UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DA COMPUTAÇÃO

Descoberta Automatizada de Associações com o Uso do Algoritmo Apriori como Técnica de Mineração de Dados

Derciley Cunha de Almeida

Goiânia-GO
2011

Derciley Cunha de Almeida

Descoberta Automatizada de Associações com o Uso do Algoritmo Apriori como Técnica de Mineração de Dados

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica e de Computação da Universidade Federal de Goiás, para obtenção do título de Mestre.

Área de Concentração: Engenharia da Computação.

Orientador: Professor Dr. Leonardo da Cunha Brito.

Goiânia-GO
2011

Dados Internacionais de Catalogação na Publicação na (CIP)
GPT/BC/UFG

Almeida, Derciley Cunha.
A447d Descoberta automatizada de associações com o uso do algoritmo apriori como técnica de mineração de dados [manuscrito] / Derciley Cunha de Almeida. - 2011.
xv, 173 f. : il., figs, tabs.

Orientador: Prof. Dr. Leonardo da Cunha Brito.
Dissertação (Mestrado) – Universidade Federal de Goiás, Escola de Engenharia Elétrica e de Computação, 2011.
Bibliografia.
Inclui lista de figuras, abreviaturas, siglas e tabelas.

1. Mineração de dados 2. Algoritmo Apriori 3. Descoberta de associações I. Título.

CDU: 004.622

Universidade Federal de Goiás
Escola de Engenharia Elétrica e de Computação
Coordenação do Programa de Pós-Graduação em Engenharia Elétrica e de Computação

FOLHA DE APROVAÇÃO

"Descoberta Automatizada de Associações com o Uso do Algoritmo Apriori como Técnica de Mineração de Dados"

DERCILEY CUNHA DE ALMEIDA

Dissertação defendida e aprovada pela banca examinadora constituída pelos senhores:

Prof. Dr. Leonardo da Cunha Brito – Orientador (FEF/UFG)

Prof. Dr. Vinícius Sebba Patto – INF/UFG

Prof. Dr. Gélson da Cruz Júnior - EEEC/UFG

Prof. Dr. Cássio Dener Noronha Vinhal – EEEC/UFG

Goiânia, 25 de fevereiro de 2011

À minha esposa Lílian e à minha filha Kimberly, meus grandes amores.

AGRADECIMENTOS

Ao Deus criador, razão de toda a existência.

A minha querida esposa pelo apoio incondicional e compreensão.

A minha filha, que traz motivação com seu simples olhar.

Aos meus familiares, por todo o carinho que recebi.

Ao professor e grande amigo Dr. Leonardo Guerra de Rezende Guedes, pelas ajudas mais que valiosas.

Ao meu orientador, o professor Dr. Leonardo da Cunha Brito, pela paciência, apoio e direcionamentos.

Aos demais professores por todo o conhecimento por mim adquirido.

Aos colegas pela amizade.

A Diretoria da Central de Medicamentos de Alto Custo Juarez Barbosa, pelo apoio durante a realização das pesquisas.

“Nós estamos nos afogando em informações, mas sofrendo a falta de conhecimento”.

John Naisbett

RESUMO

Atualmente é possível o armazenamento e o gerenciamento de grandes quantidades de dados, através de modernos sistemas informatizados. Por outro lado, a análise completa e a extração do máximo de informações desse universo de dados disponíveis passaram a ser um grande desafio, diante das limitações próprias de um ser humano. Essa dissertação aborda o tema mineração de dados, também muito conhecido pelo termo em inglês *data mining*. Trata-se da extração de informações de bases de dados de forma automatizada, com o uso de recursos tecnológicos. Uma das possibilidades que as tecnologias de *data mining* oferecem é a busca automatizada de possíveis associações existentes entre dados. As informações sobre associações entre dados podem ser muito úteis para se compreender possíveis relações de causa e efeito entre muitas variáveis envolvidas em estudos e análises de dados para tomada de decisões. Há várias técnicas de mineração de dados e muitas podem ser utilizadas para descoberta de associações. O principal objetivo deste trabalho é estudar mais especificamente o método de busca automatizada de associações conhecido como Apriori de forma a avaliar sua sistemática, capacidade e resultados.

O estudo é direcionado por um problema que está relacionado à busca pelo aprimoramento dos resultados gerados pelo algoritmo Apriori sob a premissa de que uma preparação de dados específica e direcionada para o uso do algoritmo pode aprimorar os resultados do processo de mineração de dados.

As conclusões são extraídas de um estudo de caso sobre a aplicação do algoritmo Apriori em uma base de dados com informações sobre fornecimento de medicamentos de uma unidade de saúde. São avaliados e comparados os resultados de três experimentos para se verificar a influência de uma preparação de dados no desempenho do algoritmo.

Ficou evidenciado que o algoritmo Apriori alcança resultados satisfatórios na tarefa de busca por associações entre dados, no entanto, é recomendável uma preparação específica desses dados para que a aplicação do algoritmo alcance melhores resultados ou muitas associações existentes podem não ser encontradas.

Palavras-chave: mineração de dados; descoberta de associações; Apriori; banco de dados;WEKA.

ABSTRACT

Nowadays, the use of modern information systems allows the storage and management of increasingly large amounts of data. On the other hand, the full analysis and the maximum extraction of useful information from this universe of available data present considerable challenges in view of inherent human limitations. This dissertation deals with the subject of data mining, which is the use of technology resources in order to extract information from databases in an automated way. One of the possibilities offered by data mining technologies is the automated search for possible associations within data. Information about such associations can be useful for understanding cause and effect relationships between the involved variables in data analysis for decision making. There are several data mining techniques and many of them can be used for discovering associations. The main goal of this work is to study a particular method for automated search of associations called “Apriori”, evaluating its capabilities and outcomes.

The study focuses on the problem of improving the Apriori algorithm results, taking into consideration that the results of the data mining process might be improved if the data are prepared specifically for Apriori application.

The conclusions are drawn from a case study in which the Apriori algorithm was applied to a database with information on drug distribution at a health institute. The results of two experiments are considered in order to evaluate the influence of data preprocessing on the Apriori algorithm's performance.

It was found that the Apriori algorithm yields satisfactory results on the discovery of association in data; however, for best results, it is advisable that the data be prepared in advance, specifically for the Apriori application, otherwise many associations in the database might be left undiscovered.

Keywords: data mining; association discovery; Apriori; databases; WEKA.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS E SIGLAS

CART – Classification and Regression Trees (algoritmo desenvolvido por um grupo de estatísticos, entre eles L. Breiman, J. Friedman, R. Olshen e C. Stone).

CEP – Código de Endereçamento Postal.

CID – Classificação Estatística Internacional de Doenças e Problemas Relacionados com Saúde.

CMAC – Central de Medicamentos de Alto Custo.

CRM – Customer Relationship Management (Gestão do Relacionamento com o Cliente).

DATASUS – Departamento de Informática do SUS.

ERP – Enterprise Resource Planning (Gestão dos Recursos Empresariais).

IBGE - Instituto Brasileiro de Geografia e Estatística.

ID3 – Iterative Dichotomiser (algoritmo de indução de árvores de decisão desenvolvido por J. Ross Quinlan).

IMC – Índice de Massa Corporal.

JVM – Java Virtual Machine (Máquina Virtual Java).

KDD – Knowledge Discovery in Databases (Descoberta de Conhecimento em Banco de Dados).

LME – Laudo para Solicitação/Autorização de Medicamentos de Dispensação Excepcional e Estratégicos.

OLAP – On-Line Analytical Processing (Processamento Analítico em “Tempo Real”).

OLTP – On-Line Transaction Processing (Processamento de Transações em “Tempo Real”).

OMS – Organização Mundial de Saúde.

SIMPEP – Simpósio de Engenharia da Produção.

SQL – Structured Query Language (Linguagem de Consulta Estruturada).

SUS – Sistema Único de Saúde.

UNESP – Universidade Estadual Paulista.

WEKA – Waikato Environment for Knowledge Analysis (software de mineração de dados desenvolvido pela Universidade Waikato, situada na Nova Zelândia).

XML – eXtemsible Markup Language (Linguagem de Marcação Extensível).

SUMÁRIO

1- INTRODUÇÃO

1.1- Tema

O tema desse trabalho é **Mineração de Dados,** também muito conhecido pelo termo em inglês ***data mining***.

A proposta dessa dissertação é estudar mais especificamente o método de busca automatizada por associações em base de dados conhecido como Apriori, com o objetivo de avaliar sua sistemática, capacidade e resultados.

Este tema tem provocado discussões entre os estudiosos do assunto no que se refere à sua definição e à sua abrangência, pois existe um outro termo muito utilizado em estudos relacionados, que é a Descoberta de Conhecimentos em Bancos de Dados, ou em inglês *Knowledge Discovery in Databases (KDD)*.

Vários autores entendem que *data mining* é apenas uma das etapas de todo o processo de *KDD* (Han *et al.*, 2006, p. 5; Fayyad *et al.*, 1996, p. 3; Adriaans *et al.*, 1996, p. 5; Passos *et al.*, 2005, p. 2). Alguns chegam a criticar o fato de muitos outros autores erroneamente ensinarem que os termos *data mining* e *KDD* são sinônimos. Entretanto, observa-se que várias obras que mencionam e descrevem todo um processo denominado como *KDD* são intituladas com o uso do termo *data mining*.

Neste trabalho é utilizado o termo mineração de dados (ou *data mining*), até pelo fato de ser mais conhecido, sob a ótica de que se trata de um processo completo, tal como o *KDD*, para extrair ou descobrir padrões, informações ou conhecimentos através de análise automatizada de bases de dados, com o uso de recursos tecnológicos (Witten *et al.*, 2005, p. 5; Silberschatz *et al.*, 2006, p. 496; Bigus,1996, p. 9; Tang *et al.*, 2005, p. 2; Elmasri *et al.*, 2005, p. 624; Han *et al.*, 2006, p. 5; Berson *et al.*, 2000, p. 6).

1.2- Problema

Para Barros (2004, p. 78) “todo trabalho científico nasce de uma dificuldade ou questionamento que deve ser cuidadosamente formulado”, isto é, um problema oriundo do tema geral de estudo.

Com o mesmo entendimento, Rudio (2004, p. 87) afirma que “toda pesquisa científica começa pela formulação de um problema e tem por objetivo buscar a sua solução”. Esse problema é “uma questão proposta para ser discutida”. Ainda segundo o autor, a formulação do problema “consiste em dizer, de maneira explícita, clara, compreensível e operacional, qual a dificuldade” que se pretende analisar (Rudio, 2004, p. 94).

Nesse sentido, o problema analisado neste trabalho é:

- **o algoritmo Apriori apresenta limitações e baixo desempenho para encontrar regras de associação durante o processo de mineração de dados quando são processados dados que possam apresentar muitos itens distintos, como variáveis numéricas e datas.**

1.3- Hipótese

Este trabalho adota o **método** de abordagem **hipotético-dedutivo.** Trata-se de um método considerado lógico por excelência, historicamente relacionado com a experimentação (Andrade, 2003, p. 132).

Neste método, um problema (dúvida) é definido a partir de estudos, conhecimento prévio e observação de fatos ou fenômenos relacionados a um determinado tema ou assunto e uma solução, provisória, para este problema é proposta, a hipótese (Koche, 2002, p. 70). A hipótese é testada durante o desenvolvimento da pesquisa e, ao final, pode ser rejeitada ou não (Andrade, 2003, p. 143). Novos problemas podem surgir, os quais poderão ser novamente avaliados e assim por diante (Koche, 2002, p. 70).

Uma hipótese é uma suposta, provável e provisória resposta a um problema, cuja comprovação será verificada através da pesquisa (Marconi *et al.*,

2007, p. 128). Trata-se de uma tentativa de antecipar a resposta do problema de pesquisa (Martins, 2002, p. 41).

Embora nem todos os tipos de pesquisa necessitem da formulação de hipótese, obtêm-se grandes vantagens metodológicas nas pesquisas em que há essa possibilidade, como os estudos experimentais e os estudos descritivos, pois a hipótese se torna um balizador para o pesquisador na condução do trabalho (Martins, 2002, p. 41). A hipótese fixa uma diretriz capaz de impor ordem e finalidade a todo o processo de experimentação (Ruiz, 2002, p. 54).

Assim, toda hipótese é uma tentativa de resposta ao problema e possui a função de orientar o pesquisador na coleta e análise dos dados (Barros, 2004, p. 83).

Segundo Marconi *et al.* (2007, p. 130), há várias maneiras de formular uma hipótese, mas a mais comum é a utilização de uma expressão do tipo “se x, então y”.

Assim, com base nas teorias e aplicações práticas registradas em várias bibliografias, este trabalho se baseia na seguinte hipótese básica:

- **se houver uma preparação de dados direcionada para o uso do algoritmo Apriori em um processo de mineração de dados, então os resultados gerados podem ser aprimorados e mais associações podem ser encontradas após o processamento do algoritmo.**

As hipóteses devem ser postas à prova, verificadas, aprovadas ou reprovadas pelos fatos (Ruiz, 2002, p. 55). Para Marconi *et al.* (2007, p. 163) os resultados finais alcançados com a realização da pesquisa podem comprovar ou rejeitar as hipóteses.

1.4- Objetivos

O trabalho tem como objetivo principal **identificar as principais limitações do algoritmo Apriori, com as respectivas orientações para que sejam superadas, e avaliar a influência de uma preparação dos dados disponíveis, diante da definição dos parâmetros e critérios mínimos para a**

realização da busca de associações pelo algoritmo, nos resultados encontrados.

Entre os objetivos específicos do trabalho estão:

- descrever os aspectos mais determinantes para um melhor desempenho do algoritmo Apriori;
- avaliar e comparar os resultados de uma aplicação do processo de mineração de dados em uma base de dados ao aplicar o algoritmo, sob diferentes condições de preparação de dados;
- contribuir com um direcionamento para realização de buscas automatizadas por associações em bases de dados;
- verificar a relação entre a definição de critérios mínimos para realização de busca de associações e a preparação dos dados para o aprimoramento dos resultados gerados pelo algoritmo Apriori;
- sugerir uma metodologia para a aplicação do processo de mineração de dados com o uso do algoritmo Apriori que ofereça um direcionamento para uma preparação de dados específica dos dados, antes do processamento do algoritmo, como alternativa para aumentar as chances de se encontrar mais associações entre esses dados.

1.5- Justificativa

Diante da capacidade para armazenamento de grandes quantidades de dados em razão dos avanços dos recursos tecnológicos, surgem as limitações para análise desses dados de forma a extrair o máximo de informações.

Atualmente, existem muitas ferramentas que auxiliam na análise de dados. Elas oferecem recursos para realização de consultas rápidas e complexas em bancos de dados. Todavia, muitas informações passam despercebidas durante essas consultas. A aplicação de técnicas de mineração de dados, em conjunto com essas consultas ou até mesmo como um recurso complementar, pode aumentar ainda mais o poder de extração e descoberta de informações.

Uma informação sobre a existência de uma associação específica entre dados pode ser muito útil em diversas situações. Entretanto, nem sempre ela está visível, ou seja, nem sempre é facilmente identificada e extraída de uma grande base de dados. Enquanto algumas associações são facilmente identificadas, outras podem estar “escondidas”, principalmente quando há grandes quantidades de dados envolvidos.

O interesse por ferramentas de descoberta automatizada de conhecimentos em bancos de dados tem crescido bastante. O grande crescimento do número de publicações sobre o assunto demonstra isso. Muitas publicações tratam de relatos de aplicações de ferramentas de *KDD* em diversas áreas como: negócios; governo; medicina; e ciência (Fayyad *et al.*, 1996, p. 3).

Vale destacar o fato de que há uma variedade de metodologias que podem ser utilizadas para analisar fontes de dados com o objetivo de descobrir padrões e tendências nelas existentes. Da mesma forma, existem muitas tecnologias e ferramentas disponíveis para a realização de processos de *data mining* (Westphal *et al.*, 1998, p. 6).

Com relação à busca automatizada por associações, um método de mineração de dados bastante eficiente, conhecido e utilizado é o algoritmo Apriori (Tang *et al.,* 2005, p. 230). Este método foi proposto com o objetivo específico de descobrir associações existentes em bases de dados. A técnica se baseia no princípio de que há uma grande chance de existir uma associação entre itens presentes entre os dados se estes aparecem frequentemente juntos.

Um algoritmo de busca de associações nada mais é que uma engenharia de contagem de correlações existentes (Tang *et al.,* 2005, p. 230).

O algoritmo Apriori inicialmente identifica os itens existentes entre os dados e verifica a frequência em que cada um se repete. Em seguida, separa os itens com as maiores frequências e verifica se alguma combinação entre eles também ocorre muitas vezes.

A idéia é bem original, simples e muito interessante, já que após a identificação de uma grande frequência em que uma combinação de itens se repete, pode-se investigar possíveis razões para que isso tenha ocorrido e, então, chegar-se a uma conclusão sobre a existência de algum tipo de padrão de associação entre esses itens.

O método Apriori verifica a frequência em que a combinação de dois itens aparece. Se a frequência for alta, adiciona-se um terceiro item a combinação, faz-se novamente a verificação e assim por diante.

A sistemática de análise pode ficar prejudicada se entre os dados houver uma grande quantidade de itens diferentes. Isso porque pressupõe-se que entre muitos itens diferentes há relativamente poucas repetições. As baixas frequências em que os itens originalmente aparecem não possibilitam a identificação de combinações de itens que apresentam um maior número de repetições que possam destacá-las das demais.

Determinados tipos de dados apresentam muitos itens diferentes. Por exemplo, em um período de um ano, há mais de 360 dias possíveis para eventos ocorrerem. Assim, em conjuntos de dados com informações sobre a ocorrência de diversos eventos, durante vários anos, possivelmente existem muitos itens diferentes referentes às datas em que estes eventos ocorreram. Dessa forma, apenas uma coincidência muito peculiar poderia gerar uma alta frequência de uma combinação entre uma determinada data e um evento específico de forma que pudesse ser identificada pelo método Apriori.

Essa situação também pode ocorrer com dados numéricos, principalmente com variáveis contínuas.

Em relação a variáveis numéricas, por exemplo, segundo Larose (2005, p. 190), o algoritmo Apriori apresenta uma deficiência para processar bem variáveis desse tipo, a não ser que sejam discretizadas durante uma etapa de pré-processamento de dados.

Dessa análise, surge a premissa de que de uma simples aplicação do método Apriori, diretamente em uma base de dados que contenha uma maior quantidade desses tipos de dados, não resultam muitas informações sobre associações.

Observa-se que, atualmente, é muito comum as bases de dados possuírem bastantes registros de dados numéricos e do tipo data, com muitos itens distintos. Diante disso, alguns direcionamentos para realização de uma aplicação de um processo de mineração de dados precisam ser considerados quando a busca por associações é o principal objetivo a ser alcançado. Se há a premissa de que o algoritmo Apriori apresenta limitações para lidar com esses tipos de dados, pode-se avaliar o uso de algum outro método ou considerar a possibilidade de os dados

serem transformados para que se tornem mais adequados para análise do algoritmo, pois não é interessante simplesmente deixar de analisar esses dados.

Se não é recomendável a submissão de dados com muitos itens distintos a um processamento pelo algoritmo Apriori, então é indicado que alguns procedimentos sejam realizados previamente para o agrupamento desses itens. O principal objetivo dessa transformação deve ser o de reduzir as quantidades de itens distintos presentes na base de dados, ao mesmo tempo em que se preserva a sua integridade. As mudanças devem buscar apenas abstrair a essência das informações relacionadas a esses dados.

Os dados precisam deixar de conter muitos itens distintos e com baixas frequências e passar a apresentar poucos grupos de itens com frequências maiores. Com isso, a busca passa a considerar as associações possíveis entre itens que representam um determinado grupo ou categoria e os que compõem outros agrupamentos, em vez de avaliar as combinações entre itens isolados.

Com um agrupamento a busca não seria direcionada para associações do tipo “um evento ocorre mais frequentemente em uma data”, mas que “determinada categoria de eventos ocorre dentro de um período específico”.

