JHI JOURNAL OF HEALTH INFORMATICS

Mineração de dados aplicada ao conhecimento em uma população universitária

Data Mining applied to knowledge in a university population

La minería de datos aplicada al conocimiento en población universitária

Priscyla Waleska Targino de Azevedo Simões[1], Felipe Ribeiro Sampaio[2], Jose Márcio Cassettari Júnior[3], Samuel Cesconetto[4], Maria Inês da Rosa[5]

Descritores: Inteligência Artificial; Mineração de Dados; Tabagismo; Epidemiologia Descritiva

RESUMO

Objetivo: Descrever o processo de mineração de uma base de dados de tabagismo obtida de uma população universitária do Sul de Santa Catarina. **Métodos:** Estudo de natureza aplicada (tecnológica), transversal, de campo e laboratório, e descritivo. A base de dados utilizada foi de um estudo de prevalência realizado na Universidade do Extremo Sul Catarinense no segundo semestre de 2010, resultando em 575 registros. Foi realizado pré-processamento; em seguida, mineração de dados, primeiro pela clusterização fuzzy, sucedida pela tarefa de classificação; última etapa abordou a avaliação das árvores e regras geradas. **Resultados:** Foram realizados mais de 300 experimentos, resultando em 524 regras, 339 oriundas da base completa, e 185 da clusterizada de fumantes. Na base completa obteve-se sensibilidade 98,0[IC95%=(97,0;99,0)], especificidade 87,0[IC95%=(97,0;99,0)], acurácia 98,0[IC95%=(79,0;94,0)]; a base clusterizada resultou em sensibilidade 84,0[IC95%=(78,0;90,0)], especificidade 73,0[IC95%=(61,0;86,0)], acurácia 82,0[IC95%=(74,0;89,0)]. **Conclusão:** O perfil epidemiológico dos tabagistas resultante das regras geradas em nosso estudo foi semelhante da literatura.

Keywords: Artificial Intelligence; Data Mining; Smoking; Epidemiology Descriptive

ABSTRACT

Objective: To describe the process of Data Mining of a smoking database obtained from a university population in the South of Santa Catarina. **Methodos**: Descriptive, laboratory and camp, transversal, technologic nature study. The database used was originated from a prevalence study in the second semester of 2010, at the University do Extremo Sul Catarinense, which has resulted 575 registers. In the beginning the preprocessing was performed; next, the Data Mining, first trough fuzzy clustarization, followed by the classification; last step assessed the generated rules. **Results**: More than 300 experiments were performed, resulting 524 rules, 339 originated from the complete non-clusterized database, and 185 from smoking cluster. The complete database showed sensitivity 98,0[CI95%=(97,0;99,0)], specificity 87,0[CI95%=(97,0;99,0)] and precision 98,0[CI95%=(79,0;94,0)]; the smoking clusterized database resulted sensitivity 84,0[CI95%=(78,0;90,0)], specificity 73,0[CI95%=(61,0;86,0)] and precision 82,0[CI95%=(74,0;89,0)]. **Conclusion**: The epidemiologic profile of the tobacco users resultant of the generated rules in our research was similar to the literature.

Descriptores: Inteligencia Artificial; Minería de Datos; Fumar; Epidemiología Descriptiva

RESUMEN

Objetivo: Describir el proceso de minería de una base de datos del consumo de tabaco obtenido de una población universitaria del Sur de Santa Catarina. **Metodos:** Estudio de naturaleza aplicada (tecnología), ámbito transversal y de laboratorio, y descriptivo. La base de datos utilizada fue un estudio de prevalencia realizado en Universidad del Extremo Sul Catarinense el segundo semestre de 2010, que resulta en 575 registros. Hemos llevado a cabo pre-procesamiento, a continuación, la minería de datos, primero por agrupamiento difuso, logrado por la tarea de clasificación; etapa final se dirigió a la evaluación de los árboles y las reglas generadas. **Resultados:** Realizaron más de 300 experimentos, que resulta las normas 524, 339 que surgen de la base completa, y 185 de los fumadores clúster. Base sólida la sensibilidad se obtuvo 98,0[IC95%=(97,0;99,0)], especificidad 87,0[IC95%=(97,0;99,0)], precisión 98,0[IC95%=(79,0;94,0)]; basado en clúster produjo sensibilidad 84,0[IC95%=(78,0;90,0)], la especificidad 73,0[IC95%=(61,0;86,0)], exactitud 82,0[IC95%=(74,0;89,0)]. **Conclusión:** El perfil epidemiológico de los fumadores que resultan de las normas generadas en nuestro estudio fue la literatura similar.

