

DESCOBERTA DE CONHECIMENTO EM DADOS ABERTOS SOBRE ACIDENTES DE TRÂNSITO COM CICLISTAS EM PORTO ALEGRE

DANIELE CARVALHO DOS SANTOS

*Graduanda em Administração Pública e Social
Departamento de Ciências Administrativas da Escola de Administração da
UFRGS
danielesantos1503@gmail.com*

DANIEL GARCIA VIANNA

*Graduando em Administração Pública e Social
Departamento de Ciências Administrativas da Escola de Administração da
UFRGS
dgianna30@gmail.com*

FABRÍCIO REICHERT DOS SANTOS

*Graduando em Administração Pública e Social
Departamento de Ciências Administrativas da Escola de Administração da
UFRGS
fabricioreichert@gmail.com*

IVAN KRÁS BORGES SCHARDOSIM

*Graduando em Administração Pública e Social
Departamento de Ciências Administrativas da Escola de Administração da
UFRGS
ivanschardosim@gmail.com*

DANIELA FRANCISCO BRAUNER

*DSc. em Informática pela PUC-Rio
Professora do Departamento de Ciências Administrativas da Escola de
Administração da UFRGS
danibrauner@gmail.com / daniela.brauner@ufrgs.br
ORCID <https://orcid.org/0000-0002-6379-7825>*

RESUMO

Objetivo: Este artigo descreve a aplicação da técnica de descoberta de conhecimento em bancos de dados em cima de um conjunto de dados abertos sobre acidentes envolvendo ciclistas em Porto Alegre. O principal objetivo do trabalho é obter informações a partir dos dados que permitam entender as ocorrências. Para tanto, foi escolhido como período de abrangência de acidentes aqueles ocorridos em um período de cinco anos, entre 2011 e 2015.

Design/Methodologia/Abordagem: Esse estudo foi realizado seguindo as etapas do processo de descoberta de conhecimento em bancos de dados, com a utilização de algoritmos de mineração de dados através da ferramenta Weka, bem como de visualização de resultados via dashboards criados na ferramenta Qlik Sense.

Resultados: Com o uso da mineração de dados foi possível analisar as localidades onde a ocorrência de acidentes envolvendo ciclistas foi maior, utilizando a divisão da cidade de Porto Alegre em regiões. Podendo constatar uma maior incidência de acidentes envolvendo dois ou mais ciclistas na região sul da cidade.

Originalidade/valor: Sua originalidade se justifica ao trazer uma experiência prática do uso de dados abertos sobre acidentes para descoberta de conhecimento sobre acidentes de bicicleta, demonstrando o valor no uso de dados como apoio à tomada de decisões pelo poder público.

Palavras-chave: Acidentes de trânsito. Ciclistas. Descoberta de conhecimento em bancos de dados. Mineração de dados.

DISCOVERY OF KNOWLEDGE IN OPEN DATA ON TRANSIT ACCIDENTS WITH CYCLISTS IN PORTO ALEGRE

ABSTRACT

Purpose: This paper describes the practical application of the process of knowledge discovery in databases over a set of open data on accidents involving cyclists in Porto Alegre, Brazil. The main objective was to obtain information from the data to understand the occurrences. To do so, it was chosen as the period of coverage of accidents those occurring in a period of five years between 2011 and 2015.

Design/methodology/approach: This study was carried out following the steps of the process of Knowledge Discovery in Databases (KDD), using data mining algorithms in the Weka tool, as well as visualizing results via dashboards created in the Qlik Sense tool.

Findings: Using data mining it was possible to analyze the regions of Porto Alegre city where the occurrence of accidents involving cyclists was greater. As a result, it was possible to see a greater incidence of accidents involving two or more cyclists in the south region, where there is less adequate infrastructure.

Practical implications and originality: Its originality is justified by bringing a practical experience of using open data to discover knowledge about bicycle accidents, demonstrating the value in the use of public data as support for decision making by the public power.

Keywords: *Traffic accidents. Cyclists. Knowledge discovery in databases. Data Mining.*

I INTRODUÇÃO

Cada vez mais, as pessoas buscam alternativas mais saudáveis e menos nocivas ao meio ambiente para locomover-se nas cidades (Lobo, Binatti & Ferreira, 2015). Neste sentido, estudos realizados na área de mobilidade urbana afirmam que o aumento das ciclofaixas também contribuem para a ocupação de espaços públicos, transformando-os em espaços de convivências (Caccia, 2015). Dessa forma, a criação de ciclofaixas também pode ser encarada como um investimento na segurança pública, pois reduz os espaços ociosos, em que normalmente possuem maior índice de criminalidade. No intuito de ter uma melhor qualidade de vida e bem-estar, como também diminuir os poluentes nos locais onde vivem, várias cidades do mundo possuem um sistema de utilização de bicicletas compartilhadas, incluindo o município de Porto Alegre, que possui o sistema chamado BikePOA. Tal sistema foi implementado pela Prefeitura de Porto Alegre, através da empresa Serttel Ltda, em parceria com o Banco Itaú, em 22 de setembro de 2012 (Empresa Pública de Transporte e Circulação [EPTC], 2017), no qual os ciclistas podem ter acesso a estes veículos em 39 pontos diferentes do município. Em 2017, houve um aumento de 84% no uso de bicicletas de aluguel em Porto Alegre (Scirea, 2017). Em 2015, um levantamento sobre o perfil do ciclista (Lobo et al., 2015), apontou que 62,9% dos ciclistas porto-alegrenses pedalam 5 ou mais dias na semana.

Nesse sentido, o presente artigo visa averiguar, em um determinado período de tempo, se houve ou não mudanças nos índices de acidentes de trânsito envolvendo ciclistas. Para isso, utilizamos como subsídio os dados abertos obtidos no portal #DataPoa (EPTC, 2015). O portal de dados abertos de Porto Alegre foi criado com o objetivo de “[...] convidar a comunidade a participar do desenvolvimento de soluções inteligentes para a nossa cidade [...]”. Além disso, os dados presentes neste portal possibilitam “[...] desenvolvedores criar e construir plataformas web, aplicativos e softwares que ajudarão a cidade e seus moradores como um todo, criando laços colaborativos entre governo municipal, empresas, desenvolvedores e cidadãos” (Companhia de Processamento de Dados do Município de Porto Alegre [PROCEMPA], s.d.).

Com base nestas premissas, foram encontrados neste sítio eletrônico um conjunto de dados referentes aos acidentes de trânsito com uma base amostral entre os anos 2000 e 2015. Todavia, para que nossa pesquisa tivesse maior precisão quanto a influência do aumento ou diminuição do número de acidentes relacionados com o aumento do número de ciclistas, foi definida uma amostra de 5 anos, ou seja, entre os anos de 2011 e 2015.

Neste sentido, apresentamos os resultados da análise dos dados baseada na aplicação do processo de descoberta de conhecimento em bancos de dados e no uso de algoritmos de

aprendizagem de máquina aliado a análises estatísticas simples através de *dashboards*. O objetivo é extrair conhecimento relevante sobre quais são os bairros ou zonas de Porto Alegre, como também os períodos do dia e dias da semana onde há maior incidência de acidentes envolvendo bicicletas.

