ANÁLISE DE COMPORTAMENTO RODOVIÁRIO NA REGIÃO METROPOLITANA DO RECIFE-BRASIL

UMA ABORDAGEM DE MINERAÇÃO DE DADOS

ANALYSIS OF ROAD-TRAFFIC BEHAVIOR IN THE METROPOLITAN AREA OF RECIFE-BRAZIL

A DATA MINING APPROACH

Resumo — **As rodovias federais que atravessam a Região Metropolitana de algumas cidades estão constantemente congestionadas, não apenas pela quantidade de veículos, mas por serem alvo de paralisações das mais diversas matizes, como acidentes, buracos, intempéries naturais e outros tipos de fatores de congestionamento. Para dirimir alguns destes problemas propomos um modelo de classificação para o 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 de tráfego, como retenção, redução de fluxo de tráfego. A fonte de dados dessa pesquisa provém da base de dados da Polícia Rodoviária Federal de Pernambuco (PRF) a partir de 2007 até 2016 tendo considerado veículos, traçado da via e trechos da rodovia relacionados a acidentes, dentre outros. Com base nas informações obtidas, foi realizada uma Mineração de Dados utilizando a metodologia CRISP-DM para encontrar padrões comportamentais nas rodovias e em seu entorno. Foram empregados algoritmos de aprendizagem de máquina para classificação e regressão, sendo priorizadas, Árvores de Decisão e Redes Neurais. Os valores da área sob a curva ROC (AUC) obtidos foram acima de 0.7 que reflete um bom grau de confiabilidade. Em comparação com abordagens usuais de navegação, o modelo de predição proposto representa um avanço em termos de mobilidade e gestão do transporte de cargas, uma vez que possibilita antecipar eventos e comportamentos. Assim possibilitando que sejam favorecidas rotas alternativas e ampliando o espaço temporal de escolha para determinadas rotas.**

Palavras Chave – Modelo de predição; Mineração de dados; Tráfego em rodovias; Árvores de decisão, CRISP-DM.

Abstract — **Federal highways that cross the Metropolitan Regions of some cities are constantly congested, not only by the number of vehicles, but also because they are subject to stoppages of the most diverse natures, such as accidents, holes, natural weather and other types of problems. We propose a classification model of behavioral patterns for the federal highways that cross the metropolitan area of Pernambuco, which is a state in the Northeastern region of Brazil. The proposed model allows some events anticipation, especially those that may cause constraints, retention, or reduction of traffic flow. The data source of this research is from the database of the Federal Highway Police of Pernambuco (PRF) since 2007 until 2016. We have considered vehicles, track layout and road sections related to accidents, among others. Based on the information obtained, a Data Mining was performed using the CRISP-DM methodology to find behavioral patterns on highways and in their surroundings. Machine learning algorithms were used for classification and regression, being prioritized, Decision Trees and Neural Networks. The values of the area under the ROC (AUC) curve obtained were above 0.7 which reflects a good degree of reliability. The proposed prediction model means an advance in terms of mobility and cargo transport management, since it allows anticipating events and behaviors, favoring the choice of alternative routes and increasing the time for choice for certain routes**

Keywords – Prediction model; Data mining; Road traffic; Decision tree; CRISP-DM.

I. INTRODUÇÃO

O transporte de cargas que atravessa as regiões metropolitanas das grandes cidades brasileiras é realizado principalmente pelas rodovias federais brasileiras. Essas rodovias frequentemente se encontram congestionadas em determinados dias/horários, principalmente no entorno das metrópoles. Recentemente, tem sido contabilizado aumento expressivo aumento de novos veículos incorporado à frota nacional, a cada ano. Por outro lado segundo o DataSUS – Departamento de Informática do Sistema Único de Saúde do Ministério da Saúde, foram registrados 43 075 óbitos nas vias brasileiras em 2014, número que em 2015 passou de 45 000. O estado de Pernambuco, localizado na região Nordeste do Brasil, possuía, em 2015, uma frota de 2.765.521 de veículos,

sendo que boa parte dessa frota trafega pelas onze rodovias federais que cruzam o estado [1]. Sendo a Polícia Rodoviária Federal um dos órgãos de controle público responsável para atender e registrar os acontecimentos nessas rodovias. Estas que serão o estudo de caso deste trabalho.

