

DESCOBERTA DE CONHECIMENTO EM DADOS PÚBLICOS: MODELO BUZETI APLICADO AOS SISTEMAS ALICEWEB E DATASUS

Luiz Buzeti¹, Antonio A.C. Pereira², Flávio Bortolozzi³, Marcelo P. Bernunci⁴, Iara C. de Almeida⁵

Abstract. *The present study aimed to contribute to improve the management of health knowledge, allowing the discovery of knowledge in public data of the ALICEWeb and DATASUS systems. Initially, the existing methodologies that enabled the discovery of knowledge were studied, and the Buzeti model was defined. Following the steps proposed by the Buzeti model, were performed: selection of public data, pre-processing and cleaning of data, transformation of data, data mining and, finally, analysis and cross-referencing of DATASUS data with ALICEWeb system data, with the objective of finding correlations between morbidities and drug importation. The greatest contribution of this study is the understanding of the use of a KDD methodology, that allows the creation of new knowledge from public data.*

Keywords: *data mining; knowledge discovery in databases; public data; ALICEWeb system; DATASUS system.*

Resumo. *O presente estudo objetivou contribuir para melhorar a gestão do conhecimento em saúde, permitindo descoberta de conhecimento em dados públicos dos sistemas ALICEWeb e DATASUS. Inicialmente foram estudadas as metodologias existentes que viabilizassem a descoberta do conhecimento, sendo então definida o modelo Buzeti. Na sequência foram realizadas as etapas propostas pelo modelo Buzeti: seleção dos dados públicos, pré-processamento e limpeza dos dados, transformação destes dados, mineração de dados e, finalmente, análise e cruzamento dos dados do sistema DATASUS com os dados do sistema ALICEWeb, com objetivo de encontrar correlações entre morbidades e importação de fármacos. A grande contribuição desse estudo é a compreensão do uso de uma metodologia KDD que permite a criação de novo conhecimento a partir de dados públicos.*

Palavras-chave: *mineração de dados; descoberta de conhecimento em banco de dados; dados públicos; sistema ALICEWeb; sistema DATASUS.*

¹ Analista de Sistemas; Universidade Estadual de Londrina, Londrina, Paraná; buzeti@uel.br.

² Acadêmico do curso de Medicina da Unicesumar, Maringá, Paraná; antonio-augusto_@hotmail.com

³ Bolsista do Programa Produtividade em Pesquisa do Instituto Cesumar de Ciência, Tecnologia e Inovação - ICETI. Programa de Pós-Graduação em Gestão do Conhecimento da Unicesumar, Maringá, Paraná. flavio.bortolozzi@unicesumar.edu.br.

⁴ Bolsista do Programa Produtividade em Pesquisa do Instituto Cesumar de Ciência, Tecnologia e Inovação - ICETI. Programa de Pós-Graduação em Promoção da Saúde da Unicesumar, Maringá, Paraná; marcelo.bernuci@unicesumar.edu.br

⁵ Bolsista do Programa Produtividade em Pesquisa do Instituto Cesumar de Ciência, Tecnologia e Inovação - ICETI. Programa de Pós-Graduação em Gestão do Conhecimento da Unicesumar, Maringá, Paraná. iara.almeida@unicesumar.edu.br.

1 INTRODUÇÃO

Ao longo das mudanças econômicas mundiais, os tradicionais fatores de produção - trabalho, capital e terra - vêm adquirindo importância secundária. Em oposição, o fator conhecimento têm-se tornado, cada vez mais, no principal recurso das organizações (BARR, 1981). Segundo Paton et al. (2015), a informação e o conhecimento são as armas mais competitivas para as empresas, e o acúmulo do conhecimento de qualidade proporciona um diferencial competitivo dentro do mercado em que atuam. A Gestão do Conhecimento (GC) tem como objetivo auxiliar as organizações no trato deste ativo. Conforme KIMIZ (2005, pg. 3), a GC é a coordenação deliberada e sistemática das pessoas, tecnologia, processos e estrutura organizacional para agregar valor através da reutilização e inovação. Esta coordenação é conseguida através da criação, compartilhamento e aplicação do conhecimento, bem como através da alimentação das valiosas lições aprendidas e as melhores práticas para a memória corporativa, a fim de promover a aprendizagem organizacional continuada.

Com tamanha importância para o mundo dos negócios, a necessidade de gerir conhecimento é óbvia e de característica complexa. De acordo com Davila et al. (2014), a complexidade da GC pode ser observada a partir dos ciclos que a formam que são: a captura e a criação, o compartilhamento e a disseminação, e a aquisição e a aplicação do conhecimento na organização. Importante ressaltar que - se esses ciclos forem devidamente implantados e maturados dentro de uma organização - pode-se fomentar a excelência no negócio pois o conhecimento registrado e compartilhado permite tanto maior eficiência e produtividade quanto prevenção de erros já cometidos, como também auxilia um processo de inovação.

