DOI: 10.5433/1981-8920.2015v20n3p274

USO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA A IDENTIFICAÇÃO AUTOMÁTICA DE BENEFICIÁRIOS PROPENSOS AO DIABETES MELLITUS TIPO 2

UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA LA IDENTIFICACIÓN AUTOMÁTICA DE LOS BENEFICIARIOS PROPENSOS A LA DIABETES MELLITUS TIPO 2

Deborah Ribeiro Carvalho - ribeiro.carvalho@pucpr.br
Doutora em Computação de Alto Desempenho pela Universidade Federal do Rio Janeiro (UFRJ). Professora Professor da Pontifícia Universidade Católica do Paraná (PUCPR).

Marcelo Rosano Dallagassa - mrdallagassa@gmail.com
Mestre em Tecnologia em Saúde pela Pontificia Universidade Católica do Paraná (PUCPR). Professor da PUCPR.

Sandra Honorato da Silva - sandrahonorato.honoratosilva@gmail.com
Doutora em Enfermagem pela Universidade de São Paulo (USP). Professora da Pontificia Universidade Católica do Paraná (PUCPR).

RESUMO

Introdução: As empresas de saúde armazenam uma grande quantidade de dados visando fundamentalmente o controle administrativo, pagamentos das contas médicas, etc., não havendo obrigatoriedade do preenchimento de dados clínicos epidemiológicos, entre os quais o CID – Código Internacional de Doenças. Este tipo de prática dificulta a identificação de possíveis enfermidades de seus beneficiários, a partir da utilização de técnicas de extração de informação tradicionais, e consequentemente, a implantação de programas de prevenção de doenças e de promoção da saúde.
Objetivo: Portanto, esse artigo propõe um modelo baseado em técnicas de mineração de dados para a identificação automática de beneficiários com propensão a doenças crônicas.
Metodologia: Metodologicamente esse modelo compreende as seguintes etapas: identificação inicial das variáveis e respectivas análises; seleção das variáveis a serem utilizadas e preparadas; mineração de dados e validação das

regras descobertas por especialistas. Objetivando testar o modelo proposto foi realizado um experimento voltado ao reconhecimento de indivíduos com propensão ao diabetes mellitus tipo 2.
Resultados: Para o processo de mineração de dados foram selecionadas 12 variáveis, considerando um conjunto de 43.375 beneficiários, sendo descobertas 843 regras, com uma taxa de acerto de 88,9%. Dentre essas 843 regras foram selecionadas seis para serem avaliadas por quatro especialistas.
Conclusões: Essa avaliação concluiu pela eficácia do modelo, com um grau de concordância da ordem de 89,6%.

Palavras-chave: Diabetes mellitus 2. Classificação. Descoberta do conhecimento em base de dados.

1 INTRODUÇÃO

A gestão da informação pode ser compreendida como sendo um dos núcleos da ciência da informação que atua em uma perspectiva de aperfeiçoamento do acervo da informação e do conhecimento para fins organizacionais. Constituindo um campo promissor para a formulação de novos conceitos, ágil para a compreensão de novos fenômenos e desenvolvimento de pesquisas (SEMIDÃO, 2013).

A discussão entre a gestão da informação e do conhecimento na administração oportuniza soluções e métodos alternativos para problemas de decisão no ambiente organizacional (ARAUJO, 2014). Neste contexto vale a atenção em preservar os registros de dados gerados nas organizações objetivando complementar as informações para apoio ao processo decisório a partir da descoberta de padrões.

Por outro lado, não é suficiente que os registros de dados continuem sendo preservados se não existe uma efetiva utilização. Por exemplo, Malta e Mehry (2010) afirmam a existência de uma “crise na promoção da saúde”. Esse setor encontra-se fortemente concentrado em um modelo produtor de procedimentos e seus respectivos registros, não estão sendo efetivamente considerados para determinações do processo saúde-doença. Esta não determinação produz custos

excessivos e crescentes, pois se utiliza de recursos tecnológicos centrados em exames e medicamentos. Comentam ainda sobre a crescente disponibilidade de novas tecnologias na área da saúde e a mudança no quadro atual de transição epidemiológica e demográfica que induz a uma maior incidência de Doenças Crônicas (DCs). As DCs levam o indivíduo enfermo a um longo período de latência, com progressivas manifestações da doença, especialmente se estiver sem acompanhamento adequado. Situação esta que motiva por ações que minimizem os custos da assistência à saúde, ampliando a prevenção e o seu respectivo controle (BRASIL, 2006a).

Diante deste quadro, Merhy e Franco (2008) destacam a importância de um novo modelo assistencial centrado no usuário e na qualidade da prestação dos serviços. Rodrigues e Anderson (2011) afirmam a necessidade de um modelo baseado em Medicina da Família e da Comunidade de forma a contribuir para o aprimoramento de serviços para prevenção de doenças e promoção da saúde.

No entanto, as operadoras de planos de saúde encontram dificuldades para atuar na prevenção de DCs por não disporem de dados clínicos epidemiológicos de seus beneficiários, sistematizados em suas bases de dados possibilitando a extração e informações e conhecimentos, por exemplo, por meios mais automatizados.

Entre outros motivos, isso se deve ao fato da construção dos sistemas de informação destas empresas ser orientada para o controle administrativo, visando pagamento dos prestadores de serviço e gestão de contratos.

