MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO NA ÁREA DE ONCOLOGIA

Fabrício Scheunemann[1], Fabrício Pretto[2]

Resumo: No mercado competitivo da atualidade, as organizações buscam qualificar seu gerenciamento e tomada de decisão a partir da análise das informações. O simples fato de armazenar e recuperar essa informação já proporciona grande benefício às organizações. Contudo, apenas resgatar a informação não propicia todas as vantagens possíveis. As técnicas de mineração de dados permitem que se explorem grandes conjuntos de dados a fim de estabelecer relações, associações e descobrir padrões úteis que tenham valor para a organização com o propósito de entender o fenômeno gerador dos dados. Este trabalho expõe a aplicação do algoritmo de classificação árvore aumentada do Naïve Bayes (TAN) com a descoberta não supervisionada. Utilizou-se a ferramenta de mineração de dados Weka com o intuito de descobrir conhecimento útil da especialidade médica de oncologia na base de dados de uma casa de saúde.

Palavras-chave: Mineração de dados. Descoberta de conhecimento. Oncologia.

INTRODUÇÃO

O rápido avanço na tecnologia de coleta e armazenamento de dados permitiu que as organizações acumulassem vasta quantidade de dados, principalmente na área da saúde. Muitas vezes, ferramentas e técnicas tradicionais de análise de dados não podem ser usadas devido ao tamanho do conjunto de informações ser muito grande, tornando-se necessário o desenvolvimento de novos métodos para análises de dados.

A mineração de dados é uma tecnologia que combina métodos tradicionais de análise com algoritmos sofisticados para processar grandes volumes de dados, com o objetivo de estabelecer relações, associações e descobrir padrões úteis que poderiam permanecer ignorados.

Na mineração de dados, o processo geral de conversão de dados brutos em informações úteis é chamado de descoberta de conhecimento em banco de dados *Knowledge Discovery in Databases* (KDD). Esse processo consiste de uma série de passos de transformação, do pré-processamento ao pós-processamento dos resultados da mineração de dados (TAN; STEINBACH; KUMAR, 2009).

Segundo Carvalho (2005), a mineração de dados pode ser realizada de três diferentes formas em função do nível de conhecimento que se tenha do problema estudado. Se há pouco conhecimento, faz-se a descoberta não supervisionada; se há suspeita de alguma relação interessante, faz-se a testagem em hipótese; se há muito conhecimento, faz-se a modelagem matemática da relação.

1 Graduado em Sistemas de Informação pelo Centro Universitário UNIVATES, Lajeado/RS. fabricio.scheunemann@gmail.com

2 Mestre em Ciência da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul, Brasil (2008). Professor do Centro Universitário UNIVATES, Lajeado/RS. pretto.f@gmail.com

Qualquer uma das três possíveis metodologias de mineração de dados necessita basicamente das mesmas técnicas para sua realização. As técnicas são de caráter genérico e podem ser implementadas por ferramentas diferentes, como árvores de decisão, algoritmos estatísticos, algoritmos genéticos, regras de decisão, redes neurais artificiais, redes bayesianas e lógica fuzzy (REZENDE et al., 2003).

Ao longo do tempo, percebeu-se que a velocidade de armazenamento das informações no setor da saúde era muito maior do que a velocidade de análise. Isso gera um problema e uma contradição, pois as organizações, por possuírem vasta quantidade de dados, possuem a falsa sensação de que estão bem informadas. Porém, essas informações precisam ser analisadas de forma correta e em tempo hábil.

Diante disso, este trabalho de pesquisa propõe identificar e analisar os dados para mineração de dados na descoberta de conhecimento na área de oncologia em uma casa de saúde. O local onde o estudo foi desenvolvido é uma instituição filantrópica de direito privado, sendo referência em diversas especialidades nas regiões dos vales do Taquari e Rio Pardo.

DESCOBERTA DE CONHECIMENTO

Segundo Fayyad (1996), a descoberta de conhecimento em banco de dados ou *Knowledge Discovery in Databases*, é um processo usado para a identificação de padrões válidos em análise de grandes conjuntos de dados, podendo descobrir informações úteis, conforme ilustrado na Figura 1.

Figura 1 - Etapas do processo KDD

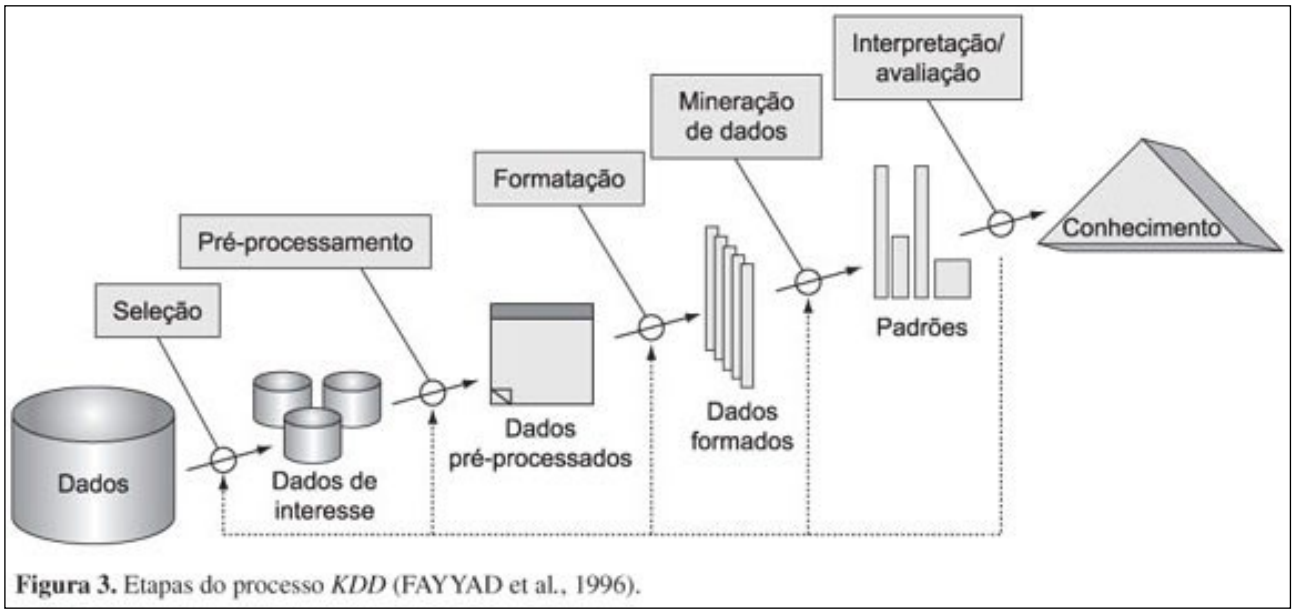


Figura 3. Etapas do processo *KDD* (FAYYAD et al., 1996).

Fonte: Dos autores.

O processo KDD é um conjunto de atividades contínuas compostas por cinco etapas:

- **seleção:** nesta fase do KDD serão decididos quais os conjuntos de dados que serão relevantes pra a tarefa de análise da base de dados;

- **pré-processamento:** nesta fase acontece a limpeza dos dados e ajustes nas informações ausentes, errôneas e inconsistentes nas bases de dados, a fim de ter uma qualidade dos dados;
- **formatação ou transformação:** nesta fase acontece a transformação dos dados, serão analisados os dados e reorganizá-los para que sejam interpretados por um software de mineração de dados;
- **mineração de dados:** nesta fase faz com que os meros dados sejam transformados em informações através de algoritmos;
- **interpretação ou avaliação:** nesta fase é onde as regras indicadas pela etapa de mineração serão interpretadas para a descoberta de conhecimento, após a interpretação poderão surgir padrões, relacionamentos e descoberta de novos fatos.

DATA WAREHOUSE

Segundo Silveira (2014), o *Data Warehouse* (DW) é um repositório de informação que congrega os dados de origem operacional e transacional de uma organização e dados externos, que são filtrados, validados e carregados no repositório, que passam a ser a fonte de informação para aplicações de análise.

Sua construção é um processo normalmente moroso e complexo, por diversos fatores, dentre os quais a grande quantidade de dados, diversas fontes de informações com bases heterogêneas e muitas vezes inconsistentes, sendo necessário o envolvimento de várias áreas da empresa para interpretação dos dados.

Embora o conceito de DW se aplique a grandes quantidades de dados, sua capacidade não é infinita, devendo ser utilizada sabiamente. Apenas dados relevantes devem constar no repositório.