Ainda assim, a transformação de dados requer cuidados para que não haja o comprometimento da integridade das informações envolvidas. A transformação de dados não pode alterar o real significado das informações que podem ser extraídas desses dados.

Nesse sentido, um agrupamento de itens não pode ser realizado de uma forma arbitrária ou sem critérios válidos que representem a realidade dos fatos relacionados aos dados. Não se pode, por exemplo, tratar determinado material como pertencente a uma categoria se não houver alguma similaridade com os demais itens desse grupo.

O ideal é que os critérios sejam bem definidos para que grupos sejam formados conforme a maior similaridade ou aproximação entre as características de seus itens. Para isso, essas aproximações precisam ser identificadas. Mais interessante seria encontrar uma forma de descobrir os agrupamentos que já ocorrem naturalmente. Algumas outras técnicas específicas de mineração de dados são capazes de descobrir agrupamentos existentes entre os dados e poderiam ser utilizadas nessa fase de preparação de dados para o uso do algoritmo Apriori.

Assim, é preciso que haja uma preparação específica dos dados, de uma forma bem orientada, para que os resultados a serem gerados pelo algoritmo Apriori sejam aprimorados. Essa preparação envolve outros procedimentos, além das transformações de dados. Alguns procedimentos como a seleção e limpeza adequadas de dados e remoção de duplicidades são muito importantes em qualquer processo de mineração de dados, seja qual for o objetivo a ser alcançado.

Um conjunto de orientações pode direcionar o processo de preparação de dados de forma a melhorar o desempenho da análise do algoritmo Apriori e, consequentemente, mais associações podem ser encontradas.

Vale destacar que algumas oportunidades também motivaram a realização deste trabalho. Não há notícias sobre aplicações do processo de mineração de dados para análise e extração de informações pela Secretaria Estadual de Saúde, onde o pesquisador atuou profissionalmente por cerca de um ano, e uma iniciativa pode despertar o interesse e a cultura do uso mais frequente deste processo. Um primeiro trabalho pode se tornar uma referência para novas aplicações em diferentes situações e para vários outros propósitos. Nesse sentido, uma base de dados de uma unidade pública de saúde jurisdicionada à Secretaria Estadual de Saúde foi escolhida para realização de um estudo de caso que pudesse verificar a hipótese desse trabalho.

O universo de dados disponíveis na Secretaria é bem maior, o que faz aumentar as oportunidades para extração de vários tipos de informações sob diversos aspectos, regiões e demais variáveis relacionadas à área da saúde. Apenas para uma melhor compreensão da dimensão desse universo de dados, vale ressaltar que há uma série de procedimentos que devem ser adotados em toda a rede de saúde pública no Brasil para registrar as diversas informações sobre os atendimentos realizados nas unidades de saúde. Essas informações são coletadas sistematicamente e formam um grande universo de dados gerenciado por um órgão do Ministério da Saúde denominado Departamento de Informática do Sistema Único de Saúde (SUS), mais conhecido como DATASUS. Este órgão sistematiza e disponibiliza esses dados para estudos e tomada de decisões de gestores públicos. Assim, a Secretaria Estadual de Saúde possui acesso a esse universo de dados, além dos que o próprio órgão coleta e gerencia.

Diante disso, é possível obter dados com informações diversas e abrangentes sobre pacientes e atendimentos realizados nas diversas unidades de

saúde pública, como: consultas; procedimentos; uso e fornecimento de medicamentos; internações; cirurgias; tratamentos; entre muitas outras.

Sabe-se das dificuldades e problemas para realização da coleta desses dados por uma série de fatores (tecnológicos e até culturais). Assim, há ressalvas sobre a consistência e integridade dessa base de dados, o que reforça uma necessidade de uma análise profunda e melhor preparação desses dados durante processos de extração de informações.

Organizações que atuam em outras áreas da Administração Pública Estadual também já possuem imensas bases de dados históricos armazenados eletronicamente. Os dados abrangem as ações realizadas pelos órgãos e entidades públicos que atuam em áreas como educação, segurança pública, arrecadação e fiscalização de tributos, entre outras. Há também os originados de diversas coletas de informações e estatísticas sobre índices e indicadores, como os de desenvolvimento econômico e humano. Esse universo de dados pode ser utilizado como valiosa fonte de informação para tomada de decisões.

Nesse sentido, há grandes oportunidades para aplicação do processo de mineração de dados nessas bases de dados.

1.6- Estrutura do trabalho

"É imprescindível correlacionar a pesquisa com o universo teórico", pois toda a pesquisa deve conter as premissas ou pressupostos teóricos sobre os quais será fundamentada a interpretação de dados, fatos e resultados (Marconi *et al.*, 2007, p. 226). Um método eficiente para se realizar essa correlação é através da pesquisa bibliográfica que "é a atividade de localização e consulta de fontes diversas de informação escrita, para coletar dados gerais ou específicos a respeito de determinado tema" (Carvalho *et al.*, 2007, p. 100).

Nesse sentido, os capítulos 2 a 6 visam apresentar a fundamentação teórica, por meio de uma pesquisa bibliográfica, sobre o tema mineração de dados.

O capítulo 2 apresenta uma visão geral sobre o processo de mineração de dados, os fatores históricos relacionados à sua origem e evolução e a motivação e os objetivos para sua aplicação. Também são apresentadas as tarefas, aplicações

práticas, etapas, ferramentas (softwares), procedimentos adotados, além das principais técnicas utilizadas na execução de todo o processo.

O capítulo 3 trata de procedimentos e técnicas que podem ser aplicados durante a etapa de preparação dos dados, conhecida como pré-processamento, para o alcance de melhores resultados com o processo de *data mining*.

Os capítulos 4 e 5 apresentam, mais detalhadamente, informações sobre os algoritmos Apriori e K-Means, respectivamente, pois são as técnicas de mineração de dados utilizadas e avaliadas neste trabalho.

No capítulo 6 é proposta uma metodologia para a aplicação completa do processo de mineração de dados para realização de uma busca automatizada de associações com o uso do algoritmo Apriori. O principal objetivo é oferecer orientações, especialmente quanto ao pré-processamento de dados, para que os resultados obtidos com o uso desse método em particular sejam otimizados.

O método de **procedimento** utilizado no trabalho é o **estudo de caso**. Segundo Severino (2007, p. 121) um estudo de caso se trata de uma pesquisa que se concentra no estudo de um caso particular, considerado representativo de um conjunto de casos análogos.

No capítulo 7 é apresentado um estudo de caso da aplicação da metodologia proposta em uma base de dados de uma unidade pública de saúde para verificar sua influência sobre os resultados do algoritmo Apriori.

Os resultados alcançados são analisados e discutidos no capítulo 8 e as conclusões apresentadas no capítulo 9.

2- MINERAÇÃO DE DADOS (DATA MINING)

2.1- Introdução

Nota-se a todo o momento pessoas que buscam, de alguma forma, registrar fatos, contatos, observações e tantas outras informações. Diversos recursos são utilizados, como agendas, diários, blocos de anotações e muitos outros. Ao que parece, existe uma sensação de que essas informações serão úteis futuramente e, por isso, devem estar organizadas de tal forma que possam estar prontamente disponíveis, quando necessário.

Para Witten *et al.* (2005, p. 4), as pessoas estão mergulhadas em dados. Computadores pessoais tornaram fácil a tarefa de guardar algumas coisas que antes seriam totalmente desconsideradas. Para a decisão sobre o que fazer com tudo isso há uma alternativa bem simples: guardar tudo em um disco de dados. Se necessário, basta adquirir um novo, pois não é caro. Assim, decisões e escolhas do dia-a-dia sobre compras, investimentos e passeios são registradas eletronicamente.

Desde 1960, tecnologias de informação e bancos de dados têm evoluído dos sistemas de processamento de um simples arquivo até os mais sofisticados e poderosos sistemas de banco de dados. Nos anos 70, já era possível gerenciar os dados em rede e surgiram os sistemas de bancos de dados relacionais com ferramentas e métodos ainda mais avançados para modelar e acessar os dados (Han *et al.*, 2006, p. 3). A figura 1, da página 27, demonstra o histórico da evolução das tecnologias de informação na visão de Han *et al.* (2006, p. 2).

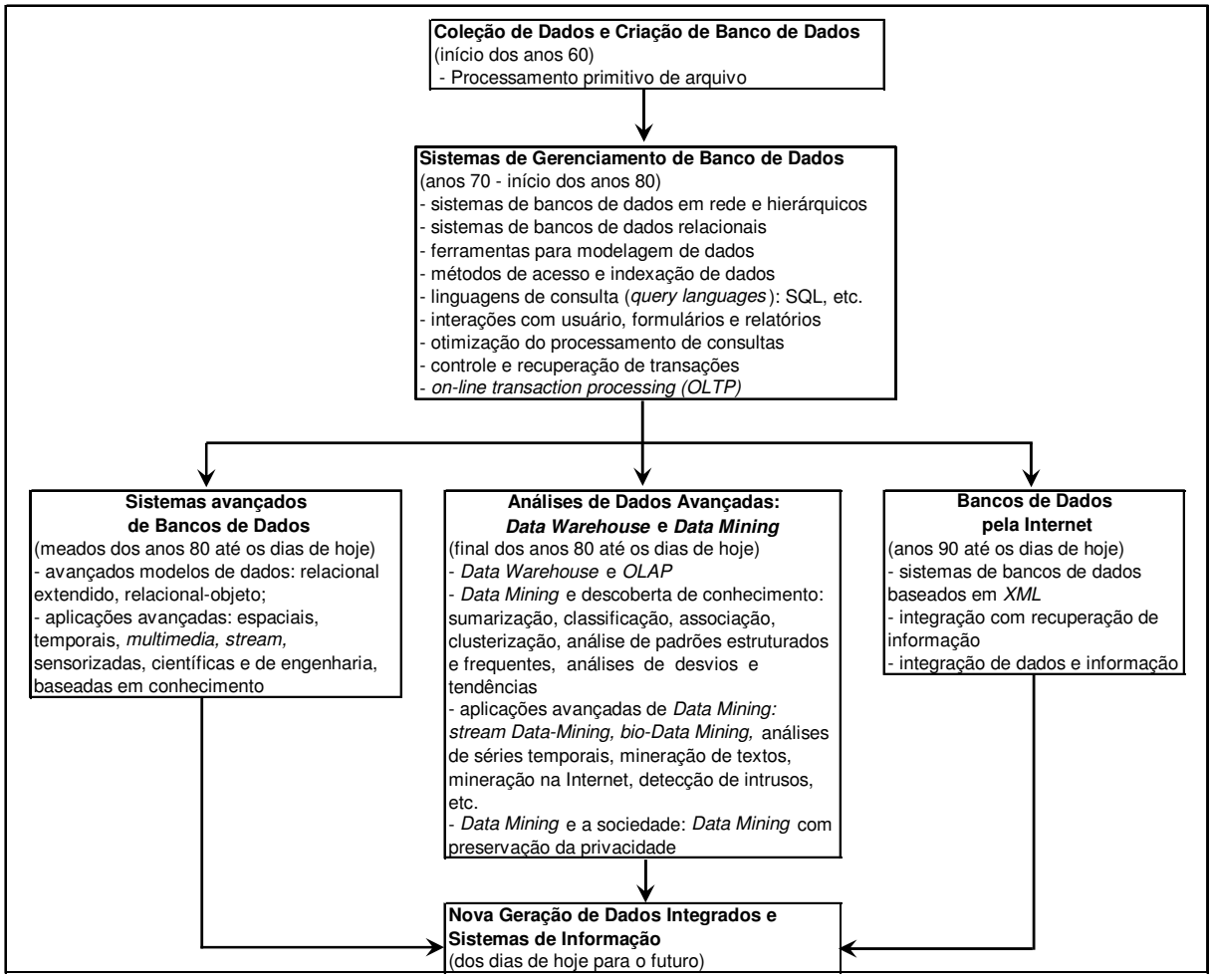


Figura 1. A evolução da tecnologia de sistema de banco de dados (Han *et al.*, 2006, p. 2, adaptado).

O incrível progresso de tecnologias da informação nas últimas três décadas contribuiu para uma maior oferta de supercomputadores e poderosos equipamentos de coleta e armazenamento de dados. Com o avanço das tecnologias de informação, tornou-se possível armazenar e organizar grandes quantidades de dados eletronicamente (Han *et al.*, 2006, p. 3). O volume de dados registrados atualmente é imenso, graças aos mais modernos recursos de softwares de bancos de dados e redes de dados e telecomunicações.

Observa-se, também, que o acesso e a manipulação dos dados estão mais otimizados, o que torna mais ágil e fácil a extração de informações e descoberta de conhecimentos. Sabe-se que já existem diversas tecnologias que oferecem recursos que tornam possível a construção de relatórios, dos mais simples aos mais complexos, com o máximo de informações e detalhamentos e sob as mais diversas dimensões e perspectivas. Diversas ferramentas permitem que analistas visualizem os dados de diferentes maneiras e são capazes de resumir grandes

quantidades de dados, com respostas rápidas a consultas (Silberschatz *et al*., 2006, p. 485). Com isso, existe a oportunidade para que um analista possa realizar as mais diversas análises com os dados disponíveis em um determinado banco de dados.

Como exemplo, as ferramentas *OLAP (On-Line Analytical Processing)* têm sido muito utilizadas para analisar grandes bancos de dados, pois apresentam facilidades para realização de consultas complexas em bases de dados multidimensionais. É um sistema interativo, capaz de atender às consultas de um analista dentro de alguns segundos (por isso o termo *on-line*) (Silberschatz *et al.*, 2006, p. 488), o que torna possível a análise de um universo de dados sob várias e diferentes perspectivas complexas e multidimensionais (Berson *et al.*, 2000, p. 69). Entretanto, essas ferramentas, "normalmente, são orientadas às consultas, ou seja, são dirigidas pelos usuários, os quais possuem hipóteses que gostariam de comprovar, ou simplesmente, executam consultas aleatórias" (Rezende *et al*., 2005, p. 308).

Para Bigus (1996, p. 5) as várias ferramentas de consulta ajudam apenas quando se sabe o que procurar. Certamente, as ferramentas mais modernas oferecem uma visão multidimensional dos dados e um processamento rápido que melhoram a capacidade de análise de dados, mas já não são suficientes se for considerado, por exemplo, o ambiente altamente competitivo que envolve o mundo dos negócios.

Por vários anos, o termo *data mining* esteve associado a vários tipos de abordagens para análise de dados. Muitas pessoas da indústria de software e empresários frequentemente se referiam à ferramenta *OLAP* como um principal componente da tecnologia de *data mining* (Kudyba *et al*., 2001, p. 23).

Mesmo que seja utilizada uma ótima ferramenta de consulta, ficam evidentes as limitações de uma análise realizada por um ser humano se o seu objeto for um grande volume de dados. É possível extrair padrões de dados com o uso de consultas *SQL* (*Structured Query Language*) em bancos de dados, no entanto, poderiam ser necessárias centenas ou até milhares de *queries* para explorar todas as combinações possíveis (Tang *et al.,* 2005, p. 230). Diante de uma enorme quantidade de possíveis combinações, associações e classificações, algumas delas podem passar despercebidas na análise e informações podem não ser identificadas.

> [...] Essa abordagem dependente do usuário pode impedir que padrões escondidos nos dados sejam encontrados de forma "inteligente", uma vez que o usuário não terá condições de imaginar todas as possíveis relações e associações existentes em um grande volume de dados.
> A evolução da computação possibilitou um aumento na capacidade de processamento e armazenamento de dados. A facilidade atual que uma aplicação científica ou comercial possui para gerar *gigabytes* ou *terabytes* de dados em poucas horas excede em muito a capacidade de pesquisadores e analistas de mercado em fazer análises sobre os mesmos. (Rezende *et al.*, 2005, p. 308)

"Muitas vezes, ferramentas e técnicas tradicionais de análise de dados não podem ser usadas devido ao tamanho do conjunto de dados ser muito grande" e por isso a extração de informação útil é um objetivo extramente desafiador (Tan *et al.*, 2009, p.1).

Para Adriaans *et al.* (1996, p. 6) as ferramentas de *data mining* e de consultas *on-line*, na verdade, são complementares. As ferramentas de mineração de dados não substituem totalmente as consultas tradicionais, mas oferecem possibilidades e capacidades adicionais.

Na prática, observa-se que muitas organizações, ainda que direcionem grandes esforços para armazenar e organizar grandes volumes de dados, não alcançam os possíveis benefícios decorrentes de todo esse empenho.

> [...] Durante a década passada, grandes volumes de dados vêm sendo acumulados e armazenados em bancos de dados. Muito desses dados procedem de softwares comerciais, tais como aplicações financeiras, Enterprise Resource Management (ERP), Customer Relationship Management (CRM), e diários virtuais. O resultado dessa coleção de dados é que organizações têm se tornado ricas em dados e pobres em conhecimento. As coleções de dados se tornaram tão vastas e estão aumentando tão rapidamente em tamanho que o uso prático desses armazéns de dados ficou limitado. [...] (Tang *et al.,* 2005, p. 2, tradução livre)[1]

A lacuna entre gerar e entender os dados tem aumentado. A quantidade de dados aumenta e, consequentemente, a capacidade das pessoas para entendê-los diminui. Estima-se que quantidade de dados armazenados no mundo dobra a cada vinte meses (Witten *et al.,* 2005, p. 4).

[1] Traduzido do texto original para o português: "[…] During the past decade, large volumes of data have been accumulated and stored in databases. Much of this data comes from business software, such as financial applications, Enterprise Resource Management (ERP), Customer Relationship Management (CRM), and Web logs. The result of this data collection is that organizations have become data-rich and knowledge-poor. The collections of data have become so vast and are increasing so rapidly in size that the practical use of these stores of data has become limited. […] " (Tang *et al.*, 2005, p. 2)

Segundo Bigus (1996, p. 4), esses dados não são mais vistos como apenas um produto do processamento de operações do dia a dia. Os dados dessas operações representam a situação atual dos negócios no momento, mas também há os dados históricos. Se combinados, é possível entender para onde se está indo e identificar o respectivo ponto de partida.

Toda essa enorme quantidade de dados supera a habilidade das pessoas para interpretar e refletir sobre esses dados, o que faz sugir a necessidade de uma nova geração de ferramentas e técnicas para análise automatizada e inteligente (Fayyad *et al.*, 1996, p. 1). Por isso, descobrir padrões, tendências e anomalias em grandes massas de dados é um dos grandes desafios da era da informação (Kantardzic *et al.*, 2005, p. 1).

Diante disso, a análise dessas grandes quantidades de dados precisa ser aprimorada para maximizar as oportunidades de extrair informações úteis e de descobrir conhecimentos que possam ser aplicados em situações práticas. Com esse propósito, também já é possível utilizar recursos da computação como um auxílio para realização desse tipo de análise. Uma alternativa é repassar todo o conhecimento necessário para a realização desse diagnóstico para uma máquina, já que não apresenta limitações no que se refere à investigação e processamento de grandes volumes de dados.

Programas de computador que podem aprender têm sido o foco de pesquisas em inteligência artificial desde o início da tecnologia computacional, por volta de 1950 (Adriaans *et al.*, 1996, p. 3).

“Os computadores têm algumas vantagens sobre os seres humanos, principalmente no que diz respeito à velocidade e consistência com que executam determinadas funções”, mas se tratam de processadores de símbolos e para torná-los capazes de desempenhar “uma tarefa tão bem quanto um especialista humano, alguém deve muni-lo de conhecimento especializado, comparável ao que um especialista humano possui” (Rezende *et al.*, 2005, p. 14).

Esse é um dos grandes desafios para os profissionais que atuam na área de tecnologia, qual seja o de criar, numa máquina, uma capacidade analítica própria de um ser humano para que se possa fazer associações, classificações e combinações de forma automatizada. Assim, um computador poderia “percorrer” todo um banco de dados e analisar dado por dado, em pouco tempo, se comparado

ao período que um ser humano levaria para fazer a mesma análise. Diante de uma análise tão minuciosa, pode-se descobrir informações importantes, porém ocultas.