[1] *Doutora em Ciências da Saúde – Universidade do Extremo Sul Catarinense - UNESC, Criciúma (SC), Brasil.*
[2] *Graduado em Medicina pela Universidade do Extremo Sul Catarinense - UNESC, Criciúma (SC), Brasil.*
[3] *Graduado em Ciência da Computação pela Universidade do Extremo Sul Catarinense - UNESC, Criciúma (SC), Brasil.*
[4] *Graduando em Medicina pela Universidade do Extremo Sul Catarinense - UNESC, Criciúma (SC), Brasil.*
[5] *Doutora em Epidemiologia – Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre (RS), Brasil.*

Autor Coorespondente: **Priscyla Waleska Targino de Azevedo Simões**
e-mail: **pri@unesc.net**

Artigo recebido: **21/06/2013**
Aprovado: **13/08/2013**

www.jhi-sbis.saude.ws

INTRODUÇÃO

Globalmente estima-se que cerca de 21% da população fumem, a Organização Mundial de Saúde (OMS) estima que mais de 6 milhões de pessoas irão morrer até o final de 2011 vítimas de doenças atreladas ao uso do cigarro, o que faz do tabagismo a principal causa de morte evitável do mundo[1].

A Descoberta de Conhecimento em Bases de Dados ou Knowledge Discovery in Databases (KDD) ocupa-se do desenvolvimento de métodos e técnicas que visam extrair conhecimento de grandes bancos de dados. O processo de KDD é definido como a identificação de padrões e modelos úteis às bases de dados, sendo formado por várias etapas, dentre as quais, o data mining (DM) é a mais importante, definida pelo conjunto de métodos e processos utilizados para executar a descoberta de conhecimento[2-3].

O DM pode ser aplicado em variadas áreas de atuação da saúde incluindo a clínica, radiologia, saúde pública, entre outras, e possui grande utilidade para a geração de hipóteses, assim, a descoberta de modelos preditivos por meio do DM pode ser útil na prática clínica podendo auxiliar ao processo de tomada de decisões[2, 4-6]. O objetivo principal da mineração de dados preditiva na informática clínica visa produzir modelos de predição de doenças por exemplo, e dessa maneira, atua para apoiar as decisões clínicas na prática médica[6]. As predições clínicas variam desde a simples estratificação de risco de uma população de pacientes, com base em fatores de risco conhecidos, como idade ou estilo de vida, até o desfecho de uma determinada doença, incluindo a previsão dos efeitos a que um tratamento ou uma droga específica podem submeter um grupo de indivíduos[2, 4].

O presente trabalho tem o objetivo de descrever o processo de mineração de uma base de dados de tabagismo obtida de uma população universitária do Sul de Santa Catarina.

MÉTODOS

Estudo de natureza aplicada (tecnológica), transversal, de campo e laboratório, e descritivo. A base de dados utilizada nesta pesquisa foi oriunda de um estudo de prevalência realizado na Universidade do Extremo Sul Catarinense (UNESC) no segundo semestre de 2010, aprovado pelo Comitê de Ética sob o protocolo 153/2009, cujo principal objetivo buscou avaliar a prevalência de tabagismo e fatores associados entre universitários, traçando um perfil destes estudantes para uma melhor abordagem com medidas educativas que visem diminuir o número de fumantes nesta população[7].

A população foi composta pelos 8400 alunos, distribuídos nas áreas de saúde e biológicas, engenharias e tecnológicas, sociais aplicadas, e, humanidades, ciências e educação, dos cursos de graduação da UNESC, correspondendo a 25% deste total de alunos da saúde e o restante das demais áreas.

Foi estimada uma amostra aleatória e probabilística, estratificada por curso e dividida em dois grandes grupos, totalizando 575 indivíduos (sendo 144 alunos da saúde e 431 alunos das outras áreas) que originaram o banco de dados que foi submetido ao processo de KDD, sendo realizado inicialmente o pré-processamento com a seleção de atributos, limpeza de dados, binarização e transformação de variáveis. A etapa do pré-processamento é importante para a otimização do desenvolvimento da etapa central do KDD (o DM) pois facilita a extração de conhecimento do banco[3].

Na seleção de atributos, a quantidade de variáveis que constituíram o banco de dados foi reduzida de 49 para 36, de maneira que as mais importantes para o tabagismo foram selecionadas, buscando diminuir a sua dimensão e a possibilidade de associações incorretas durante a extração do conhecimento. Tal seleção foi realizada com supervisão de um especialista do domínio de aplicação, sendo selecionadas as variáveis com maior associação ao tabagismo já consolidas pela literatura científica. Buscamos selecionar somente as variáveis indicadas pelo especialista visto o objetivo de nosso estudo em pesquisar associações multifatoriais ao tabagismo entre as variáveis selecionada.

