

UNIVERSIDADE FEDERAL DO PARANÁ

CARLOS HENRIQUE KURETZKI

**TÉCNICAS DE MINERAÇÃO DE DADOS APLICADAS EM BASES DE DADOS DA
SAÚDE A PARTIR DE PROTOCOLOS ELETRÔNICOS**

CURITIBA

2009

CARLOS HENRIQUE KURETZKI

**TÉCNICAS DE MINERAÇÃO DE DADOS APLICADAS EM BASES DE DADOS DA
SAÚDE A PARTIR DE PROTOCOLOS ELETRÔNICOS**

Dissertação apresentada ao Programa de Pós-graduação em Ciência, Gestão e tecnologia da Informação, Área de concentração: Gestão da Informação e do Conhecimento do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná, como parte das exigências para obtenção do título de Mestre.

Orientador: Prof. Dr. José Simão de Paula Pinto

CURITIBA

2009


Kuretzki, Carlos Henrique, 1984-
Técnicas de mineração de dados aplicadas em bases de dados da
saúde a partir de protocolos eletrônicos
/ Carlos Henrique Kuretzki. – 2009.
98 f. : il. 30 cm

Orientador: Prof. Dr. José Simão de Paula Pinto.
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de
Ciências Sociais Aplicadas, 2009.


1. Informática em saúde. 2. Mineração de dados. 3. Protocolos
eletrônicos. I. Pinto, José Simão de Paula. II. Universidade Federal do
Paraná.
Curso de Pós-Graduação Multidisciplinar em Ciência, Gestão e
Tecnologia da Informação. III. Mineração de dados.

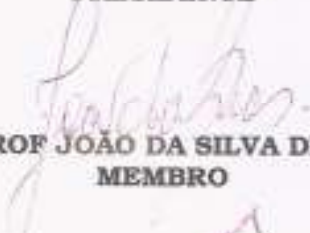
"TÉCNICAS DE MINERAÇÃO DE DADOS APLICADOS EM BASES DE DADOS DA SAÚDE A PARTIR DE PROTOCOLOS ELETRÔNICOS "


ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE MESTRE EM CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO (ÁREA DE CONCENTRAÇÃO: GESTÃO DA INFORMAÇÃO E DO CONHECIMENTO), E APROVADA EM SUA FORMA FINAL PELO CURSO DE PÓS-GRADUAÇÃO MULTIDISCIPLINAR EM CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO DA UNIVERSIDADE FEDERAL DO PARANÁ.


PROF. DR. RICARDO MENDES JUNIOR
COORDENADOR DO MESTRADO

APRESENTADO À COMISSÃO EXAMINADORA INTEGRADA PELOS PROFESSORES:


PROF JOSÉ SIMÃO DE PAULA PINTO
PRESIDENTE


PROF JOÃO DA SILVA DIAS
MEMBRO


PROFº OSVALDO MALAFAIA
MEMBRO

17 de dezembro de 2009

Dedico este trabalho aos meus pais Sérgio e Maria, ao meu avô Estefano e
minha avó Eloina e ao meu irmão Eduardo.

À minha noiva Suelen.

AGRADECIMENTO

Agradeço ao meu orientador PROF. DR. JOSÉ SIMÃO DE PAULA PINTO pelo apoio e dedicação concedida.

Ao PROF. DR. OSVALDO MALAFAIA pela oportunidade de aplicação do trabalho e pelo incentivo da pesquisa.

Ao PROF. SILVIO BORTOLETO que me iniciou na pesquisa científica em minha graduação.

A todos os professores do programa do mestrado em Ciência, Gestão e Tecnologia da Informação. Em especial a PROFA. DRA. DENISE FUKUMI TSUNODA pelo auxílio prestados no desenvolvimento da mineração de dados.

Aos amigos PROF. FARUK ABRÃO KALIL FILHO E PROF. CLEITON TREML pela parceria e contribuição no trabalho.

A PROFA. DRA. MARIA ELIANA MADALOZZO SCHIEFERDECKER pela sua atenção.

A todos os colegas do CENTRO DE ESTUDOS DE REALIDADE VIRTUAL APLICADA e do mestrado. Em especial a MARCUS GARCIA DE ALMEIDA e CLÁUDIA MULLER.

A CAPES e à UNIVERSIDADE FEDERAL DO PARANÁ, pela oportunidade de conhecimento científico e auxílio financeiro para a realização deste trabalho.

“A dúvida é o principio da sabedoria”.

Aristóteles.

RESUMO

A busca por aperfeiçoamento na análise de dados em pesquisas clínicas é constantemente requisitada por profissionais da saúde. Através deste contexto, neste trabalho propõe-se desenvolver métodos de mineração de dados em um software de pesquisas clínicas conhecido como SINPE[®]. O objetivo deste trabalho é proporcionar a profissionais de saúde, usuários de um sistema de protocolos eletrônicos, melhoria na análise dos dados coletados, por meio da utilização de técnicas de mineração de dados. Para tanto se desenvolveu cinco etapas. A primeira, verificar o aceite dos usuários para com a ferramenta proposta. Segunda, Pesquisar na literatura os métodos para a tarefa proposta. Terceira, implementar os métodos pesquisados. Quarta, testar os métodos implementados e por último validar a interface proposta para utilização dos métodos de mineração de dados. Com aceite positivo por parte dos usuários deste software foram implementados os métodos ID3 e Apriori de mineração de dados, sendo testados em dois bancos de dados de protocolos eletrônicos, validando por fim a interface proposta. A metodologia pode ser aplicada com sucesso para o desenvolvimento dos métodos propostos em software de pesquisas clínicas que utilizem a mesma plataforma do SINPE[®].

Palavras-chave: Informática em saúde. Mineração de dados. Protocolos eletrônicos.

ABSTRACT

The search for improvement in data analysis in clinical research is constantly requested by health professionals. Through this context, this work is to develop methods of data mining in a clinical research software known as SINPE ©. The objective of this work is to provide health professionals, users of a system of electronic protocols, improved analysis of data collected through the use of techniques of data mining. For that we developed five steps. First, check the users to accept the proposed tool. Second, the literature search methods for this task. Third, implement the methods studied. Fourth, test methods implemented and finally validate the interface proposed for use in methods of data mining. With positive acceptance by users of this software were implemented methods and ID3 Apriori data mining, and tested on two databases of electronic protocols, finally validating the proposed interface. The methodology can be applied successfully to the development of the proposed methods in clinical research software using the same platform SINPE ©.

Keywords: Biomedical informatics. Data Mining. Electronic protocols.

LISTA DE ILUSTRAÇÕES

QUADRO 1 - FÓRMULA DA ENTROPIA	21
QUADRO 3 - PSEUDOCÓDIGO DO ALGORITMO ID3.....	22
QUADRO 2 - FÓRMULA DO GANHO COM BASE NA ENTROPIA	22
QUADRO 4 - EXEMPLO DE UTILIZAÇÃO DO ALGORITMO ID3.....	23
FIGURA 1 - ÁRVORE DE DECISÃO GERADA PELO ALGORITMO ID3.....	26
FIGURA 2 - ÁRVORE DE DECISÃO COM PODA GERADA PELO ALGORITMO ID3.....	27
QUADRO 5 - FÓRMULA DE SUPORTE	28
QUADRO 6 - FÓRMULA DA CONFIANÇA	28
QUADRO 8 - PSEUDOCÓDIGO DO ALGORITMO APRIORI	29
QUADRO 9 - EXEMPLO DE UTILIZAÇÃO DO ALGORITMO APRIORI.....	29
QUADRO 10 - REPRESENTAÇÃO POR DVD COM CÁLCULO DE SUPORTE	30
QUADRO 11 - CÁLCULO DE SUPORTE PARA DUAS OCORRÊNCIAS	31
QUADRO 12 - CÁLCULO DE SUPORTE PARA TRÊS OCORRÊNCIAS	32
QUADRO 13 - CÁLCULO DA CONFIANÇA.....	33
QUADRO 15 - RESULTADO DO EXEMPLO UTILIZANDO O ALGORITMO APRIORI	33
FIGURA 3 - SINPE© ANALISADOR COM A FUNÇÃO DE MINERAÇÃO DE DADOS.....	39
FIGURA 4 - JANELA DE APRESENTAÇÃO DA MINERAÇÃO DE DADOS	40
FIGURA 5 - JANELA DE ESCOLHA DO MÉTODO COM A OPÇÃO DE CLASSIFICAÇÃO	41
FIGURA 6 - EXEMPLO DO MÉTODO DE CLASSIFICAÇÃO.....	42
FIGURA 7 - JANELA DE ESCOLHA DO MÉTODO COM A OPÇÃO DE ASSOCIAÇÃO	43
FIGURA 8 - EXEMPLO DO MÉTODO DE ASSOCIAÇÃO.....	44
FIGURA 9 - JANELA COM A SELEÇÃO DOS ITENS.....	45
FIGURA 10 - RESULTADO DA MINERAÇÃO DO MÉTODO ID3.....	46
FIGURA 11 - RESULTADO DA MINERAÇÃO DO MÉTODO APRIORI	47
FIGURA 12 - PARÂMETROS DO MÉTODO APRIORI	47
GRÁFICO 1 - CONHECIMENTO NA ÁREA DE INFORMÁTICA EM SAÚDE	50
GRÁFICO 2 - OPINIÃO DOS PARTICIPANTES EM TER UMA FERRAMENTA DE MINERAÇÃO DE DADOS NO SINPE.....	51
FIGURA 13 - RESULTADO DA CLASSE COM O ID3	53
FIGURA 14 - RESULTADO DO WEKA COM O ID3.....	54
FIGURA 15 - ITENS SELECIONADOS DO PROTOCOLO DE DOENÇAS DO APÊNDICE.....	57
FIGURA 16 - RESULTADO DO MÉTODO ID3 PARA O PROTOCOLO DE DOENÇAS DO APÊNDICE.....	58
FIGURA 17 - ITENS SELECIONADOS DO PROTOCOLO DE NUTRIÇÃO DOMICILIAR.....	59
FIGURA 18 - RESULTADO DO MÉTODO ID3 PARA O PROTOCOLO DE NUTRIÇÃO DOMICILIAR.....	60
FIGURA 19 - ÁRVORE GERADA PELO MÉTODO ID3 PARA O PROTOCOLO DE NUTRIÇÃO DOMICILIAR.....	61
FIGURA 20 - ITENS SELECIONADOS DO PROTOCOLO DE DOENÇAS DO APÊNDICE.....	62
FIGURA 21 - RESULTADO DO MÉTODO APRIORI PARA O PROTOCOLO DE DOENÇAS DO APÊNDICE.....	63
FIGURA 22 - ITENS SELECIONADOS DO PROTOCOLO DE NUTRIÇÃO DOMICILIAR.....	64
FIGURA 23 - RESULTADO DO MÉTODO APRIORI PARA O PROTOCOLO DE NUTRIÇÃO DOMICILIAR.....	65
GRÁFICO 3 - NÍVEL DE SATISFAÇÃO DOS ENTREVISTADOS PARA COM O MÉTODO DE CLASSIFICAÇÃO.....	66

GRÁFICO 4 - NÍVEL DE SATISFAÇÃO DOS ENTREVISTADOS PARA COM O MÉTODO DE ASSOCIAÇÃO	67
QUADRO 16 - USABILIDADE X RESULTADO	71
FIGURA 24 - DIAGRAMA DE CLASSE DO MÉTODO ID3	85
QUADRO 17 - LISTA DE FUNÇÕES.....	87

SUMÁRIO

1	INTRODUÇÃO	14
2	REVISÃO DA LITERATURA	16
2.1	CIÊNCIA DA INFORMAÇÃO	16
2.2	PESQUISA MÉDICA	17
2.3	PROTOCOLOS ELETRÔNICOS	18
2.4	SOFTWARES PARA PESQUISA EM SAÚDE	18
2.4.1	Epi Info®	18
2.4.2	Sistema Integrado de Protocolos Eletrônicos	19
2.5	MINERAÇÃO DE DADOS	20
2.5.1	Algoritmo ID3	21
2.5.2	Funcionamento do algoritmo ID3	23
2.5.3	Algoritmo Apriori	27
2.5.4	Funcionamento do algoritmo Apriori	29
2.6	USABILIDADE DE SISTEMAS DE INFORMAÇÃO	34
3	MATERIAL E MÉTODO	36
3.1	VERIFICAR O INTERESSE DOS USUÁRIOS DO SINPE® PARA COM O MÉTODO DE MINERAÇÃO E DADOS E SUA POSSÍVEL IMPLEMENTAÇÃO NESTE SOFTWARE	36
3.2	PESQUISAR NA LITERATURA OS MELHORES ALGORITMOS E TÉCNICAS DE MINERAÇÃO DE DADOS PARA UTILIZAÇÃO NA TAREFA PROPOSTA	37
3.3	IMPLEMENTAR NO MÓDULO ANALISADOR DO SINPE® OS ALGORITMOS SELECIONADOS	37
3.3.1	Preparo do SINPE® Analisador	38
3.3.2	Interface de Mineração de Dados para o usuário	39
3.4	TESTAR ESTES ALGORITMOS COM PROFISSIONAIS DA SAÚDE E SEUS PROTOCOLOS ELETRÔNICOS	47
3.5	VALIDAR A INTERFACE PROPOSTA PARA A MINERAÇÃO DE DADOS ATRAVÉS DO MÓDULO ANALISADOR DO SINPE®	48
4	RESULTADOS	49
4.1	VERIFICAR O INTERESSE DOS USUÁRIOS DO SINPE® PARA COM O MÉTODO DE MINERAÇÃO E DADOS E SUA POSSÍVEL IMPLEMENTAÇÃO NESTE SOFTWARE	49
4.2	PESQUISAR NA LITERATURA OS MELHORES ALGORITMOS E TÉCNICAS DE MINERAÇÃO DE DADOS PARA UTILIZAÇÃO NA TAREFA PROPOSTA	52

4.2.1	Método Id3	52
4.2.2	Método Apriori	54
4.3	IMPLEMENTAR NO MÓDULO ANALISADOR DO SINPE® OS ALGORITMOS SELECIONADOS	55
4.3.1	Algoritmo Id3	55
4.3.2	Algoritmo Apriori.....	55
4.4	TESTAR ESTES ALGORITMOS COM PROFISSIONAIS DA SAÚDE E SEUS PROTOCOLOS ELETRÔNICOS.....	56
4.4.1	Utilizando o método ID3 a partir do protocolo da Cirurgia do Aparelho Digestivo.....	56
4.4.2	Utilizando o método ID3 a partir do protocolo de Nutrição	58
4.4.3	Utilizando o método Apriori a partir do protocolo da Cirurgia do Aparelho Digestivo.....	62
4.4.4	Utilizando o método Apriori a partir do protocolo de Nutrição	63
4.5	VALIDAR A INTERFACE PROPOSTA PARA A MINERAÇÃO DE DADOS ATRAVÉS DO MÓDULO ANALISADOR DO SINPE®	65
5	DISCUSSÃO	68
6	CONCLUSÕES	73
7	APÊNDICE.....	77
7.1	QUESTIONÁRIO DE PRÉ-TESTE	77
7.2	QUESTIONÁRIO.....	78
7.3	QUESTIONÁRIO PÓS-IMPLEMENTAÇÃO	81
7.4	DOCUMENTAÇÃO DE IMPLEMENTAÇÃO.....	84
7.5	MANUAL DE UTILIZAÇÃO DO SINPE® ANALISADOR: MINERAÇÃO DE DADOS.....	87

1 INTRODUÇÃO

A necessidade de aperfeiçoamento de resultados obtidos através de coleta de dados tornou-se fator prioritário para com o SINPE[®] (Sistema Integrado de Protocolos Eletrônicos). Esta demanda ganhou relevância devido ao número de usuários que utilizam esta ferramenta em suas pesquisas, proporcionando novos trabalhos científicos e auxiliando a área da saúde. Visualizando a potencialidade de melhoria deste aplicativo, propõem-se uma nova funcionalidade para seu uso, nova funcionalidade que atribui qualidade e potencialidade nos resultados obtidos através de pesquisas.

Para isso implementou-se dois métodos de mineração de dados nesta ferramenta. O primeiro é o método ID3 que visa estabelecer relações entre sequência de itens como: pacientes que no exame físico apresentaram sinal de Blumberg passaram por cirurgia do tipo apendicectomia com dor abdominal, estão propensos a terem apendicite aguda. O segundo é o método Apriori que relaciona dois ou mais conjuntos de itens a fim de gerar regras como: pacientes que passaram por cirurgia de apendicectomia tiveram dor abdominal.

A justificativa da pesquisa é disponibilizar métodos para implementação da mineração de dados em um software de pesquisas clínicas, possibilitando aos pesquisadores melhor qualidade na análise de dados, de forma fácil e rápida, sem que para isso sejam necessários conhecimentos técnicos profundos de mineração de dados.

O objetivo deste trabalho é proporcionar a profissionais de saúde, usuários de um sistema de protocolos eletrônicos, melhoria na análise dos dados coletados, por meio da utilização de técnicas de mineração de dados e está subdividido em cinco objetivos específicos:

1. verificar o interesse dos usuários do SINPE[®] para com o método de mineração de dados e sua possível implementação neste software;
2. pesquisar na literatura os melhores algoritmos e técnicas de mineração de dados para utilização na tarefa proposta;
3. implementar no módulo Analisador do SINPE[®] os algoritmos selecionados;

4. testar estes algoritmos com profissionais da saúde e seus protocolos eletrônicos;
5. validar a interface proposta para a mineração de dados do módulo Analisador do SINPE[®].

2 REVISÃO DA LITERATURA

2.1 CIÊNCIA DA INFORMAÇÃO

Segundo (LIMA, cap. 22, 2005) ciência da informação é o estudo da informação por si só, envolve criação, identificação, coleta, validação, representação, reuperação e o uso das informações, tendo como base um produtor ou consumidor de informação dos quais buscam sentido e finalidade para com a informação.

Meio potencializador da informação é a tecnologia da informação. Para autores como Davenport, McGee e Prusak a tecnologia da informação deve ser considerada suporte da informação. "Portanto, informação é recurso estratégico, que necessita de atenção e tratamento gerencial adequados" (LIMA, cap. 22, 2005).

Informação em saúde, do ponto de vista da saúde, é recurso básico de qualquer sociedade e, por conseguinte, a informação em saúde é fundamental ao processo de tomada de decisões no âmago das políticas públicas, objetivando elevar a qualidade de vida dos povos. Informações extraídas com base na área da saúde são imprescindíveis ao planejamento, à implantação, à implementação e à avaliação de ações e serviços de saúde. A função macro da informação em saúde é detectar problemas individuais e coletivos do quadro sanitário de uma população, oferecendo elementos para a análise desse quadro a fim de apresentar alternativas para minimizar a situação encontrada. Este tema está intrinsecamente vinculado à pesquisa em saúde, que se constitui *per se* em fonte inesgotável de novos saberes, haja vista a dinamicidade inerente à própria ciência (TARGINO, 2009).

Informação e conhecimento são conceitos distintos, tendo em vista que para obter conhecimento é necessário informação. Para obtenção de conhecimento é necessário integrar informação com outros dados e informações, sendo possível esta obtenção (BRASIL, 2008).

2.2 PESQUISA MÉDICA

Pesquisas médicas são efetuadas diariamente, de diferentes modos, para obter os dados necessários mediante *clinical trial* que é o experimento no qual dados de interações de determinados pacientes são agrupados, ordenados e analisados a fim de aprender sobre um campo de estudo, validar a eficácia de um novo tratamento ou perceber um processo de uma determinada doença que não é de outra maneira compreendida (SHORTLIFFE e PERREAULT, 2001; SHORTLIFFE e CIMINO 2006).

O propósito do *clinical trial*, é determinar os benefícios e os danos de uma intervenção. Esta determinação criada, serve para medir os efeitos de diferentes tratamentos e resultados (SINHA e JONES *et al.*, 2008).

Questionário é um dos métodos utilizados como instrumentos de estudo baseados em populações para a pesquisa. (LCHAT, VERSTRAETEN *et al.*, 2008)

O método utilizado para coleta de dados de pacientes é ainda a de questionários em papel (LUNGENHAUSEN, LANGE *et al.*, 2007).

Questionários via correspondência e entrevistas via telefone são dois métodos efetivos para coleta de dados, combinar estes dois métodos pode reduzir a não resposta do levantamento. Reunir estes dois tipos de métodos necessita cautela, se os diferentes modos de coleção de dados obterem diferentes padrões de respostas ou respostas de diferentes partes da população (Feveile, Olsen *et al.*).

Segundo Shortliffe e Barnett (2001, p. 59) estudo retrospectivo é utilizado para investigar questões que não são assuntos de estudo no tempo em que os dados foram coletados.

A execução de *guidelines* clínicos é um dos mais interessantes tópicos de estudo em informática clínica. Eles contem parte de orientações ou princípios para ajudar nos cuidados de saúde, como cuidado sobre decisões adequadas a diagnósticos, terapêutico, ou outros procedimentos para específicas circunstâncias clínicas (ISERN e MORENO, 2008).

2.3 PROTOCOLOS ELETRÔNICOS

Segundo o *Dictionary of Medical Terms* protocolo é parte de instruções para a gestão clínica de uma condição particular, inclusive testes, cirurgia e tratamento de drogas. “Protocolo de pesquisa é projetado especificamente para coleta futura de dados.” (SHORTLIFFE e BARNETT, 2001, p. 59).

Registro de pacientes baseados em computador são repositórios que conservam eletronicamente informações sobre a existência individual da posição da saúde e de seu cuidado, assim sendo ele pode servir a vários usuários (SHORTLIFFE e BARNETT, 2001, p. 328).

Segundo Borsato (2005), com a coleta de dados utilizando protocolos é possível realizar pesquisas clínicas, quanto maior o número de coletas mais preciso será o resultado.

O uso de protocolos eletrônicos permite a promoção de informações confiáveis, acelera o processo de armazenamento e tratamento dos dados, utilizando uma matriz digitalizada, podem ser acessados a qualquer tempo para levantamentos na gestão de recursos, sendo fundamental para análise gerencial de resultados. (WALACH, 2008).

2.4 SOFTWARES PARA PESQUISA EM SAÚDE

2.4.1 Epi Info[®]

Segundo Bós (2004), o Epi Info[®] é um conjunto de softwares implementado para a plataforma Microsoft Windows. Desenvolvido pelo Centro Americano de Controle e Prevenção de Doenças, seus módulos são utilizados por profissionais da saúde que conduzem investigações de epidemias, administração de bancos de dados para vigilância de saúde pública, além de ser um banco de dados para uso

geral e aplicações estatísticas. Com o Epi Info[®] alunos, pesquisadores e profissionais da saúde podem desenvolver um questionário personalizando o processo de entrada de dados e gerenciando a análise deste banco de dados com base em seus questionários. Por se tratar de um aplicativo de domínio público ele pode ser baixado gratuitamente a partir da Internet através do endereço <<http://www.cdc.gov/EpiInfo/>>.

A última versão do Epi Info[®] lançada em 18 de agosto de 2008 é a 3.5.1, este aplicativo necessita de um sistema operacional *Microsoft Windows 98* ou superior para o seu funcionamento (EPI INFO[™], 2009).

As ferramentas agregadas ao Epi Info[®] permitem obter estatísticas epidemiológicas como tabelas e gráficos, além de produzir mapas com dados de distribuição geográficas (PINTO, 2005).

2.4.2 Sistema Integrado de Protocolos Eletrônicos

Segundo Pinto (2005), o Sistema Integrado de Protocolos Eletrônicos (SINPE[®]), possui direito registrado no Instituto Nacional de Propriedade Industrial sob número 00051543, com orientação e propriedade intelectual do Prof. Dr. Osvaldo Malafaia. Este sistema possui dezessete anos de pesquisa e desenvolvimento permitindo a construção, coleta e análise de protocolos eletrônicos na área da saúde.

