ISSN: 2237-826X
3(2), jul./dez. 2014

A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras

Data mining and the quality of extracted knowledge from police reports of Brazilian federal highways

Jefferson de Jesus Costa[1], Flávia Cristina Bernardini[1], José Viterbo Filho[1]

[1] *Universidade Federal Fluminense (UFF), Rio de Janeiro, RJ, Brasil*

Autor para correspondência/*Corresponding author*: Jefferson de Jesus Costa [jeffersoncosta@id.uff.br]

Agradecimentos/*Acknowledgments*: Agradecemos a todos os envolvidos no processo de confecção desse trabalho e também aos revisores por suas contribuições, que nos auxiliaram a melhorar o material, além de oferecerem interessantes considerações para trabalhos futuros.

Recebido/*Submitted*: 15 Nov. 2014
Aceito/*Approved*: 21 Dez. 2014

Resumo

Introdução: Apresenta e analisa os resultados encontrados com a aplicação do processo de Mineração de Dados nos boletins de ocorrências de rodovias federais brasileiras gerados pela Polícia Rodoviária Federal (PRF) em 2012. O objetivo desse trabalho é analisar a viabilidade da aplicação do processo de Mineração de Dados sobre os dados fornecidos pela PRF, a fim de identificar associações entre variáveis relacionadas aos acidentes de trânsito em todas as rodovias federais.
Método: Empregaram-se algoritmos de aprendizado supervisionado e simbólico e um algoritmo de regras de associação, ambos implementados na ferramenta Weka. Quanto à base de dados o estudo compreende os registros referentes ao ano de 2012. Sobre essa parcela da base de dados aplicou-se a etapa de pré-processamento dos dados, os quais foram utilizados para extração dos modelos e padrões na ferramenta Weka e, por último, avaliaram-se os modelos e os padrões extraídos.
Resultados: No aprendizado supervisionado, os resultados obtidos com os algoritmos J48 e PART foram considerados promissores, pois para todas as classes de causas de acidente, os valores obtidos de área sob a curva ROC (AUC) estiveram acima de 0,5. Além disso, utilizando-se o algoritmo *Apriori*, foram geradas 38 regras de associação com confiança maior que 0,8.
Conclusão: Conclui-se que é importante uma proposta de modelo para distribuição dos dados dessa base de dados, com o objetivo de utilizá-la para o processo de mineração de dados, bem como para outras tarefas de extração de conhecimento e tomada de decisão. Observa-se, ainda, a necessidade de melhoria da qualidade dos dados a serem disponibilizados desde a fase de coleta, ou seja, nos sistemas para cadastro dos dados.

Palavras-chave: Dados Governamentais Abertos. Mineração de Dados. Regras de Associação. Descoberta de Conhecimento em Bases de Dados.

Abstract

***Introduction**: This paper presents and analyzes the results obtained when applying Data Mining process in the bulletins of occurrences of the Brazilian federal highways generated by the Federal Highway Police (PRF) in 2012. The purpose of this work is to analyze the feasibility of implementing the Data Mining process on data provided by PRF in order to identify associations between variables related to transit accidents in all Brazilian federal highways.*
***Method**: It was used symbolic supervised learning algorithms, as well as an algorithm of generation of association rules, implemented in Weka tool. Regarding the database, it was used the records of 2012. On this portion of the database it was conducted the step of data preprocessing, which were used for extracting models and patterns in the Weka tool and, lastly, evaluated the models and extracted patterns.*
***Results**: In supervised learning, the results obtained with J48 and PART algorithms have been considered promising due to the fact that for all classes of accidents causes, the values of area under the ROC curve (AUC) were above 0.5. Furthermore, using the Apriori algorithm there have been generated 38 association rules with confidence greater than 0.8.*
***Conclusion**: It was concluded that is important to propose a model for data distribution of this database, in order to use it for data mining process, as well as other knowledge extraction tasks and decision making. It was noted still, the need to improve the quality of data to be provided from the initial stage of data gathering, that is, in the very systems used to record the data.*

***Keywords**: Open Government Data. Data Mining. Association Rules. Knowledge Discovery in Databases.*

INTRODUÇÃO

Diversos países têm demonstrado interesse em disponibilizar seus dados governamentais de forma pública, isto é, acessíveis a qualquer cidadão, visando aumentar a transparência nas ações governamentais e a participação popular. Segundo a descrição do Portal Brasileiro de Dados Abertos, esse movimento – denominado Open Data – teve início em 2009, sendo que o Brasil aderiu à iniciativa em 2011 (Brasil, 2014b). Dados deste Portal, de particular interesse para esta investigação, são os registros do Sistema BR-Brasil, desenvolvido pelo Departamento de Polícia Rodoviária Federal (DPRF), e de responsabilidade do Ministério da Justiça (Brasil, 2014a). Segundo o Portal, o Sistema

> [...] visa suprir todas as deficiências operacionais em termos de informatização e controle, substituindo a grande maioria dos serviços burocráticos associados às atividades da Polícia Rodoviária Federal e disponibilizando seus registros on-line em todo o país (Brasil, 2014b).

No Portal Brasileiro de Dados Abertos, e mais especificamente neste Sistema, podem ser encontrados os boletins de ocorrências em rodovias federais do País que aconteceram entre 2007 e 2013 (Brasil, 2014b). Uma tarefa interessante e relevante para a sociedade brasileira seria o de analisar os dados de acidentes rodoviários na tentativa de extrair algum padrão e encontrar os principais fatores que estejam causando esses acidentes. Tal tarefa pode auxiliar o processo de tomada de decisão, assim como futuros planejamentos, para que haja uma redução de acidentes nas rodovias federais brasileiras. Segundo Rezende, Pugliesi, Melanda e Paula (2003) e Witten e Frank (2009), as ferramentas e técnicas do processo de Mineração de Dados (MD) podem ser utilizadas para a descoberta de padrões e, neste particular, Reis (2013) apresenta uma proposta de dissertação que visa aplicar o processo de MD com o objetivo de encontrar padrões nas variáveis envolvidas em acidentes de trânsito na rodovia BR-381, no Estado de Minas Gerais, entre 2008 a 2012. Anteriormente, Balbo (2011) propôs um método de análise multivariada para análise dos acidentes da BR-277. Entretanto, até o momento, se desconhecem trabalhos que explores os dados de todas as rodovias federais brasileiras, bem como que descrevam a aplicação do processo de MD em toda ou em parte dessa base (Sistema BR-Brasil).

Uma das etapas do processo de MD é a de pré-processamento, a qual engloba o tratamento e a preparação dos dados. Para que sejam descobertos padrões de qualidade é importante que essa etapa seja cuidadosamente executada (Rezende et al., 2003; Witten & Frank, 2009). Ainda, segundo Facelli, Lorena, Gama e Carvalho (2011), o desempenho dos algoritmos de aprendizado de máquina geralmente é afetado pelo estado em que os dados se encontram, ou seja, pela qualidade dos dados disponíveis. Podem ser mencionadas algumas das tarefas incluídas nessa fase, a saber: limpeza dos dados, tratamento de ruídos, tratamento de dados faltantes, seleção e construção de atributos, dentre outras. Para este estudo, devido à significativa quantidade de dados disponibilizada no Portal, optou-se pela utilização das ocorrências registradas durante o ano de 2012. Ao estudar a base com maior profundidade, diversos problemas puderam ser observados, resultando em dificuldades no processo de descoberta de novos conhecimentos. Tais problemas são descritos oportunamente neste trabalho.

O objetivo desta investigação é, portanto, apresentar as dificuldades encontradas para aplicar o processo de Mineração de Dados na base de dados da Polícia Rodoviária Federal Brasileira, bem como descrever os resultados obtidos ao aplicar o referido processo. Deve ser observado que foram utilizados os dados de todas as rodovias federais do País, sendo ainda realizada uma discussão sobre alguns tratamentos de dados que foram necessários na base de dados. Para o desenvolvimento deste trabalho utilizou-se a ferramenta Weka para aplicação do processo de MD (Witten & Frank, 2009).

O artigo está organizado da seguinte maneira: na segunda seção é descrita uma breve fundamentação teórica sobre o processo de MD e de Aprendizado de Máquina, e também sobre os algoritmos utilizados. Na terceira seção apresenta-se o domínio da aplicação, ou seja, a base de dados de Boletins de Ocorrência da Polícia Rodoviária Federal; os problemas encontrados para minerar a base de dados; e as etapas de pré-processamento realizadas que permitiram a aplicação do processo de Mineração de Dados na base de dados. Na quarta seção são apresentados os resultados obtidos com os algoritmos PART, J48 e *Apriori*. Na quinta e última seção relatam-se as conclusões e apontam-se trabalhos futuros.

Mineração de dados e aprendizado de máquina

A Mineração de Dados pode ser definida como a exploração e a análise, através de meios automáticos ou semiautomáticos, de grandes quantidades de dados com o objetivo de descobrir padrões e regras significativas (Berry & Linoff, 1997). De acordo com Rezende et al. (2003) e Witten e Frank (2009), o processo de Mineração de Dados pode ser dividido, basicamente, em três diferentes etapas: (i) pré-processamento dos dados; (ii) extração de modelos e padrões; e (iii) avaliação dos modelos e padrões extraídos. A primeira fase – a de pré-processamento dos dados – envolve tarefas de limpeza dos dados, tais como aplicação de filtros, seleção e construção de atributos, preenchimento de valores faltantes, tratamento de ruídos, entre outras. O objetivo dessa fase é tornar os dados estatisticamente de melhor qualidade para extração de padrões. Na fase de extração de modelos e padrões podem ser utilizados diferentes métodos e técnicas de aprendizado de máquina (Rezende et al., 2003; Witten & Frank, 2009). Para utilizar algoritmos de aprendizado de máquina no processo de Mineração de Dados (MD) podem ser empregadas, basicamente, duas abordagens para descoberta de conhecimento: aprendizado preditivo e aprendizado descritivo (Facelli et al., 2011).

No aprendizado preditivo, o algoritmo de aprendizado é uma função que objetiva construir um estimador dado um conjunto de exemplos rotulados. O rótulo (ou etiqueta) toma valores em um domínio conhecido. Se o domínio dos rótulos, ou seja, o conjunto ao qual os rótulos dos dados pertencem, for um conjunto infinito e ordenado de valores (p. ex., o conjunto dos números reais), o problema é dito de regressão e o estimador é denominado "regressor". Porém, se o domínio dos rótulos é um conjunto finito e não ordenado de valores, o problema é dito de classificação, e o estimador é denominado "classificador". O aprendizado preditivo é também conhecido por aprendizado supervisionado, e é o tipo de tarefa de predição utilizado neste trabalho.

No aprendizado descritivo, as tarefas envolvem a identificação de informações relevantes nos dados sem um elemento externo para guiar o processo de aprendizado. As tarefas descritivas podem ser divididas em: sumarização, cujo objetivo é encontrar uma descrição mais simples e compacta dos dados; associação, cujo objetivo é buscar padrões frequentes de associações entre os atributos de um conjunto de dados; e agrupamento, cujo objetivo é identificar grupos nos dados de acordo com a similaridade entre os objetos. Neste trabalho, foi explorada a tarefa de associação para o conjunto de dados utilizado, para tentar identificar relações entre fatores de acidentes rodoviários.