A seguir, durante a limpeza de dados, registros com informações inconsistentes foram retificados ou excluídos. Após, na análise da necessidade de binarização, verificou-se que diversas variáveis que possuíam mais de uma resposta possível já haviam sido dicotomizadas no estudo de Cacciatori et al.[7], como por exemplo: idade, estado civil, tabagismo, uso de bebidas alcoólicas, curso, e fase; assim optamos por manter a binarização destas variáveis já dicotomizadas. No último passo relacionado à transformação das variáveis, as possíveis respostas para cada atributo foram convertidas de nominais para numéricas.

Na etapa seguinte do processo, utilizou-se o aplicativo em desenvolvimento vinculado a um projeto apoiado pela Financiadora de Estudos e Projetos (FINEP)[8], que é uma ferramenta em desenvolvimento pelos Grupos de Pesquisa em Inteligência Computacional Aplicada, Tecnologia da Informação e Comunicação na Saúde, e Redes de Comunicação da UNESC. Essa ferramenta foi usada em uma das tarefas preliminares do DM, a clusterização fuzzy, que tem por objetivo particionar os diferentes dados em grupos com características semelhantes, considerando que a lógica fuzzy auxilia neste processo pois possibilita aos elementos pertencerem a grupos distintos simultaneamente, onde os elementos de um cluster podem pertencer a outros grupos ao mesmo tempo, dependendo dos graus de pertinência envolvidos. Assim, a lógica fuzzy foi escolhida pois quando a tarefa de clusterização clássica é aplicada em grandes bases de dados, parte da informação pode estar localizada em dois clusters, nessa situação existe a possibilidade destes dados serem forçados a pertencer a um grupo no qual não possua suas características, o que prejudica a execução da tarefa e por conseqüência pode gerar ambigüidade dos resultados obtidos[9].

Dentre as formas de utilização da clusterização fuzzy, optou-se pela utilização do algoritmo Gustafson-Kessel que permite encontrar com maior precisão os grupos pré-existentes[9]. Este algoritmo foi aplicado buscando dividir a população inicial do estudo em dois grupos homogêneos (tabagistas e não tabagistas), no entanto, usando para isto as similaridades entre os registros.

A clusterização foi feita selecionando como entrada os atributos listados no Quadro 1, tendo como saída o tabagismo. Após o seu emprego, optou-se por classificar os registros resultantes do cluster com predominância dos fumantes, e, em um segundo momento, em classificar também a base de dados não clusterizada contendo todos os registros.

A tarefa de classificação encontra características de um conjunto de dados a fim de diferenciá-los uns dos outros, resultando em regras que podem ser utilizadas em outro grupo de dados[10]. Para o emprego desta tarefa optou-se pela utilização do algoritmo J48, por ser utilizado com sucesso em alguns estudos voltados à área da saúde, como o realizado por Vianna e cols[11] que buscou caracterizar a mortalidade infantil por meio deste mesmo algoritmo. Tal técnica tem como objetivo o desenvolvimento de

Quadro 1 – Atributos considerados na clusterização.