Em 2007, uma determinação do Conselho Federal de Medicina desobrigou a notificação do CID - Código Internacional de Doenças nas guias dos atendimentos ambulatoriais. A ausência de dados clínicos epidemiológicos e também do CID dificulta a possibilidade de identificação, a recorrência de natureza específica dos serviços de saúde prestados a seus associados, ampliando a complexidade inerente

à definição e à implementação de ações de prevenção de doença e promoção da saúde.

Porém estas organizações possuem grande volume de dados referentes aos atendimentos de beneficiários que se constituem fonte potencial para pesquisa, a partir do qual é possível a descoberta de novas informações e padrões para o apoio a tomada de decisões (CARVALHO; TSUNODA; ESCOBAR, 2014; LUNARDELLI; TONELLO, MOLINA, 2014).

Não obstante a ausência de dados clínicos epidemiológicos, incluindo o CID, este grande volume de dados pode dificultar o uso de técnicas tradicionais, em geral adotadas, na implementação de sistemas de apoio a decisão. Uma alternativa que pode ser explorada é a descoberta de conhecimento a partir de ferramentas da tecnologia da informação, a exemplo do DW - Data Warehouse, KDD - Knowledge Discovery in Databases e mineração de dados.

Neste contexto, o objetivo desse artigo é a elaboração de um modelo baseado no processo KDD visando a identificação automática de beneficiários com propensão a doenças crônicas (DCs), voltado para o apoio à implantação de programas de prevenção de doenças e de promoção da saúde, bem como a sua respectiva validação. Serão utilizadas bases de dados de uma operadora de saúde, que em geral não dispõe de dados clínicos epidemiológicos, nem da especificação do CID, porém contém dados históricos, coletados e armazenados, referentes às demandas por procedimentos e exames realizados pelos seus beneficiários.

Modelos desta natureza constituem interessante instrumento para a gestão de planos de saúde e subsidiam uma melhor compreensão por parte dos beneficiários sobre as variáveis que contribuem ou não para a sua própria saúde e melhores práticas de promoção e prevenção (ARAÚJO, 2004).

Desenvolver medidas e reunir informações de resultados sobre tratamentos é um dos focos mencionados por Porter e Teisberg (2007), justificando a importância para os planos de saúde, de conhecer os ciclos de atendimento: diagnóstico, gerenciamento e prevenção de doenças, como estratégia para melhorar a qualidade dos serviços e reduzir os custos (PORTER; TEISBERG, 2007).

Vale destacar que DCs, tais como: o diabetes, o câncer, as cardiovasculares, a cirrose hepática, as pulmonares obstrutivas crônicas e os transtornos mentais, são consideradas prioridades na agenda das instituições voltadas para a prestação de serviços de saúde. Em 2007, aproximadamente 72% das mortes no Brasil foram atribuídas as DCs (SILVA JUNIOR, 2009).

Para facilitar o entendimento do modelo proposto alguns conceitos serão apresentados. O DW contribui para a estruturação e armazenamento de um conjunto de diversas fontes de dados, pois permite extrair informação de diversas formas atendendo grande parte das necessidades inerentes ao processo decisório (HARJINDER; PRAKASH, 1996). Complementarmente esta estruturação potencializa a utilização dos dados armazenados a partir do processo KDD.

O processo KDD permite identificar padrões nos dados, agregando conhecimento ao gestor (FAYYAD; PIATESKY-SHAPIRO; SMYTH, 1996). A Figura 1 apresenta o KDD com as suas diversas fases e respectivas sequências.

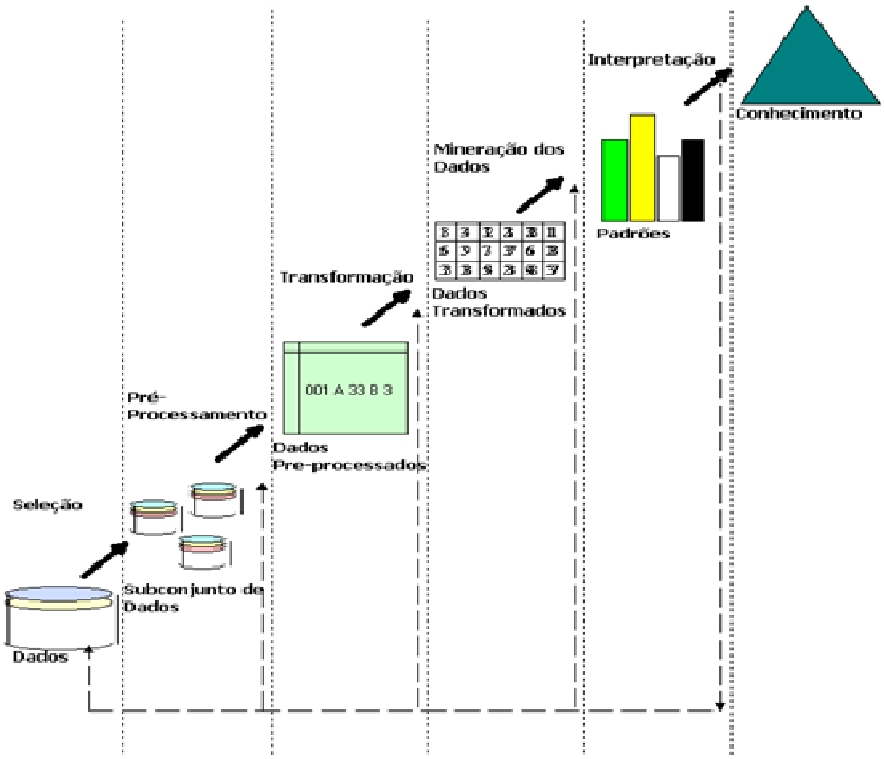
São três as etapas principais do processo de KDD: pré-processamento, mineração de dados e pós-processamento. Durante o pré-processamento as diversas fontes de dados são identificadas, selecionadas e preparadas as variáveis que serão processadas na etapa de mineração de dados.

