ATAS DE SAÚDE AMBIENTAL - ASA
Volume 3, número 1
Jan/Abr. – 2015
ISSN: 2357-7614

O USO DO *DATA MINING* NA PROMOÇÃO DE SAÚDE

Alice Monique Pacheco Souza[1*], José Eduardo Zaia[2]

RESUMO - O data mining (mineração de dados) é uma das etapas do processo Knowledge Discovery in Database, que tornou se a ferramenta mais conhecida do mesmo, pois sua metodologia visa a preparação e exploração dos dados, interpretação dos resultados e a percepção dos conhecimentos minerados. Diante o crescente número de dados na área da saúde, o data mining pode ser uma ferramenta de grande importância na extração de conhecimento dos dados e dessa forma poderá auxiliar os gestores de saúde nas tomadas de decisões voltadas a prevenção e promoção da saúde.

Palavras-chave: mineração de dados, promoção da saúde

The use of data mining in health promotion

ABSTRACT - Data mining is one of the steps of the process of Knowledge Discovery in Databases, which became the most well-known tool of this process, as it has the aim of preparing and exploring data; interpreting results and providing a perception of the mined data. Given the growing knowledge in healthcare, data mining can be a very important tool in knowledge discovery and may aid heath manager decision making in prevention and health promotion**.**

Keywords: data mining, health promotion

[1]Médica veterinária, mestre em Promoção de Saúde, Mestrado em andamento em Promoção de Saúde pela Universidade de Franca. *Autor para correspondência; e-mail: souzaalice20@yahoo.com
[2] Doutor em Ciências Biológicas pela Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Professor pesquisador da Universidade de Franca. Endereço para contato: Universidade de Franca, Universidade de Franca.

INTRODUÇÃO

A evolução dos sistemas computacionais e a facilidade de aquisição de hardwares nas últimas décadas impulsionaram o armazenamento de dados e essa prática cresce a cada dia, proporcionalmente ao desenvolvimento de novas e mais complexas estruturas de armazenamento, como: *Data Warehouses*, Bibliotecas Virtuais, entre outros (Cunha et al. 2014). Entretanto, as diferentes técnicas de gerenciamento disponíveis para análise dos dados com o objetivo de extrair conhecimento, tornaram-se inadequadas considerando o enorme número de dados armazenados, portanto, centena de milhões de reais, gastos com armazenamento dos mesmos, torna-se inviável pois apenas dados não geram conhecimento (Ebecken et al. 2007). Com isso o *Data Mining (*mineração de dados), proposto na década de 80, se tornou uma das tecnologias mais promissoras na busca pelo conhecimento. Muito utilizada em *marketing*, finanças e manufatura. Aclamada como uma das principais tecnologias para um futuro próximo, é considerada o ponto mais alto na busca de conhecimento para tomada de decisões, levando outras áreas, como por exemplo a da saúde a utilizar essa técnica (Côrtes et al. 2002).

Diversas definições para Mineração de Dados podem ser encontradas na literatura, tais como:

"É análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados" (Hand et al. 2001).

"Busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelos homens" (Weis; Indurkhya 1999).

"Processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, frequentemente desconhecidos, a partir de grandes quantidades de dados armazenados em banco de dados" (Braga et al. 2007).

Data mining ou mineração de dados é apenas parte do processo de Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database – KDD*) que é constituído por diferentes etapas, a primeira delas é definir quais os resultados deseja-se obter, após a escolha dos dados é necessário que os mesmos passem por uma preparação, que envolve as fases de limpeza, integração, seleção e transformação, que são denominadas análise descritiva ou análise prévia que consiste na observação dos dados com a finalidade de localizar inconsistências, incoerências ou erros de preenchimento. A descrição dos dados pode ser aplicada em uma análise com o objetivo de torna-la mais clara e por fim é realizada a sumarização dos mesmos para facilitar o entendimento (Côrtes et al. 2002).

