileiras. Para isto propomos uma solução multidisciplinar através da integração de diversas tecnologias disponíveis desde a análise dos dados históricos das rodovias a utilização informações de redes sociais e dados governamentais.

1.3.1 Objetivos Específicos

- Representar a problemática da logística de cargas em uma plataforma adaptável:
 1. Desenvolver uma plataforma adaptável que utilize dados de redes sociais (Twitter); analisar o contexto das rodovias por essas redes através da mineração de dados em textos e integrar ao resto do sistema através que verifica palavras chaves como: protestos, acidentes e outras.

- Desenvolver um modelo preditivo dos fenômenos que envolvem as rodovias:
 1. Para o desenvolvimento do modelo preditivo pretende-se utilizar bases de dados históricas.
 2. O modelo preditivo integra várias bases de dados, tais como: Polícia Rodoviária Federal – PRF, Batalhão de Polícia de Transito – BPRv e dados históricos do Instituto de Pesquisas Espaciais – INPE ou dos dados de precipitações pluviométricas do “European Centre for Medium-Range Weather – ECMWF.
 3. 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.
- Propor uma simulação interativa da estrutura com a dinâmica:
 1. A malha viária está representada na Internet em mapas de bases vetoriais que pretendemos integrar ao nosso sistema de informações e também vamos incorporar as informações que a Polícia Rodoviária Federal (PRF) dispõe, que controla as rodovias BRs.
 2. A simulação interativa utiliza uma plataforma baseada na API do Google Maps.
 3. Para simulação interativa da estrutura preditiva com a dinâmica real serão capturados “feeds” de redes sociais, por exemplo pelo Twitter. Essa técnica fará um arco cibernético mantendo o sistema preditivo atualizado.

1.4 Resultados Esperados

Ao final dessa pesquisa pretende-se obter um Modelo de Sistema de Suporte à Decisão adaptável, que contribua como uma ferramenta para complementar à tomada de decisão para a gestão do transporte de cargas rodoviárias. A região inicialmente escolhida será a RMR.

2

Estado da Arte

2.1 Introdução

Nesse estudo, foram realizadas três frentes de pesquisa para a revisão da literatura:

- A primeira frente relacionada a modelos preditivos de mineração de dados relacionados à pesquisa em lide;
- A segunda frente relacionada às tecnologias mineração em textos em redes sociais;
- Finalmente a última frente de pesquisa relacionadas às tecnologias de mapeamento rodoviário integrados aos sistemas de APIs

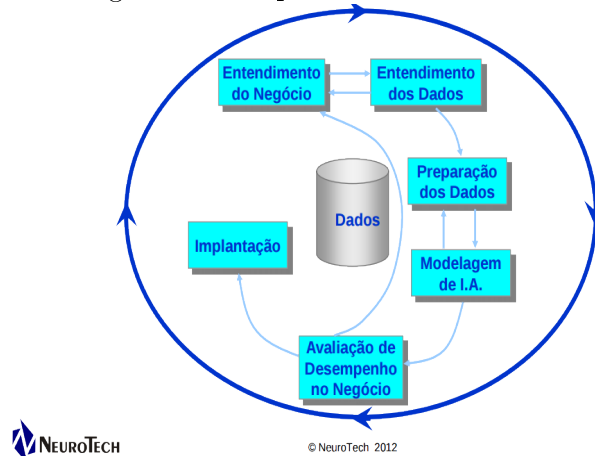
2.2 CRISP-DM

O processo de mineração “Cross Industry Standard Process for Data Mining” (CRISP-DM) (6) descreve como os especialistas em mineração de dados aplicam as técnicas em lide. 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 modelo de predição, como exemplificado na figura a seguir:

O **Entendimento do negócio** é uma fase crucial da mineração, um especialista (ou muitos) deve ser consultado. O analista de dados consegue fazer re-uso de conhecimento lendo periódicos e artigos, mas a experiência de um profissional da área é condição “sine qua non” nessa fase. Após a primeira fase, o analista de dados à fase do **Entendimento dos dados**. Nessa fase o analista “olha” para os dados com a acurácia de um especialista, procurando identificar qualidade nos dados. Dados ausentes são comuns em bases de dados não estruturadas, os “missing data” são sempre um problema a ser considerado, pois seu tratamento pode consumir muito tempo do analista de dados. A fase da **Preparação dos dados**, cobre a construção final do conjunto de dados. Preparar os dados significa criar e selecionar atributos, criar tabelas ou planilhas e registros dos dados.