Há bastante tempo, profissionais de diversas áreas de conhecimento trabalham com a idéia de que em um conjunto de dados podem existir padrões que não estejam explícitos, mas que podem ser encontrados, identificados e validados automaticamente e utilizados em tentativas de prever situações futuras. Os padrões úteis encontrados nos dão a condição de fazer previsões não triviais para os novos dados. O mundo gera dados ao mesmo tempo em que se desenvolve e se torna mais complexo e, com isso, uma análise de dados inteligente é um recurso muito valioso atualmente. Um grande objetivo da mineração de dados é armazenar dados eletronicamente e realizar buscas de forma automatizada, com o uso de um computador (Witten *et al.,* 2005, p. 4).

2.2- Definições de mineração de dados (data mining)

Pela definição de *data mining* apresentada por Witten *et al.* (2005, p. 5), o termo se refere ao "processo de descobrir padrões em dados", que pode ser "automático ou (mais usualmente) semiautomático".

Segundo Berry *et al.* (1997, p. 6), a mineração de dados só faz sentido quando há grandes volumes de dados. Por isso, os autores agregam essa condição em sua definição ao ensinar que *data mining* é a exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados com o objetivo de se descobrir padrões e regras significativos (Berry *et al.*, 1997, p. 5). Mais tarde, em outra de suas obras, Berry *et al.* (2000, p. 7) ratificam quase todo o contexto dessa definição, mas afirmam que não querem mais direcionar o foco do significado de *data mining* para os "meios automáticos ou semiautomáticos" utilizados e sim, para a efetiva exploração e análise de dados.

Nota-se que, além da ênfase para a aplicação da mineração de dados em "grandes quantidades de dados", que é comum ser encontrada nas definições de *data mining*, surgiu também a referência de que o objetivo da aplicação é descobrir padrões e regras que sejam "significativos".

No mesmo sentido, Silberschatz *et al.* (2006, p. 496) entende que "o termo mineração de dados (ou *data mining*) refere-se, em geral, ao processo de

analisar grandes bancos de dados de forma semi-automática para encontrar padrões úteis" e descobrir conhecimento.

Bigus (1996, p. 9) define *data mining* como a descoberta automatizada e eficiente de informação valiosa e que não seja óbvia em uma grande base de dados. O termo eficiente está relacionado ao custo e o benefício dessa descoberta ao considerar o valor da informação, que é determinado pela sua utilidade, principalmente em processos decisórios.

Por esta definição, a informação descoberta precisa ser valiosa e trazer benefícios que compensem os esforços aplicados no processo. É comum encontrar referências ao valor da informação descoberta, em definições de *data mining*.

Alguns estudiosos entendem que muitas informações, na verdade, estão "escondidas" entre os dados e é possível encontrá-las com a aplicação do processo de mineração de dados.

Segundo Tang *et al.* (2005, p. 2), *data mining* significa "analisar dados e encontrar padrões ocultos utilizando recursos automatizados ou semiautomatizados". A sua principal finalidade é "extrair padrões dos dados em mãos, aumentar seu valor intrínseco e transformá-los em conhecimento".

Para Elmasri *et al.* (2005, p. 624) é "a mineração ou a descoberta de novas informações em função de padrões ou regras em grandes quantidades de dados". Esses padrões, talvez, nem poderiam ser encontrados simplesmente com pesquisas e consultas na base de dados.

Garber (2001, p. 255) informa que "a idéia de que grandes massas de dados escondem valiosas informações" se intensifica e muitas ferramentas já foram desenvolvidas para identificar as informações escondidas nesses bancos de dados. É como imaginar "uma mina com vários diamantes escondidos, dentre outras pedras sem valor. Daí a denominação em inglês *data mining"*, que na verdade "é uma comparação explícita entre o processo de garimpagem e a busca de informações. Esta atividade também é chamada de *Knowledge Discovery in Database (KDD)."*

Han *et al.* (2006, p. 5), informam que muitas pessoas usam o termo *data mining* como sinônimo de *Knowledge Discovery in Databases (KDD)*, ou seja, Descoberta de Conhecimento em Banco de Dados, que é outro conceito popularmente muito utilizado. Alternativamente, outras pessoas entendem que *data mining* é apenas uma etapa muito importante dentro de todo um processo de

descoberta de conhecimento em banco de dados. Por essa perspectiva, *data mining* é apenas uma etapa presente na sequência de etapas de todo um processo.

Segundo Fayyad *et al.* (1996, p. 3), o termo *data mining* tem sido utilizado mais comumente por estatísticos, analistas de dados e pela comunidade de sistemas de informações gerenciais, enquanto que KDD é mais usado por pesquisadores de inteligência artificial e aprendizado da máquina. Para os autores, KDD "se refere a todo o processo de encontrar e interpretar padrões dos dados", enquanto *data mining* "se refere a uma classe de métodos que são usados" em alguma parte desse processo.

Adriaans *et al.* (1996, p. 5) também concordam que há uma confusão gerada pelas definições dos termos *data mining* e *KDD*. Muitos autores afirmam que são sinônimos. Adriaans *et al.* (1996, p. 5) informam que na primeira conferência mundial de *KDD*, no ano de 1995 em Montreal, foi proposto que o termo *KDD* seja empregado para descrever todo o processo de extração de conhecimentos a partir de dados, enquanto que o termo *data mining* deveria ser usado exclusivamente para o estágio de descoberta no processo de *KDD*.

Com esse entendimento, Passos *et al.* (2005, p. 2) apontam que a expressão mineração de dados (data mining) é mais popular, mas na realidade é uma das etapas da Descoberta de Conhecimento em Base de Dados (*KDD*). Assim, *data mining* é uma parte do processo de *KDD* (Elmasri *et al.,* 2005, p. 624; Fayyad *et al.*, 1996, p. 3).

Mais tecnicamente, "a mineração de dados é uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados" (Tan *et al.*, 2009, p.1). No processo, são utilizadas poderosas tecnologias analíticas para rapidamente explorar completamente toda essa "montanha" de dados para extrair informação valiosa e útil (Kantardzic *et al.*, 2005, p. 1).

Para Kantardzic *et al.* (2005, p. 1), *data mining* resolve o paradoxo de que quanto mais dados se tem, mais difíceis e mais demorados ficam uma efetiva análise e a extração de significados deles. Entretanto, Kudyba *et al.* (2001, p. 23) lembram aos usuários que *data mining* não é algum tipo de mágica computacional no estilo "caixa preta" ou "bola de cristal" que fornece conhecimentos.

2.3- Etapas da mineração de dados (data mining)

O processo completo de *data mining,* ou mesmo *KDD,* é constituído de uma seqüência de três etapas básicas: pré-processamento (ou preparação dos dados); mineração de dados; e pós-processamento (ou análise dos dados) (Passos *et al.,* 2005, p. 11; Bigus, 1996, p. 10).

A entrada do processo de *data mining* é uma série de registros de dados relacionados a fatos, pessoas, objetos, etc. Esses registros de dados também são chamados de exemplos ou instâncias. Cada registro é considerado um indivíduo independente ao se considerar o contexto da análise. Na série, cada indivíduo possui uma ou várias características predeterminadas, também chamadas de atributos. Assim, a série de registros de dados é representada por uma matriz de instâncias *x* atributos. Ao visualizar a série na forma de uma tabela, cada linha é uma instância (registro) e cada coluna é um atributo (característica) da respectiva instância (Witten *et al.,* 2005, p. 45). Isso porque os algoritmos de *data mining* requerem que os dados estejam apresentados em um formato tabular, em que as linhas representem o objeto da mineração de dados e as colunas descrevem as características dessas linhas (Berry *et al.*, 2000, p. 181).

Acontece que as bases de dados dos dias de hoje estão muito suscetíveis a apresentar sujeiras (erros e desvios), perdas e inconsistências, devido, principalmente, aos seus enormes tamanhos e pelo fato de seus dados terem origem de múltiplas e heterogêneas fontes. A baixa qualidade dos dados utilizados em um processo de *data mining* induz a baixa qualidade dos resultados da mineração (Han *et al.*, 2006, p. 47).

Muitos dos registros nas bases de dados estão incompletos ou apresentam desvios. Em uma base de dados, por exemplo, podem existir: atributos que são redundantes ou obsoletos; dados ausentes; valores não consistentes com a realidade ou senso comum; e formatos de dados não adequados para mineração de dados (Larose*,* 2005, p. 27).

Estes problemas são comuns em grandes bancos de dados e ocorrem devido a diversas razões (Han *et al.*, 2006, p. 48):

- alguns atributos não existem ou não estão disponíveis;
- outros não foram inseridos por não terem sido considerados importantes na época;

- dados relevantes podem não ter sido registrados por erros na sua interpretação ou até mesmo por falhas nos equipamentos;
- dados podem ter sido apagados por apresentarem conflitos com outros;
- o histórico de registros e modificações nos dados podem ter sido negligenciados;
- os instrumentos de coleta de dados utilizados não eram apropriados ou apresentavam defeitos;
- erros podem ter sido cometidos por pessoas ou programas de computador na entrada ou durante a transmissão de dados.

Por isso, inicialmente, os dados obtidos precisam ser organizados e tratados para que possam ser utilizados no processo de mineração de dados. Esta é a fase de pré-processamento, quando os dados são efetivamente preparados. Entre as principais funções dessa etapa estão (Passos *et al.*, 2005, p. 11):

- seleção de dados: também conhecida como redução de dados. Trata-se da identificação de quais informações, dentre as bases de dados existentes devem ser efetivamente consideradas durante o processo;
- limpeza dos dados: tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade dos fatos por eles representados, ou seja, garantir a sua veracidade e integridade, além da certificação de que estejam completos;
- codificação dos dados: os dados precisam estar formatados de forma que possam ser reconhecidos pelas ferramentas (softwares e algoritmos) que serão utilizadas no processo;
- enriquecimento dos dados: consiste em conseguir, de alguma forma, informação que possa ser agregada aos registros existentes para enriquecer os dados de maneira a contribuir com a descoberta de conhecimentos. Para isso, podem ser utilizadas bases de dados externas para obtenção de mais dados.

Uma análise prévia dos dados obtidos é extremamente importante para o sucesso da mineração de dados. Durante essa fase, muitas vezes se observa que não são todos os dados que devem ser utilizados no processo. A utilização de

determinados dados pode comprometer os resultados ou até mesmo inviabilizar a aplicação de técnicas de mineração de dados.

Vale destacar que nessa fase de pré-processamento é preciso considerar os objetivos a serem alcançados com a mineração de dados e conhecer bem a base de dados a ser manipulada. Nesse momento, o auxílio de um especialista com domínio no conhecimento a que se refere à base de dados já se torna muito importante para a seleção e purificação dos dados (Bigus, 1996, p. 11). Da mesma forma, durante o pré-processamento já devem ser considerados as tarefas, as técnicas, os algoritmos e os softwares de mineração de dados que poderão ser aplicados nas próximas etapas do processo.

Dependendo da tarefa e das ferramentas de mineração de dados que serão aplicadas, os dados precisarão estar apresentados sob determinados tipos de formatos para que um respectivo software, ao utilizar um algoritmo específico, possa realizar a leitura dos dados corretamente (Bigus, 1996, p. 12). Assim, torna-se necessário o conhecimento profundo sobre as ferramentas e tecnologias a serem aplicadas.

Um mesmo dado pode ser apresentado apenas (i) na forma de um termo ou categoria, que geralmente apenas classifica ou diferencia um registro, ou (ii) em um formato numérico, que normalmente expressa algum tipo de medida, valor monetário ou quantidade e, nesse caso, até pode ser utilizado em cálculos como média, proporções, distâncias, etc.

> Determinados algoritmos possuem restrições quanto aos tipos de variáveis existentes no conjunto de dados. Neste caso, duas alternativas podem ser consideradas: a) elimina-se do conjunto de algoritmos de Mineração de Dados todos aqueles que forem incompatíveis com os tipos de variáveis envolvidas no problema; ou b) opta-se por utilizar um determinado algoritmo de Mineração de Dados e realizar todo o pré-processamento sobre o conjunto de dados de forma a torná-lo compatível com o algoritmo desejado. (Passos *et al.*, 2005, p. 54)

Preparar os dados (entrada) para o processo de *data mining* geralmente consome a maior parte dos esforços aplicados no processo inteiro (Witten *et al.,* 2005, p. 52; Cabena *et al.*, 1998, p. 47), por outro lado, quando técnicas de pré-processamento são aplicadas antes da mineração de dados, melhora-se substancialmente a qualidade dos resultados ou até mesmo diminui-se o tempo que seria necessário para a execução de todo o processo de mineração (Han *et al.*, 2006, p. 47).

A etapa seguinte ao pré-processamento é a mineração de dados propriamente dita, ou seja, a busca efetiva por conhecimentos úteis. Nesta etapa, são aplicados os algoritmos, com auxílio de softwares, sobre os dados previamente tratados para realização de uma ou mais tarefas de mineração de dados para o alcance do resultado esperado (Passos *et al.,* 2005, p. 12).

Para Cabena *et al.* (1998, p. 55) é a fase que efetivamente ocorre a mineração de dados, já que os algoritmos de *data mining* são efetivamente aplicados nos dados pré-processados. Estes algoritmos "são fundamentados em técnicas que procuram, segundo determinados paradigmas, explorar os dados de forma a produzir modelos de conhecimento" (Passos *et al*., 2005, p. 52).

Após a mineração de dados, os resultados obtidos precisam ser avaliados para verificar se há algum conhecimento a eles associado que possa ser extraído e aplicado. Por isso há a etapa de pós-processamento, que é o tratamento do conhecimento de forma a torná-lo mais claro (para a sua melhor interpretação) e mais fácil de ser aplicado.

Nem sempre esse tratamento é necessário, pois os resultados já podem estar relativamente claros. Ainda assim, uma melhor organização das idéias em gráficos, tabelas, diagramas e relatórios pode muito contribuir para a representação do conhecimento obtido (Passos *et al.,* 2005, p. 15). Do mesmo modo pode estimular "a percepção e a inteligência humana e aumentar a capacidade de entendimento e associação de novos padrões" (Passos *et al*., 2005, p. 58). É também nesta etapa que se define novas alternativas de investigação dos dados (Passos *et al*., 2005, p. 55).

Por outro lado, tecnologias de *data mining* podem encontrar muitos padrões ou regras em uma base de dados, no entanto, nem todas são "interessantes". Segundo (Han *et al*., 2006, p. 27), os padrões são interessantes quando são: facilmente entendidos por seres humanos; válidos, ao considerar um grau mínimo de acertos; potencialmente úteis; e novos, ou seja, até então não haviam sido descobertos. Para Witten *et al.* (2005, p. 5) os padrões descobertos devem ser significativos e alcançar alguma vantagem, preferencialmente uma vantagem econômica.

Larose (2005, p. 11) argumenta que a execução da mineração de dados deve ser transparente, ou seja, os resultados precisam descrever claramente os padrões encontrados, de formar a tornar fácil e intuitiva a sua interpretação. O autor

informa, ainda, que alguns métodos utilizados em *data mining* possibilitam uma interpretação mais simples, intuitiva e agradável de seus resultados, enquanto que outros, por serem mais complexos, não são compreendidos facilmente por pessoas que não são especialistas no assunto.

Tão importante quanto descobrir um conhecimento é a tarefa de representar e descrever este conhecimento de forma que possa ser entendido pelas pessoas. Isso é muito útil para explicar o conhecimento que foi descoberto e o embasamento para previsão de novas situações. O aprendizado se completa quando se adquire o conhecimento e, ao mesmo tempo, a capacidade de utilizá-lo (Witten *et al.*, 2005, p. 9).

Nesse sentido, a etapa de pós-processamento é imprescindível, já que de nada adiantaria todo o esforço de processamento e análise de dados e de uso de técnicas e ferramentas de mineração de dados se os resultados alcançados não puderem ser compreendidos, utilizados e aplicados em situações práticas e todo esse conhecimento obtido puder ser disseminado.

Pode-se, então, visualizar as etapas do processamento de *Data MIning* conforme a figura 2, apresentada a seguir.

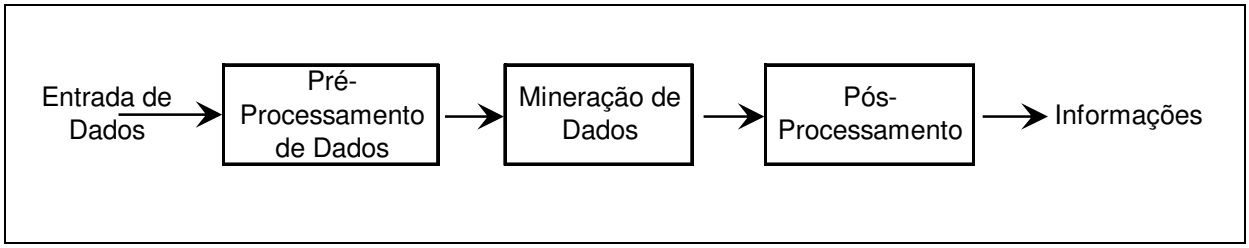


Figura 2. Etapas do processo de *data mining*.

Embora haja uma sequência de etapas no processo de mineração de dados, essa trajetória não é linear, ou seja, durante a execução, pode-se retornar a uma ou mais etapas anteriores, caso seja necessário para a continuidade do processo ou isso possa melhorar os resultados já encontrados (Adriaans *et al.*, 1996, p. 37).

A mineração de dados é um processo iterativo, por isso algumas atividades de cada etapa podem ser realizadas novamente e repetidamente após a análise dos resultados gerados em outra, principalmente para ajustes e adequações no processo para o alcance dos objetivos previamente estabelecidos (Rezende *et al.*, 2005, p. 317).

2.4- Tarefas de mineração de dados (data mining)

Há outras formas para se descobrir informações úteis. Atualmente, existem ferramentas capazes de fazer consultas altamente complexas e refinadas em bancos de dados. Também são fantásticos os mecanismos de buscas disponíveis na Internet para descobrir páginas que tratam de determinado assunto. Entretanto, vale destacar, como bem ensinam Tan *et al.* (2009, p.3), que muitas atividades, embora possam descobrir informações, não são consideradas tarefas de mineração de dados:

> [...] Por exemplo, a procura de registros individuais usando um sistema gerenciador de banco de dados ou a busca de determinadas páginas da Web através de uma consulta em um mecanismo de busca na Internet são tarefas relacionadas à área de **recuperação de dados**. Embora sejam importantes e possam envolver o uso de algoritmos e estruturas de dados sofisticadas, essas tarefas se baseiam em técnicas tradicionais da ciência da computação e em recursos óbvios dos dados para criar estruturas de índice para organizar e recuperar de forma eficiente as informações. Contudo, a mineração de dados tem sido usada para melhorar sistemas de recuperação de informações. (grifo do original)

As tarefas de mineração de dados estão correlacionadas, principalmente, à finalidade da busca na base dados e o resultado esperado.

Entre as principais tarefas de mineração de dados estão (Passos *et al.,* 2005, p. 14):

- descoberta de associação: abrange a busca que tenha como principal finalidade a de descobrir se há dois ou mais itens na base de dados que frequentemente apresentam ocorrências simultâneas, o que pode indicar uma associação entre esses itens ou até a influência de uns sobre outros;
- descoberta de sequências: é uma extensão da tarefa de associação, utilizada basicamente para as mesmas finalidades. Entretanto, a tarefa considera, além da frequência em que ocorrem as coincidências, um período determinado e a sequência em que os itens se apresentam;
- classificação: a busca visa encontrar um padrão que possa explicar as classificações que se encontram previamente

designadas para os registros da base de dados. É uma tarefa bastante aplicada, pois é muito comum a definição de classes para separar ou identificar registros em bancos de dados, como os de clientes, produtos e contas, durante o uso de sistemas de informação para auxílio à execução de determinadas atividades;

- regressão: é uma tarefa similar a de classificação, no entanto, se baseia apenas em atributos numéricos (valores reais e contínuos), relacionados aos registros, para a indicação de um padrão que possa explicar a classificação dos registros;
- clusterização[2]: trata-se de uma tentativa de encontrar semelhanças e similaridades entre as características e propriedades de alguns registros da base de dados que os tornem diferentes dos demais. A principal finalidade é verificar a possibilidade de separar e agrupar os registros em *clusters* ou subconjuntos de tal forma que seus elementos sejam similares, mas, ao mesmo tempo, apresentem características distintas de outros componentes dos demais *clusters*. Nota-se que, diferentemente do que ocorre na tarefa de classificação, os registros não estão previamente classificados. No caso da clusterização, os *clusters* identificados podem ser analisados e posteriormente classificados conforme suas principais características;
- sumarização: essa tarefa consiste em procurar, identificar e indicar quais as características que são comuns em determinados conjuntos de dados. É também denominada descrição de conceitos. Comumente, a sumarização é aplicada em cada um dos *clusters* identificados na clusterização para verificar quais são as características que lhes são peculiares;
- detecção de desvios: a finalidade principal é verificar se há registros no banco de dados que não atendam aos padrões considerados normais para o contexto analisado. Diferentemente

[2] O termo "clusterização", na verdade, significa agrupamento e deve ter sido escolhido pelos autores como referência em português para a mesma palavra "*clustering"* em inglês, utilizada para denominar a "técnica de *data mining* para fazer agrupamentos automáticos de dados" (Wikipédia, 2011).