A busca pela inovação e a consequente incorporação do conhecimento nos produtos gerados, classifica os setores industriais em quatro níveis de intensidade tecnológica: alta, meio-alta, meio-baixa e baixa. Uma das formas de identificar tais segmentos é através da análise da média de gasto em Pesquisa e Desenvolvimento sobre o faturamento. Esse gasto é um dos parâmetros que define a intensidade tecnológica de cada setor industrial, pois aumenta o nível de conhecimento incorporado aos produtos (ZAWISLAK et al., 2013). A indústria farmacêutica é classificada no segmento de alta intensidade tecnológica justamente por ter as características exigidas. Esse esforço inovador acaba resultando em externalidades do conhecimento, tanto tácito como explícito. Essas externalidades, através da produção de novos conhecimentos, resultam no desenvolvimento do país e entrada de divisas financeiras.

No entanto, no caso brasileiro, observa-se um déficit contínuo no comércio de produtos farmacêuticos entre o Brasil e o mundo. Nos dados referentes às transações exteriores (importação/exportação) de fármacos, observa-se que, no ano de 2005, o Brasil importou 3,31 bilhões de dólares. Já no ano de 2011 houve um salto, e o Brasil importou 9,50 bilhões de dólares (INTERFARMA, 2012). Entretanto, esse fenômeno não pode ser observado apenas pelo viés financeiro, mas também pelo que produz de bem-estar social.

Apesar desse reconhecimento internacional pela grande quantidade de dados brutos, o Sistema DATASUS pode ser enquadrado no que Han e Kamber (2006) descrevem como sendo “rico em dados, pobre em conhecimento”. Através da análise desse conjunto de dados, a GC (por meio do ciclo de captura e criação do conhecimento) pode contribuir na melhora da gestão da saúde pública brasileira. Essa criação e captura de novos conhecimentos podem auxiliar os gestores da saúde nos processos de inovação, planejamento e tomada de decisão.

Conforme (KIMIZ, 2005), é possível extrair, codificar e organizar novos conhecimentos de forma sistemática com a utilização de ferramentas e técnicas da GC. Neste nosso estudo, pode-se dizer que é possível utilizar ferramentas que analisem a relação entre variáveis (isto é, medicamentos, interações entre outros) dentro de base de dados de grande volume, com o propósito de extrair novos conhecimentos. Assim, o desafio em prestar serviços de saúde de qualidade à população brasileira pode ter a GC como aliada.

Com o desenvolvimento das tecnologias da informação e informação (TIC), cada vez mais o processo de extração de conhecimento de grandes volumes de dados e a sua complexidade pode ter seu custo muito elevado, dependendo das características do problema e do que se pretende obter. Nas organizações, a quantidade de dados gerada tem superado a capacidade humana de interpretar e compreender tanta informação em tempo hábil.

Atualmente, existe tecnologia para extração de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD), é uma área interdisciplinar que envolve as áreas de Bancos de Dados, Inteligência Artificial (Redes Neurais, Lógica Fuzzy, etc.) e Estatística, entre outras (FAYYAD et al., 1996). Logo, a partir de grandes quantidade de dados e informações, nas mais diversas áreas do conhecimento, gestores e profissionais da área de saúde envolvidos em tomada de decisão tem agora possibilidade para otimizar seus processos. Mais especificamente, pode-se integrar os dados oriundos das fontes públicas de saúde dos Sistemas ALICEWeb e DATASUS e assim ter uma melhor compreensão de

variáveis envolvidas. Neste estudo pretendemos verificar a relação entre ‘importação de fármacos’ e ‘índices de morbidade’ da população brasileira.

2 PREMISSAS CONCEITUAIS

No decorrer desta pesquisa foi feita uma análise dos principais conceitos; definições e técnicas utilizadas ao longo de todo o seu desenvolvimento. Destacamos: as práticas de descobertas de conhecimento em banco de dados; a mineração de dados; big data; a gestão do conhecimento; as técnicas de avaliação do site DATASUS para a obtenção dos dados; a avaliação do site ALICEWeb e conceitos estatísticos.

Levando-se em consideração esse cenário, Silva e Breternitz (2013) definem *Big Data* como sendo: “um conjunto de tendências tecnológicas que permite uma nova abordagem para o tratamento e entendimento de grandes conjuntos de dados para fins de tomada de decisões.” Conforme (Rezende, 2003), *Knowledge Discovery in Databases* (KDD) é constituído pelas seguintes etapas:

1. Seleção de dados - o conjunto de dados é escolhido com suas variáveis, atributos e registros. As fontes para esta seleção pode vir de diferentes formatos, tais como data warehouses, planilhas ou sistemas de informação;
2. Pré-processamento e limpeza dos dados - o objetivo é garantir a qualidade dos dados, eliminando dados redundantes e/ou inconsistentes, recuperando dados incompletos e avaliando dados discrepantes (*outliers*);
3. Transformação dos dados - os dados são formatados e armazenados de forma correta;
4. Data mining - permite que haja a descoberta de novos conhecimentos;
5. Interpretação e avaliação - o conhecimento gerado é avaliado por especialistas. De acordo com Prass (2007), o processo pode retornar para uma das fases anteriores se essa avaliação considerar que o conhecimento gerado não é adequado.

Conforme Lemos et al. (2005), o *Data Mining* utiliza-se de conceitos tanto da área de estatística quanto da área de Inteligência Artificial, mais especificamente da subárea Aprendizado de Máquina (AM). Segundo Michalski et al. (2013), o AM pretende que um

programa aprenda e melhore o seu desempenho dada a experiência adquirida em uma determinada prática.