Para que a população porto-alegretense utilize a bicicleta como meio de transporte ou alternativa real aos modelos já existentes é necessário que o poder público ofereça proteção com a criação de novas ciclofaixas e ciclovias adequadas para a mobilidade urbana. Por isso, o tema abordado possui relação direta com a realidade urbana de Porto Alegre, conforme comprova a matéria publicada no Jornal Correio do Povo (2017), em 15 de abril de 2017, o qual apresenta um levantamento realizado pela Coordenação de Informações de Trânsito da EPTC. Este estudo indicou que desde a implementação do BikePoa, 290 ciclistas sofreram acidentes, sendo que no ano de 2016 o número caiu para 163. Contudo, ao final de 2016 houve aumento de 25% de acidentes envolvendo ciclistas mulheres na interseção da rua Mariante com a avenida Protásio Alves. Assim, os resultados obtidos no artigo poderão subsidiar o Poder Público, como também a população envolvida, da necessidade de adoção de uma política pública voltada para melhoria dos locais de tráfego, ampliando e melhorando as ciclofaixas já existentes, bem como relativo à educação no trânsito, horários de maior conscientização e da importância do aumento da utilização desta modalidade de transporte que, além de menos poluente, é mais econômica e proporciona aos usuários exercitarem-se durante o tráfego até seus locais de trabalho, de estudo e/ou de lazer.

Para isso, o trabalho está organizado da seguinte forma: na seção 2, apresentam-se alguns trabalhos relacionados; na seção 3, descreve-se o processo de descoberta de conhecimento, explicando a metodologia utilizada na condução deste trabalho; na seção 4, apresentam-se os resultados da análise de dados e, por fim, apresentam-se as considerações finais, bem como sugestões para os trabalhos futuros.

2 TRABALHOS RELACIONADOS

Rosa, Silva e Murta (2017) mapearam a densidade dos acidentes com ciclistas na cidade de Montes Claros, estado de Minas Gerais (MG), para identificar possíveis problemas nos locais onde existam ciclovias na região. Para isso, foi utilizado um Sistema de Informações Geográficas (SIG). A identificação dos pontos de maior ocorrência de acidentes envolvendo ciclistas nas vias urbanas de Montes Claros pelo SIG mostrou a dificuldade para compatibilizar os dados entre diferentes órgãos públicos, além de apontar a concentração dos acidentes nas principais vias da cidade. O mesmo caso aconteceu ao tentarmos usar os logradouros para identificação exata dos

acidentes, porém os dados não eram padronizados e identificados, dificultando a análise através das ferramentas utilizadas no estudo relatado neste artigo.

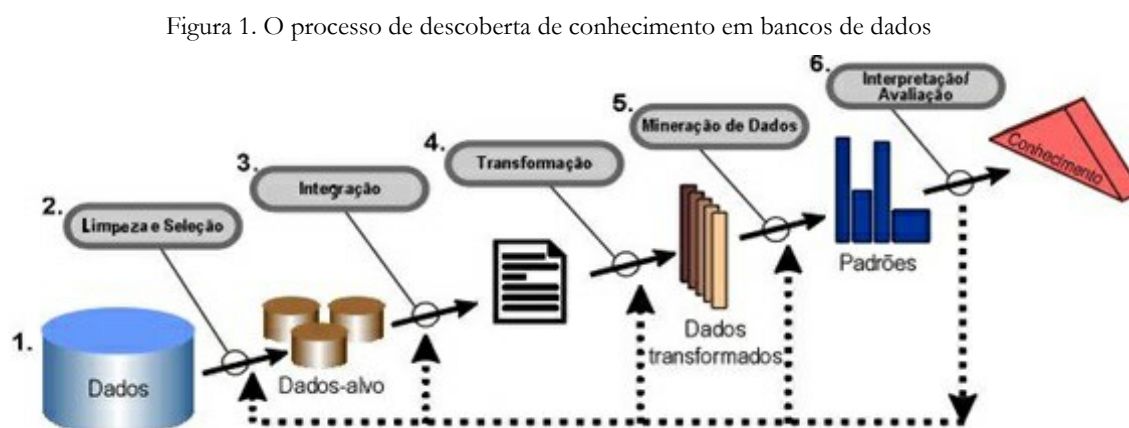
Em Waechter (2014) são utilizadas técnicas estatísticas como análise descritiva, geoestatística e multivariada para avaliar a situação dos acidentes de trânsito envolvendo ciclistas em Porto Alegre (RS), com dados dos anos de 2004 a 2013. Diferente da proposta deste artigo, Waechter focou somente nas técnicas estatísticas e aqui relatamos o uso de algoritmos de aprendizagem de máquina para identificação de padrões nos dados.

Em Flores, Steffenello e Winck (2017) é apresentado um estudo para mapear os padrões de acidentes de trânsito com vítimas fatais ocorridos no Rio Grande do Sul e em Santa Maria, entre 2007 e fevereiro de 2016, a fim de identificar os principais fatores envolvidos. Para isso, foi empregado o mesmo processo utilizado neste artigo, o processo de Descoberta de Conhecimento em Bancos de Dados, por meio da execução dos algoritmos A priori e J48, na ferramenta de mineração de dados Weka. Os resultados demonstraram que o uso do Weka foi satisfatório, porém os dados são focados no Rio Grande do Sul e com foco geral para acidentes de trânsito e não focam apenas nos acidentes de bicicleta, diferente do objetivo deste trabalho.

3 O PROCESSO DE DESCOBERTA DE CONHECIMENTO

Segundo Fayyad, Piatetsky, Smyth e Uthurusamy (1996), a descoberta de conhecimento é um processo que possui certa complexidade na extração de informações potencialmente úteis, mas que muitas vezes são desconhecidas pelo usuário, a partir dos dados armazenados em um banco de dados.

Deste modo, a descoberta de conhecimento é um processo que envolve as seguintes etapas, conforme apresentadas na Figura 1:



Fonte: De "Advances in Knowledge Discovery and Data Mining" de U. M. Fayyad, G. Piatetsky Shapiro, P. Smyth, e R. Uthurusamy, 1996, AAAI Press, The Mit Press.

Abaixo, faz-se uma descrição detalhada de cada uma das etapas do processo de descoberta do conhecimento:

1. Dados: definir o tipo de conhecimento a descobrir, compreendendo o domínio da aplicação e o tipo de decisão que tal conhecimento pode contribuir para melhorar;

2. Limpeza e Seleção: selecionar um conjunto de dados alvo ou focar num subconjunto, onde a descoberta deve ser realizada;

3. Integração: Nessa etapa o usuário deve remover ruídos quando necessário, escolher estratégias para manipular campos de dados ausentes e formatar dados para a adequá-los à ferramenta de mineração;

4. Transformação: Devemos localizar características úteis para representar os dados dependendo do objetivo da tarefa. No caso dos acidentes relacionados a ciclistas, reduzir o número de variáveis, e enriquecendo o valor da informação;

5. Mineração de dados: selecionar os métodos a serem utilizados para localizar padrões nos dados, além de buscar o melhor ajuste dos parâmetros para a tarefa.