O SINPE[®] passou por três fases (versões) de desenvolvimento. A primeira desenvolvida em Java utilizava um banco de dados centralizado com uma arquitetura cliente-servidor. Porém esta versão não teve continuidade devido a problemas logísticos. A segunda versão desenvolvida em Delphi utilizava o banco de dados Microsoft[®] Access, sendo substituída por uma terceira versão devido a alguns problemas de desempenho. Já sua terceira e última versão desenvolvida em C#, conta com ambos os tipos de conexões, local utilizando Microsoft[®] Access e remota utilizando Microsoft[®] SQL Server, permitindo que o usuário trabalhe em um ambiente descentralizado, multicêntrico e com múltiplos usuários (MALAFAIA, DIETZ *et al.*, 2006).

Este software possui dois módulos. O primeiro é o SINPE[®] operacionalizado pelo Dr. Emerson Paulo Borsato desenvolvido para construção dos protocolos de pesquisa e coleta de dados. O segundo módulo é o SINPE[®] Analisador também conhecido como Interface de Visibilização de Informação, operacionalizado pelo Prof. Dr. José Simão de Paula Pinto, para disponibilizar relatórios e gráficos através de estatística descritiva com base em protocolos e coletas de dados previamente registradas (KALIL FILHO, 2008).

O objetivo do SINPE[®] Analisador é o rápido levantamento de informações contidas nos protocolos de pesquisa proporcionados pelas coletas de dados do SINPE[®]. Não é objetivo deste módulo, concorrer com ferramentas de análise estatística ou de geração de gráficos, mas de proporcionar ao usuário deste software meios de agilizar e simplificar a obtenção de dados para pesquisas e geração de aulas. A escolha da linguagem de programação utilizada para este módulo foi a C#, este fato ocorreu devido à compatibilidade com o software desenvolvido pelo Dr. Emerson Paulo Borsato. Utilizando este módulo o pesquisador consegue obter: levantamento de seu método, gráficos e informações para constar nos resultados de sua pesquisa e o uso de estatística descritiva (PINTO, 2005).

2.5 MINERAÇÃO DE DADOS

Mineração de Dados ou *Data Mining* tem como propósito analisar uma parte de dados ou informações em ordem para origem indefinida e potencialmente padrões úteis. (FAYYAD, PIATETSKY-SHAPIRO *et al.*, 1996).

Segundo Van Bommel (1997), técnicas de Mineração de Dados e Descoberta de Conhecimento em Bases de Dados eram originalmente desenvolvidas para processos corporativo de vendas e produção de dados, mas eles são igualmente relevantes para parâmetros de cuidados da saúde.

Para Witten e Frank (2006) Mineração de Dados é definido como o processo de descoberta de padrões em dados. Estes processos devem ser automáticos ou semi-automáticos. Já os padrões descobertos devem ser significativos e que eles levem a algum benefício.

“Muitas vezes a Mineração de Dados é uma parte do processo de descoberta do conhecimento sendo um dos campos mais importantes em gestão do conhecimento. [...]Técnicas como: modelos Bayesianos, árvore de decisão, redes neurais artificiais, regras de associação e algoritmos genéticos são geralmente usados na descoberta de padrões ou de conhecimento que são anteriormente desconhecidos para o sistema e para os usuários. A Mineração de Dados pode ser usada em: aplicativos de marketing, relação com cliente, engenharia, medicina, análise de crimes, prognóstico especialista, mineração na Web e computação móvel, entre outros.” (CHEN, FULLER *et al.*, 2005).

2.5.1 Algoritmo ID3

O algoritmo ID3 é um algoritmo de classificação com árvore de decisão. A Classificação é o processo que encontra propriedades comuns entre determinado conjunto de dados a fim de ordenar e classificar estes dados (Tsunoda, 2004).

Para Quilan (1986), classificação com árvores de decisão pode ser aplicado a diagnósticos de um sintoma médico no qual as classes podem ser qualquer estágio de uma doença ou suas possibilidades de tratamento.

Árvores de decisão são facilmente executadas por sistemas e podem ser muito precisas, mas elas deixam a desejar em relação à representação de conhecimento (QUINLAN, 1986).

Para gerar árvore de decisão a partir do método ID3, é necessário se obter o ganho de informação, descrito na fórmula 2, este cálculo é dependente da “entropia”, descrito na fórmula 1, ela representa as ocorrências de cada valor do atributo distribuído para cada classe (TSUNODA, 2004).

$\text{Entropia}(C) = - \sum_{i=1}^n p_i \log_2 p_i$	1
onde	
C = é o conjunto de classes	
n = é o número de classes	
p _i = é a frequência da classe <i>i</i>	

QUADRO 1 - FÓRMULA DA ENTROPIA
FONTE: (TSUNODA, 2008)

$$\text{Ganho}(C,A) = \text{Entropia}(C) - \sum_{i=1}^n p(t_i^A) \text{Entropia}(C|t_i^A)$$

onde

Entropia(C) = é o cálculo da entropia das classes antes da participação, conforme a equação anterior;

n = é o número de classes

i = é a *i*-ésima partição, correspondente ao *i*-ésimo valor do atributo A

$p(t_i^A)$ = é a frequência da *i*-ésima partição baseada no atributo A

Entropia(C| t_i^A) = é a entropia da *i*-ésima partição

2

QUADRO 2 - FÓRMULA DO GANHO COM BASE NA ENTROPIA

FONTE: (TSUNODA, 2008)

Id3 (D, A, Atributos)

1. // com **D** o conjunto de exemplos, **A** o atributo alvo, e **Atributos** a lista de atributos que caracteriza as entidades;
2. **criar** um nó raiz da árvore;
3. **se** todos os exemplos são positivos **então**
4. rotular o nó como positivo (+) e devolver a árvore;
5. **se** todos os exemplos são negativos **então**
6. rotular o nó como negativo (-) e devolver a árvore;
7. **se** a lista de atributos está vazia **então**
8. devolver a árvore, com um rótulo idêntico ao valor assumido pela maioria dos exemplos;
9. **senão**
10. A \rightarrow o atributo da lista de atributos, que melhor classifique os exemplos;
11. usa-se o Ganho de Informação para escolher o atributo que melhor classifique os exemplos;
12. o nó recebe A, como rótulo;
13. **para cada** valor possível v_i de Valores(A) **faça**
14. adicionar um novo ramo, a partir do nó, correspondente ao teste $A = v_i$;
15. $D_{v_i} \rightarrow$ o subconjunto de D, com $A = v_i$;
16. **se** $D_{v_i} = \{\}$: **então**
17. devolver a árvore, com um rótulo idêntico ao valor assumido pela maioria dos exemplos;
18. **senão**
19. abaixo deste ramo adicionar a sub-árvore **Id3 (D_{v_i} , A, Atributos- $\{A\}$)**

QUADRO 3 - PSEUDOCÓDIGO DO ALGORITMO ID3

Fonte: (TSUNODA, 2008)

2.5.2 Funcionamento do algoritmo ID3

Para verificar o funcionamento do algoritmo ID3 teremos como base o QUADRO 4 o caso dado não é recursivo como especificado pelo algoritmo deste método, porém interpreta seu funcionamento.

Considerando os registros deste quadro identifica-se qual será as características para o lançamento de um novo produto, uma barra de cereal composta um complemento e uma fruta. Além destes dados cada registro possui a identificação do país onde foi realizada a pesquisa para o lançamento desta barra de cereal.

Quantidade de registros	País	Complemento	Fruta	Compra ou não a barra de cereal?
1	Suíça	Baunilha	Banana	Sim
2	Brasil	Chocolate	Banana	Sim
3	Brasil	Baunilha	Laranja	Não
4	Brasil	Baunilha	Banana	Não
5	Brasil	Baunilha	Laranja	Não
6	Brasil	Chocolate	Banana	Não
7	EUA	Chocolate	Laranja	Sim
8	EUA	Baunilha	Laranja	Sim
9	EUA	Chocolate	Laranja	Sim
10	Suíça	Baunilha	Laranja	Sim
11	Suíça	Chocolate	Banana	Não
12	Suíça	Baunilha	Banana	Sim

QUADRO 4 - EXEMPLO DE UTILIZAÇÃO DO ALGORITMO ID3

FONTE: O autor (2009)

Temos como primeiro passo o cálculo da entropia do conjunto. Levando em consideração os valores sim e não do atributo meta, neste caso este atributo é o campo “Compra ou não a barra de cereal?” do quadro exemplo.

Este cálculo é descrito por Shanon já apresentado neste trabalho pela formula 1.

$$S(\text{Compra a barra, sim}) = -\frac{7}{12} \log_2 \frac{7}{12}$$

$$S(\text{Compra a barra, sim}) = -0,5833(-0,7776)$$

$$S(\text{Compra a barra, sim}) = 0,4536$$

$$S(\text{Compra a barra, não}) = -\frac{5}{12} \log_2 \frac{5}{12}$$

$$S(\text{Compra a barra, não}) = -0,4167(-1,2630)$$

$$S(\text{Compra a barra, não}) = 0,5263$$

$$S(\text{Compra a barra, } \Sigma) = 0,4536 + 0,5263$$

$$S(\text{Compra a barra, } \Sigma) = \boxed{0,9799}$$

O próximo passo é calcular a entropia e o ganho de informação para cada coluna do quadro exemplo, para tanto é necessário a utilização da fórmula 2, de ganho de informação. Abaixo estão indicados os cálculos pela coluna país.

$$S(\text{Brasil, sim}) = -\frac{1}{5} \log_2 \frac{1}{5}$$

$$S(\text{Brasil, sim}) = 0,4644$$

$$S(\text{Brasil, não}) = -\frac{4}{5} \log_2 \frac{4}{5}$$

$$S(\text{Brasil, não}) = 0,2575$$

$$S(\text{Brasil, } \Sigma) = 0,4644 + 0,2575$$

$$S(\text{Brasil, } \Sigma) = 0,7219$$

$$S(\text{EUA, sim}) = -\frac{3}{3} \log_2 \frac{3}{3}$$

$$S(\text{EUA, sim}) = 0$$

$$S(\text{EUA, não}) = 0$$

$$S(\text{EUA, } \Sigma) = 0 + 0$$

$$S(\text{EUA, } \Sigma) = 0$$

$$S(\text{Suíça, sim}) = -\frac{3}{4} \log_2 \frac{3}{4}$$

$$S(\text{Suíça, sim}) = 0,3113$$

$$S(\text{Suíça, não}) = -\frac{1}{4} \log_2 \frac{1}{4}$$

$$S(\text{Suíça, não}) = 0,5000$$

$$S(\text{Suíça, } \Sigma) = 0,3113 + 0,5000$$

$$S(\text{Suíça, } \Sigma) = 0,8113$$

Uma vez calculado a entropia da coluna país para cada conjunto de registros, deve-se agora calcular o ganho de informação para esta coluna, levando em consideração cada resultado destes conjuntos.

$$G(\text{País}) = 0,9799 - 0,7219 \frac{5}{12} - 0 \frac{3}{12} - 0,8113 \frac{4}{12}$$

$$G(\text{País}) = \boxed{1,5511}$$

O mesmo procedimento deve ser efetuado para as demais colunas. O resultado do ganho da informação por coluna dará a ordem em que esta coluna será representada na árvore. A próxima coluna calculada será a coluna complemento.

$$S(\text{Baunilha, sim}) = -\frac{4}{7} \log_2 \frac{4}{7}$$

$$S(\text{Baunilha, sim}) = 0,4613$$

$$S(\text{Baunilha, não}) = -\frac{3}{7} \log_2 \frac{3}{7}$$

$$S(\text{Baunilha, não}) = 0,5239$$

$$S(\text{Baunilha, } \Sigma) = 0,4613 + 0,5239$$

$$S(\text{Baunilha, } \Sigma) = 0,9852$$

$$S(\text{Chocolate, sim}) = -\frac{3}{5} \log_2 \frac{3}{5}$$

$$S(\text{Chocolate, sim}) = 0,4422$$

$$S(\text{Chocolate, não}) = -\frac{2}{5} \log_2 \frac{2}{5}$$

$$S(\text{Chocolate, não}) = 0,5288$$

$$S(\text{Chocolate, } \Sigma) = 0,4422 + 0,5288$$

$$S(\text{Chocolate, } \Sigma) = 0,9710$$

$$G(\text{Complemento}) = 0,9799 - 0,9852 \frac{7}{12} - 0,9710 \frac{5}{12}$$

$$G(\text{Complemento}) = \boxed{0,0006}$$

A última coluna calculada é a coluna fruta.

$$S(\text{Banana, sim}) = -\frac{3}{6} \log_2 \frac{3}{6}$$

$$S(\text{Banana, sim}) = 0,5000$$

$$S(\text{Banana, não}) = -\frac{3}{6} \log_2 \frac{3}{6}$$

$$S(\text{Banana, não}) = 0,5000$$

$$S(\text{Banana, } \Sigma) = 0,5000 + 0,5000$$

$$S(\text{Banana}, \Sigma)=1$$

$$S(\text{Laranja, sim})=-\frac{4}{6}\log_2\frac{4}{6}$$

$$S(\text{Laranja, sim})=0,3900$$

$$S(\text{Laranja, não})=-\frac{2}{6}\log_2\frac{2}{6}$$

$$S(\text{Laranja, não})=0,5283$$

$$S(\text{Laranja}, \Sigma)=0,3900+0,5283$$

$$S(\text{Laranja}, \Sigma)=0,9183$$

$$G(\text{Fruta})=0,9799-1\frac{6}{12}-0,9183\frac{6}{12}$$

$$G(\text{Fruta})=\boxed{0,0207}$$

O próximo passo é montar a árvore. A prioridade dos itens identificados nela é definida com a fórmula de ganho de informação, a coluna com o maior ganho ficará em primeiro na árvore. Neste exemplo, a coluna com o maior valor é a coluna “País” tendo o resultado do ganho de 1,5511, a segunda é a coluna “Fruta” com resultado 0,0207, por fim a coluna “Complemento” tendo de resultado 0,0006.

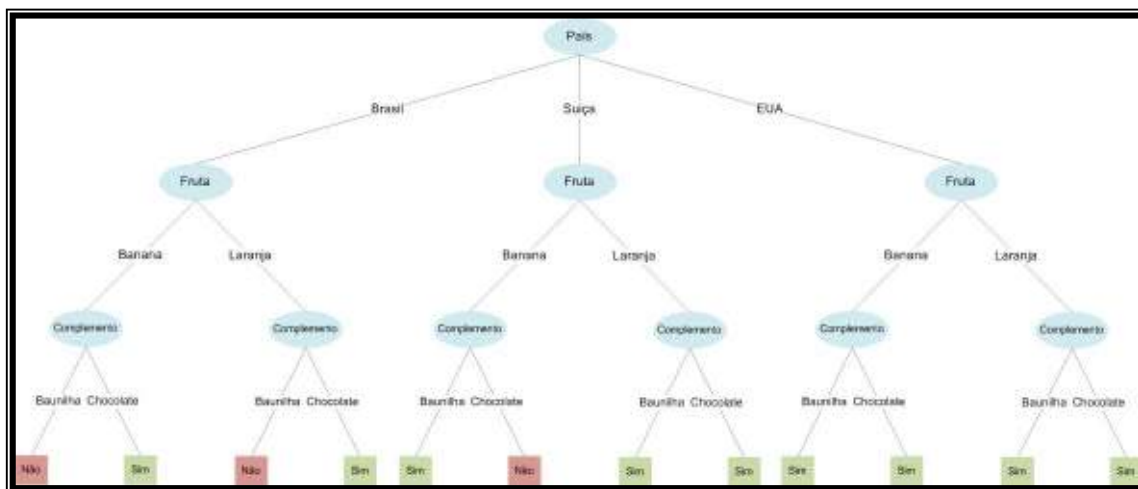


FIGURA 1 - ÁRVORE DE DECISÃO GERADA PELO ALGORITMO ID3
 FONTE: O autor (2009)

O último passo é simplificar a árvore criada, retirando os itens repetidos do atributo meta, este procedimento é conhecido como poda *prunning* de árvore em mineração de dados.

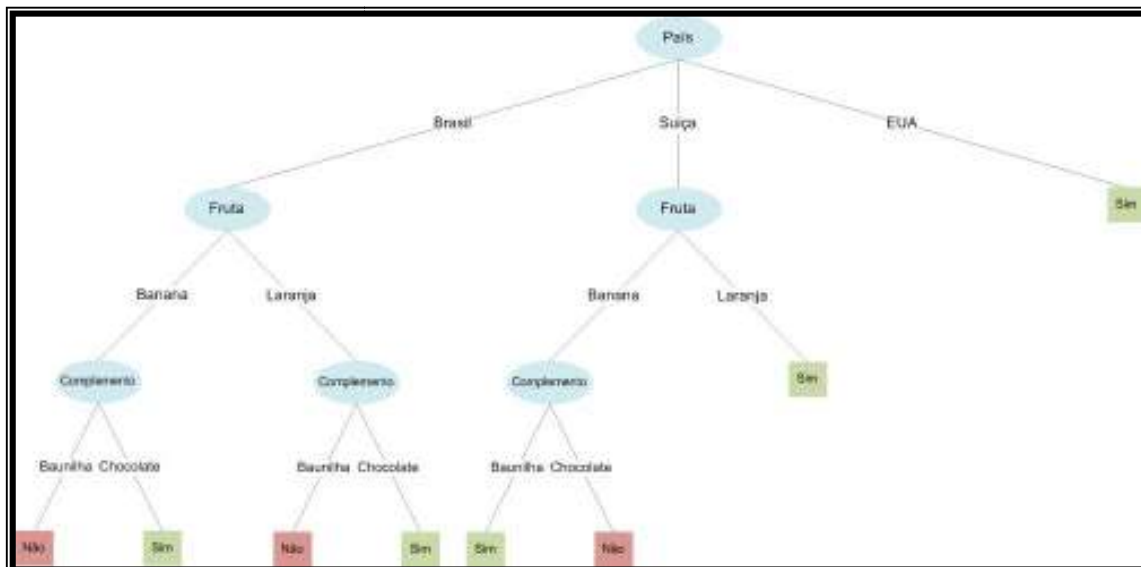


FIGURA 2 - ÁRVORE DE DECISÃO COM PODA GERADA PELO ALGORITMO ID3
 FONTE: O autor (2009)

A árvore torna-se útil ao apoio à decisão quando validada por um ou mais especialistas. Tendo em vista sua visualização e possibilidade de acompanhamento dos itens selecionados para seu resultado final.

2.5.3 Algoritmo Apriori

Mineração de dados utilizando regras de associação foi descrita por Agrawal. Regras de associação mostram resultados obtidos através da descoberta de padrões em itens (AGRAWAL, IMIELINSKI *et al.*, 1993).

Esta tarefa consiste em encontrar conjuntos de itens que ocorram simultaneamente e de forma frequente em um banco de dados (GOLDSCHMIDT E PASSOS, 2005).

Agrawal (1993) descreve suporte como uma fração de transações em determinado item X que satisfaz a união dos itens em seu consequente e antecedente. Conforme descrito na equação 3:

$$FSuporte = \frac{|X \cup Y|}{N}$$

onde

X = número de ocorrências da primeira coluna

Y = número de ocorrências da segunda coluna

N = total de registros

QUADRO 5 - FÓRMULA DE SUPORTE
 FONTE: (Tsunoda, 2008)

O suporte não pode ser confundido com a confiança, enquanto a confiança é uma medida da força da regra, suporte corresponde à significância estatística (AGRAWAL *et al.*, 1993). A equação da confiança é descrita:

$$FConfiança = \frac{|X \cup Y|}{X}$$

onde

X = número de ocorrências da primeira coluna

Y = número de ocorrências da segunda coluna

QUADRO 6 - FÓRMULA DA CONFIANÇA
 FONTE: (TSUNODA, 2008)

O Quadro 7 mostra o funcionamento e formato de pseudocódigo do algoritmo Apriori.

1. F_1 = conjuntos frequentes de tamanho 1;
2. $k = 1$;
3. **enquanto** ($F_k \neq 0$) **faça**
4. $k = k + 1$;
5. Gerar C_k (todos os candidatos de tamanho k) a partir de F_{k-1} ;
6. **para cada** transação t pertencente a base de dados **faça**
7. **para cada** candidato em C_k **faça**
8. **se** todos os itens do candidato pertencem a t **então**
9. Incrementar o contador associado ao candidato;
10. F_k = todos os candidatos pertencentes a C_k com suporte
11. mínimo maior ou igual a $SupMin$;
12. **fim-enquanto**;
13. Resposta = união de todos os conjuntos F_k ;

QUADRO 8 - PSEUDOCÓDIGO DO ALGORITMO APRIORI

Fonte: (TSUNODA, 2008)

2.5.4 Funcionamento do algoritmo Apriori

Para verificar o funcionamento do algoritmo Apriori, utiliza-se um caso.

Determinada locadora está criando pacotes de locação. O objetivo é disponibilizar estes pacotes contendo dois ou três filmes. Utilizando o Apriori, com suporte mínimo de 0,4 e confiança mínima de 0,7, quais pacotes serão sugeridos para os clientes desta locadora? Estes valores de suporte e confiança foram definidos aleatoriamente.

O Quadro 9 representa as locações efetuadas pelos clientes desta locadora, onde cada identificador representa uma locação e seus filmes locados, representado pela coluna filmes locados.

Identificador	Filmes locados
1	Rei Leão, Pequena Sereia, Monstros S/A
2	Monstros S/A, Shrek, Hercules
3	Pequena Sereia, Shrek, Tarzan, Rei Leão
4	Pequena Sereia, Shrek, Hercules, Monstros S/A, Tarzan
5	Rei Leão, Pequena Sereia, Monstros S/A, Shrek
6	Rei Leão, Monstros S/A, Hercules, Tarzan
7	Tarzan, Shrek, Monstros S/A
8	Monstros S/A, Shrek, Hercules

QUADRO 9 - EXEMPLO DE UTILIZAÇÃO DO ALGORITMO APRIORI

FONTE: O autor (2009)

O primeiro passo para a execução do algoritmo Apriori, é tabular os produtos listados no Quadro 9 relacionando-os com o registro de locações. Em outras palavras cada linha do Quadro 10 representa uma locação.

A organização apresentada pelo Quadro 10 é necessária devido aos cálculos de suporte efetuados neste algoritmo. O primeiro cálculo de suporte efetuado é referente a cada coluna do quadro, a fim de verificar quais destas colunas utiliza-se para associá-las as demais colunas.

O cálculo de suporte por coluna é dado pela seguinte formula:

$$\text{SupColuna} = \frac{X}{N}$$

Sendo que X representa o número de ocorrências de determinado produto em relação a todas locações e N representa o total de locações.

Vamos efetuar este exemplo para a coluna do primeiro produto que é o DVD Rei Leão, também representado por DVD1. O número total de locações para este DVD é de 4 em um total de 8 registros, portanto tem-se o seguinte resultado:

$$\text{SupColuna} = \frac{4}{8} = 0,5$$

Após efetuar este mesmo cálculo para todas as colunas deste exemplo, nota-se que todas colunas possuem suporte acima de 0,4, conforme descrito no enunciado. Devido a este fato todas as colunas serão utilizadas nos demais cálculos.

	DVD1	DVD2	DVD3	DVD4	DVD5	DVD6
Identificador	Rei Leão	Pequena Sereia	Monstros	Shrek	Hercules	Tarzan
1	x	x	x			
2			x	X	x	
3	x	x		X		x
4		x	x	X	x	x
5	x	x	x	X		
6	x		x		x	x
7			x	X		x
8			x	X	x	
Suporte	0,50	0,50	0,87	0,75	0,50	0,50

QUADRO 10 - REPRESENTAÇÃO POR DVD COM CÁLCULO DE SUPORTE
 FONTE: O autor (2009)

O próximo passo é agrupar em pares as colunas que passaram pelo teste de suporte, calculando novamente o suporte com as ocorrências destes pares. Este agrupamento deve contemplar todas as possibilidades " 2^n " referente às colunas que possuem suporte maior ou igual ao exposto pelo enunciado, conforme representado pelo QUADRO 11.