O objetivo dessa pesquisa, foi propor um modelo de predição de comportamento das rodovias, de modo a tornar possível a escolha de dias, horários e locais para trafegar com menos interrupções devido ao fluxo do trânsito intenso. O mesmo interesse neste assunto foi despertado em outros pesquisadores. Costa, Bernardini, Lima e Viterbo [11] destacam algumas a **falta de atenção** do condutor a principal causadora dos acidentes. Eles utilizaram árvores de decisão e o algoritmo gerador de regras de associação Apriori proposto por Agrawal, Imielinki e Swami(1993). De forma análoga pretendemos minerar os registros para antecipar os problemas e assim melhor apoiar decisões de roteamento rodoviário.

II. FUNDAMENTAÇÃO TEÓRICA

Nessa sessão apresentaremos algumas pesquisas correlatas: que analisam o tráfego em vias e rodovias e uso de técnicas de inteligência artificial, particularmente, árvores de decisão e redes neurais, dentre outras. No Brasil no sul-sudeste, algumas pesquisas têm sido conduzidas nesse campo. Corrêa [18] desenvolveu um estudo exploratório, em que buscou identificar pesquisas no Brasil que tratassem da utilização de redes neurais aplicadas ao setor de transporte, comparando com aquelas produzidas, sob a mesma temática, em países desenvolvidos. Essa autora identifica a produção de estudos nesse campo desde os anos 90, corroborando a breve análise que fizemos, utilizando como ferramenta de busca o google acadêmico. Nos EUA, uma pesquisa foi conduzida com dados obtidos a partir do Sistema de Relatório de Análise de Fatalidades (FARS), no Alabama, Shanthi e Ramani [20] buscou analisar em que medida os algoritmos de classificação de mineração de dados são eficazes para previsão dos fatores que influenciam acidentes de trânsito, levando em conta a gravidade da lesão decorrente do acidente. Nessa pesquisa foi comparado o desempenho de algoritmos como C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes e Árvore de randomização, para modelar a gravidade da lesão ocorrida no acidente. A acurácia foi avaliada com base na precisão e valores de recall e os resultados apontaram para a eficácia desses algoritmos, com destaque para as árvores de randomização. As pesquisas desenvolvidas no Brasil têm focado na identificação de problemas relacionados a congestionamento de vias urbanas ou rodovias no interior do estado, propondo, a partir da utilização de técnicas de IA, como redes neurais, árvores de decisão, lógica fuzzy, identificar o comportamento dos veículos nas vias, como a pesquisa de Ferreira [21], que buscou utilizar combinação de técnicas de IA para previsão do comportamento do tráfego na cidade de São Paulo.

A técnica de extração do conhecimento de grandes bases de dados é conhecida como "Knowledge Discovery Databases" (KDD). No processo de extração do conhecimento o KDD se caracteriza pela aplicação de algoritmos específicos para descoberta de padrões e/ou comportamentos em grandes bases de dados, também conhecidas como repositórios de dados. 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 [1].

A. O CRISP-DM

O "CRoss Industry Standard Process for Data Mining" CRISP-DM é um processo de mineração de dados que descreve como especialistas nesse campo aplicam as técnicas de mineração para obter os melhores resultados [3]. O CRISP-DM é um processo recursivo, em que 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.

O contexto da aplicação do CRISP-DM [3] 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 aplicação das técnicas de mineração de dados identifica padrões ocultos nos dados, inacessíveis pelas técnicas tradicionais, como por exemplo, consultas a banco de dados, técnicas estatísticas, dentre outras. Além disso, possibilita analisar um grande numero de variáveis simultaneamente, o que não acontece com o cérebro humano [4]. Fayyad [5] destaca a natureza interdisciplinar do KDD que contempla a intersecção de campos de pesquisa tais como Aprendizado de Máquina (Machine Learning), Reconhecimento de Padrões, Inteligência Artificial, estatística, computação de alto desempenho e outros. Propõe, ainda que o objetivo principal seja extrair um conhecimento de alto nível a partir de dados de baixo nível.