Entre as diversas tarefas da mineração destaca-se a classificação, que consiste na construção de um modelo a partir do qual

é possível prever o valor de um determinado atributo de interesse, denominado classe (TAN; STEINBACH; KUMAR, 2005).

Os classificadores podem ser representados, entre outras formas, por árvores de decisão que constituem estruturas hierárquicas compostas por um nó denominado raiz a partir do qual derivam várias ramificações, compostas pelos atributos considerados “mais relevantes” para a construção do modelo. Este conjunto de ramificações é concluído pelos nós folhas que representam os valores do atributo que caracteriza a classe a ser predita (HAN; KAMBER, 2006).

Figura 1 - Etapas do processo de descoberta de conhecimentos.



Fonte: Fayyad, Piatesky-Shapiro e Smyth (1996)

Entre os algoritmos que descobrem padrões e que representam na forma de árvores de decisão, tem-se o J48 disponível no ambiente WEKA (UNIVERSITY OF WAIKATO, 2007) que constitui uma versão do algoritmo C4.5 (QUINLAN, 1993).

Na etapa de pós-processamento os padrões descobertos são avaliados. Por exemplo, a representação dos padrões descobertos a partir de árvores de decisão permite interpretar cada percurso entre o nó raiz e o nó folha como uma regra do tipo “Se – Então”. Este fato possibilita que o gestor compreenda e analise a relação representada pela regra, entre os atributos e a classe predita, buscando, assim, aplicar sobre a situação problema que originou o processo por busca de padrões ou por novos conhecimentos (STEINER et al., 2006).

Para a avaliação da capacidade preditiva do classificador descoberto existem várias formas, uma das mais utilizadas é a taxa de acerto - a quantidade de registros classificados corretamente pelo número do total de registros Esta medida é calculada sobre o(s) subconjunto(s) segmentado(s) do conjunto original disponível não considerado(s) durante a construção da árvore de decisão.

É possível identificar na literatura alguns relatos do uso da tarefa de classificação para apoiar processos decisórios envolvendo DCs, como por exemplo, Toussi et al. (2009) apresentam um modelo baseado no algoritmo de árvore de decisão C5.0, para a análise das diretrizes francesas para o diabetes mellitus tipo 2. O conjunto de dados utilizado por esses autores foi composto por dados não apenas administrativos, mas também antropométricos e clínicos. Este estudo, além de destacar a capacidade preditiva do classificador descoberto e representado na forma de árvore de decisão, também valoriza a possibilidade de compreensão permitindo a sua análise frente às diretrizes francesas.

Soni et al. (2011) apresentam os resultados de uma pesquisa que avaliou algumas das técnicas de mineração de dados que vem sendo utilizadas particularmente para a predição de doenças do coração. Os resultados obtidos revelaram que árvore de decisão superou o naive bayes quanto ao tempo necessário para a sua execução, apesar da qualidade preditiva ser semelhante entre as duas técnicas. Os resultados também apontam que estes dois métodos (árvore de decisão

e naive bayes) superam os resultados obtidos pelas técnicas KNN, Redes Neurais e Classificação baseada em agrupamentos.

Marinov et al. (2011) apresentaram uma revisão sistemática envolvendo 17 estudos aplicando técnicas de mineração de dados para questões relacionadas ao diabetes, dos quais 10 utilizaram, entre outras, a tarefa de classificação, sendo adotada a representação do classificador por árvore de decisão em seis deles. Entre os algoritmos utilizados estão o CART, C4.5 e o C5.0.

2 MATERIAIS E MÉTODOS

O modelo proposto baseia-se no processo KDD, adequando-o para a identificação de beneficiários com potencial indicativo de doenças crônicas, sendo adotado o Diabetes Mellitus tipo 2. Foram observadas as seguintes etapas do processo KDD: análise inicial da situação problema, seleção das variáveis, preparação dos dados para a aplicação, mineração de dados, preparação das regras descoberta e avaliação e validação das regras por especialistas.

A escolha do Diabetes Mellitus tipo 2 deve-se a sua crescente prevalência em países em desenvolvimento, associado ao fato de vários estudos comprovarem a possibilidade da sua redução a partir de ações preventivas, a exemplo de mudanças no estilo de vida (FERREIRA et al., 2005). Desta forma se justifica a proposta de novas estratégias que facilitem a identificação destes indivíduos objetivando a prevenção.

