O PROCESSO DE ELEGIBILIDADE DE PACIENTES PARA PROGRAMAS DE PROMOÇÃO À SAUDE

Marcelo Rosano Dallagassa (PUCPR) E-mail: mrdallagassa@gmail.com
Jair Biguetti D. Colmanetti (PUCPR) E-mail: jbdc_jaja@yahoo.com.br
Deborah Ribeiro Carvalho (PUCPR) E-mail: ribeiro.carvalho@pucpr.br

Resumo: O artigo apresenta o potencial do processo *Knowledge Discovery in Database* na seleção de participantes para programas de promoção à saúde. São apresentados dois experimentos, em cenários distintos, tendo como mesmo foco a elegibilidade de participantes aos programas de promoção à saúde. Entretanto, em um dos cenários, estava disponível a Classificação Internacional de Doenças. Em relação às estratégias de mineração de dados, no primeiro, foi adotada apenas a classificação, enquanto, no segundo, a classificação e descoberta de regras de associação temporal. Para o primeiro cenário, foram selecionadas 12 variáveis, considerando um conjunto de 43.375 beneficiários, sendo descobertas 843 regras. Dentre estas, foram selecionadas seis para serem avaliadas por quatro especialistas. Em relação ao segundo experimento, foram consideradas 170 variáveis sobre 1.617 colaboradores. Foram descobertas 14 regras de produção e outras 108 regras temporais, todas avaliadas por cinco especialistas. A partir dos resultados das experimentações, considerando os dois cenários distintos, foi possível identificar não apenas os fatores que indicam a seleção dos participantes, mas também a janela temporal de sua ocorrência.

Palavras-chave: Descoberta do conhecimento em base de dados, Classificação, Prevenção e promoção à saúde, Regras de associação temporal, Diabetes mellitus 2, Ortopedia.

.

THE PATIENT ELIGIBILITY PROCESS FOR HEALTH PROMOTION PROGRAMS

Abstract: This article presents the potential of the KDD process in the participants selection for health promotion programs. Two experiments are presented in different scenarios, focusing on the eligibility of participants to health promotion programs. One of them has the International Classification of Diseases but in the second scenario the ICD is not available. Regarding the Data Mining strategies, in the first one only the classification was adopted, while in the second classification and discovery of temporal association rules were adopted. For the first scenario, 12 variables were selected, considering a set of 43,375 beneficiaries, with 843 rules being discovered. Among these 843 rules, six were selected to be evaluated by four experts. In relation to the second experiment, 170 variables were considered about 1,617 employees. We have discovered 14 production rules and another 108 time rules - all of which were evaluated by five experts. From the results of the experiments, considering the two distinct scenarios, it was possible to identify not only the factors that indicate the selection of the participants, but also the temporal window of their occurrence

Keywords: Data Mining; Classification; Health Promotion; Temporal Association Rules; Diabetes Mellitus; Orthopedics

1. Introdução

A mudança da transição epidemiológica e demográfica do Brasil agrava a incidência de Doenças Crônicas (DCs), gerando um aumento dos custos da saúde, em função da crescente utilização de materiais de alto custo, de tratamentos mais onerosos e de complicações inerentes às doenças (MORAIS; BURMESTER, 2014)

As DCs acometem o indivíduo enfermo durante um longo período de latência, com progressivas manifestações, especialmente se estiver sem um acompanhamento adequado. Tal situação motiva ações dirigidas das organizações de saúde, no sentido de minimizar os custos da assistência, ampliando a prevenção, seu respectivo controle e o gerenciamento de casos.

Miranda (2003) comenta sobre a necessidade de se pensar em um novo modelo

assistencial, focado na saúde e não na doença, e enfatiza que, segundo a Organização Mundial da Saúde, entre os diversos fatores que fazem com que um indivíduo ultrapasse os 65 anos, apenas 10% estão ligados à assistência médica. O estilo de vida é responsável por 53%, o meio ambiente, por 20% e a herança genética, pelos demais 17%. Esse fato evidencia o baixo alcance do modelo vigente para a melhoria da qualidade de vida e de saúde da população e impulsiona a adoção de um novo paradigma assistencial.

A medicina moderna é praticada por aqueles que utilizam o conhecimento, na medida de sua capacidade, para desenvolver a saúde, evitar a doença e a invalidez e prolongar a vida com qualidade. Araújo (2004) apresenta três níveis de prevenção para a promoção à saúde:

a) Prevenção primária: compreende ações que permitem a redução da ocorrência de doenças. Inclui não somente campanhas de vacinação, mas também investimentos em saneamento básico, sugestões sobre os hábitos de vida e alimentares, campanhas antitabagismo, entre outras que apontem para ganhos em qualidade de vida.

b) Prevenção secundária: envolve ações que objetivam a redução ou a eliminação de consequências para a saúde, decorrentes de doenças crônicas como câncer, diabetes, doenças cardiovasculares, entre outras. Recebe também a denominação gerenciamento de doenças.

c) Prevenção terciária: envolve ações que permitem minimizar o sofrimento causado pelas limitações impostas às pessoas já acometidas por doenças crônicas. Recebe também a denominação gerenciamento de caso.