Aprendizado supervisionado

No problema padrão de aprendizado de máquina supervisionado, a entrada do algoritmo consiste de um conjunto de exemplos S, com N exemplos T_i , $i = 1, \ldots, N$, escolhidos de um domínio X com uma distribuição D fixa, desconhecida e arbitrária, da forma $\{(\mathbf{x}1, y1), \ldots, (\mathbf{x}_N, y_N)\}$ para alguma função desconhecida $y = f(x)$. Os $\mathbf{x}_i$ são tipicamente vetores da forma $(x_{i1}, x_{i2}, \ldots, x_{iM})$ com valores discretos ou numéricos. x_{ij} refere-se ao valor atributo j, denominado X_j, do exemplo T_i. Neste trabalho, também se denominam os atributos X_j como atributos de descrição do domínio. Os valores de y_i referem-se ao valor do atributo Y, frequentemente denominado atributo classe. Os valores de y em problemas de classificação, como é o caso neste trabalho, são tipicamente pertencentes a um conjunto discreto de classes $C = \{C_v\}$, $v = 1, \ldots, N_{Ci}$, ***i.e*** $y \in \{C_1, \ldots, C_{NCl}\}$. O objetivo de um algoritmo de aprendizado supervisionado para problemas de classificação é construir um classificador **h**, que tem como entrada um exemplo **x**, não classificado, ou seja, um vetor de valores de atributos discretos e/ou contínuos, e a sua saída é um valor discreto, ou seja, a classe a ser predita (Domingos, 2012). O classificador **h** é também denominado hipótese da desconhecida e verdadeira função **f**, tal que, para todo $\mathbf{x} \in X$, $\mathbf{f}(\mathbf{x}) = y$ (Mitchell, 1997).

Para avaliar os conhecimentos gerados a partir da base de dados de interesse neste trabalho, utilizando aprendizado supervisionado para problemas de classificação, foram definidos os algoritmos de aprendizado de máquina PART (Witten & Frank, 2009) e J48, que é uma implementação do algoritmo C4.5 (Quinlan, 1993; Witten & Frank, 2009). Ambos os algoritmos oferecem como saída conjuntos de regras facilmente interpretáveis por seres humanos. O objetivo do algoritmo PART é induzir um classificador composto por regras de decisão. Já o J48 tem como finalidade gerar uma árvore de decisão baseada no conjunto de dados de treinamento.

As árvores de decisão possuem um custo computacional baixo, por isso têm sido largamente utilizadas em problemas de classificação. Além disso, são fáceis de entender, fato que aumenta a confiabilidade neste tipo de estrutura. A ideia por trás dos algoritmos de indução de árvore de decisão é decompor a classificação em um conjunto de escolhas sobre cada variável em etapas, iniciando na raiz da árvore e percorrendo as folhas, onde ocorre a classificação. Os diversos algoritmos de árvore de decisão existentes utilizam basicamente o mesmo princípio: a árvore é construída de maneira gulosa, começando pela raiz, escolhendo o atributo com mais informação a cada iteração (Quinlan, 1988). O algoritmo C4.5 (Quinlan, 1993), escrito na linguagem C e usado para gerar árvores de decisão, deu origem ao algoritmo J48, que é uma implementação *open source* em Java do C4.5 para o *software* Weka (Witten & Frank, 2009). O objetivo do J48 é gerar uma árvore de decisão baseada em um conjunto de dados rotulados. O J48 envolve variáveis qualitativas contínuas e discretas presentes na base de dados, por isso a sua expressiva utilização no processo de descoberta de conhecimento e de geração de árvores de decisão. Além disso, é considerado o algoritmo que apresenta o melhor resultado na indução de árvores de decisão, a partir de um conjunto de dados de treinamento. Para induzir uma árvore de decisão, o J48 utiliza a abordagem de dividir-para-conquistar, ou seja, divide um problema complexo em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe (Witten & Frank, 2009). Uma característica

interessante da árvore de decisão está relacionada a cada caminho da árvore gerar uma regra de decisão, e entre as regras não existe intersecção de cobertura de exemplos. Em outras palavras, não existe sobreposição dessas regras no espaço de descrição dos exemplos (Baranauskas & Monard, 2000).

O PART foi desenvolvido tendo como base o algoritmo J48. Ele gera uma lista de decisão e, assim como o J48, também usa a técnica de dividir-para-conquistar. O algoritmo constrói uma árvore de decisão C4.5 parcial a cada iteração e coloca a melhor folha dentro de uma regra (Witten & Frank, 2009). O processo de geração das regras de associação acontece da seguinte maneira: as regras são induzidas a partir de uma árvore e posteriormente são refinadas. Para cada regra criada, é estimada a sua cobertura das instâncias da base de dados. Isso acontece repetidas vezes até que todas as instâncias estejam cobertas. As regras com coberturas mais altas são mantidas e apresentadas para o usuário e as demais são descartadas (Frank & Witten, 1998). A diferença de um algoritmo de indução de regras de decisão em relação a um algoritmo de árvore de decisão reside no fato de que as regras de decisão são induzidas para cobrir um conjunto de exemplos e dessa maneira pode haver sobreposição das regras construídas no espaço de descrição dos exemplos (Baranauskas & Monard, 2000). Dessa maneira, os conceitos aprendidos com esses diferentes algoritmos podem ser bastante distintos, tendo sido utilizados – para fins deste estudo – dois algoritmos de indução de classificadores simbólicos.

Aprendizado de regras de associação

Neste trabalho foi utilizado o algoritmo de construção de regras de associação *Apriori* (Borgelt & Kruse, 2002), cujas regras produzidas associam atributos do domínio de descrição dos exemplos. O algoritmo *Apriori* foi proposto por Agrawal, Imielinski e Swami (1993), e consiste na busca por padrões que indicam o relacionamento entre conjuntos de itens. O Apriori é um dos algoritmos mais utilizados para a descoberta de regras de associação, pois executa diversas leituras na base de dados de transações, sendo capaz de trabalhar com um número grande de atributos. Como resultado, o algoritmo obtém várias alternativas combinatórias entre eles. Ainda assim, devido ao processo de otimização para a geração das regras, o algoritmo consegue ter um bom desempenho em termos de processamento. O *Apriori* também utiliza a técnica de dividir-para-conquistar, com o objetivo de encontrar regras de associação para todas as expressões possíveis.

Seja $A=\{a_1,\ldots,a_q\}$ o universo de q itens. Os itens podem ser produtos, ou valores específicos de atributos, de um conjunto de dados (p. ex., "leite" e "pão" são itens em um domínio de supermercado, ou "idade = jovem", se idade for um dos atributos de descrição do domínio). Um conjunto de itens I é um subconjunto de A, ou seja, $I \subseteq A$. As regras de associação são definidas como: "Se I_i então I_j" ou "I_i => I_j", onde I_i e I_j são conjuntos de itens, $I_i \cap I_j = \emptyset$, I_i é o antecedente da regra, e I_j é o consequente da regra. No *Apriori*, dado um conjunto de transações, ou conjunto de dados, S_{trans}, a busca pelas regras de associação é realizada em duas fases: geração e poda. Na primeira fase, o algoritmo percorre todo o conjunto de dados para gerar todas as combinações de valores possíveis. Em seguida, são mantidas apenas as combinações com uma frequência maior que um valor mínimo pré-determinado, denominado suporte. O suporte de um conjunto de itens I, denominado sup(I), é definido pela Equação 1, onde o símbolo # significa "número de" (Carvalho, Sampaio, & Mongiovi, 1999).

$$\sup(I) = \frac{\#\ transações\ que\ contém\ os\ os\ elementos\ do\ conjunto\ de\ itens\ I}{\#\ total\ de\ transações} \quad (1)$$

Na segunda fase, as regras são construídas a partir dos conjuntos de itens, e é utilizada outra medida para seleção das regras consideradas relevantes – o fator de confiança de uma regra R: $I_i \rightarrow I_j$. A confiança de uma regra *conf*(R) é definida pela Equação 2:

$$\mathrm{conf}(R) = \frac{\#\ transações\ que\ possuem\ I_i\ e\ I_j}{\#\ transações\ que\ possuem\ somente\ I_i} \quad (2)$$

Além de utilizar a confiança como parâmetro de seleção das regras, essa medida também é usada para avaliar a qualidade das regras construídas.

Deve ser observado que, como o *Apriori* espera que cada atributo de descrição do domínio possua itens para serem relacionados, é necessário que o domínio de cada atributo seja discreto, ou seja, possua um número limitado de valores possíveis.

Avaliação dos classificadores

Para avaliar um classificador **h**, inicialmente é necessário coletar informações das decisões tomadas pelo classificador em um conjunto de teste S_{te}, não utilizado na fase de treinamento de **h**. Para isso, é construída uma matriz bidimensional, cujas dimensões são denominadas classe verdadeira e classe predita. A essa matriz dá-se o nome de matriz de confusão, mostrada na Tabela 1. Cada elemento $M(C_i,C_j)$ da matriz, definido pela Equação 3, indica o número de exemplos que pertencem à classe C_i e foram preditos como pertencentes à classe C_j. Nessa equação, $||\mathbf{h}(\mathbf{x}) = C_j||$ é igual a 1 se a igualdade $\mathbf{h}(\mathbf{x}) = C_j$ for verdadeira, ou é igual a 0 se a igualdade for falsa. O número de predições corretas para cada classe são os números apresentados na diagonal principal da matriz de confusão, ou seja, os valores associados a $M(C_i,C_j)$. Todos os outros elementos da matriz $M(C_i,C_j)$, para $i \neq j$, são referentes ao número de erros cometidos em cada classe. Para cada classe *Cv*, $v = 1,\ldots, NCl$, pode-se calcular as taxas de verdadeiros positivos (*TP*, do inglês *True Positives*), verdadeiros negativos (*TN*, do inglês *True Negatives*), falsos positivos (*FP*, do inglês *False Positives*), e falsos negativos (*FN*, do inglês **False Negatives**). A ideia é considerar cada classe C_v como sendo a classe positiva e todas as outras como compondo a classe negativa em relação à C_v. Assim, $TP_{Cv} = M(C_v,C_v)$; $TN_{Cv} = \sum_{\forall Ci \forall C|i \neq v} M(C_i,C_i)$; $FP_{Cv} = \sum_{\forall Ci \forall C|i \neq v} M(C_v,C_i)$; e $FN_{Cv} = \sum_{\forall Ci \forall C|i \neq v} M(C_i,C_v)$. Neste trabalho, utiliza-se a taxa de erro de **h** – *err*(**h**) –, que é a soma de todos os valores $M(C_i,C_j)$ tais que $i \neq j$, bem como outras medidas para avaliar o comportamento do classificador nas classes. Tais medidas são precisão – *Prec*(**h**), definida pela Equação 6; sensibilidade, ou *recall* – *Rec*(**h**), definida pela Equação 7; F – *F*(**h**), definida pela Equação 8; e área sob a curva ROC – *AUC*(**h**). A curva ROC é um gráfico que trata a relação entre as taxas de TP e FP, pois o ideal é que TP seja 1, e FP seja 0, Assim, quanto mais próximo um classificador possua o par (*TP*, *FP*) ao ponto (1,0), melhor o classificador. Deve ser observado que classificadores cujos pares (*TP*, *FP*) tais que *TP* = *FP* são modelos considerados aleatórios. Daí, a área sob a curva ROC (*AUC*, do inglês *Area Under Curve*) pode ser calculada – quanto mais próximo de 1 o valor da *AUC*, melhor o modelo construído. A média das *AUC*s, calculada sobre a *AUC* para cada classe C_v, pode ser então calculada.