O cálculo para o suporte destas combinações é diferente do cálculo de suporte por coluna, uma vez que é necessário contemplar as ocorrências das colunas envolvidas nestas combinações, conforme a equação 3.

Exemplificando esta fórmula de suporte com combinações, toma-se como base a primeira combinação "DVD1, DVD2". As ocorrências desta combinação aparecem nas linhas 1, 3 e 5 do QUADRO 10. Tendo a seguinte resolução:

$$FSuporte = \frac{3}{8} = \boxed{0,38}$$

Esta mesma fórmula deve ser aplicada para as demais combinações possíveis, conforme apresentado no QUADRO 11.

Combinações	Suporte
DVD1, DVD2	0,38
DVD1, DVD3	0,38
DVD1, DVD4	0,25
DVD1, DVD5	0,13
DVD1, DVD6	0,25
DVD2, DVD3	0,38
DVD2, DVD4	0,38
DVD2, DVD5	0,13
DVD2, DVD6	0,25
DVD3, DVD4	0,63
DVD3, DVD5	0,50
DVD3, DVD6	0,38
DVD4, DVD5	0,38
DVD4, DVD6	0,38
DVD5, DVD6	0,25

QUADRO 11 - CÁLCULO DE SUPORTE PARA DUAS OCORRÊNCIAS
FONTE: O autor (2009)

Após efetuado os cálculos para todas as ocorrências, os valores que atenderem o suporte devem ser destacados, no exemplo tem-se: DVD3, DVD4 e DVD3, DVD5, pois apresenta suporte maior do que 0,4, conforme estipulado no enunciado deste exemplo. Caso nenhum dos itens agrupados tenham o suporte suficiente, o algoritmo é encerrado não passando nenhuma regra do resultado. Como no exemplo destacou-se ocorrências, a mesma formula deve ser aplicada para grupos de três colunas, como descrito no QUADRO 12.

Para a primeira ocorrência contendo DVD1, DVD2, DVD3:

$$FSuporte = \frac{2}{8} = 0,25$$

Combinações	Suporte
DVD1, DVD2, DVD3	0,25
DVD1, DVD2, DVD4	0,25
DVD1, DVD2, DVD5	0
DVD1, DVD2, DVD6	0,12
DVD1, DVD3, DVD4	0,12
DVD1, DVD3, DVD5	0,12
DVD1, DVD3, DVD6	0,12
DVD1, DVD4, DVD5	0
DVD1, DVD4, DVD6	0,12
DVD1, DVD5, DVD6	0,12
DVD2, DVD3, DVD4	0,25
DVD2, DVD3, DVD5	0,12
DVD2, DVD3, DVD6	0,12
DVD2, DVD4, DVD5	0,12
DVD2, DVD4, DVD6	0,25
DVD2, DVD5, DVD6	0,12
DVD3, DVD4, DVD5	0,37
DVD3, DVD4, DVD6	0,25
DVD3, DVD5, DVD6	0,25
DVD4, DVD5, DVD6	0,25

QUADRO 12 - CÁLCULO DE SUPORTE PARA TRÊS OCORRÊNCIAS
FONTE: O autor (2009)

Caso ainda exista algum registro do grupo de três elementos que possua o suporte maior ou igual ao definido, deve-se agrupá-los em quatro colunas e assim

sucessivamente. Neste exemplo nenhum registro com três colunas apresentou suporte.

O último passo é calcular a confiança para os registros que passaram pelo suporte estipulado, este cálculo é dado pela fórmula:

Tendo em vista que o X representa o número de ocorrências da primeira coluna e o Y representa as ocorrências da segunda coluna, obtendo a seguinte resolução para DVD3, DVD4:

$$F\text{Confiança} = \frac{5}{7} = 0,7142$$

Esta fórmula deve ser aplicada para todas ocorrências que apresentaram resultado maior que o suporte estipulado, representado no QUADRO 13.

Combinações	Suporte	Confiança
DVD3, DVD4	0,63	0,71
DVD4, DVD3	0,63	0,83
DVD3, DVD5	0,50	0,57
DVD5, DVD3	0,50	1,00

QUADRO 13 - CÁLCULO DA CONFIANÇA
FONTE: O autor (2009)

Recordando que DVD3 representa Monstros S/A, DVD4 é Shrek e DVD5 é Hércules, os pacotes sugeridos para os clientes desta locadora são apresentados no Quadro 14.

Combinações	Suporte	Confiança
Monstros S/A e Shrek	0,63	0,71
Shrek e Monstros S/A	0,63	0,83
Hercules e Monstros S/A	0,50	1,00

QUADRO 15 - RESULTADO DO EXEMPLO UTILIZANDO O ALGORITMO APRIORI
FONTE: O autor (2009)

Estes três pacotes do QUADRO 15, atendem às condições impostas, tendo suporte maior ou igual 0,4 e confiança maior ou igual a 0,7, sendo que Monstros S/A e Shrek e Shrek e Monstros S/A atendem ambas condições independentes de sua ordem.

2.6 USABILIDADE DE SISTEMAS DE INFORMAÇÃO

O sistema de informação (SI) coleta, processa, armazena, analisa e dissemina informações com um determinado objetivo. Como qualquer outro sistema, o sistema de informação inclui *inputs* (dados, instruções) e *outputs* (relatórios, cálculos). Ele *processa* os *inputs* e produz *outputs*, que são enviados para o usuário ou para outros sistemas. Um sistema de informação baseado em computador é um sistema de informação que utilize tecnologia da computação para executar algumas ou todas as tarefas desejadas. (TURBAN; MCLEAN; WETHERBE, 2004, p. 39)

Um sistema de informação pode ser definido tecnicamente como um conjunto de componentes inter-relacionados que coletam (ou recuperam), processam, armazenam e distribuem informações destinadas a apoiar a tomada de decisões, a coordenação e o controle de uma organização. (LAUDON E LAUDON, 2007).

Sistema especialista é uma área de pesquisa em Inteligência Artificial, que consiste em prover o computador de mecanismos e meios a que ele aja como um especialista de determinada área. Assim sendo o computador estaria apto a: fazer diagnósticos médicos; descobrir jazidas minerais; jogar xadrez; ensinar e aprender; entre muitas outras tarefas (CUNHA; RIBEIRO, 1987, p. 6).

A interface do usuário para o sistema especialista pode facilitar ou restringir o uso do sistema. Idealmente, esta interface deveria ser também modelada para o usuário e para o ambiente deste sistema (WARNER; SORENSON; BOUHADDOU, 1997, p. 12).

“Interação Homem Computador (IHC), é uma área de pesquisa preocupada com o design, avaliação e implementação de interatividade de sistemas computacionais para usuários humanos, sendo ela mais focada em fenômenos circundantes destes sistemas. IHC é associado com usabilidade e fatores humanos em um contexto do sistema.” (Scandurra, 2007).

Métodos de usabilidade são usados para estimar vastas tecnologias de informática médica (DANSKY, PALMER *et al.*, 2001).

Segundo (Nielsen, 1993), a usabilidade é contemplada em cinco atributos:

1. *learnability*: o sistema deve ser relativamente fácil de aprender;
2. *efficiency*: um usuário experiente pode atingir um alto nível de produtividade;
3. *memorability*: funcionalidade suportadas pelo sistema devem ser de fácil memorização após aprendidos pela primeira vez;

4. *errors*: o sistema deve ser desenhado para minimizar e detectar erros e apoiar na sua recuperação;
5. *satisfaction*: a experiência do usuário deveria ser subjetivamente satisfatória.

3 MATERIAL E MÉTODO

Quanto ao propósito à pesquisa é exploratória, quanto à natureza dos dados é quantitativa, quanto ao delineamento ou forma é não-experimental com um estudo de caso, quanto à obtenção dos dados ela é uma pesquisa de campo e de laboratório, quanto sua aplicabilidade é aplicada e quanto ao método é metodológica.

O planejamento para o início dos trabalhos foi dividido em cinco etapas conforme os objetivos específicos descritos no trabalho.

3.1 VERIFICAR O INTERESSE DOS USUÁRIOS DO SINPE[®] PARA COM O MÉTODO DE MINERAÇÃO E DADOS E SUA POSSÍVEL IMPLEMENTAÇÃO NESTE SOFTWARE

Em primeiro de janeiro do ano de 2009, foi elaborado um questionário pré-teste conforme Apêndice 7.1 , com o objetivo de levantar o interesse dos usuários para utilização de métodos de mineração de dados no SINPE[®], sua ferramenta de pesquisa. O critério de inclusão para esta pesquisa pré-teste de que os participantes fossem alunos do programa de Pós-Graduação em Clínica Cirúrgica da Universidade Federal do Paraná, da linha de pesquisa em Protocolos Eletrônicos e usuários do SINPE[®].

A avaliação deste questionário e de seus resultados ocorreu através da disciplina Seminário de Dissertação do programa de pós-graduação em Ciência Gestão e Tecnologia da Informação, ministradas pelas professoras Dra. Leilah Santiago Bufrem e Dra. Helena de Fátima Nunes Silva. Como contribuição desta mesma disciplina foi proposto um novo questionário conforme Apêndice 7.2 . Este questionário possui questões fechadas utilizando a escala *Likert* e uma questão parcialmente aberta.

A aplicação deste questionário ocorreu logo após sua proposta. O critério de inclusão de participantes deste novo questionário de que os entrevistados fossem usuários do SINPE[®], desconsiderando a relação de aluno regular do programa de

Pós-Graduação em Clínica Cirúrgica da Universidade Federal do Paraná, da linha de pesquisa em Protocolos Eletrônicos. O questionário disponível no Apêndice 7.2 foi encaminhado por email para dezenove usuários ativos do SINPE[®] no ano de 2009.

Para a coleta dos dados dos questionários utilizou-se as ferramentas do SINPE[®], software de pesquisa da área da saúde já abordado neste trabalho.

3.2 PESQUISAR NA LITERATURA OS MELHORES ALGORITMOS E TÉCNICAS DE MINERAÇÃO DE DADOS PARA UTILIZAÇÃO NA TAREFA PROPOSTA

Definiu-se que o SINPE[®] Analisador tivesse uso de bibliotecas ou classes compatíveis com sua plataforma, descartando assim a possibilidade de comunicação com outra ferramenta de Mineração de Dados, como o Weka. Caso a biblioteca ou classe encontrada não atendesse as necessidades para a implementação no aplicativo SINPE[®] Analisador seria desenvolvido manualmente o método de mineração.

Efetuiu-se busca exaustiva em ferramentas de pesquisa na *web* e em sites relacionados à Mineração de Dados a fim de localizar bibliotecas ou classes que tenham implementado os algoritmos propostos. Para tanto selecionou-se os algoritmos ID3 de classificação e o algoritmo Apriori de associação na pesquisa. Os algoritmos ID3 e Apriori foram levados em consideração devido ao conhecimento, funcionamento e manuseio deles por parte do desenvolvedor.

3.3 IMPLEMENTAR NO MÓDULO ANALISADOR DO SINPE[®] OS ALGORITMOS SELECIONADOS

A implementação dos algoritmos propostos no módulo SINPE[®] Analisador foi dividida em três fases. A primeira consistiu em documentar o diagrama de classe e a lista de funções para a tarefa proposta, conforme Apêndice 7.4. A segunda em preparar o SINPE[®] Analisador para a plataforma do Microsoft Visual Studio 2008

tendo em vista que esta é a ferramenta conhecida pelo desenvolvedor uma vez que a versão original deste módulo utilizava a plataforma SharpDevelop versão 1.1. Na terceira e última etapa foram desenvolvidos os protótipos de tela e estabelecido o local onde a mineração de dados ficaria disposta no módulo.

3.3.1 Preparo do SINPE[®] Analisador

Definiu-se como a Mineração de Dados ficaria disposta para o usuário. Optou-se em incorporá-la no SINPE[®] Analisador em vez de criar um aplicativo à parte, isto porque o SINPE[®] Analisador já é uma ferramenta para a visualização dos resultados de coletas.

O SINPE[®] Analisador foi desenvolvido na linguagem de programação CSharp (C#) para acesso ao banco de dados Microsoft Access. Sua IDE de desenvolvimento foi SharpDevelop versão 1.1 que utilizava a versão 1.0 do Microsoft Framework. Devido ao não conhecimento desta IDE, optou-se em utilizar o Microsoft Visual Studio 2008 com o Microsoft Framework versão 3.5.

Para importar o projeto do SINPE[®] Analisador para o Microsoft Visual Studio 2008 foi necessário:

- 1) exportar o projeto do SharpDevelop versão 1.1 para o Microsoft Visual Studio 2003;
- 2) exportar o projeto do Microsoft Visual Studio 2003 para o Microsoft Visual Studio 2008.

Após esta etapa abre o projeto no Microsoft Visual Studio 2008 e corrigiram-se todas as classes do projeto SINPE[®] Analisador que não possuía a importação das bibliotecas padrões utilizadas pelas classes da plataforma. Para isso inseriu-se as seguintes linhas em cada classe:

- using System.Drawing;
- using System.Drawing.Imaging;
- using System.Windows.Forms.

Foi compilado e testado a versão do SINPE[®] Analisador no Microsoft Visual Studio 2008. Não necessitando de nenhum outro ajuste, pois o aplicativo funcionou

perfeitamente como na versão desenvolvida no SharpDevelop, partiu-se para a pesquisa das bibliotecas e classes a serem utilizadas.

3.3.2 Interface de Mineração de Dados para o usuário

Inseriu-se de um botão que possibilita a mineração de dados a partir do SINPE[®] Analisador, representado pela Figura 3. Esta opção abrirá às funcionalidades inerentes à mineração de dados.

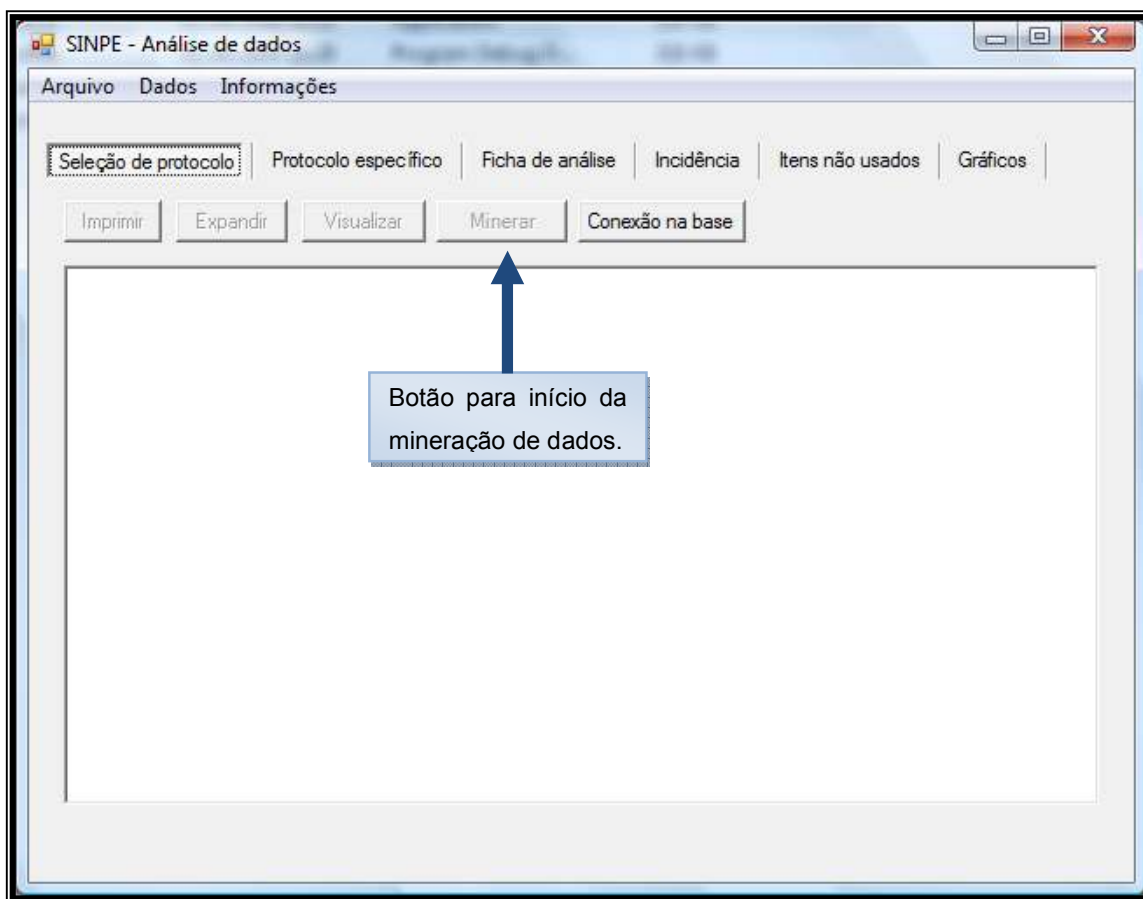


FIGURA 3 - SINPE[®] ANALISADOR COM A FUNÇÃO DE MINERAÇÃO DE DADOS

A Figura 4 representa a janela de boas vindas da mineração de dados do SINPE[®] Analisador. Ela apresenta as etapas que o usuário terá até chegar ao seu resultado. Esta janela também descreve qual é a função desta ferramenta.

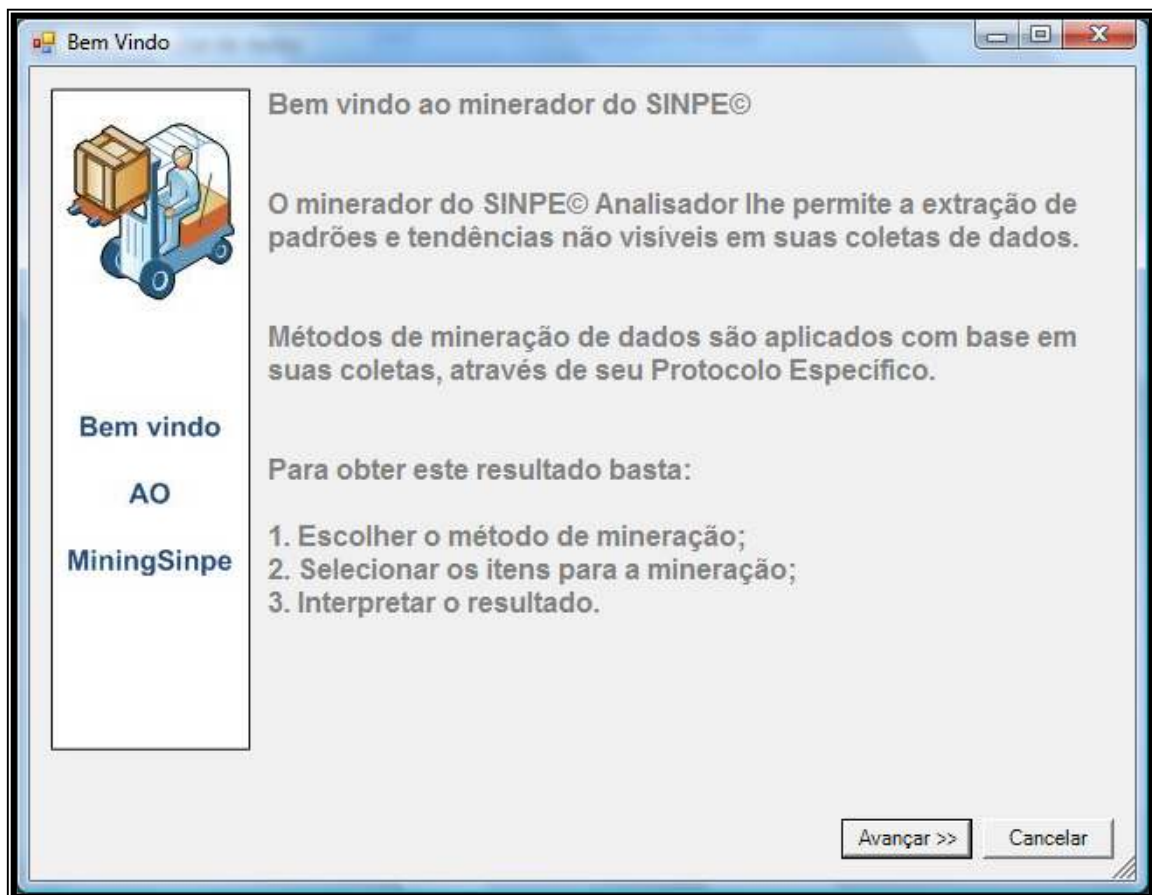


FIGURA 4 - JANELA DE APRESENTAÇÃO DA MINERAÇÃO DE DADOS

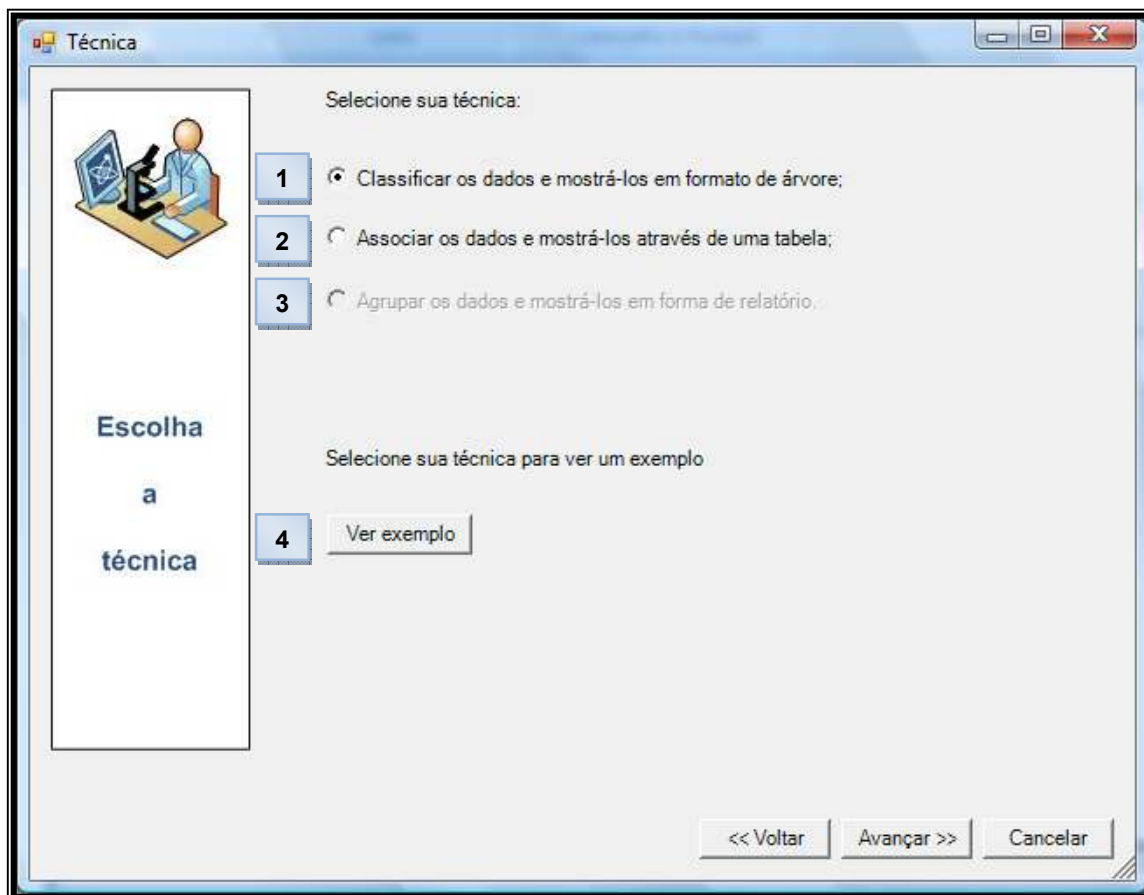


FIGURA 5 - JANELA DE ESCOLHA DO MÉTODO COM A OPÇÃO DE CLASSIFICAÇÃO

Na Figura 5 a opção "Classificar os dados e mostrá-los em formato de árvore" selecionada, indicada como 1 e clicando no botão "Ver exemplo" indicado como 4, abrirá a janela com a exemplificação deste método descrevendo suas funcionalidades, conforme Figura 6.