Este estudo é de natureza quantitativo, retrospectivo, de cunho descritivo que se utilizou de uma base de dados de uma operadora de plano de saúde do estado do Paraná, com autorização da instituição e submissão ao comitê de ética em pesquisa da PUC-PR, segundo o protocolo de nº 0001638/08.

A etapa de análise inicial foi realizada a partir de retrospectiva histórica para a investigação das principais características, consideradas

como importantes para a identificação de beneficiários com potencial de desenvolvimento de doença crônica (ALMEIDA FILHO; ROUQUAYROL, 2006).

Para essa análise foram criados dois grupos de beneficiários que passaram por internamentos no terceiro trimestre de 2007:

- Grupo 1, todos os beneficiários que tiveram internamento, pelo CID de diabetes Mellitus 2, totalizando 59 indivíduos com idade ≥ 25;
- Grupo 2 (pareado), selecionou-se 59 indivíduos de forma aleatória que tiveram internamento clínico também com idade ≥ 25, porém sem a ocorrência do CID referente ao diabetes.

Este período de tempo se justifica, pois ainda não estava em vigor a não obrigatoriedade da notificação do CID, o que permitiu a construção do Grupo 1.

Construídos esses dois grupos de beneficiários (Grupo 1 e 2), foram recuperados seus respectivos atendimentos referentes às consultas por especialidade médica e solicitações de exames de forma retrospectiva aos cinco anos anteriores, permitindo a comparação da demanda entre os dois grupos para a definição do quadro de variáveis do estudo.

A partir da análise comparativa foi possível identificar as variáveis (consultas por especialidades médicas e exames) que se diferenciaram nos dois grupos, resultando na seleção daquelas mais relevantes para a construção do conjunto de dados a ser submetido a mineração de dados. O critério adotado para caracterizar a “diferença” entre as varáveis foi:

- Considerando especialidades médicas – dentre aquelas, do Grupo 1, que apresentaram um diferencial de frequência

superior a 30% em relação a respectiva frequência obtida no Grupo 2, foram selecionadas as três com maior frequência;

- Considerando exames - dentre aqueles, do Grupo 1, que apresentaram um diferencial de frequência superior a 25% em relação a respectiva frequência obtida no Grupo 2, foram selecionados os dois com maior frequência. Para complementar a relação dos exames foram consideradas sugestões de especialistas e de referências bibliográficas.

Para a construção do conjunto de dados para a mineração de dados, foram sistematizados os atendimentos dos beneficiários ao longo de seis anos, entre janeiro de 2007 a dezembro de 2012, considerando os 43.375 beneficiários ainda ativos em dezembro de 2012. Este intervalo de seis anos foi adotado em função da disponibilidade de dados no DW da operadora. Os dados 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.

Complementarmente a cada beneficiário foi atribuída a sua respectiva classe identificadora, construída conforme os seguintes critérios:

- Se existência de atendimento com CID de diabetes ou número de solicitações de exame de hemoglobina glicosilada <= 2 então classe = “sem indicativo para DM2”;
- Se existência de atendimento com CID de diabetes ou número de solicitações de exame de hemoglobina glicosilada > 2 então classe = “com indicativo para DM2”;

Esses critérios avaliam a quantidade de atendimentos considerando o CID do diabetes e o número de exames de hemoglobina glicosilada referentes ao período de seis anos, conforme Diretrizes da Sociedade Brasileira do Diabetes – SBD. Esta diretriz recomenda que a frequência do exame de hemoglobina glicosilada, seja de pelo menos duas vezes ao ano para todos os diabéticos e quatro vezes ao ano para aqueles submetidos às alterações no esquema terapêutico (SBD, 2007).

Para a etapa de mineração de dados foi selecionado um algoritmo que descobre classificadores representados na forma de árvore de decisão, por facilitar a compreensão e a análise das relações entre as variáveis envolvidas (STEINER et al., 2006). O algoritmo utilizado foi o J48 (UNIVERSITY OF WAIKATO, 2007). Para validar a taxa de acerto foi adotada a segmentação do conjunto original 66% para a construção do classificador e 34% para teste, distribuição aleatória já disponibilizada pelo próprio algoritmo.

A árvore de decisão descoberta foi transformada pelo algoritmo que pós-processa árvore de decisão (PAD), em regras no formato “Se (condições) Então (valor da classe predita).” (CARVALHO; MILANI;, 2013).

Para a seleção das regras, a serem avaliadas pelos especialistas, foi adaptado o método proposto por Carvalho (2005). Este método propõe ranquear todas as regras descobertas por dois critérios, conforme a seguinte precedência: (1) primeiro em relação a cobertura (número de registros atendidos pelo antecedente da regra) e (2) a taxa de acerto. A partir do conjunto de regras ranqueado foram selecionadas as três regras com melhor ranqueamento respectivamente predizendo cada uma das três classes (sem indicativo para DM2, com indicativo para DM2 e com forte indicativo para DM2), totalizando assim nove regras selecionadas.

Na etapa de avaliação o objetivo foi verificar o quanto as regras agregaram valor ao que os especialistas já conheciam. Sendo assim foi

possível perceber se a mineração de dados descobriu regras que oportunizaram novos elementos em relação aquele conhecimento prévio dos especialistas.