O Caderno de atenção básica sobre o diabetes mellitus, do Ministério da Saúde, evidencia que mudanças no estilo de vida reduziram 58% da incidência de diabetes em três anos. Essas mudanças visavam à discreta redução de peso (5-10%), manutenção do peso perdido, aumento da ingestão de fibras, restrição energética moderada, restrição de gorduras, especialmente as saturadas, e aumento de atividade física regular, além de intervenções farmacológicas. Alguns medicamentos utilizados no tratamento do diabetes, como a metformina, também foram eficazes, reduzindo em 31% a incidência da doença em três anos. Esse efeito foi mais acentuado em pacientes com Índice de Massa Corpórea (IMC) maior que 35 kg/m2.4

Miranda (2005) corrobora e complementa que o gerenciamento de doença está baseado no entendimento de que é possível atuar na rede causal relacionada com determinadas doenças que ocorrem com significativa magnitude em termos de morbimortalidade, propiciando a intervenção num momento mais precoce da sua história natural, de forma a reduzir a ocorrência de suas manifestações e complicações, tendo, por consequência, uma melhor qualidade de vida para os beneficiários, com menor custo para o sistema.

Os programas de promoção à saúde passam, inicialmente, pela elegibilidade dos indivíduos, em geral de natureza crônica ou de predisposição a tal, com o objetivo de abordar os diversos graus de necessidades de saúde, contemplando a gestão de fatores ou comportamentos de risco, por meio de uma equipe de profissionais da saúde, para evitar ou minimizar os agravamentos das suas condições, promovendo, dessa maneira, a detecção precoce e um plano de tratamento preventivo e adequado das doenças e suas complicações.

Desenvolver um processo assistencial, que contemple programas de promoção à saúde, permite, além da redução de custos, uma melhoria da qualidade de vida dos pacientes,

influenciando na redução das necessidades de atenção à saúde que demandem ações de maior complexidade assistencial. Contudo, existe o desafio da elegibilidade dos pacientes, dado o volume de dados processados, a partir de estratégias da tecnologia da informação "ditas" tradicionais. O desafio amplia-se pela não disponibilidade de dados clínicos epidemiológicos, em função da privacidade e sigilo de informações garantidos ao indivíduo, inclusive pela não mais obrigatoriedade da informação da Classificação Internacional de Doenças (CID) nas transações entre os prestadores e pagadores.

Para que se implantem programas de promoção à saúde, é fundamental a busca das organizações por alternativas, como o processo *Knowledge Discovery in Database* (KDD), o qual compreende em uma de suas fases a mineração de dados, em que ocorre a aplicação de algoritmos com a finalidade específica de identificar padrões válidos, novos, potencialmente úteis e compreensíveis (FAYYAD et al., 1996). Portanto, este artigo tem como objetivo apresentar potencialidades do uso do processo KDD na saúde, na elegibilidade dos indivíduos com propensão às DCs, voltada para o apoio a programas de promoção à saúde.

Modelos dessa natureza constituem interessante instrumento para a gestão das organizações de atenção à saúde e subsidiam uma melhor compreensão por parte dos pacientes sobre as variáveis que contribuem ou não para a sua própria saúde e melhores práticas de promoção e prevenção.

O processo KDD foi originalmente desenvolvido para apoiar a gestão empresarial, mas tem sido amplamente adotado por outras áreas, inclusive na saúde. Compreende três etapas: o pré-processamento, a mineração de dados e o pós-processamento (Figura 1).

Figura 1 – Etapas do processo de descoberta de conhecimento. Fonte: Adaptação de Fayyad et. al. (1996).

O pré-processamento é uma etapa, em geral, trabalhosa, em função dos dados disponíveis não estarem organizados de forma a permitir a aplicação direta dos algoritmos de mineração. A etapa de extração de padrões (mineração de dados) é a mais direcionada ao cumprimento dos objetivos, buscando por padrões que apoiem o processo decisório, ou seja, o problema de gestão que motivou o processo KDD.

No modelo proposto por Fayyad et al. (1996), são descritas algumas tarefas para o processo de mineração de dados, entre elas: a classificação, o agrupamento e a descoberta de regras de associação.

A classificação consiste em classificar um novo registro como pertencente a determinada classe entre várias previamente definidas. Cada classe, também chamada atributo-meta, corresponde a um padrão único de valores dos atributos previsores (demais atributos que caracterizam o registro). Esse padrão único também pode ser interpretado como uma descrição da classe.