Variáveis de entrada da clusterização	**Significado da variável**
SEXO	Categorização do gênero: 1. masculino, 2. feminino.
FASE_DICOT	Fase dicotomizada: 1. fase inicial, 2. fase final.
FILHOS	Você tem filhos? 1. sim, 2. não.
QUEM_MORA	Com quem você mora? 1. pais, 2. conjugê, 3. familiares, 4. só, 5. amigos, 6. conjugê e familiares, 7. pais e familiares.
FINANC	Como se dá sua manutenção financeira? 1. atividade acadêmica, 2. trabalho, 3. rendas, 4. mesada, 5. atividade acadêmica e trabalho, 6. trabalho e mesada, 7. atividade acadêmica e mesada, 8. rendas e mesada, 9. trabalho e rendas.
SALARIO	Qual o seu rendimento mensal? 1. 1 salário mínimo, 2. 2 salários mínimos, 3. 3 salários mínimos, 4. 4 salários mínimos ou mais, 97. menos de 1 salário mínimo
SAL_FAM*	Qual a faixa salarial da sua família?
FUM_CASA	Em sua casa, existe algum fumante? 1. sim, 2. não.
QUEM_FUMA	Quem é fumante? 1. pai, 2. mãe, 3. irmão, 4. outros, 5. pai e irmão, 6. mãe e outros, 7. pai e mãe, 8. pai, mãe e irmão, 9. mãe e irmão, 10. pai, mãe e outros, 11. pai, mãe, irmão e outros.
DROGAS	Você é usuário de drogas? 1. sim, 2. não.
ALGUMA_DC	Você tem alguma doença? 1. sim, 2. não.
QUAL_DC	Qual doença?
MEDICACOES	Usa medicações? 1. sim, 2. não.
QUAIS_MED*	Quais medicações?
FUM_QNTOS_CIG*	Quantos cigarros fuma por dia? 96. nenhum, 97. 2 a 3 por semana.
QTS_ANOS*	Há quantos anos fuma?
OQ_FUMA*	O que você fuma? 1. cigarros comercializados, 2. charuto, 3. cachimbo, 4. cigarros feitos em casa com fumo, 5. cachimbo e charuto, 6. cigarros comercializados e charuto, 7. cigarros comercializados, charuto, cachimbo e cigarros feitos em casa com fumo.
ALG_MED	Algum médico já lhe aconselhou a parar de fumar? 1. sim, 2. não.
PRETEN_DEIXAR	Pretende deixar de fumar? 1. sim, 2. não.
TENT_PARAR	Já tentou deixar de fumar? 1. sim, 2. não.
FUMA_EM_CASA	Você fuma dentro de sua casa? 1. sim, 2. não, 3. de vez em quando.
ALGUEM_PROB	Alguém na sua casa tem problemas respiratórios? 1. sim, 2. não.
FUMA_LOC_PROIB	Você fuma em locais proibidos? 1. sim, 2. não.
INFLUENCIAS	Como você começou a fumar? 1. influência de amigos, 2. influência dos pais, 3. vontade própria, 4. por modismo, 5. por efeito de propaganda, 6. influência de ídolos, 7. por rebeldia, 9. por vontade própria e influência de ídolos, 10. por influência de amigos, vontade própria e rebeldia.
PQ_FUMA	Atualmente por que você fuma? 1. por prazer, 2. para relaxar, 3. porque não consegue ficar sem fumar, 4. para parecer mais importante, 5. para encaixar-se no grupo de amigos, 6. outros, 7. por prazes e para relaxar, 8. para relaxar e porque não consegue ficar sem fumar.
EX_QNTO_TEMPO*	Fumou por quanto tempo? 1. 4 anos, 2. 2 anos, 3. 1 ano, 4. 10 anos, 5. 7 anos, 6. 15 anos, 7. 8 anos, 8. 20 anos, 9. 3 anos.
PAROU_TEMPO*	Parou há quanto tempo? 1. 6 meses, 2. 1 ano, 3. 2 meses, 4. 7 anos, 5. 5 anos, 6. 6 anos, 7. 8 anos, 8. 9 anos, 9. 11 anos, 10. 2 anos, 11. 13 anos, 12. 19 anos.
COMO_PAROU*	Como conseguiu parar? 1. por vontade própria, 2. discos de nicotina, 3. grupo de apoio, 4. chicletes de nicotina, 5. acupuntura, 6. por doença, 7. outros.
OUTROS_PARAR*	Outros motivos pelos quais parou? 1. gravidez.
ESTCIV_DIC	Estado civil dicotomizado: 1. solteiro, 2. união estável.
FUMO_DICOT	Fumo dicotomizado: 1. sim, 2. não.
INICIO_FUMO*	Com que idade começou a fumar?
CURSO_DICOT	Curso dicotomizado: 1. área da saúde, 2. não área da saúde.
IDADE_DICOT	Idade dicotomizada: 1. < 21 anos, 2. > 22 anos.
ALCOOL_DICOT	Álcool dicotomizado: 1. sim, 2. não.
IDADE_INIC_DICOT	Idade de início dicotomizada: 1. < 21 anos, 2. > 22 anos.

regras e modelos preditivos, sendo utilizado principalmente para estratificação de risco e previsão de desfechos clínicos. Esse algoritmo busca ainda gerar um modelo classificador na forma de uma árvore de decisão, sendo considerado a evolução de outros algoritmos de classificação, com o diferencial relacionado às mudanças na análise do ganho de informação dos atributos considerados, além de implementar a simplificação da árvore de decisão por meio de medidas estatísticas, onde avalia-se a significância de algumas regras geradas pela árvore. Aquelas que não acrescentam conhecimento são podadas, resultando em uma árvore mais objetiva que pode representar uma classificação mais correta[12].

As árvores de decisão permitem, através da disposição hierárquica das variáveis de entrada, a geração de regras que futuramente podem ser utilizadas como modelos preditivos, auxiliando nas decisões clínicas[11].

Nesse contexto, buscando classificar as bases supracitadas pelo algoritmo J48, as mesmas foram exportadas da Shell Weka para o *software* Waikato Environment for Knowledge Analysis (WEKA) versão 3.6.4[13], que foi escolhido para a realização desta etapa: por ser uma ferramenta de *software* livre regida por Licença Pública Geral; pela facilidade de uso; possuir interface amigável; apresentar uma grande quantidade de tarefas, métodos e algoritmos de DM, incluindo a classificação pelo algoritmo J48; e, por ser citada em variados estudos nacionais e internacionais como o realizado por Vianna[11].

