

Utilização de mineração de dados na análise de dados sobre o comportamento do mosquito *Aedes Aegypti*

Elias Rocha Gonçalves Júnior (UCAM-Campos) eliasrgjunior1@gmail.com
Virgínia Siqueira Gonçalves (UCAM-Campos) virginiasideiragoncalves@gmail.com
Julio Cezar Negri Ramos (UCAM-Campos) julio.negri@gmail.com

Resumo:

Neste artigo, foi demonstrado como a técnica de mineração de dados é eficiente no descobrimento de conhecimentos antes ocultos sobre o comportamento do mosquito *Aedes Aegypti*. As informações foram mineradas de um banco de dados gerado por um questionário distribuído digitalmente e que obteve 341 respostas. O foco dessa pesquisa era apenas a cidade de Campos dos Goytacazes, no Rio de Janeiro, e trazia perguntas sobre hábitos de indivíduos e sua relação com a transmissão de doenças. Os dados foram minerados utilizando a ferramenta WEKA, no qual foi aplicada a técnica de clusterização com o algoritmo *SimpleKMeans*. Os conhecimentos obtidos demonstraram uma relação entre indivíduos do sexo feminino, jovens (entre 20 e 39 anos) que moram no Centro da cidade e ficam menos de 12h por dia em casa, sendo que esse grupo estava associado à questão de terem contraído dengue. Também foi percebida uma associação clara entre o Centro da cidade e esses indivíduos.

Palavras chave: Mineração de Dados, WEKA, *Aedes Aegypti*.

Use of data mining in the analysis of data on the behavior of the *Aedes Aegypti* mosquito

Abstract

In this article, it was demonstrated how the technique of data mining is efficient in the discovery of previously hidden knowledge about the behavior of the *Aedes Aegypti* mosquito. The information was mined from a database generated by a digitally distributed questionnaire that obtained 341 responses. The focus of this research was only the city of Campos dos Goytacazes, in Rio de Janeiro, and it asked questions about individuals' habits and their relation with the transmission of diseases. The data were mined using the WEKA tool, in which the clustering technique was applied with the algorithm *SimpleKMeans*. The obtained knowledge showed a relation between female individuals, young people (between 20 and 39 years) who live in the Center of the city and stay less than 12 hours a day at home, being that group was associated with the question of having contracted dengue fever. It was also perceived a clear association between the City Center and these individuals.

Key-words: Data Mining, WEKA, *Aedes Aegypti*

1. Introdução

A mineração de dados tem sido uma técnica que tem mostrado potencial uso para gestores, na tomada de decisões e na construção e descoberta de conhecimentos que estariam ocultos ou passariam despercebidos sob uma grande camada de dados aleatórios.

A análise de dados era realizada desde seus primórdios por meios estatísticos, nesse processo, os dados têm sido sistematicamente coletados e armazenados eletronicamente. Somente esse

método matemático não proporciona a descoberta de informações desconhecidas em uma base de dados. Segundo Bothorel, Serrurier e Hurter (2011), a Mineração de Dados visa extrair o máximo de conhecimento a partir de bancos de dados com grande volume de informação.

As principais motivações para o uso das técnicas de mineração de dados são o aumento da capacidade de processamento e armazenamento de dados, o baixo custo do processo, já que se usa uma ferramenta de software livre, a grande quantidade de dados que têm sido armazenados nos últimos anos e a inviabilidade de se analisar esses dados de forma manual.

Fayyad (1996) expõe a Mineração de Dados como um processo para identificação, nos dados utilizados como base, padrões válidos, úteis e compreensíveis. Nesse contexto, surge a questão de como tirar proveito do conhecimento subjacente a todo o patrimônio digital que vem sendo acumulado. Portanto, a mineração de dados pode ser vista como uma ferramenta aplicável nesse contexto.

A dengue é uma infecção viral sistêmica cujo principal vetor é o mosquito *Aedes aegypti*, com boa adaptação ao clima tropical e subtropical. A doença pode se manifestar de formas variadas, desde as oligossintomáticas até formas graves, representando um importante problema de Saúde Pública em nível mundial (BHATT *et. al.*, 2013). Segundo estimativa da World Health Organization (2009), 50 milhões de casos da doença ocorrem a cada ano, em média, levando a 500 mil hospitalizações e mais de 20 mil óbitos.