O modelo CRISP-DM contempla seis fases para um projeto de mineração de dados, sendo assim determina-se um ciclo de vida compreendido para cada uma dessas fases: A primeira fase, conhecida como Entendimento do negócio, ou "fase de entendimento dos objetivos é dos requerimentos sob a perspectiva do negócio" segundo CHAMPAN [6] e uma fase crucial da mineração. A fase Entendimento dos dados caracteriza-se pelo exame acurado dos dados, procurando identificar a sua qualidade. Dados ausentes "missing data" – são comuns em base de dados não estruturados, configurando-se como um problema a ser considerado. A terceira fase, Preparação dos dados diz respeito sobre a construção final do conjunto dos dados, cada conjunto de dados é "explicado" por um atributo, para selecionar quais dados serão mais relevantes, para variáveis numéricas calcula-se o coeficiente de correlação entre os atributos (variáveis). Outra forma de qualificar os dados é calculando a quantidade de informação que cada atributo possui. A máxima entropia de cada atributo pode

fornecer informações sobre a qualidade da variável quando esta estabelece ganho de informação [10], vide a equação da Entropia (1).

$$H_x = - \sum_{\forall x \in X} P(x) log_2 P(x) \quad (1)$$

Onde H_x é a medida de entropia, x um atributo do conjunto X de variáveis. H_x pode variar entre 0 e 1, quando o valor da entropia tende para 1 significa que maior ganho de informação. A entropia condicional, formalizada na equação (2), é a entropia restante dos atributos de Y no valor y quando o atributo X é dado como x [17]:

$$H_{Y|X} = - \sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y|x) log_2 P(y|x) \quad (2)$$

Na quarta fase do CRISP-DM, Modelagem de I.A. a tecnologia deve ser escolhida de forma criteriosa, baseada na experiência do analista de dados. Na fase cinco, Avaliação de desempenho um ou mais modelos devem ter sido construídos e testados de forma que o modelo esteja adequado aos objetivos do negócio. A sexta e última fase, Implementação caracteriza-se pela conclusão do processo, pode ser necessário retomar todo o ciclo até que o modelo esteja adequado as necessidades especificas determinadas previamente. A figura 1 ilustra o domínio das técnicas aplicadas à mineração de dados, note-se a generalidade do CRISP-DM.

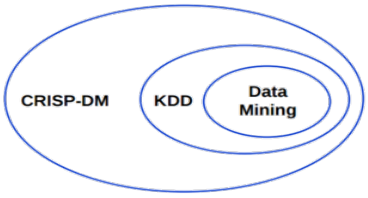


Figura 1: Domínio das técnicas aplicadas à mineração de dados e generalidade do CRISP-DM. Fonte Neurotech S.A. [7]

B. *Aprendizagem de Máquina*

Aprendizagem de Máquina ou "Machine Learning" são técnicas para analisar dados de forma automatizada e interativa. Segundo Shalev-Shwartz & Bem-Davis [8] o termo Aprendizagem de maquina refere-se à detecção automática de padrões de dados.

Para Nilsson [9], 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 forma que melhora o desempenho futuro. Sistemas que executam tarefas de inteligência artificial, tais como Reconhecimento de padrões, Diagnósticos, Controle de Robôs, Predição e outros, precisam ser modificados para executarem "Machine Learning".

C. *Aprendizagem Bayesiana*

Dado um conjunto de variáveis aleatórias Ω = {BR, Km, RestVisibilidade, TipoVeiculo, TipoAcidente, CausaAcidente, TraçadoVia, Período, Gravidade, DiaSemana, Delegacia}. a variável aleatória H (hipótese) denota o tipo de Ω, com valores possíveis para $h_1,..,h_{11}$. 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 o espaço de variáveis para uma predição sobre a parte desconhecida de X temos [17]:

$$P(h_i|d) = \sum_i P(X|d, h_i) P(h_i|d) = \sum_i P(X|h_i) P(h_i|d) \quad (3)$$

Onde cada hipótese h_i determina uma distribuição de probabilidades sobre a variável X.

D. *Classificação e Regressão para análise preditiva.*

Classificação é um processo para encontrar um modelo que descreve e distingue classes de dados. Esse modelo tem como base de análise um conjunto de treinamento (i.e. objetos de dados para os quais serão encontrados rótulos que os classifiquem) e é utilizado para predizer quais rótulos de classes terão os objetos desconhecidos [5].