Na fase de **Modelagem de I.A.** a tecnologia deve ser escolhida com critério baseado em experiência do analista de dados. Em sistemas de suporte à decisão, uma tecnologia

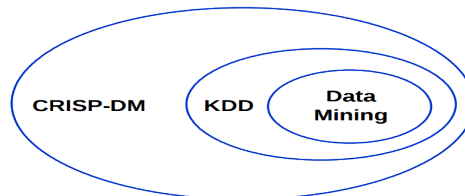
Figura 2.1: O padrão CRISP-DM



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.

Na fase da **Avaliação de desempenho** um ou muitos modelos devem ter sido construídos e testados, apresentando alta qualidade da perspectiva da análise dos dados. Criar um modelo geralmente não é o fim do processo, contudo é um modelo de entendimento do negócio. Até que se obtenha as respostas satisfatórias, o modelo deverá ser refeito várias vezes até sua **Implantação**. A próxima figura descreve o domínio das técnicas aplicadas à mineração de dados:

Figura 2.2: Domínio das técnicas aplicadas a mineração de dados



Fonte: Neurotech – 2012

Quando são desenvolvidos sistemas de predição e análise de diagnóstico, há que se avaliar 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 (7), 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 (8), 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 (9) em pacientes, e na radiologia para classificação de imagens.

Essas métricas são amplamente utilizadas na classificação binária de resultados contínuos. Para isso ser construído, a Matriz de Contingência 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:

A matriz da Tabela 2.3 sintetiza a matriz da Tabela 2.4, portanto as duas tabelas são equivalentes. De acordo com as probabilidades condicionais temos:

Tabela 2.1: Matriz de Confusão

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

Tabela 2.2: 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(Y)	—

$$P(X, Y) = P(X|Y).P(Y) = P(Y|X).P(X) \quad (2.1)$$

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 ($\text{tpr} = P(Y|X)$) com a taxa dos falsos positivos ($\text{fpr} = P(Y, \bar{X})$).

2.3 Data Mining

Técnicas de mineração de dados trabalham com dados estruturados, preenchidos em sua totalidade (sem *missing data*), para poder extrair informações relevantes. Um dos maiores problemas na extração de informações é que os dados não estão estruturados, ou não estão no grão adequado, ou ainda faltam dados (“missing data” – dados ausentes). Para contornar o problema de dados ausentes existem várias técnicas, como preenchimento dos dados através de técnicas de inteligência artificial.

O caminho da extração dos dados até sua mineração e, por fim, extração de conhecimento é longa. Na figura a seguir temos um exemplo desse caminho:

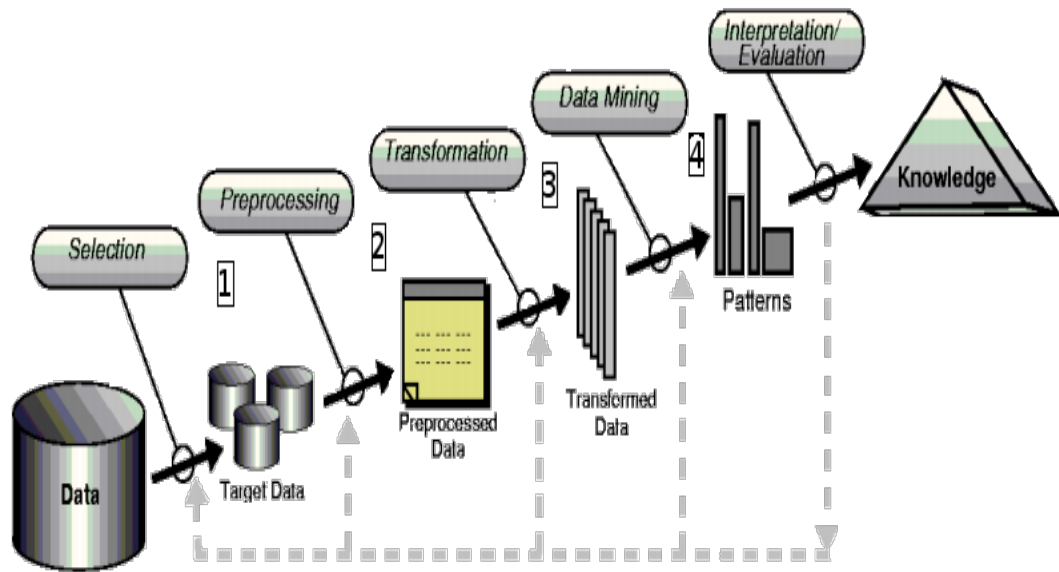
O Big data está representado na figura onde se lê “Data” e está repleto de *missing data* e/ou dados inconsistentes, conhecidos como dados não estruturados. O balão onde se lê “Selection” representa a coleta das informações ou a seleção dos dados no *Big Data*. Em nossa pesquisa esses dados são provenientes das mais diversas fontes, tais como, redes sociais, câmaras de trânsito, informações de satélites meteorológicos e outras fontes.

Armazenar qualquer quantidade de dados nessa etapa pode ser um grande problema, devido à sua extensão, porém os dados relevantes podem ser armazenados em “Target Data” com a tecnologia Hadoop, utilizando as técnicas de “Map” e “Reduce” para criar *cluster* de informações e ler os fluxos de dados (stream data). Algumas técnicas de IA podem ser aplicadas nessa etapa como, “Data Mining Swarm Robotics” através de Botnets¹ 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 (10). Esses dados normalmente são coletados

¹Botnet é citado no sentido da coleta de informações

Figura 2.3: Minerando dados no Big Data



(Excerto de Fayyad et al., - 1996)

por técnicas de Mineração de Textos, também conhecidas como Mineração de Dados em Textos, técnicas de IA como “Machine Learning” têm sido muito utilizadas. Em “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 (11). As Redes Neurais, dependendo da quantidade de variáveis de entrada, poderão 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 prever 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).

²Metaheurística são heurísticas aplicadas em problemas onde os custos computacionais não são tratáveis em tempo polinomial, devido às explosões combinatórias geradas pelo grande número de tentativas. Metaheurísticas bioinspiradas metaforizam o comportamento de animais sociais, tais como formigas, pássaros, peixes e outros

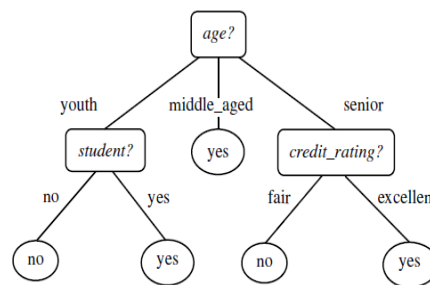
2.4 Machine Learning

O desafio dos 3 V's (Velocidade, Variedade e Volume) pode ser um fator determinístico na escolha da ferramenta mais adequada para analisar dados e extrair informação. As árvores de decisão 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, (11) isso pode representar o desafio maior para implementar esta técnica.

2.4.1 Árvore de Decisão

Han e Kamber (13) 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. A seguir, a árvore de decisão que explicita se um cliente, de acordo com sua idade, irá efetuar ou não a compra de um computador:

Figura 2.4: Árvore de decisão



Para Ian e Frank (14), 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.

O algoritmo *C4.5* é considerado um exemplo clássico de método de indução de árvores de decisão. O *C4.5* (15) foi inspirado no algoritmo *ID3* (16), 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 (17).

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

2.4.2 Regressão Logística

Os modelos de regressão (linear e logística) são técnicas para analisar o relacionamento entre variáveis. No entanto, a regressão linear é utilizada para problemas de natureza

contínua, sendo que a regressão logística é semelhante, contudo, a variável dependente não é contínua, é discreta ou categórica (18).