Para esta avaliação foi elaborado um formulário apresentado a quatro especialistas da área médica em DCs. Todos os especialistas foram esclarecidos sobre o objetivo da pesquisa, receberam orientação necessária para o devido preenchimento e concordaram em participar do estudo. O prazo para preenchimento e devolução do formulário foi de uma semana.

O 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 equivoco ou erro (escore = 1).
- Discordo “D” = A regra contraria o conhecimento, mas apresenta uma ou mais condições no antecedente da regra, que representa equivoco ou erro (escore = 0).

A partir da avaliação e identificação dos respectivos escores para as nove regras foi aplicado o Índice de Validade de Conteúdo (IVC), proposto por Waltz, Strickland e Lenz (1991), que estima a validade de um item de mensuração, com base no julgamento dos especialistas. O IVC é obtido a partir da seguinte expressão:

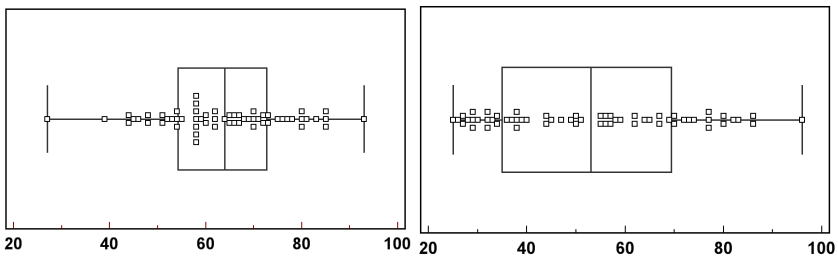
IVC = Total dos Pontos Obtidos / Total Máximo Possível de Pontos * 100.

3 RESULTADOS

Os resultados obtidos a partir do experimento realizado para validar a metodologia proposta são apresentados a seguir.

A partir da figura 2 é possível constatar a diferença entre os dois grupos. No Grupo 1, o primeiro quartil é de 54 anos, a média de 66 anos e o terceiro quartil é de 75 anos. No Grupo 2, tem-se o primeiro quartil de 35 anos, a média de 53 e o terceiro quartil de 69 anos, dessa maneira percebe-se a diferenciação da variável idade em relação aos dois grupos.

Figura 2 - Comparativo da idade entre os Grupos 1 e 2



Fonte: Produzida pelos autores

Os resultados da análise inicial confirmam os fatores de risco ao DM2 apresentados no Caderno de Atenção Básica do Ministério da Saúde sobre o DM2, incluindo idade superior a 45 anos (BRASIL, 2006b) e com as diretrizes da Sociedade Brasileira de Diabetes que referenciam que, em geral, o DM2 é diagnosticado após os 40 anos de idade (SBD, 2007).

Na comparação da quantidade de atendimentos, realizados por especialidade médica, entre Grupo 1 e o Grupo 2 (Tabela 1), destaca-se:

- a endocrinologia, como especialidade característica do tratamento do DM2; e

- a cardiologia, pela identificação de problemas cardiovasculares e cerebrovasculares como complicações do DM2 (SBD, 2007).

A especialidade oftalmologia foi mantida no estudo, pois a literatura refere que, em algumas vezes, o diagnóstico do DM2 é realizado a partir de suas complicações crônicas como, por exemplo, a neuropatia e a retinopatia (BRASIL, 2006b).

Tabela 1 – Número de consultas realizadas pelos Grupos 1 e 2, segundo especialidades

Especialidades	Grupo 1	Grupo 2
Clínica Médica	432	365
Cardiologia	236	153
Endocrinologia	177	94
Oftalmologia	166	189
Ginecologia	84	191
Nefrologia	75	1
Psiquiatria	66	19
Gastroenterologia	54	23
Geriatria	54	5
Ortopedia	54	151
Neurologia	46	68

Fonte: Produzida pelos autores

No que se refere aos exames realizados pelos beneficiários (Tabela 2) observam-se diferenças na comparação do Grupo 2, internamento clínico, com o Grupo 1 de beneficiários internados por DM2, evidenciando a importância e a caracterização do exame Hemoglobina Glicada ou Hemoglobina Glicosilada como um exame de avaliação inicial para o diagnóstico do diabetes (BRASIL, 2006b). Esta variável foi utilizada como critério para a construção do atributo classe para a construção do classificador.

Tabela 2 – Número de exames realizados pelos grupos 1 e 2, segundo respectivos tipos

Exames Realizados	Grupo 1	Grupo 2
Glicose	1512	704
Creatinina	887	659
Hemograma Completo	840	838
Potássio	613	430
Sódio	496	378
Ureia	459	415
Colesterol Total	447	343
Hemoglobina Glicosilada	444	57
Rotina de Urina	398	314
Triglicerídeos	379	328

Fonte: Produzida pelos autores

A partir dos resultados foram identificadas 12 variáveis (Quadro 1) a serem consideradas para a construção do classificador, agrupadas em exames laboratoriais e especiais, consultas realizadas e outras.

Quadro 1 - Atributos selecionados para o processo de mineração de dados

Exames Laboratoriais	Exames Especiais	Consultas Especialidades	Outras
Glicose Creatinina Microalbuminúria Colesterol Total	Curva Glicêmica Mapeamento de Retina	Oftalmologia Endocrinologia Nefrologia Cardiologia	Idade CID

Fonte: Produzida pelos autores

São preconizados os exames de fundo de olho (mapeamento de retina), de Glicose (glicemia de jejum), hemoglobina glicosilada (A1C), creatinina e microalbuminuria para a avaliação clínica inicial de pacientes com diabetes (BRASIL, 2006b), além do próprio exame físico para o DM2. Kobus (2006) também referência a microalbuminuria, hemoglobina glicosilada e o mapeamento de retina como procedimentos específicos e indicativos do DM2.