O principal objetivo da construção de um classificador é descobrir algum tipo de relação entre os atributos previsores e o atributo que caracteriza a classe (FREITAS; LAVINGTON, 1998). Segundo Breiman et al. (1984), um classificador extraído de um conjunto de dados atende a dois propósitos: predição do valor da classe e entendimento da relação existente entre os atributos previsores e a classe. Para cumprir o segundo, é exigido do classificador que ele não apenas classifique, mas também explicite o

conhecimento extraído de forma compreensível.

Entre as formas de representação do classificador descoberto, a árvore de decisão constitui uma representação simples e de fácil compreensão. Em geral, o processo de descoberta está baseado na indução, a partir de um conjunto de exemplos de dados para os quais as classes são previamente conhecidas. A estrutura da árvore é organizada de tal forma que:

- Cada nó interno (não folha) é rotulado com um dos atributos previsores.
- Os ramos (ou arestas) originados a partir de um nó interno são rotulados com operadores relacionais (>, >=, <, <=, =) e valores do domínio do respectivo atributo.
- Cada nó-folha é rotulado com um valor de domínio do atributo que caracteriza a classe.

A fim de contribuir para a compreensibilidade da árvore descoberta, esta pode ser traduzida em regras "se... (condições), então... (classe)", cuja interpretação é: "se" os valores dos atributos satisfazem as condições da regra, "então" o exemplo pertence à classe prevista por ela.

A tarefa de agrupamento consiste na identificação de um conjunto finito de classes ou clusters, baseado nos atributos de objetos não previamente classificados (CHEN; HAN; YU, 1996). Um cluster é basicamente um conjunto de objetos agrupados em função de sua similaridade ou proximidade, de forma que as similaridades intraclusters (dentro de um mesmo cluster) sejam maximizadas e as interclusters (entre clusters diferentes), minimizadas.

Uma vez definidos os agrupamentos, cada registro é associado ao correspondente grupo e as características comuns dos registros de um mesmo grupo podem ser sumarizadas para formar a descrição do agrupamento. Por exemplo, um conjunto de pacientes pode ser identificado em vários grupos (clusters), baseado nas similaridades dos seus sintomas; os sintomas comuns aos pacientes de cada grupo (clusters) podem ser usados para descrever o grupo (cluster) a que um novo paciente pertencerá. Assim, um novo paciente X poderia ser atribuído a determinado grupo no qual os demais teriam sintomas os mais parecidos possíveis. Dessa forma, a tarefa de agrupamento, cujo resultado é a identificação deste, pode ser realizada como pré-processamento para realização da tarefa de classificação (KUBAT, 1998).

A tarefa de descoberta de regras de associação objetiva descobrir relações, que são expressões do tipo X → Y, lidas como: SE (X) ENTÃO (Y), sendo X e Y conjuntos de itens, que, em geral, atendem à seguinte propriedade: X ∩ Y = Ø, representando que os conjuntos de itens X e Y são, em geral, encontrados simultaneamente em uma mesma instância. Um exemplo de regra do tipo X → Y poderia ser: <SE> implantação de stent convencional <ENTÃO> consulta de emergência, com suporte de 42% e confiança de 82%, ou seja, 42% dos indivíduos que integram a base de dados implantaram stent convencional, dos quais 82% também tiveram consulta de emergência.

Formalmente, confiança (2) e suporte (1) são definidos da seguinte forma (AGRAWAL; IMIELINSKI; SWAMI, 1993; SRIKANT; AGRAWAL, 1995):

$$\text{Suporte} = \frac{|X \cup Y|}{N} \quad (1) \qquad \text{Confiança} = \frac{|X \cup Y|}{[X]} \quad (2)$$

Em que: N é o número total de exemplos e |X| denota a cardinalidade do conjunto X.

A partir dos padrões descobertos pela etapa de mineração de dados, existem vários critérios para avaliar a respectiva qualidade. Os três mais usados são a precisão preditiva, a compreensibilidade e o grau de interesse do conhecimento descoberto. A taxa de acerto é normalmente medida pelo número de exemplos de teste classificados corretamente dividido pelo número total de exemplos do conjunto de teste.

Existem várias estratégias propostas na literatura para a atribuição de grau interesse aos padrões descobertos. As diversas medidas descritas, em geral, são organizadas em dois grupos, ditas *user-driven* e *data-driven* (SILBERSCHATZ; TUZHILIN, 1996; FREITAS, 1998; CARVALHO; FREITAS; EBECKEN, 2003). A ideia básica das primeiras é que o usuário especifique suas crenças ou conhecimento prévio sobre o domínio da aplicação e o sistema utilize esse tipo de informação para selecionar regras "interessantes". Uma regra é considerada interessante se representa alguma novidade com relação às crenças ou conhecimento prévio do usuário (SIONARA, 2006). Em contrapartida, as medidas ditas *data-driven* tentam estimar o quanto as regras podem ser surpreendentes ao usuário de uma forma mais automática e indireta, a partir de formulações aplicadas sobre os dados, sem exigir uma prévia especificação de crenças ou conhecimento (TAN; KUMAR; SRIVASTAVA, 2002).