6. Interpretação e avaliação: após uma retomada aos passos anteriores, refletir sobre possíveis conclusões dos resultados encontrados;

7. Conhecimento: Nessa etapa deve ser implementado o conhecimento descoberto, incorporando o conhecimento à performance do sistema, ou documentá-lo e reportá-lo às partes interessadas.

Neste artigo, aplicaremos o processo de descoberta de conhecimento, apresentando as etapas realizadas nas seções subsequentes.

3.1 COLETA DE DADOS

O conjunto de dados utilizado neste trabalho foi obtido através do portal #DataPoa (EPTC, 2015). O #DataPoa é um portal de dados abertos com informações da cidade de Porto Alegre, que compila uma quantidade considerável de dados e indicadores de diversas áreas, tais como, saúde, educação, meio ambiente, limpeza urbana, entre outras. Para o presente trabalho, selecionamos os dados referentes à mobilidade urbana, mais especificamente, os relativos a acidentes de trânsito envolvendo ciclistas na cidade. Nesta categoria, haviam informações que abrangiam o período de 2000 a 2015, não apenas referentes a acidentes envolvendo bicicletas, mas também outros meios de transporte da cidade. Os arquivos de dados foram obtidos em sua totalidade para posterior seleção dos dados necessários para análise.

3.2 LIMPEZA E SELEÇÃO

Ao escolhermos trabalhar com os dados referentes a acidentes de trânsito, que abarcavam informações de 2000 a 2015, decidimos utilizar apenas os dados disponíveis dos últimos 5 anos, de 2011 até 2015, por apresentarem as informações mais atualizadas e corresponder ao período de início da implementação de ciclovias e do Bike POA na cidade. Percebemos que nestes anos houve um crescimento no número de ciclovias, ainda que não o suficiente e conforme planejado inicialmente pela Prefeitura de Porto Alegre, e de pessoas que passaram a utilizar a bicicleta como meio de transporte alternativo, bem como para outros fins (Correio do Povo, 2017).

Como existiam variáveis que não estavam presentes em todos os 5 anos que abrangiam o estudo, foram excluídas colunas não coincidentes para padronizá-los e poder manipulá-los em uma única tabela, sendo que, em seguida, foram compiladas as informações de todo o período selecionado em apenas um arquivo (2011-2015), em formato .csv. Após a padronização em um arquivo único, foi constatado que algumas informações seriam desnecessárias para a análise dos dados, sendo que possivelmente não trariam nenhuma informação relevante, tais como, a fonte do dado, o número do boletim de ocorrência, dentre outras. Os dados utilizados são representados através do dicionário de dados presentes na Tabela 1.

Tabela 1 - Dicionário de dados do arquivo utilizado neste artigo.

Coluna	Tipo de dados	Descrição
TIPO_ACID	Texto	Tipo de acidente
DIA_SEM	Texto	Dia da semana em que ocorreu o acidente
FERIDOS	Numérico	Número de feridos
FATAIS	Numérico	Número de vítimas fatais
AUTO	Numérico	Número de automóveis envolvidos
TAXI	Numérico	Número de táxis envolvidos
LOTACAO	Numérico	Número de lotações envolvidas
ONIBUS_URB	Numérico	Número de ônibus de Porto Alegre envolvidos
ONIBUS_MET	Numérico	Número de ônibus metropolitanos envolvidos
ONIBUS_INT	Numérico	Número de ônibus (outros tipos) envolvidos
CAMINHAO	Numérico	Número de caminhões envolvidos
MOTO	Numérico	Número de motos envolvidas
BICICLETA	Numérico	Número de bicicletas envolvidas
TEMPO	Texto	Situação do tempo no momento do acidente
NOITE_DIA	Texto	Período do dia em que ocorreu o acidente (noite/dia)
REGIAO	Texto	Região da cidade em que ocorreu o acidente
MES	Numérico	Mês em que ocorreu o acidente
ANO	Numérico	Ano em que ocorreu o acidente
UPS	Numérico	Unidade de severidade padrão do acidente (1-acidente com danos materiais, 5-acidente com ferido, 13-acidente com morte)

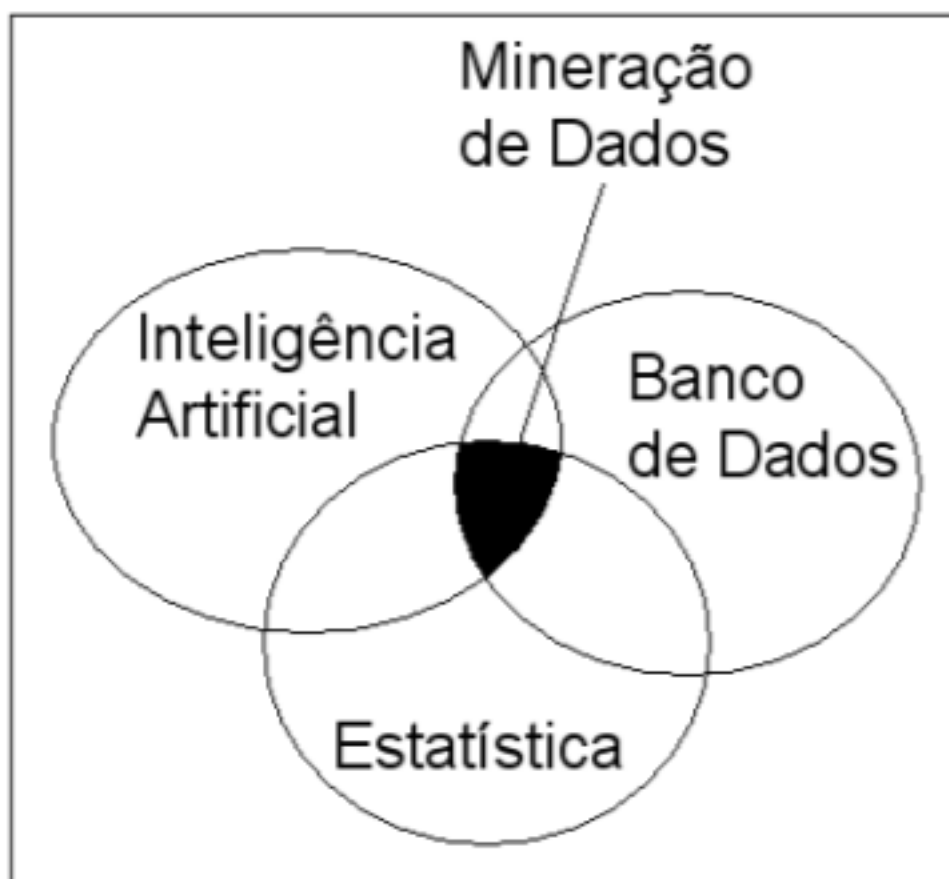
Fonte: Autores (2017)

Inicialmente, esse novo arquivo foi salvo no formato .csv, para utilização na ferramenta Qlik Sense. Logo após, para utilização no software livre Weka, foi feita a conversão do arquivo para o formato .arff, padrão da ferramenta, no intuito de facilitar a leitura por este software na geração de uma árvore de decisão.