Sobre o conjunto de 43.375 beneficiários foram selecionados os dados referentes as 12 variáveis compondo assim o subconjunto de dados para a etapa de mineração. A partir deste subconjunto foi

descoberta a árvore de decisão composta por 843 ramificações que apresentou a seguinte a matriz de confusão (Tabela 3).

Tabela 3 - Matriz de Confusão

Previsto		**Real**
Sem indicativo	Com indicativo	
12.296	475	Sem Indicativo
1.154	822	Com Indicativo

Fonte: Produzida pelos autores

A classe “sem indicativo para DM2” obteve uma cobertura de 12.771, com 12.296 registros classificados corretamente, o que corresponde a 96,3% de acerto. Já a taxa de acerto individual para a classe “com indicativo para DM2” foi 42%, sendo a taxa geral 88,9%.

A tabela 4 apresenta as seis regras selecionadas e apresentadas para avaliação aos especialistas, bem como o escore atribuído respectivamente para cada regra.

Tabela 4 - Validação das regras pelos quatro especialistas

Regra	**A**	**B**	**C**	**D**	**Total**
1 – SE Exame Microalbuminuria = 0 E Exame Glicose <= 7 ENTAO Sem Indicativo	2	2	2	2	8
2 – SE Exame Microalbuminuria = 0 E Curva Glicêmica <=0 E Exame Glicose > 7 E <=9 E Sexo = ‘F’ E Consulta Cardiologia <=6 ENTAO Sem Indicativo	1	2	2	2	8
3 – SE Exame Microalbuminuria = 0 E Curva Glicêmica <=0 E Exame Glicose > 9 E <=11 E Idade > 38 E Exame Creatinina <=12 E Sexo = ‘F” E Consulta Oftalmologia > 0 E Consulta Endocrinologista <= 2 ENTAO Sem Indicativo	1	2	2	2	7
4 – SE Exame Microalbuminuria > 0 E Exame Glicose > 12 E Consulta Endocrinologista > 1 ENTAO Com Indicativo	2	2	2	2	8
5 – SE Exame Microalbuminuria > 0 E Consulta Endocrinologista <= 1 E Sexo=’M’ E Exame Glicose > 14 ENTAO Com Indicativo	1	2	2	2	7

6 – SE Exame Microalbuminuria = 0 E Exame Glicose > 14 E <= 17 E Idade > 40 E Consulta Endocrinologista > 4 E Sexo 'F' E Exame Creatinina < = 15 E Exame Curva Glicêmica <= 0 ENTAO Com Indicativo	1	1	2	2	6

Fonte: Produzido pelos autores
Legenda (A, B, C e D): (2 – Concorda); (1 – Concorda Parcialmente); (0 – Discorda).

A análise do IVC (Waltz *et al.* ,1991) sobre a avaliação dos especialistas, em relação as seis regras selecionadas, obteve concordância, em relação ao seu conteúdo, da ordem de 89.6%.

4 DISCUSSÃO

O processo KDD foi aplicado sobre dados de uma operadora de plano de saúde, sendo extraídas regras a partir da mineração de dados, que puderam ser avaliadas por especialistas da área médica, para a classificação de beneficiários em relação ao diabetes mellitus tipo 2.

A partir da construção do classificador foi obtida uma taxa de acerto geral de 88,9% referente as classes "sem indicativo ao DM2" e "com indicativo ao DM2" (Tabela 3). Sobre o fato da diferença entre as taxas de acerto individuais para as classes "sem indicativo ao DM2" (96,3%) e "com indicativo ao DM2" (42%) não invalida os resultados tendo em vista que os beneficiários não classificados como "sem indicativo ao DM2" são encaminhados para o programa de gerenciamento de casos. Situação que também corroborada para o IVC=89,6%, obtido a partir da avaliação dos especialistas.

Este modelo proposto e testado considerando a DM2, foi posteriormente replicado, na mesma operadora de saúde, para a descoberta de regras que permitissem identificar beneficiários propensos a outras DCs, tais como hipertensão, isquemias do coração, neoplasias, doenças pulmonares obstrutivas crônicas, renal crônica, obesidade e psiquiátricas.

Vale destacar que além da contribuição científica da proposta apresentada neste artigo, também representa uma contribuição social, pois a partir dos resultados obtidos pelos experimentos – original e diversas replicações - foi criado um "portal" para a operadora de plano de saúde em questão, a partir do qual é possível selecionar, de forma automática, os beneficiários para encaminhamento aos diversos programas de prevenção de doença e promoção da saúde.

Adicionalmente esses resultados da aplicação do modelo proposto, sobre os dados da operadora de plano de saúde, apresentarem ser eficientes para a elegibilidade de beneficiários com potencial para evolução para DCs, se comparados às estatísticas disponíveis. Por exemplo, a partir das regras descobertas, para a identificação de beneficiários como elegíveis para o programa de diabéticos, foram indicados 5.953 beneficiários, representando 5,7% do total de beneficiário da carteira. Este resultado está compatível com o manual do VIGITEL de 2012 (BRASIL, 2012) que apresenta 5,6% de adultos com diagnostico médico referido para o diabetes.