É possível identificar na literatura alguns relatos do uso da tarefa de mineração de dados para apoiar processos decisórios envolvendo DCs. Por exemplo, Toussi et al. (2009) apresentam um modelo baseado no algoritmo de árvore de decisão C5 (KUHN,2014), para a análise das diretrizes francesas do diabetes mellitus tipo 2. O conjunto utilizado foi composto por dados administrativos, antropométricos e clínicos. Este estudo, além de destacar a capacidade preditiva do classificador descoberto e representado na forma de árvore de decisão, valorizou a possibilidade de compreensão, permitindo sua análise frente às diretrizes.

Soni et al. (2011) trazem os resultados de uma pesquisa que avaliou algumas das técnicas de mineração de dados que vêm sendo utilizadas, particularmente para a predição de doenças do coração. Os resultados obtidos revelaram que a árvore de decisão superou o *naive bayes* quanto ao tempo necessário para a sua execução, apesar de a qualidade preditiva ser semelhante. Os resultados também apontam que esses dois métodos (árvore de decisão e *naive bayes*) superam os resultados obtidos pelas técnicas de Redes Neurais Artificiais (RNAs) e classificação baseada em agrupamentos.

Marinov et al. (2011) apresentam uma revisão sistemática envolvendo 17 estudos aplicando técnicas de mineração de dados para questões relacionadas ao diabetes, dos quais dez utilizaram, entre outras, a tarefa de classificação, sendo adotada a representação por árvore de decisão em seis deles. Entre os algoritmos empregados, estão o CART, C4.5 e C5.0 (BREIMAN, 1984; QUINLAN, 1993; KUHN, 2014)

Chae et al. (2001) trabalharam com um banco de dados da Korea Medical Insurance, composto por 13.689 usuários, para a obtenção de regras de associação a ser utilizadas para o desenvolvimento de políticas públicas de promoção à saúde da população.

Como exemplo para a tarefa de descoberta de regra de associação, pode-se citar uma aplicação em gestão da saúde voltada à identificação de usuários com doenças cardiovasculares (KOBUS, 2006) também para programas de promoção à saúde. Em seu trabalho, descreve diversas regras associando algumas tecnologias de saúde, que podem ser avaliadas em relação aos registros de atendimentos associados a determinadas doenças.

Para o experimento relatado por (KOBUS, 2006), foi utilizada uma base de dados de

uma prestadora de serviços de saúde com 1.168.983 registros de atendimentos a 55.814 usuários. As variáveis selecionadas foram: sexo, data nascimento, data procedimento e código procedimento. Neste experimento, foram descobertas 639 regras a partir do algoritmo APRIORI, (BORGELT, 1998), sob a perspectiva da revascularização do miocárdio.

2. Método

São apresentados dois experimentos, objetivando a seleção de participantes para o programa de promoção à saúde, a partir da adoção do processo KDD. O experimento 1, proposto por Carvalho, Dallagassa e Silva, (2015) tem como foco o diabetes mellitus tipo 2 e o experimento 2, proposto por Colmanetti (2016), é sobre afastamentos do trabalho por questões relacionadas à ortopedia. Ambos foram aprovados pelo Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Paraná (PUCPR), sob os números 0001638/08 e 1.183.459, respectivamente.

Para a construção do conjunto de dados (experimento 1), foram sistematizados os atendimentos dos beneficiários ao longo de seis anos, entre janeiro de 2007 e dezembro de 2012, considerando os 43.375 beneficiários ainda ativos em dezembro de 2012. Esse intervalo de seis anos foi adotado em função da disponibilidade no Data Warehouse (DW) da operadora, os quais foram organizados por beneficiário, porém sem a respectiva identificação, não apenas por questões éticas, mas também por não ser relevante para o processo em questão.

O experimento 2 foi desenvolvido em uma empresa de logística de abrangência nacional, porém considerando dados relativos ao estado do Paraná. Foram estudados os colaboradores que tiveram afastamentos relativos à ortopedia, entre 2008 e 2013, bem como selecionado, de forma aleatória, igual número de colaboradores sem afastamento pelo mesmo motivo no período referido.

Além de ambos possuírem focos distintos de investigação, o primeiro dispunha do dado referente ao CID e o segundo não, tendo em vista não ser mais de notificação obrigatória. Ademais, apesar de estarem fortemente baseados nas etapas do KDD, existem algumas diferenças que podem ser identificadas em azul, a partir da Figura 2.