Pinheiro e seus colaboradores (2016) afirmam que os Sistemas de Informações em Saúde "ainda não atingiram todo o seu potencial, pois são utilizados de maneira incipiente pela gestão da saúde para o processo decisório". Sabe-se que gestão eficaz do conhecimento pode ser feita através do uso da tecnologia da informação. Além disto, grande volume de dados e informações disponíveis podem se tornar um grande aliado para os gestores, pois estes são convertidos em conhecimento, ativo essencial para melhoria dos sistemas e das organizações.

O sistema ALICEWeb do Ministério do Desenvolvimento, Indústria e Comércio Exterior, foi desenvolvido de forma a modernizar as formas de acesso às estatísticas brasileiras de exportações e importações brasileiras. Este sistema é atualizado mensalmente com dados do mês anterior e as fontes de dados são oriundas do Ministério do Desenvolvimento, Indústria e Comércio Exterior. Assim, toda transação comercial entre Brasil e o outros países do mundo fica devidamente registrado no Ministério, mais precisamente no sistema ALICEWeb.

Já o Departamento de Informática do SUS (DATASUS) permite acesso às bases de informações em saúde entre outros. Atualmente, os sistemas do DATASUS são considerados pela Organização Mundial da Saúde (OMS) e Organização Pan-Americana de Saúde (OPAS), um dos mais completos do mundo (LIMA et al., 2015). Por meio do site oficial do DATASUS (2017b), o Ministério da Saúde disponibiliza dados para subsidiar a análises da situação sanitária, para a tomadas de decisão baseadas em evidências e para a elaboração de programas de ações de saúde. Os dados disponíveis são organizados e classificados na forma de Morbidade (ocorrências que levam às internações), Incapacidade (ocorrências que levam à incapacidade), entre outros. Neste estudo serão utilizados os dados de morbilidade.

3 METODOLOGIA

A metodologia utilizada na elaboração deste estudo foi de natureza aplicada. O procedimento foi exploratório com abordagem quantitativa, para fins descritivos, através de fontes secundárias de dados. Realizou-se também procedimentos de pesquisa bibliográfica e experimental, aplicando o estudo de caso nos sistemas ALICEWeb e DATASUS.

4 RESULTADOS OBTIDOS

O objetivo geral deste estudo foi a criação do conhecimento utilizando técnicas do KDD e tendo como fonte de dados os sistemas DATASUS e ALICEWeb. Para tal, foram elaboradas as seguintes questões de pesquisas:

I - Quais são as premissas e os conceitos de Gestão do Conhecimento e tecnológicos que serão necessários para fundamentar a pesquisa?

II - Como propor um modelo para solucionar o problema proposto?

III - Como cruzar e analisar as informações nos dois sistemas (DATASUS x ALICEWEB).