As manifestações clínicas da doença dependem da susceptibilidade do hospedeiro, da genética viral, da resposta imune e de possíveis reações cruzadas advindas de infecções prévias. Estas características parecem estar relacionadas com o desenvolvimento de formas graves da dengue (GUZMAN *et. al.*, 2010).

Alguns fatores influem na densidade de mosquitos, como as condições de saneamento e os aspectos socioeconômicos e culturais das comunidades. Donalísio e Glasser (2002) afirmam que o conhecimento desses fatores é fundamental para a compreensão das epidemias de dengue e o direcionamento das ações de controle.

Segundo Rigau-Pérez (1999), o sistema de vigilância epidemiológica da dengue tem um papel primordial nas atividades de prevenção e controle da doença: ser capaz de detectar precocemente o aumento de casos e epidemias, além dos casos graves, e a alteração no perfil epidemiológico. Para tanto, torna-se necessário uma informação consistente e oportuna, diagnóstico laboratorial otimizado, critérios de definição de caso claros e objetivos e profissionais de saúde com conhecimento clínico da doença.

Neste âmbito, o objetivo deste trabalho é utilizar as técnicas de mineração de dados para descobrir hábitos e comportamentos do mosquito *Aedes Aegypti*, além de buscar determinar o perfil principal de pessoas que contraíram alguma doença transmitida pelo mesmo. Uma pesquisa com 14 questões foi estruturada e distribuída eletronicamente. As respostas foram processadas através do software WEKA e os resultados serão apresentados ao longo deste artigo.

2. Metodologia

No processo de tratamento dos dados, objeto deste trabalho, será utilizado o conceito de descoberta de conhecimento em bases de dados, KDD. Tal processo foi definido por Fayyad, Piatetsky-shapiro e Smyth (1996), como sendo: “... o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados”.

Para levantar os dados para a mineração, foi elaborado um questionário com 14 perguntas utilizando a ferramenta online *Google Forms* (forms.google.com), representada na Figura 1, e

distribuído eletronicamente através de redes sociais, encaminhamentos via e-mail, programas de compartilhamento de mensagens e SMS.