As variáveis mineradas pela classificação, utilizando o algoritmo de árvore de decisão J48, foram escolhidas visando-se caracterizar o perfil epidemiológico dos fumantes da população em estudo.

Em um primeiro momento foi minerado o cluster dos tabagistas – o qual totalizou 45 registros, seis registros de fumantes a menos que na base contendo todos os elementos da pesquisa – utilizando-se todos os atributos pré-processados, presentes no Quadro 1 como variáveis independentes. Posteriormente foi minerado o cluster contendo todos os registros, dentre os quais de tabagistas e de não tabagistas, novamente utilizando os mesmos atributos pré-processados como variáveis independentes (Quadro 1).

A classificação foi realizada inicialmente considerando todos os atributos apresentados no Quadro 1. Tal processo foi iterativo, e, em cada repetição foram sendo excluídos os atributos com maior ganho de informação buscando explicitar novas associações.

Após o DM, dentre as regras que tinham sido geradas, foram analisadas somente aquelas que representaram uma associação na amostra acima de 5% e com taxa de registros classificados corretamente maior ou igual a 70%. Esses valores foram escolhidos a partir de outros estudos, no entanto, tais pesquisas não explicam o porquê destes valores de corte, pois ainda não existe um consenso sobre estas medidas no DM visto que a classificação depende da aplicação em que o DM é usado.

Nesta etapa foram realizados aproximadamente 300 experimentos e destes, foram selecionados os melhores para avaliação. Na avaliação das regras utilizou-se também a ferramenta Weka sendo consideradas as instâncias classificadas correta e incorretamente, o coeficiente kappa, o erro absoluto médio, a sensibilidade e a especificidade, a área da curva ROC, e a acurácia.

RESULTADOS

Esta seção apresenta os resultados obtidos com a classificação: da base original, que não foi clusterizada; da base de tabagistas clusterizada pelo algoritmo Gustafson-Kessel; e, ao final, são apresentadas algumas características da população resultante das regras geradas.

A classificação pelo algoritmo J48 das bases clusterizada e não clusterizada resultou em mais de 300 árvores e 524 regras oriundas das árvores de decisão, e dessas, as 19 mais relevantes são apresentadas a seguir.

A avaliação das árvores que originaram as 13 regras encontra-se disponível na Tabela 1.

Características da população

Na mineração do cluster de tabagistas, 66,66% responderam ter outros fumantes em casa, sendo mais comum o fumante ser o pai (16,66%) ou a mãe (16,66%), e demais membros da família e amigos (36,66%).

Dos tabagistas que pretendem abandonar o hábito, 8,88% nunca tentaram parar e têm alguém com problemas respiratórios em casa, enquanto 40,00% dos que não têm alguém com problemas pulmonares em casa não pretendem deixar de fumar.

No nosso estudo, 8,86% da população entrevistada foram formados por fumantes. Dentre todos os indivíduos, 30,26% responderam que têm tabagistas em casa, sendo o fumante o pai (29,88%) ou a mãe (29,88%).

Dos tabagistas desta base, 58,82% declararam ter iniciado o hábito de fumar com idade menor que 21 anos, enquanto 41,17% com idade maior de 21 anos. Quando questionados a respeito do que os motivou a começar a fumar, 27,45% declararam a influência de amigos e 49,01% por vontade própria. Sobre o motivo de fumar,

Tabela 1 – Avaliação das árvores

Descrição	Cluster Fumantes (n=6 árvores)		Base Completa (n=13 árvores)	
	Média(±DP)	IC 95%	Média(±DP)	IC 95%
Instâncias classificadas corretamente	84,07(±12,91)	(78,27-89,88)	98,14(±1,83)	(97,32-98,96)
Instâncias classificadas incorretamente	15,93(±12,91)	(10,12-21,73)	1,86(±1,83)	(1,04-2,68)
Sensibilidade	84,07(±12,92)	(78,26-89,88)	98,15(±0,02)	(0,97-0,99)
Especificidade	73,47(±27,63)	(61,04-85,89)	86,81(±0,17)	(0,79-0,94)
Kappa	0,84(±0,13)	(0,78-0,90)	0,98(±0,02)	(0,97-0,99)
Área da curva ROC	0,77(±0,20)	(0,68-0,86)	0,91(±0,11)	(0,87-0,96)
Acurácia	81,82(±16,85)	(74,24-89,40)	98,12(±0,02)	(0,97-0,99)

39,21% responderam que fumam para relaxar e 29,41% por prazer.

Ainda considerando os que fumam, 21,66% declararam fumar em casa regularmente, 11,76% eventualmente e 54,90% nunca; 23,52% relataram a presença de problemas pulmonares em outros moradores da casa onde vivem; 58,82% declararam que pretendem deixar de fumar; 41,17% já tentaram parar de fumar; 50,98% nunca foram aconselhados a abandonar o hábito por algum médico.