As técnicas de mineração de dados devem ser escolhidas de acordo com os objetivos a serem alcançados, de forma a obter resultados mais precisos. Basicamente existem duas metodologias para aplicação dessas técnicas, que são a abordagem *top-down* (teste de

hipótese) que testa uma hipótese pré-estabelecida a fim de comprova-la ou refuta-la e a abordagem *botton-up* (busca de conhecimento) que consiste em um processo de exploração de dados na tentativa de obter alguma descoberta (Côrtes et al. 2002).

O processo de *data mining* envolve uma série de etapas, desde a preparação dos dados, a descoberta de padrões até a avaliação do quanto esses padrões agregam valor aos conhecimentos do pesquisador sobre o problema em questão**.** Diferentes técnicas podem ser utilizadas para alcançar os resultados pretendidos, entretanto, cada técnica possui sua característica e exige profissionais capacitados a interpretar seus resultados (Hand et al. 2001). As mais citadas na literatura são:

- Árvores de decisão: É um fluxograma semelhante a uma estrutura de árvore e cada "parte" dessa árvore representa as variáveis, os resultados obtidos e a distribuição desses dados (Côrtes et al. 2002).
- Regras de associação: Expõe características e tendências dos dados estudados, através de um processo de interconexão dos mesmos. Entende-se que a presença de um item implica necessariamente na existência do outro (Hand et al. 2001).
- Redes Neurais: São programas computacionais que implementam detecções sofisticadas de padrões e algoritmos, para construir modelos principalmente de grandes bancos de dados históricos (Côrtes et al. 2002).
- Algoritmos genéticos: São algoritmos de otimização e busca que trabalham com um conjunto de possíveis soluções simultaneamente. Tem alta capacidade de solucionar problemas em paralelo e fornece uma poderosa ferramenta para mineração de dados (Côrtes et al. 2002).

A finalidade da mineração de dados é encontrar regras de associação entre as variáveis, assim como a busca de padrões e a extração de conhecimento dos dados analisados, ou seja, mineração de dados utilizados para descoberta de informações. Dentre as diferentes técnicas citadas algumas são mais descritas na literatura, tais como: Árvores de decisão e Redes Neurais (Hand et al. 2001).

A qualidade e complexidade de informações na área da saúde e o tempo limitado dos profissionais têm determinado a necessidade do desenvolvimento de processos que propiciem percursos mais sucintos até os resultados provenientes da pesquisa, portanto a revisão sistemática é um recurso importante na prática baseada em evidências (Galvão et al. 2004).

Revisão sistemática é uma forma de pesquisa que utiliza como fonte de dados a literatura sobre um tema estabelecido, no qual os estudos selecionados são avaliados com rigor metodológico, os métodos e os resultados dos mesmos são coletados, categorizados, avaliados, sintetizados e interpretados, com a finalidade de averiguar se são suficientemente válidos para serem considerados (Galvão et al. 2004, Sampaio; Mancini 2007).

A promoção da saúde demanda uma ação coordenada entre governo, setor saúde e outros setores sociais e econômicos, organizações voluntárias e não governamentais, autoridades locais, indústria e mídia. As pessoas, em todas as esferas da vida, devem

envolver-se neste processo como indivíduos, famílias e comunidades. As estratégias e programas na área da promoção da saúde devem se adaptar as necessidades locais e as possibilidades de cada país e região, bem como levar em conta diferenças em seus sistemas sociais, culturais e econômicos (Otawa 1986).

Atualmente o Brasil enfrenta o desafio de desenvolver novos modelos de gestão para a saúde, que conciliem recursos finitos para uma demanda crescente de recursos. De forma que os dados referentes aos atendimentos dos usuários possam ser observados e prontamente analisados pelos gestores, permitindo um melhor acompanhamento e gestão dos recursos disponíveis aos usuários (Miranda 2003).

Desde a metade do século XX, as ciências industriais vêm se desenvolvendo, com o objetivo de assessorar o homem na recuperação de dados, na elaboração de informações e na descoberta de conhecimentos que orientem as tomadas de decisões, a partir da identificação de padrões ou associações (Rezende 2005).