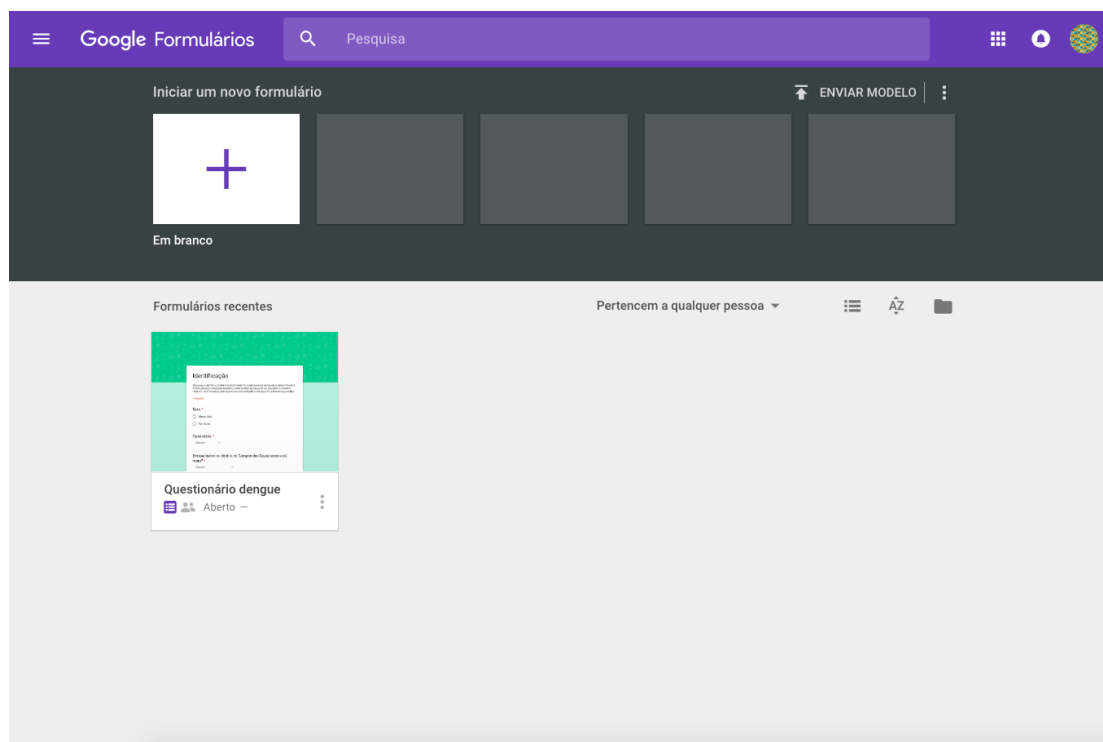


Figura 1 – Tela inicial da ferramenta online *Google Forms*

O questionário ficou disponível durante os meses de maio e junho de 2017 e continha perguntas que visavam respostas que ajudariam a construir uma espécie de perfil das vítimas do *Aedes Aegypti*. As seguintes perguntas estruturavam o questionário:

- a) Sexo
- b) Faixa etária
- c) Em que bairro ou distrito de Campos dos Goytacazes você mora?
- d) Suas atividades te permitem ficar em casa:
 - a. Mais de 12h por dia
 - b. Menos de 12h por dia
- e) Em qual período você fica mais tempo em casa?
- f) Já teve dengue?
- g) Sua casa tem mosquitos?
- h) Os mosquitos preferem picar você na maioria das vezes?
- i) Em que parte da casa ocorrem as picadas?
- j) Onde se escondem?
- k) Quais objetos/locais mais atraem?
- l) Onde os mosquitos mais atacam?
- m) Horários que mais atacam?

n) Como você se previne?

O questionário obteve 341 respostas e, após a limpeza dos dados, foi preparado para ser minerado a fim de se buscar resultados relevantes que gerassem conhecimentos. Segundo Elmasri e Navalhe (2005) citado por Nascimento (2012), a mineração de dados se refere à descoberta de novas informações em função de padrões em grandes quantidades de dados. Seu objetivo é a extração de conhecimentos dessas grandes quantidades de dados que seriam impossíveis de serem extraídos manualmente. A mineração é um dos processos de KDD (Knowledge Discovery in Databases). Segundo Amo (2003), é comum existir uma confusão entre os termos KDD e mineração de dados, mas na verdade o segundo é uma das etapas do primeiro:

“Afinal, o que é Mineração de Dados? Falando simplesmente, trata-se de extrair ou minerar conhecimento de grandes volumes de dados. Muitas pessoas consideram o termo Mineração de Dados como sinônimo de Knowledge Discovery in Databases (KDD) ou Descoberta de Conhecimento em Banco de Dados.”.

Ainda de acordo com Amo (2003), o KDD é um processo amplo que consiste de várias etapas, conforme demonstrado na Figura 2:

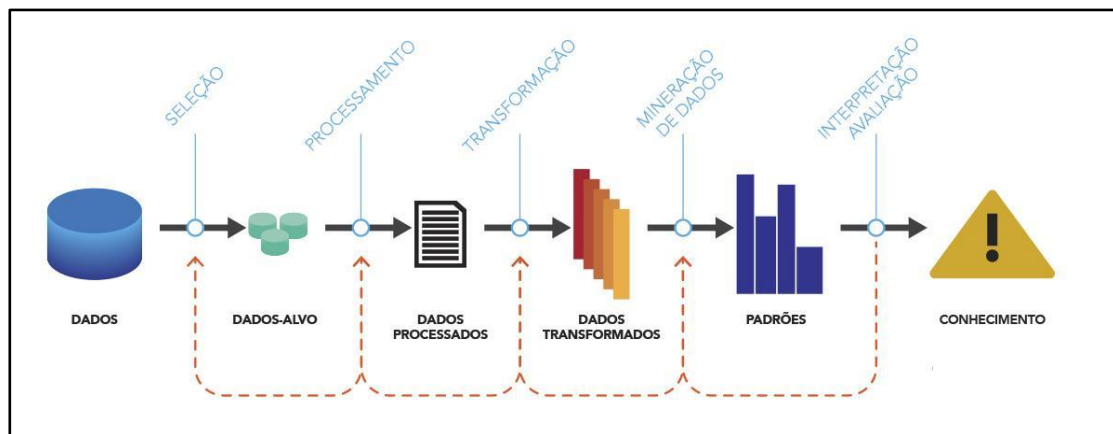


Figura 2 – O ciclo do processo de KDD

Amo (2003) relata as etapas deste processo, que inclui:

- a) Limpeza dos dados: etapa na qual são eliminados ruídos e dados inconsistentes;
- b) Integração dos dados: etapa na qual diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados;
- c) Seleção: etapa na qual são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como endereço e telefone não são de relevantes para decidir se um cliente é um bom comprador ou não;
- d) Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação);
- e) Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse;
- f) Avaliação ou Pós-processamento: etapa na qual são identificados os padrões interessantes de acordo com algum critério do usuário;
- g) Visualização dos Resultados: etapa na qual são utilizadas técnicas de representação de

conhecimento a fim de apresentar ao usuário o conhecimento minerado.