Sobre o uso de drogas, todos os indivíduos arrolados no estudo que referiram usar drogas fumam, correspondendo a 35,29% dos fumantes.

Classificação da base original (não clusterizada)

Na classificação da base de dados completa contendo fumantes e não fumantes foram geradas 339 regras, sendo as 6 mais relevantes apresentadas a seguir.

Os fumantes que iniciaram o hábito por vontade própria, que não têm pessoas com problemas pulmonares em casa e que fumam por prazer não pretendem deixar de fumar (7 ocorrências, 2 erros). Já os que iniciaram por vontade própria, que frequentam os cursos da área da saúde, possuem alguém que vive na mesma casa com problemas respiratórios (5 ocorrências).

Dos que não fumam em casa, também não costumam fumar em locais proibidos (34 ocorrências, 10 erros).

Em relação aos tabagistas que nunca tentaram parar de fumar, também não pretendem deixar de fumar (15 ocorrências, 2 erros); por outro lado, aqueles que já tentaram parar de fumar alguma vez, pretendem deixar de fumar definitivamente (30 ocorrências, 11 erros).

Sobre os problemas pulmonares, dos fumantes que não têm pessoas com problemas pulmonares em casa, a maioria está iniciando o curso universitário, relatam fumar para relaxar, não estão em união estável e não fumam em locais proibidos (9 ocorrências).

Classificação da base clusterizada de fumantes

Na classificação da base de dados contendo o cluster de tabagistas, foram geradas 185 regras, sendo as 7 mais relevantes apresentadas a seguir.

Os indivíduos com menos de 21 anos e que estão começando o curso universitário têm intenção de parar de fumar (15 ocorrências, com 1 erro), já aqueles com mais de 21 anos, estão terminando o curso universitário (24 ocorrências, 5 erros).

Dos tabagistas que possuem alguma doença, e que estão realizando algum tipo de tratamento medicamentoso, relataram que pretendem deixar de fumar (7 ocorrências, 1 erro).

Entre os que não pretendem deixar de fumar, relataram a inexistência de pessoas com problemas pulmonares na casa onde vivem, e nunca tentaram abandonar o cigarro (18 ocorrências, 6 erros). Dos tabagistas que pretendem deixar de fumar, a maioria já tentou parar (21 ocorrências, 2 erros).

Sobre os fumantes que têm como fonte financeira o próprio trabalho, não costumam fumar em locais proibidos (27 ocorrências, 4 erros); já aqueles que têm como fonte de renda a mesada fornecida por terceiros, e que não usam medicações, costumam fumar em locais proibidos (6 ocorrências, 1 erro).

DISCUSSÃO

O KDD vem sendo utilizado amplamente em diversas áreas do conhecimento, entre elas a saúde, com o surgimento de extensas bases de dados, buscando a descoberta de conhecimento não trivial, sutil ou exaustiva aos métodos relacionados à estatística descritiva e analítica. Na Medicina, o DM é usado principalmente no desvelamento de novas relações entre os dados que constituem um banco, podendo gerar novas hipóteses e modelos de predição que possam ser usados para apoiar as decisões clínicas[2].

No Brasil, o DM já foi usado em trabalhos que visaram identificar, entre outros temas, o perfil dos pacientes portadores de beta talassemia heterozigota[14], características da mortalidade infantil[11], e até mesmo o perfil dos beneficiários de planos de saúde suplementar[15]. No tocante ao perfil epidemiológico dos tabagistas o nosso trabalho parece ser o primeiro no Brasil a abordá-lo através das técnicas de DM.

No mundo alguns estudos associando ao DM e tabagismo já foram realizados[16-21] dentre os quais podemos citar um estudo realizado nos Estados Unidos que teve por objetivo descobrir relação entre alguns biomarcadores de possível dano à exposição a fumaça do cigarro, baseado no modelo de Regressão Multivariada com Spline Adaptado[17].

Outro estudo também realizado nos Estados Unidos resultou no perfil dos indivíduos que cessaram o hábito de fumar, realizado pela tarefa de classificação do DM, através do método de redes neurais[18]. Apesar do algoritmo J48 de árvores de decisão ser amplamente utilizado em estudos relacionados à Medicina, não encontramos pesquisas que o tenham utilizado no Tabagismo.

O fato do tabagismo representar a principal causa de morte evitável no mundo pode justificar a realização de estudos epidemiológicos em grupos de tabagistas, como o nosso, que possibilitem o desenvolvimento de novas estratégias de prevenção e abandono do vício.