Classe Verdadeira	Predita C_1	Predita C_2	...	Predita C_{NCl}
C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	...	$M(C_1, C_{NCl})$
C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	...	$M(C_2, C_{NCl})$
...	...	...	...	...
C_{NCl}	$M(C_{NCl}, C_1)$	$M(C_{NCl}, C_2)$	...	$M(C_{NCl}, C_{NCl})$

Tabela 1. Matriz de Confusão.
Fonte: autoria própria.

$$M(C_i, C_j) = \sum_{\forall (\mathbf{x},y) \in S_{te} | y = C_i} ||\mathbf{h}(\mathbf{x}) = C_j|| \quad (3)$$

$$Prec(\mathbf{h}) = \frac{\sum_{v=1}^{NCl} TP_{C_v}}{\sum_{v=1}^{NCl} TP_{C_v} + FP_{C_v}} \quad (4)$$

$$Rec(\mathbf{h}) = \frac{\sum_{v=1}^{NCl} TP_{C_v}}{\sum_{v=1}^{NCl} TP_{C_v} + FN_{C_v}} \quad (5)$$

$$F(\mathbf{h}) = \frac{2 \times Prec(\mathbf{h}) \times Rec(\mathbf{h})}{Prec(\mathbf{h}) + Rec(\mathbf{h})} \quad (6)$$

A fim de estimar as medidas previamente descritas, existem diversas técnicas para construir o conjunto de treinamento e teste. É comum utilizar a técnica de validação cruzada para estimar a taxa de erro de um classificador, bem como as outras medidas. Na explicação a seguir, detalha-se a estimativa da medida *err*(**h**); de maneira análoga, podem ser estimadas as outras medidas. Na técnica de validação cruzada com *K* partições, o conjunto de dados *S* é dividido aleatoriamente em *K* partições $S_1,\ldots,S_K$, disjuntas, sendo todas as partições de conjuntos de dados de aproximadamente o mesmo tamanho. Após, são executadas *K* iterações de indução e teste de um classificador. Na primeira iteração, é induzido o classificador $\mathbf{h}_1$ utilizando os conjuntos de dados $S_2,\ldots,S_K$. Daí, $\mathbf{h}_1$ é testado com o conjunto S_1, obtendo assim a taxa de erro $err(\mathbf{h}_1)$. Na segunda iteração, é induzido o classificador $\mathbf{h}_2$ utilizando os conjuntos de dados S_1 e $S_3,\ldots,S_K$. Daí, $\mathbf{h}_2$ é testado com o conjunto S_2, obtendo assim a taxa de erro $err(\mathbf{h}_2)$, e assim sucessivamente. Então, a média $m_{err}(\mathbf{h})$ e o erro padrão $se_{err}(\mathbf{h})$ dessas taxas de erro são calculados, definidos respectivamente pelas Equações 7 e 8, do modelo final. Esse

modelo final é construído utilizando todos os exemplos disponíveis. Estimados a média e o erro padrão da taxa de erro do classificador **h**, pode ser utilizado o teste t de *Student* para comparar o poder de predição dos dois algoritmos de aprendizado de máquina para um mesmo conjunto de dados (Baranauskas & Monard, 2000).

$$m_{Err}(\mathbf{h}) = \frac{1}{K}\sum_{k=1}^{K} err(\mathbf{h}_k) \tag{7}$$

$$se_{Err}(\mathbf{h}) = \sqrt{\frac{1}{K-1} \times \frac{1}{K}\sum_{k=1}^{K} (err(\mathbf{h}_k) - m_{Err}(\mathbf{h}))^2} \tag{8}$$

No Weka se implementa a validação cruzada estratificada com *K* partições. A diferença está na maneira em que é feita a divisão do conjunto de dados original em *K* partições. Neste tipo de técnica com *K* partições, as partições são feitas de modo que seja respeitada a distribuição dos exemplos nas classes. Ou seja, se no conjunto de dados original existem 20% dos exemplos na classe C_1 e 80% dos exemplos na classe C_2, então em cada partição construída com a técnica de validação cruzada estratificada com *K* partições, existem aproximadamente 20% dos exemplos pertencentes à classe C_1 e 80% dos exemplos pertencentes à classe C_2.

Pré-processamento de dados

Um conjunto de dados pode conter diversos tipos de ruídos e/ou imperfeições, como valores incorretos, inconsistentes, duplicados ou ausentes. Frequentemente são utilizadas técnicas de pré-processamento de dados para melhorar a qualidade dos mesmos, essas técnicas podem ser de eliminação ou minimização dos problemas citados. Dados processados – onde estão presentes apenas atributos relevantes para o domínio – levam à indução de conceitos mais precisos e mais enxutos, o que também implica em uma maior facilidade na interpretação dos padrões extraídos. Técnicas de pré-processamento são úteis também para tornar os dados mais adequados para um determinado algoritmo, como por exemplo, a substituição do domínio de um atributo contínuo por um domínio discreto, tarefa necessária quando se utiliza o algoritmo *Apriori* (Facelli et al., 2011).

Nem todos os atributos do conjunto de dados original são necessários para determinada tarefa de aprendizado de máquina como, por exemplo, um atributo que possua o mesmo valor para todas as instâncias. Quando um atributo não contribui para a estimativa do valor do atributo classe, ele é considerado irrelevante. (Facelli et al., 2011)

Na seção a seguir é descrita a base de dados de interesse para o desenvolvimento deste trabalho.

Base de dados: boletins de ocorrências em rodovias federais brasileiras da Polícia Rodoviária Federal

Os boletins de ocorrências em rodovias federais brasileiras, disponíveis no Portal Brasileiro De Dados Abertos (Brasil, 2014b), são caracterizados como Dados Abertos Governamentais (DAG) ou dados públicos, pois são disponibilizados na internet para livre utilização pela sociedade (Agune, Gregorio Filho, & Bolliger, 2010). A comunidade envolvida com os DAG afirma que, para que os dados sejam definidos como tal, eles devem seguir oito princípios listados a seguir (The Annotated 8 principles of Open Government Data, 2014):

a) Completos. Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo – mas não se limitando a – documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, regulados por estatutos;
b) Primários. Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada;
c) Atuais. Os dados são disponibilizados o quão rapidamente seja necessário para preservar seu valor;
d) Acessíveis. Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis;
e) Processáveis por máquina. Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado;
f) Acesso não discriminatório. Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro;

g) Formatos não proprietários. Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo;
h) Livres de licenças. Os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

Deve ser observado que o segundo princípio determina que os dados abertos devam ser publicados como coletados na fonte, princípio este que pode ser percebido nos boletins de ocorrências da PRF. Entretanto, observa-se que esse princípio dificultou a etapa de pré-processamento para o processo de Mineração de Dados, especialmente devido: a) ao volume significativo de dados; b) a uma grande quantidade de dados faltantes; e c) aos diversos problemas de dados errôneos encontrados (corrigidos manualmente quando possível). Vários desses problemas são descritos adiante.

Os dados analisados neste trabalho foram obtidos no referido portal de dados abertos e fazem parte da base de dados da Polícia Rodoviária Federal. Primeiramente, foi analisada qual parcela de dados seria considerada útil para analisar as ocorrências que foram registradas durante todo o ano de 2012, pois este era o ano mais recente e com os dados dos dois semestres publicados. Infelizmente, os dados do segundo semestre de 2013 ainda não tinham sido publicados, o que reforçou a opção por se utilizarem os dados de 2012. Pode-se observar que este fato fere o terceiro princípio dos DAG, que diz respeito à atualidade dos dados. Nessa porção da base de dados foi identificado que algumas tabelas e diversos campos possuem informações consideradas irrelevantes e/ou desnecessárias para o processo de mineração de dados, incluindo a aquisição de novos conhecimentos. Um exemplo são os atributos: (i) *ocotipo*, da tabela '*OCORRENCIA*', que registra o tipo da ocorrência. Neste, todos os registros contêm o valor "1" (que significa acidentes rodoviários). Ao se considerar que esse atributo apresenta um dado redundante, procedeu-se a retirada; (ii) *ocostatus*, também presente na tabela de ocorrências, que registra o *status* da ocorrência. Todos os registros contêm o valor "S", que significa que a ocorrência já foi encerrada, e por isso o atributo também foi removido; (iii) o campo *oacdanoterc*, presente na tabela '*OCORRENCIAACIDENTE*', registra se aconteceu dano a terceiro. A maioria dos registros (384211 de 390974 registros totais) contém o valor 'N', e, por esse motivo, esse campo também foi removido; e (iv) na tabela '*PESSOA*', os campos *pesteccodigo*, que identifica o estado civil do envolvido em uma ocorrência, e o campo *pestgicodigo*, que identifica o grau de instrução da pessoa, foram desconsiderados, pois ambos eram campos do tipo chave estrangeira para outra tabela, que não estão disponíveis no portal de dados abertos. Questionou-se, ainda, a utilidade de alguns dados, devido à significativa quantidade de dados faltantes em alguns atributos. O Diagrama de Entidade e Relacionamento (DER), que pode ser visualizado na Figura 1, ilustra as entidades utilizadas neste trabalho para a extração dos dados[1]. Uma descrição detalhada deste DER pode ser encontrada em Anexo.

Dentre os problemas identificados, durante o pré-processamento dos dados, é importante destacar:

a) alguns atributos estão presentes na base de dados, porém nem todos são úteis para a descoberta de conhecimento e para o próprio sistema BR-Brasil, pois foram descontinuados nas alterações de versões do sistema, ou seja, alguns campos utilizados em determinada versão do sistema foram retirados em outras;
b) significativa quantidade de dados faltantes – nos 390.973 registros totais, foram encontradas 181.428 ocorrências de dados faltantes, que tiveram que ser substituídos por '?';
c) cidades e códigos de rodovias federais inexistentes e/ou repetidos;
d) dicionário de dados incompleto e de difícil compreensão, pois alguns atributos não são descritos e, por outro lado, outros, como o atributo *oacgirofundo* (tabela de ocorrências dos acidentes), não possuem uma identificação de sua finalidade;
e) algumas tabelas, como a que identifica o modelo de pista, não estão no conjunto de dados publicados;
f) DER incompleto e desatualizado;
g) alguns campos possuem diferentes opções de escolha, mas em todos os registros somente um valor é escolhido, como acontece no atributo que registra o tipo de envolvido no acidente: dentre 16 opções, apenas duas – passageiro e condutor – são utilizadas;

[1] O DER completo de toda a base da PRF pode ser visualizado em http://migre.me/iehd4.

h) o campo que demonstra o estado físico do acidentado não agrega valor ao processo de descoberta de conhecimento, pois não é preenchido sempre, e quando o é, apresenta erros de digitação; além disto, cada pessoa envolvida em um acidente pode ter um tipo de estado físico diferente;
i) na tabela '*TIPOVEICULO*' encontram-se valores repetidos e/ou similares, como por exemplo, "reboque", "semi-reboque e reboque" ou "semi-reboque";
j) pelos valores observados no atributo município, acredita-se que o preenchimento do nome do município onde ocorreu o acidente (bem como da BR) são campos de texto, ou seja, o usuário pode digitar sem uma padronização (valores predeterminados), o que implica em vários erros de digitação;
k) campos que poderiam agregar valor ao processo de MD, tais como o registro do estado da pista na hora do acidente, ou a ocorrência de danos ao ambiente, não são preenchidos na maioria das vezes;
l) a separação das ocorrências por semestre é feita com base na data da finalização da mesma, o que implica que ocorrências de 2001, por exemplo, podem ser encontradas nos dados de 2012, pois sua finalização ocorre nesse ano; e
m) falta de padronização nas entradas dos dados, tais como valores de textos em campos numéricos e vice-versa, nomes de cidades grafados erroneamente e/ou digitados de maneiras distintas, códigos de rodovias federais inexistentes, dentre outros.