A ferramenta utilizada para fazer a mineração de dados é o software WEKA (Waikato Environment for Knowledge Analysis), desenvolvido pela Universidade de Waikato da Nova Zelândia.

Com o WEKA, foi realizado o processo de clusterização com o algoritmo *SimpleKMeans*, que é um dos mais usados para se trabalhar com regras de associação em bancos de dados, pois ele detecta os *itemsets* frequentes (L1), que são os conjuntos de itens com um padrão repetitivo (AGRAWAL *et. al.* 1993).

Para Salvador *et. al.* (2009), a clusterização, uma das técnicas da mineração de dados, busca agrupar os dados de tal maneira que seja capaz de potencializar a similaridade dos objetos de um mesmo grupo e/ou diferença entre grupos distintos.

De acordo com Tan, Steinbach e Kumar (2009), o objetivo é aprender uma função alvo F para mapear cada grupo de atributos X para um dos rótulos de classes Y pré-determinados.

Com os dados prontos, o WEKA foi configurado para utilizar apenas 9 dos 15 atributos, conforme Figura 3, e foi realizada a etapa da mineração denominada agrupamento ou clusterização. Nesta etapa, o software identifica grupos de objetos através de algumas características em comum ou critérios de similaridade. Um cluster se trata de um conjunto de objetos similares entre si e diferentes dos objetos que pertençam a outros grupos.