> das demais tarefas, em que a repetição de padrões é uma característica fundamental na busca por conhecimento, a detecção de desvios procura identificar padrões com pouca incidência e que sejam suficientemente distintos dos valores normalmente registrados.

A análise de associações consiste no processo de avaliação das probabilidades de um evento específico ocorrer ao considerar a ocorrência de outros eventos (Kudyba *et al.*, 2001, p. 11). A descoberta de associação busca encontrar ligações entre registros ou série de registro em um banco de dados (Cabena *et al.*, 1998, p. 68).

Um evento que ocorre muito frequentemente é considerado um padrão frequente, como os itens que aparecem juntos em diversos registros. Os padrões frequentes são indícios de possíveis associações (Han *et al.*, 2006, p. 23).

Os padrões encontrados referentes a associações entre dois ou mais itens, por exemplo, podem ser resultantes de algum tipo de ligação entre eles que precisa ser investigada e esclarecida. Talvez, haja algum padrão de comportamento dos agentes responsáveis pela inserção dos dados ou fatores relacionados à operação que originou os registros que expliquem essa ligação. Passos *et al.* (2005, p. 14) apresentam um exemplo clássico e didático de descoberta de associações através da mineração de dados:

> [...] uma grande rede de mercados norte-americana descobriu que um número razoável de compradores de fralda também comprava cerveja na véspera de finais de semana com jogos transmitidos pela televisão. Com uma análise mais detalhada sobre os dados, pode-se perceber que tais compradores eram, na realidade, homens que, ao comprarem fraldas para seus filhos, compravam também cerveja para consumo enquanto cuidavam das crianças e assistiam aos jogos na televisão durante o final de semana. Este exemplo ilustra a associação entre fraldas e cervejas. Esta empresa utilizou o novo conhecimento para aproximar as gôndolas de fraldas e cervejas na rede de mercados, incrementando assim a venda conjunta dos dois produtos.

Uma base de dados pode apresentar muitas associações e, por isso, esses resultados precisam ser filtrados pelas maiores frequências e níveis precisão. Em seguida, deve-se examiná-las cuidadosamente para verificar se são significativas ou não (Witten *et al.,* 2005, p. 43). O fato de dois eventos ocorrerem simultaneamente não garante que a possível relação entre eles seja importante ou significativa (Westphal *et al.*, 1998, p. 189).

Também é muito comum verificar em bases de dados registros que estão classificados sob determinados rótulos, geralmente designados durante a utilização dos respectivos sistemas de informação. Clientes podem, por exemplo, estar identificados em duas classes como: adimplentes e inadimplentes. Isto porque durante as várias operações realizadas em um sistema de cobrança esses rótulos foram atribuídos aos clientes conforme seu comportamento em relação aos pagamentos realizados. Como a tarefa de classificação é um processo que busca encontrar um modelo que descreva as diferentes classes de dados predeterminadas (Elmasri *et al.*, 2005, p. 634), um propósito dessa tarefa poderia ser a descoberta de um modelo (uma sistemática) capaz de identificar a razão porque clientes de uma instituição financeira geralmente estão adimplentes ou inadimplentes. A descoberta poderia se basear nos diversos dados coincidentes (padrões) presentes na maioria ou em grande parte dos registros históricos referentes às situações de adimplência e inadimplência.

Para Han *et al.* (2006, p. 24) a classificação é o processo que visa encontrar um modelo que descreve e diferencia os registros atuais conforme suas classificações. A proposta é usar esse modelo para prever a classe de um futuro registro. O maior estímulo para o uso de *data mining* é a sua capacidade de construir modelos que possibilitam prever situações futuras, principalmente porque auxiliam em tomada de decisões (Berson *et al.*, 2000, p. 33).

Assim, a tarefa de classificação tem como objetivo prever, sob a premissa de que cada registro de dados pertence a uma das várias classes previamente definidas, a classe de um novo registro, que ainda não é conhecida. Para isso, é feita uma análise de seus atributos em comparação com os dos registros anteriormente analisados (Silberschatz *et al.*, 2006, p. 497).

No caso de classificações adotadas para clientes, um padrão pode ser identificado através de diagnósticos de grande frequência em que ocorrem coincidências ou proximidades em determinados dados, como: renda, despesas, tipo de residência, sexo, quantidade de filhos, etc. Existem muitas possibilidades de descobertas de novo conhecimento sobre padrões de compra quando se analisa esses tipos de variáveis (Elmasri *et al.*, 2005, p. 625). Pela análise das diversas operações ocorridas no passado, pode-se, por exemplo, estabelecer uma sistemática que possa mostrar como foi identificada a maioria ou grande parte dos

clientes inadimplentes e utilizá-la, inclusive, para projetar as chances de inadimplência para as futuras negociações.

Nesse mesmo sentido, Witten *et al.* (2005, p. 42) ensinam que a classificação é utilizada quando a base de dados apresenta séries de dados já classificados e o principal objetivo é descobrir quais os critérios, até então desconhecidos, foram utilizados na definição de cada classe.

A tarefa de regressão, também conhecida como estimativa, é similar a classificação, mas é utilizada quando os atributos em estudo são numéricos. A tentativa de prever novas situações se baseia nos valores históricos dos mesmos atributos em períodos anteriores (Larose*,* 2005, p. 12). “A regressão lida com a previsão de um valor, em vez de uma classe” (Silberschatz *et al.*, 2006, p. 501).

Para essa tarefa, são utilizados indicadores de probabilidades, que na verdade são as chances de algum evento previsto ocorrer. Um exemplo dessa tarefa seria calcular a probabilidade de um cliente responder a uma oferta promocional (Cabena *et al.*, 1998, p. 65).

Em muitos casos, ainda que não se perceba, muitos registros de um banco de dados apresentam características muito semelhantes, de forma que poderiam ser separados em grandes grupos. Uma varredura nas compras realizadas por vários clientes durante um longo período, por exemplo, poderia demonstrar que quase todos eles se enquadrariam em algum dos perfis de compra identificados pela proximidade de fatores relacionados ao valor total da compra, quantidade de itens, utilização de créditos, tipos de produtos, entre outros.

O objetivo principal do processo de *clustering* (clusterização) é dividir os registros em grupos, de tal forma que os registros de um grupo sejam similares aos demais do mesmo grupo e diferentes daqueles de outros grupos (Elmasri *et al.*, 2005, p. 637; Kogan *et al.*, 2006, p. 127).

A clusterização consiste na busca em grandes bases de dados para identificar e distinguir diferentes grupos através da verificação de variáveis que são estatisticamente similares para um mesmo conjunto de dados (Kudyba *et al.*, 2001, p. 10).

Segundo Witten *et al.* (2005, p. 43), a clusterização é utilizada quando não há classes predeterminadas, mas os registros parecem se agrupar naturalmente. O sucesso dessa tarefa é medido subjetivamente pela clareza e utilidade dos resultados gerados para um usuário.

A tarefa de clusterização não busca classificar, estimar ou prever novas situações, mas segmentar a base de dados em estudo, dividindo-a em grupos homogêneos (Larose, 2005, p. 16). É a busca pelo agrupamento de registros, ou casos, em classes de objetos similares. Diferentemente da classificação, por exemplo, não há classes predeterminadas (Larose, 2005, p. 147).

Frequentemente, a tarefa de clusterização é realizada como uma etapa preliminar de um processo de *data mining*, ou seja, os seus resultados (grupos) são submetidos à outra tarefa de mineração de dados, com diferentes técnicas, para melhor avaliação. Até mesmo pelos enormes tamanhos das atuais bases de dados dos dias de hoje, a clusterização pode ser uma alternativa para a escolha de dados que serão submetidos ao processo de mineração de dados (Larose, 2005, p. 148).

Por exemplo, já que foram identificados grupos muito coesos, no que se refere às características de seus elementos, uma boa atitude seria a investigação sobre o que torna os registros tão semelhantes. Uma tentativa de sumarização poderia mostrar, por exemplo, que clientes de determinado perfil de compra são geralmente do sexo feminino, cursam alguma faculdade atualmente e têm idade entre 20 e 24 anos.

A sumarização busca resumir as características gerais de uma classe de dados (Han *et al.*, 2006, p. 22). É uma tarefa que “envolve métodos para encontrar uma descrição compacta” de uma parte da base de dados (Fayyad *et al.*, 1996, p. 15).

Em algumas situações, o objetivo da mineração de dados é simplesmente descrever o que acontece em uma complicada base de dados, de uma forma a melhorar o entendimento sobre as pessoas, produtos e processos que produziram esses dados (Berry *et al.*, 1997, p. 55; Berry *et al.*, 2000, p. 11).

Há, ainda, casos em que especialistas de determinado domínio de conhecimento conseguem definir uma forma de estabelecer um limite do que seria uma situação normal em determinado contexto. Com isso, uma busca minuciosa em uma base de dados poderia detectar os desvios que ocorreram em algumas situações, para que sejam investigados. Uma operadora de cartão de crédito poderia definir uma forma de traçar um perfil de compra para cada cliente e ao mesmo tempo identificar aquelas negociações que estiverem fora do padrão encontrado.

Um desvio é a variação em relação a alguma expectativa ou algum comportamento conhecidos (Cabena *et al.*, 1998, p. 69). Também pode ocorrer pela

presença de valores extremos, que geralmente estão posicionados próximos aos limites que denotam o intervalo dos dados, ou quando há dados que não condizem com a tendência ou direção estabelecida pelos demais (Larose, 2005, p. 34).

A análise para detecção de desvios é realizada com a finalidade de buscar casos raros que apresentam comportamentos muito diferentes dos demais. Também busca identificar se houve alguma mudança "brusca" de comportamento no mesmo caso observado anteriormente. É muito utilizada em detecção de fraudes em instituições financeiras (Tang *et al.,* 2005, p. 10).

Desvios podem ser descobertos pela avaliação das diferenças encontradas entre os valores esperados ou considerados normais (principais características dos objetos de um grupo, por exemplo) e os presentes na base de dados (Han *et al.*, 2006, p. 458).

Nota-se que a escolha da tarefa depende do resultado final esperado. Conforme ensinam Witten *et al.* (2005, p. 43), mesmo que uma base de dados ofereça todas as condições para realização de uma classificação, talvez seja possível optar, também, pela busca de associações.

Vale destacar, que a base de dados deve ser preparada conforme a tarefa escolhida. Por exemplo, o processamento e as respectivas regras resultantes da tarefa de associação envolvem, geralmente, apenas dados não numéricos e por isso a base de dados precisa ser analisada e preparada para esse tipo de tarefa. Talvez, haja a necessidade de conversão ou retirada de alguns dados para a execução do processo.

2.5- Técnicas de mineração de dados (data mining)

A mineração de dados não é por si só uma técnica. Qualquer técnica que ajude a extrair informações de dados é útil; por isso, técnicas de *data mining* formam um grande grupo heterogêneo. Nesse sentido, várias técnicas diferentes são utilizadas para diferentes propostas e contextos (Adriaans *et al.*, 1996, p. 47).

"As técnicas de mineração de dados são organizadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis e recentes que poderiam, de outra forma, permanecer ignorados" (Tan *et al.*, 2009, p. 3).

Essas técnicas são implementações de algoritmos que são utilizados para executar operações de *data mining* (Cabena *et al.*, 1998, p. 63).

Uma técnica de mineração de dados se refere a qualquer teoria que possa fundamentar a implementação de determinado método durante o processo. Nesse sentido, um método é a implementação de uma operação específica de mineração de dados e corresponde a um algoritmo em particular (Passos *et al.*, 2005, p. 17).

O grande desafio é o desenvolvimento de metodologias capazes de lidar com milhares de atributos e milhões de registros (Kantardzic *et al.*, 2005, p. 3).

Data mining envolve uma integração de técnicas de diversas disciplinas do conhecimento, como: tecnologias de banco de dados; estatística; aprendizado da máquina; inteligência artificial; computação de alto desempenho; reconhecimento de padrões; redes neurais; recuperação e visualização de dados; processamento de imagens e sinais; e análise espacial e temporal de dados (Han *et al.*, 2006, p. 9; Kantardzic *et al.*, 2005, p. 1).

Nesse sentido, Tan *et al.* (2009, p. 7) também informam que pesquisadores de diversas disciplinas e áreas do conhecimento estão envolvidos em estudos para o desenvolvimento de técnicas de mineração de dados. Nos estudos, são aplicadas diversas idéias e metodologias, como: amostragem, estimativa e teste de hipóteses a partir de estatísticas; algoritmos de busca; técnicas de modelagem; teorias de informação, de aprendizagem da inteligência artificial, de reconhecimento de padrões e de aprendizagem de máquina; otimização; computação evolutiva; processamento de sinais; visualização e recuperação de informações.

Elmasri *et al.* (2005, p. 625 e p. 641) também asseguram que, além das técnicas de áreas como aprendizado de máquinas, estatística, redes neurais, lógica *fuzzy,* inteligência artificial e algoritmos genéticos que podem ser utilizadas em mineração de dados, há, ainda, as regras e árvores de decisão, que são técnicas baseadas em procedimentos que buscam encontrar e representar regras e padrões estruturados em bases de dados durante a realização de tarefas como classificação ou clusterização.

Apenas para exemplificar o conceito de técnica de mineração de dados, pode-se citar a técnica de indução de árvore de decisão como um exemplo, já que é bastante utilizada na tarefa de classificação (Silberschatz *et al.*, 2006, p. 498).

Um método de classificação, por exemplo, é atrativo quando envolve a construção de uma “árvore de decisão” que represente os padrões encontrados, entre os registros de dados, que possibilitam classificar esses registros. Na verdade, trata-se de uma coleção de “nós de testes”. Pode ser representada, inclusive, por um diagrama que, comparativamente, possui a aparência de uma árvore. Essa árvore pode ter vários “nós” conectados por “ramos”, desde a “raiz” (nó raiz) até suas folhas (nós de folha), onde é o seu término. Em cada nó, é realizado um teste ou tomada de alguma decisão. Por convenção, a raiz, que é a decisão (ou teste) inicial, fica no topo do diagrama de árvore de decisão (como se uma árvore estivesse de cabeça para baixo). A partir da raiz surgem os ramos (nós filhos), que são as decisões (testes) intermediárias. Os testes são realizados em cada ramo, que podem desenvolver outros (filhos), ou seja, a necessidade de novos testes. Assim, a árvore cresce até chegar ao seu tamanho máximo, onde estão as suas folhas. As pontas da árvore estão no último nível de decisão. De fato, as folhas não representam mais decisões ou testes e, sim, os resultados (Larose, 2005, p. 107).

Apenas como exemplo, Larose (2005, p. 107) ilustra uma árvore de decisão construída a partir da análise da classificação de clientes sob a perspectiva de risco para concessão de empréstimos, conforme a figura 3 a seguir.

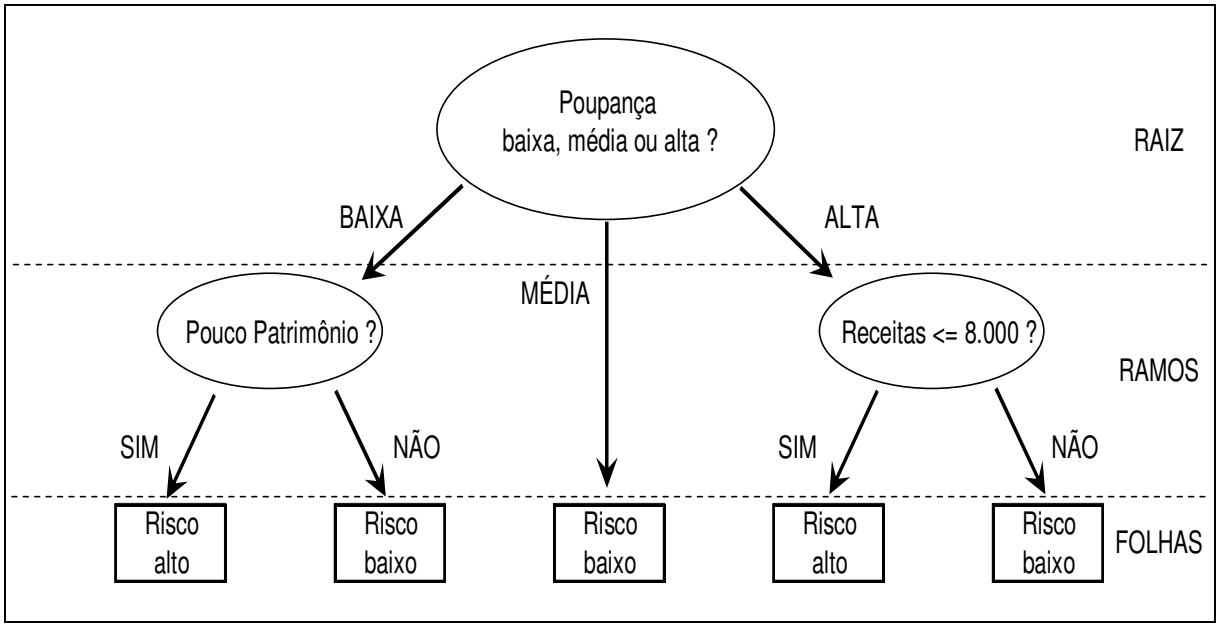


Figura 3. Exemplo de um diagrama de árvore de decisão simples (Larose, 2005, p. 107, adaptado).

A técnica de indução de árvore de decisão é muito eficiente em termos de tempo de processamento e fornece um método muito intuitivo para análise dos resultados (Cabena *et al.*, 1998, p. 72).

Um dos aspectos mais atrativos do uso de árvores de decisão é a facilidade para sua interpretação, especialmente quanto à criação de regras de decisão. Uma regra de decisão pode ser construída simplesmente ao percorrer um caminho e realizar os devidos testes desde a raiz até uma folha (Larose, 2005, p. 121).

O grande desafio, então, é como construir a árvore, isto é, como buscar os padrões dos dados e criar as conexões entre as regras de decisão que possibilitam realizar as classificações. Vários métodos e cálculos podem ser implementados com essa finalidade.

Ao final de 1970 e início de 1980, um pesquisador de aprendizado da máquina, J. Ross Quinlan, desenvolveu um algoritmo de indução de árvores de decisão que ficou conhecido como ID3 (do inglês, *Iterative Dichotomiser*). Mais tarde, o mesmo pesquisador apresentou o algoritmo C4.5, como o sucessor do ID3. O C4.5 se tornou uma referência e é utilizado como parâmetro para desenvolvimento e comparação de novos algoritmos de classificação por indução de árvores de decisão. Na mesma época, um grupo de estatísticos, entre eles L. Breiman, J. Friedman, R. Olshen e C. Stone, desenvolveram o algoritmo CART (*Classification and Regression Trees*). Embora tenham sido desenvolvidos de forma independente, os algoritmos se baseiam na mesma abordagem de “cima para baixo” de criar árvores de decisão através de uma recursiva iteração do tipo “dividir e conquistar”. (Han *et al.*, 2006, p. 292).