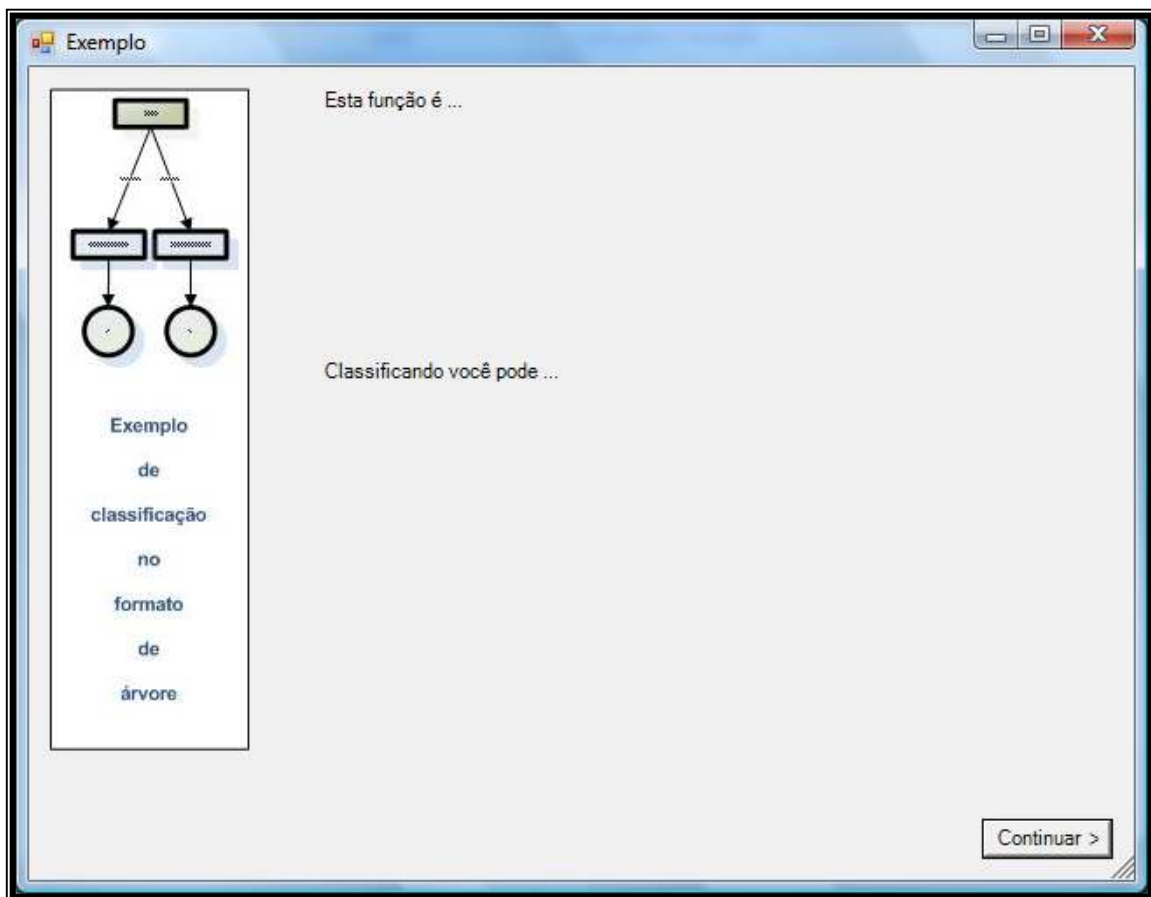


FIGURA 6 - EXEMPLO DO MÉTODO DE CLASSIFICAÇÃO

A mesma exemplificação do método ocorre para a opção “Associar os dados e mostrá-los através de uma tabela”, conforme a Figura 7 e Figura 8.

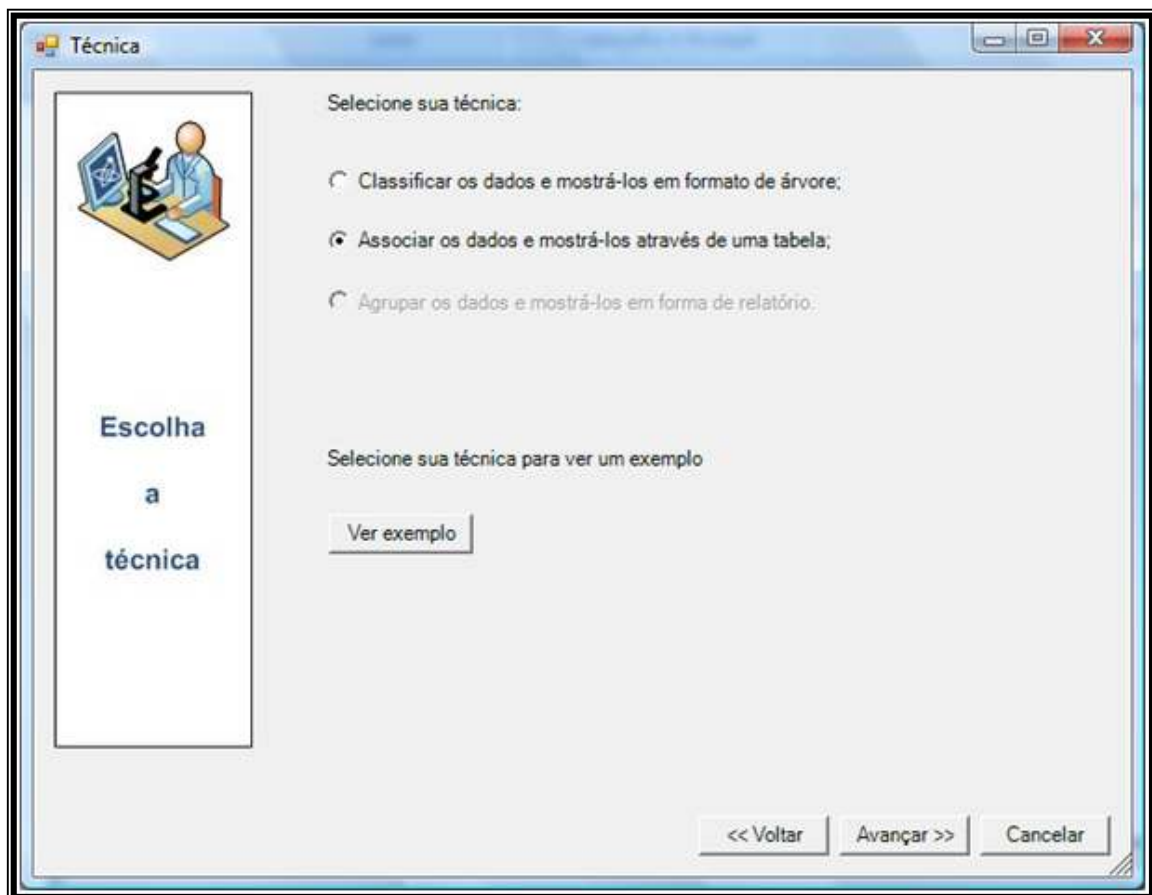


FIGURA 7 - JANELA DE ESCOLHA DO MÉTODO COM A OPÇÃO DE ASSOCIAÇÃO

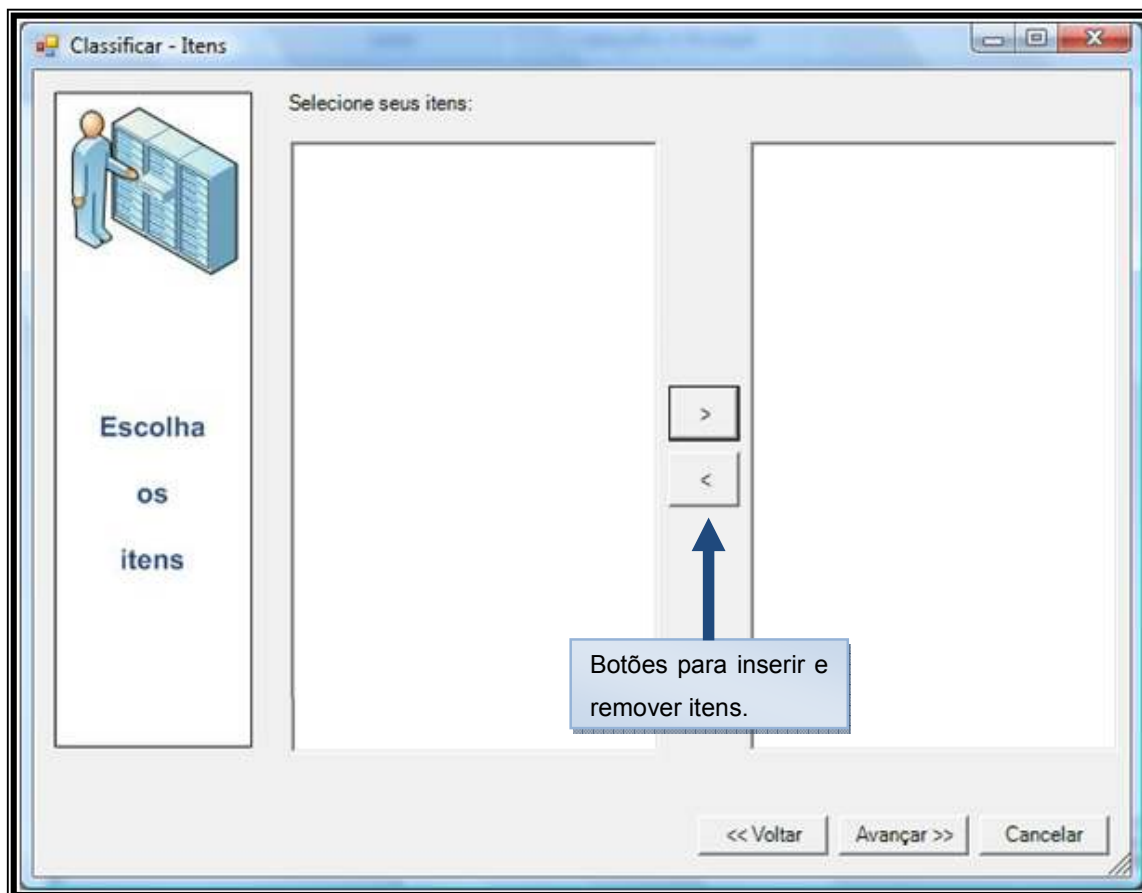


FIGURA 9 - JANELA COM A SELEÇÃO DOS ITENS

A Figura 10 representa a tela contendo os resultados do método ID3. Sendo que: 1 – são os registros utilizados para a mineração de dados; 2 – é a descrição do método utilizado; 3 – tempo que levou para obter o resultado; 4 – mostra a árvore de decisão em formato de texto.

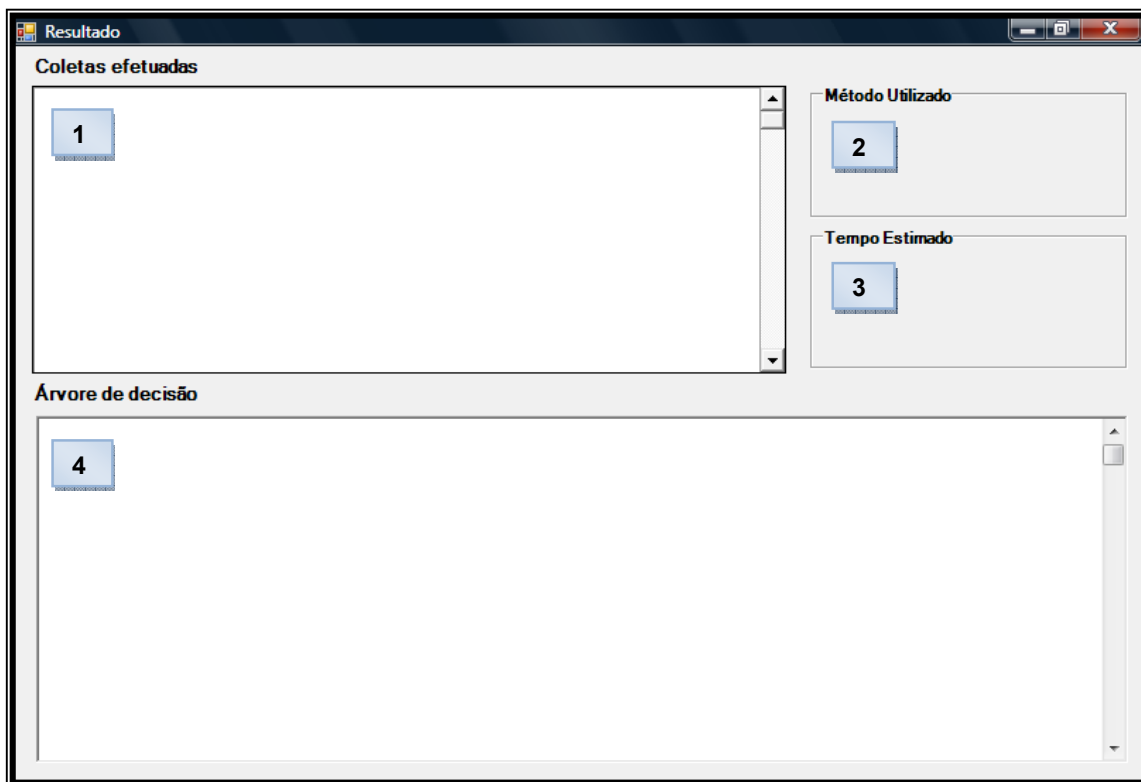


FIGURA 10 - RESULTADO DA MINERAÇÃO DO MÉTODO ID3

A Figura 11 representa a tela contendo os resultados do método ID3. Sendo que: 1 – são os registros utilizados para a mineração de dados; 2 – é a descrição do método utilizado; 3 – mostra as regras criadas pelo método. O botão “Alterar valores” dá a possibilidade do usuário alterar os valores de suporte e confiança para este método, quando acionado este botão mostra as informações referentes a FIGURA 12.

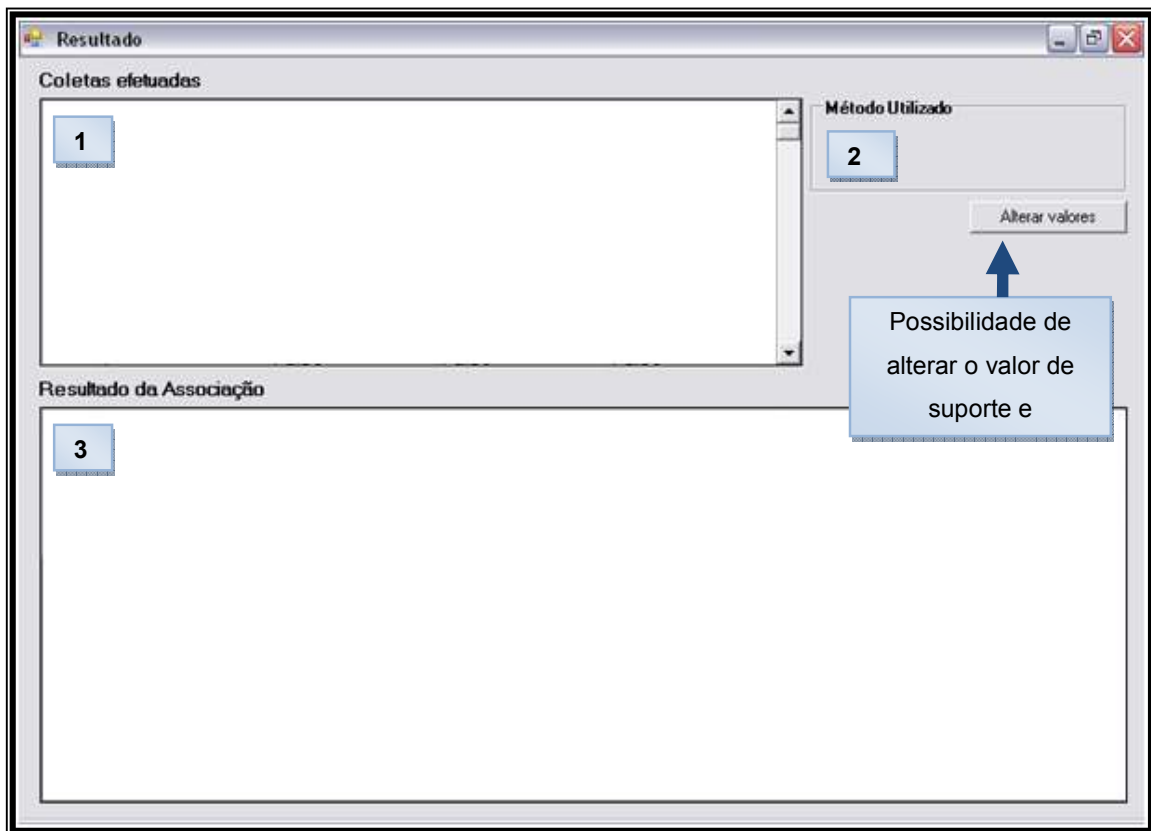


FIGURA 11 - RESULTADO DA MINERAÇÃO DO MÉTODO APRIORI

Valores Utilizados	
Suporte	<input type="text" value="0"/> %
Confiança	<input type="text" value="0"/> %
<input type="button" value="Analisar"/>	

FIGURA 12 - PARÂMETROS DO MÉTODO APRIORI

3.4 TESTAR ESTES ALGORITMOS COM PROFISSIONAIS DA SAÚDE E SEUS PROTOCOLOS ELETRÔNICOS

No final do mês de abril de 2009, foi realizada a mineração de dados a partir do Protocolo Mestre de doenças do cólon selecionando o Protocolo Específico de doenças do apêndice, havendo 66 coletas realizadas no Hospital de Clínicas da UFPR de forma prospectiva durante o período de outubro de 2007 a outubro de 2008, sendo 33 pacientes do sexo masculino e 33 do feminino. Este procedimento de teste contou com a participação e supervisão do aluno de doutorado Faruk Abrão Kalil Filho.

No início do mês de maio de 2009, foi realizada a mineração de dados a partir do Protocolo Mestre de nutrição enteral domiciliar selecionando o Protocolo Específico de atendimento nutricional domiciliar, havendo 111 coletas realizadas nas Unidades de Saúde da cidade de Curitiba de forma prospectiva durante o período de março à agosto de 2008, sendo 53 pacientes do sexo masculino e 58 do feminino. Este procedimento de teste contou com a participação e supervisão da aluna de doutorado Maria Eliana Madalozzo Schieferdecker do programa Clínica Cirúrgica da UFPR.

A seleção destes dois protocolos ocorreu devido à proximidade dos mesmos para com o projeto, além de suas bases de dados se encontrarem com bom número de coletas.

3.5 VALIDAR A INTERFACE PROPOSTA PARA A MINERAÇÃO DE DADOS ATRAVÉS DO MÓDULO ANALISADOR DO SINPE[®]

Para validar a interface proposta da mineração de dados, elaborou-se questionário conforme Apêndice 7.3 , com base nas questões abordadas por Nielsen (1993) sobre usabilidade de sistemas. A pesquisa contou com a participação de quatro usuários do software de pesquisas, sendo aplicada durante o período de 19 de outubro de 2009 a 26 de outubro de 2009. Para auxiliar na utilização da ferramenta de mineração de dados, os entrevistados utilizaram o Manual de utilização do SINPE[®] Analisador com o recurso de mineração de dados, conforme Apêndice 7.5

4 RESULTADOS

4.1 VERIFICAR O INTERESSE DOS USUÁRIOS DO SINPE[®] PARA COM O MÉTODO DE MINERAÇÃO E DADOS E SUA POSSÍVEL IMPLEMENTAÇÃO NESTE SOFTWARE

Contando com a participação de seis alunos da Pós-Graduação em Clínica Cirúrgica, obteve-se os seguintes resultados a partir do questionário do estudo exploratório disposto no Apêndice 7.1 : três entrevistados estão cursando o doutorado e três mestrado. Todos possuíam conhecimentos em Informática Médica; apenas um aluno possuía conhecimento em Mineração de Dados. Todos entrevistados utilizavam o SINPE[®] Analisador como ferramenta para avaliação de seus resultados; um destes também utilizava o Microsoft Excel como ferramenta complementar. Todos os participantes da pesquisa gostariam de ter uma ferramenta de Mineração de Dados embutida no SINPE[®].

A pesquisa definitiva contou com a participação de 13 usuários entre eles médicos, enfermeiras, nutricionistas e fisioterapeutas. Estes entrevistados utilizavam o software em suas pesquisas de mestrado, doutorado ou pós-doutorado.

Dos pesquisadores entrevistados, seis possuíam sua pesquisa na área de medicina, quatro em fisioterapia, dois em gestão em saúde e um na área de nutrição. Entre os entrevistados um deles utilizou outro software complementar para a coleta de dados de sua pesquisa.

Todos entrevistados utilizavam o SINPE[®] Analisador como ferramenta de análise de dados de sua pesquisa, dois o Microsoft Excel e um o software Bioestat 5.0.

O Gráfico 1 mostra o conhecimento dos entrevistados perante a área de informática em saúde.

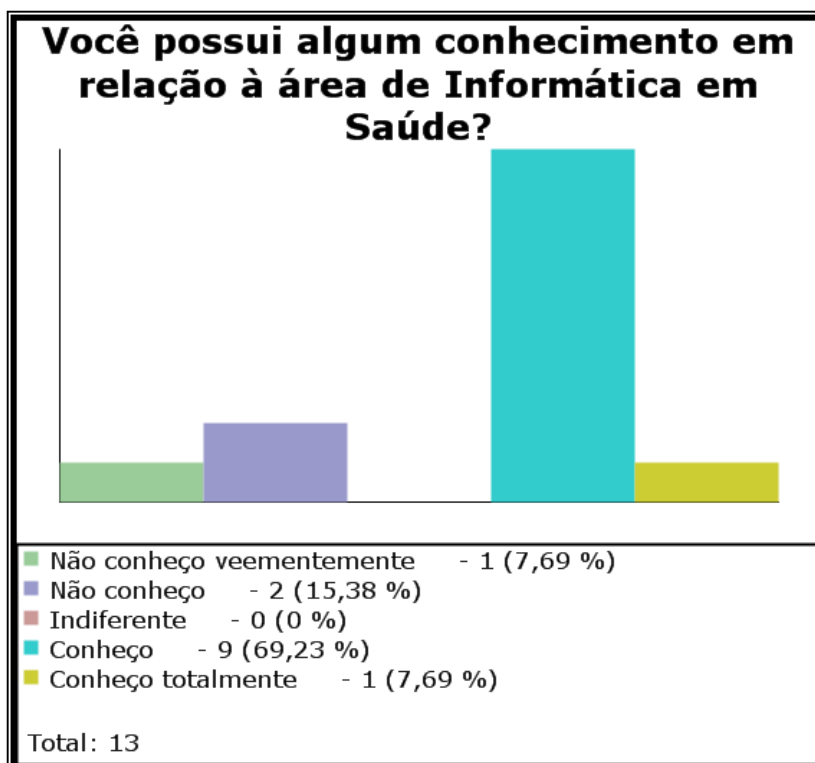


GRÁFICO 1 - CONHECIMENTO NA ÁREA DE INFORMÁTICA EM SAÚDE

A pesquisa também abordou a opinião por parte dos entrevistados em relação a interface do software pesquisado. Dois participantes apontaram que a interface do software é excelente, nove muito bom e dois bom.