O uso de mineração de dados e análise epidemiológica, nos dados da saúde, proporcionaria o mapeamento da situação da saúde, assim como a produção de estatísticas de acordo com a incidência e prevalência de doenças e riscos a saúde, tanto atual, quanto futura dos usuários, baseado nessas estatísticas, os gestores seriam capacitados a tomar decisões relacionadas aos serviços a serem prestados a população, de acordo com suas reais necessidades, e ainda possibilitaria promover ações em promoção da saúde (Kobus et al. 2006).

Considerando a vasta utilização do *data mining* em diferentes áreas, esse artigo objetiva apresentar um panorama do uso de ferramentas de mineração de dados em estudos relacionados a promoção de saúde.

MATERIAL E MÉTODOS

Foi realizada uma revisão bibliográfica sistemática, método que utiliza como fonte de dados a literatura, consistindo numa forma de sintetizar as informações disponíveis em dado momento sobre determinado assunto. Tem como princípios gerais a exaustão na busca de estudos analisados, a seleção justificada dos mesmos por critérios de inclusão e exclusão, e a avaliação da qualidade metodológica (Sampaio; Mancini 2007).

Para a busca, foram utilizados: Portal CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) http://www.periodicos.capes.gov.br/ e Scielo (*Scientific eletronic library*) http://www.scielo.br/ entre abril e maio de 2014. Foram utilizados como descritores *health promotion*, *data mining*, mineração de dados e promoção da saúde. Como critério de inclusão, foram considerados artigos que efetivamente utilizaram quaisquer umas das diversas técnicas de mineração de dados como método, bem como aqueles com data de publicação igual ou posterior ao ano 2000.

RESULTADOS

Foram obtidos, de acordo com a estratégia definida 151 artigos (127 dos periódicos CAPES, 18 da biblioteca Scielo e seis estudos identificados nas referências bibliográficas destes trabalhos). De acordo com os objetivos do estudo e critérios de inclusão sete artigos foram selecionados, dentre eles dois são artigos internacionais, realizados na Coreia do Sul (Chae et al. 2001) e Estados Unidos (Shaw et al. 2012), os outros nacionais realizados em diferentes regiões do país, dentre os artigos analisados, o maior número de publicações foi concentrado nos anos de 2006 a 2012. Os artigos utilizados nessa revisão estão descritos na Tabela 1.

Dentre os sete artigos analisados, apenas um relatou a aplicação da técnica de mineração de dados em um banco de dados pequeno, com apenas 16 indivíduos (Carvalho et al. 2012), todos os outros aplicaram as técnicas em grandes bancos de dados compostos por diferentes variáveis, porém, o fato de o banco de dados ser pequeno não influenciou nos resultados esperados (Carvalho et al. 2012).

A maioria dos autores (Chae et al. 2001, Kobus et al. 2006, Dalagassa 2009, Shaw et al. 2012, Vianna et al. 2010), relatam dificuldades com banco de dados não preenchidos adequadamente, com ausência de variáveis importantes e em alguns casos informações incorretas, o que prejudica os resultados dos estudos.

Todos os autores aplicaram árvore de decisão, como técnica de mineração de dados, alguns justificam essa escolha por ser um método preditivo, uma vez que desempenham inferência nos dados com o objetivo de fornecer previsões ou tendências, além de informações especificas sobre os fatores de risco e grupo alvo que podem ser usados em uma análise política, para as tomadas de decisões (Chae et al. 2001, Dalagassa 2009). Sterner et al. (2006), aplicou em seu estudo, duas técnicas de mineração: árvore de decisão e redes neurais, com o intuito de comparar ambas, e então o autor concluiu que as redes neurais dificultam a compreensão dos dados, já que no estudo realizado por ele, esta técnica precisou em média de 1000 interações para cada situação a ser testada, enquanto que a árvore de decisões apresentou resultados que induziram no processo de decisão, além de facilitar a compreensão dos dados analisados, por conseguinte o autor observou nesse estudo, maiores vantagens na aplicação de árvore de decisão, considerando que essa técnica não assume nenhuma distribuição particular para os dados e ainda os mesmos podem ser qualitativos ou quantitativos.