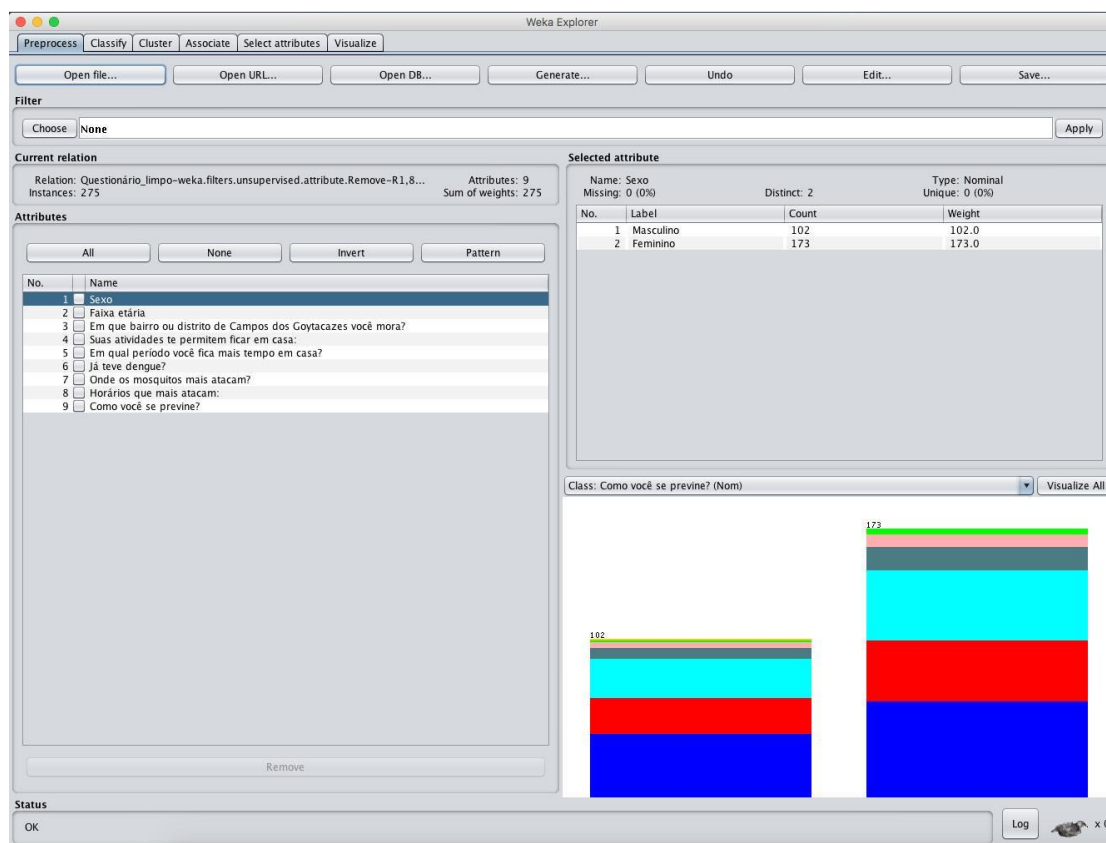


Figura 3 – Tela de processamento do WEKA

Como o objetivo do trabalho é a obtenção de conhecimento sobre o comportamento do mosquito *Aedes Aegypti* e a transmissão da dengue, foram gerados cinco clusters, como indicado na Figura 4, utilizando o algoritmo *SimpleKMeans*, sendo selecionados apenas os clusters que indicavam que as pessoas daquele grupo tinham contraído dengue. A partir disso, foram

analisados os clusters que apresentavam resultados indicando pessoas que contraíram dengue anteriormente.

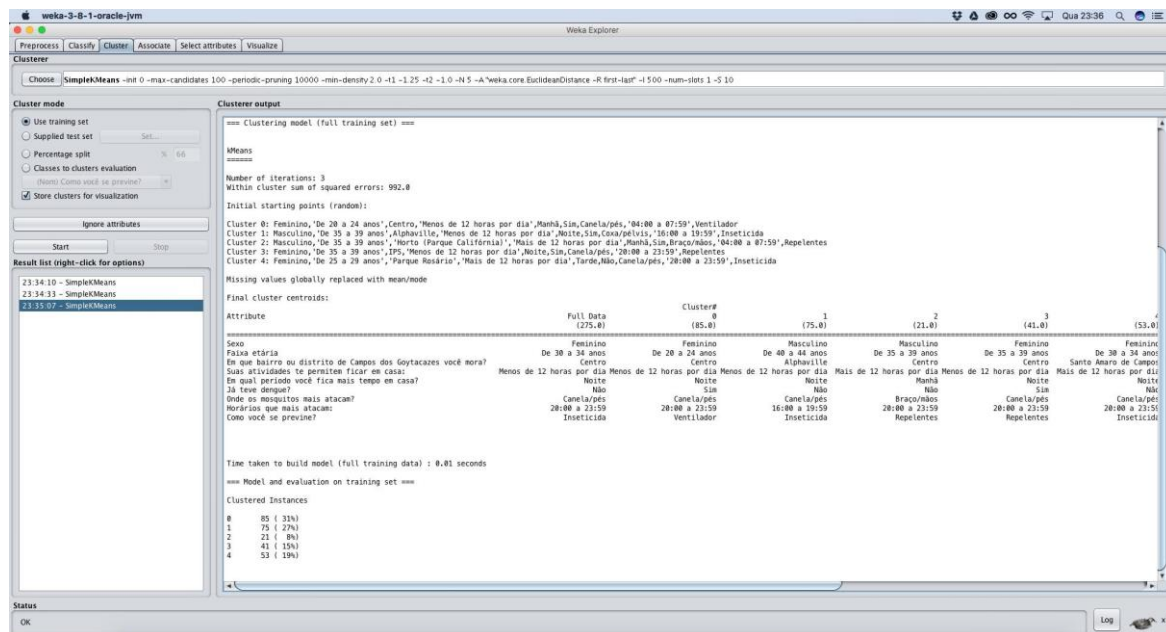


Figura 4 – Tela de resultados do WEKA mostrando os resultados do processo de clusterização

3. Resultados e Discussão

Na análise dos dados minerados pelo WEKA, foram destacados apenas os clusters #0 e #3, pois eram aqueles que associavam informações com usuários que declararam na pesquisa já terem contraído dengue pelo menos uma vez.

O Cluster #0, vide Figura 5, indica uma relação entre pessoas do sexo feminino da faixa etária entre 20 e 24 anos, moradoras do Centro e que permaneciam por menos de 12 horas por dia contraíram dengue, mais especificamente na parte da noite.