Para aplicar classificação em dados IA utiliza-se de algoritmos de Aprendizagem de Maquina tais como: Árvores de decisão, Naive Bayes, Redes Neurais Artificiais, dentre outros algoritmos especializados em descobrir padrões nos dados, geralmente pretende-se com isso adquirir novos conhecimentos a partir do entendimento desses dados, esse processo de descoberta de novos conhecimentos é conhecido como KDD – Knowledge Discovery Database [1]. Após escolher a origem dos dados (inputs), o KDD é dividido em fases atinentes: *(i)Seleção* – dados alvos são escolhidos a partir do(s) conjunto(s) de dados a serem minerados; *(ii)Preprocessamento* – os dados devem passar por um tratamento retirando-se as inconsistências, dados ausentes "missing data"; *(iii) Transformação* – nessa fase são criadas novas variáveis a partir dos dados para adequarem-se aos processos de mineração; *(iv)Mineração de dados* – essa fase aplicam-se técnicas de I.A. com os algoritmos de Aprendizagem de Maquina; *(v)Interpretação/Avaliação* – penúltima fase, interpretam-se os resultados estando de acordo passa-se a próxima fase, não estando de acordo retornam-se as fases anteriores; *(vi)Descoberta de conhecimento* – extraem-se conhecimentos para por exemplo predizer resultados futuros a partir de novos dados dessa natureza.

E. *Árvores de Decisão*

Tem como entrada um conjunto de atributos (variáveis) para retornar como saída um decisão. O valor esperado da saída deve estar de acordo com o que foi dada a entrada.

Árvores de decisão [10] são algoritmos rápidos que produzem regras de indução após uma classificação hierárquica. A fase de extração dos dados é fortemente influenciada pelas variáveis escolhidas, isso pode representar um desafio maior para implementar essa técnica. Outro problema frequentemente encontrado em algoritmos de aprendizagem é o "overfitting" ou superadaptação aos modelos. Segundo RUSSEL E NORVIG [10] o "overfitting" ocorre quando o numero de atributos e grande e o algoritmo "deixa" de aprender.

F. *Naïve Bayes*

Esta 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 predição. Este classificador tem sido muito empregado para

classificar documentos e detectar spam em mensagens[19]. A probabilidade condicional pode ser explicada por um vetor x = (x_1,...,x_n) que representa *n* características (variáveis independentes) que se atribui a esta instancias de probabilidades $p(C_k \mid x_i,...x_n)$ para cada *K* possível ter vindo da classe C_k. Aplicando o teorema de Bayes da probabilidade condicional temos:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (4)$$

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 tambem pode ser entendida como: $p(posteriori) = \frac{p(prori) * verossimilhança}{evidencia}$ (5)

III. IDEIA PROPOSTA

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

A. *Aplicação do 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 [11] destacam a não padronização e não aceitação dos dados pela comunidade internacional. EAVES, D. [12] 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

- BR – Nomenclatura da rodovia (i.e. BR 101);
- Km – Quilômetro em que deu a ocorrência;
- Tipo de veículo – envolvido na ocorrência, ex.: carro, motocicleta, caminhão, etc.;
- Tipo de acidente – colisão lateral, frontal, traseira, etc.; atropelamento: com ou sem morte, envolvendo pessoas e/ou animais;
- Horário e data da ocorrência; dentre outros que serão apresentados mais adiante.

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).

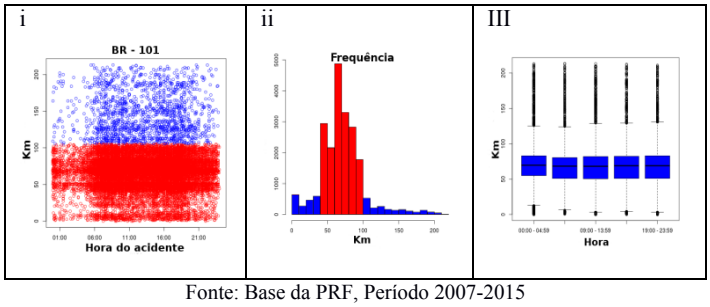
Mineração de dados: O algoritmo escolhido para a pesquisa foi Árvore de decisão que possibilita uma interpretação imediata e de fácil compreensão. Como ferramentas, foram escolhidas o Knime [13] e R [14] e o Weka [15], com objetivo de estabelecer uma comparação entre ambos, cuja intenção era produzir um classificador mais preciso. Nessa direção, a técnica Ensamble de classificadores [13] 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.