Tabela 1. Descrição dos objetivos, métodos, resultados e conclusões dos estudos selecionados

Referências	Objetivo	Método	Resultados e Conclusões
(Chae et al. 2001)	Encontrar associações entre as diferentes variáveis do banco de dados, a fim de obter os resultados de saúde dos usuários.	Utilizou o banco de dados do *Korea Medical Insurance*, composto por 13.689 usuários, aplicou a técnica de Árvore de decisão.	Obteve as características dos indivíduos, através das associações feitas pelo KDD e concluiu que esses resultados podem subsidiar políticas públicas e promover a saúde da população.
(Kobus et al. 2006)	Encontrar um modelo de identificação de padrões, que contribua para a indicação de usuários com doenças cardiovasculares elegíveis para o ingresso em programas de gerenciamento de caso.	Nesse estudo foi utilizada a base de dados composta por usuários do Instituto Curitiba de Saúde (operadora de saúde responsável pelos atendimentos à saúde dos funcionários da Prefeitura de Curitiba). Período de 2001 a 2005, 401.041 registros referentes a 8.457 usuários. Foi utilizada a técnica de árvore de decisão.	Os objetivos propostos foram alcançados, de modo que foram encontradas regras de associação consideradas relevantes pelos autores, de maneira que as mesmas possam subsidiar a melhora da qualidade de vida desses usuários e contribuir para utilização dos recursos de financiamento de modo eficiente.
(Sterner et al. 2006)	O presente trabalho objetiva mostrar a influência da análise exploratória dos dados no desempenho das técnicas de Mineração de Dados (*Data Mining*) quanto à classificação de novos padrões por meio da sua aplicação a um problema médico (icterícia) e comparar duas técnicas de *data mining*.	Foi utilizado um banco de dados de um hospital, composto por 118 pacientes, com 14 variáveis diferentes; foram aplicadas as técnicas de árvore de decisão e redes neurais.	Os autores relatam que apenas a técnica de árvore de decisão, foi eficaz em relacionar as variáveis, e ressaltam que especialistas das mais diversas áreas poderiam avaliar os resultados fornecidos pelos métodos abordados (e/ou outros métodos adicionais utilizados para classificação) e validar (ou não) a plausibilidade das predições realizadas por eles, tendo-se, pois, uma ferramenta auxiliar para as suas tomadas de decisão.
(Dalagassa 2009)	Identificar beneficiários com indicativo de *Diabetes mellitus* tipo 2, para ingresso e programas de promoção da saúde, prevenção de doenças e gerenciamento de caso.	O estudo foi desenvolvido utilizando a técnica de árvore de decisão e o software WEKA (*Waikato Environment for Knowledge Analysis*) sobre a base de dados da Data Warehouse da Unimed Federação do Estado do Paraná, estrutura que possibilita gerar as informações de atendimentos de beneficiários ativos do estado.	O autor conclui que através dessa metodologia, é possível estabelecer indicadores que monitorem todo o processo de prestação de serviços, considerando combinar a visão de indicadores, focado na estratégia das operadoras com ênfase nas ações de prevenção e promoção da saúde.