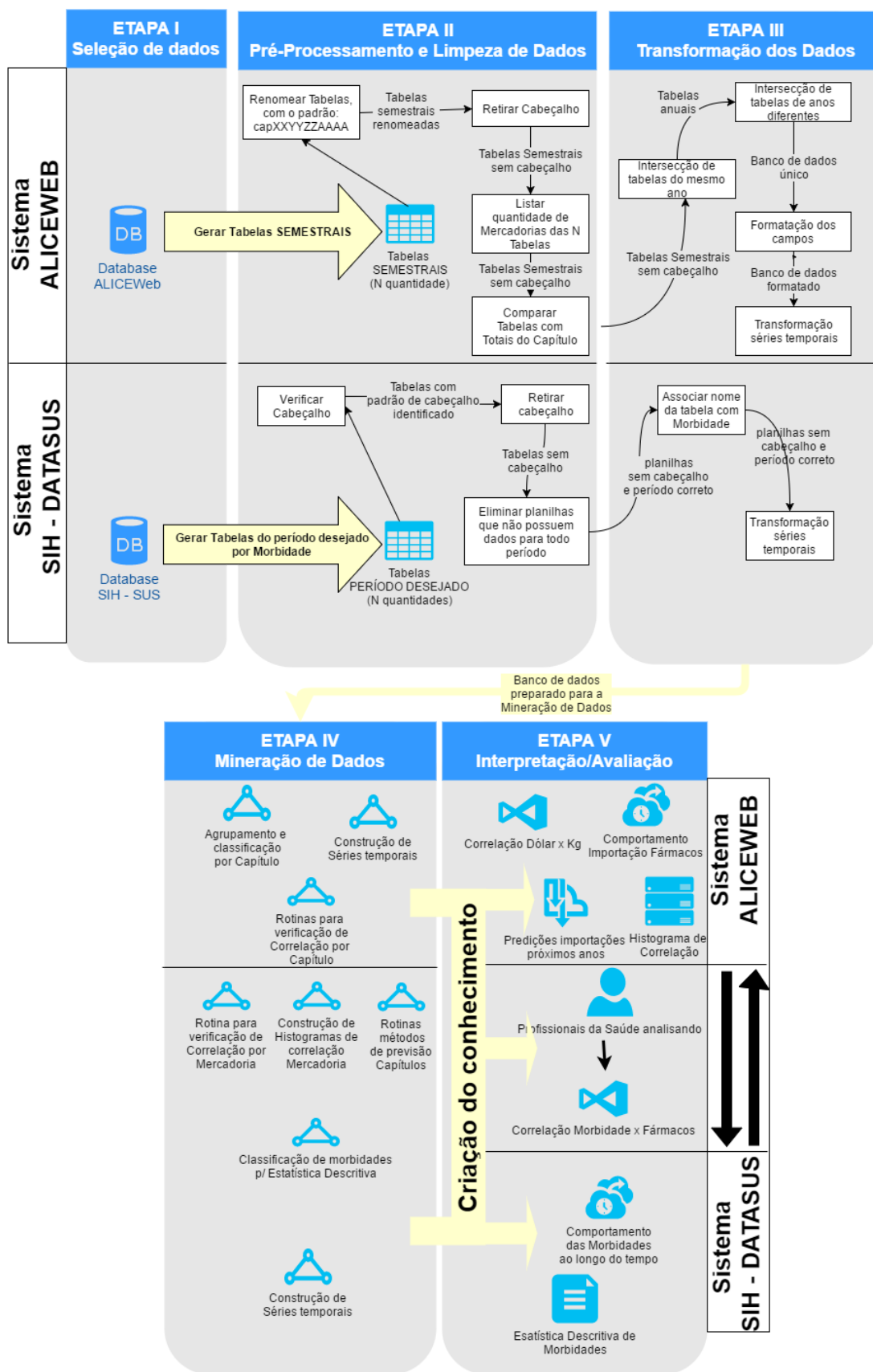
Para responder a questão de pesquisa I, foram realizadas buscas em bases de dados eletrônicas, tanto brasileiras quanto internacionais, utilizando as palavras ‘gestão do conhecimento’, ‘datasus’, ‘mineração de dados’, ‘*knowledge management*’, ‘*data mining*’ no período de 2009 até 2015. Os motores de busca utilizados foram o Google Acadêmico, LILACS e MEDLINE, com um total de 106 estudos retornados. Após, foi feita a filtragem desses estudos, o critério é deveriam fazer a mineração de dados públicos de saúde.

Para responder às questões de pesquisa II, foi proposto um Modelo de KDD baseado em Fayyad et al. (1996), que denominamos Modelo Buzeti. O modelo Buzeti possui 5 etapas, que correspondem às fases de descoberta de conhecimento: seleção de dados, pré-processamento e limpeza de dados, transformação dos dados, e mineração de dados e interpretação/Avaliação (conforme apresentado na Figura 1). Na sequência, serão apresentadas as quatro primeiras fases, aplicadas tanto para o Sistema ALICEWeb quanto para o Sistema DATASUS e, após, será apresentada a avaliação do cruzamento dos dados advindos dos dois sistemas.

4.1 SELEÇÃO DE DADOS

O objetivo do sistema ALICEWeb é retirar os dados do Sistema Integrado de Comércio Exterior (SISCOMEX), no qual os próprios exportadores-importadores fornecem as informações relativas às operações. Para o Sistema ALICEWeb, foram explorados os dados do capítulo 28 denominado como ‘Produtos químicos inorgânicos, compostos inorgânicos ou orgânicos de metais preciosos, de elementos radioativos, de metais das terras raras ou de isótopos’.

Figura 1 – Modelo Buzeti de Mineração



Para o Sistema DATASUS, o processo de seleção de dados foi mais complexo pois foi necessário a aplicação de filtros para selecionar os dados. Além disto, foi necessário acessar o aplicativo TABNET (2017), um tabulador genérico de domínio público que permite gerar informações das bases de dados do Sistema Único de Saúde. Para acessar o sistema, deve-se acessar o site oficial do DATASUS (DATASUS, 2017b), escolher a opção ‘informações de Saúde TABNET’. Após, escolher a opção ‘Epidemiológicas e Morbidades’ e depois a opção ‘Morbidade Hospitalar do SUS (SIH/SUS)’. Finalmente, deve-se escolher, na sequência, as opções ‘Geral, por local de internação, a partir de 2008’ e ‘Abrangência geográfica: Brasil por município’.

Salienta-se que, na Figura 2, o filtro 1 representa a abrangência das internações, ou seja, no Brasil (em todo território nacional); o filtro 2 é uma série temporal e, como tal, possui os meses e anos da observação; o filtro 3 representa os campos nos quais o usuário tem interesse; o filtro 4 representa o período de janeiro de 2009 à junho de 2016; e o o filtro 5 representa a opção das doenças do Capítulo CID-10. Após esta filtragem, foram geradas 312 tabelas com as morbididades e suas respectivas internações, óbitos, valores etc.