A regressão logística está definida como o logarítmo a seguir:

$$\log \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (2.2)$$

onde $\pi(x)$ é definido como:

$$\pi(x) = \frac{1}{1 + e^{-\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}} \quad (2.3)$$

e x_1, x_2, \dots, x_p são as variáveis a serem exploradas.

A aplicação da regressão logística foi utilizada, pela primeira vez, com sucesso na oferta de crédito nos anos seguintes ao fim da 2ª guerra mundial, para tomar decisão de oferecer crédito a terceiros (19). A regressão logística é comumente aplicada para problemas de classificação binária (ou booleano).

2.4.3 Redes Neurais

2.5 Redes sociais

As redes sociais são um arcabouço de informações sobre todo tipo de assunto vivenciado no cotidiano das pessoas, inclusive situações que dizem respeito ao nosso ambiente de pesquisa. O cenário abaixo, encontrado numa rede social, exemplifica a sequência de informações retiradas do Twitter, um microblog onde os usuários escrevem num pequeno espaço (cerca de 140 caracteres), os mais diversos assuntos, e os usuários conectam por uma multiplicidade de dispositivos: computadores, tablets e celulares, formando uma grande rede social mundial. A ideia inicial do Twitter, segundo seus fundadores, era que essa rede se comportasse como um “SMS da Internet” (20). As informações são enviadas aos usuários, conhecidas como twittes, em tempo real e também enviadas aos usuários seguidores que tenham assinado para recebê-las.

"Em 2010 empresas e usuários armazenaram mais de 13 exabytes de novos dados"(21).

Tabela 2.3: 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}

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



A Polícia Rodoviária Federal de Santa Catarina, 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 trânsito 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://vejaovivo.com.br> dentre outros.

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

2.5.1 Data Mining - Text Data Mining

Minerar dados em texto nas redes sociais não é uma tarefa atômica, devendo ser dividida 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” (10). 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 2.5: Mapa mental da Mineração em textos



3

Metodologia

A metodologia utilizada no modelo preditivo é o CRISP – DM. de Suporte a Decisão que forneça informação suficiente a um gestor para decidir quando e por onde enviar uma frota de caminhões 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 providas de GPS, contudo não analisam dados históricos dessas rodovias nem fazem previsões futuras sobre o comportamento delas.

3.1 Plano geral da metodologia

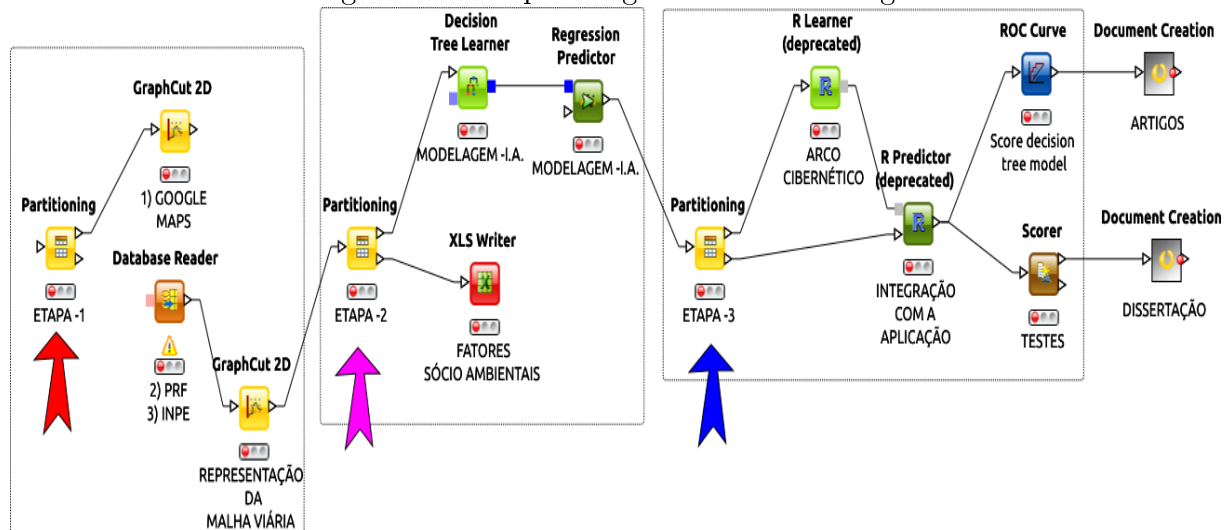
A metodologia contempla um plano em três etapas, cada uma dividida em fases atinentes. A figura a seguir ilustra essa metodologia descrita graficamente, onde as três etapas são representadas por retângulos.