Continuação Tabela 1

Referência	Objetivo	Método	Resultados e conclusões
(Carvalho et al. 2012)	Os autores tiveram como objetivo, exemplificar e discutir a exploração de dados oriundos de acompanhamentos dos pacientes da fisioterapia, utilizando Mineração de Dados e o pós-processamento dos padrões (regras) descobertos.	Foi utilizado um conjunto de dados, que continham registros do acompanhamento de 16 pacientes de uma clínica de fisioterapia, os registros foram selecionados aleatoriamente. Foi utilizado a técnica de árvore de decisão e programa computacional C4.5.	Os autores concluíram que as associações encontradas, apresentam características que devem ser mais exploradas para efeito de adequações de rotinas de trabalho ou fomento a comportamentos de autocuidado, além de outras intervenções, nos níveis primários, secundários ou terciários, na área da saúde.
(Shaw et al. 2012)	Os autores tiveram como objetivo nesse estudo, comparar a prevalência de alegação de assedio no ambiente de trabalho e comparar com outras variáveis, como sexo, idade, raça, etnia ou deficiência, e identificar quais as combinações de características, estão mais relacionadas às queixas de assédio ou outras formas de discriminação.	Foi realizado um estudo retrospectivo, que utilizou dados secundários, para avaliar as características dos indivíduos, foram analisadas também características do local de estudo, como a atividade e o porte da empresa estudada. A técnica de mineração de dados utilizada foi a árvore de decisão.	Os autores concluem que os resultados obtidos fornecem subsídios aos profissionais da área, para que haja um melhor direcionamento a sensibilização das políticas de trabalho, voltadas ao assédio, trabalhando de forma a desenvolver consciência e resiliência nos trabalhadores, podendo dessa forma diminuir os efeitos altamente negativos de um ambiente de trabalho hostil.
(Vianna et al. 2010)	Os autores tiveram por objetivo relatar a experiência da integração das bases de dados de três diferentes sistemas de informação e demonstrar a viabilidade e a possibilidade de replicação em saúde pública da Mineração de Dados (MD), retirando, dessa forma, a subjetividade do analisador, e verificar se é possível utilizar de modo produtivo o resultado obtido com as regras geradas pela ferramenta utilizada.	Foram utilizados os bancos de dados do SINASC, SIM e SINI, de 2000 a 2004. Foram registrados nesses bancos, nesse período 9.372 óbitos infantis. A técnica de mineração de dados utilizada foi a árvore de decisão e o Software utilizado foi o WEKA (*Waikato Environment for Knowledge Analysis*).	Com a análise das regras geradas pelo *data mining*, obteve-se o perfil da mortalidade infantil no Estado do Paraná, no período de 2000 a 2004. As características geradas pelo DM são importantes ainda para a construção de mapas de risco que podem ser utilizados como estratégia de ação para a redução dos óbitos, com a consolidação da integralidade da atenção à saúde e com o fortalecimento de ações intra e intersetoriais no território. Este trabalho demonstra que é possível utilizar o DM em Saúde Pública, obtendo-se conclusões relevantes.

DISCUSSÃO

Foi possível observar no presente estudo que algoritmo correspondente a árvore de decisão foi aplicada em todos os estudos analisados, o que corrobora com Dalagassa (2009), quando afirma que essa técnica é um padrão tanto na área acadêmica, quanto na comercial e consiste em um conjunto hierárquico de conceitos. Sua principal característica está relacionada à compreensibilidade do resultado final, evidenciando em cada resultado da classificação, um indicativo dos registros classificados corretamente em relação aos classificados incorretamente.

Assim como Dalagassa (2009) e Vianna et al. (2010), relatam que os resultados obtidos após a aplicação de técnicas de *data mining* (mineração de dados), corroboram com resultados observados em outros estudos que utilizaram outras ferramentas estatísticas. O *software* WEKA utilizado pelos autores acima citados é uma ferramenta amistosa ao uso por profissionais de saúde, tendo em vista que é um *software* de acesso livre podendo ser utilizado sem custo e com grande facilidade pelas secretarias de saúde, é eficaz, rápido e efetivo.

Os trabalhos estudados visam estabelecer relações e padrões, entre as diferentes variáveis, facilitando a construção de mapas de risco, ou outras formas de observação dos resultados que facilitem o redirecionamento ou a implementação de estratégias de ação em prevenção e Promoção de Saúde.

As ações de Promoção de Saúde têm como objetivo reduzir a iniquidade do estado de saúde da população e assegurar recursos e oportunidades igualitárias na capacitação das pessoas a realizar completamente seu potencial de saúde, para isso, é de suma importância que os gestores de saúde tenham conhecimento dos ciclos de atendimentos e diagnósticos dos usuários, para que possa desenvolver estratégias a fim de melhorar a qualidade dos serviços e reduzir os custos dos mesmos, considerando que a informação em saúde é fundamental ao processo de tomada de decisões no âmbito das políticas públicas, objetivando a qualidade de vida dos povos.

Observou se nesse estudo a relevância e a necessidade da orientação de profissionais da área da saúde, quanto a importância da realização completa da coleta de dados, já que a maioria dos autores descrevem a dificuldade na organização do banco de dados e o quanto os mesmos são incompletos, o que torna os resultados obtidos menos confiáveis, além do tempo gasto na preparação dos mesmos.