Embora a técnica de indução de árvore de decisão seja uma das técnicas mais recentes utilizadas em *data mining*, após o seu surgimento em meados dos anos 80, outras técnicas, principalmente as baseadas em estatísticas, já são estudadas por décadas, algumas até por séculos, e agora aplicadas como ferramenta de mineração de dados (Tang *et al.,* 2005, p. 11).

> É importante notar que as técnicas descrevem um paradigma de extração de conhecimento e vários algoritmos podem seguir esse paradigma. [...]
> Entre as técnicas mais utilizadas em Mineração de Dados [...] estão as Regras e Árvores de Decisão, as Redes Neurais que apesar de não gerarem conhecimento explícito são empregadas para diversas tarefas de Mineração de Dados, aplicação de Algoritmos Genéticos que fazem parte

> da computação evolutiva, e Lógica Fuzzy. A combinação dessas técnicas, que constitui os Sistemas Híbridos também tem obtido êxito na Extração de Conhecimentos em Base de Dados. (Rezende *et al.*, 2005, p. 327)

Segundo Passos *et al.* (2005, p. 17) existem diversos tipos de técnicas e de algoritmos para mineração de dados e esses tipos podem ser classificados em técnicas tradicionais, técnicas específicas e técnicas híbridas.

Para os autores, as técnicas tradicionais são aquelas que utilizam tecnologias que existem independentemente do contexto da mineração de dados, ou seja, também são aplicadas em outros contextos e com outras finalidades. Entretanto, produzem bons resultados quando utilizadas em algoritmos de *data mining*. Alguns exemplos dessas tecnologias são:

- redes neurais: uma técnica computacional que constrói um modelo matemático inspirado em um sistema neural biológico simplificado, com capacidade de aprendizado, generalização, associação e abstração. Através de repetidas apresentações de dados às redes, elas buscam aprender padrões automaticamente, ou seja, por sua "própria" experiência. Isso porque uma rede neural procura por relacionamentos, constrói modelos matemáticos automaticamente e os corrige de modo a diminuir seu próprio erro;
- lógica nebulosa (*fuzzy logic*): É uma técnica que permite construir sistemas que lidem com informações imprecisas ou subjetivas. Os sistemas computacionais tradicionais tratam os dados por meio de uma lógica mais exata e linear. Diferentemente da lógica clássica, a lógica nebulosa oferece flexibilidade na definição e na avaliação de conceitos;
- algoritmos genéticos: são modelos computacionais inspirados na evolução natural e na genética, aplicados a problemas complexos de otimização. Buscam, a partir de uma solução inicial, evoluir para a melhor solução possível de um determinado problema, com base nas transformações observadas na natureza em seus processos de evolução;
- estatística: diversos modelos matemáticos fornecem opções para análise, exploração e interpretação de dados.

As técnicas específicas são desenvolvidas com o propósito de realizar tarefas de mineração de dados. O algoritmo conhecido por Apriori, proposto por R. Agrawal e R. Srikant em 1994 (Han *et al.*, 2006, p. 235; Witten *et al.,* 2005, p. 141), é um exemplo de técnica específica, pois foi desenvolvido com o intuito de realizar a tarefa de descoberta de associação. Do mesmo modo, o algoritmo K-Means, criado por J. B. MacQueen, em 1967, é uma técnica muito popular – tem sido muito utilizado na prática –, desenvolvida especificamente para realização de tarefas de clusterização (Kogan *et al.*, 2006, p. 128; Berry *et al.*, 1997, p. 192).

As técnicas híbridas utilizam mais de uma técnica no processo de mineração de dados. Uma grande vantagem se deve ao sinergismo obtido pela combinação de duas ou mais técnicas de modelagem, o que torna a análise mais poderosa, no que se refere à interpretação, aprendizado, estimativa de parâmetros, generalização, dentre outros aspectos, e com menos deficiências.

Vale destacar, que mesmo que vários algoritmos sejam utilizados para uma determinada tarefa de mineração de dados, isto não significa que sejam iguais, pois cada um pode apresentar pontos fortes e fracos (Cabena *et al.*, 1998, p. 63).

2.6- Execução do processo de mineração de dados (data mining)

Segundo Adriaans *et al.* (1996, p. 79) a premissa para realização de um processo de mineração de dados é a de que existem mais informações escondidas na base de dados, além das que foram identificadas em uma primeira análise. Na verdade, neste processo é possível extrair quatro níveis de conhecimento à medida que a análise dos dados se aprofunda:

- superficial: as informações podem ser facilmente extraídas com uma simples consulta;
- multi-dimensional: a extração requer o uso de ferramentas que possibilitem de forma mais rápida a realização de consultas mais complexas para visualização e análise dos dados sob diversas dimensões de agrupamentos, ordenamentos e associações;
- escondido: as informações podem ser encontradas com o uso de algoritmos capazes de reconhecer padrões. Também poderiam ser

encontradas com a utilização de ferramentas de consulta, mas possivelmente o tempo consumido seria bem maior;

- profundo: é o tipo de informação que existe na base de dados que somente é encontrada quando surge um indício ou vestígio que possa direcionar a análise para esse descobrimento. Talvez uma mudança do espaço de busca ou até mesmo da própria técnica de investigação seja determinante.

A mineração de dados é normalmente utilizada para o alcance de algum objetivo, que pode ser a descoberta de alguma aplicação prática ou descrição do conhecimento a ser obtido (Elmasri *et al.*, 2005, p. 625). As principais perguntas a serem respondidas são: o que se quer saber e o que se quer fazer com este novo conhecimento (Adriaans *et al.*, 1996, p. 81).

Antes de se iniciar um processo de mineração de dados, primeiramente é preciso pensar em como extrair o máximo de informações desse processo. Uma boa prática a ser adotada em um processo de *data mining* seria identificar os tipos de padrões que se espera encontrar antes mesmo do primeiro registro de dados ser processado (Westphal *et al.*, 1998, p. 25).

A execução do processo de mineração de dados deve ser iniciada com um exame criterioso da base de dados. A partir desta análise e de entrevistas junto a especialistas no domínio da aplicação são definidos os objetivos a serem alcançados ao longo da execução do processo. As atividades podem ser orientadas, por exemplo, para validar alguma hipótese inicialmente postulada ou tentar descobrir algum conhecimento pela análise dos dados existentes. Por outro lado, a expectativa quanto aos resultados pode estar relacionada à possibilidade de criação de um modelo que permita, a partir de um histórico de dados, prever ou estimar valores de atributos em novas situações ou simplesmente possa descrever o conhecimento existente naquela base de dados (Passos *et al.*, 2005, p. 15).

Os objetivos geralmente estão inseridos nas principais classes de objetivos da mineração de dados que podem ser (Rezende *et al.*, 2005, p. 317):

- preditivas: como as classes de classificação e regressão;
- descritivas: como as classes de associação, clusterização e sumarização.

Veja-se que o objetivo estabelecido é determinante para a definição da tarefa de mineração de dados a ser realizada, ao observar que TAN *et al.* (2009, p.

8) adotam a mesma classificação para as tarefas de mineração de dados, que, posteriormente, são escolhidas de acordo com cada objetivo, dividindo-as em duas categorias:

- tarefas de previsão: têm o objetivo de prever o valor de um atributo com base nos valores de outros. A classificação e a regressão são dois tipos de tarefas de previsão;
- tarefas descritivas: o objetivo é o de derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumam os relacionamentos dos dados.

Também, Han *et al.* (2006, p. 21) ensinam que as tarefas descritivas caracterizam as propriedades gerais dos dados enquanto que as preditivas estabelecem inferências sobre os dados com o objetivo de se fazer previsões.

Assim como acontece com qualquer implementação de tecnologias de informação, se o processo de *data mining* não for aplicado corretamente, os resultados gerados, juntamente com todo o investimento realizado, serão inúteis. A chance de uso inadequado de *data mining* é potencialmente ainda maior que de outros tipos de aplicações de análise de dados, já que envolve tecnologias de natureza mais complexa (Kudyba *et al.*, 2001, p. 43).

Kudyba *et al.* (2001, p. 43) recomendam, ainda, algumas dicas sobre como conduzir o processo de mineração de dados:

- minerar dados corretos. Um processo deficiente de coleta de dados direcionará o trabalho para o fracasso. Os dados precisam conter informações corretas e relevantes. Nem todas as grandes bases de dados estão adequadas para a mineração de dados. A relevância da informação está associada ao objetivo a ser alcançado, ou seja, o que se quer fazer com os dados;
- pensar primeiro, minerar depois. Trata-se de uma boa preparação antes de iniciar efetivamente os procedimentos. Alguns passos simples podem determinar o sucesso da aplicação, como: decidir o que se quer saber; escolher a forma de avaliar os resultados, se estão bons ou ruins; inspecionar os dados visualmente, por meio de gráficos; revisar os resultados;

- não torturar os dados para obter a resposta que se quer. Nem todas as tentativas de mineração de dados geram bons resultados. Os fracassos geralmente se devem aos dados coletados e não a condução do processo. Tentativas de usar ferramentas, algoritmos ou softwares distintos não alcançarão resultados muito diferentes;
- suspeitar de estatísticas perfeitas. Os resultados com altos índices de certeza, devem ser analisados com cautela. Nesses casos, é preciso, também, avaliar os dados coletados.

Tang *et al.* (2005, p. 13) demonstram a execução de um projeto de *data mining* em oito passos:

1- coleta de dados: reunião de dados relevantes presentes em um ou nos mais diversos bancos de dados disponíveis para formar a base de dados para a utilização no processo. Algumas empresas, por exemplo, podem ter até centenas de bancos de dados descentralizados para armazenamento de fatos relacionados às suas operações. Outras, já possuem todos os dados disponíveis reunidos em uma única base de dados. Ainda assim, dados de fontes externas podem também ser coletados, como informações demográficas e econômicas, disponibilizadas por instituições especializadas, que podem contribuir para o melhor desempenho dos resultados;

2- limpeza e transformação dos dados: remoção de “sujeiras” e informações irrelevantes. A transformação dos dados, às vezes se faz necessária para o processo de mineração de dados, como a alteração do formato de sua apresentação;

3- planejamento e construção do modelo de mineração de dados: inicialmente há a definição dos objetivos e das tarefas da mineração de dados a serem realizadas. Torna-se necessária a participação de um ou mais especialistas com domínio do conhecimento relacionado a esses objetivos. Em seguida, há a escolha dos algoritmos mais apropriados ao objetivo e à tarefa de mineração de dados;

4- avaliação do modelo: trata-se da análise do desempenho do modelo construído. Tecnologias podem ser utilizadas para testar o modelo e avaliar os resultados gerados. Mais de um algoritmo podem ter sido

utilizados e, por isso, os seus resultados podem ser comparados. Essa avaliação também abrange a análise que verifica se os padrões de dados encontrados possuem algum significado. Pode-se concluir, então, sobre o nível de precisão (desempenho) do modelo e até mesmo sobre a qualidade da base de dados em termos de gerar informações significativas. Assim, os passos anteriores podem ser novamente realizados, assim como novos testes, caso o modelo apresente um nível de desempenho considerado baixo. Da mesma forma, a análise dos resultados por especialistas com grandes conhecimentos do contexto avaliado pode contribuir muito para a validação dos resultados;

5- relatório: formatação para a apresentação dos resultados. É preciso que os conhecimentos encontrados estejam sistematizados e apresentados de forma que facilite o entendimento das pessoas interessadas que, talvez, nem conheçam bem o processo de mineração de dados, mas se beneficiarão dos resultados gerados;

6- previsões: parte-se da idéia de que encontrar padrões de dados é apenas uma parte do processo. O objetivo final é utilizá-los para se fazer previsões de novas situações. Deve-se criar uma sistemática para realização constante de testes, com base nesses padrões, na tentativa de prever novas situações ao mesmo tempo em que se verifica se as previsões anteriores se concretizam com a entrada de novos dados;

7- integração do modelo a operações e atividades: o modelo pode ser integrado aos atuais sistemas de informação utilizados e até passar a fazer parte de alguma etapa operacional ou decisória dentro de um contexto de planejamento ou plano de ação;

8- gestão do modelo: ajustes do modelo em consequência de mudanças. Em determinados casos, os padrões de dados se mantêm mais estáveis, enquanto que em outros, em decorrência de mudanças nos sistemas de informação e nos cenários e ambientes de execução, o modelo precisa ser aperfeiçoado para lidar com novos padrões. Conclui-se, então, que um modelo tem uma duração limitada, já que uma nova versão precisa ser criada frequentemente.

Nota-se que esses passos de execução estão em sintonia e se encaixam nas etapas da mineração de dados, previamente apresentadas: pré-processamento (1 e 2), mineração de dados (3 e 4) e pós-processamento (5 a 8).

A tecnologia é essencial para o processo de *data mining*, no entanto, não significa que é suficiente para o seu sucesso. O sucesso depende muito mais de um foco bem estabelecido e do empenho do elemento humano que qualquer outro aspecto, inclusive a tecnologia (Cabena *et al.*, 1998, p. 90).

Observa-se que seres humanos estão envolvidos em muitas, se não todas as atividades e fases de todo o processo (Fayyad *et al.*, 1996, p. 41). A figura 4, a seguir, demonstra bem a visão de Fayyad *et al.* (1996, p. 42).

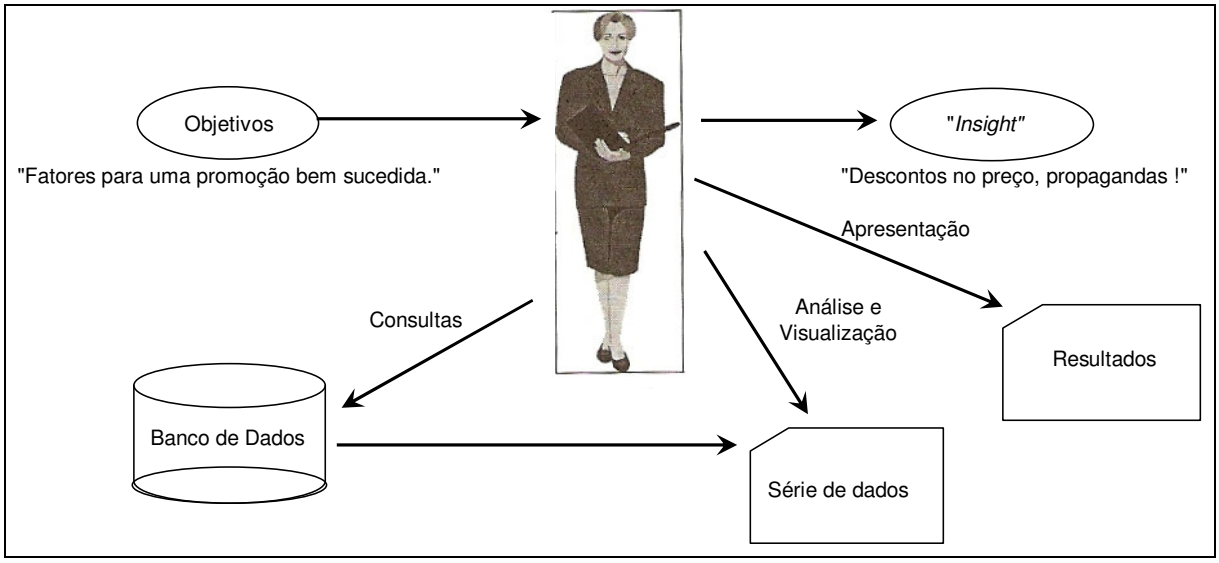


Figura 4. Visão das tarefas do analista (Fayyad *et al.*, 1996, p. 42, adaptado).

2.7- Ferramentas para aplicação de mineração de dados (data mining)

O processo de mineração de dados é relativamente complexo, pois além de envolver fatores operacionais como manipulação de grandes volumes de dados, tratamento dos resultados e aplicação de algoritmos específicos, há também fatores relacionados ao controle e condução das atividades a serem realizadas (Passos *et al.*, 2005, p. 20).

> O processo de Extração de Conhecimentos em Base de Dados é facilitado consideravelmente se for usada uma ferramenta que ofereça suporte para uma variedade de técnicas, com diferentes algoritmos disponíveis e voltadas para várias tarefas de Mineração de Dados.

> O progresso da área de Extração de Conhecimentos em Base de Dados e sua utilização nos mais variados domínios e pelas mais diversas organizações têm motivado o desenvolvimento de várias ferramentas comerciais além da elaboração de muitos protótipos de pesquisa. (Rezende *et al.*, 2005, p. 328)

Atualmente, existem diversas ferramentas (softwares) disponíveis que facilitam a execução do processo de mineração de dados e minimizam, principalmente, os problemas decorrentes dos fatores operacionais. Em geral, essas ferramentas reúnem vários métodos e técnicas de mineração de dados que podem ser aplicados em diversas tarefas (Passos *et al.*, 2005, p. 20).

Por haver diversos profissionais de diferentes disciplinas de conhecimento envolvidos em pesquisas na área de mineração de dados, é esperada a existência de uma grande variedade de ferramentas e sistemas de *data mining* (Han *et al.*, 2006, p. 29).

O poder da tecnologia da ferramenta a ser utilizada é realmente uma parte essencial em um projeto de mineração de dados. Os aspectos mais importantes de uma ferramenta de *data mining* são a facilidade para preparação de dados, a disponibilidade de algoritmos, a sua escalabilidade, o seu desempenho e a praticidade para visualização de resultados (Cabena *et al.*, 1998, p. 90).

Para escolher o sistema de *data mining* mais apropriado para uma tarefa, é importante ter uma visão multidimensional dessas ferramentas. Em geral, as ferramentas devem ser avaliadas com base nos seguintes recursos (Han *et al.*, 2006, p. 660; Berry *et al.*, 1997, p. 425):

- tipos de dados: quais os formatos de dados que o sistema é capaz de processar;
- requisitos de sistema: em quais sistemas operacionais a ferramenta funciona bem; arquiteturas de bancos de dados exigidas; entre outros;
- fontes de dados: em quais fontes de dados o sistema é capaz de buscar dados para realização da tarefa de mineração de dados;
- funções e métodos de mineração de dados: essas funções formam o núcleo de uma ferramenta de *data mining*, e por isso é preciso verificar as técnicas disponíveis no sistema;

- ligação com sistemas de banco de dados ou de *Data Warehouse*[3]: alguns sistemas de banco de dados ou de *Data Warehouse* já oferecem recursos de *data mining* neles integrados. Dessa forma, algumas ferramentas são comercializadas apenas em conjunto, como em um pacote fechado, enquanto que outras podem ser adquiridas separadamente;
- escalabilidade: a capacidade de a ferramenta processar maior ou menor quantidade de registros e atributos;
- ferramentas de visualização: formas de visualização dos dados através de gráficos, tabelas e diagramas;
- interface gráfica com o usuário: recursos gráficos para facilitar o uso do sistema pelo usuário.

Alguns exemplos de sistemas de *data mining* (Han *et al.*, 2006, p. 663) estão listados na tabela 1, da página 58:

[3] Um *Data Warehouse* ou "armazém de dados", ou ainda depósito de dados, é um sistema de computação utilizado para armazenar informações relativas às atividades de uma organização em bancos de dados, de forma consolidada. O desenho da base de dados favorece os relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão.
[...]
Os *Data Warehouse* surgiram como conceito acadêmico na década de 80. Com o amadurecimento dos sistemas de informação empresariais, as necessidades de análise dos dados cresceram paralelamente. Os sistemas transacionais não conseguiam cumprir a tarefa de análise com a simples geração de relatórios. Nesse contexto, a implementação do data warehouse passou a se tornar realidade nas grandes corporações. O mercado de ferramentas de data warehouse, que faz parte do mercado de Business Intelligence, cresceu e ferramentas melhores e mais sofisticadas foram desenvolvidas para apoiar a estrutura do data warehouse e sua utilização.
[...]
O *Data Warehouse* possibilita a análise de grandes volumes de dados, coletados dos sistemas transacionais. [...] Por definição, os dados em um data warehouse não são voláteis, ou seja, eles não mudam, salvo quando é necessário fazer correções de dados previamente carregados. Os dados estão disponíveis somente para leitura e não podem ser alterados.
Atualmente, por sua capacidade de sumarizar e analisar grandes volumes de dados,o data warehouse é o núcleo dos sistemas de informações gerenciais e apoio à decisão das principais soluções de business intelligence do mercado (Wikipédia, 2011).