4.2 PRÉ-PROCESSAMENTO E LIMPEZA DE DADOS

Para o Sistema ALICEWeb, por meio da opção filtrar pelo Capítulo 28 foram geradas 42 tabelas diferentes. As características nessa opção é que o sistema permite no máximo 6 períodos por consulta, assim, como o estudo terá um período de abrangência de 84 meses, foi necessária a repetição da consulta por 14 vezes, o que resultou em 42 consultas/tabelas. Após o recebimento dos 42 e-mails, foi necessário salvar as 42 tabelas que estavam em anexo. Para uma melhor identificação e também para facilitar o processo de agrupamento de todas as tabelas em um único banco de dados, foi adotado o seguinte padrão para nomeação das tabelas - "capXXYYZZAAAA", onde XX é o número do Capítulo, YY é o mês inicial, ZZ é o mês final e AAAA é o ano.

Já para o Sistema DATASUS, foram geradas 312 planilhas com extensão .CSV, sendo que cada uma das planilhas representa uma morbidade do CID - 10. Essas planilhas exigem menos pré-processamento que o ALICEWeb mas apresentaram problemas de formatação, tais como os cabeçalhos que não apresentam um padrão para identificar a morbidade.

4.3 TRANSFORMAÇÃO DOS DADOS

Para o Sistema ALICEWeb, dada a diferença entre os semestres, optou-se por manter apenas as mercadorias que pertenciam à intersecção das tabelas do primeiro e do segundo semestre. Por exemplo, para o processo sobre as tabelas "cap2801062009" e "cap2807122009". A coluna "Mercadorias Originais" exibe as quantidades originais geradas pelo ALICEWeb: 339 mercadorias para o primeiro semestre; 344, para o segundo. Fazendo a intersecção destas tabelas, foram eliminadas 12 mercadorias em "cap2801062009" e 17 registros na tabela "cap2807122009". Já para o Sistema DATASUS, um total de 32 planilhas apresentaram problemas de falta de meses ou de possuir inconsistência nos dados, portanto foram eliminadas.

4.4 MINERAÇÃO DE DADOS

No Sistema ALICEWeb, foram selecionadas e aplicadas várias técnicas de modelagem utilizando o software R. Inicialmente foi utilizada a estatística descritiva para a comparação entre os capítulos do NCM. Após, foram realizadas técnicas de análise de séries temporais com linha de tendência e de previsão para gastos em dólar por meio de treino e teste dos dados. Nesta fase, os resultados foram comparados/avaliados de forma a encontrar os melhores modelos.

No Sistema DATASUS foi utilizada a estatística descritiva para a comparação entre as morbidades com maior número de internações, gastos e óbitos. Foram analisados, por meio de séries temporais, os comportamentos das morbidades que mais se destacaram na análise anterior.

Em ambos os sistemas, foram considerados modelos que resultaram na criação de novos conhecimentos que possam contribuir para a gestão em saúde. Contudo, as estruturas de dados não foram aglutinadas em um único banco de dados.

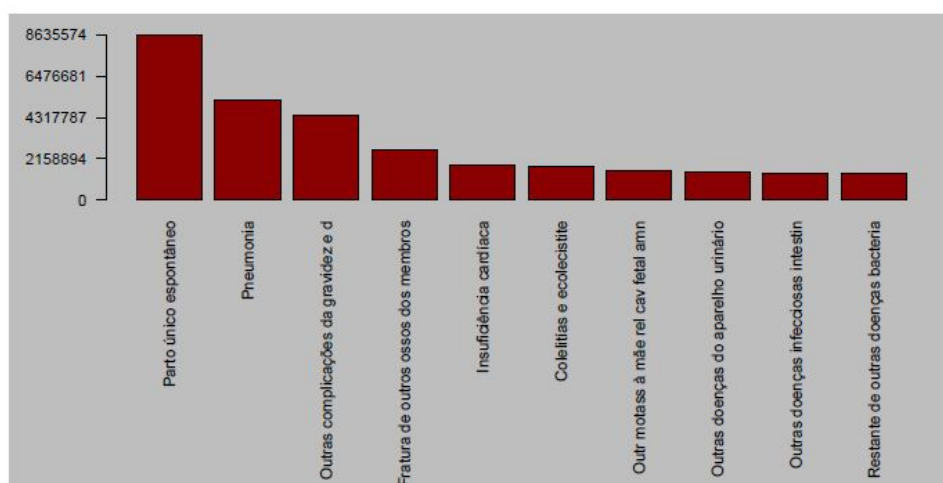
4.5 ANÁLISE E INTERPRETAÇÃO DAS INFORMAÇÕES DO SISTEMA DATASUS

Nesta fase, procurou-se determinar a relação entre as variáveis internações e óbitos ocorridos em função das internações do período de Jan/2009 até Jun/2016. Estas variáveis

foram analisadas tanto separadamente quanto em conjunto. Primeiro foram analisados quais são as morbidades que mais geraram internações pelo SUS.

Observe no gráfico da Figura 2 que a morbidade com maior número de internações é o ‘Parto’, seguida pela ‘Pneumonia’. As morbidades ‘Outras complicações da gravidez’ e ‘fratura de outros ossos do membro’ também estão em um nível maior. Já entre as morbidades ‘insuficiência cardíaca’ e ‘Restante de outras doenças de bactéria’ apresentam números muito próximos das internações.