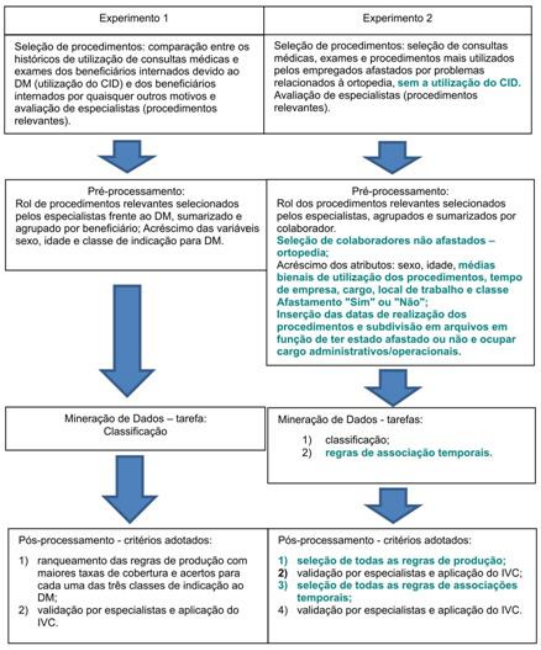

Figura 2 – Encaminhamentos experimentos 1 e 2. Fonte: Carvalho et. al. (2012) e Colmanetti (2016).

No caso do experimento 1, foi criado o atributo "classe" para cada registro, conforme os seguintes critérios:

- Se existe atendimento com CID de diabetes ou número de solicitações de exame de hemoglobina glicosilada <= 2, então classe = "sem indicativo para DM2".
- Se existe atendimento com CID de diabetes ou número de solicitações de exame de hemoglobina glicosilada > 2, então classe = "com indicativo para DM2".

Já para o experimento 2, o atributo "classe" decorreu naturalmente do fato de se tratar ou não de pessoa afastada por problemas ortopédicos.

Para atender às sugestões de trabalhos futuros propostos por Kobus (2006), Carvalho, Dallagassa e Silva (2015) e Escobar (2015), sobre sequenciamento de eventos demandados por usuários de serviços de saúde, foi também adotada a tarefa de descoberta de regras de associação temporal. Para sua implementação, foi aplicado o algoritmo AssocTemp (SOKOLOSKI, CARVALHO e DALLAGASSA, 2014) considerando as seguintes janelas temporais: 0-360 dias, 361-720 dias e 721-1.080 dias, as quais se justificam em função da periodicidade imposta pela legislação trabalhista brasileira para realização de exames periódicos. Dessa forma, possibilitou a análise dos especialistas frente às demandas dos exames periódicos, sendo que, na empresa foco do estudo, estes ocorrem anualmente ou a cada dois anos. Também foram agrupados em conjuntos distintos ocupantes de cargos operacionais e administrativos.

A justificativa para trabalhar de forma diferenciada com os dois grupos de cargos decorre de a literatura preconizar que, em função das atividades desempenhadas, se podem ou não potencializar problemas musculoesqueléticos (BRASIL, 2009). Vale destacar que, na empresa foco do experimento 2, as atividades desempenhadas podem ser divididas considerando os grupos de cargos administrativos e operacionais. Entre os afastados, 820 colaboradores executam atividades relativas ao grupo de cargos operacionais e 323, ao grupo de cargos administrativos. Por sua vez, o conjunto dos não afastados por ortopedia foi composto por 322 colaboradores que executam atividades relativas àquele grupo e 152, a este. Dessa forma, os dados foram agrupados em arquivos distintos, para cada um dos grupos de cargos

3. Resultados

Para avaliação dos padrões descobertos, foi adotada a mesma estratégia para ambos os experimentos. Cada especialista, ao analisar cada regra, atribuiu um grau de concordância, que posteriormente foi traduzido para o respectivo escore. O conjunto de opções para o grau de concordância foi estabelecido da seguinte forma:

- Concordo "C" = a regra confirma o conhecimento (escore = 2).
- Concordo parcialmente "CP" = a regra contraria o conhecimento, mas não apresenta uma ou mais condições no antecedente da regra que representa equívoco ou erro (escore = 1).
- Discordo "D" = a regra contraria o conhecimento, mas apresenta uma ou mais condições no antecedente da regra, que representa equívoco ou erro (escore = 0).

A partir da etapa de seleção dos procedimentos, foram obtidos os seguintes resultados:

- Experimento 1: foram identificadas 12 variáveis a ser consideradas envolvendo exames laboratoriais e especiais, consultas realizadas e outras.
- Experimento 2: foram identificados 82 procedimentos envolvendo questões

vinculadas à ortopedia.

A partir da etapa de pré-processamento, obtiveram-se os seguintes resultados:

- Experimento 1: conjunto de 43.375 registros referentes às 12 variáveis, mais o atributo "classe".
- Experimento 2: conjunto de 1.617 registros envolvendo as variáveis "idade", "sexo", "tempo de empresa", "cargo", "local de trabalho", "quantidade utilizada de cada um dos 82 procedimentos relevantes", bem como suas respectivas "médias bienais de utilização", mais o atributo "classe" indicando afastamento "sim" ou "não".