Ferramenta de *data mining*	Desenvolvedor	Tarefas que podem ser realizadas
Intelligent Miner	IBM	associação, classificação, regressão, detecção de desvios, clusterização.
Microsoft SQL Server	Microsoft	associação, classificação, regressão, detecção de desvios, clusterização. Suporta integração com algoritmos desenvolvidos por terceiros.
MineSet	Purple Insight	várias funções de mineração de dados, inclusive associação e classificação. Seu ponto forte é uma série robusta de ferramentas gráficas.
Oracle data mining	Oracle	associação, classificação, regressão, clusterização.
Clementine	SPSS	associação, classificação, clusterização e ferramentas de visualização.
Enterprise Miner	SAS Institute	associação, classificação, regressão, clusterização, análise de séries temporais e pacotes de análises estatísticas.
Insightful Miner	Insighful Inc.	limpeza de dados, classificação, clusterização e pacotes de análises estatísticas com ferramentas de visualização.
CART	Saltford Systems	Classificação e regressão.
See5 e C5.0	RuleQuest	classificação por indução de árvores de decisão.
Weka	Universidade de Waikato, situada na Nova Zelândia (código livre)	desenvolvido em java com uma série de algoritmos, funções e tarefas de *data mining*, como: pré-processamento, associação, classificação, regressão, clusterização e visualização.

Tabela 1. Exemplos de sistemas de *data mining* (Han *et al.*, 2006, p. 663).

2.8- Aplicações práticas da mineração de dados (data mining)

Para Witten *et al.* (2005, p. 4), não há nada de novo no que se refere à busca por padrões existentes em dados. Pessoas têm procurado padrões em dados desde o início da vida humana. Os comportamentos dos seres humanos, animais e plantas são observados em diferentes situações. O trabalho de um cientista, por exemplo, é verificar o sentido dos dados disponíveis para descobrir padrões que influenciam em como o mundo físico funciona e reunir essas informações para

formular teorias que possam ser usadas para prever o que pode acontecer em novas situações. Nesse sentido, o processo de mineração de dados tem o objetivo de “resolver problemas pela análise de dados já presentes em banco de dados”.

Ainda, segundo Witten *et al.* (2005, p. 26), a maioria das aplicações de *data mining* está direcionada para as áreas de *marketing* e vendas. As instituições bancárias foram as primeiras a adotar a mineração de dados. Atualmente, os bancos têm utilizado essa ferramenta para detectar mudanças nos padrões individuais dos clientes, que podem estar relacionadas, por exemplo, a alterações em seu estilo de vida, como a transferência para outra cidade. A mineração de dados pode indicar os grupos de clientes para os quais determinados serviços são mais apropriados, assim como aqueles que oferecem maior lucratividade.

A mineração de dados pode ser aplicada em qualquer situação em que haja razoáveis volumes de dados históricos sobre algum assunto (Passos *et al.*, 2005, p. 159).

Passos *et al.* (2005, p. 159) ainda apresentam aplicações de mineração de dados vivenciados pelos próprios autores. Alguns exemplos estão, resumidamente, apresentados a seguir:

- um projeto em uma grande empresa do ramo de telecomunicações teve como principal objetivo classificar clientes de acordo com seu potencial de compra de serviços. Inicialmente, foi selecionada uma amostra no banco de dados de clientes para que pudessem preencher um questionário fornecido pela empresa. Foi realizado um processo de clusterização na amostra e, com as classes definidas, foi gerado um modelo classificador que foi aplicado na base de dados completa para verificar o potencial de compras de todos os clientes. Com isso, ações de marketing específicas foram realizadas conforme o perfil de cada classe de clientes;
- na área social, foi realizado um projeto que tinha como principal objetivo auxiliar as atividades de reintegração de pessoas de rua no Estado do Rio de Janeiro. Em síntese, o governo do Estado acolhe pessoas de rua e as submete a diferentes programas de reintegração social, conforme o seu perfil. Foram utilizadas as bases de dados sobre os perfis dessas pessoas juntamente com os resultados de reintegração de cada programa. A mineração de

dados teve como meta caracterizar o perfil dessas pessoas para viabilizar um processo melhor de triagem, com o objetivo de direcioná-las aos programas de reintegração mais adequados ao seu perfil;

- na educação, uma grande base de dados gerada com respostas de todas as escolas do Estado do Rio de Janeiro a mais de 600 perguntas referentes a sua gestão no ano de 2001. A aplicação da mineração de dados teve como objetivo buscar caracterizar perfis de escolas de forma a descobrir, dentre várias questões, por que determinadas escolas têm uma procura maior que outras, por que algumas escolas têm alto índice de evasão, e assim por diante. Com isso, o governo pode estudar medidas para solucionar os problemas detectados;
- na área financeira, um projeto teve como principal objetivo gerar um modelo classificador para identificar os clientes que pagam em dia, os que pagam em atraso e os que nem pagam os seus créditos. Foi considerado um histórico de pagamentos de uma instituição financeira durante um período de tempo definido. O padrão encontrado foi incorporado a um sistema especialista que funciona como apoio à decisão para análise de novas solicitações de empréstimos recebidas na central de atendimento.

Rezende *et al.* (2005, p. 469) demonstram uma aplicação de mineração de dados em um conjunto de dados de cerca de cinco mil clientes de uma companhia de telefonia celular, com base em vinte e um atributos relacionados a suas atividades e características. Como objetivo principal, buscava-se uma classificação para os clientes. Foram empregadas técnicas de árvores de decisão, árvores de regressão e redes neurais. As metodologias aplicadas permitem, por exemplo, identificar o perfil de clientes que devem receber um tratamento especial pela empresa com o objetivo de evitar que esses consumidores troquem o serviço prestado pela companhia por outro de um concorrente, assim como podem auxiliar na construção de um ranking de estratégias de marketing para decisão sobre aquelas que devem ser executadas. Ao comparar os resultados encontrados com a aplicação de cada técnica, concluiu-se que estavam muito próximos. Dessa forma, a técnica considerada como mais indicada para o caso foi a árvore de decisão, pela

maior simplicidade de uso e por apresentar menor custo computacional para uma precisão relativamente adequada, além de estar disponível em várias ferramentas de mineração de dados. Como aplicação prática dos conhecimentos obtidos, a análise dos resultados mostrou que se houvesse um catálogo de promoções, 90% entre os 10% dos melhores clientes responderiam satisfatoriamente a essas promoções.

Elmasri *et al.* (2005, p. 640) informam que as tecnologias de mineração de dados podem ser aplicadas em grande variedade de contextos de tomada de decisão, como:

- *marketing* – análises do comportamento do consumidor baseadas em padrões de consumo;
- finanças – análise de crédito de clientes e detecção de fraudes;
- produção – otimização de processos de fabricação;
- saúde – descoberta de padrões em imagens radiológicas e análise de efeitos colaterais de remédios e da efetividade de certos tratamentos.

Tarefas de *data mining,* como a classificação, por exemplo, podem ser aplicadas (Kudyba *et al.*, 2001, p. 30):

- no *marketing* – para definir quais clientes há maior expectativa para responder a um campanha de *marketing*;
- nas estratégias de vendas – para identificar clientes que estão mais propensos a comprar determinado produto;
- na detecção de fraudes – para monitorar comportamentos de potenciais fraudadores em operações de cartões de crédito e contratos de seguros.

Nesse mesmo sentido, Bigus (1996, p. 16), em sua obra, demonstra, com vários exemplos que há muitas aplicações de *data mining* nas áreas de *marketing*, finanças, produção, saúde e medicina, entre outras.

Além de mostrar exemplos de aplicações nessas principais áreas, Han *et al.* (2006, p. 654) ainda demonstram aplicações para análise de dados biológicos, que são muito importantes em pesquisas que trazem grandes benefícios para a área de saúde. Os autores também afirmam que há experiências bem sucedidas de

várias outras aplicações em áreas como meteorologia, astronomia, ciências geográficas, engenharia química e mecânica, entre outras.

3- PROCEDIMENTOS E TÉCNICAS DE PRÉ-PROCESSAMENTO DE DADOS

Para Tan *et al.* (2009, p. 4), o propósito da etapa de pré-processamento é transformar os dados em um formato apropriado para análises subsequentes do processo de *data mining* e abrange atividades para a realização de fusão de múltiplas fontes de dados, a limpeza dos dados para remoção de ruídos e retirada de informações duplicadas e a seleção de registros e características que sejam relevantes à tarefa de mineração de dados. Ainda, segundo os autores, talvez seja a etapa mais trabalhosa e demorada de todo o processo, já que os dados podem ser coletados ou terem sido armazenados de diversas formas.

Preparar os dados para um processo de *data mining* frequentemente consome boa parte dos esforços investidos em todo o processo, pois é muito comum os dados disponíveis para análise serem de baixa qualidade e um processo de avaliação criteriosa toma muito tempo (Witten *et al.,* 2005, p. 52).

No início de um trabalho de mineração de dados, um dos primeiros procedimentos a serem adotados é juntar os dados envolvidos em uma única série de registros (Witten *et al.,* 2005, p. 52).

Segundo Passos *et al.* (2005, p. 26), inicialmente, os dados precisam estar organizados em uma única e, possivelmente muito grande, estrutura tabular bidimensional, pois a maioria dos métodos de mineração de dados pressupõe que os dados estejam estruturados em uma tabela. Neste momento já há uma etapa de seleção de dados, pois os autores ensinam que a extração dos dados da base de dados transacional para esta tabela pode ocorrer de duas formas:

- junção direta: todos os atributos e registros da base de dados transacional são incluídos nesta nova tabela, sem uma análise crítica quanto a que variáveis e casos podem realmente contribuir para o processo;
- junção orientada: são escolhidos apenas os atributos e os registros com algum potencial para influir no processo de análise. É importante a parceria com um especialista com conhecimento profundo do contexto de origem dos dados, pois devem ser desconsiderados apenas os atributos e registros sobre os quais se

> tenha uma visão clara da inexistência de potencial de contribuição para o processo.

Observa-se que o processo de seleção pode ter dois enfoques distintos: a escolha de atributos ou a escolha de registros a serem considerados no processo (Passos *et al.*, 2005, p. 27).

A seleção de registros é caracterizada pela escolha de casos, isto é, apenas uma parte dos dados é utilizada. Pode ocorrer quando há interesse de se analisar apenas algum segmento de dados ou uma amostra aleatória ou avaliar os dados sem a presença de determinados registros (Passos *et al.*, 2005, p. 27).

Já a seleção de atributos tem o objetivo de retirar ou substituir os atributos do conjunto de dados de tal forma que a informação original seja preservada (Passos *et al.*, 2005, p. 29).

Entre os principais procedimentos para a seleção de atributos destaca-se a eliminação direta. Trata-se da retirada dos atributos cujo conteúdo não seja relevante ao processo. Essa eliminação depende do conhecimento prévio sobre o objetivo e o contexto de todo o processo. Basicamente são eliminados os atributos com valores constantes nos registros de dados e aqueles que são apenas identificadores, isto é, com valores usados apenas para identificar unicamente cada registro da base de dados (Passos *et al.*, 2005, p. 31).

Para a seleção dos dados deve-se verificar quais informações presentes na base de dados podem efetivamente ser úteis e não comprometer os resultados. Como o principal objetivo da mineração de dados é encontrar padrões de dados associados a conhecimentos, não se pode, por exemplo, selecionar dados que simplesmente individualizam e identificam determinada pessoa ou objeto na base dados.

Ao imaginar uma análise de riscos realizada por uma operadora de seguros de vida é possível verificar que dados cadastrais como nomes e documentos não são considerados e nem sequer influenciam no cálculo de uma respectiva apólice de seguro de vida. Nesse sentido, em um projeto de mineração de dados destinado a atender a esse tipo de análise, dados desse tipo seriam, da mesma forma, desconsiderados, pois, além de não influenciar na análise de riscos, raramente apresentam algum tipo de padrão útil (Passos *et al.*, 2005, p. 11).

A seleção de dados também é conhecida por redução de dados, pois, na prática, o que ocorre é a retirada de atributos e registros que não contribuirão para o

processo de mineração de dados ou, em determinados casos, é necessária para diminuir o tamanho da base de dados (Passos *et al.*, 2005, p. 11).

É muito incomum uma situação em que os dados disponíveis para aplicação de um processo de mineração de dados tenham sido coletados e armazenados já com esse propósito (Berry *et al.*, 1997, p. 67; Witten *et al.,* 2005, p. 59). No mundo real, os dados não estão prontos para um processo de *data mining*. Muitas transformações são necessárias para preparar os dados em um formato para a mineração de dados (Berry *et al.*, 2000, p. 181).

> Normalmente, os dados disponíveis para análise não estão em um formato adequado para Extração de Conhecimento [...]. Dessa maneira, torna-se necessária a aplicação de métodos para tratamento, limpeza e redução do volume de dados antes de iniciar a etapa de Extração de Padrões.
> É importante salientar que a execução das transformações deve ser guiada pelos objetivos do processo de extração a fim de que o conjunto de dados gerado apresente as características necessárias para que os objetivos sejam cumpridos. (Rezende *et al.*, 2005, p. 314)

O principal objetivo da transformação de dados é modificá-los para diferentes formatos em termos de tipos de dados e valores (Tang *et al.,* 2005, p. 13). As transformações de dados necessárias e específicas dependem da técnica escolhida e do software utilizado como ferramenta para aplicação do processo de *data mining.* Algumas ferramentas, por exemplo, precisam que todas as variáveis contínuas sejam divididas em intervalos, já outras exigem que todos os valores sejam normalizados em um intervalo específico entre 0 e 1 (Berry *et al.*, 1997, p. 67). Por exemplo, pode-se dividir os dados sobre idades em cinco grupos de idades pré-definidos e também é comum o uso de técnicas de normalização para transformar e mapear todos dados numéricos de um atributo em um número de 0 a 1 para assegurar que os valores mais altos não exerçam domínio sobre os mais baixos (Tang *et al.,* 2005, p. 14).

Frequentemente, surgem situações em que os dados necessários à análise não estão disponíveis ou mesmo não existem ainda, mas podem ser gerados para serem utilizados como informação complementar durante uma análise. Dados extras podem contribuir muito com valores mais significativos para o processo de mineração de dados. O processo de *data mining* é iterativo. Novos dados podem ser envolvidos em qualquer etapa deste processamento. Isto porque durante os avanços no processo de busca por padrões, comumente se nota que há certas informações que estão ausentes (Westphal *et al.*, 1998, p. 2).

Usualmente, torna-se necessário enriquecer os dados com atributos adicionais derivados de outros existentes, principalmente nos casos de busca de possíveis relacionamentos existentes entre dados de diferentes atributos ou de padrões de dados que aparecem nos registros ao mesmo tempo. Alguns atributos podem ser criados com base na experiência ou conhecimento sobre relações existentes entre variáveis em determinadas situações. Um atributo pode ser criado através da realização de cálculos que envolvam valores de vários atributos. Por exemplo, uma densidade demográfica pode ser calculada pela divisão da população pela respectiva área geográfica. Um atributo do tipo “idade” pode ser criado a partir de um atributo referente à data de nascimento de uma pessoa. Ao adicionar novos atributos, aumentam-se as chances para que o processo de descoberta de conhecimento gere resultados mais significativos (Berry *et al.*, 1997, p. 75; Berry *et al.*, 2000, p. 162; Passos *et al.*, 2005, p. 49).

Pode acontecer, principalmente se forem criados muitos atributos novos, que se tenha muitas variáveis que se referem praticamente a mesma informação, mas não é recomendável que haja variáveis fortemente correlacionadas. Os algoritmos de mineração de dados tipicamente precisam de apenas uma variável para identificar alguma característica específica entre os dados. Assim, apenas uma deve ser selecionada (Berry *et al.*, 2000, p. 162). É muito comum a substituição dos atributos originais pelos respectivos atributos derivados deles (Passos *et al.*, 2005, p. 49).

Muito da arte da mineração de dados se refere à criação de novos atributos (Berry *et al.*, 2000, p. 181). Mesmo outras fontes de dados podem ser utilizadas para se obter dados adicionais que possam contribuir para o sucesso dos resultados da mineração de dados. Esses novos dados devem ser escolhidos cuidadosamente (Passos *et al.*, 2005, p. 12).

Ao manipular dados referentes a indivíduos e objetos e suas características com os valores expressos exatamente como observado na realidade, é importante ter em mente que informações podem ser abstraídas para um nível mais baixo de detalhamento. Pode-se combinar as situações que se enquadram em características similares, como pesos, datas e lugares. Além de possibilitar uma visão dos intervalos que representam as diversidades existentes entre os dados, este tipo de abstração permite o processamento de mais dados em um processo de mineração de dados. Esta técnica é muito útil em processos analíticos em que se

busca extrair um entendimento inicial de classes de relacionamentos e padrões existentes em uma série de dados (Westphal *et al.*, 1998, p. 35).

Em algumas situações, há muitos valores distintos em um atributo e torna-se necessário agrupar esses valores em poucos grupos para reduzir a complexidade da análise (Tang *et al.,* 2005, p. 14) e a abstração pode reduzir o número de valores distintos em determinados atributos, o que pode proporcionar um melhor desempenho a diversos algoritmos de mineração de dados. Com menos valores, menos comparações são feitas e o tempo de processamento desses algoritmos tende a ser menor (Passos *et al.*, 2005, p. 33).

Nas buscas por associações, por exemplo, se há muitos itens distintos, Tang *et al.* (2005, p. 237) recomendam agrupá-los em categorias.

Valores de atributos, especialmente os numéricos, podem ser agrupados em categorias. Estes agrupamentos ocorrem mais frequentemente em séries de dados com valores discretos e contínuos, mas podem, na verdade, ser realizados em qualquer tipo de dados. Os tamanhos dos intervalos dependem dos valores constantes na série de dados, mas o número de categorias não pode ser pequeno e nem tão grande. Assim, se for estabelecido um tamanho maior para os intervalos, poucos serão os grupos formados, entretanto, aumenta-se o risco para que padrões importantes não sejam encontrados. Pode-se, então estabelecer um critério de agrupamento que seja o mais razoável e natural possível (Westphal *et al.*, 1998, p. 36).

O mapeamento de intervalos, também conhecido como discretização, pode ser realizado através de alguns procedimentos que podem ser adotados para a divisão dos valores de um atributo numérico em intervalos, com o uso de alguns critérios, como (Passos *et al.*, 2005, p. 41):

- divisão em intervalos com comprimentos definidos pelo usuário;
- divisão em intervalos com igual comprimento;
- divisão em intervalos por meio de um processo automatizado de clusterização.

Observa-se, então, que já na própria etapa de pré-processamento de dados pode ser realizada a tarefa de clusterização, com o uso de técnicas próprias e específicas para essa tarefa. Dessa forma, a divisão dos dados em grupos obedece a um critério já estabelecido naturalmente entre os próprios dados.

Este mesmo princípio também pode ser aplicado em dados qualitativos. Há muitos casos em que há níveis naturais de abstração entre os dados, tal como acontece com as informações sobre regiões geográficas, ou seja, a divisão já é reconhecida e aceita por conta de estudos geográficos (Westphal *et al.*, 1998, p. 37).

Outra situação muito comum é observar que muitos dos atributos disponíveis na base de dados, ainda que analisados isoladamente, contêm uma riqueza de informações (Berry *et al.*, 2000, p. 164).

Nesse sentido, em um processo de mineração de dados é muito importante a utilização de metadados, que na verdade são “dados dentro de dados”. Os metadados podem adicionar valores mais significativos para uma análise. Um analista que atua com processos de *data mining,* geralmente busca desenvolver uma boa habilidade para ser capaz de reconhecer informações adicionais que podem estar envolvidas em um mesmo dado e de identificar oportunidades para aplicá-las. O principal objetivo do uso de metadados é dar mais sentido às informações e possibilitar que os resultados gerados em um processo de *data mining* possam ser interpretados. Seria impossível listar todos os diferentes tipos de informações e os metadados relacionados aos dados, mas há alguns casos, em particular, em que os metadados são mais comumente utilizados, por exemplo, os dados sobre datas e endereços (Westphal *et al.*, 1998, p. 37).