ALGORITMO TAN

O algoritmo utilizado no estudo chama-se *tree augmented naïve Bayes* (TAN), o que, traduzido textualmente, significa árvore aumentada do Naïve Bayes (no entanto a tradução fica apenas o esclarecimento). O classificador TAN foi criado por Friedman e Goldszmidt, com o objetivo de melhorar o Naïve Bayes, sendo uma estrutura parecida, mas que permite a dependência entre os atributos.

Para descobrir a dependência entre atributos no método TAN, é utilizado o algoritmo Chow e Lui, em que cada nodo pode ter no máximo um pai e deve-se encontrar os atributos que tenha maior correlação. O algoritmo realiza o cálculo de relação de acordo com os valores obtidos em X e Y:

$$Ip(X,Y) = \sum_{x,y} P(x,y)\frac{P(x,y)}{P(x)P(y)} \quad (1)$$

Na fórmula acima, o valor Ip(X;Y) é a informação que X exerce sobre Y ou vice-versa, sendo essa informação calculada para todos os pares de atributos. A partir do algoritmo de Chow e Lui, Friedman adaptou-o para que todos os atributos sejam dependentes da classe. O objetivo é obter a árvore de dependências que maximize o peso das informações mútuas entre os atributos.

$$Ip(X,Y \mid C) = \sum_{x,y,c} P(x,y,c)\frac{P(x,y \mid c)}{P(x \mid c)P(y \mid c)} \quad (2)$$

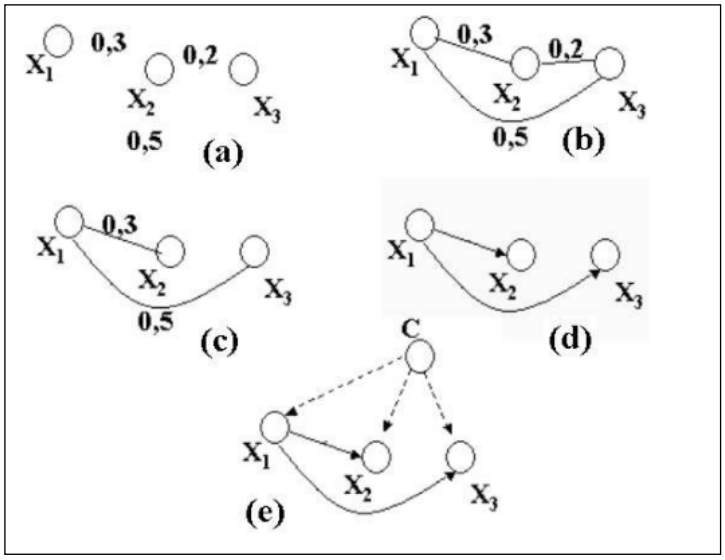
A construção da árvore de dependências é o que diferencia o TAN do Naïve Bayes, pois, teoricamente, é devido à dependência entre os atributos que o TAN melhora o desempenho em relação ao Naïve Bayes.

Para construir o grafo de dependências baseado no método TAN, deve-se utilizar a fórmula Chow e Lui adaptada por Friedman, que serão citados em cinco passos:

- 1º passo: obtém a informação mútua entre cada par de nodos;
- 2º passo: desenha o grafo com os nós e as ligações entre eles, verificando o custo de cada ligação;
- 3º passo: calcula o grafo que maximiza a informação mútua entre os atributos de forma acíclica;
- 4º passo: define o nodo raiz com as informações mútuas mais altas;
- 5º passo: adiciona a classe pai de todos os atributos;

Após esses cinco passos, o grafo utilizando o método TAN é criado, conforme ilustrado na Figura 3, em que existem as relações entre o nodo pai com os atributos, conforme método de Naïve Bayes e entre os atributos de acordo com método TAN.

Figura 2 - Definição de dependências do algoritmo TAN



Fonte: Dos autores.

WEKA

Existem, disponíveis no mercado, ferramentas gratuitas e pagas para mineração de dados. Essas ferramentas são capazes de executar as etapas do processo de mineração.

O *software* de mineração de dados a ser utilizado neste estudo chama-se WEKA (*Waikato Environment for Knowledge Analysis* ou Ambiente para a Análise do Conhecimento), por ser um *software* de distribuição gratuita, desenvolvido em Java, que se consolidou como a ferramenta de mineração de dados mais utilizada no meio acadêmico. Grande parte de seus componentes de *software* é resultante de teses e dissertações de grupos de pesquisa da Universidade de Waikato, na Nova Zelândia.

Por meio de sua interface gráfica, conhecida como Weka Explorer, é possível conduzir processos de mineração de dados de forma simples, realizando a avaliação dos resultados obtidos e a comparação de algoritmos.

TRABALHOS CORRELATOS

Ao abordar trabalhos correlatos cujo foco seja a aplicação de técnicas e algoritmos de mineração de dados na área da saúde, verificou-se que minerações de dados envolvendo técnica de mineração de classificadores como Naïve Bayes e algoritmos de classificação TAN estão sendo bastante empregadas, uma vez que essa técnica possui grande aplicabilidade à área da saúde, devido a sua natureza de apoio a decisão, além de ser facilmente validada por especialistas e usuários envolvidos.

Abicalaffe (2000) apresentou em seu artigo elaborado na Pontifícia Universidade Católica do Paraná (PUC/PR) proposta de desenvolvimento de um software aplicando a rede bayesiana na prevenção da gestação de alto risco.

Em sua monografia defendida na Universidade Católica de Goiás, Filho (2006) apresentou proposta de desenvolvimento de uma ferramenta para o auxílio no diagnóstico de anomalias cromossômicas para a Síndrome de Turner.

Redeker (2010), em sua monografia prodizida no Centro Universitário UNIVATES, desenvolveu estudo de descoberta de conhecimento na área de cardiologia de uma casa de saúde utilizando o algoritmo TAN com o intuito de descobrir relações entre características de pacientes.

Já Sarabando (2010) apresentou em sua monografia elaborada na Universidade do Porto, estudo da aplicação de redes bayesianas ao prognóstico da sobrevivência no câncer de próstata.

Em sua monografia defendida na Universidade Vila Velha, Pachiarotti (2012) apresentou projeto implementando técnicas de mineração de dados como classificação (algoritmo J48 de árvores de decisão, redes neurais e Naïve Bayes), associação (algoritmo Apriori), clusterização (k-means), em um cenário de atendimento médico para a descoberta de padrões e comportamentos que possibilitem melhor tomada de decisões para os gestores de operadoras médicas, de forma a otimizar recursos e prover maior qualidade no atendimento ao público.

SELEÇÃO DOS DADOS

O processo para descoberta de conhecimento foi iniciado com a etapa de seleção dos dados. Após o reconhecimento dos atributos dos dados da base do sistema SISRHC, foi gerado modelo de dados no formato de planilha, exportando os registros relevantes para o experimento de pacientes de primeira consulta oncológica entre o período de 2011 a 2014, totalizando 3.267 registros e 10 atributos, conforme apresentado na Tabela 1.

Tabela 1 - Atributos de identificação do arquivo do SISRHC na etapa de seleção dos dados

ATRIBUTO	DESCRIÇÃO
SEXO	Indica o sexo do paciente
RACA	Indica a raça do paciente
IDADENEO	Indica a idade do paciente quando constatada a neoplasia
DTPRICON	Indica o ano da primeira consulta
CIRURGIA	Indica se o paciente fez cirurgia
EMTRAT	Indica há quanto tempo está em tratamento o paciente
HISTFAM	Indica se há histórico familiar de câncer

ATRIBUTO	DESCRIÇÃO
ALCOOL	Indica se há histórico de consumo de bebida alcoólica
TABAG	Indica se há histórico de consumo de tabaco
CID	Indica o tumor conforme Cadastro Internacional de Doenças para Oncologia (CID-O)

Fonte: Dos autores.

A fim de estabelecer o enriquecimento do experimento, utilizou-se outra fonte de dados, o sistema de gestão TASY. Para tanto, foi realizada consulta SQL na base de dados do TASY, separando as informações relacionadas com os pacientes do SISRHC complementando os dados com mais seis atributos, como pode ser visualizado na Tabela 2.