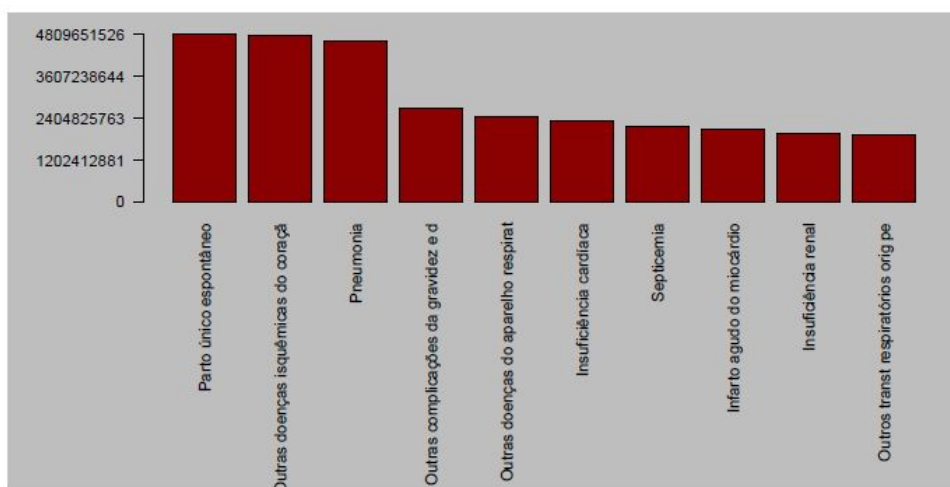
Figura 2 – 10 Morbidades com maior número de internações



Fonte: Autores

Já no gráfico da Figura 3 pode-se analisar quais são as morbidades que mais geram gastos a partir de suas internações: o ‘Parto’ também é o de maior gasto e as ‘Outras doenças isquêmicas do coração’, que não aparece entre as 10 maiores em casos, aparece como a segunda em gastos pelo SUS.

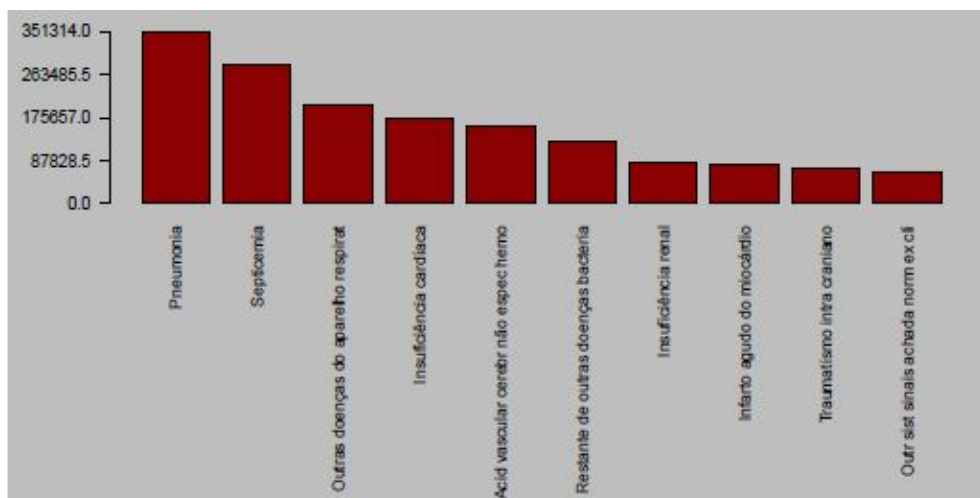
Figura 3 – Morbidades com maiores gastos em reais



Fonte: Autores

No gráfico da Figura 4 pode-se verificar quais são as morbidades com maior número de óbitos. A ‘Pneumonia’ que aparece nas primeiras posições na classificação por ‘Gastos’ e na ‘Quantidade de Internações’, aparece também como a morbidade com maior número de óbitos no Brasil, seguido por ‘Septicemia’.

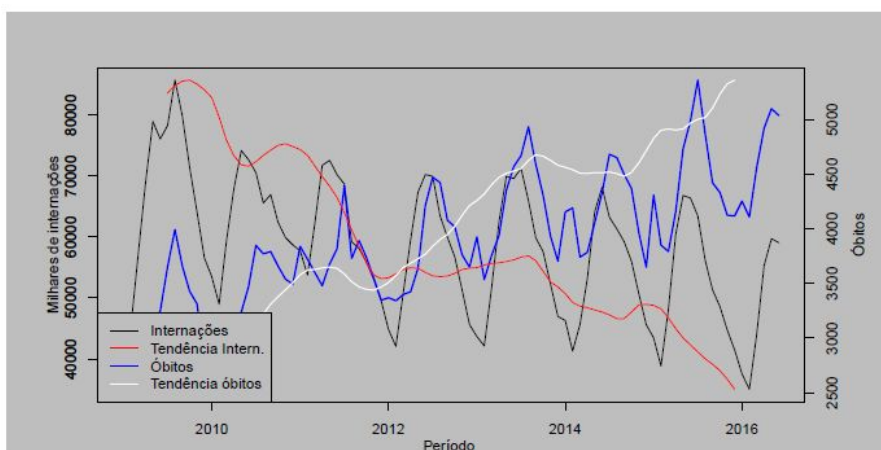
Figura 4 – Morbidades que geram o maior número de óbitos



Fonte: Autores

Na Figura 5, o gráfico apresenta a relação de Internações versus Óbitos para a morbidade pneumonia. É possível também detectar a sazonalidade, isto é diminuiu os casos no verão e aumentam no inverno; as internações estão com tendência de queda, ou seja, uma diminuição de casos (em vermelho); os óbitos estão com tendência de alta, ou seja, com um aumento de casos (em branco).