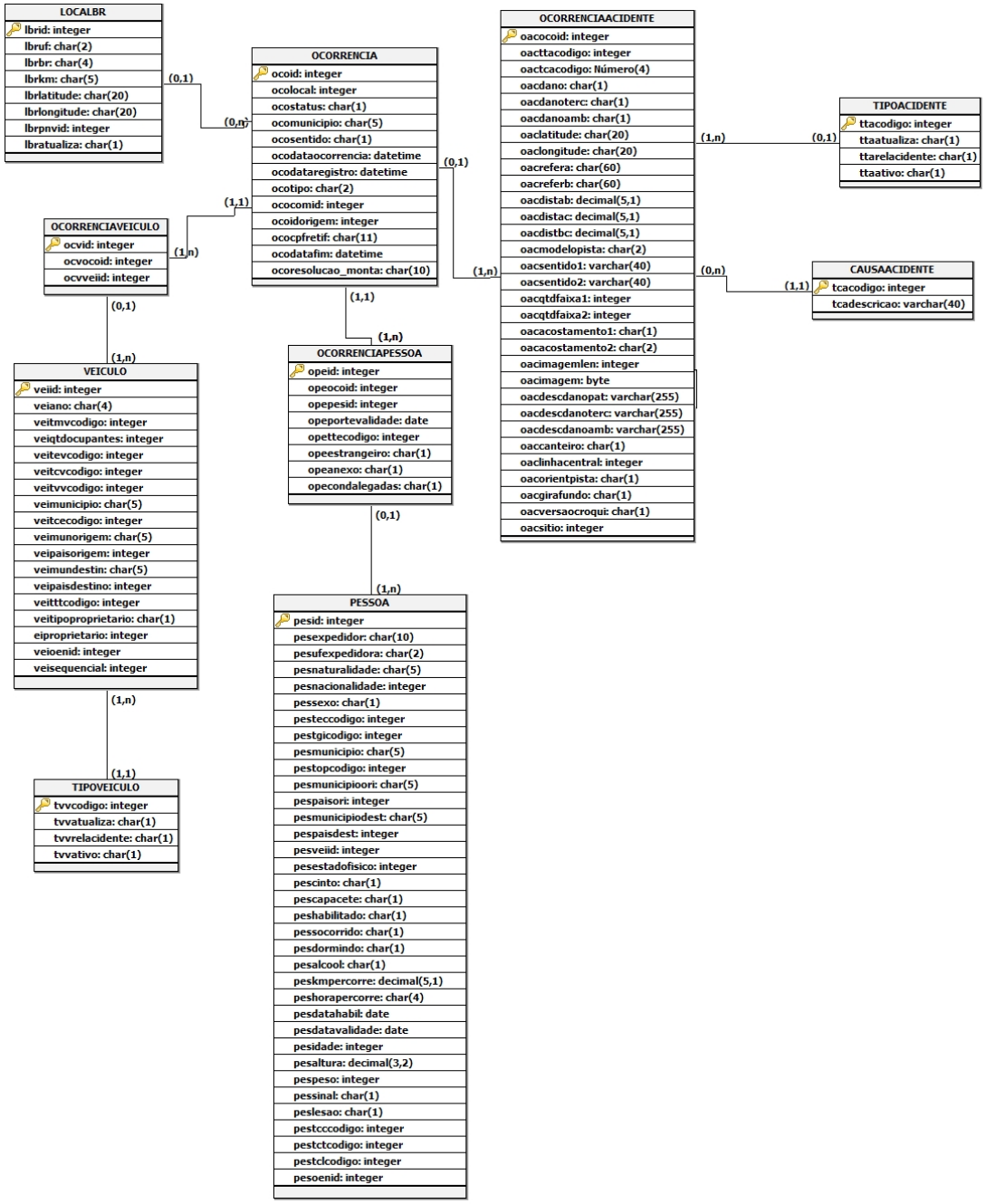

Figura 1. Diagrama de entidade e relacionamento da base de dados utilizada.
Fonte: autoria própria.

Descrição do estudo: aplicação de ferramentas de mineração de dados nos dados de boletins de ocorrências da PRF em 2012

Para esse estudo foram utilizados os dados do ano de 2012, presentes na base de dados de ocorrências em rodovias federais, da PRF. Em seguida, realizou-se a limpeza desses dados, principal tarefa da etapa de pré-processamento. Utilizou-se o *software* Weka (Witten & Frank, 2009) que requer que os dados estejam em formato atributo-valor. Efetivaram-se, ainda, outras pequenas transformações, devido a restrições da ferramenta e do uso do algoritmo *Apriori* (esse, por sua vez, exige que todos os atributos de descrição do domínio possuam um domínio discreto). Realizaram-se as seguintes substituições:

a) os atributos do tipo data foram alterados para dia da semana;
b) o horário do acidente foi transformado para o período do dia (manhã, tarde, noite ou madrugada) referente ao acidente;
c) a data de fabricação do veículo serviu para categorizarmos o veículo em novo, seminovo, usado ou com mais de 10 anos de uso;
d) a data de nascimento foi substituída pela faixa etária da pessoa (criança, adolescente, jovem, adulto ou idoso);
e) a data de vencimento da Carteira Nacional de Habilitação (CNH) foi substituída por um campo que informa se o motorista estava com a habilitação vencida ou não;
f) os dados faltantes foram substituídos por '?'.

Na Tabela 2 são exibidas as características da base de dados processada, na qual # At. Discretos e # At. Contínuos são, respectivamente, o número de atributos de descrição do domínio que são discretos e contínuos; # Exs. é o número total de exemplos presentes na base de dados utilizada; # Classes (*NCl*) é o número de classes presentes na base de dados; e Distribuição dos Exs. nas Classes apresenta as classes presentes na base de dados, o número de exemplos presente em cada classe (# Exs.) e o percentual de exemplos em cada classe (% Exs.). Deve ser observado que na coluna # Exs. existem dois valores. O primeiro valor é referente a todos os exemplos presentes no conjunto de dados. No entanto, dos 390.973 exemplos, somente 294.480 são rotulados com a classe "Causa do Acidente". Portanto, somente esses exemplos foram utilizados nos algoritmos de predição (J48 e PART), e considerados para computar a distribuição dos exemplos nas classes, exibida na respectiva coluna da Quadro 1.

# At. Discretos	# At. Contínuos	# Exs.	# Classes (*NCl*)	Distribuição dos Exs. nas Classes		
				Classe	# Exs.	% Exs.
				Animais na Pista	7211	2,4%
				Defeito Mecânico no Veículo	13088	4,4%
				Defeito na Via	4346	1,5%
				Desobediência à Sinalização	19266	6,5%
8	0	390.973 / 294.480	10	Dormindo	8638	2,9%
				Falta de Atenção	136342	46,3%
				Ingestão de Álcool	16007	5,4%
				Não Guardar Distância de Segurança	47558	16,1%
				Ultrapassagem Indevida	11904	4,0%
				Velocidade Incompatível	30120	10,2%

Quadro 1. Características da base de dados processada – ano de 2012.
Fonte: autoria própria.

No Quadro 2, é apresentada uma descrição do conteúdo de cada atributo e, no Quadro 3, exibidos os atributos e os valores possíveis nos atributos.

Atributo	Conteúdo
Tipo de Veículo	Tipo de veículo envolvido na ocorrência, como por exemplo, automóvel, motocicleta, etc.
Ano do Veículo	Categorização do veículo de acordo com o seu ano de fabricação. Na base original esse valor é numérico e representa o ano de fabricação do veículo, mas para que seja possível extrair conhecimento, esses números foram transformados em variáveis que agregam valor. A categorização foi feita da seguinte maneira: veículos fabricados há menos de três anos, foram considerados seminovos; entre três e 10 anos, veículos usados e o restante foi classificado como veículos com mais de 10 anos de fabricação.
Estado Físico da Pessoa	Estado físico em que a pessoa se encontrava quando os agentes da PRF chegavam ao local do acidente.
Faixa Etária da Pessoa	Idade das pessoas envolvidas na ocorrência. Na base original, esse valor pode ser encontrado através da data de nascimento do envolvido, a partir dessa data calculou-se a idade. Em seguida, categorizou-se a idade em faixas etárias: entre 0 e 12 anos, atribuiu-se o valor 'criança'; entre 13 e 17 anos, 'adolescente'; entre 18 e 25, 'jovem; entre 26 e 59 anos, 'adulto'; pessoas com mais de 60 anos, 'idoso';
Tipo de Acidente	Tipo de acidente da ocorrência como, por exemplo, "atropelamento de pessoa".
Modelo da Pista	Modelo da pista do local do acidente. Por exemplo, se o acidente foi em uma reta ou em uma curva acentuada.
Período do Dia	Na base de dados original é disponibilizada a hora em que o acidente ocorreu, assim como outros campos descritos acima. Esse valor numérico não agrega valor aos algoritmos que foram utilizados no processo de descoberta de conhecimento. Assim, categorizou-se a hora do acidente da maneira como segue: ocorrências que aconteceram entre 6h e 11h receberam o valor 'manhã'; entre 12h e 18h, 'tarde'; entre 19h e 23h, 'noite' e as que ocorreram entre 0h e 5h, foram substituídas pelo valor 'madrugada'.
Dia da Semana	Nos boletins de ocorrências, o dia da semana do fato não é registrado, mas sim a data (dia, mês e ano) da ocorrência. Com essa informação, foi possível descobrir o dia da semana em que ocorreu o acidente.

Quadro 2. Atributos de descrição e uma descrição sobre seu conteúdo.
Fonte: autoria própria.

Atributo	Domínio do Atributo
Tipo de Veículo	Bicicleta, Ciclomotor, Motoneta, Motocicleta, Triciclo, Quadriciclo, Automóvel, Micro-ônibus, Ônibus, Bonde / Trem, Reboque, Semireboque, Charrete, Caminhão, Carroça, Carro de Mão, Caminhonete, Utilitário, Caminhão Trator, Trator de Rodas, Trator de Esteiras, Trator Misto, Camioneta, Caminhão Tanque, Não Identificado.
Ano do Veículo	Seminovo, Usado, Mais de 10 Anos.
Estado Físico da Pessoa	Ileso, Lesões Leves, Lesões Graves, Morto, Ignorado.
Faixa Etária da Pessoa	Criança, Adolescente, Jovem, Adulto, Idoso.
Tipo de Acidente	Atropelamento de Animal, Atropelamento de Pessoa, Capotamento, Colisão com Bicicleta, Colisão com Objeto Fixo, Colisão Frontal, Colisão Lateral, Colisão Traseira, Incêndio, Colisão Transversal, Tombamento, Saída de Pista, Derramamento de Carga, Colisão com Objeto Móvel, Queda de Motocicleta / Bicicleta / Veículo, Danos Eventuais.
Modelo da Pista	Reta, Curva Aberta, Ponte, Bifurcação, Bifurcação com Rotatória, Curva Acentuada, Curva Acentuada a Direita, Curva Diamante, Curva 180º, Sinuosa, Cruzamento, Cruzamento com Rotatória, Cruzamento com Viaduto, Cruzamento com Canteiro, Retorno 1, Retorno 2, Início / Fim de Pista Dupla, Vicinal, Vicinal Dupla, Saída, Cruzamento com viaduto 2.
Período do Dia	Manhã, Tarde, Noite, Madrugada.
Dia da Semana	Domingo, Segunda-feira, Terça-feira, Quarta-feira, Quinta-feira, Sexta-feira, Sábado.