3.3 MINERAÇÃO DE DADOS

Na etapa de mineração de dados são utilizadas técnicas e algoritmos de diferentes áreas do conhecimento, principalmente inteligência artificial (especialmente aprendizagem de máquina), banco de dados (recursos para manipular bases de dados) e estatística (comumente na avaliação e validação de resultados), conforme a Figura 2 (Silva, 2004).

Figura 2. Mineração de dados e recursos de diversas áreas.



Fonte: Recuperado do artigo acadêmico “Mineração de Dados - Conceitos, Aplicações e Experimentos com WEKA” de M. Silva, 2004, Universidade do Estado do Rio Grande do Norte - UERN, Natal, RN, Brasil, <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>

No presente excerto, utilizamos no processo de descoberta do conhecimento tanto a algoritmos de inteligência artificial implementados na ferramenta Weka, manipulação de bancos de dados em planilhas eletrônicas, como também análises estatísticas e construção de dashboards interativos através da ferramenta Qlik Sense.

O sistema Weka é uma ferramenta de código aberto, desenvolvida pela Universidade de Waikato que contém uma coleção de algoritmos de aprendizagem de máquina pré-implementados para execução em tarefas de mineração de dados (University of Waikato [UW], n.d.). Nela, os algoritmos podem ser aplicados diretamente numa base de dados ou invocados através de chamadas programadas em linguagem de programação Java.

Dentre os algoritmos implementados no Weka, temos os algoritmos de aprendizagem supervisionada (Witten & Frank, 2000). Estes algoritmos aprendem um modelo de forma supervisionada, ou seja, é fornecida uma classe à qual cada amostra no treinamento pertence e de forma preditiva ao qual o algoritmo pode desempenhar inferências nos dados com o intuito de fornecer previsões ou tendências, obtendo informações não disponíveis a partir dos dados originais. Logo, os algoritmos de classificação são exemplos de algoritmos supervisionados (com ênfase na precisão da regra) em que é possível determinar o valor de um atributo através dos valores de um subconjunto dos demais atributos da base de dados. Para tal tarefa, as formas mais comuns de representação de conhecimento dos algoritmos de classificação são regras e árvores. Para representá-los, os algoritmos Id3, C45, J48, por exemplo, geram como resultado árvores de classificação, enquanto que outros como Prism, Part, OneR geram regras de classificação (Silva, 2004). Neste processo, foi utilizado o algoritmo J48 para representação do conhecimento em forma de árvore de decisão, cujos resultados serão apresentados na sequência.

De posse do arquivo de dados, executamos o algoritmo J48 no Weka e observamos a dificuldade por parte do software em gerar uma árvore de decisão relevante, devido a quantidade de informações existentes no arquivo e a baixa quantidade de incidência de cada um dos atributos. Com isso, a árvore ficou cheia de ramificações e com margem de incidência pequena, não apresentando bons resultados na classificação das instâncias de teste. Deste modo, foi necessário realizar uma limpeza dos dados mais adequada, deixando somente variáveis relacionadas diretamente a bicicletas, retirando outros meios de transporte (táxi, automóvel, lotação, ônibus, caminhão e moto), bem como o mês da ocorrência, o ano e o grau de severidade (UPS) de cada acidente. A retirada destas informações se deu em razão de não apresentarem relação direta com os acidentes envolvendo bicicleta, ou seja, não havendo necessidade da relação deste meio de transporte com outros. Além disso, retiraram-se as variáveis ano e mês por não

haver grande variação nestes pontos. Após esse processo, os dados restantes estão demonstrados no dicionário de dados, conforme Tabela 2.

Tabela 2 - Dicionário de dados do arquivo utilizado neste artigo.

Coluna	Tipo de dados	Descrição
DIA_SEM	Texto	Dia da semana em que ocorreu o acidente
FERIDOS	Numérico	Número de feridos
FATAIS	Numérico	Número de vítimas fatais
BICICLETA	Numérico	Número de bicicletas envolvidas
TEMPO	Texto	Situação do tempo no momento do acidente
NOITE_DIA	Texto	Período do dia em que ocorreu o acidente (noite/dia)
REGIAO	Texto	Região da cidade em que ocorreu o acidente

Fonte: Autores (2017)

Deste modo, utilizamos os algoritmos baseados na atividade de classificação, em que é possível usar como base um subconjunto de dados para determinar o valor de um atributo. Em seguida, executamos os algoritmos DecisionStump, J48, Hoeffdingtree, LMT, RandomForest, RandomTree, REPTree. Após testes com estes algoritmos na ferramenta Weka, apresentam-se os resultados de classificação nas Figuras 3 (Decision Stump), 4 (HoeffdingTree), 5 (LMT), 6 (RandomForest), 7 (RandomTree), 8 (REPTree) e 9(J48), sendo esse último aquele que apresentou maior margem de acertos.

Figura 3. Algoritmo Decision Stump

```

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      762           61.4516 %
Incorrectly Classified Instances    478           38.5484 %
Kappa statistic                    0.0457
Mean absolute error                 0.4716
Root mean squared error             0.4863
Relative absolute error             99.5309 %
Root relative squared error         99.9205 %
Total Number of Instances          1240
    
```

Fonte: Autores (2017)

Figura 4. Algoritmo HoeffdingTree

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      768          61.9355 %
Incorrectly Classified Instances    472          38.0645 %
Kappa statistic                    0.0163
Mean absolute error                 0.4725
Root mean squared error             0.486
Relative absolute error             99.7288 %
Root relative squared error         99.8592 %
Total Number of Instances          1240
```

Fonte: Autores (2017)

Figura 5. Algoritmo LMT

```
Time taken to build model: 0.98 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      766          61.7742 %
Incorrectly Classified Instances    474          38.2258 %
Kappa statistic                    0.0441
Mean absolute error                 0.4704
Root mean squared error             0.4897
Relative absolute error             99.2869 %
Root relative squared error         100.6105 %
Total Number of Instances          1240
```

Fonte: Autores (2017)

Figura 6. Algoritmo RandomForest

```

Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      716          57.7419 %
Incorrectly Classified Instances    524          42.2581 %
Kappa statistic                    0.0151
Mean absolute error                 0.4498
Root mean squared error             0.4921
Relative absolute error             94.9281 %
Root relative squared error        101.0996 %
Total Number of Instances          1240
    
```

Fonte: Autores (2017)

Figura 7. Algoritmo RandomTree

```

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      737          59.4355 %
Incorrectly Classified Instances    503          40.5645 %
Kappa statistic                    0.0376
Mean absolute error                 0.444
Root mean squared error             0.4989
Relative absolute error             93.6977 %
Root relative squared error        102.5136 %
Total Number of Instances          1240
    
```

Fonte: Autores (2017)

Figura 8. Algoritmo REPTree

```

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      754      60.8065 %
Incorrectly Classified Instances    486      39.1935 %
Kappa statistic                    0.0226
Mean absolute error                 0.4653
Root mean squared error             0.4885
Relative absolute error             98.2115 %
Root relative squared error         100.3737 %
Total Number of Instances          1240
    