Para visualizar o conhecimento do público para com a tarefa proposta, perguntou-se seus conhecimentos para com os métodos de mineração de dados, divididos em: métodos de classificação, associação e agrupamento. Dos entrevistados sete conhece métodos de classificação, cinco não conhece e um não conhece veementemente, sendo que a mesma distribuição ocorreu para o método de associação. A distribuição de resposta obtida pelo método de agrupamento foi diferente, onde quatro conhece, sete não conhece e dois não conhece veementemente.

Um dos entrevistados não utilizaria um software de mineração de dados em sua pesquisa ele conhece a área de informática em saúde bem como os métodos de mineração de dados; sua justificativa que a ferramenta não é objeto de sua pesquisa. Doze participantes utilizariam um software de mineração de dados em sua pesquisa, sendo que seis relataram excelência para um programa com esta função, cinco muito bom e um bom.

O Gráfico 2 mostra o interesse atribuído por parte dos usuários em uma ferramenta de mineração de dados utilizada a partir do SINPE[©].

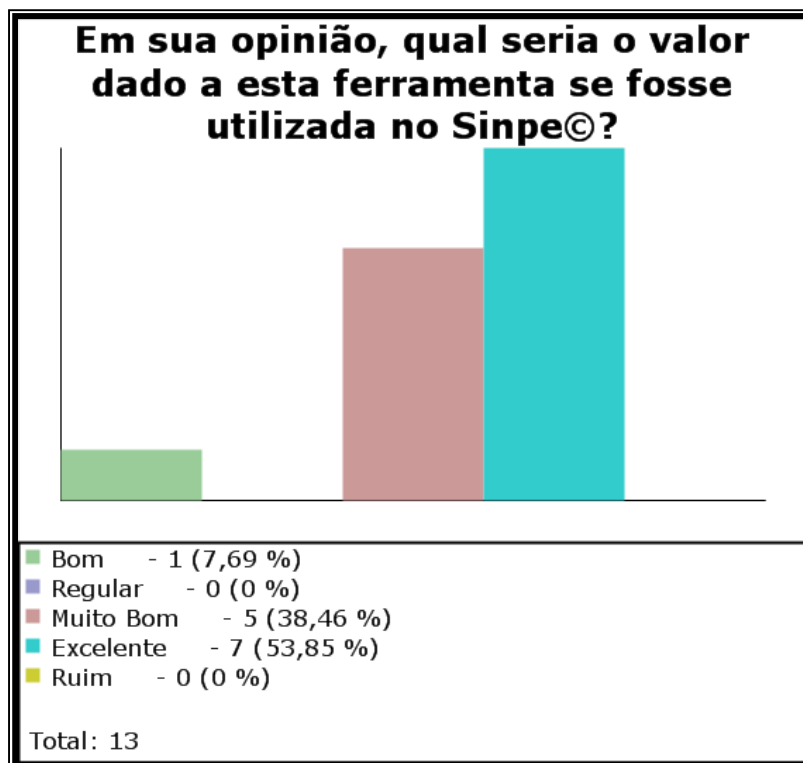


GRÁFICO 2 - OPINIÃO DOS PARTICIPANTES EM TER UMA FERRAMENTA DE MINERAÇÃO DE DADOS NO SINPE

Após levantamento dos dados realizou-se um cruzamento das informações perante os entrevistados que conhece os métodos de classificação e associação. O método de agrupamento foi descartado nesta etapa, pois ele não foi implementado na pesquisa.

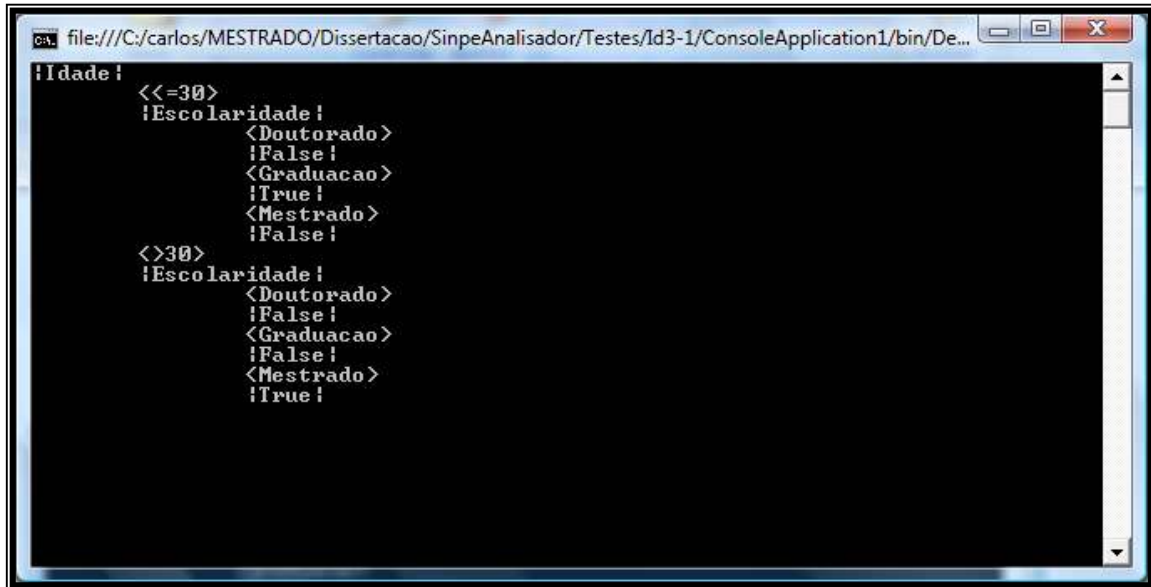
Neste cruzamento de dados obteve-se que sete entrevistados conhece métodos de classificação e de associação. Entre eles um não utilizaria a mineração de dados em sua pesquisa; seis utilizariam sendo que quatro excelente, dois muito bom e um bom. Considerando este mesmo grupo que conhece os métodos três relataram excelente a utilização de métodos de mineração de dados a partir do SINPE[©], três muito bom e um bom.

4.2 PESQUISAR NA LITERATURA OS MELHORES ALGORITMOS E TÉCNICAS DE MINERAÇÃO DE DADOS PARA UTILIZAÇÃO NA TAREFA PROPOSTA

4.2.1 Método Id3

Durante a pesquisa foi localizado uma classe contendo o algoritmo ID3, desenvolvido na mesma linguagem de programação do SINPE Analisador.

Antes de sua implementação no SINPE[©] Analisador efetuou-se um teste a fim de validar o funcionamento da classe localizada. Partindo de uma mesma base de dados, executou-se a classe através do Microsoft Visual Studio 2008, e foi obtido o resultado apresentado na Figura 13.



```
file:///C:/carlos/MESTRADO/Dissertacao/SinpeAnalizador/Testes/Id3-1/ConsoleApplication1/bin/De...
!Idade!
<=<=30>
!Escolaridade!
<Doutorado>
!False!
<Graduacao>
!True!
<Mestrado>
!False!
<>30>
!Escolaridade!
<Doutorado>
!False!
<Graduacao>
!False!
<Mestrado>
!True!
```

FIGURA 13 - RESULTADO DA CLASSE COM O ID3

Após esta etapa, adaptou-se a mesma base de dados para o arquivo de leitura do aplicativo Weka, um arquivo (.arff). Este software gratuito de Mineração de Dados está disponível em <<http://www.cs.waikato.ac.nz/ml/weka/>>. O resultado obtido está disposto na Figura 14. Tal aplicativo foi utilizado devido sua disponibilidade no laboratório e em especial, por ser bastante estável e confiável.

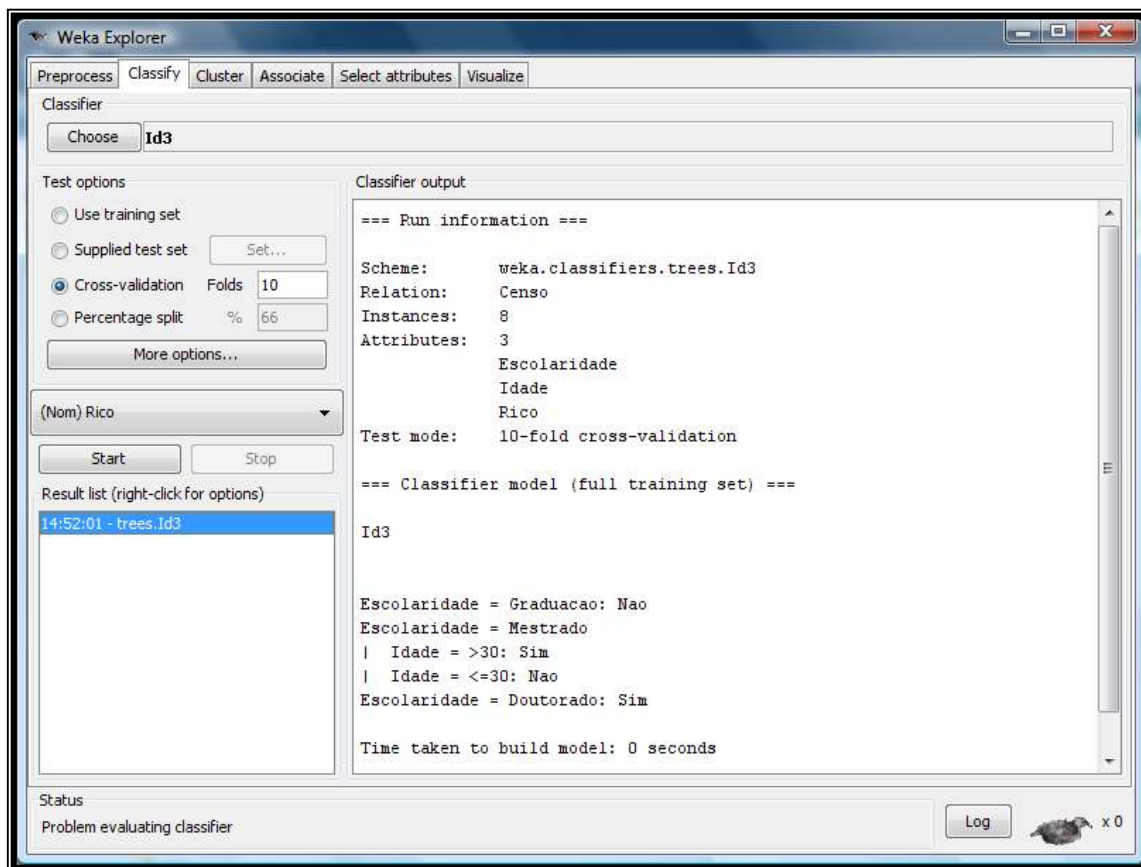


FIGURA 14 - RESULTADO DO WEKA COM O ID3

O resultado gerado por esta classe é equivalente ao gerado pelo Weka utilizando o mesmo método com os mesmos parâmetros. Confrontados os resultados, concluiu-se o funcionamento correto da implementação realizada.

4.2.2 Método Apriori

A única classe compatível e implementada utilizando o método Apriori possuía muitos erros de programação. Para testar esta classe era necessário ter um arquivo no formato “.csv”, que representa a base de dados com o conteúdo a ser minerado; mas o desenvolvedor desta classe não disponibilizou um modelo deste arquivo dificultando ainda mais os testes.

Devido a estes empecilhos foi decidido programar o método manualmente no SINPE[®] Analisador.

4.3 IMPLEMENTAR NO MÓDULO ANALISADOR DO SINPE[®] OS ALGORITMOS SELECIONADOS

4.3.1 Algoritmo Id3

Localizada a classe compatível com o SINPE[®] Analisador, ela foi implantada nele, foram alterados os seguintes itens do código fonte com alterações pontuais relativas aos parâmetros para a mineração de dados.

1. na função *getBestAttribute*: esta função retorna os atributos que possuem o melhor ganho de informação, a fim de montar a árvore de decisão; quando não existia mais atributos válidos, esta função retornava *null*, porém ela não criava uma instância deste valor para a função *Attribute*, gerando assim um erro; esta função foi corrigida com a chamada da instância dos valores com *null*;
2. outra alteração ocorreu na função *calcEntropy*: função que calcula a entropia do conjunto; a divisão dos valores positivos e negativos pelo total de registros não possuía a conversão para o tipo de variável em questão, ambos do tipo *double*; esta conversão foi implantada.

Estes itens do código fonte não foram identificados no teste inicial pois a base de dados testada possuía poucos registros, sendo que estes erros foram apresentados apenas na fase de implementação da classe para o SINPE Analisador.

4.3.2 Algoritmo Apriori

Como não foi localizada uma classe ou biblioteca contendo o método Apriori de mineração de dados, compatível com o SINPE[®] Analisador, este método foi

implementado pelo autor deste trabalho, levando em consideração apenas os resultados positivos para gerar as regras de associação.

4.4 TESTAR ESTES ALGORITMOS COM PROFISSIONAIS DA SAÚDE E SEUS PROTOCOLOS ELETRÔNICOS

4.4.1 Utilizando o método ID3 a partir do protocolo da Cirurgia do Aparelho Digestivo

A Figura 15 mostra os itens selecionados a partir do Protocolo Específico de Doenças do Apêndice na caixa à direita.

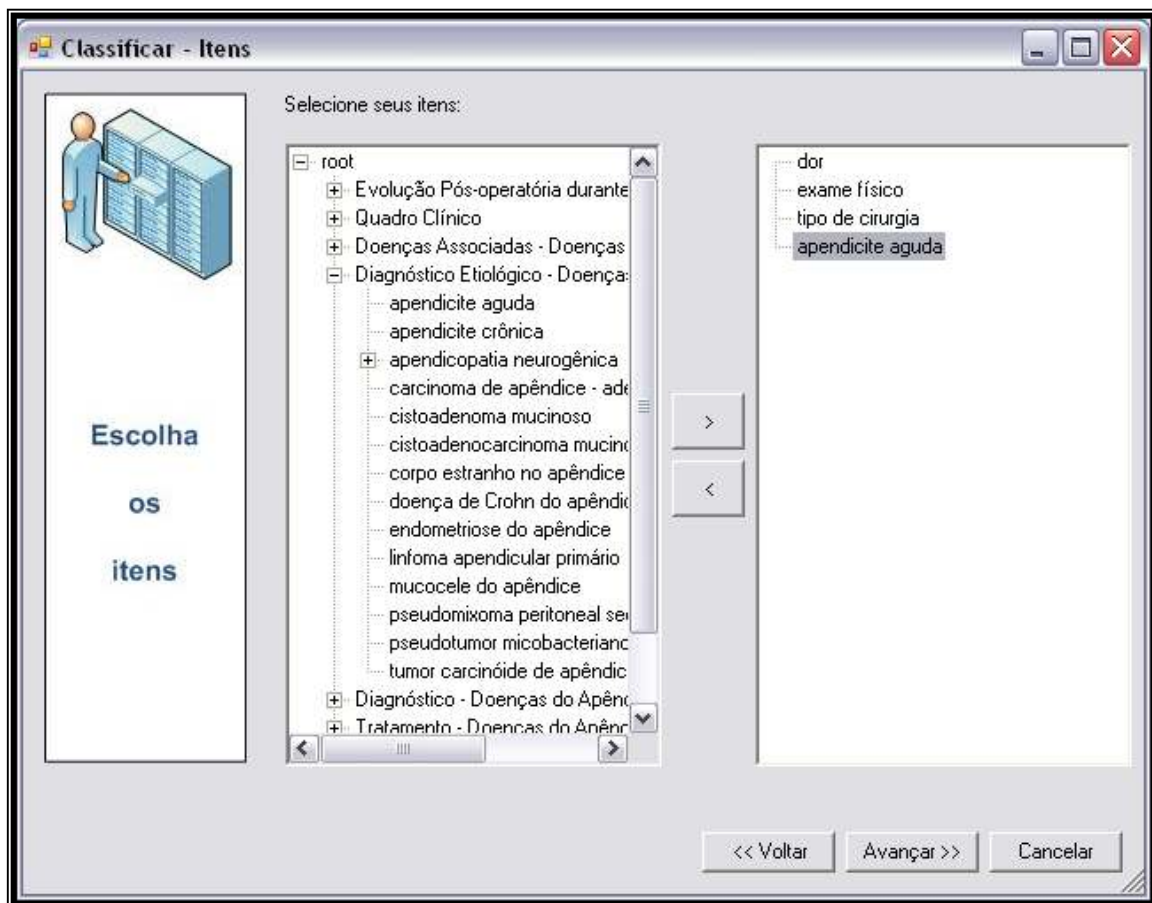


FIGURA 15 - ITENS SELECIONADOS DO PROTOCOLO DE DOENÇAS DO APÊNDICE

O resultado obtido através da seleção dos itens deste Protocolo Específico é representado pela FIGURA 16. Este procedimento foi realizado com o auxílio do aluno de doutorado em Clínica Cirúrgica da UFPR Faruk Abrão Kalil Filho.

Segundo este pesquisador o resultado desta mineração condiz com a realidade apresentada nas coletas clínicas realizadas no serviço de Cirurgia do Aparelho Digestivo do hospital de clínicas da UFPR.

The screenshot shows a window titled "Resultado" with the following content:

Coletas efetuadas

Atributo1	Atributo2	Atributo3	Atributo4
abdominal	Não possui coletas	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal de Blumberg	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal Rovsing	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal Rovsing	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal de Blumberg	apendicectomia	<input checked="" type="checkbox"/>
abdominal	Não possui coletas	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal de Blumberg	apendicectomia	<input checked="" type="checkbox"/>
abdominal	Não possui coletas	apendicectomia	<input checked="" type="checkbox"/>

Método Utilizado
Classificação com árvore de decisão
Algoritmo ID3

Tempo Estimado
00:00:13

Árvore de decisão

```

[exame físico]
  <Não possui coletas>
    [tipo de cirurgia]
      <apendicectomia>
        [dor]
          <abdominal>
            [True]
          <plastrão/ massa palpável em fossa ilíaca direita>
            [tipo de cirurgia]
              <apendicectomia>
                [dor]
                  <abdominal>
                    [True]
                  <sinal de Blumberg>
                    [tipo de cirurgia]
                      <apendicectomia>
                        [dor]
                          <abdominal>
                            [True]
                          <sinal Rovsing>
                            [tipo de cirurgia]
                              <apendicectomia>
                                [dor]
                                  <abdominal>
                                    [True]

```

FIGURA 16 - RESULTADO DO MÉTODO ID3 PARA O PROTOCOLO DE DOENÇAS DO APÊNDICE

4.4.2 Utilizando o método ID3 a partir do protocolo de Nutrição

A Figura 17 mostra os itens selecionados a partir do Protocolo Específico de Atendimento Nutricional Domiciliar na caixa à direita.

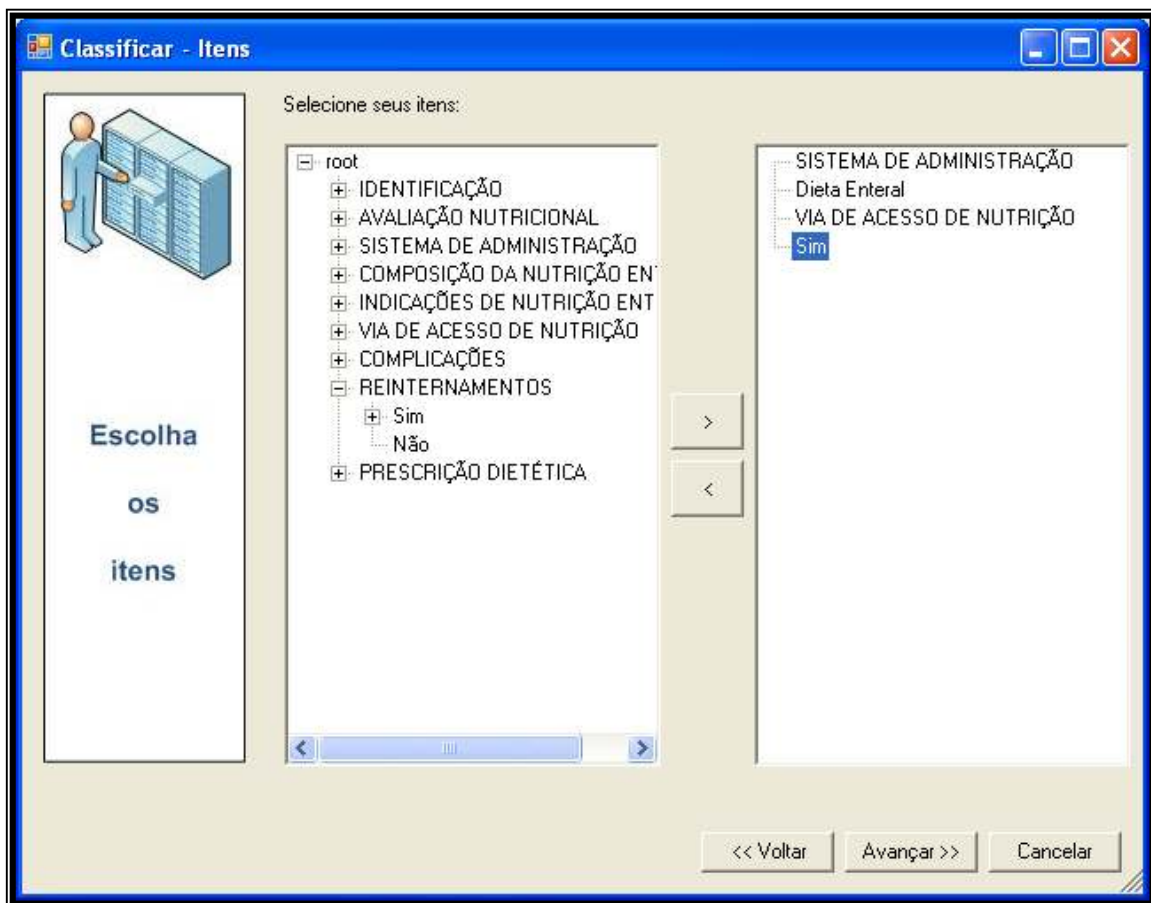


FIGURA 17 - ITENS SELECIONADOS DO PROTOCOLO DE NUTRIÇÃO DOMICILIAR

A Figura 18 mostra o resultado desta mineração de dados com base nos itens selecionados na Figura 17.

4.4.3 Utilizando o método Apriori a partir do protocolo da Cirurgia do Aparelho Digestivo

Para análise deste método, foi utilizado o mesmo banco de dados já abordado no método ID3, contendo o mesmo número de coletas e o mesmo Protocolo Específico de Doenças do Apêndice.

Os itens selecionados foram alterados, pois este método no SINPE permite a seleção de qualquer item do protocolo.