Quadro 3. Atributos de descrição e seus respectivos domínios da base de dados processada – ano de 2012.
Fonte: autoria própria.

Com os dados pré-processados e transformados, realizou-se a etapa de extração de padrões. Foram utilizados os algoritmos PART (indução de regras de conhecimento), J48 (árvores de decisão) e *Apriori* (construção de regras de associação), descritos anteriormente[2]. Optou-se pela utilização de algoritmos de aprendizado simbólico, já que os classificadores induzidos por tais algoritmos podem ser transformados em conjuntos de regras proposicionais R = B → H, que são mais facilmente interpretadas por seres humanos (Bernardini, 2006).

Resultados e análise – algoritmos PART e J48

Na Tabela 2 são exibidos os resultados obtidos com os algoritmos J48 e PART para cada uma das medidas $mea(\mathbf{h}) \in \{err(\mathbf{h}), prec(\mathbf{h}), rec(\mathbf{h}), F(\mathbf{h}), AUC(\mathbf{h})\}$. Deve ser observado que todas as medidas foram estimadas utilizando a técnica de validação cruzada estratificada com 10 partições, descrita anteriormente. Como a classe majoritária conjunto de dados é "Falta de Atenção", com 46,3% dos exemplos, o erro majoritário da base de

[2] Deve ser observado que o Erro Majoritário de uma base de dados é o limite superior que a taxa de erro de um classificador deve atingir. Se a taxa de erro de um classificador for maior que o erro majoritário, o classificador é menos eficiente que predizer, para os exemplos futuros, a classe majoritária.

dados utilizada é de 53,7%, e a média das taxas de erro obtidas para os algoritmos J48 e PART são menores ou iguais a esse valor – $m_{err}(\mathbf{h}_{J48}) = 45{,}88\%$ e $m_{err}(\mathbf{h}_{J48}) = 46{,}36\%$ –, tais fatos indicam que os classificadores obtidos são melhores preditores do que classificadores dos exemplos na classe majoritária.

Na Tabela 3, foram marcados com o símbolo ▲ os casos em que o algoritmo apresenta melhor resultado em relação ao outro, com 95% de confiança segundo o teste t de *Student*, e com o símbolo ○ os casos em que ambos os algoritmos não apresentaram diferença estatística, também segundo o teste t de *Student*.

Algoritmo		$err(\boldsymbol{h})$		$prec(\boldsymbol{h})$		$rec(\boldsymbol{h})$		$F(\boldsymbol{h})$		$AUC(\boldsymbol{h})$	
J48	$m_{mea}(\mathbf{h}_{J48})$	45,88%	▲	46,2%	▲	8,9%		15%		83,4%	○
	$se_{mea}(\mathbf{h}_{J48})$	0,04%		1,0%		0,2%		0,4%		0,2%	
PART	$m_{mea}(\boldsymbol{h}_{PART})$	46,36%		42,2%		14,7%	▲	21,8%	▲	83,3%	○
	$se_{mea}(\boldsymbol{h}_{PART})$	0,07%		0,7%		0,3%		0,4%		0,2%	

Tabela 2. Resultados dos algoritmos J48 e PART.
Fonte: autoria própria.

Pode-se observar, nessa tabela, que o algoritmo J48 apresenta melhor comportamento para a base de dados utilizada para as medidas *err*(**h**) e *prec*(**h**). Já o algoritmo PART apresenta melhor comportamento em relação ao J48 para as medidas *rec*(**h**) e *F*(**h**). Ainda assim, para ambos os casos, observa-se que os valores das medidas *rec*(**h**) e *F*(**h**) são baixos, o que indica uma melhor análise em relação às predições nas classes. Sendo assim, observam-se também as medidas em cada uma das classes. Nas Tabelas 3 e 4 são mostradas a taxa de *TP* e de *FP* e as medidas $prec(C_v)$, $rec(C_v)$, $F(C_v)$ e $AUC(C_v)$ para os algoritmos J48 e PART, respectivamente, e para cada uma das classes $C_v \in C$, gerou-se um gráfico com esses dados, que podem ser visualizados nas Figuras 2 e 3.

Classe	Taxa de *TP*	Taxa de *FP*	$prec(C_v)$	$rec(C_v)$	$F(C_v)$	$AUC(C_v)$	% Ex.
Animais na Pista	0,758	0,001	0,943	0,758	0,84	0,924	2,4%
Defeito Mecânico no Veículo	0,34	0,014	0,536	0,34	0,416	0,777	4,4%
Defeito na Via	0,074	0,002	0,327	0,074	0,121	0,725	1,5%
Desobediência à Sinalização	0,089	0,007	0,462	0,089	0,15	0,834	6,5%
Falta de Atenção	0,918	0,67	0,542	0,918	0,681	0,675	2,9%
Ingestão de Álcool	0,107	0,012	0,346	0,107	0,163	0,675	46,3%
Motorista Dormindo	0,22	0,008	0,462	0,22	0,298	0,772	5,4%
Não Guardar Distância de Segurança	0,019	0,005	0,425	0,019	0,036	0,824	16,1%
Ultrapassagem Indevida	0,242	0,012	0,455	0,242	0,316	0,805	4,0%
Velocidade Incompatível	0,496	0,046	0,55	0,496	0,522	0,778	10,2%

Tabela 3. Resultados do algoritmo J48 em cada classe.
Fonte: autoria própria.

Classe	Taxa de *TP*	Taxa de *FP*	$prec(C_v)$	$rec(C_v)$	$F(C_v)$	$AUC(C_v)$	% Ex.
Animais na Pista	0,761	0,002	0,925	0,761	0,835	0,925	2,4%
Defeito Mecânico no Veículo	0,35	0,016	0,503	0,35	0,413	0,8	4,,4%
Defeito na Via	0,092	0,003	0,299	0,092	0,141	0,742	1,5%
Desobediência à Sinalização	0,147	0,014	0,421	0,147	0,218	0,833	6,5%
Falta de Atenção	0,839	0,576	0,557	0,839	0,669	0,688	2,9%
Ingestão de Álcool	0,166	0,019	0,337	0,166	0,223	0,772	46,3%
Motorista Dormindo	0,255	0,011	0,417	0,255	0,317	0,811	5,4%
Não Guardar Distância de Segurança	0,165	0,046	0,41	0,165	0,235	0,839	16,1%
Ultrapassagem Indevida	0,258	0,014	0,438	0,258	0,325	0,809	4,0%
Velocidade Incompatível	0,482	0,046	0,544	0,482	0,511	0,796	10,2%

Tabela 4. Resultados do algoritmo PART em cada classe.
Fonte: autoria própria.

Pode-se observar, nas Tabelas 3 e 4, e nos gráficos das Figuras 2 e 3, que, para as medidas Taxa de *FP*, $prec(C_v)$, $rec(C_v)$ e $F(C_v)$, somente a classe "Animais na Pista" apresentou valores altos, diferente das outras classes, que apresentaram valores baixos. Observa-se também que todas as classes apresentaram valores acima de 0,5 na medida de $AUC(C_v)$, indicando que houve aprendizado em cada uma das classes, apesar dos valores estarem abaixo de 0,8 em alguns casos. Ainda, apesar de "Defeito Mecânico no Veículo" ter somente 4,4% dos exemplos da base de dados, as taxas apresentadas para essa classe são promissoras, já que as medidas de *prec* e *AUC* apresentam valores maiores que 0,5. Por outro lado, a classe "Ingestão de Álcool" possui 46,3% dos exemplos da base de dados, mas ainda assim possui baixa taxa de *TP*, *prec*, *rec* e *F*, o que indica que, para essa base, a distribuição dos dados não tem impacto direto nos resultados obtidos. Adiante, ainda nesta seção, é apresentado um estudo mais aprofundado em relação ao aprendizado de classificadores simbólicos para cada uma das classes.

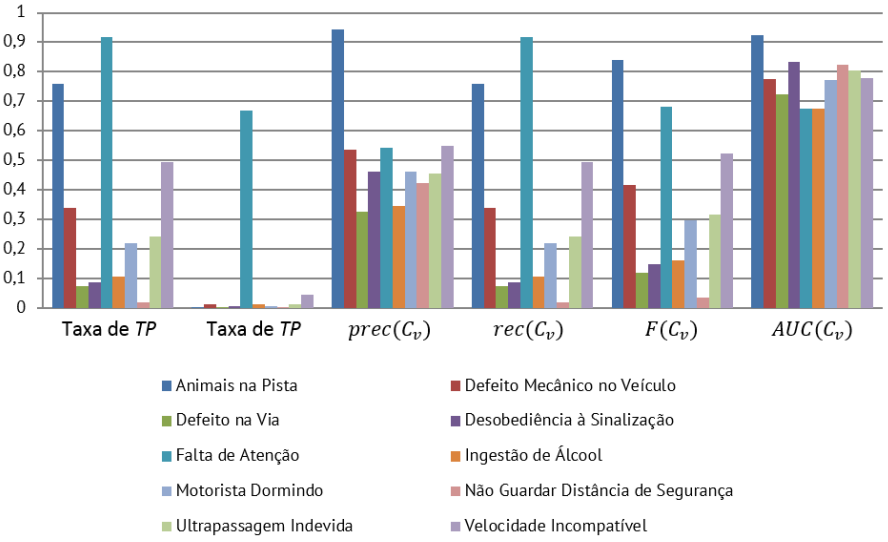

Figura 2. Valor das medidas por classe - algoritmo J48.
Fonte: autoria própria.

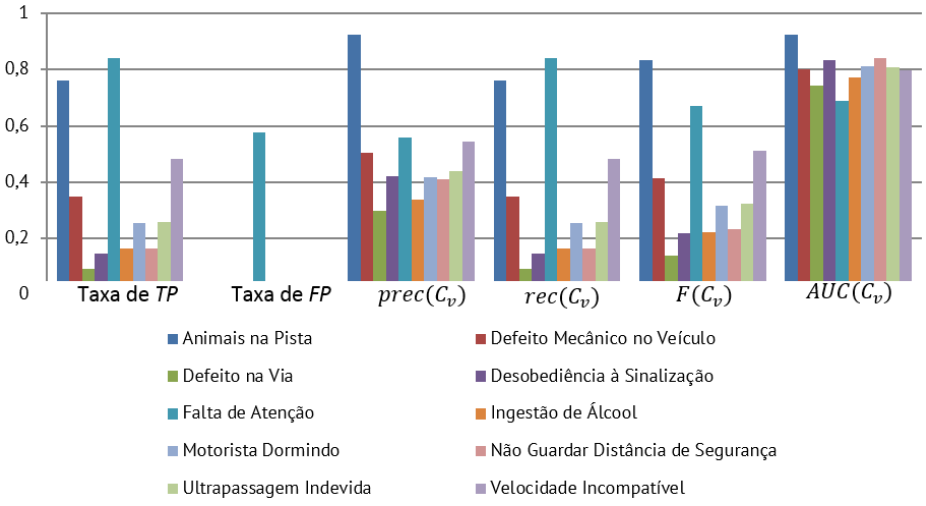

Figura 3. Valor das medidas por classe – algoritmo PART.
Fonte: autoria própria.