Tabela 2 - Atributos de identificação do arquivo do TASY na etapa de seleção dos dados

ATRIBUTO	DESCRIÇÃO
PESO	Indica o peso do paciente
ALTURA	Indica a altura do paciente
OBITO	Indica se o paciente foi a óbito
PESQUISA	Indica se o paciente participa do protocolo de pesquisa nacional e/ou internacional
MUNIBGE	Indica o código do município no IBGE
COORDEN	Indica a microrregião do município no IBGE

Fonte: Dos autores.

A ideia desse enriquecimento foi complementar os dados, pois o SISRHC não possuía essas informações, ou, se as possuía, não eram completas ou confiáveis.

As informações das duas fontes de dados foram relacionadas conforme o número do prontuário do paciente, campo que ambas as fontes de dados possuíam. Ressalta-se que o atributo prontuário, não será considerado na análise para preservar a privacidade dos pacientes com neoplasias e por não ter utilidade para análise.

PRÉ-PROCESSAMENTO

Na etapa de pré-processamento foram verificados os atributos que necessitavam de correção e ajustes de formatação nos dados, a fim de eliminar problemas e tornando mais adequados para uso dos algoritmos na mineração de dados.

Alguns erros ocorrem por falha humana, como o atributo "SEXO", que primeiramente havia sido selecionado da fonte de dados do SISRHC, porém, como foram encontrados 64 registros de dados incorretos, optou-se por cruzar os valores das duas fontes de dados para obter a real informação desse atributo. No final, foi constatado que a fonte de dados do TASY estava completamente correta, por isso foi utilizado o atributo "SEXO" dessa fonte de dados.

Alguns erros são sistemáticos e mais fáceis de detectar e corrigir, como o atributo "ALTURA", que não se tinha um padrão no TASY, sendo cadastrado em alguns momentos em centímetros e em outros em metro. Assim, foram convertidos todos os valores de centímetros para metro.

Nos atributos "PESO" e "ALTURA", foi constatado que havia 26 registros com ausência de valores. Assim, foi conversado com a equipe da oncologia e constatado que esses registros poderiam ser resgatados nos prontuários manuais, pois essa informação era necessária para o tratamento

oncológico. Então, foram separadas as pastas dos pacientes e preenchidos para não eliminarmos os pacientes por falta de registro.

O atributo "OBITO" foi necessário extrair da fonte de dados TASY, devido à falha no processo de lançamento da informação no SISRHC. Como no TASY a informação estava em formato de data, foi convertido caso estivesse preenchido no valor (S), caso contrário no valor (N).

O atributo "RACA" foi preciso extrair da fonte de dados TASY. Primeiramente, o atributo havia sido selecionado da fonte de dados do SISRHC, porém, como foram encontrados 122 registros de dados incorretos, optou-se por cruzar os valores das duas fontes de dados para termos a real informação desse atributo. Como ocorreu com o atributo "SEXO", foi constatado que a fonte de dados do TASY estava completamente correta. Por isso, foi utilizado o atributo "RACA" dessa fonte de dados.

No final da etapa de pré-processamento, do montante de 3.267 pacientes, foram desconsiderados 59 pacientes por apresentarem inconsistência nos dados, totalizando 3.208 pacientes com registros íntegros. Na Tabela 3 estão os atributos atualizados na etapa do pré-processamento da fonte de dados do SISRHC.

Tabela 3 - Atributos de identificação do arquivo do SISRHC na etapa de pré-processamento

ATRIBUTO	DESCRIÇÃO
IDADENEO	Indica a idade do paciente quando constatada a neoplasia
DTPRICON	Indica o ano da primeira consulta
CIRURGIA	Indica se o paciente fez cirurgia
EMTRAT	Indica há quanto tempo está em tratamento o paciente
HISTFAM	Indica se há histórico familiar de câncer
ALCOOL	Indica se há histórico de consumo de bebida alcoólica
TABAG	Indica se há histórico de consumo de tabaco
CID	Indica o tumor conforme Cadastro Internacional de Doenças para Oncologia (CID-O)

Fonte: Dos autores.

Na Tabela 4 estão os atributos que foram atualizados na etapa do pré-processamento da fonte de dados do TASY.

Tabela 4 - Atributos de identificação do arquivo do TASY na etapa de pré-processamento

ATRIBUTO	DESCRIÇÃO
SEXO	Indica o sexo do paciente
RACA	Indica a raça do paciente
PESO	Indica o peso do paciente
ALTURA	Indica a altura do paciente
OBITO	Indica se o paciente foi a óbito
PESQUISA	Indica se o paciente participa do protocolo de pesquisa nacional e/ou internacional
MUNIBGE	Indica o código do município no IBGE
COORDEN	Indica a microrregião do município no IBGE

Fonte: Dos autores.

FORMATAÇÃO

Após serem selecionados, limpos e pré-processados, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados.

Para os atributos "IDADENEO", "PESO" e "ALTURA", optou-se por normalizar os dados, que consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos.

O atributo "IDADENEO" foi transformado em dois novos atributos, o atributo "FETARIA" que agrupa a idade de quatro em quatro anos utilizando como referência a faixa etária do IBGE, e o atributo "EETARIA", que agrupa a idade em três estruturas etárias (IBGE, 2000).

Os atributos "PESO" e "ALTURA" foram utilizados para compor um único atributo, o índice de massa corporal (IMC), aplicando a fórmula:

$$IMC = \frac{\text{PESO (quilograma)}}{\text{ALTURA}^2 \text{ (metros)}} \quad (3)$$

Dessa forma, com o valor do IMC foi possível relacionar os resultados aos pacientes, conforme apresenta a Tabela 5.

Tabela 5 - Resultados do índice de massa corporal

IMC	Resultado
Abaixo de 18,49	Subnutrido
Entre 18,5 e 24,99	Peso saudável
Entre 25 e 29,99	Sobrepeso
Entre 30 e 34,99	Obesidade I
Entre 35 e 39,99	Obesidade II
Acima de 40	Obesidade III

Fonte: Dos autores.

As informações modificadas foram agrupadas em um repositório único no formato de planilha, com 16 atributos de ambas as fontes de dados, conforme Tabela 6.

Tabela 6 - Atributos de identificação do arquivo de ambas as fontes

ATRIBUTO	DESCRIÇÃO
SEXO	Indica o sexo do paciente
FETARIA	Indica a faixa etária do paciente conforme Faixa etária IBGE
EETARIA	Indica a estrutura etária
IMC	Indica o resultado do IMC
RACA	Indica a raça do paciente
OBITO	Indica se o paciente foi a óbito
DTPRICON	Indica o ano da primeira consulta
PESQUISA	Indica se o paciente participa do protocolo de pesquisa nacional e/ou internacional
CIRURGIA	Indica se o paciente fez cirurgia
EMTRAT	Indica quanto tempo está em tratamento o paciente
HISTFAM	Indica se há histórico familiar de câncer
ALCOOL	Indica se há histórico de consumo de bebida alcoólica

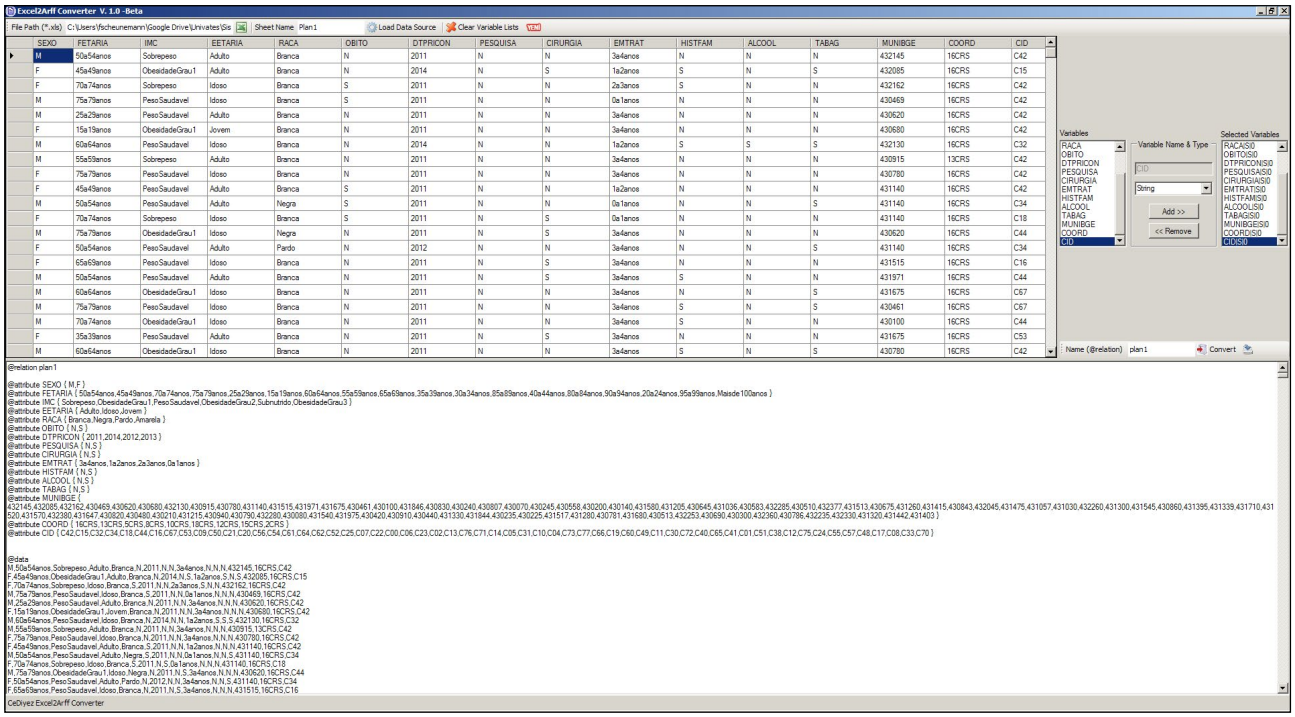
ATRIBUTO	DESCRIÇÃO
TABAG	Indica se há histórico de consumo de tabaco
CID	Indica o Cadastro Internacional de Doenças para oncologia (CID-O)
MUNIBGE	Indica o código do município no IBGE
COORD	Indica a microrregião do município no IBGE