5 CONCLUSÃO

As operadoras de planos de saúde possuem uma grande quantidade de dados relativos aos atendimentos realizados pelos seus beneficiários, porém pela forma como são disponibilizados não facilitam a extração de novos conhecimentos a partir de estratégias "ditas" tradicionais. Esta dificuldade se agrava pela não disponibilidade de dados clínicos epidemiológicos dos beneficiários.

Esse fato se justifica pela privacidade garantida aos usuários, inclusive pela não mais obrigatoriedade da informação do CID. Porém, para que essas operadoras implantem programas de promoção da saúde e prevenção de doenças é preciso recorrer às tecnologias de informação alternativas.

Este artigo apresentou e testou um modelo baseado no processo KDD que, a despeito das dificuldades inerentes à ausência de dados relevantes (CID, p. ex.), permitiu identificar beneficiários com propensão a DM2. Apesar de ter sido apresentado e testado para a DM2, o modelo também foi replicado para outras DCs, demonstrando o seu potencial de generalização na aplicação.

O modelo mostrou-se eficiente podendo ser aplicado para a seleção de beneficiários para programas de prevenção de DCs, programas estes que proporcionam a redução de custos para a operadora de planos de saúde, promoção da saúde e a melhoria da qualidade de saúde e de vida para os seus beneficiários.

A vantagem competitiva de uma organização possui dependência direta com a sua capacidade de tomada de decisões, da adoção de estratégias e ações diárias, nas quais seus colaboradores contribuam para a geração de resultados. Sendo a informação oportuna matéria prima essencial para que este objeto seja alcançado (ANDRADE; BARRETO, 2015).

REFERÊNCIAS

ALMEIDA FILHO, Naomar de, ROUQUAYROL, Maria Zélia. **Introdução à epidemiologia**. Rio de Janeiro: Guanabara Koogan; 2006.

ANDRADE, Antonio Rodrigues de; BARRETO, Aldo de Albuquerque. Alinhamento estratégico nas organizações a informação como elemento integrador de propósito, processos e pessoas. **DataGramaZero Revista de Informação**, Rio de Janeiro, v. 16, n. 1, fev. 2015.

ARAÚJO, Carlos Alberto Ávila. O que é ciência da informação? **Informação & Informação**, Londrina, v.19, n. 1, p.1-30, 2014.

ARAÚJO, Mário Luiz Cardoso de. **Gerencia de assistência à saúde no setor de saúde suplementar**: uma experiência. 2004. Dissertação (Mestrado em Saúde Pública) - Escola Nacional de Saúde Pública, fundação Oswaldo Cruz, Rio de Janeiro, 2004.

BRASIL. Agência Nacional de Saúde Suplementar. **Manual técnico de promoção da saúde e prevenção de riscos e doenças na saúde suplementar**. Rio de Janeiro: ANS; 2006a.

BRASIL. Ministério da Saúde**. Diabetes mellitus**. Brasília, 2006b. (Cadernos de Atenção Básica, n. 16).

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Vigitel Brasil 2011**: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico. Brasília, 2012. (Série G. Estatística e Informação em Saúde).

CARVALHO, Debora Ribeiro. **Algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados**. 2005.Tese (Doutorado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

CARVALHO, Deborah Ribeiro, TSUNODA, Denise, ESCOBAR, Leandro Fabian. Pontos de atenção para o uso da mineração de dados na saúde. **Informação & Informação**, Londrina, v. 19, n. 3, p. 249-273, 2014.

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery: an overview**. Cambridge: AAAI Press; 1996.

FERREIRA Sandra. et al. Intervenções na prevenção do diabetes mellitus tipo 2: é viável um programa populacional em nosso meio? **Arquivos Brasileiros de Endocrinologia & Metabologia**, São Paulo, v. 49, n. 4, p. 479-484, 2005.

HAN, Jiawei, KAMBER, Micheline. **Data mining**: concepts and techniques. 2nd ed. California: Morgan Kaufmann Publishers; 2006.

HARJINDER, Hill. S.; PRAKASH, Rao. C. **The official guide to data warehousing.** Indianopolis: QUE Corporation; 1996.

LUNARDELLI, Rosane Suely Alvares, TONELLO, Izângela Maria Sansoni, MOLINA Letícia Gorri. A constituição da memoria dos procedimentos em saúde no contexto do prontuário eletrônico do paciente. **Informação & Informação**, Londrina, v. 19, n. 3, p. 107-124, 2014.

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.

MALTA, Deborah Carvalho; MERHY, Emerson Elias. O percurso da linha do cuidado sob a perspectiva das doenças crônicas não transmissíveis. **Interface – Comunicação Saúde Educação**, Botucatu, v. 14, n. 34, p. 593-605, 2010.

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.

MERHY, Emerson Elias; FRANCO, Túlio Batista. Reestruturação produtiva em saúde. In: FIOCRUZ. **Dicionário da educação profissional em saúde**. 2. ed. Rio de Janeiro: EPSJV, 2008. p. 348-252.