Pode-se observar também que, em relação ao número de regras criadas por ambos os algoritmos[3], o PART construiu um classificador com todos os exemplos de treinamento, que possui um total de 12.600 regras de decisão. Já o J48 construiu uma árvore de decisão com 14.311 nós folhas, ou seja, 14.300 regras de decisão.

Dentre as regras geradas pelo algoritmo J48, são mostradas, nos Quadros 4 e 5, as que foram julgadas mais interessantes. O critério de escolha foi considerar regras que traziam novos conhecimentos a respeito dos acidentes rodoviários e que, do ponto de vista dos autores, poderiam ser utilizadas pelas autoridades, e pela própria população, para a diminuição de acidente em rodovias federais. Entretanto, deve ser observado que

[3] Cada caminho do nó raiz até um nó de decisão, ou nó folha, de uma árvore de decisão pode ser reescrito como uma regra de decisão.

não houve uma validação junto aos órgãos competentes quanto a essa validade. Nos resultados do PART, foram selecionadas algumas regras que possuem alto valor de precisão e cobertura da regra, que são listadas na Figura 6. Nessas figuras, #Cob é o número de casos cobertos corretamente pela regra e #Incorr é o número de exemplos incorretamente cobertos pela regra.

- SE Tipo de Acidente = Capotamento E Modelo da Pista = Reta e Tipo de Veículo = Automóvel E Período do Dia = Manhã E Dia da Semana = Domingo E Estado Físico da Pessoa = Lesões Graves E Ano do Veículo = Usado ENTÃO Causa do Acidente = Ingestão de Álcool. (#Cob = 3,52; #Incorr = 2,52)
- SE Tipo de Acidente = Colisão com Objeto Fixo E Modelo da Pista = Reta E Tipo de Veículo = Motocicleta E Estado Físico da Pessoa = Lesões Graves E Período do Dia = Noite E Dia da Semana = Terça-feira ENTÃO Causa do Acidente = Falta de Atenção. (#Cob = 16; #Incorr = 1)
- SE Tipo de Acidente = Colisão com Objeto Fixo E Modelo da Pista = Reta E Tipo de Veículo = Automóvel E Estado Físico da Pessoa = Lesões Graves E Período do dia = Madrugada E Ano do Veículo = Mais de 10 Anos ENTÃO Causa do Acidente = Ingestão de Álcool. (#Cob = 8,94; #Incorr = 2,94)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Tipo de Veículo = Motocicleta ENTÃO Causa do Acidente = Falta de Atenção. (#Cob = 16; #Incorr = 7)

Quadro 4. Regras selecionadas do classificador induzido pelo algoritmo J48 (PARTE 1).
Fonte: autoria própria.

- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Período do Dia = Manhã E Dia da Semana = Segunda-feira E Tipo de Veículo = Automóvel E Estado Físico da Pessoa = Morto ENTÃO Causa do Acidente = Ultrapassagem Indevida. (#Cob = 19.25; #Incorr = 11)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Período do Dia = Noite E Dia da Semana = Sábado E Tipo de Veículo = Automóvel E Estado Físico da Pessoa = Morto E Ano do Veículo = Usado ENTÃO Causa do Acidente = Ultrapassagem Indevida. (#Cob = 16,13; #Incorr = 9,13)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Período do Dia = Noite E Dia da Semana = Quinta-feira E Tipo de Veículo = Micro-ônibus ENTÃO Causa do Acidente = Ultrapassagem Indevida. (#Cob = 27; #Incorr = 4)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Curva 180º ENTÃO Causa do Acidente = Velocidade Incompatível. (#Cob = 287,87; #Incorr = 75,04)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Cruzamento E Tipo de Veículo = Motocicleta ENTÃO Causa do Acidente = Defeito Mecânico no Veículo. (#Cob = 13,11; #Incorr = 4,04)

Quadro 5. Regras selecionadas do classificador induzido pelo algoritmo J48 (PARTE 2).
Fonte: autoria própria.

Algumas regras obtidas trazem informações interessantes como, por exemplo a terceira regra do Quadro 6, que revela que acidentes com incêndio, ocorridos na quarta-feira de manhã, aconteceram por que o motorista dormiu ao volante. A quarta regra apresentada também chama a atenção para um fato muito comum em acidentes rodoviários: a falta de atenção. De acordo com os resultados da última regra, a maioria dos atropelamentos de pessoas em retas é causada por falta de atenção do motorista.

- SE Tipo de Acidente = Atropelamento de Animal E Modelo da Pista = Reta E Tipo de Veículo = Automóvel E Período do Dia = Noite ENTÃO Causa do Acidente = Animais na Pista. (#Cob = 1036; #Incorr = 17)
- SE Tipo de Acidente = Incêndio E Estado Físico da Pessoa = Ileso E Modelo da Pista = Reta ENTÃO Causa do Acidente = Defeito Mecânico no Veículo. (#Cob = 628; #Incorr = 7)
- SE Tipo de Acidente = Incêndio E Dia da Semana = Quarta-feira E Período do Dia = Manhã ENTÃO Causa do Acidente = Motorista Dormindo. (#Cob = 16)
- SE Tipo de Acidente = Atropelamento de Pessoa E Modelo da Pista = Reta E Período do Dia = Tarde e Tipo de Veículo = Automóvel ENTÃO Causa do Acidente = Falta de Atenção. (#Cob = 343; #Incorr = 85)

Quadro 6. Regras selecionadas da árvore de decisão induzida pelo algoritmo PART.
Fonte: autoria própria.

Devido ao significativo número de classes, foi avaliado também o comportamento dos algoritmos PART e J48 para cada uma das classes $C_v \in C$, em cada uma das medidas previamente mencionadas. Analogamente aos experimentos anteriores, utilizamos a técnica de validação cruzada estratificada com 10 partições. Foi gerado um conjunto de dados para cada classe $C_v \in C$, no qual os exemplos $\mathbf{x} \in C_v$ são rotulados como positivos, e os exemplos $\mathbf{x} \notin C_v$ são rotulados como negativos. Essa abordagem está relacionada à técnica de divisão de problemas de classificação denominada "um-contra-todos" (Facelli et al., 2011). Nas Tabelas 5 e 6 são apresentados os resultados obtidos para os algoritmos J48 e PART para cada uma das classes. Em cada uma dessas tabelas, é apresentada a média $m_{mea}(\mathbf{h})$ e o erro padrão $se_{mea}(\mathbf{h})$ para cada um dos algoritmos, onde $mea(\mathbf{h}) \in \{err(\mathbf{h}), prec(\mathbf{h}), rec(\mathbf{h}), F(\mathbf{h}), AUC(\mathbf{h})\}$. Na última coluna das tabelas, é apresentado também o número médio de regras dos classificadores induzidos em cada iteração da validação cruzada estratificada com 10 partições. Nessas tabelas, também foram marcados com o símbolo ▲ os casos em que o algoritmo apresenta melhor

resultado em relação ao outro, com 95% de confiança segundo o teste t de *Student*, e com o símbolo ○ os casos em que ambos os algoritmos não apresentaram diferença estatística, também segundo o teste t de *Student*. Ainda, foram marcados com * os resultados que estão iguais ou piores que o erro majoritário, indicando que o algoritmo não conseguiu aprender os conceitos nas classes minoritárias.

Nas Tabelas 5 e 6, pode-se observar que as taxas geradas por ambos os algoritmos são praticamente iguais. Com base nas taxas geradas pelo J48, pode-se observar que, para a classe "Animais na Pista", foram obtidos bons resultados em todas as medidas, já que todas as medidas apresentaram valores acima de 50%. Tal fato reforça o resultado para essa classe exibido para o algoritmo J48 considerando todas as classes em conjunto, cujo resultado é exibido na Tabela 6. Em relação à classe "Falta de Atenção", observa-se que, apesar dos exemplos dessa classe corresponderem a 46,3% dos exemplos do conjunto de dados, a taxa de erro é a mais alta dentre todas as taxas de erro. No entanto, todas as demais taxas estão acima de 50% para essa classe. É importante observar, na Tabela 6 (J48), as classes cujo número médio de regras foi aproximadamente 1. Isso indica que o classificador gerado é aquele que classifica todos os exemplos na classe majoritária, ou seja, não foi induzido nenhum conhecimento. Daí, para as classes "Defeito na Via", "Desobediência à Sinalização", "Ingestão de Álcool" e "Não Guardar Distância de Segurança", considera-se que o algoritmo J48 não é capaz de aprender conhecimento.

Classe C_v		$err(\boldsymbol{h})$		$prec(\boldsymbol{h})$		$rec(\boldsymbol{h})$	$F(\boldsymbol{h})$		$AUC(\boldsymbol{h})$	# **Reg.**
Animais na Pista	$m_{mea}(\mathbf{h}_{J48})$	0,68%	○	97,80%	▲	73,86%	84,15%	○	92,98%	16
	$se_{mea}(\mathbf{h}_{J48})$	0,01%		0,23%		0,45%	0,34%		0,21%	
Defeito Mecânico no Veículo	$m_{mea}(\mathbf{h}_{J48})$	3,87%	▲	76,60%	▲	18,74%	30,11%		77,22%	530
	$se_{mea}(\mathbf{h}_{J48})$	0,01%		0,67%		0,21%	0,29%		0,21%	
Defeito na Via	$m_{mea}(\mathbf{h}_{J48})$	1,48%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Desobediência à Sinalização	$m_{mea}(\mathbf{h}_{J48})$	6,54%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Falta de Atenção	$m_{mea}(\mathbf{h}_{J48})$	35,18%		63,92%	▲	55,15%	59,21%		69,16%	4541
	$se_{mea}(\mathbf{h}_{J48})$	0,07%		0,09%		0,26%	0,15%		0,08%	
Ingestão de Álcool	$m_{mea}(\mathbf{h}_{J48})$	5,44%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Motorista Dormindo	$m_{mea}(\mathbf{h}_{J48})$	2,78%	▲	76,74%	▲	7,81%	14,16%		72,53%	352
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		1,43%		0,29%	0,46%		0,42%	
Não Guardar Distância de Segurança	$m_{mea}(\mathbf{h}_{J48})$	16,15%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Ultrapassagem Indevida	$m_{mea}(\mathbf{h}_{J48})$	3.87%		67.47%	▲	8.07%	14.40%		80,53%	622
	$se_{mea}(\mathbf{h}_{J48})$	0,01%		0,74%		0,31%	0,49%		0,26%	
Velocidade Incompatível	$m_{mea}(\mathbf{h}_{J48})$	8,40%	▲	67,36%	▲	34,65%	45,77%		76,38%	2479
	$se_{mea}(\mathbf{h}_{J48})$	0,06%		0,52%		0,30%	0,38%		0,20%	

Tabela 5. Resultados do algoritmo J48 para cada classe C_v.
Fonte: autoria própria.