A etapa 1 contempla a fases da coleta das bases de dados históricas e proposição de um modelo de representação único.

O retângulo central é a etapa 2, que consiste nas fases de Identificação dos fatores sócio ambientais na bases históricas, a modelagem do sistema preditivo e aplicação das técnicas de IA. Propomos inicialmente a Regressão Logística e Árvores de decisão, testes iniciais e depuração do modelo.

A última etapa da nossa metodolia contempla a coleta das bases vetoriais, esta contém um arco cibernético construído com os dados de redes sociais, como por exemplo a API do Twitter, este “per si” permite uma contraposição ao modelo preditivo na perspectiva do usuário gestor. Isso pois, modelos preditivos com o passar do tempo tendem a desfazer-se. Uma proposta algorítmica para substituir a API das redes sociais poderá ser desenvolvida e testada numa fase complementar, o algoritmo Ant-Miner poderá vir a ser um candidato de adaptação.

Figura 3.1: Etapas da gerais da metodologia



A representação da malha viária que acopla a estrutura dinâmica com a outra estrutura será o “front-end” da metodologia proposta estará na terceira etapa.

3.2 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. Essas bases de dados são integradas gerando um único e complexo modelo preditivo que será acoplado a estrutura dinâmica.

3.3 Reflexão sobre as tecnologias utilizadas no modelo preditivo

Não existe uma técnica de mineração que generalize os mais diversos ambientes preditivos, mas sim um “pool” dessas técnicas onde uma complementa outra. As técnicas preditivas tradicionais que contemplam análise de grandes massas de dados como base heterogêneas são possíveis quando adaptadas para uma forma comparável à que foram inicialmente concebidas, por que as variáveis em uma base de dados a priori guardam pouca relação as variáveis de outra base de dados. neste caso essas variáveis ou são excluídas ou são transformadas a fim de “guardarem” um correlação com a outra base de dados. Quando há uma proximidade entre as bases de dados, na fase de transformação de dados, onde são criadas novas variáveis, a proximidade entre as bases se estreita. Nesta pesquisa, bases heterogêneas foram integralizadas num única grande base, onde as variáveis independentes foram em sua maioria preservadas e/ou 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

3.3.1 As variáveis do modelo preditivo

Algumas técnicas de IA são altamente sensíveis a dados ausentes os “missing data” à dados com pouca consistência e outros tipos de dados comuns em bases mantidas sem um bom critério de inserção dos dados. A variável dependente foi designada como **gargalo** e as variáveis independentes (ou explicativas) são:

Tabela 3.1: Variáveis do modelo preditivo

KM	Numeração do quilômetro
BR	Numeração da Br
condPista	Condição da pista: seca, molhado, ...
restVisibili	Restrição de visibilidade: inexistente, neblina, .., outros
tipoAcident	Tipo de Acidente: atropelamento, colisão, paralisação,...
tipoDano	Tipo de Dano: leve, médio, grave
Município	Localidade onde ocorreu
Ano	Ano que ocorreu o acidente
Mês	Mês que ocorreu o acidente
Dia	Dia que ocorreu o acidente
Hora	Hora que ocorreu o acidente

À base de dados da PRF, relativas a interdições das vias, por motivos diversos, não haviam variáveis tais como; visibilidade, condições da via, gravidade paralisação e outras. Foram incorporadas à essa base essas novas variáveis, para populá-las, adotou-se a lógica; presumivelmente protestos são realizados com boa visibilidade, em condições de via razoáveis e a gravidade da paralisação foi considerada leve.