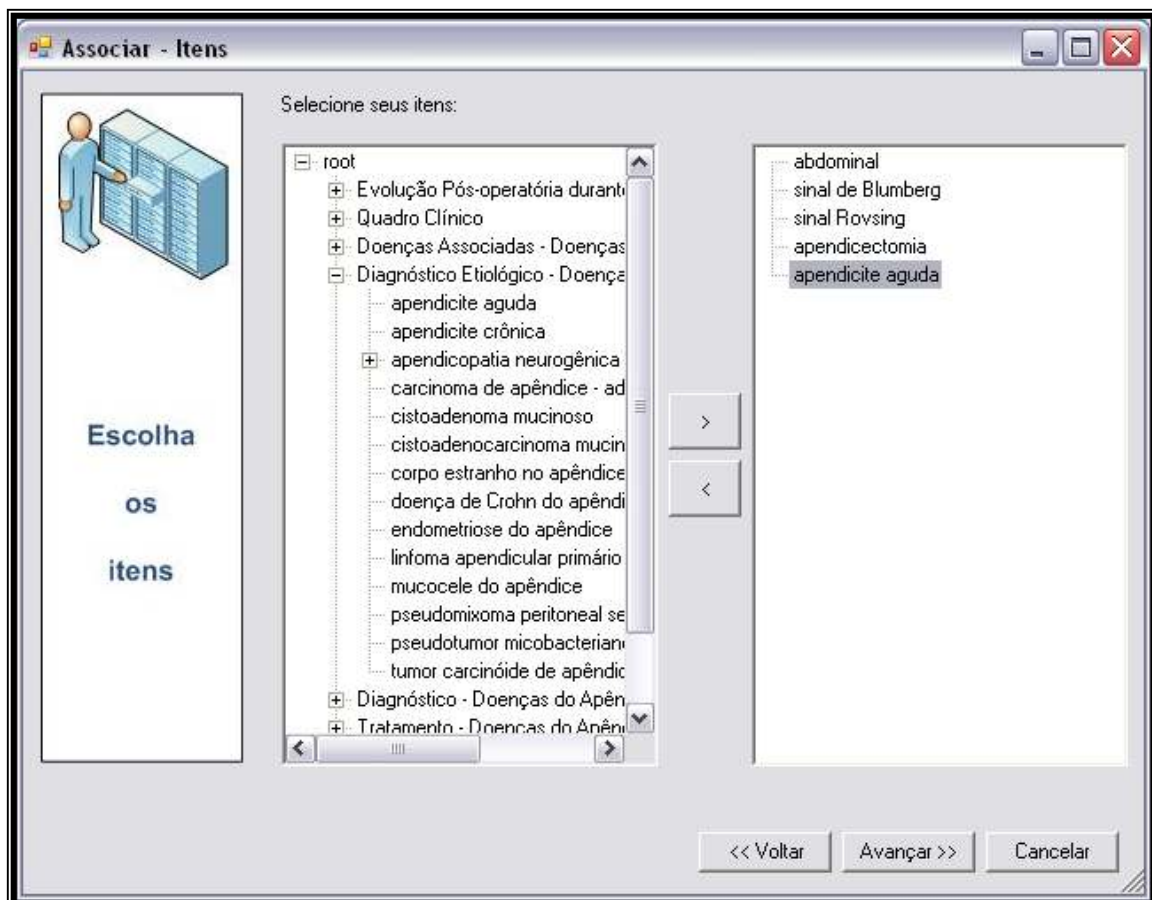


FIGURA 20 - ITENS SELECIONADOS DO PROTOCOLO DE DOENÇAS DO APÊNDICE

O resultado obtido através da seleção dos itens deste Protocolo Específico é representado pela FIGURA 21. Este procedimento foi realizado com o auxílio do aluno de doutorado em Clínica Cirúrgica da UFPR Faruk Abrão Kalil Filho.

Resultado

Coletas efetuadas

Id	1	2	3	4	5
9	True	True	False	True	True
10	True	True	False	True	True
11	True	False	False	True	True
12	True	False	False	True	True
13	True	True	False	True	True
14	True	True	False	True	True
15	True	True	False	True	True
16	True	True	False	True	True

Método Utilizado
Regras de Associação
Algoritmo Apriori

Valores Utilizados
Suporte: 10 %
Confiança: 50 %
Analisar

Resultado da Associação

Valor	Suporte em %	Confiança em %
abdominal está para apendicectomia	100	100
sinal de Blumberg está para abdominal	60,606060606060609	100
sinal de Blumberg está para apendicectomia	60,606060606060609	100
sinal Rovsing está para abdominal	10,606060606060606	100
sinal Rovsing está para apendicectomia	10,606060606060606	100
apendicectomia está para abdominal	100	100
apendicite aguda está para abdominal	87,878787878787875	100
apendicite aguda está para apendicectomia	87,878787878787875	100
sinal de Blumberg está para apendicite aguda	60,606060606060609	90
abdominal está para apendicite aguda	100	87,878787878...
apendicectomia está para apendicite aguda	100	87,878787878...
sinal Rovsing está para apendicite aguda	10,606060606060606	71,428571428...
apendicite aguda está para sinal de Blumberg	87,878787878787875	62,068965517...
abdominal está para sinal de Blumberg	100	60,606060606...
apendicectomia está para sinal de Blumberg	100	60,606060606...
sinal Rovsing está para sinal de Blumberg	10,606060606060606	57,142857142...

FIGURA 21 - RESULTADO DO MÉTODO APRIORI PARA O PROTOCOLO DE DOENÇAS DO APÊNDICE

4.4.4 Utilizando o método Apriori a partir do protocolo de Nutrição

Para análise deste método, foi utilizado o mesmo banco de dados já abordado no método ID3, contendo o mesmo número de coletas e o mesmo Protocolo Específico de Atendimento Nutricional Domiciliar, sendo que os itens

selecionados para este método foram alterados (Figura 22). Os valores selecionados referentes ao Índice de Massa Corpórea são relativos a idosos do sexo feminino.

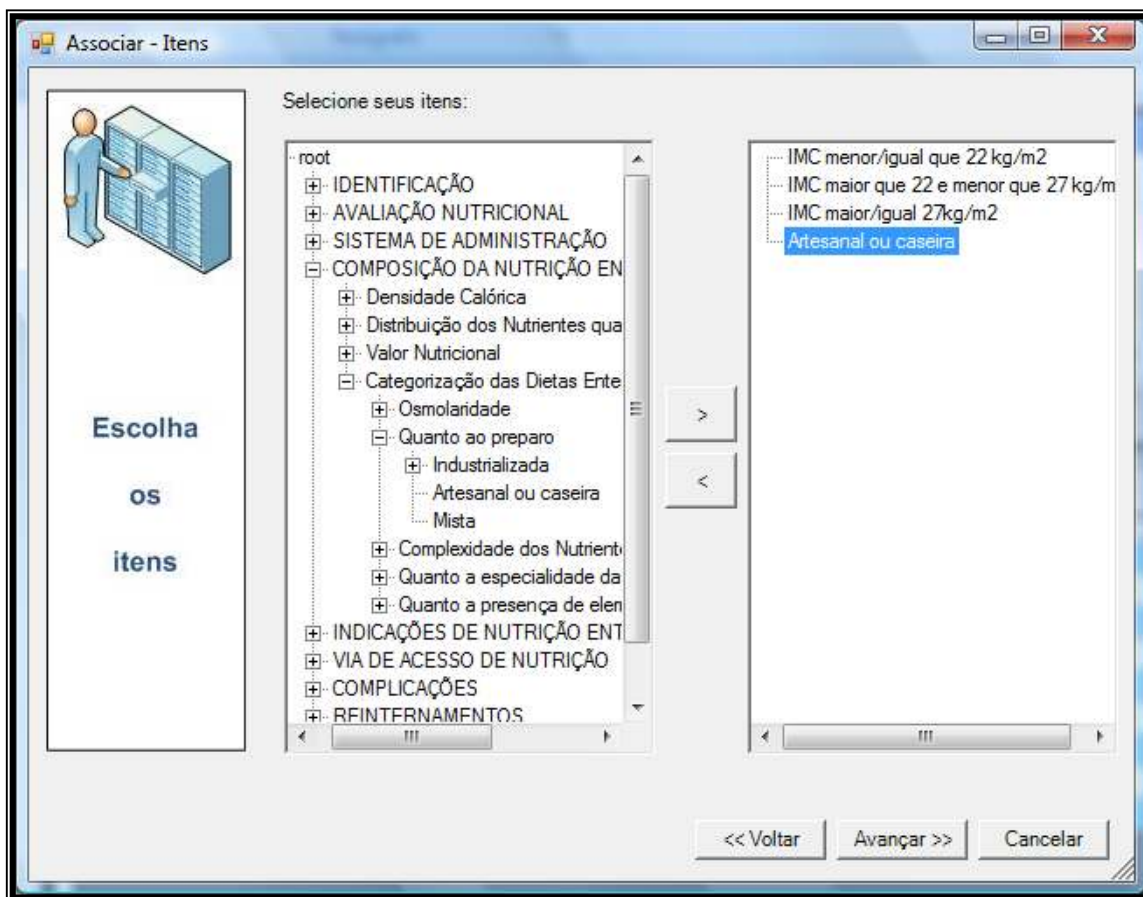


FIGURA 22 - ITENS SELECIONADOS DO PROTOCOLO DE NUTRIÇÃO DOMICILIAR

O resultado obtido através da seleção dos itens deste Protocolo Específico é representado pela Figura 23. Este procedimento foi realizado com o auxílio da mesma pesquisadora do método anterior. Para este método foram atribuídos 10% de suporte e 50% de confiança.

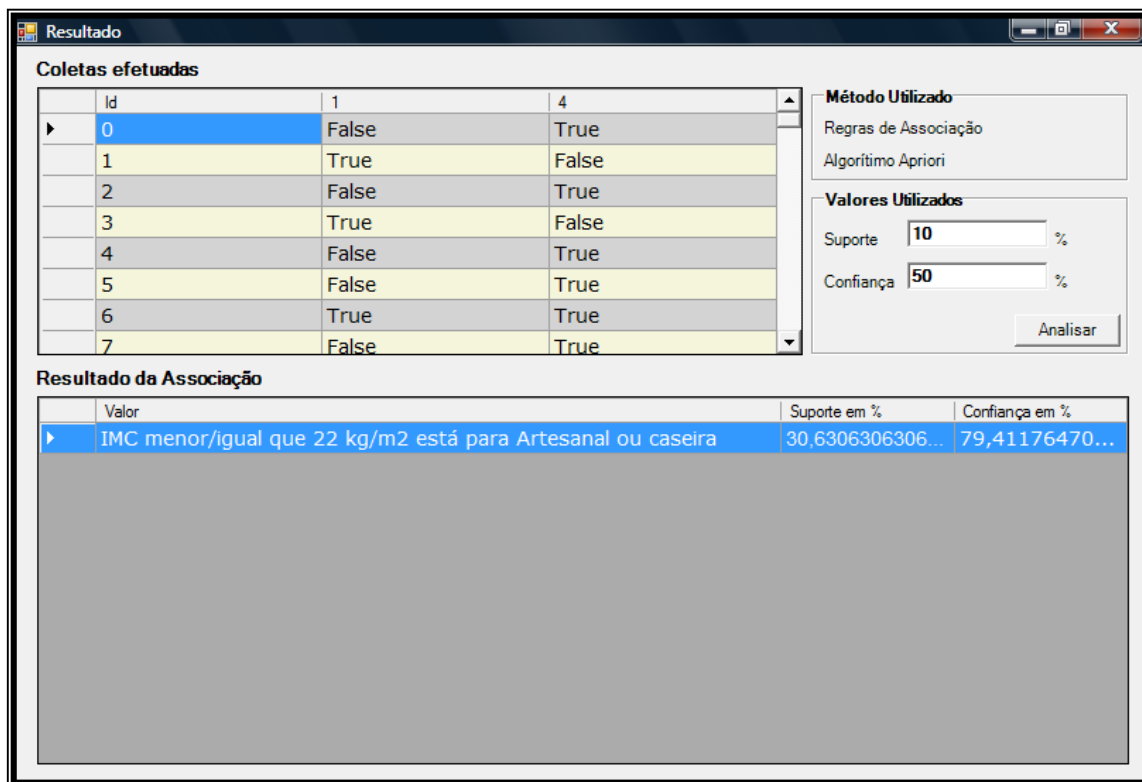


FIGURA 23 - RESULTADO DO MÉTODO APRIORI PARA O PROTOCOLO DE NUTRIÇÃO DOMICILIAR

4.5 VALIDAR A INTERFACE PROPOSTA PARA A MINERAÇÃO DE DADOS ATRAVÉS DO MÓDULO ANALISADOR DO SINPE[®]

A pesquisa contou com a participação de cinco entrevistados, onde todos participantes relatam que o módulo Mineração de Dados do SINPE[®] é fácil de aprender. Dois que o software é muito produtivo e três que ele é produtivo.

Entre os entrevistados dois apontavam que utilizariam o software novamente sem a necessidade de auxílio de especialistas; outros três apontavam que conseguem utilizar novamente. Todos os entrevistados relataram que não foi apresentado nenhum erro de aplicativo com o software em questão. Todos também relataram que ele é fácil de aprender.

Sobre a interface do novo módulo, o Gráfico 3 mostra o nível de satisfação perante os entrevistados para com o método de classificação.

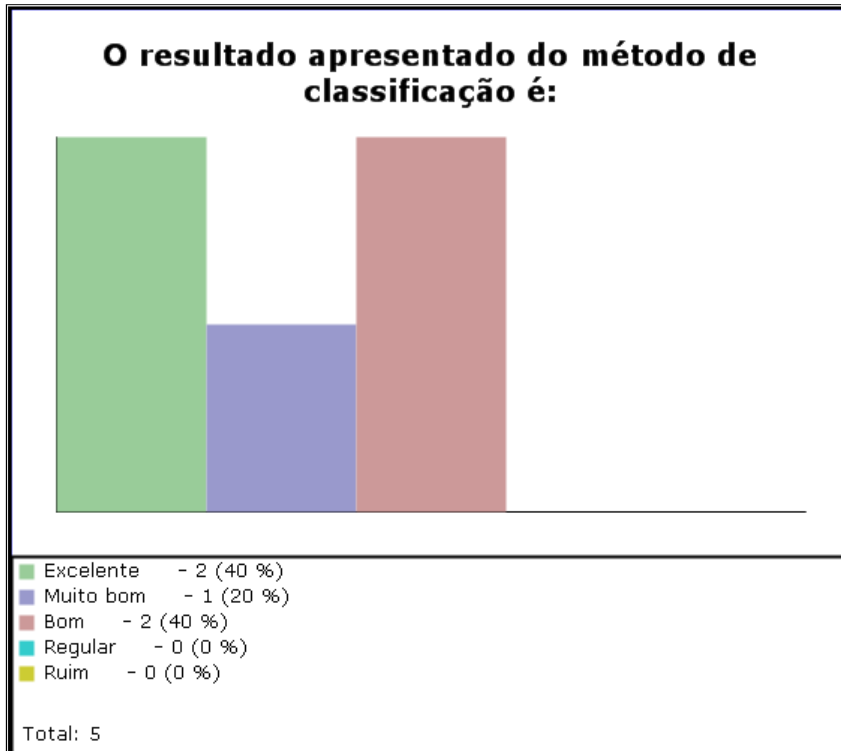


GRÁFICO 3 - NÍVEL DE SATISFAÇÃO DOS ENTREVISTADOS PARA COM O MÉTODO DE CLASSIFICAÇÃO

O Gráfico 4 mostra o nível de satisfação perante os entrevistados para com o método de associação.

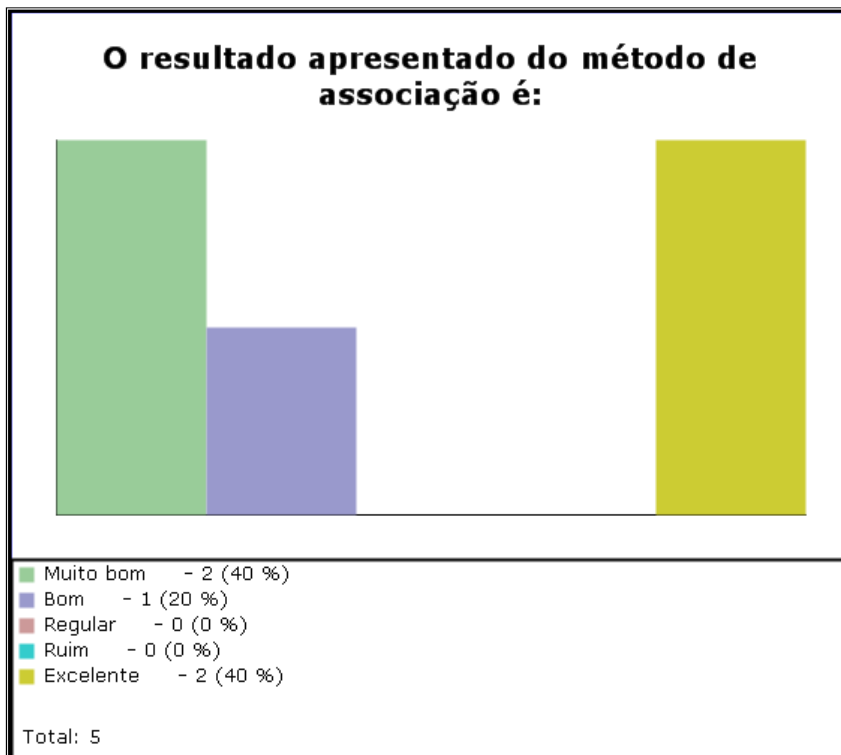


GRÁFICO 4 - NÍVEL DE SATISFAÇÃO DOS ENTREVISTADOS PARA COM O MÉTODO DE ASSOCIAÇÃO

Perante os entrevistados os exemplos apresentados pelo software são “excelentes” para dois e “bom” para três entrevistados. A janela de seleção de itens é “muito bom” para dois entrevistados e “excelente” para três.

5 DISCUSSÃO

Informática médica (*Medical Informatics*) é o ramo da ciência preocupado com o uso de computadores, comunicação e tecnologia para adquirir, armazenar, analisar, comunicar e mostrar informações e conhecimentos médicos para facilitar o entendimento e melhorar a precisão e a confiabilidade de fabricação de decisão (WARNER, SORENSON *et al.*, 1997).

A primeira aplicação prática e relevante da computação para a medicina, desenvolvida por Herman Hollerith e seu sistema de processamento de dados utilizando cartão perfurado, utilizado no censo de 1890 nos Estados Unidos. O método foi adaptado para estudos epidemiológicos e pesquisas públicas em saúde, iniciando a era da tecnologia eletromecânica do cartão perfurado, que foi muito adotada durante as décadas de 1920 e 1930 (SHORTLIFFE e CIMINO, 2006, p. 24).

O termo Informática médica foi originalmente introduzido na Europa. Este termo apareceu no ano de 2000 sendo muito utilizado nos Estados Unidos, embora algumas pessoas não gostem desse nome, tendo em vista que a palavra “médica” foca em profissionais médicos e o termo abrange outros profissionais da saúde. Desta forma o termo informática em saúde (*health informatics* ou *health care informatics*) foi ganhando popularidade, embora tenha a desvantagem de excluir aplicações em biologia. A relação entre informática médica e bioinformática tornou-se incerta e gradualmente o termo informática médica da lugar a *biomedical informatics*. Este termo é um campo da ciência que lida com *biomedical information*, dados e conhecimento, seu armazenamento, recuperação, solução de problemas e o auxílio à decisão (SHORTLIFFE e CIMINO, 2006).

Segundo Shortliffe e Barnett (2001, p. 5), “O registro médico em papel é totalmente inadequado para o encontro das necessidades da medicina moderna”. A informatização na coleta de dados tem importância fundamental na melhoria da qualidade das informações, e também na possibilidade de interligação de computadores e instituições, o que possibilitaria a ampliação da coleta e armazenamento de dados de uma maneira mais confiável e multicêntrica. A informação clínica utilizada na assistência, ou seja, no contato com o paciente, faz do registro eletrônico a maior fonte de informações corretas, desde que adequadamente coletado, fornecendo também o controle da informação clínica,

potencializando a existência de dados apurados quando coletados na forma prospectiva. O protocolo eletrônico facilita o acesso à informações do paciente, de maneira segura, rápida e eficiente (KALIL FILHO, 2008).

Devido à escassez de recursos, pesquisadores e cientistas brasileiros, em saúde ou em outras áreas, carecem de condições mínimas para a sua atuação, como: facilidade de acesso à informação, infra-estrutura adequada, possibilidade de comunicação com os pares em caráter permanente e sistemático e oportunidades para a divulgação ampla dos seus trabalhos. A divulgação de resultados é etapa intrínseca e não complementar à pesquisa científica (TARGINO, 2009).

O uso de computadores em informática em saúde gerou uma explosão na qualidade de dados eletrônicos (CHEN, FULLER *et al.*, 2005).

Devido a este e outros fatores desenvolveu-se a ferramenta de mineração de dados no SINPE[®] Analisador, possibilitando o usuário pesquisador complementar seus resultados obtidos através de sua pesquisa, pesquisa esta que também conta com as funcionalidades de estatística descritiva através de gráficos disponíveis no módulo do SINPE[®] Analisador.

Tema central em informática médica é o uso de dados para a descoberta de conhecimento, processo conhecido como mineração de dados. O propósito deste termo é identificar significantes padrões em dados que não poderiam ser detectados de outras formas (CHEN *et al.*, 2005, p.153-4).

Conceitos de Mineração de Dados e Estatística, muitas vezes se sobrepõem, entretanto mantêm diferenças que muitas vezes podem não ser muito claras. O termo Mineração de Dados em estatística é utilizado há muito tempo referindo-se à análise de dados sem formular claramente uma hipótese (BRASIL, 2008).

Métodos de estatística analítica poderão ser utilizados em conjunto com a ferramenta de mineração de dados desenvolvida. Pretende-se também implementar métodos de estatística analítica nesta ferramenta futuramente.

Durante o desenvolvimento notou-se a necessidade de alterar o nome dos métodos implementados. Pensando na interpretação por parte dos usuários deste software, alterou-se a descrição dos métodos de mineração de dados. O método ID3 está implementado na opção “Classificar os dados e mostrá-los em formato de árvore”, indicado como 1 na Figura 5. O método Apriori é representado pela opção “Associar os dados e mostrá-los através de uma tabela”, indicado como 2. E por fim

a opção “Agrupar os dados e mostrá-los em formato de relatório”, não implementada, que representa o método K-Means indicado como 3.

Após implementação testou-se com usuários do software a ferramenta proposta. Para a pesquisadora Maria Eliana Madalozzo Schieferdecker, o resultado obtido pelo método ID3 certamente contribui para a conduta e a prescrição clínica. Ela destaca a opção representada na FIGURA 19, quando a opção “Gastrostomia” é descrita, vinculando as formas de dieta enteral, sendo que se o paciente tem esta dieta de forma “Artesanal” ele tende a ser reinternado, mas se tem a forma “Industrializada” ele não tende a ser reinternado. Isso possibilita melhores práticas e uma pesquisa com mais detalhes sobre o vínculo destes itens. Já para o método Apriori a mesma pesquisadora relatou que os pacientes apresentaram quadro de desnutrição, sendo que a forma de preparo dos alimentos “Artesanal ou caseira” possuía mistura de alimentos in natura, enquanto a industrializada era quimicamente definida, possibilitando recuperação do seu quadro nutricional.

Segundo o pesquisador Faruk Abrão Kalil Filho, o resultado do método Apriori condiz com a realidade apresentada nas coletas clínicas realizada no Serviço de Cirurgia do Aparelho Digestivo bem como para o método ID3.

Como já citado por Pinto (2005), o módulo SINPE[®] Analisador não visa concorrer com ferramentas de análise estatística ou de geração de gráficos disponíveis no mercado, mas de oferecer agilidade na obtenção dos dados coletados pelos usuários desta ferramenta. Este ponto é mantido para a funcionalidade de mineração de dados implementada no SINPE[®] Analisador, não tendo a visão de concorrência com outras ferramentas de mineração disponíveis como o Weka, e sim possibilitar ao usuário do SINPE[®] subsídios de complementar os resultados obtidos na coleta de dados.

A usabilidade de sistemas de informação é tema abordado nesta pesquisa. Visando suprir os atributos abordados por Nielsen sobre este tema, efetuou-se comparativo com os atributos pontuados como fundamentais pelo autor para que o sistema tenha usabilidade. Estes atributos são: learnability, efficiency, memorability, errors e satisfaction. O QUADRO 16 mostra o comparativo deste atributos para com o resultado obtido através do questionário pós-implementação da funcionalidade de mineração de dados no software em questão.

Atributo apresentado por Nielsen	Questão para os entrevistados	Resposta	Resultado
<i>learnability</i>	Qual valor você atribui para a utilização do módulo Mineração de dados no SINPE?	Fácil de aprender	100%
<i>efficiency</i>	Qual foi o nível de produtividade que você obteve com o software?	Muito produtivo Produtivo	40 % 60 %
<i>memorability</i>	Após a primeira utilização você diz que:	Utilizaria novamente Consegue utilizar	40 % 60 %
<i>errors</i>	Os erros apresentados pelo módulo são:	Nenhum erro apresentado	100 %
<i>satisfaction</i>	Qual valor você atribui para o módulo?	Excelente Muito bom	40 % 60 %

QUADRO 16 - USABILIDADE X RESULTADO
 FONTE: O autor (2009).