A partir da etapa de mineração de dados, aplicando a tarefa de classificação, foram obtidos os seguintes resultados:

- Experimento 1: foi descoberta a árvore de decisão, composta por 843 ramificações, apresentando uma taxa de acerto de 88,9%.
- Experimento 2: foi descoberta a árvore de decisão, composta por 14 ramificações, apresentando uma taxa de acerto de 74,1%. Do total de colaboradores afastados (1.143), o algoritmo classificou corretamente 1.034 registros, apresentando a taxa de acerto para esse subconjunto de 90,5%, enquanto a taxa de acerto para os colaboradores não afastados foi de apenas 34,8%, ou seja, dos 474 colaboradores não afastados, 165 foram classificados corretamente. No entanto, essa taxa de acerto não inviabiliza a pesquisa, em face do intuito da elegibilidade de fatores entre os afastados.

Os atributos considerados nas árvores de decisão para os dois experimentos foram:

- Experimento 1: exame glicose, exame creatinina, exame microalbuminuria, exame colesterol total, exame curva glicêmica, exame de mapeamento de retina, consulta oftalmologia, consulta endocrinologia, consulta nefrologia, consulta cardiologia, sexo e idade.
- Experimento 2: tempo de empresa, procedimento imobilizações não gessadas, consulta ortopedia em consultório, RX articulação escapuloumeral (ombro), RM articular (por articulação), consulta ortopedia em pronto-socorro, RX pé, RM coluna cervical, lombar ou dorsal, RX patela e RX punho.

Sobre a etapa de validação dos padrões descobertos, foram obtidos os seguintes resultados:

- Experimento 1: das 843 regras descobertas, foram selecionadas seis para validação pelos especialistas, obtendo um Índice de Validade de Conteúdo (IVC) de 89,6%.
- Experimento 2: após a validação das 14 regras, foi obtido um IVC de 77,9%.

Sobre o IVC obtido em relação ao experimento 2, vale destacar que, diante da argumentação dos especialistas, houve tendência a valorizar o conhecimento disponível e não necessariamente o grau de "novidade" que o padrão poderia demonstrar. Nesse sentido, quando as regras não consideraram procedimentos de alto custo ou vários procedimentos associados, o IVC não atingiu o valor mínimo de 80%. Foi observado também que as opções "concordo", "concordo parcialmente" e "discordo", disponibilizadas no instrumento de apoio para a avaliação, contribuíram para esses resultados. Assim sendo, para experimentações futuras, fica a sugestão de substituir as

opções de resposta, como, por exemplo, a regra “surpreende” o especialista em função da adoção da regra “confirma”.

A despeito de o experimento 1 dispor da variável “CID”, a complexidade do classificador descoberto (843 regras) foi bem maior que aquela descoberta no escopo do experimento 2 (14 regras), sem ter disponível esse dado. Essa situação pode ser explicada não apenas pela disponibilidade de dados sociodemográficos complementares, mas também por se tratar de um grupo de participantes com maior homogeneidade, ou seja, todos são colaboradores de uma mesma empresa, situação não presente no experimento 1, que teve como participantes beneficiários de um plano de saúde.

O modelo desenvolvido no experimento 1 apresentou resultados eficientes para a elegibilidade de beneficiários com potencial para evolução para DCs, se comparados às estatísticas disponíveis. Por exemplo, a partir das regras descobertas, foram indicados 5.953 beneficiários, representando 5,7% do total de beneficiários da carteira. Esse resultado está compatível com o manual do Vigitel de 2012 33, que apresenta 5,6% de adultos com diagnóstico médico referido para o diabetes.

Esse modelo (experimento 1), considerando o diabetes, foi posteriormente replicado para a descoberta de regras que permitissem identificar beneficiários propensos a outras DCs, como hipertensão, isquemias do coração, neoplasias, doenças pulmonares obstrutivas crônicas, doença renal crônica, obesidade e doenças psiquiátricas. Vale destacar que, além da contribuição científica, o experimento 1 representa uma contribuição social, pois, a partir dos seus resultados, bem como daqueles replicados considerando outras DCs, foi criada uma aplicação possibilitando selecionar, de forma automática, os beneficiários para os diversos programas de prevenção de doença e promoção à saúde.

Dado o objetivo do experimento 2, evidencia-se a possibilidade de selecionar os colaboradores que atendem aos seguintes critérios: (i) tempo de empresa superior a cinco e inferior a 37 anos, com mais de duas consultas de ortopedia em consultório; (ii) tempo de empresa superior a cinco e inferior a 37 anos, com menos de três consultas de ortopedia em consultório, associadas a um ou mais exames de ressonância magnética articular (por articulação) e com até um raio X de articulação escapuloumeral (ombro); (iii) tempo de empresa superior a cinco e inferior a 37 anos, com menos de três consultas de ortopedia em consultório, associadas a mais de um exame de raio X de articulação escapuloumeral (ombro); (iv) tempo de empresa superior a cinco e inferior a 37 anos, com menos de três consultas de ortopedia em consultório, associadas a mais de cinco consultas de ortopedia em pronto-socorro, com até um raio X de articulação escapuloumeral (ombro) e, ainda, a um ou mais exames de ressonância magnética da coluna cervical, lombar ou dorsal.