Já na Tabela 6, cujos resultados são referentes ao algoritmo PART, pode-se observar que as taxas de erro estão levemente mais altas que as do J48, isso significa que o algoritmo PART conseguiu gerar regras, porém regras "ruins".

Classe C_v		$err(\boldsymbol{h})$		$prec(\boldsymbol{h})$		$rec(\boldsymbol{h})$		$F(\boldsymbol{h})$		$AUC(\boldsymbol{h})$		# Reg.
Animais na Pista	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,67%	○	97,22%		74,94%	▲	84,63%	○	93,90%	▲	211
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,01%		0,27%		0,42%		0,30%		0,24%		
Defeito Mecânico no Veículo	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	3,90%		67,18%		23,88%	▲	35,22%	▲	80,98%	▲	1255
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,01%		0,58%		0,20%		0,20%		0,17%		
Defeito na Via	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	1,48%	*	47,07%	▲	2,46%	▲	4,67%	▲	76,37%	▲	424
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,01%		3.53%		0,26%		0,47%		0,52%		
Desobediência à Sinalização	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	6,60%	*	47,39%	▲	8,54%	▲	14,47%	▲	84,14%	▲	1584
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,02%		0,83%		0,24%		0,37%		0,13%		
Falta de Atenção	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	2,77%	▲	62,74%		56,95%	▲	59,70%	▲	68,93%	▲	6413
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,01%		0,10%		0,13%		0,09%		0,07%		
Ingestão de Álcool	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	35,59%	*	43,52%	▲	5,57%	▲	9,86%	▲	79,49%	▲	1434
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,08%		1.02%		0,21%		0,33%		0,19%		
Motorista Dormindo	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	5,53%		61,89%		14,67%	▲	23,70%	▲	84,86%	▲	906
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,02%		0,86%		0,34%		0,48%		0,26%		
Não Guardar Distância de Segurança	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	16,35%	*	42,96%	▲	3,82%	▲	7,02%	▲	84,19%	▲	1521
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,01%		0,52%		0,13%		0,22%		0,08%		
Ultrapassagem Indevida	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	3,84%	▲	60,71%		13,80%	▲	22,48%	▲	82,62%	▲	1066
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,02%		0,78%		0,39%		0,55%		0,30%		
Velocidade Incompatível	$m_{mea}(\mathbf{h}_{\mathrm{PART}})$	8,52%		64,59%		36,95%	▲	47,01%	▲	81,21%	▲	2613
	$se_{mea}(\mathbf{h}_{\mathrm{PART}})$	0,05%		0,45%		0,22%		0,26%		0,14%		

Tabela 6. Resultados do algoritmo PART para cada classe C_v.
Fonte: autoria própria.

Resultados e análise – algoritmo APRIORI

Foram geradas 38 regras de associação com confiança maior que 0,8. No Quadro 7 são listadas aquelas com maior valor de confiança. Valores de confiança menores que esse valor gerava um número de regras bastante grande para serem analisadas. Por outro lado, aumentando o valor de confiança para 90%, apenas duas regras, listadas na Quadro 8, foram geradas. Em ambos os casos, o suporte mínimo utilizado foi de 0,1.

- SE Tipo de Veículo = Automóvel E Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Traseira E Modelo da Pista = Reta ENTÃO Estado Físico da Pessoa = Ileso. (Conf = 93%)
- SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Tipo de Acidente = Colisão Traseira. (Conf = 88%)
- SE Ano do Veículo = Seminovo E Tipo de Acidente = Colisão Traseira ENTÃO Estado Físico da Pessoa = Ileso. (Conf = 86%)
- SE Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Lateral ENTÃO Estado Físico da Pessoa = Ileso (Conf 86%)
- SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Estado Físico da Pessoa = Ileso (Conf 86%)
- SE Tipo de Acidente = Colisão Traseira E Período do Dia = Manhã ENTÃO Estado Físico da Pessoa = Ileso (Conf = 85%)
- SE Tipo de Acidente = Colisão Traseira E Modelo da Pista = Reta E Período do Dia = Tarde ENTÃO Estado Físico da Pessoa = Ileso (Conf = 85%)
- SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Modelo da Pista = Reta (Conf = 82%)

Quadro 7. Resultados do algoritmo *Apriori* com confiança maior que 0,8.
Fonte: autoria própria.

- SE Tipo de Veículo = Automóvel E Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Traseira E Modelo da Pista = Reta ENTÃO Estado Físico da Pessoa = Ileso (Conf = 93%)
- SE Tipo de Veículo = Automóvel E Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Traseira ENTÃO Estado Físico da Pessoa = Ileso (Conf = 93%)

Quadro 8. Resultados do algoritmo *Apriori* com confiança maior que 0,9.
Fonte: autoria própria.

Nas regras obtidas pelo *Apriori*, pode-se observar algumas interessantes, como, p. ex., a regra "SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Tipo de Acidente = Colisão Traseira", que relaciona a falta de distância de segurança com uma pista reta em acidentes. Deve ser observado que tal relação está presente em 82% dos casos de acidentes nas rodovias federais do País.

CONCLUSÕES

Tendo em vista que o papel principal dos Dados Abertos Governamentais (DAG) é auxiliar no processo de criação de um governo mais transparente e participativo - tornando as informações mais compreensíveis e próximas dos cidadãos – é necessário que tais dados sejam disponibilizados seguindo um padrão aceito internacionalmente e que possibilite sua ampla reutilização, tanto por máquinas quanto por humanos. Entretanto, observam-se diversos problemas nos dados utilizados neste trabalho, considerados de extrema importância não somente para a população em geral, como para novas medidas e decisões governamentais relação aos acidentes rodoviários que acontecem diariamente nas rodovias federais brasileiras. Revisar o processo de inserção e de coleta desses dados, assim como o modelo dos mesmos, pode melhorar o resultado do processo de MD, uma vez que os diversos erros encontrados fizeram com que a confiabilidade e a qualidade dos resultados gerados pelos algoritmos não tenham sido de todo satisfatórias (ainda que regras interessantes tenham sido observadas). Porém, dada a natureza dos dados e os resultados gerados, conclui-se que a experimentação de novos algoritmos, como os que consideram a incerteza – por exemplo, aqueles que utilizem redes Bayesianas – pode ser válida para a obtenção de melhores resultados.

É importante ressaltar que os dados disponíveis no Portal Brasileiro de Dados Abertos e, consequentemente, os dados utilizados nesse trabalho, seguem o segundo princípio dos DAG, o que determina que tais dados sejam disponibilizados em sua forma primária, ou seja, tais como são coletados na fonte. Porém, essa forma de disponibilizar os dados, os torna mais difíceis de serem entendidos por homem e por máquina. Dados sem qualidade e sem padrões dificultam o processo de MD e, até mesmo, a sua reutilização em outros segmentos. Baseado nos resultados encontrados sugere-se que a forma como os dados estão sendo disponibilizados seja amplamente revista e discutida, bem como seja criado ou utilizado um padrão na forma de publicá-los. Uma solução simples seria a disponibilização destes dados de duas formas diferentes: a primeira respeitando os princípios dos DAG; e, como segunda solução, disponibilizando os dados públicos em um formato padronizado e amplamente utilizado (tais como triplas RDF, por exemplo).

Foi possível observar que alguns resultados obtidos com os algoritmos J48 e PART são promissores em relação à classificação das causas de acidentes. Os valores obtidos de área sob a curva ROC (AUC) estiveram acima de 0,5 e, ao se utilizar o algoritmo *Apriori*, foram geradas 38 regras de associação com confiança maior que 0,8. Porém, a baixa taxa de precisão dos classificadores gerados indica que há a necessidade de maior exploração nos dados para tentar extrair melhores resultados no processo. Ainda, ferramentas de visualização das estatísticas dos dados também podem ser interessantes e enriquecer os resultados encontrados, facilitando a sua reutilização e o entendimento de todos os cidadãos, o que auxiliaria a alcançar uma audiência mais ampla e, consequentemente, os objetivos dos DAG.

Observa-se também que o acesso e a interpretação desses dados não é uma tarefa trivial para o cidadão. Uma maneira de contornar esse problema é disponibilizar os dados de tal maneira que sistemas computacionais possam interpretá-los, bem como a construir interfaces mais simples para a visualização destes dados. Esse é um dos princípios da *Web* Semântica, uma extensão da *web* atual que provê um framework, composta por diversas tecnologias, dentre elas o *Resource Description Framework* (RDF), o OWL, e o SPARQL, para permitir que dados sejam compartilhados e reutilizados por aplicações, empresas e comunidades (Breitman, 2005). A *Web* Semântica fornece um ambiente onde uma aplicação pode consultar esses dados, realizar inferências usando vocabulários específicos de domínio, extrair padrões, etc. Como trabalho futuro, pretende-se propor um método para coletar os dados da base da PRF, em seu formato original, e disponibilizá-los em triplas RDF utilizando uma ontologia para modelagem.

REFERÊNCIAS

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM Sigmod Conference*. Retirado de http://www.it.uu.se/edu/course/homepage/infoutv/ht08/agrawal93mining.pdf

Agune, R. M., Gregorio Filho, A. S., & Bolliger, S. P. (2010). Governo aberto SP: disponibilização de bases de dados e informações em formato aberto. *Congresso Consad de Gestão Pública*. Retirado de http://www.prefeitura.sp.gov.br/cidade/secretarias/upload/controladoria_geral/arquivos/C3_TP_GOVERNO%20ABERTO%20SP%20DISPONIBILIZACAO%20DE%20BASES%20DE%20DADOS.pdf

Balbo, F. A. N. (2011). *Análise multivariada aplicada aos acidentes da BR-277 entre janeiro de 2007 e novembro de 2009*. (Dissertação de Mestrado em Métodos Numéricos em Engenharia). Universidade Federal do Paraná. Retirado de http://www.ppgmne.ufpr.br/arquivos/diss/239.pdf

Baranauskas, J. A., & Monard, M. C. (2000). *Reviewing some machine learning concepts and methods*. Relatórios Técnicos do ICMC/USP, 102.

Bernardini, F. C. (2006). *Combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéricos*. (Tese de Doutorado em Ciências – Ciências de Computação e Matemática Computacional). Universidade de São Paulo/São Carlos. Retirado de http://www.teses.usp.br/teses/disponiveis/55/55134/tde-29092006-110806/

Berry, M. J. A., & Linoff, G. (©1997). *Data mining techniques: For marketing, sales, and customer support*. New York: John Wiley & Sons.

Borgelt, C., & Kruse, R. (2002). Induction of association rules: Apriori implementation. *15th Conference on Computational Statistics*. Retirado de http://www.borgelt.net/papers/cstat_02.pdf

Brasil. Ministério da Justiça. (2014a). *Sistema BR-Brasil: boletins de ocorrências em rodovias federais*. Retirado de http://dados.gov.br/dataset/acidentes-rodovias-federais

Brasil. Portal Brasileiro de Dados Abertos. (2014b). *O que são Dados Abertos? 2014*. Retirado de http://www.governoeletronico.gov.br/acoes-e-projetos/Dados-Abertos

Breitman, K. (2005). *Web semântica: a Internet do futuro*. Rio de Janeiro: LTC.