No nosso estudo 8,86% da população de universitários entrevistada foi formada por fumantes, inferior aos 11,1% dos americanos que terminaram algum curso superior e também inferior aos 17,5% da prevalência da população nacional e aos 19% da região Sul apuradas pelo IBGE no ano de 2008. Tais dados corroboram ao que foi observado por outros autores, que sugerem associação entre a escolaridade e o hábito de fumar[22-23].

Observou-se predominância de idade precoce no que diz respeito ao início do hábito de fumar, uma vez que 58,82% dos tabagistas declararam que iniciaram o hábito de fumar com idade igual ou inferior a 21 anos. Esta característica corrobora aos achados do IBGE de 2008 que evidenciaram que pelo menos 50,0% dos fumantes experimentaram o cigarro pela primeira vez na adolescência, sendo que a idade mais prevalente para o

início do uso diário ficou entre 17 e 19 anos, apesar de o nosso estudo não ter discriminado quais desses indivíduos fumaram pela primeira vez na adolescência[22, 24].

Na mineração da base de dados contendo todos os registros observa-se que para aqueles indivíduos que responderam que têm algum tabagista em casa, o fumante era o pai ou mãe na maioria das vezes, o que pode sugerir uma possível influência dessas figuras na iniciação do hábito de fumar dos entrevistados. No entanto, ao minerarmos o cluster de tabagistas observamos que as figuras paterna e materna não foram as mais citadas para esse questionamento, neste grupo específico outros indivíduos que não familiares ou amigos são aqueles que fumam em casa. Tal característica diverge de outros estudos que sugerem que as figuras paterna e materna, quando tabagistas, representam importantes fatores de risco para a iniciação do hábito de fumar de seus filhos, uma vez que o fato de um dos pais fumar pode significar a aprovação de tal hábito[23].

Onze tabagistas declararam fumar em casa regularmente e doze relataram que algum membro morador da mesma casa na qual residem possuem algum problema respiratório, sugerindo que o fumo passivo possa ter papel atuante no desenvolvimento de patologias pulmonares para essa população. Confirmando essa característica da população, outros estudos sugerem que o fumo possa representar um importante fator de risco para o desenvolvimento de doenças do trato respiratório[1, 25].

Apesar da alta escolaridade da nossa amostra, apenas 58,82% dos fumantes declararam que pretendem deixar de fumar, desses, apenas 41,17% já tentaram efetivamente abandonar o tabagismo. Tal característica sugere que o prazer propiciado pelo hábito supera os riscos atrelados a ele, considerando os que não pretendem deixar de fumar.

Todos os indivíduos que se declararam usuários de drogas em nosso estudo são tabagistas, sugerindo que o hábito de fumar possa representar uma porta de entrada para a experimentação de drogas ilícitas e drogadicção.

Sobre a população de tabagistas, as principais associações encontradas pelo DM, revelaram que indivíduos jovens que cursam fases iniciais têm intenção de parar de fumar; e os que estão em tratamento por algum tipo de doença também pretendem parar de fumar.

CONCLUSÃO

O perfil epidemiológico dos tabagistas resultante das regras geradas em nosso estudo foi semelhante ao encontrado na literatura, apesar de nossa pesquisa resultar em uma grande quantidade de associações, característica essa associada ao DM, que visa estabelecer hipóteses, assim, sugere-se que futuras pesquisas investiguem as associações propostas em nosso estudo.

No entanto não foi possível encontrar relações óbvias entre algumas variáveis como o fumo e o sexo, por exemplo, o que pode ter ocorrido em detrimento da pequena quantidade de registros e em específico, da baixa proporção de tabagistas na amostra. Tal característica pode ter influenciado nos resultados obtidos, visto que o KDD foi desenvolvido para a descoberta de conhecimentos em extensas bases de dados, ou seja, para aquelas nas quais a bioestatística clássica seria muito trabalhosa e pouco sensível.

Esta pesquisa possibilitou a validação interna, mas para a externa talvez seja necessária uma amostra maior buscando uma maior precisão dos algoritmos de DM, visto que tanto a base clusterizada de fumantes quanto a base completa foram consideradas pequenas se comparadas aos bancos de dados dos estudos citados em nossa pesquisa.

Sugere-se ainda aplicar outras tarefas, métodos e algoritmos de DM, visando aprofundar o conhecimento gerado, comparando-os com resultados encontrados em nossa pesquisa.

AGRADECIMENTOS

A professora Merisandra Côrtes de Mattos Garcia e ao acadêmico Gustavo Pasquali Moretti, pela contribuição durante as primeiras etapas do projeto.

FONTE FINANCIADORA

Financiadora de Estudos e Projetos (FINEP), Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC), Conselho Nacional de Pesquisa (CNPq) e Universidade do Extremo Sul Catarinense (UNESC).