B. *Dados antes da mineração*

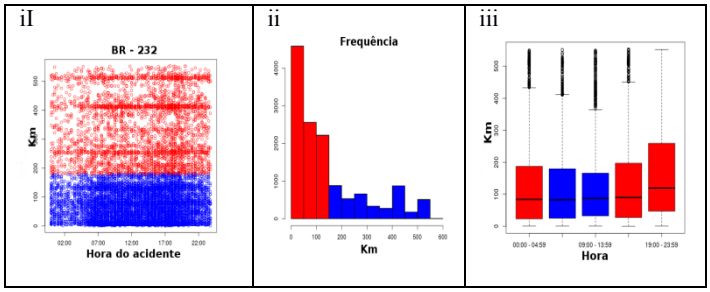
Como já sabido os condutores dos automóveis têm um papel preponderante na causa dos acidentes. As três tabelas a seguir, extraídas das bases da PRF, ajudam a elucidar os efeitos das conduções e acidentes das rodovias federais do nordeste brasileiro.

Tabela I: Gráficos de acidentes BR X Hora BR101.



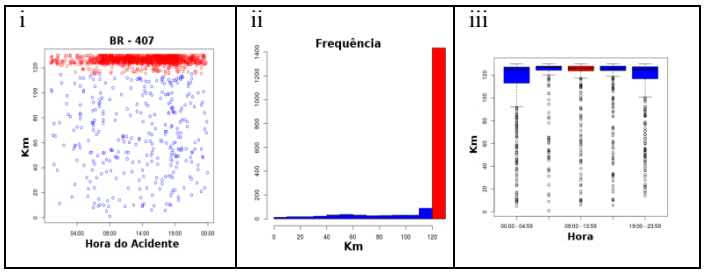
Fonte: Base da PRF, Período 2007-2015

Tabela II: Gráficos de acidentes BR X Hora BR232



. Fonte: Base da PRF, Período 2007-2015

Tabela III: Gráficos de acidentes BR X Hora BR407



Fonte: Base da PRF, Período 2007-2015

As Tabelas I, II e III contém gráficos de três BRs importantes para o nordeste brasileiro. As duas primeiras são as mais utilizadas no Estado de Pernambuco (BR 101, BR 232), com tráfego intenso. A terceira, BR 407 possui tráfego menos intenso. Em cada uma das três tabelas, os gráficos i representam os acidentes que ocorreram a cada hora (abcissa) em cada Km (ordenada) nos últimos nove anos. Os gráficos ii correspondem a frequência do local onde ocorreram esses acidentes. É possível perceber que há determinados locais, em cada rodovia, onde se concentram os acidentes. Os gráficos iii, tipo 'boxplot', exibem 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 nos gráficos 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. Na BR 232 há um padrão nos acidentes nos Km 0, 90, 110, 260, 410 e 500, e na BR 407 os acidentes se concentram na altura do Km 130.

C. Dados 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 nas Tabela 1, 2 e 3). A seguir se resume a classificação realizada.

TP: True Positive; FP: False Positive; Prec.: Precison; F-Me: F-measure ou f-score; AUC: Area Under Curve (Roc)

Prec = TP/(TP +FP)

Recall = TP/ (TP + FN)

F-Measure = 2 * Precison * Recall / (Precision + Recall)

Métrica dos classificadores

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

Instâncias Corretamente Classificadas	7987	47.6324 %
Instâncias Incorretamente Classificadas	8781	52.3676 %
Erro medio absoluto	0.0786	
Erro quadratico medio da raiz	0.2083	