As técnicas como Redes Neurais Artificiais (MLP) [CITAR], Árvores de decisão (CART) [CITAR], Regressão logística (MLR) [CITAR] fornecem visão generalizada dos fatores preponderantes, levantando padrões ocultos nos dados. Esta fase é conhecida como Aprendizagem de Máquina (acrônimo de Machine Learning)

- Redes Neurais Artificiais 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.
- Árvores de decisão para classificar acidentes 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* com nível de acurácia próximo aos 80%
- 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 Extração do conhecimento

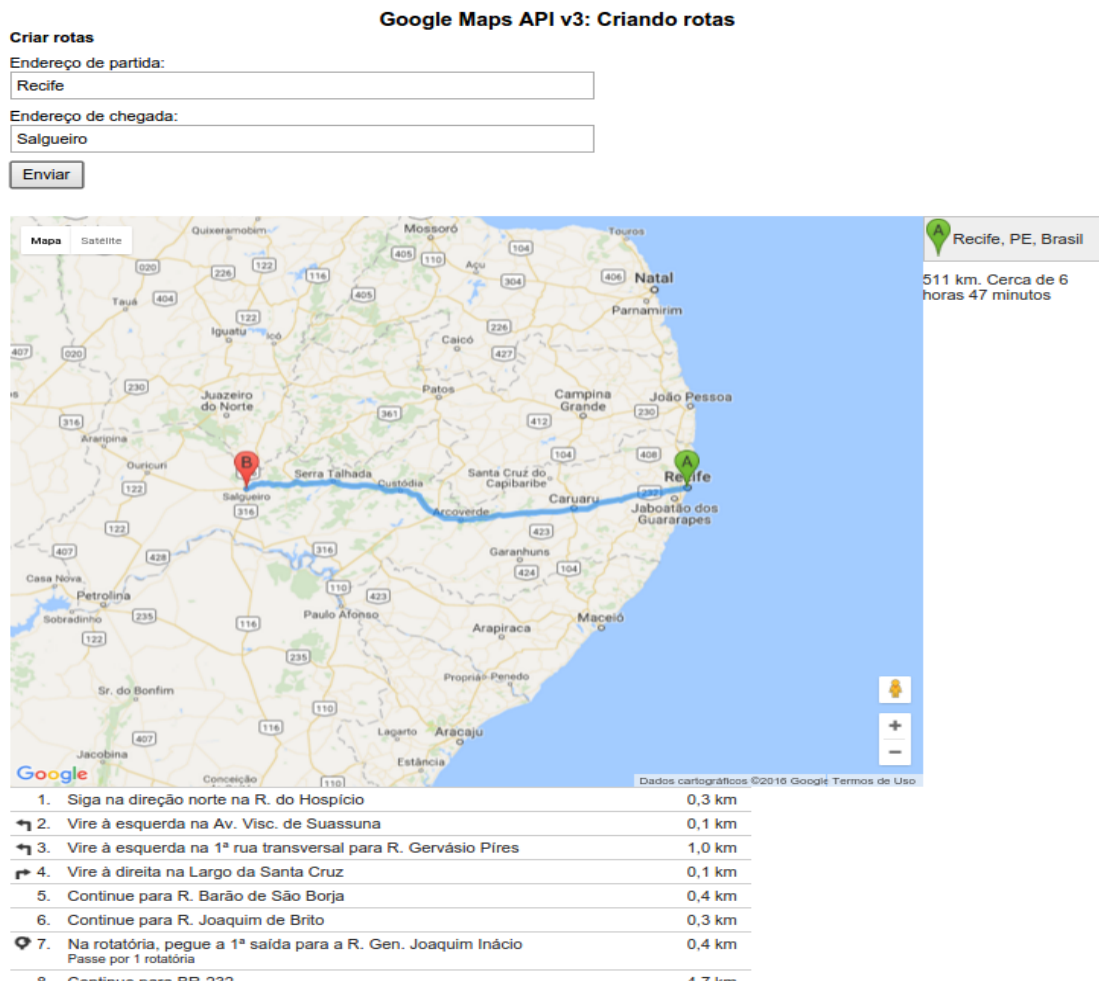
3.5 Acoplamento com a estrutura dinâmica