4. Conclusão e discussão

Em geral, as empresas que desempenham atividades de gestão em saúde dispõem de uma grande quantidade de dados relativos aos procedimentos demandados pelos usuários, porém sem a respectiva potencialização de uso, seja por desconhecimento, seja por acreditar que apenas as estratégias “ditas” tradicionais são suficientes para a extração de informação para apoio à decisão.

Este artigo apresentou dois experimentos baseados no processo KDD, objetivando demonstrar o potencial de exploração de bases de dados com foco na seleção de participantes para programas de promoção à saúde. As principais diferenças decorrem

do conjunto de dados disponível e das tarefas de mineração de dados adotadas.

No experimento 1, foi possível considerar a CID, sendo adotada a tarefa de classificação. No experimento 2, não estava disponível a CID e foram adotadas as tarefas de classificação e descoberta de regras de associação temporais.

As variáveis determinantes para a seleção, considerando o primeiro cenário, foram: exame glicose, exame creatinina, exame microalbuminuria, exame colesterol total, exame curva glicêmica, exame de mapeamento de retina, consulta oftalmologia, consulta endocrinologia, consulta nefrologia, consulta cardiologia, sexo e idade. Já no segundo, foram: tempo de empresa, procedimento imobilizações não gessadas (qualquer segmento), consulta ortopedia em consultório, exame RX articulação escapuloumeral (ombro), exame RM articular (por articulação), consulta ortopedia em pronto-socorro, exame RX pé, exame RM coluna cervical ou dorsal ou lombar, exame RX patela e exame RX punho.

Complementando as regras descobertas pela tarefa de classificação, com o intervalo de tempo da reutilização do mesmo procedimento (regras de associação temporais), tem-se que, se o tempo de empresa for superior a cinco e inferior a 37 anos, também deverá ser observado se a quantidade de consultas de ortopedia em consultório supera duas, em um intervalo de até 12 meses, devendo o respectivo colaborador ser selecionado.

As duas experimentações demonstram a eficiência da adoção do processo KDD para a seleção de participantes de programas de promoção à saúde, os quais proporcionam a redução de custos para as empresas gestoras, a promoção da saúde e a melhoria da qualidade de saúde e de vida.

5. Agradecimentos

Os autores agradecem o apoio e financiamento, dessa pesquisa, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudo.

Referências

AGRAWAL, Rakesh, IMIELINSKI, Tomasz, SWAMI, Arun. Mining Associations between Sets of Items in Massive Databases. Proc. of the ACM-SIGMOD 1993 **Int'l Conference on Management of Data**. Washington D.C., p.207-216,1993.

ARAÚJO, Mário Luiz Cardoso. **Gerência de Assistência à Saúde no Setor de Saúde Suplementar: Uma Experiência**. 58 p. Dissertação (Mestrado). Escola Nacional de Saúde Pública – ENSP, Rio de Janeiro, 2004.

BORGELT, Christian. Apriori. Espanha: European Center for Soft Computing, 1998. Disponível em: <http://www.borgelt.net/apriori.html>. Acesso em: 20 jan. 2010.

BRASIL. Constituição Federal. Decreto Nº 6.856, DE 25 DE MAIO DE 2009. Disponível em <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2009/decreto/d6856.htm> . Acesso 22 de dez.2015.

BREIMAN, Leo, FRIEDMAN, Jerome, OLSHEN, Richard, STONE, Charles. **Classification and Regression Trees**. Wadsworth and Brooks, Monterey, Ca. 1984.

CARVALHO, Deborah Ribeiro, DALLAGASSA, Marcelo Rosano, SILVA, Sandra Honorato. Uso de técnicas de mineração de dados para a identificação automática de

beneficiários propensos ao diabetes mellitus tipo 2. **Informação & Informação**, Londrina, v. 20, n. 3, p.274-296, 2015.

CARVALHO, Deborah Ribeiro. **Árvore de decisão: algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados**. 160p. Tese (Doutorado) - Programa De Pós-Graduação em Computação de Alto Desempenho/Sistemas Computacionais – UFRJ, Rio de Janeiro, 2005.

CARVALHO, Deborah Ribeiro; FREITAS, Alex Alves; EBECKEN, Nelson Francisco Favilla. A critical review of rule surprisingness measures. **Proc. Data Mining IV - Int. Conf. on Data Mining**, Rio de Janeiro, Brazil: WIT Press, Dec. p.545-556, 2003.

CHAE, Young Moon. Data mining approach to policy analysis in a health insurance domain. **International journal of medical informatics**, v. 62, n. 2, p. 103-111, 2001.