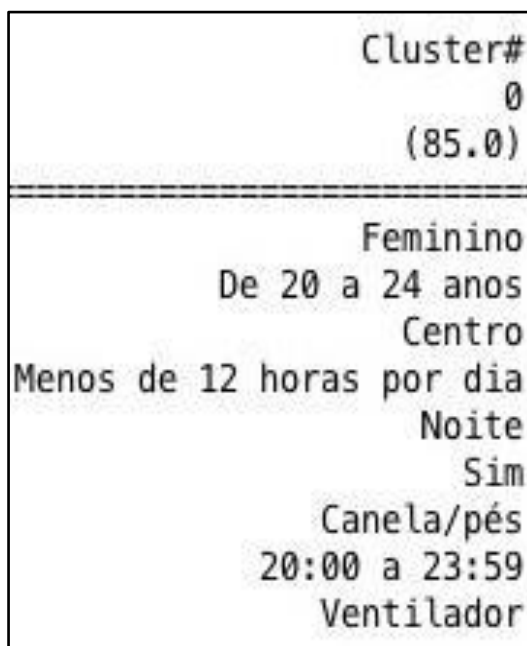


Figura 5 – Cluster #0 gerado pelo aplicativo SimpleKmeans

Houve também uma associação relativa à parte do corpo em que teria ocorrido a picada, sendo verificada a opção como principal alvo dos mosquitos. O horário em que as picadas eram mais frequentes é nos horários entre 20:00 e 23:59. O modo de prevenção de picadas mais utilizado por este grupo é o ventilador.

Já o Cluster #3, representado na Figura 6, indica que pessoas do sexo feminino da faixa etária entre 35 e 39 anos, que moram no Centro e ficam em casa menos de 12h por dia na parte da noite contraíram dengue. A parte do corpo indicada como a de preferência dos mosquitos para atacar seria também a opção “canela/pés”, sendo que esse grupo também percebe que a maioria das picadas ocorre entre 20:00 e 23:59, no entanto o modo de prevenção utilizado é repelente.

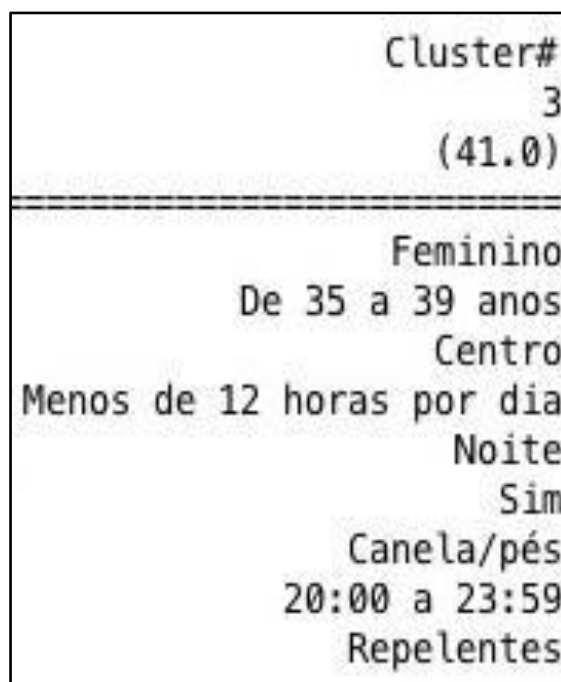


Figura 6 – Cluster #3 gerado pelo aplicativo SimpleKmeans

Há a possibilidade de ser elaborada uma relação entre as vítimas do mosquito da dengue e seu local de trabalho, pois em ambos os clusters indicam que os indivíduos ficam fora de casa mais de 12h por dia. Porém, essa informação pode ser contestada, visto que o Centro é um bairro de grande densidade populacional, tanto residencial como comercial, ou seja, muitas pessoas moram e trabalham no Centro.

Nota-se também que, nos cluster associados às pessoas que contraíram da dengue, o horário percebido das picadas foi entre as 20:00 e 23:59, contrariando o que costuma ser apontado por conhecimento popular, de que o mosquito *Aedes Aegypti* tem hábitos diurnos.

A magnitude da atual epidemia de dengue em grande parte do território nacional permite inferir que as medidas que visavam o combate ao vetor, tais como campanhas de informação e de mobilização de pessoas, fortalecimento da vigilância epidemiológica e entomológica, trabalho de campo, atuação multissetorial e utilização de instrumentos legais, não surtiram o efeito desejado.