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 também tem a possibilidade de atualizar o modelo preditivo, contudo o objetivo desta é fazer um Arco cibernético, retroalimentando todo o sistema com novas informações, pensamos que isso permite que o primeiro módulo (preditivo) seja atualizado de tempos em tempos, quando for

A API Google Maps portanto é o “front-end” do Sistema e uma futura aplicação que poderá ser desenvolvida para ser executada em um aparelho celular, “Smartphones”, com capacidade para executar aplicativos mais complexos.

Figura 3.2: Etapas da metodologia



4

Resumo, Discussão & Trabalhos futuros

4.1 Conclusão

As rodovias federais brasileiras que cruzam as regiões metropolitanas das grandes cidades se apresentarão sempre como um gargalo no fluxo do transporte de cargas devido ao crescimento do número de veículos que nelas trafegam.

A solução proposta visa reduzir e dirimir os atuais gargalos burocráticos e tecnológicos para obtenção das bases de dados históricas de entidades públicas, os direitos autorais para utilização de APIs e de tecnologias específicas necessárias para apropriação via Internet. Entendemos ser normal esse cuidado por se tratar de informações de órgão que devem primar pelas informações de seus usuários. Mas tentaremos suprir demandas reais e importantes sem que isso represente algum risco à privacidade dos geradores de dados. Em suma, nossa solução pretende mitigar o gargalo da logística de transporte de cargas, oferecendo uma solução possível à gestão de frotas de veículos que trafegam em rodovias, notadamente no caso do entrono metropolitano do Recife.

4.2 Discussão

Algumas propostas para a RMR vêm sendo amplamente difundidas pelas mídias, tais como o arco metropolitano. Arcos metropolitanos, para além dos transtornos de se contruir um, são muito caros, requerem constantes manutenções e com o passar dos anos, com o crescimento populacional no

seu entorno, tornam-se novamente um novo gargalo para o transporte de cargas.

Gerir como as rodovias são utilizadas é a maneira mais racional, elas estão aí para auxiliar no transporte de pessoas, mercadorias e para serviços, portanto é de todos e todos têm o dever de contribuir preservando-as e respeitando o direito dos outros.

4.3 Trabalhos futuros

O modelo preditivo poderá vir a ser retro-alimentado, para uma futura predição ...



Preprocessamento

A.1 Preprocessamento dos dados coleta 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* (22), 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.(23) As bases de dados do INPE e do base de dados do IBGE apresentaram boa qualidade o que justificou serem coletados diretamente da Internet.