CHEN, Ming-Syan; HAN, Jiawei, YU, Philip. Data Mining: An Overview from Database Perspective. **IEEE Transactions on Knowledge and Data Engineering**, v.8, n.6, p.866-883, 1996.

DALLAGASSA, Marcelo Rosano. **Concepção de uma metodologia para identificação de beneficiários com indicativos de diabetes mellitus tipo 2**. 105p. Dissertação (Mestrado) - Pontifícia Universidade Católica do Paraná-PUCPR, Curitiba, 2009.

ESCOBAR, Leandro Fabian. **Pós-processamento de padrões para identificação de beneficiários de alto custo em operadores de saúde**. 119 p. Dissertação (Mestrado) - Universidade Federal do Paraná, Curitiba, 2015.

FAYYAD, Usuama, PIATETSKY-SHAPIRO, Gregory, SMYTH, Padhraic, UTHURUSAMY, Ramasamy. Advances in Knowledge Discovery and Data Mining. **American Association for Artificial Intelligence**. Menlo Park, CA: MIT Press, 625p. 1996.

HUSSAIN, Farhad, LIU, Huan, SUZUKI, Einoshin; LU, Hongiun. Exception Rule Mining with Relative Interestingness Measure. **PAKDD**, v.1805, n.1, p.86-97, 2000.

KOBUS, Luciana Schleder Gonçalves. **Aplicação da descoberta de conhecimentos em bases de dados para identificação de usuário com doenças cardiovasculares elegíveis para programas de gerenciamento de caso**. 2006. Dissertação (Mestrado em Tecnologia em Saúde) - Pontifícia Universidade Católica do Paraná, Curitiba, 2006.

KUBAT, Miroslav, BRATKO, Ivan, MICHALSKI, Riszard. A Review of Machine Learning Methods. In: MICHALSKI, R. S.; BRATKO, I.; KUBAT, M. (Eds.). **Machine Learning and Data Mining: Methods and Applications**. London: John Wiley & Sons, p.3-69, 1998.

KUHN, Max, WESTON, Steve, COULTER, Nathan, QUINLAN, Ross. C50: C5.0 decision trees and rule-based models. **R package version** 0.1. 0-19. 2014.

MARINOV, Miroslav. Data mining technologies for diabetes: a systematic review. **Journal of Diabetes Science and Technology**, Thousand Oaks, v. 5, n. 6, p.1549-1556, 2011.

MORAIS, Marlus Volney, Burmester Haino. **Auditoria em Saúde**. São Paulo, Editora Saraiva,172p, 2014.

QUINLAN, Ross. **C4.5 Programs for Machine Learning**. San Diego, CA: Morgan Kaufmann Publishers, 1993.
SILBERSCHATZ, Avi; TUZHILIN, Alexander. What makes patterns interesting in knowledge discovery systems. **IEEE Transactions on Knowledge and Data Engineering**, v.8, n.6, 1996.

SILVA, Luiz Sérgio, PINHEIRO, Tarcísio Márcio Magalhães, SAKURAI, Emília. Perfil do absenteísmo em um banco estatal em Minas Gerais: análise no período de 1998 a 2003. **Ciência e Saúde Coletiva**, Rio de Janeiro, v.13, n.2, p.2049-2058, 2008.

SIONARA, Roberta. **Identificação de regras de associação interessantes por meio de análises com medidas objetivas e subjetivas**. 105p. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e Computação – ICMC – USP, São Paulo, 2006.

SOKOLOSKI, Willian Felipe, CARVALHO, Deborah Ribeiro, DALLAGASSA, Marcelo Rosano. Regra de associação temporal. **XIV Congresso Brasileiro em Informática em Saúde – CBIS**, Santos, Brasil, 2014.

SONI, Jyoti. Predictive data mining for medical diagnosis: an overview of heart disease prediction. **International Journal of Computer Applications**, New York, v. 17, n. 8, p. 43-48, 2011.

SRIKANT, Ramakrishnan; AGRAWAL, Rakesh. Mining Generalized Association Rules. Proc. 21 Int. **Conf. Very Large Databases**, p.407-419, 1995.

TAN, Pang-ning, KUMAR, Viping, SRIVASTAVA, Jaideep. Selecting the Right Interestingness Measure for Association Patterns. Proc of the Eighth ACM **SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD-2002)**, p.32-41, 2002.

TAN, Pang-ning, STEINBACH Michael, KUMAR, Vipin. **Introduction to Data Mining**. Boston, USA: Longman, 2005

TOUSSI, Massoud, LAMY Jean-Baptiste, LE Toumelin Philippe, VENOT Alain. Using data mining techniques to explore physician's therapeutic decision when clinical guidelines do not provide recommendations: methods an example for type 2 diabetes. **BMC Medical Informatics and Decision Making**, London v. 9, n. 28, p.9-28, 2009.