Figura 5 – Comparativo de Internações e Óbitos de Pneumonia

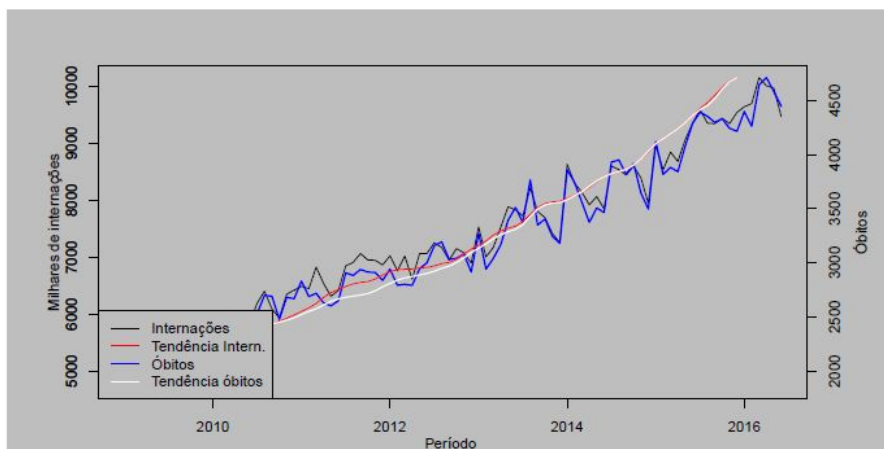


Fonte: Autores

No gráfico da Figura 6 pode-se verificar a ‘Septicemia’, as linhas de cor preta e azul representam, respectivamente, o comportamento das internações e os óbitos ao longo do

período. Além dessas características, apresenta a tendência das internações (cor vermelha) e dos óbitos (cor branca). Note a alta tanto para os casos de internações quanto para os óbitos em função das internações.

Figura 6 – Comparativo de Internações e Óbitos de Septicemia



Fonte: Autores

Através dessas análises feitas nos gráficos, pode-se verificar que a criação do conhecimento a partir das bases de dados do SIH - SUS pode colaborar com a Gestão em Saúde.

5 ANÁLISE E INTERPRETAÇÃO DO CRUZAMENTO DAS INFORMAÇÕES DAS INFORMAÇÕES NOS SISTEMAS DATASUS E ALICEWEB

Para atender à terceira questão de pesquisa, foi realizada a quinta etapa do Modelo Buzeti. Nesta etapa foram feita a análise e interpretação do cruzamento das informações dos sistemas DATASUS e ALICEWeb. O cruzamento resultou em milhares de correlações, as quais foram avaliadas se eram espúrias ou não um por profissional da área da saúde.

A lógica do cruzamento consistiu em verificar, para cada uma das morbidades (em um total de 312), qual o valor da correlação em referência a todas as mercadorias (em um total de 1363). Foram feitas 425.256 análises de correlação entre as variáveis Morbidades e Mercadorias. Note que o comportamento de uma importação pode não refletir em relação às morbidades imediatamente, ou seja, no mesmo mês, portanto considerou-se uma correlação com seis meses de diferença, isto é, comparou-se Morbidades ocorridas em Jun/2009 com importações ocorridas em Jan/2009. Esses cruzamentos produziram um total de 850.512

análises de correlação entre as variáveis Morbidades e Mercadorias. O gráfico da Figura 7 apresenta a morbidade Cólera, pois esta possui correlação significativa com seis fármacos, que estão representados também na Figura 7. Para exemplificar, a Cólera possui, em uma escala de -1 até 1, uma correlação positiva significativa de 0.8117 em relação ao Ácido tioglicólico e seus sais. Os apêndices seguem estes padrões, aparecendo a morbidade no cabeçalho e os fármacos relacionados nas linhas abaixo.

Figura 7 – Morbidade e Fármacos correlacionados

| Morbidade/Fármaco | Correlação |
|--|-------------------|
| Cólera | |
| Silicato de potássio | 0.766850796059747 |
| Nitrobenzeno | 0.787694108118047 |
| Mononitroetano; nitrometanos | 0.795949997172857 |
| Guaifenesina | 0.728151560593157 |
| Ácido tioglicólico e seus sais | 0.811751779757372 |
| Sais de fenciclidina, fenoperidina ou fentanil | 0.769202105537151 |

Fonte: Autores

Com relação aos resultados desses cruzamentos, por meio do auxílio de especialistas da área da Saúde, notou-se que houve correlações significativas, tanto positivas quanto negativas, entre os compostos e as morbidades. No entanto, não foi encontrado, nas duas formas de cruzamentos, correlações que fizessem sentido. Assim, todas as correlações podem ser consideradas espúrias, ou seja, essas correlações são meras coincidências estatísticas.