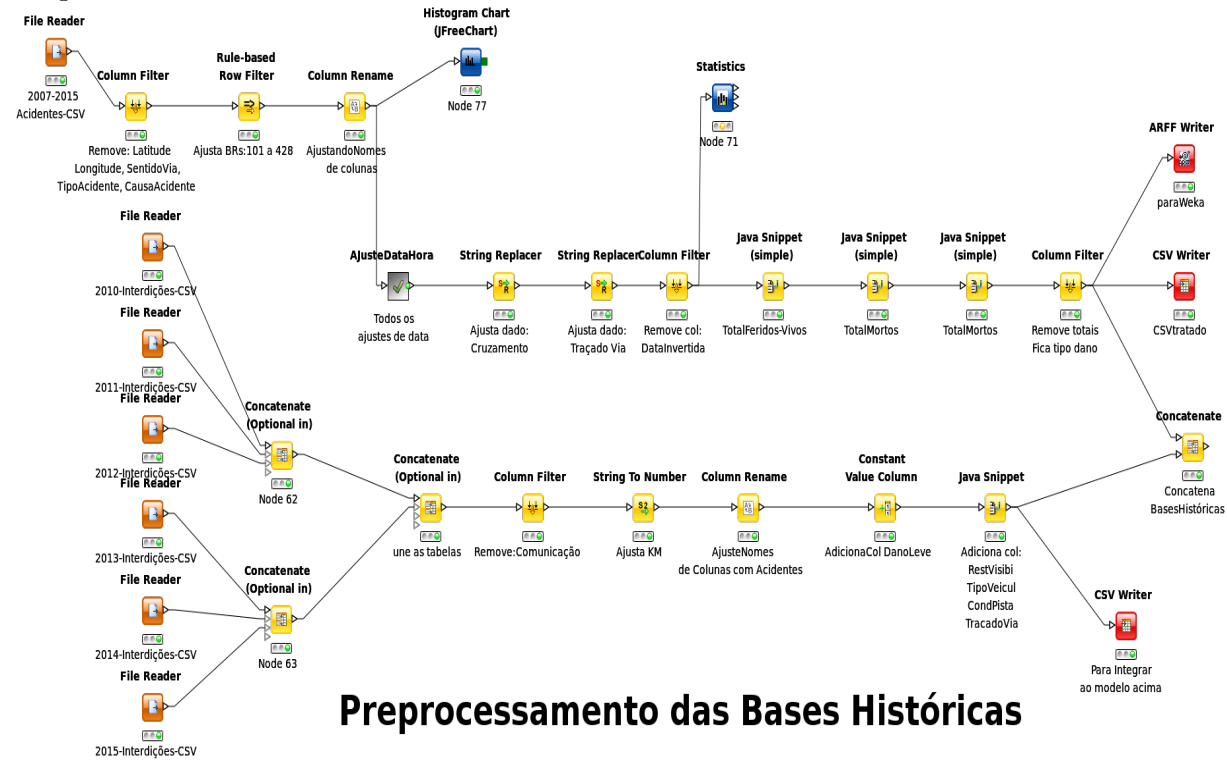
Tabela A.1: Variáveis originais da base de acidentes

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 QUADROS, C. I. D. Dez Anos Depois do. *Intercom*, v. 1995, p. 65–69, 2005.
- 2 DILSIZIAN, L. E. S. E. Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. p. 1.
- 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 WIRTH, R. Crisp-dm 1.0 – step-by-step data mining guide. 2000.
- 7 EGAN, J. P. Signal detection theory and roc analysis. New York, USA: Academic Press, 1975.
- 8 ANAESTHETIST, T.
- 9 SOUZA, C. R.
- 10 ARANHA, C.; PASSOS, E. *A Tecnologia de Mineração de Textos*. 2006. 1–8 p.
- 11 SRIVASTAVA, V. K. A.; SINGH, N. Review of decision tree algorithm: Big data analytics. *International Journal of Informative & Futuristic Research*, v. 2, n. 10, p. 3644–3654, 2015.

- 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 HAN, J.; KAMBER, M. Data mining: Concepts and techniques. Elsevier, San Francisco, v. 2 edition, 2006.
- 14 WITTEN, I. H.; FRANK, E. Data mining: Practical machine learning tools and techniques. *Elsevier, San Francisco*, v. 2 edition, 2005.
- 15 QUINLAN, J. R.
- 16 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.
- 17 HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.: s.n.], 2006.
- 18 WEST, D. Neural network credit scoring models – computers and operations research. p. 1131 – 1152, 2000.
- 19 M, H. J. Logistic regression models. Chapman and Hall – CRC Press, 2009.
- 20 DORSEY E. WILLIAMS, B. S. J.; GLASS, N. Twitter. July 2015. Disponível em: <<https://pt.wikipedia.org/wiki/Twitter>>.
- 21 JAGADISH J. GEHRKE, A. L. Y. P. J. M. P. R. R. H. V.; SHAHABI, C. Exploring the inherent technical challenges in realizing the potential of big data. *Communication of the ACM*, v. 57, n. 7, p. 86–96, July 2014.
- 22 2016. Disponível em: <<http://dados.gov.br/dados-abertos/>>.
- 23 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>>.