Fonte: Dos autores.

Nessa etapa, foi gerado arquivo com a extensão *Attribute Relation File Format* (ARFF) um dos padrões aceitáveis pelo Weka, com o auxílio do sistema *open source* Excel2ArffConverter obtido na Sourceforge.

Esse *software* converte arquivos de planilha no formato XLS para ARFF, em que os dados são armazenados em duas seções distintas. A primeira seção contém informações de cabeçalho, enquanto a segunda mantém informação dos dados, como ilustrado na Figura 3.

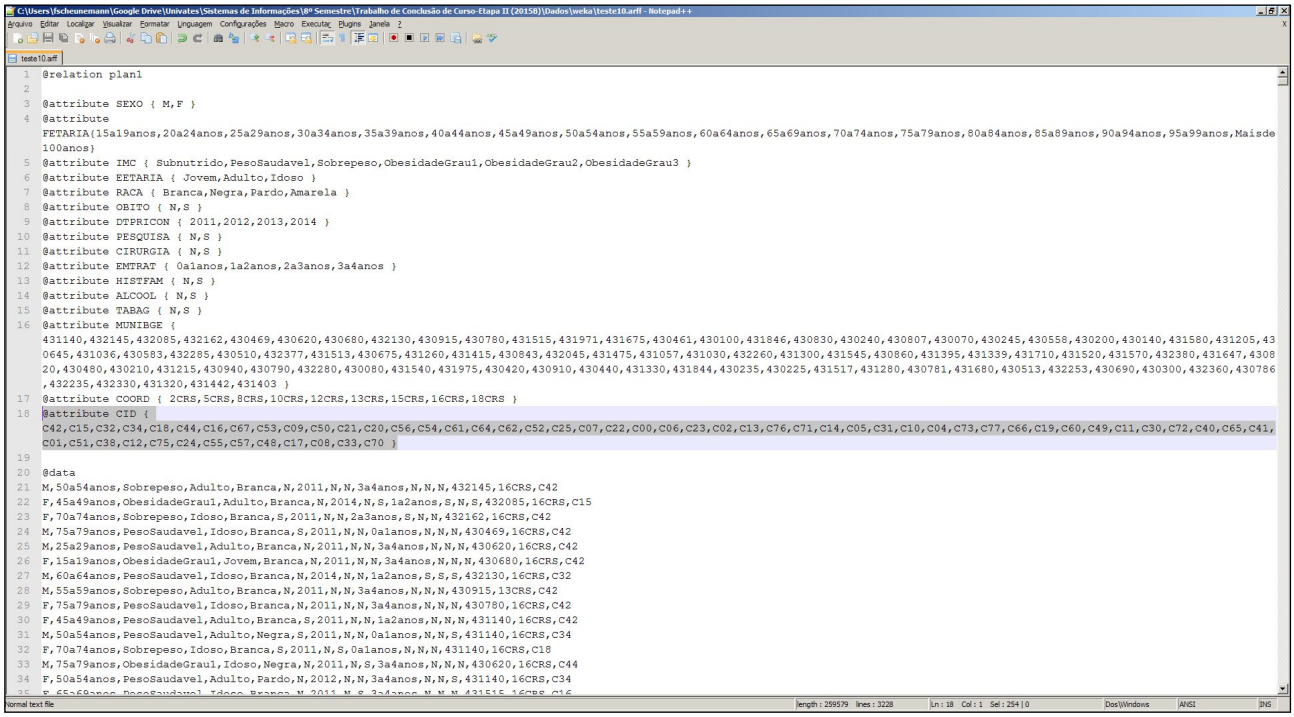
Figura 3- Tela principal do sistema Excel2ArffConverter



Fonte: Dos autores.

Para que o nodo seja escolhido como nodo classe, em suma, representa o atributo mais significativo entre todos os outros atributos. No arquivo com extensão ARFF, o atributo classe deve estar na última linha dos atributos, conforme ilustrado na Figura 4.

Figura 4 - Exemplo CID Atributo Classe



```
@relation plan1

@attribute SEXO { M,F }
@attribute
FETARIA{15a19anos,20a24anos,25a29anos,30a34anos,35a39anos,40a44anos,45a49anos,50a54anos,55a59anos,60a64anos,65a69anos,70a74anos,75a79anos,80a84anos,85a89anos,90a94anos,95a99anos,Maisde
100anos}
@attribute IMC { Subnutrido,PesoSaudavel,Sobrepeso,ObesidadeGrau1,ObesidadeGrau2,ObesidadeGrau3 }
@attribute EETARIA { Jovem,Adulto,Idoso }
@attribute RACA { Branca,Negra,Pardo,Amarela }
@attribute OBITO { N,S }
@attribute DTPRICON { 2011,2012,2013,2014 }
@attribute PESQUISA { N,S }
@attribute CIRURGIA { N,S }
@attribute EMTRAT { 0a1anos,1a2anos,2a3anos,3a4anos }
@attribute HISTFAM { N,S }
@attribute ALCOOL { N,S }
@attribute TABAG { N,S }
@attribute MUNIBGE {
431140,432145,432085,432162,430469,430620,430680,432130,430915,430780,431515,431971,431675,430461,430100,431846,430830,430240,430807,430070,430245,430558,430200,430140,431580,431205,43
0645,431036,430583,432285,430510,432377,431513,430675,431260,431415,430843,432045,431475,431057,431030,432260,431300,431545,430860,431395,431339,431710,431520,431570,432380,431647,4308
20,430480,430210,431215,430940,430790,432280,430080,431540,431975,430420,430910,430440,431330,431844,430235,430225,431517,431280,430781,431680,430513,432253,430690,430300,432360,430786
,432235,432330,431320,431442,431403 }
@attribute COORD { 2CRS,5CRS,8CRS,10CRS,12CRS,13CRS,15CRS,16CRS,18CRS }
@attribute CID {
C42,C15,C32,C34,C18,C44,C16,C67,C53,C09,C50,C21,C20,C56,C54,C61,C64,C62,C52,C25,C07,C22,C00,C06,C23,C02,C13,C76,C71,C14,C05,C31,C10,C04,C73,C77,C66,C19,C60,C49,C11,C30,C72,C40,C65,C41,
C01,C51,C38,C12,C75,C24,C55,C57,C48,C17,C08,C33,C70 }

@data
M,50a54anos,Sobrepeso,Adulto,Branca,N,2011,N,N,3a4anos,N,N,N,432145,16CRS,C42
F,45a49anos,ObesidadeGrau1,Adulto,Branca,N,2014,N,S,1a2anos,S,N,S,432085,16CRS,C15
F,70a74anos,Sobrepeso,Idoso,Branca,S,2011,N,N,2a3anos,S,N,N,432162,16CRS,C42
M,75a79anos,PesoSaudavel,Idoso,Branca,S,2011,N,N,0alanos,N,N,N,430469,16CRS,C42
M,25a29anos,PesoSaudavel,Adulto,Branca,N,2011,N,N,3a4anos,N,N,N,430620,16CRS,C42
F,15a19anos,ObesidadeGrau1,Jovem,Branca,N,2011,N,N,3a4anos,N,N,N,430680,16CRS,C42
M,60a64anos,PesoSaudavel,Idoso,Branca,N,2014,N,N,1a2anos,S,S,S,432130,16CRS,C32
M,55a59anos,Sobrepeso,Adulto,Branca,N,2011,N,N,3a4anos,N,N,N,430915,13CRS,C42
F,75a79anos,PesoSaudavel,Idoso,Branca,N,2011,N,N,3a4anos,N,N,N,430780,16CRS,C42
F,45a49anos,PesoSaudavel,Adulto,Branca,S,2011,N,N,1a2anos,N,N,N,431140,16CRS,C42
M,50a54anos,PesoSaudavel,Adulto,Negra,S,2011,N,N,0alanos,N,N,S,431140,16CRS,C34
F,70a74anos,Sobrepeso,Idoso,Branca,S,2011,N,S,0alanos,N,N,N,431140,16CRS,C18
M,75a79anos,ObesidadeGrau1,Idoso,Negra,N,2011,N,S,3a4anos,N,N,N,430620,16CRS,C44
F,50a54anos,PesoSaudavel,Adulto,Pardo,N,2012,N,N,3a4anos,N,N,S,431140,16CRS,C34
```