Muitos registros de dados possuem informações sobre datas. Uma data é uma específica instância de tempo que denota quando um evento ou série de eventos ocorreu. É possível notar que há mais informações contidas em uma data que as que inicialmente os olhos percebem. De fato, é possível extrair diferentes tipos de metadados de uma data, que podem ser utilizados para ajudar a correlacionar e refinar os dados para descoberta de padrões (Westphal *et al.*, 1998, p. 38).

Saber uma data não é suficiente. Ela possui características adicionais que podem ser armazenadas em novos atributos separados na base de dados. Algumas informações muito interessantes podem ser extraídas de uma data, como: dia da semana; mês do ano; trimestre do ano; se é feriado ou não; entre outras (Berry *et al.*, 2000, p. 165).

Na página 69, a figura 5 demonstra um exemplo de como é possível extrair vários metadados de uma data, conforme ensina Westphal *et al.* (1998, p. 38).

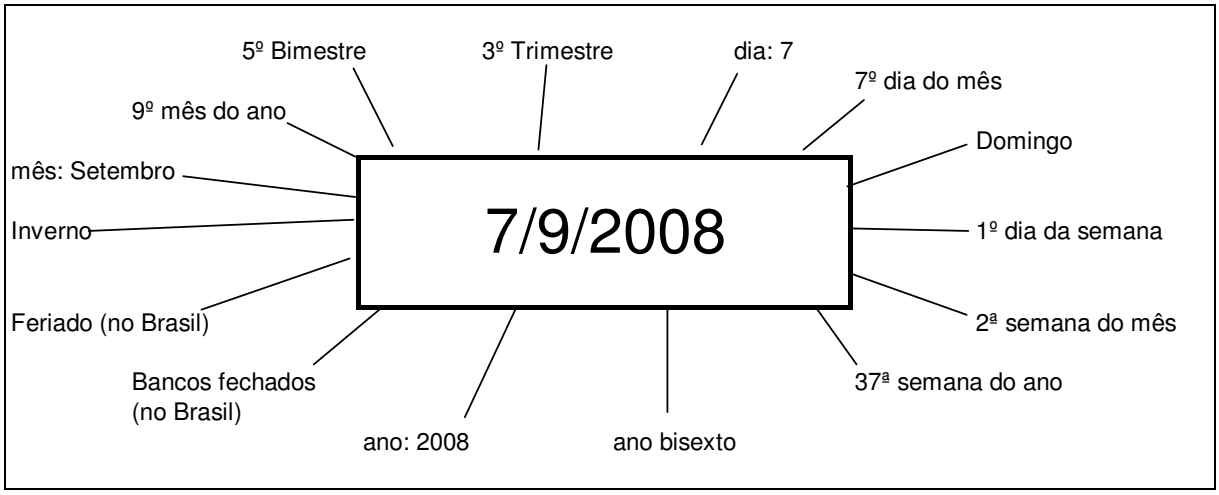


Figura 5. Metadados extraídos de uma informação sobre uma data. (Westphal *et al.*, 1998, p. 38, adaptado).

Assim, há alguns tipos de metadados, mesmo que óbvios, ao serem extraídos de uma data, ajudam na busca por padrões, como o mês e o ano, por exemplos (Westphal *et al.*, 1998, p. 39). A questão prática sobre o nível de detalhamento a ser utilizado no processo de mineração de dados tal como qual o período considerado, se uma semana, um mês, um trimestre ou ano, é frequentemente um fator crítico para o sucesso do processo de mineração de dados (Witten *et al.,* 2005, p. 53). Também para exemplificar, se há entre os dados diferentes datas relacionadas a vários eventos, é muito interessante realizar cálculos para se verificar o número de dias decorridos entre eventos importantes (Berry *et al.*, 2000, p. 165).

Outra variável comum em base de dados se trata de horários de ocorrência de eventos. Um horário específico não fornece tantas informações como uma data, mas é possível identificar se o evento ocorreu pela manhã, à tarde, à noite ou de madrugada (Berry *et al.*, 2000, p. 165).

Dados de um endereço, como rua, cidade e estado, também podem fornecer metadados valiosos. Localidades podem ser comparadas com pontos de referência e, com base em informações geográficas, pode-se verificar a sua distância em relação a eles (Westphal *et al.*, 1998, p. 41).

Outros tipos de dados possuem informações muito valiosas que podem ser extraídas e utilizadas. Mesmo os códigos de endereçamento postal – no Brasil são conhecidos como CEP – geralmente possuem um tipo de padrão de hierarquia

que denota a região onde se encontra determinado endereço e essa informação pode ser extraída. Da mesma forma ocorre com os números de telefone, já que parte do número revela a região onde o telefone está instalado. Até mesmo endereços de páginas da Internet oferecem informações sobre localidades. Das placas de identificação dos veículos também é possível extrair a informação de sua origem (Berry *et al*., 2000, p. 166).

Um problema a ser avaliado durante a preparação dos dados é que há sempre "sujeira" em uma base de dados (Berry *et al*., 2000, p. 177) e uma forma de retirar essa sujeira e outras informações irrelevantes é realizar uma limpeza nesses dados (Tang *et al.,* 2005, p. 13).

Mesmo entre os dados que são úteis para o processo de *data mining*, pode haver registros de informação incoerentes ou incompatíveis com sua finalidade. Comparativamente, se esses registros são considerados uma "sujeira" então essa base de dados precisa passar um processo de "limpeza". Conforme já mencionado, por diversos motivos os dados podem não ter sido registrados corretamente ou representarem uma grande exceção ou desvio dentro do universo de dados considerado. Para garantir que esses registros não comprometam o processo de mineração de dados ou não induzam a um falso padrão de dados, esses registros são adequados ou, em último caso, retirados do processo.

Um tipo de sujeira entre dados é a sua própria ausência, que pode ocorrer por várias razões, desde a falhas na entrada de dados até a sua real inexistência (Berry *et al*., 2000, p. 178). É preciso avaliar cuidadosamente a razão e a consequência dessas ausências, já que podem ser decorrentes, por exemplo, do mau funcionamento de equipamentos, mudanças e desenvolvimentos experimentais ocorridos no momento do registro dos dados ou junção de várias bases de dados similares, mas não idênticas. Assim, há os casos em que os dados são desconhecidos, os que não puderam ser registrados na época por algum motivo, os considerados irrelevantes e os que realmente não existem, isto é, essas informações não estão presentes na base de dados, ou seja, simplesmente os registros não mostram os valores desses dados (Witten *et al.,* 2005, p. 59).

A alternativa mais simples para tratar a ausência de dados em um processo de *data mining* é simplesmente omitir os atributos ou os registros com valores ausentes, no entanto, isso pode ser "perigoso". A ausência de dados pode ser um padrão sistemático e simplesmente deletar registros poderia direcionar os

resultados para uma subsérie ou amostra dos dados tendenciosa. Da mesma forma, parece ser um desperdício desconsiderar as informações de todos os demais atributos, por conta da ausência de um valor (Larose*,* 2005, p. 31).

Diante da ausência de dados, pode-se estudar, entre outras, algumas ações, como (Berry *et al.*, 2000, p. 178):

- não fazer nada: principalmente se a ausência não for muito representativa, os resultados não serão significativamente afetados;
- filtrar os registros com os dados ausentes: o que pode ser uma má idéia se a ausência ocorrer na maioria dos registros, pois os poucos registros resultantes podem influenciar os resultados, embora não sejam representativos;
- ignorar o atributo: se apenas poucos atributos estão maculados com ausências de dados, frequentemente faz sentido ignorá-los;
- prever os valores: é possível, com base em valores de outros atributos, determinar os dados que estão ausentes através de análises específicas para cada caso.

Outro tipo de sujeira se trata de valores incorretos, isto é, dados que não correspondem a um valor válido para aquele atributo. Muitas são as causas desse problema, embora a principal seja a entrada de dados dos sistemas informatizados utilizados nas respectivas atividades realizadas (Berry *et al.*, 2000, p. 180).

A verificação de inconsistências pode ser realizada através de consultas automáticas ao banco de dados (*queries*). Diante das inconsistências encontradas, pode-se avaliar a possibilidade de substituir os valores errôneos, manualmente ou automaticamente (com o uso de *queries*), ou eliminar os registros que apresentem essas incoerências (Passos *et al.*, 2005, p. 39).

Dados duplicados também são decorrentes de possíveis erros ocorridos durante a entrada de dados. Muitas das ferramentas com recursos de aprendizagem da máquina produzirão diferentes resultados se alguns dos registros da base de dados estiverem duplicados, pois essas duplicidades influenciam nos resultados (Witten *et al.,* 2005, p. 59).

Vale destacar que o conhecimento dos dados manipulados é insubstituível. Nesse sentido, a visualização dos dados em gráficos é muito útil na identificação de incoerências e erros existentes entre os dados, além de possibilitar

um melhor entendimento sobre as informações contidas nesses dados (Witten *et al.,* 2005, p. 60).

Também vale ressaltar que para se tornarem interessantes para o processo de *data mining*, os dados dos atributos devem ser apresentados em categorias, intervalos ou números. Tipos de dados mais complicados, como textos, geralmente não devem ser usados, embora seja possível extrair características deles (Berry *et al.*, 2000, p. 181).

4- ALGORITMO APRIORI

O algoritmo a Apriori foi proposto por R. Agrawal e R. Srikant em 1994 para mineração de séries de itens frequentes em bases de dados. O nome do algoritmo é baseado no fato de que o seu método se utiliza das características de um padrão frequente já encontrado anteriormente (prior) para buscar mais padrões (Han *et al.*, 2006, p. 235; Witten *et al.*, 2005, p. 141).

Cada valor presente em um atributo é tratado pelo algoritmo como um "item" e uma combinação de itens é denominada uma série de itens, ou em inglês, *itemset* (Witten *et al.*, 2005, p. 113; Tang *et al.,* 2005, p. 231). Usa-se o termo *K-itemset* para se fazer referência ao conjunto de itens com *K* elementos (Passos *et al.*, 2005, p. 61).

Por princípio, para se buscar associações existentes em um conjunto de dados, deve-se realizar (1) os processamentos necessários para que sejam identificadas todas as combinações de itens possíveis nesse conjunto (Witten *et al.*, 2005, p. 112) e (2) a contagem das repetições de cada combinação.

Uma medida conhecida como suporte é utilizada para mensurar essa frequência. O suporte de uma série de dois itens, por exemplo, é a quantidade total de registros da base de dados em que os dois itens aparecem conjuntamente. Ele pode ser analisado em termos percentuais ao se considerar o total de registros dessa base (Tang *et al.,* 2005, p. 232).

Acontece que, à medida que se aumenta os itens presentes em uma série de dados, aumentam exponencialmente as combinações possíveis entre esses itens, o que leva a mais processamentos de computação para se verificar todas. Se forem realizados todos esses processamentos através de um algoritmo, o processo completo de análise será bem demorado. A solução, então, é reduzir a quantidade de itens envolvidos e, consequentemente, as combinações a serem consideradas em cada etapa de processamento. Esta é uma técnica também conhecida como *pruning*, isto é, "poda". Nesse sentido, o algoritmo desconsidera certas combinações que não atendam a algum critério estabelecido (Berry *et al.*, 1997, p. 144).

O mecanismo mais comum de *pruning* é a definição de um suporte mínimo, o que significa que uma regra, para ser considerada, precisa abranger um número mínimo de registros previamente definido. Por exemplo, se há 1 milhão de

registros e se define um suporte mínimo de 1%, apenas as regras que envolvem mais que 10.000 desses registros interessam (Berry *et al.*, 1997, p. 144).

A definição de um suporte mínimo gera um efeito "cascata". Ao imaginar uma regra do tipo "se A, B e C, então D", isto é, quando os itens A,B e C aparecem juntos, então há grandes chances do item D também se fazer presente, e for definido um suporte mínimo de 10.000 registros, esta regra será considerada apenas se (Berry *et al.*, 1997, p. 145):

- "A" aparecer no mínimo em 10.000 registros; e
- "B" aparecer no mínimo em 10.000 registros; e
- "C" aparecer no mínimo em 10.000 registros; e
- "D" aparecer no mínimo em 10.000 registros.

Isto porque a combinação de 2 itens (2-*itemset*) só atingirá a frequência mínima definida se cada item (1-*itemset*) presente na combinação atingir essa frequência mínima e assim por diante (Witten *et al.,* 2005, p. 117) .

A definição de um suporte mínimo elimina os itens que não aparecem em um número de registros suficiente para alcançar o critério mínimo previamente definido. A cada etapa de computação de uma combinação, há a verificação do seu suporte e esta combinação pode ser eliminada do processo se não atingir o suporte mínimo, reduzindo, assim, a quantidade de combinações que passarão a ser consideradas nas próximas fases (Berry *et al.*, 1997, p. 145).

No mesmo sentido, para a utilização do algoritmo Apriori, deve-se definir previamente um suporte mínimo para a realização da busca por regras de associação. Esse suporte mínimo é utilizado como critério de seleção dos itens que serão considerados durante essa busca e influencia no tempo de processamento do algoritmo. Na definição desse suporte mínimo, devem ser considerados o tamanho do conjunto de dados e, principalmente, o contexto e o objetivo da análise de forma a avaliar a quantidade mínima razoável de coincidências suficiente para justificar uma possível associação.

A definição de um suporte mínimo previamente à realização da busca significa que há o interesse apenas sobre as séries de itens e regras de associação com suporte que alcance pelo menos a esse critério mínimo estabelecido (Tang *et al.,* 2005, p. 232).

Um analista da área de marketing, por exemplo, pode considerar somente associações em situações de compra com suporte acima de 20%, enquanto que um

analista de fraudes ou de casos de ações terroristas pode reduzir esse suporte ao se interessar pelos casos acima de 1%, já que embora sejam tipos de situações consideradas exceções se comparadas a todos os casos possíveis, estas apresentam graves problemas (Larose, 2005, p. 184).

Vale destacar que a quantidade de cálculos e processamentos necessários para gerar regras de associação depende muito da definição do suporte mínimo a ser considerado no processo de busca (Witten *et al.,* 2005, p. 119).

O processamento do algoritmo Apriori se inicia com a busca dos itens isoladamente (*1-itemsets*) que aparecem com mais frequência na base de dados apenas, com base no suporte mínimo previamente definido (Han *et al.*, 2006, p. 235; Witten *et al.,* 2005, p. 117; Tang *et al.,* 2005, p. 234).

Em seguida, utiliza-se apenas esses itens frequentes (os demais são descartados) para gerar as combinações possíveis entre eles e verificar a frequência de coincidências em que cada um deles aparece junto com algum outro item (*2-itemsets*). Os *2-itemsets* mais frequentes (os que alcançaram o suporte mínimo definido) são utilizados para se buscar os *3-itemsets* mais frequentes e assim por diante (Han *et al.*, 2006, p. 235; Witten *et al.,* 2005, p. 117; Tang *et al.,* 2005, p. 234).

O algoritmo se baseia basicamente em duas propriedades (Han *et al.*, 2006, p. 235; Passos *et al.*, 2005, p. 62; Tang *et al.,* 2005, p. 235):

- todas as subséries de um *itemset* frequente deve também ser frequente. Um conjunto de itens (*k-itemset)* somente pode ser frequente se todos os seus subconjuntos (“*k-1”-itemset)* forem frequentes;
- a frequência de um conjunto de itens nunca cresce ao ser adicionado mais um elemento. Pode, na melhor hipótese, permanecer com a frequência igual, ou simplesmente diminuir ao considerar as coincidências existentes com este novo elemento.

Se um item (ou série de itens) é frequente, então ele é um bom candidato a participar da próxima busca (Han *et al.*, 2006, p. 235). Em cada busca, o algoritmo gera os itens (ou séries de itens) candidatos a participar de uma outra próxima tentativa (Fayyad *et al.*, 1996, p. 311).

A figura 6, da página 76, mostra um exemplo que Han *et al.* (2006, p. 237) utilizaram para demonstrar o princípio de busca de associações do algoritmo Apriori, conforme a prévia definição de um grau mínimo de suporte.

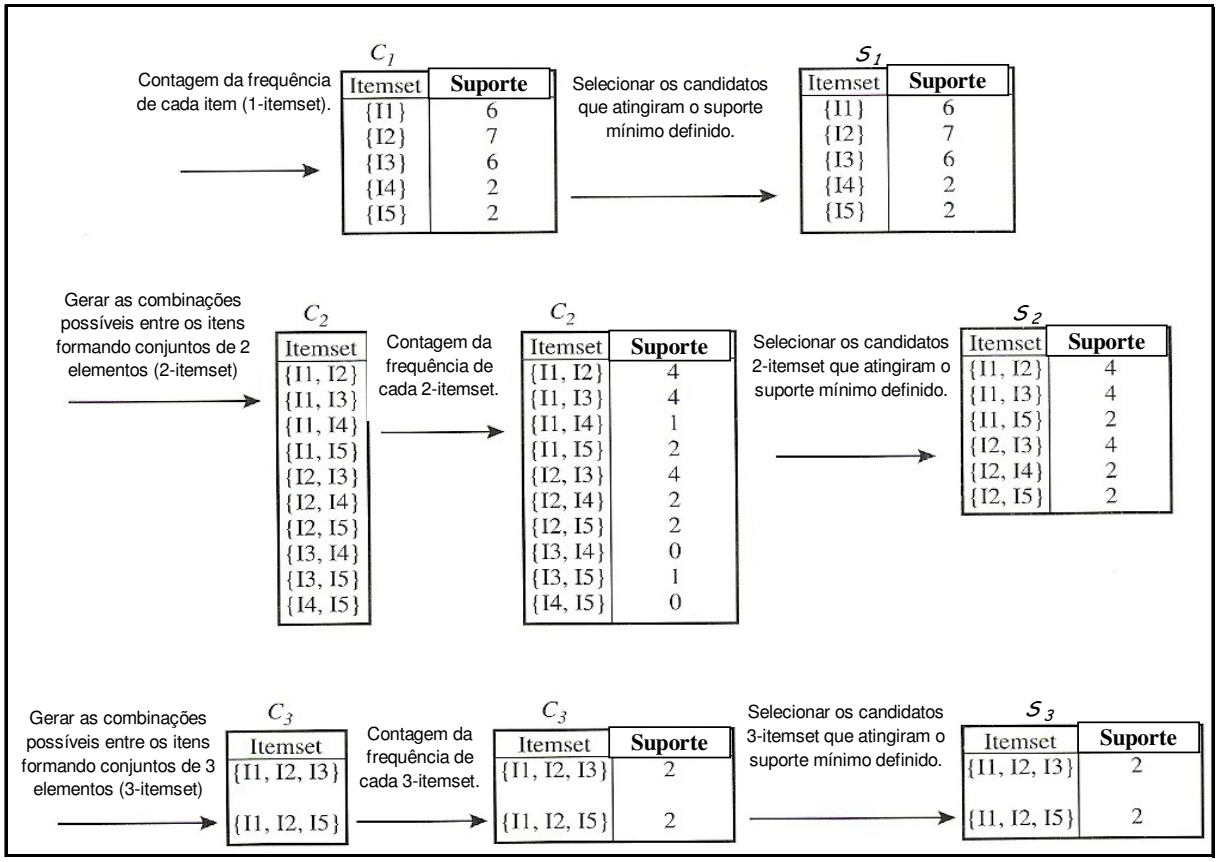


Figura 6. Exemplo de geração de série de itens durante o processamento do algoritmo Apriori com dados de 5 itens, com base em um suporte mínimo de 2 coincidências apenas (Han *et al.*, 2006, p. 237, adaptado).

No mesmo sentido, Tang *et al.* (2005, p. 235) ilustram uma situação de busca de itens frequentes ao considerar um total de 1.000 registros de compras realizadas. Supondo que o suporte mínimo definido seja de 25%, então a frequência mínima exigida para cada item (produto) deve ser maior que 250 para que as séries de itens (itemsets) sejam considerados, conforme se observa na figura 7, da página 77.