```

Fonte: Autores (2017)

Figura 9. Algoritmo J48

```

J48 pruned tree
-----

FERIDOS = 0: DIA (111.0/32.0)
FERIDOS = 1: DIA (1003.0/380.0)
FERIDOS = 2
|  REGIAO = CENTRO: NOITE (9.0/3.0)
|  REGIAO = LESTE: DIA (19.0/5.0)
|  REGIAO = NORTE: DIA (32.0/13.0)
|  REGIAO = SUL: NOITE (51.0/17.0)
FERIDOS = 3
|  REGIAO = CENTRO: DIA (0.0)
|  REGIAO = LESTE: DIA (2.0)
|  REGIAO = NORTE: DIA (2.0)
|  REGIAO = SUL: NOITE (7.0/2.0)
FERIDOS = 4: NOITE (2.0)
FERIDOS = 5: DIA (1.0)
FERIDOS = 15: NOITE (1.0)

Number of Leaves   :    13

Size of the tree   :    16

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      777      62.6613 %
Incorrectly Classified Instances    463      37.3387 %
Kappa statistic                    0.0598
Mean absolute error                 0.4659
Root mean squared error             0.4852
Relative absolute error             98.3348 %
Root relative squared error         99.6895 %
Total Number of Instances          1240
    
```

Fonte: Autores (2017)

Conforme os testes com o conjunto de dados foram sendo feitos, ficou constatado que a representação do atributo “feridos”, dentro do arquivo .arff, alterava a representação da árvore de decisão. Por isso, foram utilizadas duas formas de representação, um arquivo utilizando o atributo como forma numérica na Figura 10 e outro elencando as variáveis encontradas no atributo, conforme Figura 11.

Figura 10. Atributo como forma numérica.