Fonte: Dos autores.

No cabeçalho desse arquivo são listados seus atributos de identificação, sempre iniciados por "@attribute", sendo, após, informado o nome do atributo correspondente e, por fim, o conjunto de dados identificadores associados a ele.

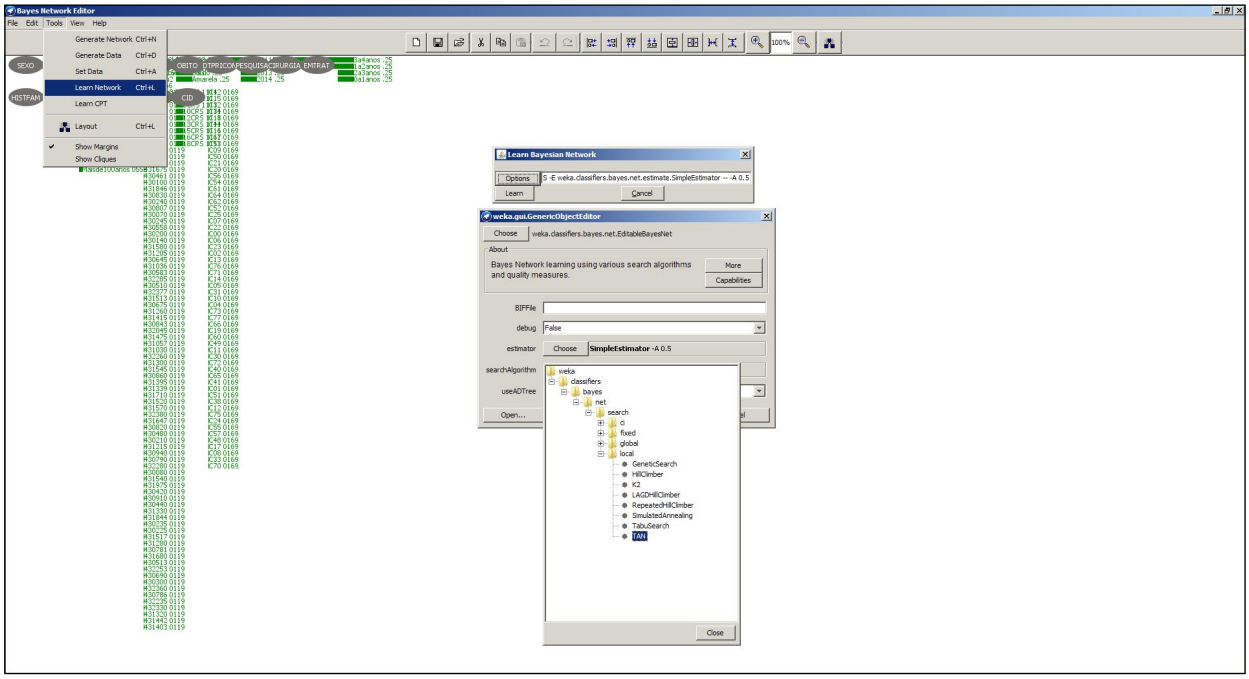
Abaixo do cabeçalho de atributos, iniciada pela marcação "@data", estão listados os 3.208 registros. Durante o processo de importação, o WEKA realiza pré-processamento das informações garantindo sua integridade. Dessa forma, o sistema verifica se para todas as informações existe um atributo cadastrado. Caso não exista, o sistema retorna um erro e não realiza a importação dos dados.

MINERAÇÃO DE DADOS

Após importados os dados para a ferramenta Weka, é possível iniciar a etapa de mineração de dados. O Weka possui recurso chamado Bayes Net Editor, no qual podem ser geradas minerações de dados com base na classificação de Bayes.

Para obter a estrutura da rede deve ser definido o algoritmo de mineração de dados. Neste estudo foi parametrizado o algoritmo TAN, que permite obter rede com uma estrutura de melhor visualização das relações entre atributos, como ilustrado na Figura 5.

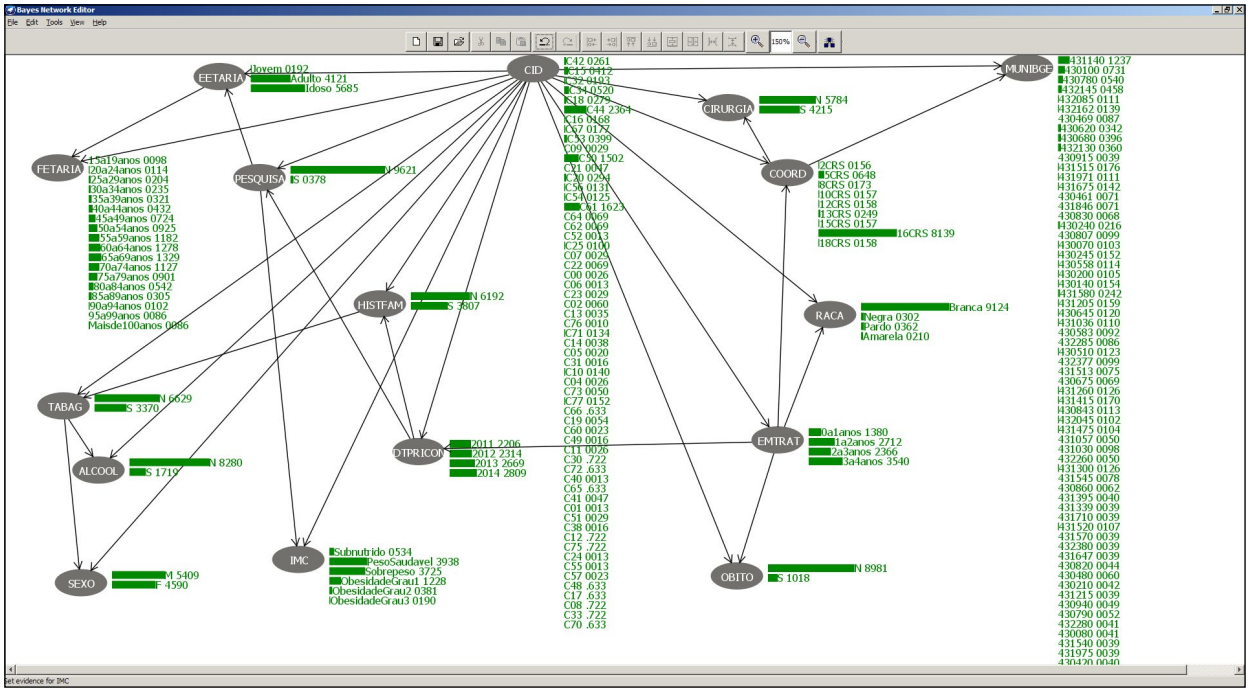
Figura 5 - Algoritmo Naïve Bayes aumentado em árvore (TAN)



Fonte: Dos autores.

Após a escolha do algoritmo TAN, foi selecionada a opção *Learn*, na qual foi exibido grafo com as relações encontradas e as probabilidades dos valores de cada atributo, ilustrado conforme observa-se na Figura 6.