Inovações em tecnologia garantem que a usabilidade e o design da interface será um alvo perpetuamente em movimento (SHORTLIFFE e CIMIRO, 2006).

Este trabalho oferece alternativa para a utilização de métodos de mineração de dados para usuário do SINPE[®], permitindo obter de forma rápida e simples material de pesquisa complementar para resultados obtidos através desta ferramenta. Funcionalidades como validação cruzada para os métodos poderão ser implementadas em futuras versões.

Também se destacam funcionalidades como: viabilizar de forma gráfica os resultados obtidos pelo método ID3; tornar o método Apriori recursivo e disponibilizar

mais atributos para o método ID3 (futuras implementações possíveis para a ferramenta).

A implementação de outros métodos de mineração de dados para com o SINPE[®], trarão novas funcionalidades e agregarão novos resultados para as pesquisas que utilizam esta ferramenta. O método não desenvolvido da opção “Agrupar os dados e mostrá-los em formato de relatório”, possibilitará ao usuário visualizar quais atributos “itens do protocolo de pesquisa” são próximos a outros através de grupos.

6 CONCLUSÕES

As atividades propostas neste trabalho foram divididas em cinco objetivos específicos, conforme:

Primeiro, verificar o interesse dos usuários do SINPE[®] para com o método de mineração e dados e sua possível implementação neste software, obteve conformidade através a aplicação do questionário pré-teste e questionário definitivo.

Segundo, pesquisar na literatura os melhores algoritmos e técnicas de mineração de dados para utilização na tarefa proposta, obteve conformidade através da revisão de literatura e do método proposto.

Terceiro, implementar no módulo Analisador do SINPE[®] os algoritmos selecionados, está relacionado as atividades desenvolvidas através de pesquisas utilizando classes e bibliotecas dos métodos selecionados.

Quarto, testar estes algoritmos com profissionais da saúde e seus protocolos eletrônicos, etapa suprida através de testes com usuários do SINPE[®], no protocolo de Cirurgia do Aparelho Digestivo e de Nutrição Enteral Domiciliar.

Validar a interface proposta para a mineração de dados do módulo Analisador do SINPE[®], etapa suprida através do questionário pós-implementação.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining Association Rules between Sets of Items in Large Databases. **SIGMOD**, v. 22, n. 2, p. 207-216, 1993.

BEMMEL, J. H. V.; MUSEN, M. A. **Handbook of Medical Informatics**. Springer, 1997. Disponível em: < [http://www.mieur.nl/mihandbook/r 3 3/handbook/home.htm](http://www.mieur.nl/mihandbook/r33/handbook/home.htm) >.

BORSATO, E. P. **Modelo multicêntrico para elaboração, coleta e pesquisa de dados em protocolos eletrônicos**. 2005. 86 (Doutor). Ciências da Saúde, UFPR, Curitiba.

BÓS, A. J. G. **Epi Info® sem mistérios: um manual prático**. Porto Alegre: EDIPUCRS, 2004. ISBN 85-7430-489-1.

BRASIL, L. M. **Informática em saúde**. Londrina: Eduel, 2008.

CHEN, H. et al. **Medical Informatics: Knowledge Management and Data Mining in Biomedicine**. New York: Springer, 2005.

CUNHA, H.; RIBEIRO, S. **Introdução aos Sistemas Especialistas**. Rio de Janeiro, Livros Técnicos e Científicos Editora S.A., 1987.

DANSKY, K. H. et al. Cost analysis of telehomecare. **Telemed J E Health**, v. 7, n. 3, p. 225-232, 2001.

Dictionary of Medical Terms. London: A & C Black Publishers Ltd 2005.

EPI INFO™. 2009. Disponível em: < <http://www.cdc.gov/epiinfo/> >. Acesso em: 15/6/2009.

FAYYAD, U. M. et al. **Advances in knowledge discovery and data mining**. Menlo Park: The MIT Press, 1996. ISBN 978-0262560979.

FEVEILE, H.; OLSEN, O.; HOGH, A. A randomized trial of mailed questionnaires versus telephone interviews: Response patterns in a survey. **BMC Medical Research Methodology**, v. 7, 2007. Disponível em: < <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=1925106> >.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

HARTMANN, V. C. e COELHO, F. E. S. Banco de dados em saúde. BRASIL, L. M. (Org.). **Informática em saúde**. Brasília: Universa, 2008. p. 57-80.

ISERN, D.; MORENO, A. Computer-based execution of clinical guidelines: A review. **Int J Med Inf**, v. 77, p. 787-808, 2008.

KALIL FILHO, F. **PROTOCOLO ELETRONICO DE COLETA DE DADOS CLINICOS EM FISIOTERAPIA RESPIRATORIA PARA DOENCAS PULMONARES**. 2008. (Mestre). Clinica Cirurgica, Universidade Federal do Parana, Curitiba.

LACHAT, C. et al. Validity of two physical activity questionnaires (IPAQ and PAQA) for Vietnamese adolescents in rural and urban areas. **International Journal of Behavioral Nutrition and Physical Activity**, v. 5, 2008. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=2494550>>.

LAUDON, K. C.; LAUDON, J. P. **Sistemas de informação gerenciais**. 2007. ISBN 978-85-7605-089-6.

LIMA, E. J. L. **Gestão estratégica da informação e inteligência competitiva**. São Paulo: Saraiva, 2005.

LUNGENHAUSEN, M. et al. Randomised controlled comparison of the Health Survey Short Form (SF-12) and the Graded Chronic Pain Scale (GCPS) in telephone interviews versus self-administered questionnaires. Are the results equivalent? **BMC Medical Research Methodology**, v. 7, 2007. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=2216025>>.

MALAFAIA, O. et al. **Elektronische Datenerfassung im Bereich der klinischen Forschung, Lehre und ärztlichen Verwaltung in dem Krankenhaus der Zukunft**. 2006.

NIELSEN, J. **Usability Engineering**. Boston: Academic Press, 1993.

PINTO, J. S. **Interfase de visibilização de informações para o sistema integrado de potocolos eletrônicos**. 2005. 111 (Doutor). Ciências da Saúde, UFPR, Curitiba.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81-106, 1986. ISSN 0885-6125.

SCANDURRA, I. **Building Usability into Health Informatics. Development and Evaluation of Information Systems for Shared Homecare**. Uppsala: Acta Universitatis Upsaliensi, 2007. ISBN 978-91-554-7065-4.

SHORTLIFFE, E. H. et al. **Medical informatics: computer applications in health care**. Addison-Wesley Longman Publishing Co., Inc., p.715ed. 1990.

SHORTLIFFE, E. H.; CIMINO, J. J. **Biomedical informatics. Computer applications in health care and biomedicine**. 3. ed. New York: Springer, 2006.

SHORTLIFFE, E. H.; PERREAU, L. E. **Medical Informatics: Computer Applications in Health Care and Biomedicine**. 2. ed. New York: Springer, 2001. 854 ISBN 0-387-98472-0.

SINHA, I. et al. A Systematic Review of Studies That Aim to Determine Which Outcomes to Measure in Clinical Trials in Children. **PLoS Medicine**, v. 5, n. 4, p. 569-578, 2008. Disponível em: < <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=2346505> >.

TARGINO, M. D. G. Informação em saúde: potencialidades e limitações **Inf. Inf.**, v. 14, n. 1, p. 52-81, 2009.

TSUNODA, D. F. **Abordagens evolucionárias para a descoberta de padrões e classificação de proteínas**. 2004. 177 (Doutorado). Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, CEFET-PR, Curitiba.

TURBAN, E. *et al.* **Tecnologia da informação para gestão: transformando os negócios na economia digital**. Porto Alegre: Bookman, 2004,

_____. **Notas de aula, disciplina Tecnologias da Informação e da Comunicação**. Curitiba. 2008

WALACH, V. R. **Aplicabilidade do Sistema Integrado de Protocolos Eletrônicos - SINPE - como ferramenta gerencial na análise de custos do procedimentos da revascularização do miocárdio no Hospital de Clínicas da Universidade Federal do Paraná**. 2008. (Mestre). Universidade Federal do Paraná, Curitiba.

WARNER, H. R.; SORENSON, D. K.; BOUHADDOU, O. **Knowledge Engineering in Health Informatics**. New York: Springer, 1997. ISBN 0-387-94901-1. Disponível em: < Biblioteca da Universidade Positivo >.

WITTEN, I. H.; FRANK, E. **Data Mining: practical machine learning tools and techniques**. San Francisco: Elsevier, 2005. ISBN 0-12-088407-0.

7 APÊNDICE

7.1 QUESTIONÁRIO DE PRÉ-TESTE

Você possui conhecimento em Informática Médica?

1. Sim
2. Não

Você possui conhecimento em Data Mining?

1. Sim
2. Não

Qual programa é utilizado para análise dos dados em sua pesquisa?

1. Excel
2. SINPE[®] Analisador
3. Outro

Em sua pesquisa é utilizado estatística descritiva?

1. Sim
2. Não

Para sua pesquisa você utilizaria um software que indique tendências ou explore grande quantidade de dados?

1. Sim

1.1. Se Sim, qual o valor que você atribui para um programa com esta função?

- 1.1.1. Ótimo

1.1.2.Bom

1.1.3.Regular

1.1.4.Ruim

2. Não

Você gostaria de ter esta ferramenta no SINPE®?

1.Sim

2.Não

7.2 QUESTIONÁRIO

Em qual área se enquadra sua pesquisa?

1. Medicina

1.1. Cirurgia do Aparelho Digestivo

1.2. Cirurgia Cardiovascular

1.3. Cirurgia Plástica

1.4. Cirurgia Bariátrica

1.5. Cirurgia Torácica

1.6. Otorrinolaringologia

1.7. Neurocirurgia

1.8. Oftalmologia

1.9. Ortopedia e Traumatologia

2. Fisioterapia

2.1. Ortopedia e Traumatologia

2.2. Respiratória

2.3. Cirurgia Bariátrica

3. Enfermagem

4. Nutrição

5. Gestão em saúde

Em sua pesquisa você utiliza outro software de coleta de dados além do SINPE[®]?

1. Sim

2. Não

Você possui algum conhecimento em relação à área de Informática em Saúde?

1. Não conheço veementemente

2. Não conheço

3. Indiferente

4. Conheço

5. Conheço totalmente

Você conhece algum método de Mineração de Dados (Data Mining)?

Métodos de classificação

1. Não conheço veementemente

2. Não conheço

3. Indiferente

4. Conheço

5. Conheço totalmente

Métodos de associação

1. Não conheço veementemente
2. Não conheço
3. Indiferente
4. Conheço
5. Conheço totalmente

Métodos de agrupamento

1. Não conheço veementemente
2. Não conheço
3. Indiferente
4. Conheço
5. Conheço totalmente

Qual programa é utilizado para análise dos dados em sua pesquisa?

1. SINPE[®] Analisador
2. Excel
3. Outro

Você utilizaria um software de mineração de dados em sua pesquisa?

Se sim, qual o valor que você atribui para um programa com esta função?

1. Ruim
2. Regular
3. Bom
4. Muito bom
5. Excelente

Se não, justifique

Em sua opinião, qual seria o valor dado a esta ferramenta se fosse utilizada no SINPE[®]?

1. Ruim
2. Regular
3. Bom
4. Muito bom
5. Excelente

Em sua opinião, qual é o valor atribuído a interface do aplicativo SINPE[®]?

1. Ruim
2. Regular
3. Bom
4. Muito bom
5. Excelente

7.3 QUESTIONÁRIO PÓS-IMPLEMENTAÇÃO

Qual valor você atribui para a utilização do módulo Mineração de dados no SINPE[®]?

1. Muito difícil de aprender
2. Difícil de aprender
3. Indiferente
4. Fácil de aprender
5. Muito fácil de aprender

Qual foi o nível de produtividade que você obteve com o software?

1. Muito improdutivo
2. Improdutivo
3. Indiferente
4. Produtivo
5. Muito produtivo

Após a primeira utilização você diz que:

1. Sem condições de utilizar novamente
2. Não consegue utilizar novamente
3. Indiferente
4. Consegue utilizar novamente
5. Utilizaria novamente sem necessidade de auxílio

Os erros apresentados pelo módulo são:

1. Inúmeros erros apresentados
2. Muitos erros apresentados
3. Indiferente
4. Poucos erros apresentados
5. Nenhum erro apresentado

Qual valor você atribui para o módulo?

1. Ruim
2. Regular

3. Bom
4. Muito bom
5. Excelente

Sobre a interface do módulo

Os exemplos dos itens são:

1. Ruim
2. Regular
3. Bom
4. Muito bom
5. Excelente

A janela de seleção dos itens é:

1. Ruim
2. Regular
3. Bom
4. Muito bom
5. Excelente

O resultado apresentado do método de classificação é:

1. Ruim
2. Regular
3. Bom
4. Muito bom
5. Excelente

O resultado apresentado do método de associação é:

1. Ruim

2. Regular
3. Bom
4. Muito bom
5. Excelente

7.4 DOCUMENTAÇÃO DE IMPLEMENTAÇÃO

A implementação dos métodos de mineração de dados com o SINPE[®] Analisador ocorreu através do diagrama de classe representado pelas Figura 24 e Figura 25.

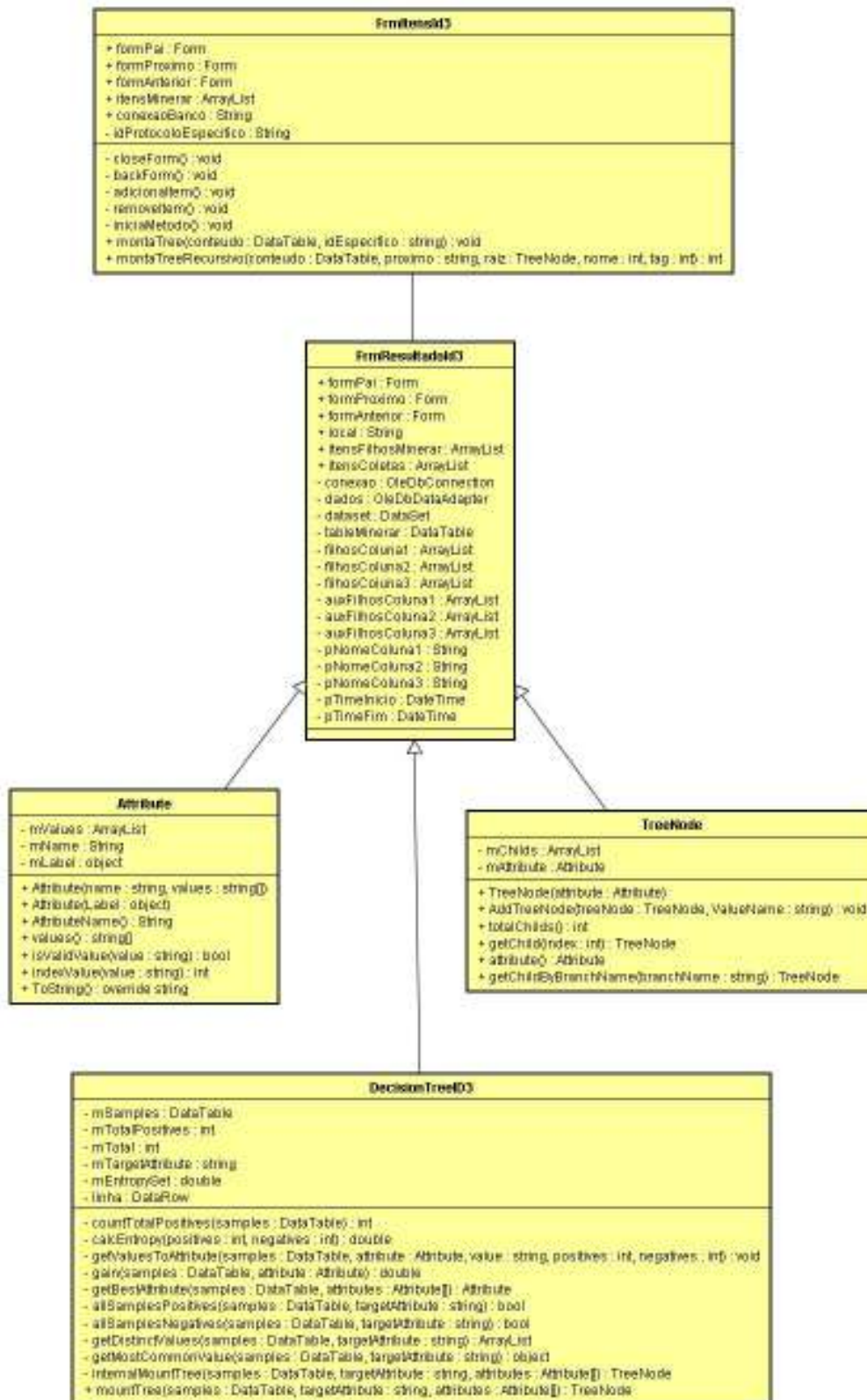


FIGURA 24 - DIAGRAMA DE CLASSE DO MÉTODO ID3

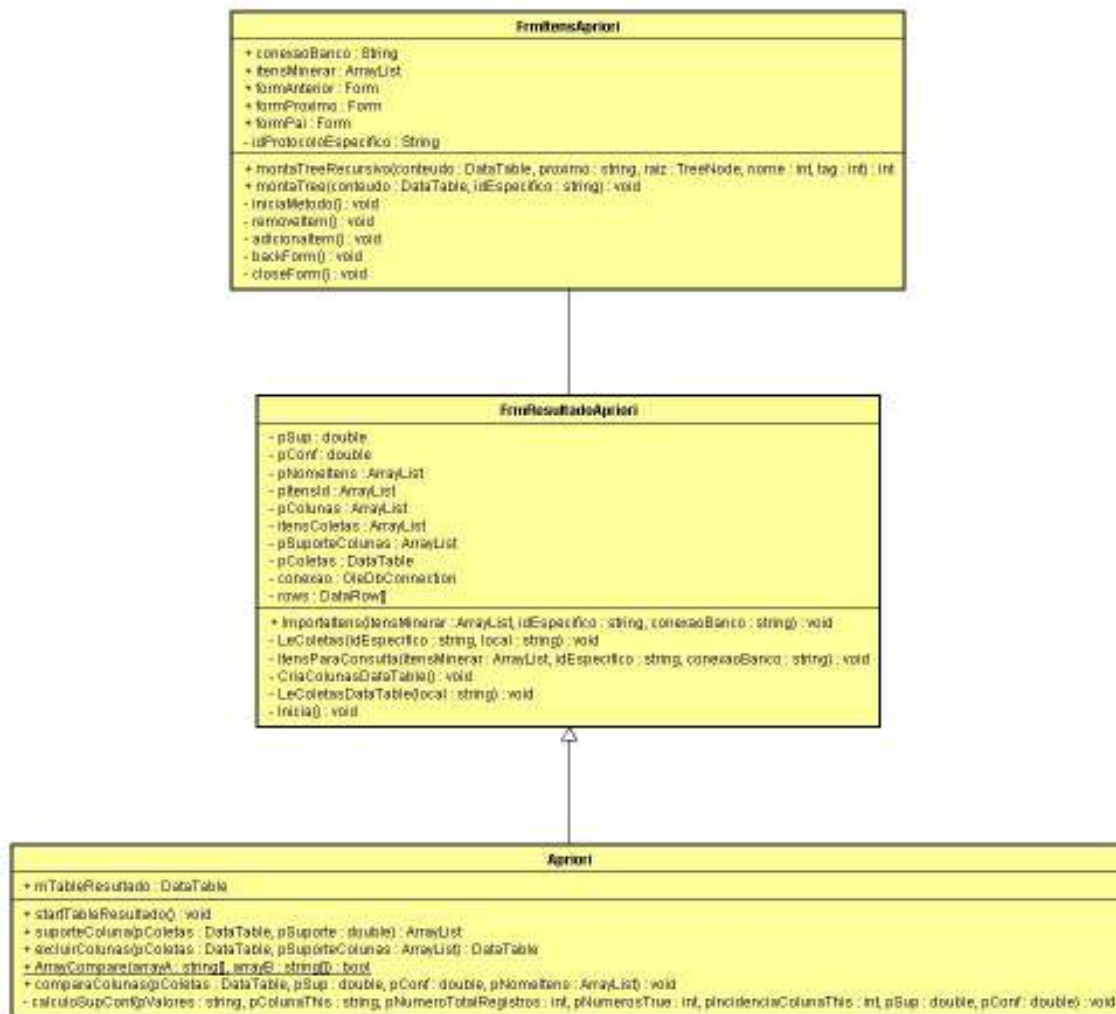


FIGURA 25 - DIAGRAMA DE CLASSE DO MÉTODO APRIORI

A lista com as principais funções das classes ID3 e Apriori são descritas no QUADRO 17.

NOME DA CLASSE	MÉTODO	FINALIDADE
Attribute	ID3	Classe que representa um atributo utilizado na classe de decisão.
TreeNode	ID3	Classe que representará a árvore de decisão montada.
DecisionTreeID3	ID3	Classe que implementa uma árvore de decisão usando o algoritmo ID3.
Importaltens	ID3	Prepara os itens para a utilização.
LeAtributoClasse	ID3	Prepara o atributo meta para a utilização pelo método ID3.
LeColetas	ID3	Classe com a consulta ao banco de dados, conforme o item selecionado.
LeColetasEfetuadas	ID3	Prepara os dados para a utilização pelo método ID3 conforme os itens informados pelo usuário.
suporteColuna	Apriori	Calcula o suporte por coluna e adiciona os que não passaram em um ArrayList.
excluirColunas	Apriori	Exclui as colunas que não tem o suporte maior que o informado pelo usuário.
comparaColunas	Apriori	Compara as colunas do Data Table para criar as regras.
calculoSupConf	Apriori	Calcula o suporte e confiança para o resultado.

QUADRO 17 - LISTA DE FUNÇÕES

7.5 MANUAL DE UTILIZAÇÃO DO SINPE© ANALISADOR: MINERAÇÃO DE DADOS

Introdução

Data Mining ou Mineração de dados é uma tecnologia voltada para a descoberta de conhecimento. Esta tecnologia é uma das etapas do KDD (*Knowledge Discoverys in Databases*), que visa à descoberta de conhecimento em base de dados, onde é composto por seis fases: seleção de dados, limpeza, enriquecimento, transformação ou codificação, mineração de dados e construção de relatórios e apresentação da informação descoberta. (HARTMANN, V. C. e COELHO, F. E. S., 2008)

Segundo Hartmann e Coelho 2008, a importância para utilização da mineração de dados, para ser útil na prática, precisa ser realizada em grandes arquivos e banco de dados.

HARTMANN, V. C. e COELHO, F. E. S. Banco de dados em saúde. BRASIL, L. M. (Org.). *Informática em saúde*. Brasília: Universa, 2008. p. 57-80.

Ferramenta

Na ferramenta de mineração de dados desenvolvida no SINPE[®] Analisador programou-se dois métodos: um de classificação de dados e outro de associação de dados.

O método de classificação de dados consiste em seqüência de itens. Exemplo: pacientes que no exame físico apresentaram sinal de Blumberg passaram por cirurgia do tipo apendicectomia com dor abdominal, estão propensos a terem apendicite aguda. O algoritmo utilizado para o método é o ID3.

Já o método de associação visa relacionar dois ou mais conjuntos de itens. Exemplo: pacientes que passaram por cirurgia de apendicectomia tiveram dor abdominal. O algoritmo utilizado para o método é o Apriori.

Como utilizar a ferramenta com o método de classificação

Abra o SINPE[®] Analisador, clique no botão “Conexão na base”. Selecione o protocolo específico que será utilizado em sua mineração e clique no botão “Minerar”.