```
@attribute DIA_SEM { DOMINGO, SEGUNDA_FEIRA, TERCA_FEIRA, QUARTA_FEIRA, QUINTA_FEIRA, SEXTA_FEIRA, SABADO }
@attribute FERIDOS numeric
@attribute FATAIS { 0, 1 }
@attribute BICICLETA { 1, 2, 15 }
@attribute TEMPO { BOM, CHUVOSO, NUBLADO }
@attribute NOITE_DIA { DIA, NOITE }
@attribute REGIAO { CENTRO, LESTE, NORTE, SUL }
@data
```

Fonte: Autores (2017)

Figura 11. Atributo com variáveis explícitas.

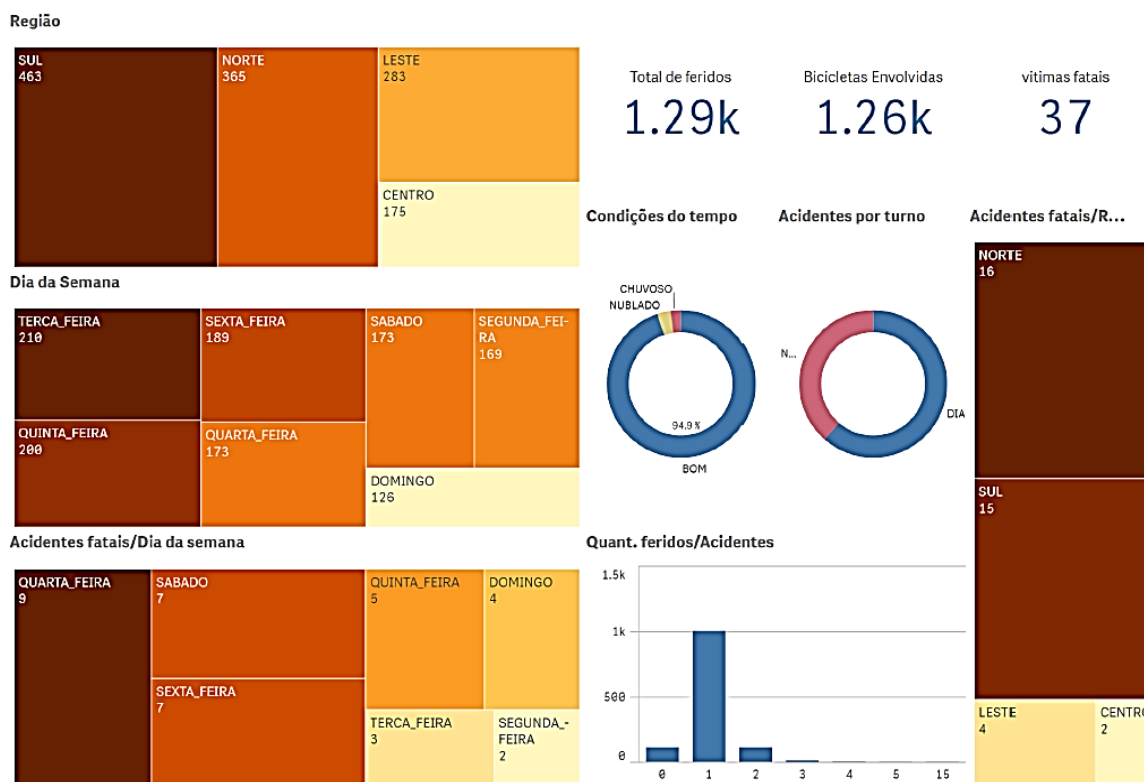
```
@attribute DIA_SEM { DOMINGO, SEGUNDA_FEIRA, TERCA_FEIRA, QUARTA_FEIRA, QUINTA_FEIRA, SEXTA_FEIRA, SABADO }
@attribute FERIDOS { 0, 1, 2, 3, 4, 5, 15 }
@attribute FATAIS { 0, 1 }
@attribute BICICLETA { 1, 2, 15 }
@attribute TEMPO { BOM, CHUVOSO, NUBLADO }
@attribute NOITE_DIA { DIA, NOITE }
@attribute REGIAO { CENTRO, LESTE, NORTE, SUL }
@data
```

Fonte: Autores (2017)

Para a utilização da ferramenta Qlik Sense, foi necessária apenas a utilização do arquivo em formato .csv, com as modificações após os testes na ferramenta Weka. Assim, optou-se por utilizar os mesmos dados para ambas ferramentas por entender ser essa a melhor forma de se chegar aos melhores resultados.

O painel de análise na ferramenta Qlik Sense foi elaborado, conforme a Figura 12, em que constam diversos cruzamentos dos dados.

Figura 12. Painel de análise na ferramenta Qlik Sense.

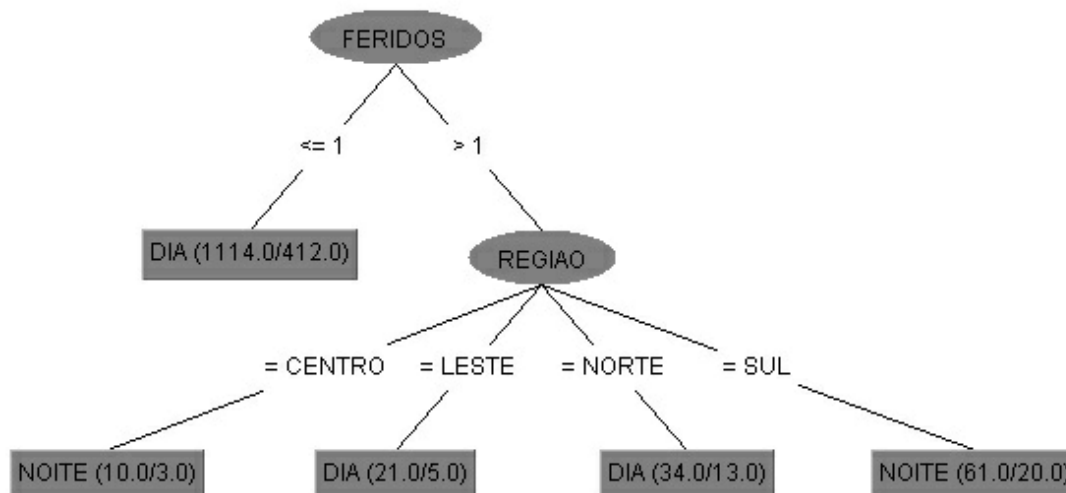


Fonte: Autores (2017)

4 INTERPRETAÇÃO, RESULTADOS E AVALIAÇÃO

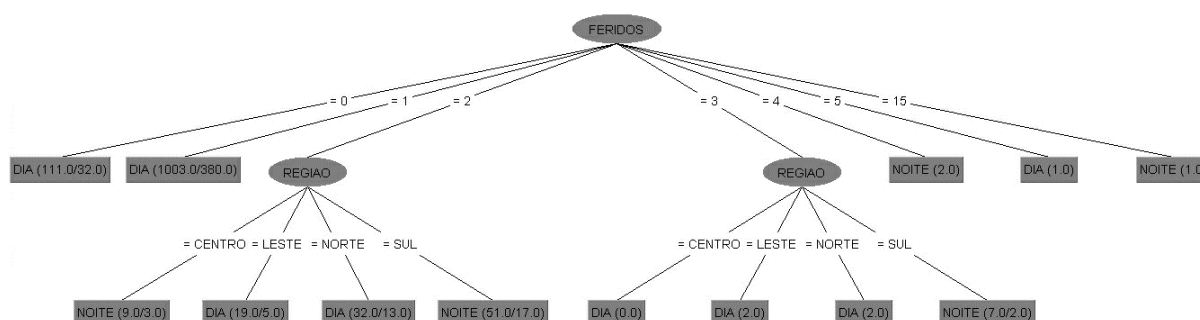
Com os dois arquivos obtidos após a otimização dos dados, foi possível manipulá-los no Weka e gerar duas árvores de decisão, conforme Figuras 13 e 14.

Figura 13. Árvore de decisão 1



Fonte: Autores (2017)

Figura 14. Árvore de decisão 2



Fonte: Autores (2017)

Com essas duas árvores de decisão, pode-se perceber que a incidência de acidentes no período do dia é muito maior envolvendo um ferido na ocorrência. Essa incidência é correspondente ao perfil de pessoas que utilizam a bicicleta não apenas como um instrumento de lazer, mas para os deslocamentos do dia-a-dia, servindo como modo de locomoção.

Quando temos a presença de mais de um ferido, conforme a Figura 13, na região sul, no período da noite, o número ocorrências aumenta. Ao analisar a Figura 14 podemos adentrar na análise envolvendo mais de um ferido, observando maior incidência de casos com dois feridos na zona sul de Porto Alegre, no período da noite.

A predominância de ocorrências na região sul no período da noite, com acidentes envolvendo mais de um ferido, pode ser explicada pela grande área arborizada dessa região, convidativa para passeios acompanhados, por essa região do município de Porto Alegre contar com uma topografia de vias mais planas do que a maior parte das demais regiões, como a região central que é caracterizada por vários declives, e por apresentar nessa mesma região baixo número de ciclovias, conforme Plano Diretor Cicloviário de Porto Alegre - PDCI (EPTC, 2009).

Além disso, merece destaque um dado que foge dos parâmetros, a ocorrência de um caso envolvendo 15 feridos, conforme a Figura 14, por ser um caso atípico o grupo decidiu pesquisar especificamente sobre esse dado e encontramos notícias sobre um atropelamento de ciclistas que participavam de um evento promovido pelo Movimento “Massa Crítica”. Conforme repórter Felipe Truda, o ocorrido foi descrito da seguinte forma: “Ao ter a passagem bloqueada por ciclistas na Rua José do Patrocínio, bairro Cidade Baixa, ele avançou com o carro sobre pelo menos 15 ciclistas.” (Truda, 2015).

Após a realização da análise utilizando o Weka, utilizou-se os mesmos dados no sistema Qlik Sense, que possui uma interface mais dinâmica com o usuário, apresentando assim os dados de um modo mais interativo e atrativo, facilitando o entendimento para o usuário final.

Após importação dos dados no Qlik Sense, foi possível perceber o pequeno número de acidentes fatais em relação ao total de feridos nos acidentes envolvendo bicicletas, conforme pode-se constatar pela Figura 15, bem como a maior quantidade de acidentes nas regiões sul e norte, conforme Figura 16.

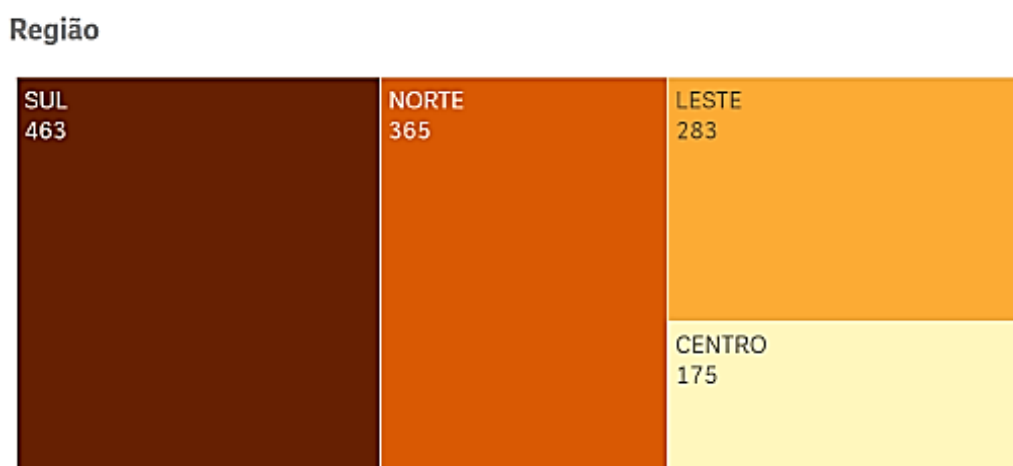
Figura 15. Feridos e vítimas fatais



Fonte: Autores (2017)

Na Figura 15, percebe-se a baixa incidência de vítimas fatais, a qual pode ser explicada em razão da maior conscientização dos ciclistas com uso de equipamentos de segurança, e pela expansão, ainda que pequena, das ciclovias em Porto Alegre.

Figura 16. Acidentes envolvendo bicicleta nas regiões da cidade de Porto Alegre

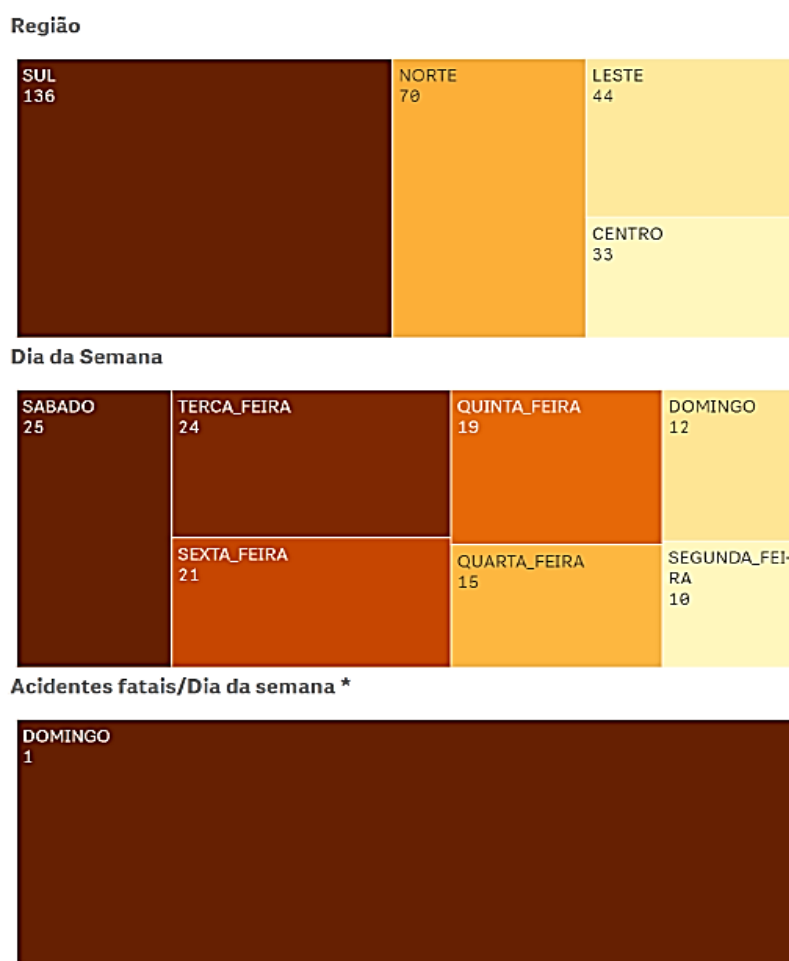


Fonte: Autores (2017)

Uma expectativa apontada pelo grupo de pesquisa após o levantamento dos dados com a ferramenta Weka, era ratificar suas informações com a ferramenta Qlik Sense quanto à relação entre a quantidade de feridos e a quantidade de bicicletas envolvidas. Para isso, selecionamos os dados contendo mais de um ferido e o resultado foi o demonstrado na Figura 17, em que temos maior ocorrência de acidentes na região sul, no sábado, com uma vítima fatal no domingo. Esse dado atesta os resultados demonstrados pela árvore de decisão nas Figuras 13, em que a maior prevalência envolvendo mais de um ferido ocorre na região sul de Porto Alegre.