Figura 6 - Grafo de distribuição por CID



Fonte: Dos autores.

O *software* Weka permite que sejam geradas simulações, incluindo ou excluindo relações (arestas) entre os atributos do arquivo.

Além disso, é possível evidenciar valores. Essa ação implica em selecionar um atributo (nodo) e determinar que apenas valores iguais ou diferentes a um certo critério devem ser exibidos. Na mineração foi utilizada a opção *Set evidence*, que filtra um único valor no nodo escolhido. Em consequência, os valores dos nodos relacionados são atualizados automaticamente, calculando os pesos (percentuais) de todos os nodos.

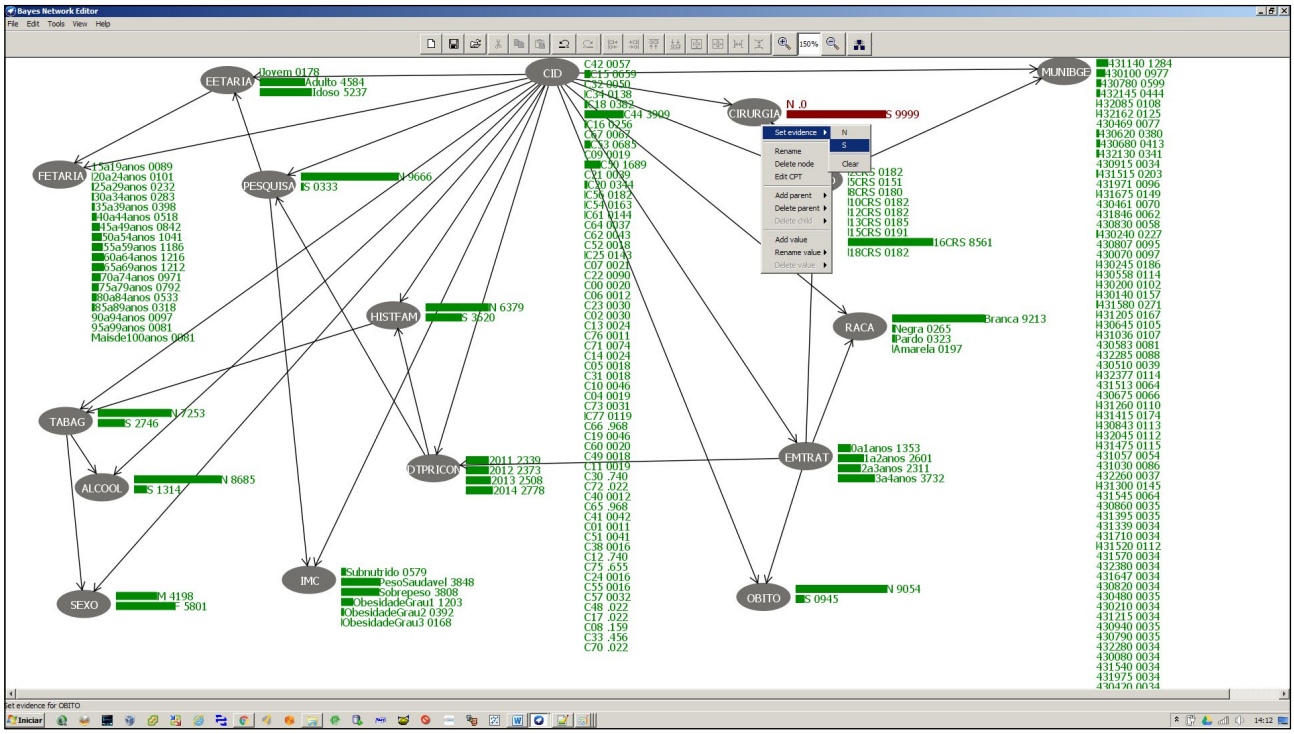
O atributo classe CID foi executado em todos os experimentos e minerado com o algoritmo TAN, sendo os resultados da mineração apresentados em grafos pela ferramenta Weka.

RESULTADOS

A quinta e última etapa foi a interpretação e a avaliação dos resultados. Essa etapa consiste em validar o conhecimento extraído na etapa de mineração de dados.

O primeiro experimento conclusivo foi iniciado evidenciando o nodo "CIRURGIA" com o valor (S). Ao analisar os resultados apresentados, baseado nas afirmações abaixo que os que os cânceres de esôfago (C15), pele (C44), mama (C50) e colo de útero (C53) são os tipos de câncer que os pacientes mais buscam por tratamento cirúrgico, como ilustrado na Figura 7.

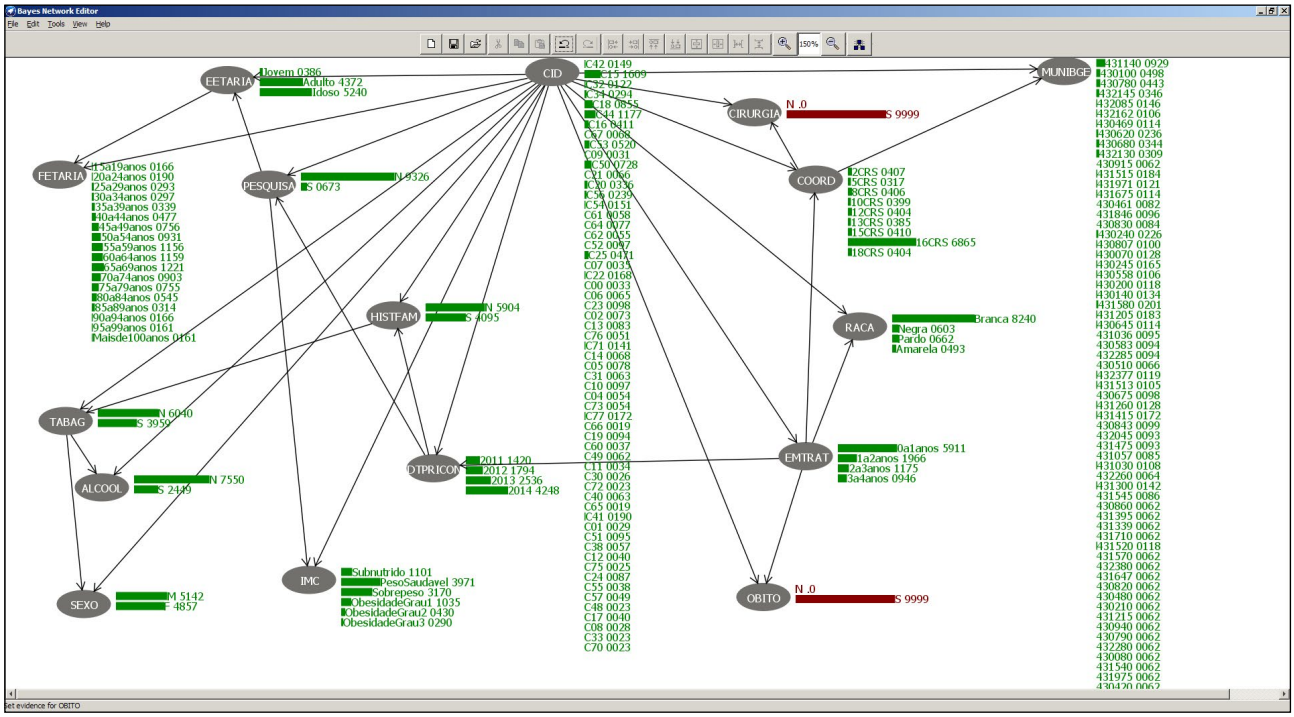
Figura 7 - Grafo evidenciando o nodo "CIRURGIA" (S)



Fonte: Dos autores.

Dentre esses tipos a maior probabilidade de ir a óbito são pacientes que realizam cirurgias com câncer de esôfago (C15) uma das hipóteses citadas pela equipe de especialistas, por ser uma cirurgia geralmente de grande porte e potencialmente contaminada, conforme ilustrado na Figura 8.

Figura 8 - Grafo evidenciando o nodo "CIRURGIA" (S) e nodo "OBITO" (S)



Fonte: Dos autores.