Será apresentada a janela de boas vindas à ferramenta de mineração de dados, conforme Figura 26.

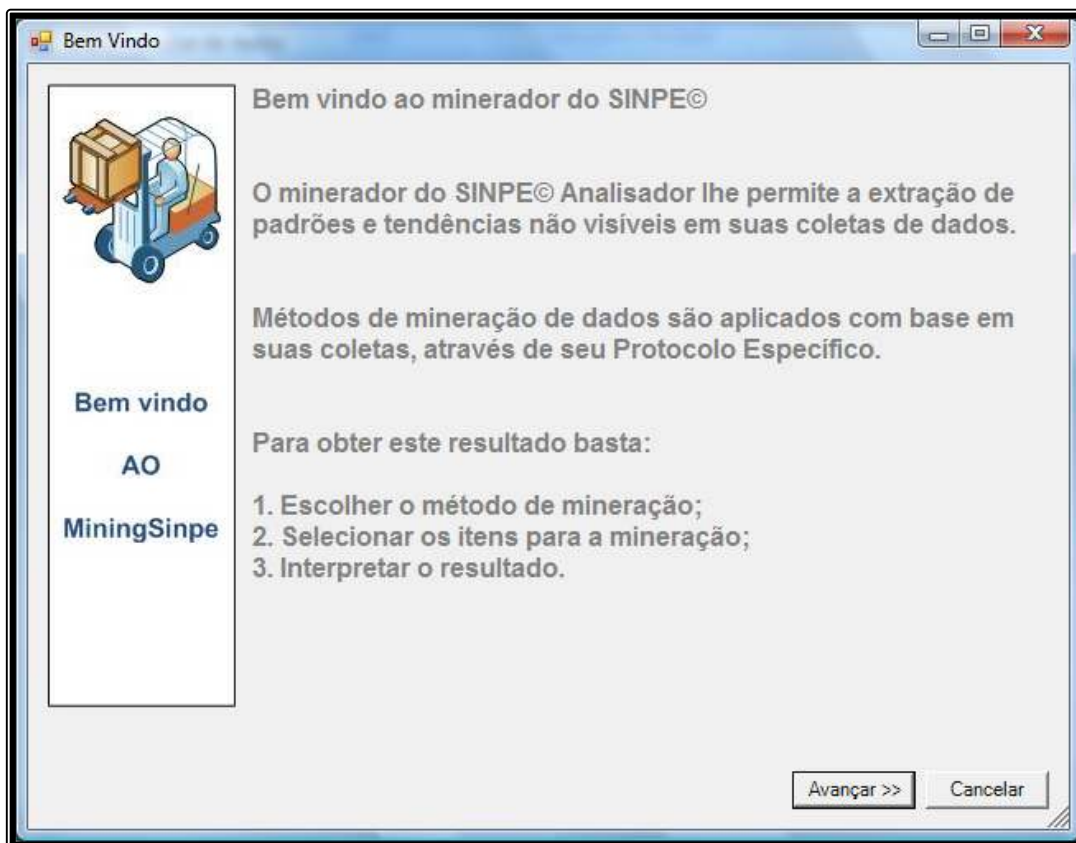


FIGURA 26 - JANELA DE APRESENTAÇÃO DA MINERAÇÃO DE DADOS

Clique no botão “Avançar”. Será apresentada a janela com a escolha do método “técnica”, Figura 27. Escolha a opção “Classificar os dados e mostrá-los em formato de árvore” e clique no botão “Avançar”.

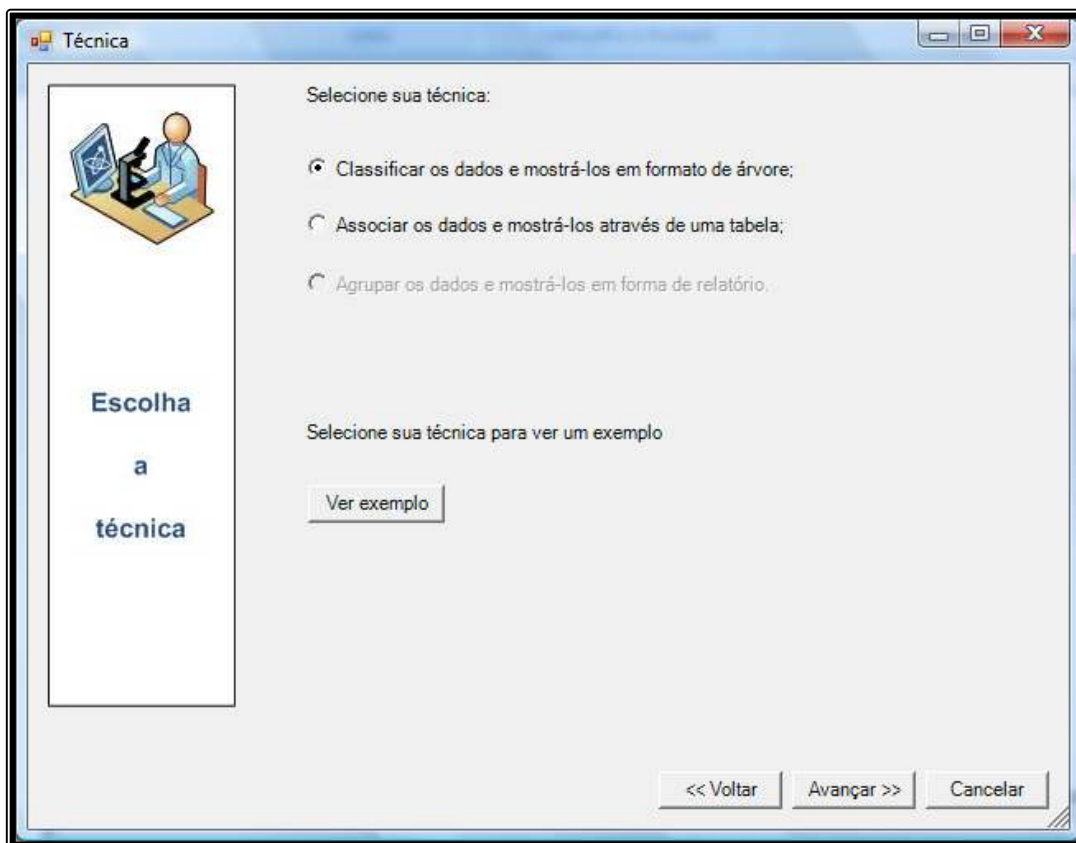


FIGURA 27 - JANELA DE ESCOLHA DO MÉTODO COM A OPÇÃO DE CLASSIFICAÇÃO

A próxima janela, (Figura 28) apresenta a seleção dos itens da mineração de dados. Para isso, basta selecionar três itens necessários para pesquisa; eles devem obrigatoriamente conter filhos (para relacioná-los com os demais itens selecionados) e um item que responderá uma questão científica. Na lista à esquerda clique no item desejado e no botão ">", os itens selecionados serão indicados na lista à direita. Caso necessite remover o item, clique nele na lista à direita e clique no botão "<". Depois de selecionados os três itens com filhos e o item da pergunta científica, clique no botão "Avançar".

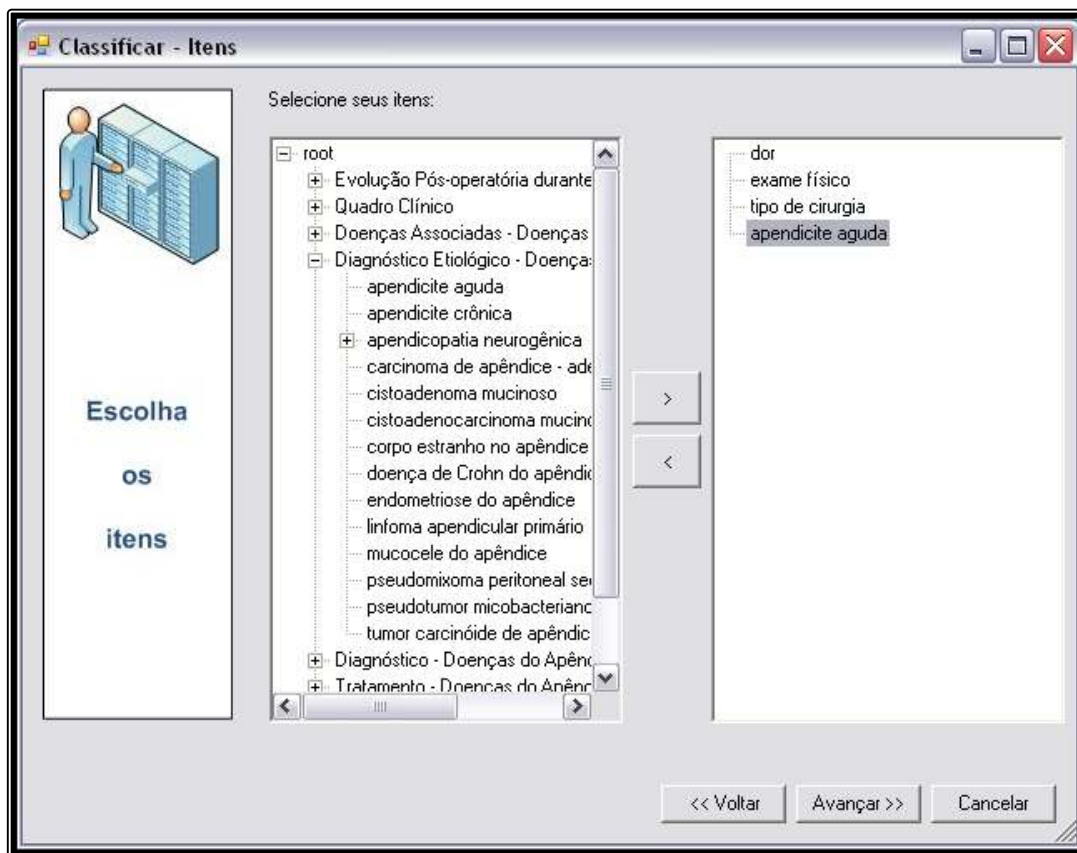


FIGURA 28 - ITENS SELECIONADOS DO PROTOCOLO ESPECÍFICO

Será apresentada uma mensagem informando que o procedimento pode levar alguns minutos. A janela de resultado abrirá (Figura 29). A parte superior esquerda mostra uma tabela contendo as coletas efetuadas. Na parte superior à direita o método utilizado e o tempo estimado da mineração de dados. Por fim, é apresentado o resultado da árvore em formato de texto indicado no campo "Árvore de decisão".

The screenshot shows a window titled "Resultado" with two main sections. The top section, "Coletas efetuadas", contains a table with four columns: "Atributo1", "Atributo2", "Atributo3", and "Atributo4". The table lists eight rows of data, all with "apendicectomia" in the third column. The second and fourth columns contain various physical exam findings like "Não possui coletas", "sinal de Blumberg", and "sinal Rovsing". The bottom section, "Árvore de decisão", displays a decision tree structure with nodes such as "[exame físico]", "<Não possui coletas>", "[tipo de cirurgia]", and "[dor]".

Atributo1	Atributo2	Atributo3	Atributo4
abdominal	Não possui coletas	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal de Blumberg	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal Rovsing	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal Rovsing	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal de Blumberg	apendicectomia	<input checked="" type="checkbox"/>
abdominal	Não possui coletas	apendicectomia	<input checked="" type="checkbox"/>
abdominal	sinal de Blumberg	apendicectomia	<input checked="" type="checkbox"/>
abdominal	Não possui coletas	apendicectomia	<input checked="" type="checkbox"/>

Método Utilizado
Classificação com árvore de decisão
Algoritmo ID3

Tempo Estimado
00:00:13

Árvore de decisão

```
[exame físico]
<Não possui coletas>
 [tipo de cirurgia]
  <apendicectomia>
   [dor]
    <abdominal>
     [True]
    <plastrão/ massa palpável em fossa ilíaca direita>
     [tipo de cirurgia]
      <apendicectomia>
       [dor]
        <abdominal>
         [True]
        <sinal de Blumberg>
         [tipo de cirurgia]
          <apendicectomia>
           [dor]
            <abdominal>
             [True]
            <sinal Rovsing>
             [tipo de cirurgia]
              <apendicectomia>
               [dor]
                <abdominal>
                 [True]
```

FIGURA 29 - JANELA DE RESULTADO

O próximo passo é interpretar o resultado obtido pelo método. Para este procedimento utilize a Figura 30 como base.

Os itens visualizados com o caracter “[” como indicado por exemplo no “[exame físico]” estão relacionados aos itens selecionados na mineração. Já os itens visualizados entre “<” como indicado por exemplo no “<sinal de Blumberg>”, são filhos dos itens selecionados. O último item apresentado pode ser “[True]” ou “[False]” que indica verdadeiro ou falso respectivamente.

A leitura da sequência apresentada na Figura 30 é definida da seguinte maneira: pacientes que efetuaram o exame físico de Blumberg passando por cirurgia do tipo apendicectomia com dor abdominal, está propenso a ter apendicite aguda.

Um fato importante, “apendicite aguda” neste exemplo é representado pelo valor “[True]”, que por sua vez faz referência ao último item selecionado que responde à uma pergunta científica.

```

|exame fisico|
  <signal de Blumberg>
    |tipo de cirurgia|
      <apendicectomia>
        |dor|
          <abdominal>
            |True|

```

FIGURA 30 - COMO INTERPRETAR O RESULTADO

Como utilizar a ferramenta com o método de associação

Abra o SINPE[®] Analisador e clique no botão “Conexão na base”. Selecione o protocolo específico que será utilizado em sua mineração de dados e clique no botão “Minerar”.

Será apresentada a janela de boas vindas à ferramenta de mineração de dados (Figura 31 - JANELA DE APRESENTAÇÃO DA MINERAÇÃO DE DADOS Figura 31).



FIGURA 31 - JANELA DE APRESENTAÇÃO DA MINERAÇÃO DE DADOS

Clique no botão “Avançar”. Será apresentada a janela com a escolha do método “técnica” (Figura 32). Escolha a opção “Associar os dados e mostrá-los através de uma tabela” e clique no botão “Avançar”.

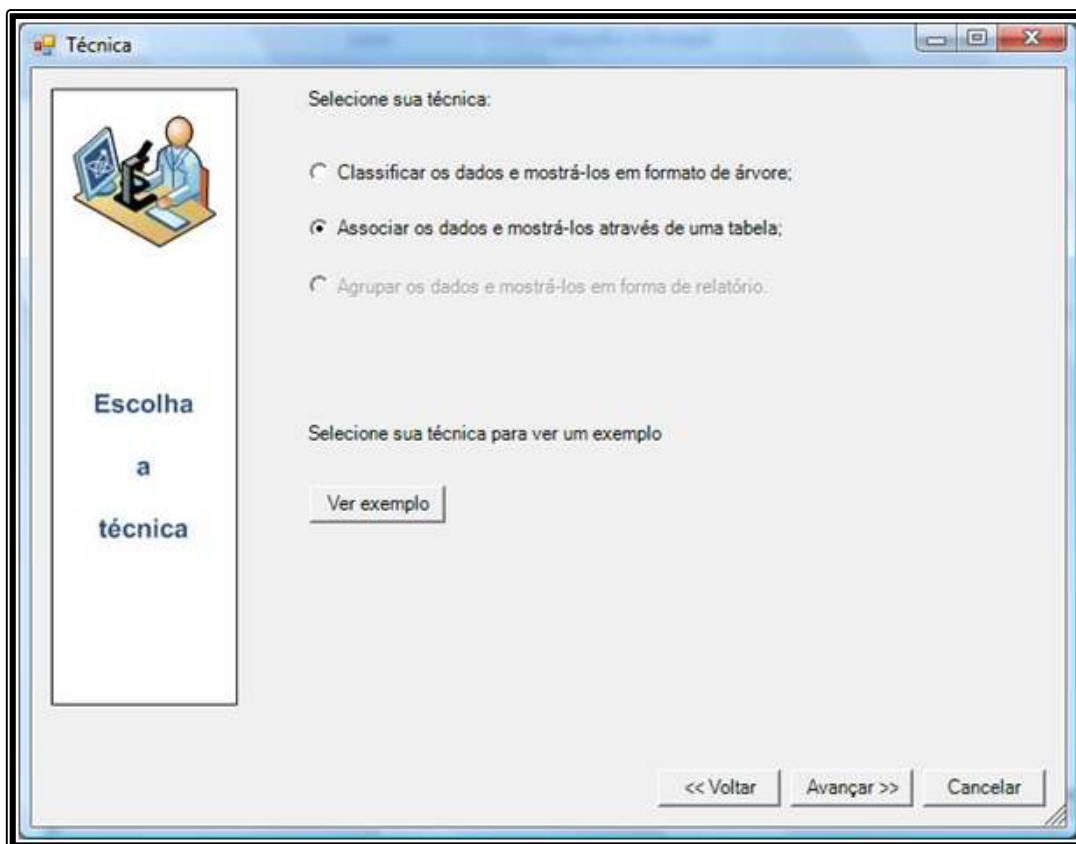


FIGURA 32 - JANELA DE ESCOLHA DO MÉTODO COM A OPÇÃO DE ASSOCIAÇÃO

A próxima janela (Figura 33), apresenta a seleção dos itens da mineração de dados. Para isso, basta selecionar o número de itens que você achar necessário e continuar o processo. Na lista à esquerda clique no item desejado e no botão “>”, os itens selecionados serão indicados na lista à direita. Caso necessite remover o item, clique no item na lista à direita e clique no botão “<”. Depois de selecionados os três itens com filhos e o item da pergunta científica clique no botão “Avançar”.

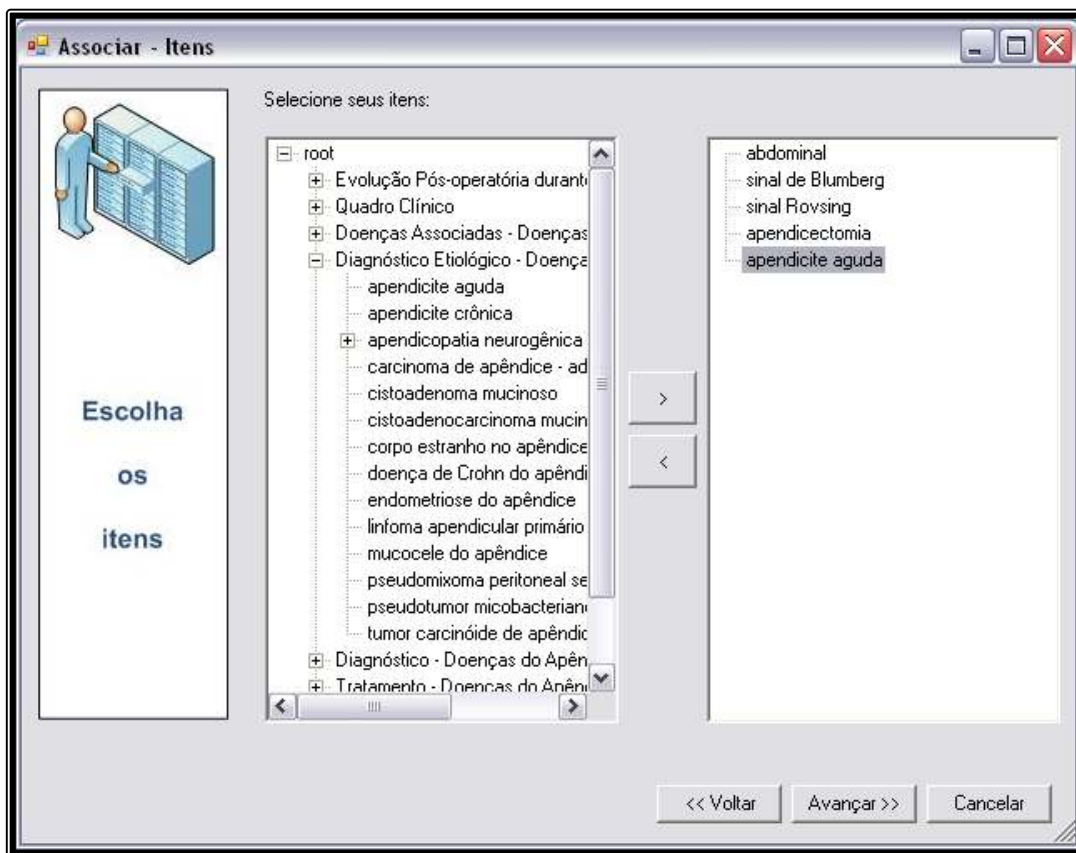


FIGURA 33 - ITENS SELECIONADOS DO PROTOCOLO ESPECÍFICO

Será apresentada uma mensagem informando que o procedimento pode levar alguns minutos. A janela de resultado abrirá (Figura 34). A parte superior esquerda mostra uma tabela contendo as coletas efetuadas. Na parte superior à direita o método utilizado e o botão “Alterar os valores”. Por fim, é apresentado o resultado das regras em formato de tabela indicado no campo “Arvore da Associação”.

Resultado						
Coletas efetuadas						
Id	1	2	3	4	5	
9	True	True	False	True	True	
10	True	True	False	True	True	
11	True	False	False	True	True	
12	True	False	False	True	True	
13	True	True	False	True	True	
14	True	True	False	True	True	
15	True	True	False	True	True	
16	True	True	False	True	True	

Resultado da Associação		
Valor	Suporte em %	Confiança em %
abdominal está para apendicectomia	100	100
sinal de Blumberg está para abdominal	60,606060606060609	100
sinal de Blumberg está para apendicectomia	60,606060606060609	100
sinal Rovsing está para abdominal	10,606060606060606	100
sinal Rovsing está para apendicectomia	10,606060606060606	100
apendicectomia está para abdominal	100	100
apendicite aguda está para abdominal	87,878787878787875	100
apendicite aguda está para apendicectomia	87,878787878787875	100
sinal de Blumberg está para apendicite aguda	60,606060606060609	90
abdominal está para apendicite aguda	100	87,878787878...
apendicectomia está para apendicite aguda	100	87,878787878...
sinal Rovsing está para apendicite aguda	10,606060606060606	71,428571428...
apendicite aguda está para sinal de Blumberg	87,878787878787875	62,068965517...
abdominal está para sinal de Blumberg	100	60,606060606...
apendicectomia está para sinal de Blumberg	100	60,606060606...
sinal Rovsing está para sinal de Blumberg	10,606060606060606	57,142857142...

FIGURA 34 - JANELA DE RESULTADO

O próximo passo é interpretar o resultado obtido pelo método. Para este procedimento, utilize a Figura 35 como base.

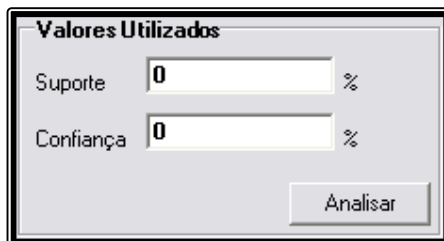
A leitura de cada regra apresentada é definida da seguinte maneira: para a primeira linha, pacientes tiveram dor abdominal passaram por cirurgia do tipo apendicectomia; esta regra é 100% confiável. Para a segunda linha, pacientes que apresentaram no exame físico sinal de Blumberg tinham dor abdominal; esta regra é 100% confiável.

Valor	Suporte em %	Confiança em %
abdominal está para apendicectomia	100	100
sinal de Blumberg está para abdominal	60,606060606060609	100

FIGURA 35 - COMO INTERPRETAR O RESULTADO

Para refinar sua pesquisa clique no botão “Alterar os valores”. Será apresentada janela conforme a Figura 36. Os valores de suporte e confiança possibilitam o descarte de regras pouco relevantes e menos frequentes. Altere os

parâmetros e clique no botão “Analisar”, o filtro será aplicado e apenas as regras que se adequem aos parâmetros informados serão apresentadas.



Valores Utilizados

Suporte %

Confiança %

Analisar

FIGURA 36 - PARÂMETROS DO MÉTODO APRIORI