Tabela IV: 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	Col. transversal.
0.026	0.012	0.066	0.026	0.038	0.684	Col. objeto fixo
0.925	0.003	0.920	0.925	0.923	0.980	Atrop.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	QuedaMoto/bicla
0.322	0.002	0.528	0.322	0.400	0.744	Col. 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	Atrop. 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	Col. 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	Derram. 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 V: Matriz de confusão para a variável Tipo de acidente

a	b	c	d	e	f	g	h	Classificadas
527	7	2	385	483	46	2	24	Col. transversal.
16	14	0	69	154	15	0	47	Col. objeto fixo
8	0	483	16	14	0	0	0	Atrop.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	QuedaMoto/bicla
8	0	0	22	38	3	38	1	Col. com bicicleta
28	34	5	85	236	1	2	120	Capotamento
…	…	…	…	…	…	…	…	...

Os valores restantes da Tabela V foram omitidos por não representar valores interessantes para esta variável. As variáveis de classe são as mesmas da Tabela IV.

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

Instâncias Corretamente Classificadas	12110	72.2209 %
Instâncias Incorretamente Classificadas	4658	27.7791 %
Erro medio absoluto	0.3816	
Erro quadratico medio da raiz	0.4368	

Tabela VI: Detalhe da acurácia por classe

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 VII: Matriz de confusão para a variável Gravidade

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

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

Instâncias Corretamente Classificadas	13507	80.5522 %
Instâncias Incorretamente Classificadas	3261	19.4478 %
Erro medio absoluto	0.0469	
Erro quadratico medio da raiz	0.1656	

Tabela VIII: 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
0.457	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
0.213	0.001	0.357	0.213	0.267	0.816	BR110

Tabela IX: Matriz de confusão para a variável BRajustada

a	b	c	d	e	f	g	h	Classificadas
6960	0	0	625	0	0	0	0	BR101
0	1071	0	156	0	0	0	0	BR104
0	0	85	0	90	11	0	0	BR116
970	9	0	3178	1	0	1	0	BR232
0	0	27	11	377	7	0	0	BR316
0	0	15	0	3	675	0	0	BR428
0	39	0	0	0	0	449	92	BR423
0	0	0	0	0	0	172	154	BR424
0	0	0	0	0	95	0	0	BR407
643	0	0	66	0	0	0	0	BR408
0	0	0	0	0	0	26	11	BR110

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 quanto melhor será o classificador. Portanto um numero de verdadeiros positivos acima de 80% e o numero de falsos positivos próximo a 1% combinado com a área da curva AUC dão maior confiabilidade aos testes.

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

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 90% em quase todos as 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 "BRajustada" foram:

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

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

Essa variedade de causas explica que o condutor dessa região não respeita 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.

Baseado nas informações encontradas pelos classificadores foi construída uma matriz onde se pode prever onde ocorrem os mais graves acidentes (com óbito ou envolvendo muitos veículos).

Tabela X: Matriz de Mortos 2D

MatrizMortos2D

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

Na Tabela X, as colunas correspondem às horas do dia, as linhas correspondem aos Km's na rodovia BR 101 (início km 0 e fim km 213). Assim podemos "navegar" conforme as condições favoráveis no cruzamento Hora X Km, calculando uma velocidade e desviando dos "obstáculos" correspondentes aos locais onde ocorrem os acidentes mais graves. Uma terceira dimensão pode ser adicionada a essa matriz que conteria o dia da semana em que ocorrem esses acidentes e uma quarta dimensão com os dias do mês do ano. No Brasil meses como Dezembro, Fevereiro e Junho são atípicos, pois concentram muitas festas, afetando essa matriz com mais ocorrências que os meses normais.

IV. Conclusões

Os resultados do estudo apontam a eficácia da aplicação da I.A. em especial a Árvore de Decisão que obteve elevada acurácia, também permitiu levantar instantaneamente os locais onde ocorrem os acidentes e explicar porque estes ocorrem. Uma desvantagem dessa técnica é o elevado número de nós, mesmo quando se faz podas à árvore. Para analisar questões relativas ao tráfego de veículos órgãos federais de controle de estradas podem se beneficiar de estudos dessa natureza. Por outro lado, essa pesquisa corrobora com os resultados encontrados em estudos semelhantes, seja no Brasil ou em outros países, conforme citado no decorrer do artigo, sugerindo que há um padrão de comportamento em rodovias que pode ser analisado, de maneira a facilitar o transporte de cargas e pessoas em seu curso. Outra contribuição da pesquisa a ser destacada é de cunho metodológico-prático. Do ponto de vista metodológico, pela contribuição pela aplicação do processo CRISP-DM, usado para construir o modelo preditivo. A perspectiva prática, se dá pela proposição de um modelo que integre predição à API de mapas de posicionamento global, fornecendo informação suficiente a um gestor para decidir quando enviar, por exemplo, uma frota de caminhões por determinada rodovia que apresente retenções crescentes de logística de cargas.