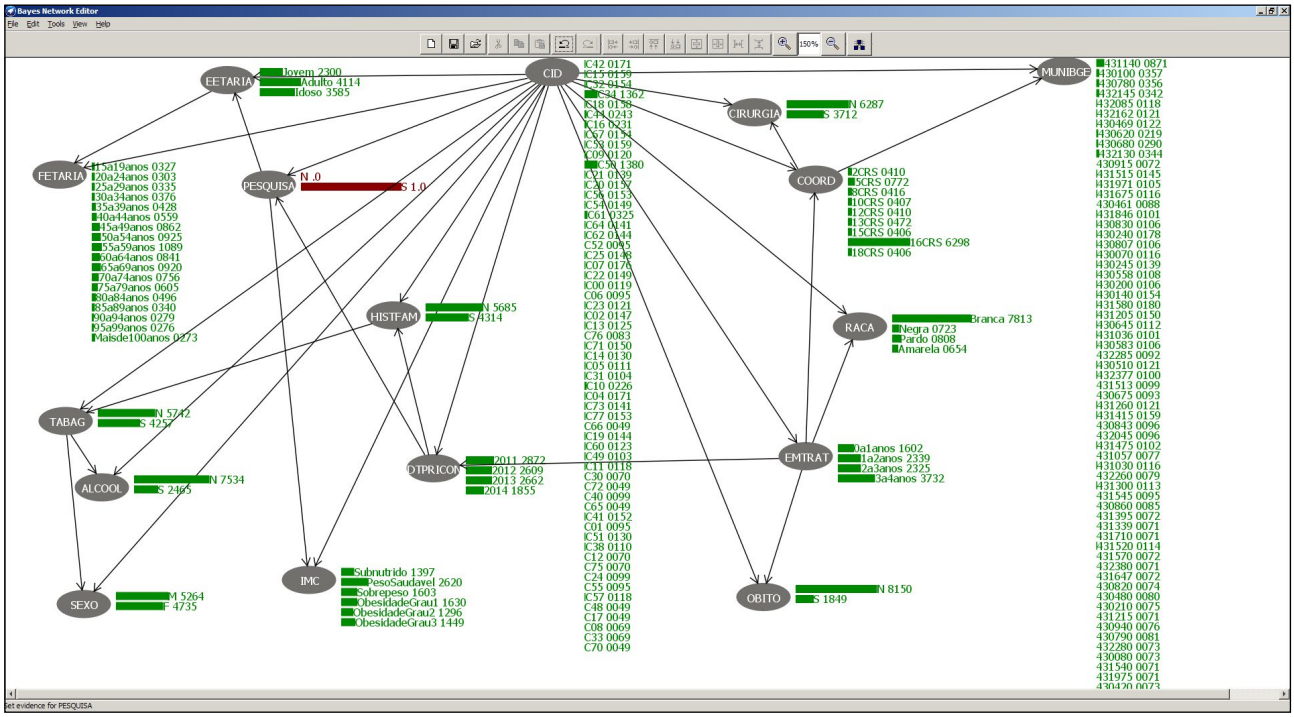
Foram selecionadas também informações importantes para esse experimento extraídas na mineração de dados:

- pacientes oncológicos de primeira consulta apresentam 42,15% de probabilidade de realizar procedimento cirúrgico;
- maior ocorrência em pacientes do sexo feminino, com 58,01% dos casos analisados;
- maior ocorrência em pacientes com peso saudável e sobrepeso;
- maior ocorrência em pacientes que buscam tratamento cirúrgico da estrutura etária idoso, seguidos apenas pelos pacientes com estrutura etária adulto.

Do ponto de vista dos gestores, as informações apresentadas sobre o primeiro experimento são relevantes, pois trazem informação consolidada extraída de dois sistemas (TASY e SISRHC) de pacientes oncológicos que optaram por tratamento cirúrgico do centro de oncologia. Essas informações podem ser trabalhadas em conjunto com o centro cirúrgico para tomada de decisão visando ao bem-estar do paciente.

O segundo experimento conclusivo foi iniciado evidenciando o nodo "PESQUISA" com o valor (S). Ao analisar os resultados apresentados, nota-se que o protocolo de pesquisa está focado em dois tipos de cânceres, o câncer de mama (C50) e o câncer de brônquios e pulmões (C34). Analisando o nodo "IMC", há probabilidade maior em pacientes com peso saudável representando 26,20% dos casos de pacientes do protocolo de pesquisa, conforme ilustrados na Figura 9.

Figura 9 - Grafo evidenciando o nodo "PESQUISA" (S)

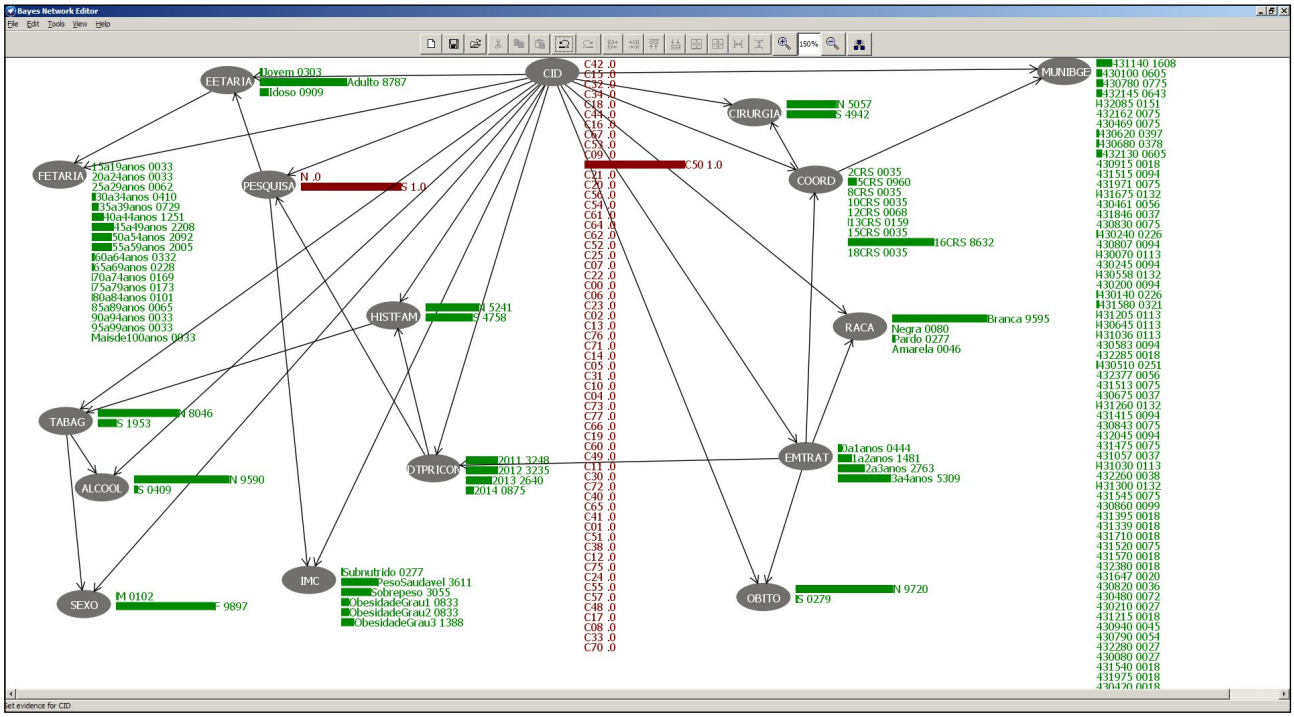


Fonte: Dos autores.

Após, foi evidenciado o nodo "CID" com o valor C50 e analisados os resultados do nodo "OBITO". Constatou-se a probabilidade de 97,20% de não irem a óbito os pacientes de pesquisa com câncer de mama (C50). Uma das hipóteses é que o tratamento de pesquisa pode proporcionar ocasião única de mudar o curso da doença, quando já não respondem aos tratamentos tradicionais.

Pacientes com estrutura etária adulto representam 87,87% dos tratamentos de pesquisa nos casos de neoplasia de mama (C50) e pacientes que participaram do protocolo de mama (C50) de pesquisa estão entre três faixas etárias, de 45 a 49 anos, de 50 a 54 anos e de 55 a 59 anos, conforme ilustrado na Figura 10.