Explicações para o relativo fracasso dessas medidas são diversas, desde pouca colaboração da população na prevenção e eliminação de criadouros do mosquito e ocupação desordenada do espaço urbano, até mudanças de hábito. Além disso, incluem-se insuficiência de campanhas oficiais e recursos humanos capacitados para as ações de campo, além da diminuição do

investimento nas ações de prevenção de forma geral.

A melhoria das condições higiênico-sanitárias das habitações, coleta apropriada do lixo e fornecimento de água encanada são medidas também importantes no controle da dengue.

4. Considerações Finais

A aplicação do KDD utilizando o software WEKA 3.8.1 sobre a base de dados meteorológicos mostrou-se uma ferramenta útil para auxiliar na descoberta de conhecimentos relevantes sobre o comportamento do mosquito *Aedes Aegypti*, que podem ser utilizados para futuras ações de prevenção à doença.

Segundo as análises realizadas, pode-se afirmar que há alguns indicativos quanto ao padrão do comportamento do mosquito *Aedes Aegypti*. Observou-se que a maioria das pessoas vítimas da dengue eram mulheres de uma faixa etária mais jovem (entre 20 e 39 anos). Foi percebido também que o Centro foi o bairro que apareceu em ambos os Clusters, bem como o a parte do corpo que sofreu a picada “canela/pés”. Ambos os grupos indicam uma relação com pessoas que ficam menos de 12h por dia em casa, sempre no período da noite, entre 20:00 e 23:59.

Como proposta para estudos futuros, sugere-se aumentar o número de dados coletados para que se tenha um conhecimento mais real das condições da região nos outros meses do ano, para determinar se há sazonalidade no período de predominância de picadas e/ou na parte do corpo na qual esta ocorre. Isso porque, para compreender uma doença multicausal como a dengue, são necessários mais elementos explicativos, relativos não apenas às características da população, como também do ambiente urbano.

Referências

AGRAWAL, R.; IMIELINSKI, T.; SWAMI A. *Mining Association Rules between Sets of Items in Large Databases*. 1993. Disponível em: <<http://arbor.ee.ntu.edu.tw/~chyun/dmpaper/agrama93.pdf>>. Acesso em: 20/07/2017.

AMO, S. *Técnicas de Mineração de Dados*. Universidade Federal de Uberlândia, Faculdade de Computação. 2003.

BHATT, S. et. al. *The global distribution and burden of dengue*. Nature Vol 496, n. 7446, p 504-507, 2013.

BOTHOREL, G.; SERRURIER, M.; HURTER, C. *Utilisation d'outils de Visual Data Mining pour l'exploration d'un ensemble de règles d'association*. Sophia Antipolis, France: Ihm'11, 2011.

CAROSIA, A.E.; CIFERRI, C.D.A. *Mineração de Dados*. Disponível em <<http://wiki.icmc.usp.br/images/3/3c/BDAMineracao.pdf>>. Acesso em: 19/07/2017.

DONALÍSIO, M.R.; GLASSER, C.M. *Vigilância entomológica e controle de vetores do dengue*. Rev bras epidemiol Vol 5, n. 3. P 259-279, 2002.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. *From Data Mining to Knowledge Discovery in Databases*. AI magazine. P. 37-54, 1996

GUZMAN, M.G. et. al.. *Dengue: a continuing global threat*. Nat Rev Microbiol Vol 8, n. 12, p 7-16, 2010.

NASCIMENTO, S.Q. *Utilização da Mineração de Dados para Análise da Tonalidade Musical em Diversos Estilos Musicais*. XIX SIMPEP. 2012.

RIGAU-PÉREZ, J.G. *Surveillance for an emerging disease: dengue hemorrhagic fever in Puerto Rico, 1988-1997*. P R Health Sci J Vol 18, n. 4, p 337-345, 1999.

SALVADOR, H.G.; CUNHA, A.M.; CORRÊA, C.S. *Vedalogic: um método de Verificação de Dados Climatológicos Apoiado em Modelos Minerados*. Revista brasileira de Meteorologia Vol 24, n. 4, p 448-460, 2009.

TAN, P.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining – Mineração de Dados*. Rio de Janeiro: Editora Ciência Moderna Ltda. 2009.

WORLD HEALTH ORGANIZATION. *Dengue: guidelines for diagnosis, treatment, prevention and control*. Geneva: World Health Organization, 2009.



VII CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO

Ponta Grossa, PR, Brasil, 06 a 08 de dezembro de 2017