Uma extrapolação à Matriz de Mortos poderá ser feita ao marcarem-se os pontos críticos produzidos por esta matriz a API dos mapas de georeferenciamento. Argumentamos que em adição às funcionalidades das 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, com a nossa abordagem dados históricos de rodovias podem auxiliar predições sobre seu comportamento.

Trabalhos futuros incluem a incorporação desta proposta em modelos formais de decisão, por exemplo de roteamento rodoviário metropolitanos.

Referências Bibliográfica

[1] Pernambuco, Evolução anual da frota de veículos, por região. Departamento Estadual de Transito – Detran /PE, 2016

[2] L. Castanheira, "Aplicação de técnicas de Mineração de Dados em Problemas de Classificação de Padrões", 2008, pp553.

[3] R. Wirth, "CRISP-DM 1.0 – Step-by-step data mining guide", 2000, pp. 7– 10.

[4] B. Possas, M. Carvalho, R. Rezende, and W. Meira jr., "Data mining: técnicas para exploração de dados", Universidade Federal de Minas Gerais, 1998.

[5] P. Fayyad, U. Piatetsky-Suapiro, and G. Smyth, "From data mining to knowledge discovery in databases", Advances in Knowledge Discovery and Data Mining, 3rd ed, vol.17, 1996, pp.1– 36.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Thomas, C. Shearer, and R. Wirth.

[7] Neurotech SA – 2012, http://www.neurotech.com.br/ acessado em: 20/01/2017

[8] S. Ben-David, and S. Shalev-Shwartz, "Understanding, Machine Learning: From Theory to Algorithms, booktitle: Understanding Machine Learning: From Theory to Algorithms', 2014, pp. 449.

[9] J. N. Nilsson, " Introduction to Machine Learning" vol. 56, pp. 387 – 389, 2005.

[10] S. Ruseel and P. Norvig, "Inteligência Artificial." .Elsevier, Rio de Janeiro, 3rd ed, pp. 716 – 721, 2004.

[11] J. Costa, F. Bernardini, F. FILHO, "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, p. 1–26, 2015. Disponível em: http://ojs.c3sl.ufpr.br/ojs/index.php/atoz/rt/printerFriendly/41346/25356.

[12] D. EAVES, "The three laws of open government data.", v. 30, 2009.

[13] Knime.org AG, https://www.knime.org/ acessado em 10/02/2017

[14] R-cran, https://www.r-project.org/ acessado em 10/02/2017

[15] Weka, http://www.cs.waikato.ac.nz/ml/index.html, acessado em 10/02/2017

[16] F. Bernardini, "Combinação de classificadores simbólicos utlizando medidas de regras de conhecimento e algoritmos genéticos", Instituto de Ciências e Matemática Computacional/USP, tese de doutorado, 2006.

[17] A. Srivastava, V. Katiyar, N. Singh, "Review of Decision Tree Algorithm: Big Data Analytics", v. 2, 2015.

[18] F. Corrêa e Outros, "Aplicação de redes neurais artificiais no setor de transportes no Brasil", Universidade Federal de São Carlos, 2015.

[19] A. Bifet, and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", 6332 LNAI, p. 1–15, 2010. Disponível em: http://doi.org/10.1007/784211841_'

[20] S. Shanthi, and R. Geetha, "Feature relevance analysis and classification of road traffic accident data through data mining techniques", v. 1, p. 24-26, 2012.

[21] R. Ferreira Pinto e outros, "Combinação de técnicas da Inteligência artificial para previsão do comportamento do tráfico veicular urbano na cidade de São Paulo, Universidade Nove de Julho, 2011.