CONCLUSÕES

Diante o exposto conclui-se que as técnicas de mineração de dados, se aplicadas em programas de promoção e prevenção da saúde podem auxiliar na identificação de necessidades da saúde dos usuários, bem como a organização dos serviços de saúde necessários para supri-las. As maiores dificuldades podem residir na construção de bancos de dados baseados em prontuários incompletos ou ilegíveis, nas informações contraditórias

ou na ausência de preenchimento de campos importantes, desse modo vale ressaltar que é necessária reorientação dos profissionais de saúde quanto à necessidade de um preenchimento adequado dos bancos de dados, de modo a aumentar a confiabilidade dos mesmos e melhor direcionar a tomada de decisão dos gestores. Não restam dúvidas que essa é uma área muito promissora e ainda apresenta muitos desafios. Esse estudo demonstra, que se utilizadas de forma adequada estas ferramentas podem auxiliar na tomada de decisões e elaboração de ações, principalmente de órgãos públicos, para a promoção de saúde e consequente melhoria da qualidade de vida da população.

REFERÊNCIAS

Braga A, Ludermir T, Carvalho APLF. Redes neurais artificiais-teoria e aplicações. 2. ed. Rio de Janeiro: LTC; 2007. 104-111p.

Carvalho DR, Moser AD, Silva VA, Dalagassa MR. Mineração de dados aplicados a fisioterapia. Rev Fisioter Mov. 2012;25(3):595-605.

Chae YM, Ho SH, Cho KW, Lee DH, Ji SH. Data mining approach to policy analysis in a health insurance domain. Int J Med Inform. 2001; 62(2-3):103-111.

Cunha F, Ribeiro N, Pereira H. Técnicas de gerenciamento de informações em uma rede de hospitais. Perspectivas em Ciência da Informação. 2014;19(1):22-36.

Côrtes S, Porcaro R, Lifschitz S. Mineração de dados- Funcionalidades Técnicas e abordagens [Tese]. Rio de Janeiro: PUC; 2002.

Dalagassa M. Concepção de uma metodologia para identificação de beneficiários com indicativos de diabetes mellitus tipo 2 [Dissetação]. Curitiba: PUC; 2009.

Ebecken NFF, Lopes MCS, Costa MCA, Rezende SO. Sistemas Inteligentes: fundamentos e aplicações. São Paulo: Manole, 2007.

Galvão M, Sawada N, Trevizan M. Revisão Sistemática: recurso que proporciona a incorporação das evidências na prática da enfermagem. Rev. Latino-am Enfermagem. 2004;12(3):549-56.

Hand D, Mannila H, Smyth P. Principles of data mining. 5ed. London: MIT Press; 2001. 93-102p.

Kobus L, Silva S, Dias J. Aplicação da descoberta de conhecimento em base de dados para identificação de usuários com doenças cardiovasculares elegiveis para programas de gerenciamento de caso [Dissertação]. Curitiba: PUC; 2006.

Miranda CRM. Gerenciamento de custos em planos de assistência a saúde. [Internet] Brasilia;2003 [Acesso em 2014 junho 28]. Disponivel em: <http://www.ans.gov.br/portal/upload/biblioteca/TT_AS_20_ClaudioMiranda Gerenciamento de Custo.pdf>

Otawa. Primeira Conferência Internacional sobre Promoção da Saúde; 1986.

Sampaio R, Mancini M. Estudos de revisão sistemática: um guia para a síntese criteriosa da evidência científica. Rev Bras Fisioter. 2007;11(1):83-89.

Shaw L, Chan F, Mcmahon B. Intersectionality and disability harassment: the interactive effects of disability, race, age and gender. RCB. 2012;55(2):82-91.

Sterner M, Soma NY, Shimizer T, Nievola JC, Sterner Neto PJ. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise dos dados. G&P. 2006;13(2):325-337.

Vianna RCXF, Moro CMCB, Moysés SJ, Carvalho D, Nievola JC. Mineração de dados e mortalidade infantil. Cad Saúde Pública. 2010;26(3):535-542.

Weis N, Indurkhya N. Predict Data Mining. Canadá: Morgan Kaufmann Publishers, 1999.