Figura 10 - Grafo evidenciando o nodo "PESQUISA" (S) e nodo "CID" (C50)



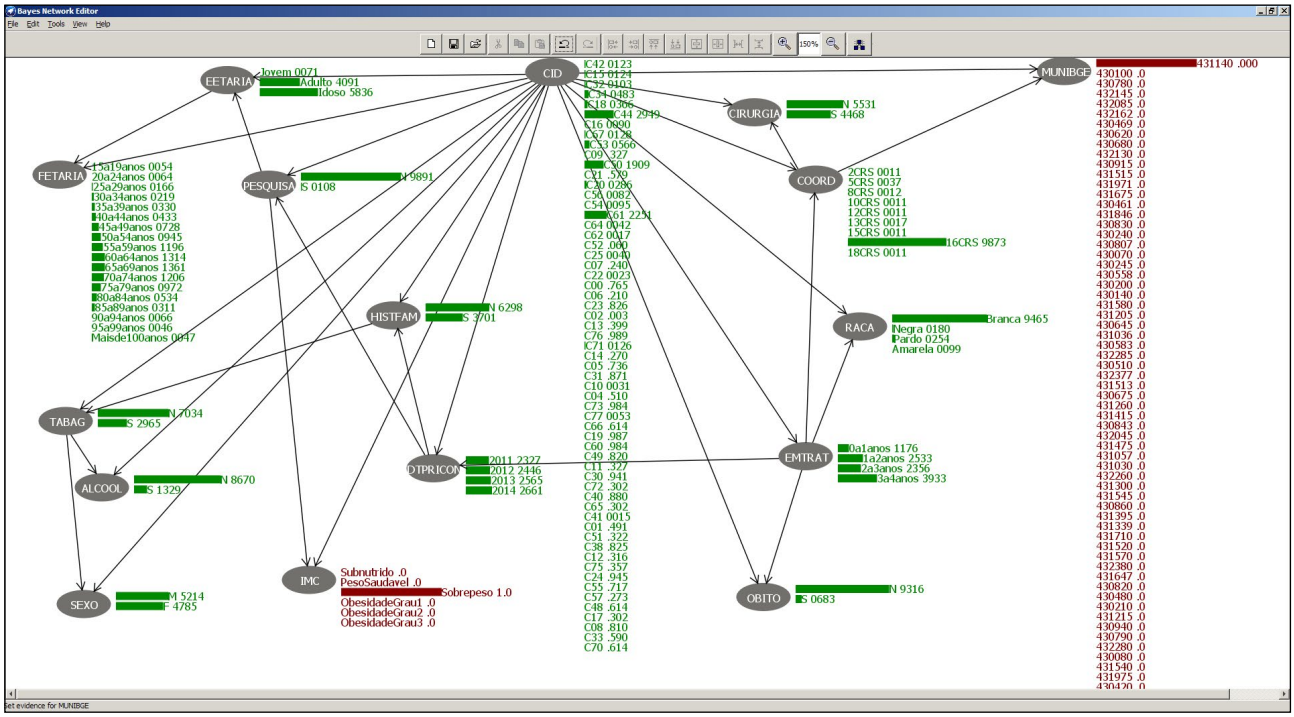
Fonte: Dos autores.

Do ponto de vista dos gestores, as informações apresentadas sobre o segundo experimento confirmam suas deduções e trazem riqueza de informações para a pesquisa clínica em oncologia.

O terceiro experimento conclusivo foi iniciado evidenciando o nodo "IMC" com o valor sobrepeso. Ao analisar os resultados apresentados, conclui-se no nodo "MUNIBGE" (431140) possui uma predominância maior que correspondente ao município IBGE de Lajeado. Após, foi evidenciado o nodo "MUNIBGE" com o valor 431140. Observa-se que os cânceres de pele (C44), de próstata (C61) e de mama (C50) se destacam, umas das hipóteses é de que as pessoas têm optado por alimentos práticos, como comidas semiprontas, que, por sua vez, são alimentação pobre em nutrientes vitais ao nosso organismo.

Analisando o nodo "EETARIA", os pacientes com estrutura etária idoso representam 58,36%. Analisando o nodo "FETARIA", destacam-se duas faixas etárias, dos 60 a 64 anos e dos 65 a 69 anos, como ilustrado na Figura 11.

Figura 11 - Grafo evidenciando o nodo "IMC" (sobrepeso) e nodo "MUNIBGE" (431140)



Fonte: Dos autores.

Foram selecionadas também informações importantes para este experimento extraídas na mineração de dados:

- o câncer de mama (C50) destaca-se em pacientes com estrutura etária adulto e os cânceres de pele (C44) e de próstata (C61) em pacientes com estrutura etária idosos no município IBGE (431140) correspondente a Lajeado.

Do ponto de vista dos gestores, as informações apresentadas sobre o terceiro experimento confirmam o que está sendo comentado nos últimos tempos sobre desvantagens dos alimentos práticos. Uma sugestão vinda do autor do estudo, para realização de um projeto interno em parceria com a Nutrição e o Centro de Oncologia, é a formulação de cartilha aos colaboradores da instituição sobre os riscos da má alimentação, podendo expandir para a comunidade em geral.

CONSIDERAÇÕES FINAIS

Com o desenvolvimento deste estudo, foi possível destacar a importância do uso de técnicas de mineração de dados para a descoberta de conhecimento, especialmente quando aplicadas no caso proposto.

Ao analisar os dados coletados, percebe-se que o algoritmo TAN contribuiu para a descoberta de conhecimento, pois, além de relacionar o nodo "classe" com os demais nodos do grafo, a relação entre os nodos ajudou a verificar as características dos pacientes oncológicos.

Verificou-se que os resultados alcançados no estudo atendem aos objetivos do trabalho e que o método de descoberta não supervisionada e a ferramenta WEKA mostraram-se eficientes e precisos na apuração das informações.

REFERÊNCIAS

ABICALAFFE, C. L.L.; AMARAL, V. F.; DIAS, J. S. **Aplicação Da Rede Bayesiana na Prevenção da Gestação de Alto Risco.** Paraná: Dissertação de Pós-Graduação da Pontifícia Universidade Católica do Paraná, 2000.

BERKA, Pert, RAUCH, Jan, ZIGHED Djamel A., **DM and Medical Knowledge Management: Cases and Applications.** New York: Hershey, 2009.

CARDOSO O. N. P.; MACHADO R. T. M. **Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras.** Obtida via internet. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122008000300004>. Acesso em: 19 mai. 2015.

CARVALHO, L. A. V. - **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração** - Ciência Moderna - RJ, 2005.

FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory e SMYTH, Padhraic. **From Data Mining to Knowlegde discovery**. American Association for Artificial Intelligence. 1996.

FILHO, H. P.F. **Aplicabilidade de Memória Lógica como Ferramenta Coadjuvante no Diagnóstico das Doenças Genéticas.** Goiás: Dissertação de Mestrado da Universidade Católica de Goiás, 2006.

PACHIAROTTI, J.F.B. **Aplicação de Técnicas de Mineração de Dados no aprimoramento do Atendimento Médico em um Cenário de Plano de Saúde.** Espírito Santo. Dissertação de graduação da Universidade Vila Velha, 2012.

PHILIPS CLINICAL INFORMATICS. **Tasy Prestador**. Obtida via internet. Disponível em: <http://www.cilatam.philips.com.br/solucoes/13/tasy-prestador/>. Acesso em: 02 out. 2015.

REDEKER, G. A.. **Descoberta de Conhecimento na área de Cardiologia.** Lajeado. Dissertação de graduação do Centro Universitário UNIVATES, 2010.

REZENDE, S.O. ; PUGLIESI, J. B. ; MELANDA, E. A.; PAULA, M. F., **Mineração de Dados. In Solange Oliveira Rezende. (Org.). Sistemas Inteligentes – Fundamentos e Aplicações.** Barueri, SP. Editora Ltda, 2003.

ROSSO, Leandro – **Oncomining - protótipo utilizando técnicas Data Mining.** Obtida via internet. Disponível em: <http://www.unochapeco.edu.br/saa/tese/2502/TCC%202%20-%20LEANDRO%20ROSSO.pdf>. Acesso em: 19 mai. 2015.

SARABANDO, A. C. L.; **Um estudo do comportamento de Redes Bayesianas no prognóstico da sobrevivência no cancro da próstata.** Porto: Dissertação de Mestrado da Universidade do Porto, 2010.

SILVEIRA, Vinicius. **O que é Data Warehouse?** Obtida via internet. Disponível em: <http://blog.intuitivus.com.br/pt/o-que-e-data-warehouse/>. Acesso em: 06 jun. 2015.

TAN, P-N; STEINBACH, M.; e KUMAR, V. - **Introdução ao Data Mining - Mineração de Dados.** – Ciência Moderna Ltda - RJ, 2009.