REFERÊNCIAS

1- World Health Organization. WHO report on the global tobacco epidemic: warning about dangers of tobacco. Italy: Who; 2011.
2- Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81-97.
3- Leung C. Mining uncertain data. Wiley Interdiscip Rev Data Min Knowl Discov. 2011;1(4):316-29.
4- Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34(2):113-27.
5- Holmes JH, Durbin DR, Winston FK. Discovery of predictive models in an injury surveillance database: an application of data mining in clinical research. Proceedings of the AMIA Annual Symp. 2000 Nov 4-8; Los Angeles, California: Hanley & Belfus; 2000. p.359-63.
6- Li J, Fu AW, Fahey P. Efficient discovery of risk patterns in medical data. Artif Intell Med. 2009;45(1):77-89.
7- Caciatori, JFF. Uso de tabaco entre universitários no Sul do Brasil [Trabalho de Conclusão de Curso]. Criciúma: Universidade do Extremo Sul de Santa Catarina, Curso de Medicina; 2011.
8- Azevedo Simões PW de, Martins PJ, Casagrandre RA, Madeira K, de Mattos MC, Manenti SA, Rosa MI da, Dal-Pizzol F, Venson R, Coral LN, Souza GS de, Pandini JC, Cassettari Junior JM, Moretti GP, Cesconetto S. Using a model of parallel distributed processing associated with data mining in the characterization of sexuality in a university population. Stud Health Technol Inform. 2013;192:1135.
9- Gustafson DE, Kessel WC. Fuzzy clustering with a fuzzy covariance matrix. Proceedings of the IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes. 1978 Jan 10-12; San Diego, California

.1978;17:761-6.
10- Han J, Kamber M. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann; 2001.
11- Vianna RCXF, Moro CMCB, Moysés SJ, Carvalho D, Nievolo JC. Mineração de dados e características da mortalidade infantil. Cad. Saúde Pública. 2010; 26(3):535-42.
12- Quinlan JR. C4.5:Programs formachine learning. San Mateo: Morgan Kaufmann Publishers; 1993.
13- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor. 2009;11(1):10-8.
14- Domingos ALB, Granzotto LA, Belini Junior E, Oliveira TYK, Domingos ACB, Bonini-Domingos CR. Perfil de beta talassemia heterozigota obtido a partir de análise data mining em banco de dados. Rev Bras Hematol Hemoter. 2010;32(1):78-9.
15- Barros EF, Romão W, Constantino AA, Souza CL. Pré-processamento para mineração de dados sobre beneficiários de planos de saúde suplementar. J. Health Inform. 2011;3(1):19-26.
16- Kartelj A. Classiûcation of Smoking Cessation Status Using Various Data Mining Methods. MB-New Series. 2010;24(3-4):199-205.
17- Warner JH, Liang Q, Sarkar M, Mendes PE, Roethig HJ. Adaptive regression modeling of biomarkers of potential harm in a population of U.S. adult cigarette smokers and nonsmokers. BMC Med Res Methodol. 2010;10:19.
18- Poynton MR, McDaniel AM. Classification of smoking cessation status with a backpropagation neural network. J Biomed Inform. 2006;39(6):680-6.
19- Kupelian V, Link CL, McKinlay JB. Association between smoking, passive smoking, and erectile dysfunction: results from the Boston Area Community Health (BACH) Survey. Eur Urol. 2007;52(2):416-22.
20- Johnson CA, Cen S, Gallaher P, Palmer PH, Xiao L, Ritt-Olson A, et al. Why smoking prevention programs sometimes fail. Does effectiveness depend on sociocultural context and individual characteristics? Cancer Epidemiol Biomarkers Prev. 2007;16(6):1043-9.
21- Comandini A, Marzano V, Curradi G, Federici G, Urbani A, Saltini C. Markers of anti-oxidant response in tobacco smoke exposed subjects: a data-mining review. Pulm Pharmacol Ther. 2010;23(6):482-92.
22- Instituto Brasileiro de Geografia e Estatística. Pesquisa nacional por amostra de domicílios 2008. Rio de Janeiro: IBGE; 2009.
23- Centers for Disease Control and Prevention (CDC). Vital signs: current cigarette smoking among adults aged >or=18 years. United States, 2009. MMWR Morb Mortal Wkly Rep. 2010;59(35):1135-40.
24- Gilman SE, Rende R, Boergers J, Abrams DB, Buka SL, Clark MA, et al. Parental smoking and adolescent smoking initiation: an intergenerational perspective on tobacco control. Pediatrics. 2009;123(2):e274-81.
25- Kauffmann F, Dockery DW, Speizer FE, Ferris BG Jr. Respiratory symptoms and lung function in relation to passive smoking: a comparative study of American and French women. Int J Epidemiol. 1989;18(2):334-44.