Além disso, outro fator importante a observar é o aumento do número de ocorrências no sábado, em que podemos inferir ser o dia da semana em que há uma maior circulação de pessoas em grupo andando de bicicleta como forma de lazer, o que acentua a probabilidade de acidentes, devido à inexperiência destas pessoas em circular entre os demais veículos circunstancialmente, em razão da inexistência de uma malha cicloviária adequada, levando os ciclistas a dividir o mesmo espaço de trânsito nessa região.

Figura 17. Acidentes com mais de 1 ferido



Fonte: Autores (2017)

Ao realizar uma comparação dos dados obtidos após a mineração de dados com o mapa da Rede Ciclovária Estrutural de Porto Alegre presente na Figura 18, observou-se que grande parte da malha ciclovária da região Sul não possui ciclovias implementadas, aparecendo no mapa apenas como Projeto da Malha Ciclovária.

Figura 18. Mapa da rede ciclovária de Porto Alegre



Fonte: Recuperado de “Mapa do plano ciclovário de Porto Alegre”. Empresa Pública de Transporte e Circulação, s.d., http://lproweb.procempa.com.br/pmpa/prefpoa/epc/usu_doc/mapa_rede_ciclovitaria_estrutural.pdf

Esse projeto previa que em todas as vias em vermelho fossem implantadas ciclovias, mas somente as vias em azul contam atualmente com ciclovias em funcionamento, muitas vezes não há continuidade em uma mesma via, o que dificulta o deslocamento dos ciclistas, tanto os que utilizam a bicicleta como lazer quanto para os que a usam como forma de alternativa de deslocamento.

Deste modo, pode-se deduzir que a falta de infraestrutura para locomoção de bicicletas nessa região contribui para maior incidência de acidentes envolvendo ciclistas, corroborando com os dados trazidos por Lobo et al. (2015) de que 27,2% das pessoas entrevistadas em 2015 responderam que a falta de infraestrutura cicloviária é um dos problemas do dia-a-dia enfrentados pelos ciclistas.

5 CONSIDERAÇÕES FINAIS

Neste artigo, foi relatado o uso de dados abertos encontrados no portal #Datapoa sobre acidentes de trânsito para análise de acidentes envolvendo bicicletas na cidade de Porto Alegre, usando o processo de descoberta de conhecimento em bancos de dados. O objetivo foi identificar, através dos dados, as regiões do município, os períodos do dia e os dias da semana com maior número de casos envolvendo este meio de transporte.

Nesse sentido, após a realização da mineração dos dados selecionados para análise referente ao período de 2011 a 2015, no qual houve ampliação da malha cicloviária em Porto Alegre e, conseqüentemente, do número de ciclistas, encontramos algumas informações importantes: do total de acidentes que envolveram bicicletas (em torno de 1260), a maior incidência foi observada no turno diurno (762), principalmente nas terças-feiras (210). Infere-se que esta informação corrobora com o resultado de um levantamento realizado em 2015, o qual informa que 85,8% dos destinos dos ciclistas de Porto Alegre são o trabalho (Lobo et al., 2015).

Em relação às ferramentas de mineração de dados utilizadas, o software Weka, através da árvore de decisão, indicou melhor correlação dos dados através do classificador J48, mostrando informações que em uma primeira análise não haviam sido cogitadas. Com a ferramenta Qlik Sense, houve dificuldades de correlacionar os dados, pois a sua utilização requer conhecimento prévio das informações que se quer obter, sendo apenas utilizado nesta pesquisa para melhor visualização dos dados. Apesar disso, conseguimos avançar em uma análise bastante proveitosa e, visto o pouco incentivo à utilização de softwares deste tipo no dia-a-dia, percebemos que o Weka, embora requeira uma dedicação maior para aprendizado de suas funcionalidades, se apresentou como uma alternativa interessante para desenvolvedores de softwares em organizações públicas, por ser um software livre e gratuito.

Para próximas pesquisas a serem realizadas sobre esta temática, sugere-se o uso de dados mais atualizados, já que em nosso caso foi possível o acesso apenas a dados sobre acidentes de trânsito até o ano de 2015. Além disso, é recomendável relacionar os dados disponíveis dos logradouros em que ocorrem os acidentes envolvendo bicicletas com os locais em que existam ou não ciclovias, pois facilitaria intervenções e investimentos inteligentes por parte do poder público ao se entrecruzarem estas informações. Nos dados adquiridos no sítio eletrônico do #DataPOA tais dados estavam presentes, mas não foram utilizados por não possuírem identificadores consistentes que pudessem ser relacionados nas ferramentas utilizadas. Logo, para que isso seja possível, faz-se necessária a utilização de técnicas mais avançadas e/ou de uma modificação no sistema de arquivamento de dados por parte da plataforma #DatapPoa. No caso da cidade de Porto Alegre, o uso de dados de logradouros poderia subsidiar decisões políticas para a implementação de novos trechos de malha cicloviária, a reestruturação ou modificação dos já existentes, como também a aplicação de outras medidas, como intervenções públicas educativas em pontos de alta ocorrência de acidentes de bicicletas na cidade e de alto fluxo de trânsito.