Carvalho, J. V., Sampaio, M. C., & Mongiovi, G. (1999). Utilização de técnicas de "Data Mining" para o reconhecimento de caracteres manuscritos. *14º Simpósio Brasileiro de Bancos de Dados*, 235-249. Retirado de http://www.dsc.ufcg.edu.br/~sampaio/Artigos/reconhecimentocaracteresmanuscritos.pdf

Domingos, P. A. (2012). Few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87. Retirado de http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Facelli, K., Lorena, A. C., Gama, J., & Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC.

Frank, E., & Witten, I. H. (1998). *Generating accurate rule sets without global optimization*. Hamilton, New Zealand: University of Waikato.

Mitchell, T. (1997). *Machine Learning*. New York: McGraw Hill.

Quinlan, J. R. (1988). Decision trees and multi-valued attributes. In: Hayes, J. E., Michei, D., & Richards, J. (Orgs.). *Machine Intelligence*, 11. New York: Oxford University. Retirado de http://aitopics.org/sites/default/files/classic/Machine_Intelligence_11/MI11-Ch13-Quinlan.pdf

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufmann.

Reis, C. V. R. (2013). *O uso da descoberta de conhecimento em Banco de Dados nos acidentes da BR-381*. (Projeto de pesquisa – Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento). Universidade FUMEC. Retirado de http://www.fumec.br/revistas/sigc/article/view/1508

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., & Paula, M. D. (2003). Mineração de dados. In: REZENDE, S.O. (Org.). *Sistemas inteligentes: Fundamentos e aplicações*. São Paulo: Manole.

The Annotated 8 principles of Open Government Data. (2014). Retirado de http://opengovdata.org/

Witten, I. H., & Frank, E. (2009). *Data Mining: Practical machine learning tools and techniques with java implementations*. Burlington, Massachusetts: Morgan Kaufmann.

Como citar este artigo (ABNT):

COSTA, J. de J.; BERNARDINI, F. C.; VITERBO FILHO, J. 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, Curitiba, v. 3, n. 2, p. 139-157, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this article* (APA):**

Costa, J. J., Bernardini, F. C., & Viterbo Filho, J. (2014). 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*, 3(2), 139-157. Retrieved from http://www.atoz.ufpr.br

ANEXO – Descrição das tabelas do DER da base de dados utilizada

LOCALBR: armazena o local onde ocorreu a ocorrência, ou seja, qual a BR onde foi registrado o acidente. Além disso, identifica o estado onde ocorreu o acidente. Algumas informações dessa tabela (tais como em qual quilômetro da BR ocorreu o acidente, altitude e longitude da ocorrência), seriam úteis para o processo de descoberta de conhecimento. Porém estes são campos que nem sempre são preenchidos. Esta é uma tabela de domínio e está associada à tabela OCORRENCIAACIDENTE e UNIDADEOPERACIONAL. Colunas:

- lbrid - identifica a chave primaria da tabela;
- lbruf - identifica a UF da ocorrência;
- lbrbr - identifica a BR da ocorrência;
- lbrkm - identifica o KM da ocorrência;
- lbrlatitude - identifica a latitude da ocorrência;
- lbrlongitude - identifica a longitude da ocorrência;
- lbrpnvid - identifica a da ocorrência;
- lbratualiza - registra a atualização do local da ocorrência;

OCORRENCIA: essa tabela contém o registro de ocorrências confirmadas a partir das comunicações recebidas. A maioria das tabelas do sistema tem algum tipo de relacionamento com essa tabela. Colunas:

- ocoid - identificação única da ocorrência;
- ocolocal - identificação do local da BR onde aconteceu a ocorrência;
- ocostatus – registra o status da Ocorrência (Aberta, Encerrada, Anulada, Estatística, Retificada e Em Processo)
- ocomunicipio - identificação do município da ocorrência;
- ocosentido - identificação do sentido da via (Crescente ou Decrescente);
- ocodataocorrencia - data da ocorrência;
- ocodataregistro - data do registro da ocorrência;
- ocotipo - tipo da ocorrência (Acidentes Rodoviários, Retenção, Apreensão e Recuperação de Veículos; Pessoa Detenção/Auxílio; Apreensão de CNH; Apreensão de Documento; Apreensão de Carga; Interdição de Rodovias; Ocorrências);
- ococomid - chave estrangeira que identifica a comunicação;
- ocoidorigem – FK (foreign key) da ocorrência retificada;
- ococpfretif - CPF que identifica o executar da retificação;
- ocodatafim - data de fim da ocorrência.

OCORRENCIAVEICULO: cadastro do veículo da pessoa envolvida na ocorrência. Colunas:

- ocvid - identificação única da ocorrência com veículo;
- ocvocoid - chave estrangeira que identifica a ocorrência;
- ocvveiid - chave estrangeira que identifica o veículo;

VEICULO: tabela que contém o cadastro dos dados do veículo. Colunas:

- veiid - identificação única do veículo;
- veiano - ano do veículo;
- veiqtdocupantes - quantidade de ocupantes do veículo;
- veimunicipio - município do veículo;
- veimunorigem - município de origem do veículo;
- veipaisorigem - país de origem do veículo;
- veimundestino - município de destino do veículo;
- veipaisdestino - país de destino do veículo;
- veiproprietario - identificação do tipo de proprietário;
- tvvcodigo - identifica a chave primária da tabela tipoveiculo (chave desligada);

TIPOVEICULO: tabela que identifica os tipos de veículos envolvidos nos acidentes. Esta é uma tabela de domínio e está associada à tabela VEICULO. Colunas:

- tvvcodigo - identificador único do tipo do veículo;
- tvvatualiza - identifica se o registro pode ser atualizado (manutenção de histórico);
- tvvativo – indica se o veículo está ativo.

OCORRENCIAPESSOA: registro das pessoas envolvidas no acidente. Colunas:

- opeid - identificador único da ocorrência de pessoas;
- opeocoid - chave estrangeira que identifica a ocorrência;
- opepesid - chave estrangeira que identifica a pessoa;
- opeportevalidade - validade do porte de arma;
- opettecodigo - chave estrangeira que identifica o tipo de envolvido;
- opeanexo - declaração em anexo('S','N');
- opecondalegadas - condições alegadas para a ocorrência.

PESSOA: tabela que registra os dados da pessoa envolvida na ocorrência. Colunas:

- pesid - identificador único de pessoa;
- pesexpedidor - órgão expedidor;
- pesufexpedidora - UF expedidora;
- pesdatanascimento – data de nascimento;
- pesnaturalidade – naturalidade;
- pesnacionalidade – nacionalidade;
- pessexo – sexo;
- pesteccodigo - chave estrangeira que identifica o estado civil;
- pestgicodigo - chave estrangeira que identifica o grau de instrução das pessoas;
- pesmunicipio – munícipio de residência;
- pescep – CEP;
- pestopcodigo - chave estrangeira que identifica a ocupação principal da pessoa;
- pesmunicipioori - município de origem da pessoa;
- pespaisori – país de origem da pessoa;

- pesmunicipiodest - município de destino da pessoa;
- pespaisdest – país de destino da pessoa;
- pesveiid - chave estrangeira que identifica o veículo da pessoa;
- pescinto - identifica se a pessoa estava utilizando cinto;
- pescapacete - identifica se a pessoa estava usando capacete;
- peshabilitado - identifica se a pessoa é habilitada;
- pessocorrido - identifica se a pessoa foi socorrida;
- pesdormindo - identifica se a pessoa estava dormindo;
- pesalcool - identifica se a pessoa estaca alcoolizada;
- peskmpercorre - indica quantos quilômetros ela percorreu;
- peshorapercorre - identifica o tempo que ela percorreu a quilometragem;
- pescategcnh - identifica a categoria da CNH;
- pesregistrocnh - numero do registro da CNH;
- pesufcnh - UF em que tirou a habilitação;
- pespaiscnh – país onde tirou a CNH;
- pesdatahabil – data da habilitação;
- pesdatavalidade – validade da habilitação;
- pesapelido - apelido atribuído a pessoa;
- pesidade – idade da pessoa;
- pesaltura – altura;
- pespeso – peso da pessoa;
- pescicatriz - identifica se possui cicatriz;
- pestatuagem - identifica se possui tatuagem;
- pessinal - identifica se a pessoa possui sinal;
- peslesao - identifica se a pessoa possui lesão;
- pestcccodigo - chave estrangeira que identifica a cor cabelo;
- pestctcodigo - chave estrangeira que identifica a cor da cútis;
- pestclcodigo - chave estrangeira que identifica a cor do olho;
- pespertences - descreve os pertences das pessoas no local da ocorrência;
- pesdadoscomplement - dados complementares das pessoas;
- vestigios_drogas - indicador de vestígios de droga;
- descricao_cicatriz – descrição da cicatriz da pessoa;
- descricao_tatuagem - descrição da tatuagem da pessoa
- descricao_sinal - descrição dos sinais da pessoa
- descricao_deficiencia – descrição da deficiência física da pessoa.

OCORRENCIAACIDENTE: cadastro da ocorrência envolvendo veículos. Colunas:
- oacocoid - identificador único do acidente;
- oacttacodigo - chave estrangeira que identifica o tipo de acidente;
- oactcacodigo - chave estrangeira que identifica a causa do acidente;
- oacdano - dano causado no acidente;
- oacdanoterc - dano causado a terceiro;
- oacdanoamb - dano causado ao acidente;
- oacrefera - referência ponto A (descreve o ponto de referencia);
- oacreferb - referência ponto B (descreve o ponto de referencia);
- oacdistab - distância entre os pontos A e B;
- oacdistac - distância entre os pontos A e C;
- oacdistbc - distância entre os pontos B e C; oacmodelopista - identificação do modelo de pista;
- oacsentido1 - descrição do sentido 1;
- oacsentido2 - descrição do sentido 2;
- oacqtdfaixa1 - quantidade de faixas no sentido 1;
- oacqtdfaixa2 - quantidade de faixas no sentido 2;
- oacacostamento1 - indicador de acostamento no sentido 1;
- oacacostamento2 - indicador de acostamento no sentido 2;
- oacimagem - indicador da existência de imagem;
- oacdescdanopat - descrição do dano causado ao patrimônio;
- oacdescdanoterc - descrição dos danos causados a terceiros;
- oacdescdanoamb - descrição dos danos causados ao ambiente;
- oaccanteiro - descreve se o local da ocorrência possui ou não canteiro;
- oaclinhacentral - descreve se a pista possui ou não linha central;
- oacorientpista - descreve se o acidente aconteceu no sentido crescente ou decrescente da pista de rolamento;
- oacversaocroqui - informa se foi realizado ou não um croqui para ocorrência.

CAUSAACIDENTE: qualifica as causas do acidente. Colunas:
- tcacodigo – identificador da causa do acidente;
- tcadescricao – descrição da causa do acidente (Velocidade incompatível, Ultrapassagem indevida, Ingestão de álcool, Desobediência à sinalização, Defeito mecânico, Defeito na via, Falta de atenção, Dormindo, Animais na pista, Não guardar distância de segurança, Outras, Não informado)

TIPOACIDENTE: qualifica os tipos de acidente. Colunas:
- ttacodigo - identificador único do tipo de acidente;
- ttaatualiza - indica se o registro permite atualização;
- ttaativo - indica se o registro está ativo.