6 CONCLUSÕES

Neste estudo foram abordadas as fases que formam a metodologia KDD, destacando a implantação de cada fase. Uma primeira contribuição da pesquisa foi a criação do modelo Buzeti que tornou possível a preparação dos dados públicos relacionados à importação de fármacos e de dados públicos sobre a morbidade, a qual possibilitou a sua análise. Este modelo propôs uma formulação teórica e um caminho para a extração de novos conhecimentos a partir dos sistemas ALICEWeb e DATASUS. Assim, esta pesquisa apresentou a complexidade das bases de dados desses sistemas, e desenvolveu um procedimento para a construção de um banco de dados único, a partir da junção de centenas de tabelas geradas pelos dois sistemas.

Como o foco do estudo foi a criação de novos conhecimentos, utilizando as técnicas do KDD, tendo como fonte de dados os sistemas ALICEWeb e DATASUS, destacamos que a quantidade de dados públicos analisados é de grande volume e com considerável complexidade. Portanto como contribuição o estudo comprova que a metodologia utilizada e o Modelo Buzeti proposto é aplicável. Além disso, o cruzamento dos dados dos dois sistemas resultou na criação de conhecimento cuja análise foi feita por profissionais de saúde. Contudo, essa análise demonstrou que não há correlação entre mercadorias importadas e internações por morbidades.

Salientamos que este estudo contribui para futuras pesquisas que necessitem seguir uma metodologia para aquisição de novo conhecimento advindo de dados públicos. Cada uma das fontes de dados, os sistemas ALICEWeb e DATASUS, assim como o cruzamento entre essas fontes, que geraram novos conhecimentos, oferecem mais possibilidades de pesquisas futuras. No sistema ALICEWeb, as Tabelas 26, 28 e 30, em função de não haver uma correlação significativa entre Quilogramas e Dólar, pode-se realizar um estudo específico das principais mercadorias, verificando quais os motivos pelos quais isso poderia estar ocorrendo. Já nas previsões para o ano de 2016, em relação aos gastos em dólar nas importações de mercadorias, faz-se necessário o confronto entre a projeção indicada pelo método escolhido e o que ocorreu com os dados reais de importação.

REFERÊNCIAS

- ALICEWEB. (2016). Recuperado em <http://aliceweb.mdic.gov.br/menu/index/item/metodologia>.
- BARR, A. Feigenbaum. ea, "the handbook of artificial intelligence," volumes i and ii, william kaufmann. Inc., Los Altos, Ca, 1981.
- Cohen, P. R., & Feigenbaum, E. A. (Eds.). (2014). *The handbook of artificial intelligence* (Vol. 3). Butterworth-Heinemann.
- Kimiz, D. (2005). Knowledge management in theory and practice. *McGill University*.
- DATASUS. Tabuladores (2017). Recuperado em <http://www2.datasus.gov.br/DATASUS/index.php?area=0608>>
- Davila, G. A., Fraga, B. D., Diana, J. B., & Spanhol, F. J. (2014). O Ciclo de Gestão do Conhecimento na Prática: um estudo nos núcleos empresariais catarinenses. *International Journal of Knowledge Engineering and Management (IJKEM)*, 3(7), 43-64.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27-34.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

- INTERFARMA. Balanço das políticas industriais para o setor farmacêutico. Interdoc VOLUME III, 2012.
- LEMOS, E. P.; STEINER, M. T. A.; NIEVOLA, J. C. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. *Revista de Administração*, v. 40, n. 3, p. 225–234, 2005.
- Prezepiorski Lemos, E., Arns Steiner, M. T., & Nievola, J. C. (2005). Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. *Revista de Administração-RAUSP*, 40(3).
- Lima, A. C., Januário, M. C., Lima, P. T., & de Moura, W. (2015). DATASUS: o uso dos Sistemas de Informação na Saúde Pública. *REFAS: Revista FATEC Zona Sul*, 1(3), 4.
- Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.
- Paton, C., Mastelari, D. A., Nogueira, J. A., Teixeira, S. A., Martins, V. F., & da Fonseca Veiga, W. (1999). O USO DO BALANCED SCORECARD COMO UM SISTEMA DE GESTÃO ESTRATÉGICA. In *Anais do Congresso Brasileiro de Custos-ABC*.
- Pinheiro, A. L. S., Andrade, K. T. S., Silva, D. D. O., Zacharias, F. C. M., Gomide, M. F. S., & Pinto, I. C. (2016). HEALTH MANAGEMENT: THE USE OF INFORMATION SYSTEMS AND KNOWLEDGE SHARING FOR THE DECISION MAKING PROCESS. *Texto & Contexto-Enfermagem*, 25(3).
- Prass, F. S. KDD–UMA VISAL GERAL DO PROCESSO., 2007.
- Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.
- Breternitz, V. J., & Silva, L. A. (2013). Big data: Um novo conceito gerando oportunidades e desafios. *Revista Eletrônica de Tecnologia e Cultura*, 2(2).
- Zawislak, P. A., Fracasso, E. M., & Tello-Gamarra, J. (2013). Intensidade Tecnológica e Capacidade de Inovação de Firms Industriais. *Proceedings of the ALTEC*.