Por fim, é essencial o incentivo a estudos que contribuam à tomada de decisão do Poder Público nesta área, visto o aumento do uso de bicicletas em Porto Alegre (Scirea, 2017). Este trabalho evidenciou, através de dados reais, que é essencial a implementação do restante das ciclovias e ciclofaixas, de forma a melhorar a segurança dos ciclistas junto ao trânsito de Porto Alegre. Além disto, as experiências apresentadas neste trabalho reforçaram a utilidade do uso de dados abertos para geração de conhecimento.

REFERÊNCIAS

- Caccia, L. S. (2015). Mobilidade Urbana: Políticas Públicas e Apropriação do Espaço em Cidades Brasileiras. (Dissertação de Mestrado). Universidade Estadual do Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brasil. Recuperado de <https://www.lume.ufrgs.br/bitstream/handle/10183/133191/000984211.pdf>
- Companhia de Processamento de Dados do Município de Porto Alegre (s.d.). Sobre o Data Poa. Porto Alegre: Autor. Recuperado de <http://www.datapoa.com.br/about>
- Correio do Povo (2017, abril 15). Ciclovias ainda são insuficientes em Porto Alegre. Porto Alegre: Autor. Recuperado em 26 julho, 2017, de www.correiodopovo.com.br/Noticias/Geral/Transito/2017/4/615213/Ciclovias-ainda-sao-insuficientes-em-Porto-Alegre

Empresa Pública de Transporte e Circulação (2009). Plano diretor cicloviário – PDCI de Porto Alegre. Porto Alegre: Autor. Recuperado de http://lproweb.procempa.com.br/pmpa/prefpoa/eptc/usu_doc/pdci_relatorio_final.pdf

Empresa Pública de Transporte e Circulação (2015, junho 21). Acidentes de trânsito em Porto Alegre. Porto Alegre: PROCempa. Recuperado de <http://datapoa.com.br/dataset/acidentes-de-transito>

Empresa Pública de Transporte e Circulação (s.d.). Bike PoA. Porto Alegre: Autor. Recuperado de http://www2.portoalegre.rs.gov.br/eptc/default.php?p_secao=228

Empresa Pública de Transporte e Circulação (s.d.). Mapa do plano cicloviário de Porto Alegre. Porto Alegre: Autor. Recuperado de http://lproweb.procempa.com.br/pmpa/prefpoa/eptc/usu_doc/mapa_rede_cicloviaria_estrutural.pdf

Fayyad, U. M., Piatetsky Shapiro, G., Smyth, P. & Uthurusamy, R (1996). *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence – AAAI Press, The MIT Press.

Flores, J., Jr., Steffenello, L. & Winck, A. (2017, abril). Mapeamento de Padrões de Acidentes de Trânsito com Vítimas Fatais a partir de Dados Públicos do Governo do Estado do Rio Grande do Sul. Anais do XIII Escola Regional de Informática de Banco de Dados - ERIBD, Passo Fundo, RS, Brasil. Recuperado de http://www.upf.br/_uploads/Conteudo/erbd2017/anais_ERBD2017_final_pos.pdf

Lobo, Z., Binatti, G. & Ferreira, G. (2015). Perfil do ciclista Brasileiro. Parceria Nacional Pela Mobilidade Por Bicicleta. Rio de Janeiro: PROURB. Recuperado de www.ta.org.br/perfil/perfil.pdf

Revista Bicicleta - O Portal da Bicicleta (s.d.). Porto Alegre ganha novas estações de aluguel de bike. Rio do Campo, SC, Brasil: Autor. Recuperado de http://revistabicicleta.com.br/bicicleta_noticia.php?porto_alegre_ganha_novas_estacoes_de_aluguel_de_bike&id=251#

Rosa, B. O., Silva, A., & Murta, A. L. S (2017). Análise da densidade de acidentes com ciclistas, utilizando SIG: Estudo de caso em Montes Claros (MG). *Revista UNIABEU*, 10(25), 245-258. Recuperado de <http://revista.uniabeu.edu.br/index.php/RU/article/view/2801>

Silva, M. (2004). Mineração de dados - Conceitos, aplicações e experimentos com Weka (Artigo acadêmico). Universidade Estadual do Rio Grande do Norte – UERN, Mossoró, RN, Brasil. Recuperado de <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>

Scirea, B (2017, fevereiro 10). Após melhorias, uso de bicicletas de aluguel aumenta 84% em Porto Alegre. Porto Alegre: Zero Hora. Recuperado de <https://gauchazh.com.br/rs/porto-alegre/noticia/2017/02/apos-melhorias-uso-de-bicicletas-de-aluguel-aumenta-84-em-porto-alegre-9717983.html>

Truda, F. (2015, fevereiro 2). Motorista que atropelou ciclistas há 4 anos no RS ainda aguarda julgamento. Porto Alegre: G1 RS. Recuperado de <http://g1.globo.com/rs/rio-grande-do-sul/noticia/2015/02/motorista-que-atropelou-ciclistas-ha-4-anos-no-rs-ainda-aguarda-julgamento.html>

University of Waikato (n.d.). Weka 3 – Machine learning software in Java. Waikato, NZ: Author. Recuperado de <http://www.cs.waikato.ac.nz/ml/weka>

Waechter, L (2014). Avaliação estatística dos acidentes de trânsito com bicicletas na cidade de Porto Alegre (Monografia de bacharelado). Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brasil. Recuperado de <http://www.lume.ufrgs.br/bitstream/handle/10183/109915/000951627.pdf?sequence=1>

Witten, I. H. & Frank, E. (2000). Data mining – Practical machine learning tools. San Francisco: Morgan Kaufmann.