CARVALHO, Deborah Ribeiro; MILANI, Cristian Simioni. Pós-Processamento em KDD. **Revista de Engenharia e Tecnologia**, Ponta Grossa, v. 5, n. 1, p. 151-162. 2013.

PORTER, Michael E.; TEISBERG, Elizabeth Olmsted. **Repensando a saúde**: estratégia para melhorar a qualidade e reduzir os custos. Porto Alegre: Bookman; 2007.

QUINLAN J. Ross. **C45**: programs for machine learning. San Mateo, California: Morgan Kaufmann; 1993.

RODRIGUES, Ricardo Donato; ANDERSON, Maria Inez Padula. Saúde da família: uma estratégia necessária. **Revista Brasileira de Medicina de Família e Comunidade**, Florianópolis, v. 6, n. 18, p. 21-24, 2011.

SBD - SOCIEDADE BRASILEIRA DE DIABETES. **Diretrizes da Sociedade Brasileira de Diabetes**. Rio de Janeiro: DIAGRAPHIC; 2007.

SEMIDÃO, Rafael Aparecido Moron. Dados, informação e conhecimento: elementos de análise conceitual. **DataGramaZero – Revista de Informação**, Rio de Janeiro, v. 14, n. 4, ago. 2013.

SILVA JUNIOR, Jarbas Barbosa. As doenças transmissíveis no Brasil: tendências e novos desafios para o Sistema Único de Saúde. In: BRASIL. Ministério da Saúde. **Saúde Brasil 2008**: 20 anos de Sistema Único de Saúde (SUS) no Brasil. Brasília: Ministério da Saúde, 2009.

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.

STEINER, Maria Terezinha Arns et al. Abordagem de um problema médico por meio do uso do processo KDD com ênfase em análise exploratória dos dados. **Revista Gestão e Produção**, Florianópolis, v. 13, n. 2, p. 325-337, 2006.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to data mining**. Boston: Pearson Addison-Wesley Longman Publishing Co.; 2005.

TOUSSI, Massoud et al. Using data mining techniques to explore physicians 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, 2009.

UNIVERSITY OF WAIKATO. **Weka 3**: machine learning software in java. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 7 mar. 2007.

WALTZ, Carolyn Feher; STRICKLAND, Ora Lea; LENZ, Elizabeth R. **Measurement in nursing and health research**. Philadelphia USA: F.A. Davis Company, 1991.

Title

The use of Data Mining techniques to Automated Detection of Beneficiaries With Indicative of Diabetes Mellitus 2

Abstract

Introduction: The Health Industry companies store a vast amount of data in order to support administrative tasks like payment of medical bills, but filling out epidemiological data (International Classification of Diseases - ICD) is not mandatory. This makes it difficult to identify the persons' illness using standard data extraction techniques as well as implementing preventive programs.
Objective: This paper proposes a data mining model that identifies automatically the patients with chronic illnesses.
Method: The proposed method is comprised of the following steps: initial identification of the variables and their analysis; variable selection; data mining and rule validation by experts. An experiment, for identifying the patients with propensity for diabetes type 2, was designed to validate the methodology.
Results: For the data mining process, 12 variables were selected, targeting 43.375 patients: 843 rules were discovered, with a 88,9% success rate.

Conclusion: From the 843 rules, six were selected to be evaluated by four experts: they considered the model efficient, with an 89.6% rate of positive results.

Keywords: Diabetes Mellitus 2. Classification. Knowledge discovery in databases.

Titulo

Utilizando técnicas de minería de datos para la identificación automática de los beneficiarios propensos a la diabetes mellitus tipo 2

Resumen

Introducción: Las empresas de salud almacenan una gran cantidad de datos objetivando fundamentalmente el control administrativo, pagos de las facturas médicas, etc., no habiendo obligatoriedad de completar datos clínicos epidemiológicos, entre ellos el CID – Código Internacional de Enfermedades. Este tipo de práctica dificulta la identificación de posibles enfermedades de sus beneficiarios, a partir de la utilización de técnicas de extracción de informaciones tradicionales, y consecuentemente, la implantación de programas de prevención de enfermedades y de promoción de la salud.
Objetivo: Por lo tanto, este artículo propone un modelo basado en la utilización de técnicas de minería de datos para la identificación automática de beneficiarios con propensión a enfermedades crónicas.
Metodología: Metodológicamente ese modelo comprende las siguientes etapas: identificación inicial de las variables y respectivos análisis; selección de las variables a ser utilizadas y preparadas; minería de datos y validación de las reglas descubiertas por especialistas. Objetivando probar el método propuesto fue realizado un experimento dirigido al reconocimiento de individuos con propensión a la diabetes mellitus tipo 2.
Resultados: Para el proceso de minería de datos fueron seleccionadas 12 variables, considerando un conjunto de 43.375 beneficiarios, y descubiertas 843 reglas, con una tasa de acierto de 88,9%.
Conclusión: De entre ellas 843 reglas fueron seleccionadas nueve para ser evaluadas por cuatro especialistas. Esa evaluación concluyó por la eficacia del modelo, con un grado de concordancia del orden de 89,6%.

Palabras-clave: Diabetes Mellitus 2. Clasificación. Descubrimiento del conocimiento en base de datos.

Recebido em: 20/08/2014
Aceito em: 25/05/2015