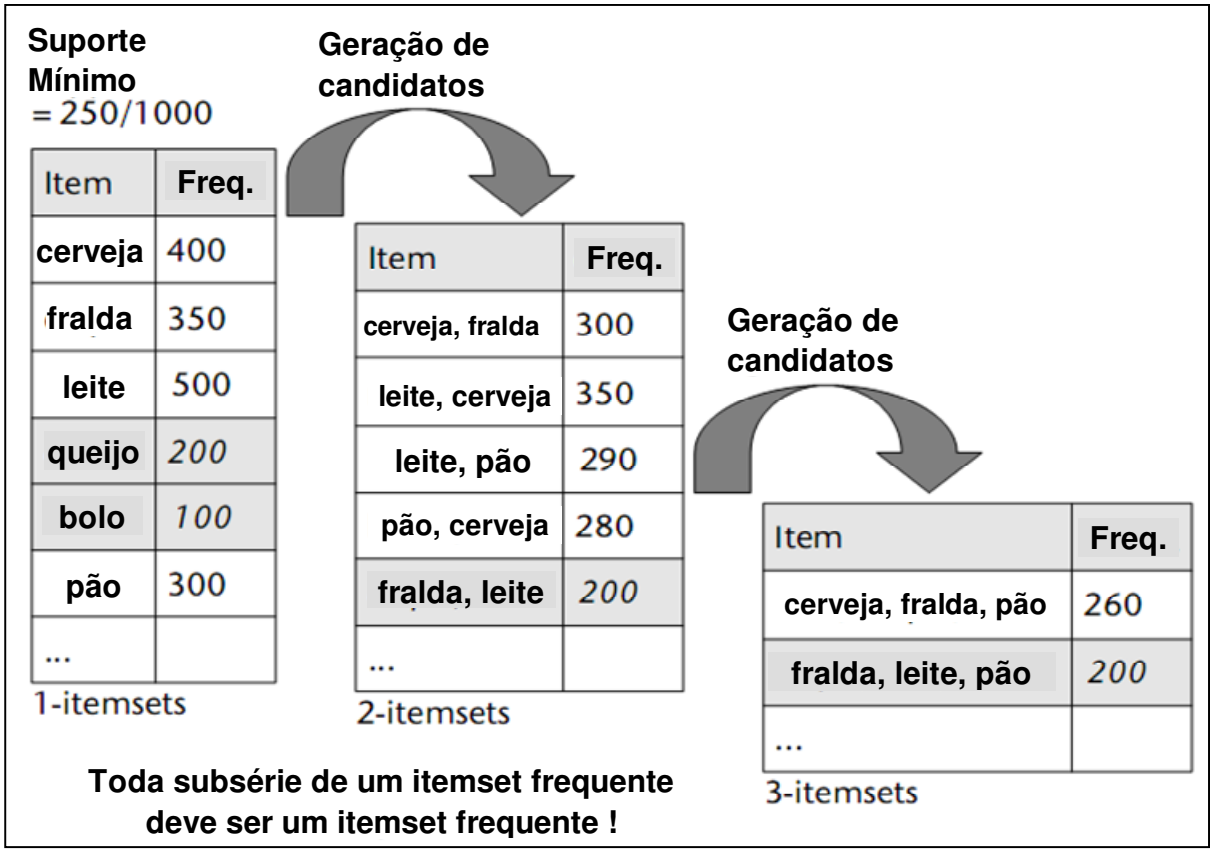


Figura 7. Exemplo de busca de séries de itens (Tang *et al.*, 2005, p. 236, adaptado).

A geração de *itemsets* e a contagem de suas respectivas correlações demandam maior tempo de processamento e uso de memória nos casos em que são formadas inúmeras séries. Isso ocorre quando a quantidade de itens distintos é muito grande. Por exemplo, ao imaginar uma quantidade de 10.000 itens existentes entre os registros e se for estabelecido um suporte muito baixo, o algoritmo de busca poderá chegar a gerar mais de 10^7 *2-itemsets* (Tang *et al.,* 2005, p. 237).

Em síntese, todo o processo é baseado na análise de que se um item (ou uma seqüência de itens) não apresenta uma frequência acima da mínima exigida (suporte mínimo), então, não há chances deste item (ou itens), quando associado a qualquer outro, apresentar essa frequência acima desse mínimo exigido. O fato é que se um item apresenta certa frequência, esta se consolida como o limite máximo a ser alcançado caso seja considerada uma associação deste item com algum outro. Se um *itemset* já não é frequente, adicionar um outro item neste *itemset* não o tornará frequente (Han *et al.*, 2006, p. 235; Larose, 2005, p. 184).

Nesse sentido, são encontradas automaticamente todas as combinações de itens possíveis que possuam frequências acima do suporte mínimo exigido. Entretanto, para cada combinação de itens frequente encontrada, o método Apriori

realiza uma avaliação de forma a verificar se dessa série de itens frequente é possível abstrair uma possível associação. Com esse objetivo, o algoritmo busca mensurar a real influência de um item sobre os outros, ou seja, medir a "força" dessa possível associação.

Adriaans *et al.* (1996, p. 64) ensinam que é preciso buscar as associações que são mais interessantes. Isto porque em uma base de dados há a possibilidade de existir inúmeras associações diferentes. Assim é preciso ter uma idéia dos tipos de associação que se busca, pois não há qualquer algoritmo que mostre o que é interessante ou não. Se por um lado, um algoritmo que encontra muitas associações pode apresentar vários resultados que são inúteis, por outro, um algoritmo que busca apenas um número limitado de associações, possivelmente deixará de apresentar muitas informações interessantes.

Assim, um outro princípio também é adotado pelo algoritmo Apriori, o qual está relacionado à probabilidade de ocorrer uma associação. Essa probabilidade também é chamada de confiança (Tang *et al.,* 2005, p. 232).

Enquanto o suporte é a medida referente ao tamanho da parte dos registros de dados que satisfaz a uma determinada regra, a confiança é a medida da frequência em que a associação se confirma, considerando todas as ocorrências de cada item analisado (Silberschatz *et al.*, 2006, p. 502).

Enquanto o suporte de uma regra de associação entre os itens A e B é a proporção do total de registros da base de dados que contêm ambos os itens, a confiança dessa regra é a mensuração da sua exatidão, isto é, ao considerar todos os registros em que consta o item A, quantos desses apresentam também o item B (Larose, 2005, p. 184). Matematicamente, a confiança da série {A, B} é calculada pela divisão do suporte da série (2-*itemset* {A,B}) pelo suporte do item A (1-*itemset* {A}) (Tang *et al.,* 2005, p. 232).

Antes de iniciar o processo de busca de associações, também pode-se definir critérios relacionados à probabilidade das regras ocorrerem. Da mesma forma que ocorre com a medida de suporte, é muito comum estabelecer um grau de confiança mínimo para a realização da busca por associações. Isso significa que apenas as regras com probabilidades maiores que o grau de confiança mínimo estabelecido interessam (Tang *et al.,* 2005, p. 232).

Geralmente, a preferência recai sobre aquelas associações com maiores índices de suporte ou de confiança ou ambos. Assim, pode-se optar pela

identificação apenas das associações que apresentarem índices superiores ao previamente estabelecidos, isto é, as associações consideradas mais fortes (Larose, 2005, p. 184).

Basicamente, o processamento do algoritmo Apriori pode, então, ser dividido em duas etapas (Passos *et al.*, 2005, p. 106; Witten *et al.,* 2005, p. 117; Tang *et al.,* 2005, p. 230; Larose, 2005, p. 184):

- encontrar todas as combinações de itens consideradas frequentes, ou seja, que satisfaçam a condição de suporte mínimo;
- a partir das combinações frequentes, gerar as regras de associação, com base no grau de confiança mínimo definido.

Tang *et al.* (2005, p. 231) ilustra bem essas etapas, conforme se observa através da figura 8, a seguir.

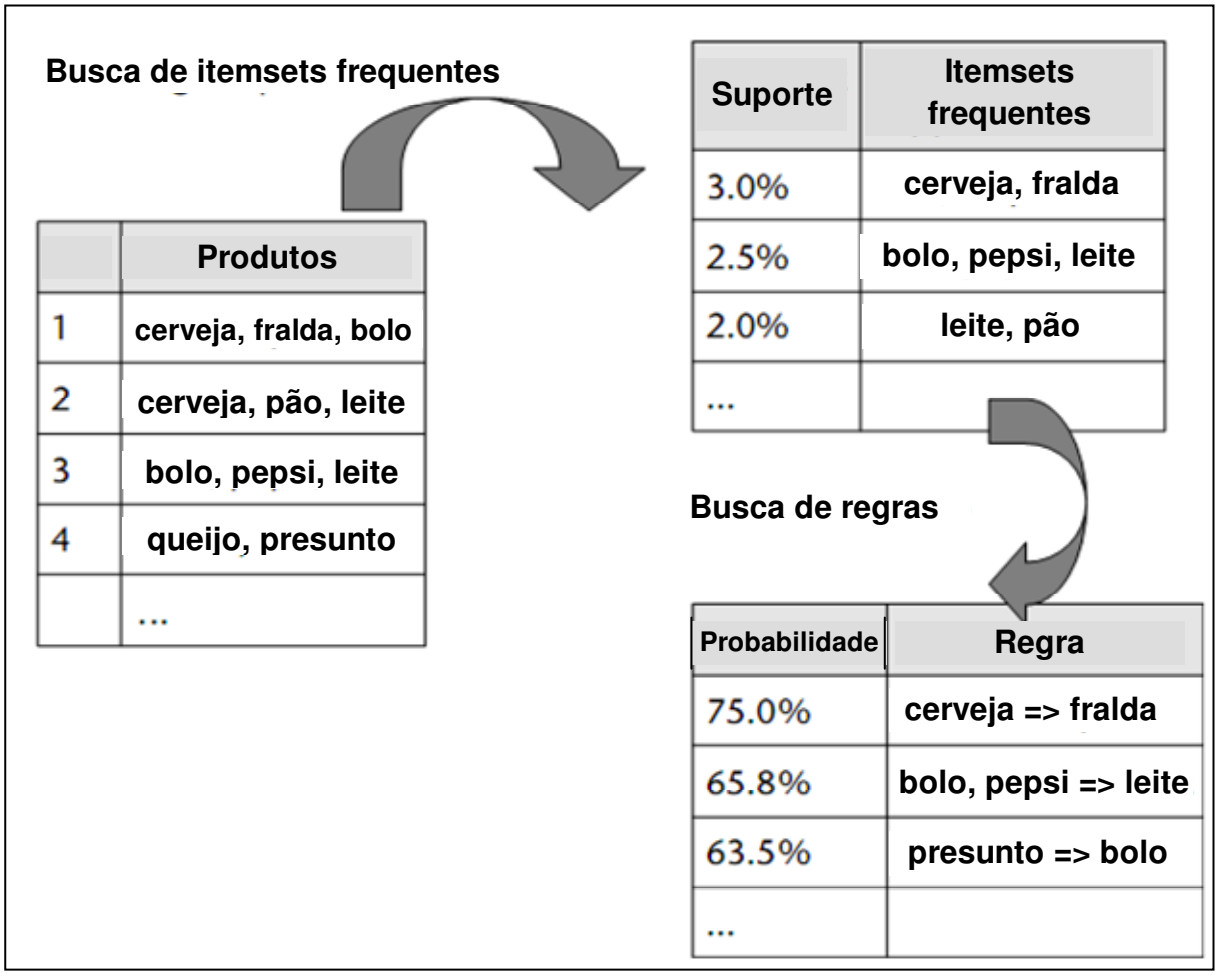


Figura 8. As duas etapas do processo do algoritmo de associação (Tang *et al.*, 2005, p. 231, adaptado).

Pela compreensão da sistemática adotada pelo método Apriori, pode-se abstrair que o algoritmo apresenta uma estrutura tal como demonstrada na figura 9, a seguir.

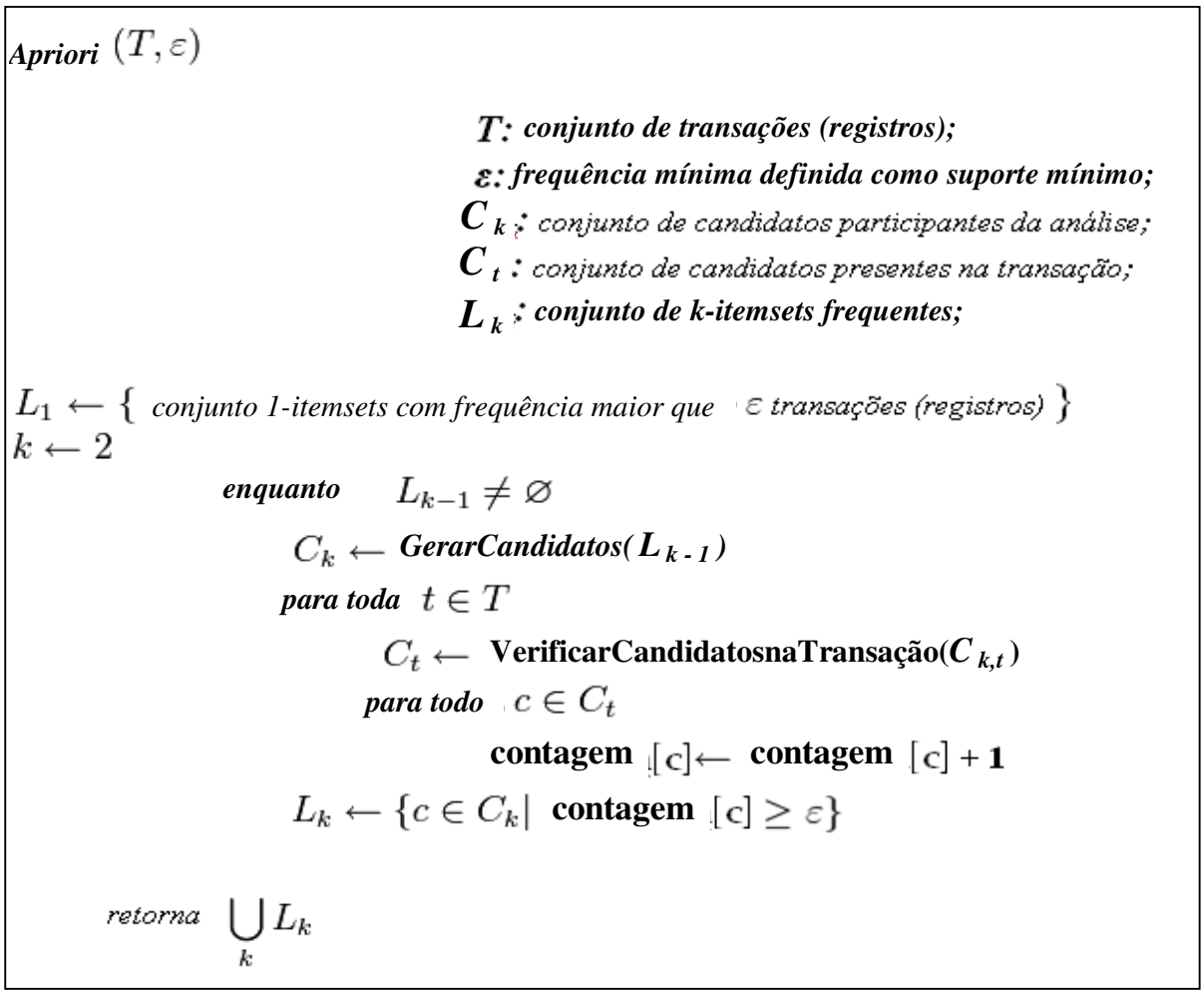


Figura 9. Pseudo-código do algoritmo Apriori (Wikipedia, 2010, adaptado).

Infelizmente, o algoritmo Apriori apresenta a deficiência de não processar bem variáveis numéricas, a não ser que sejam discretizadas durante a etapa de pré-processamento de dados (Larose, 2005, p.190).

Esse aspecto pode ser facilmente compreendido ao observar os princípios de busca de associações do algoritmo Apriori. As variáveis numéricas geralmente apresentam muitos valores distintos, o que acarreta o baixo suporte dos itens. Com suportes abaixo do mínimo definido, a grande maioria dos itens é desconsiderada e por isso os valores dessas variáveis quase não aparecem entre as regras de associação.

A alternativa de discretização das variáveis numéricas reduz a quantidade de itens distintos, pois os itens são agrupados sob algum critério. Com isso, os grupos, que passam a ser tratados como itens, podem alcançar o suporte mínimo pela união dos suportes de todos os valores agrupados. Assim, as chances dos valores da variável aparecer entre as regras de associação aumentam.

5- ALGORITMO K-MEANS

Técnicas de clusterização são utilizadas para a descoberta de grupos naturais em séries de dados sem que se tenha conhecimento sobre as características desses dados (Kogan *et al.*, 2006, p. 127).

Essas técnicas têm sido desenvolvidas por cerca de meio século e são ricas e diversificadas. Podem variar conforme a principal metodologia empregada e áreas de aplicação (Pedrycz, 2005, p. 6).

Uma técnica de clusterização muito popular é o algoritmo K-Means (Kogan *et al.*, 2006, p. 38). O método K-Means de detecção de grupos, foi criado por J. B. MacQueen, em 1967, e é um dos mais comumente utilizados na prática, para realização da tarefa de clusterização (Berry *et al.*, 1997, p. 192). A popularidade e a grande utilização do algoritmo se devem ao fato de ser facilmente entendido e implementado (Kogan *et al.*, 2006, p. 41).

O algoritmo parte do princípio de que, ao considerar a posição dos itens em um espaço geométrico, cada grupo presente em uma base de dados possui um ponto central (médio), chamado de centróide, que é a média entre as posições dos seus elementos. Assim, os grupos podem ser identificados e representados pelos seus centróides (Kogan *et al.*, 2006, p. 37).

A formação dos grupos pelo algoritmo K-Means ocorre através de um método iterativo (Kogan *et al.*, 2006, p. 101) que busca encontrar, a cada iteração, os valores para os centróides que melhor representem os grupos. Essa busca se baseia no princípio de que a distância de todos componentes dos grupos em relação aos seus respectivos centróides deve ser a mínima possível.

O conceito de distância é o componente essencial de qualquer processo de clusterização. Pelo cálculo da distância de dois itens é possível ter uma idéia do grau de proximidade entre eles e, conforme essa proximidade, alocá-los no mesmo grupo (Pedrycz, 2005, p. 2).

Por esta abordagem, em síntese, o primeiro passo é escolher o número de grupos que se quer formar. Este número é o “K” relativo ao nome do algoritmo. Em seguida, “K sementes” são escolhidas, entre os registros, como convidadas para serem os “centróides” (as posições centrais) iniciais dos grupos. A partir disso, cada registro de dados é associado a um grupo preliminar, conforme sua maior

proximidade (distância) do respectivo centróide provisório previamente encontrado. Assim que os grupos são formados, um novo centróide de cada grupo é calculado com base na média entre as distâncias dos componentes pertencentes aos grupos. Com base nestes novos centróides encontrados, verifica-se novamente a distância de todos os valores dos registros em relação aos novos centróides. Assim, como as posições dos centróides mudaram, pode acontecer que um registro que originalmente fazia parte de um grupo passe a fazer parte de outro, pois um outro centróide está mais próximo. Uma outra iteração é realizada para se calcular os novos centróides e verificar se há mudanças quanto aos elementos dos grupos. Novas iterações são realizadas para cálculo de novos centróides e o processo termina quando os centróides param de se modificar ou após ser atingido um número máximo de iterações previamente estabelecido (Berry *et al.*, 1997, p. 191; Berry *et al.*, 2000, p. 104; Passos *et al.*, 2005, p. 102; Witten *et al.,* 2005, p. 137; Tang *et al.,* 2005, p. 191; Larose*,* 2005, p. 153; Kogan *et al.*, 2006, p. 102).

Esses processamentos podem ser melhor visualizados através do exemplo apresentado através da figura 10, da página 84.

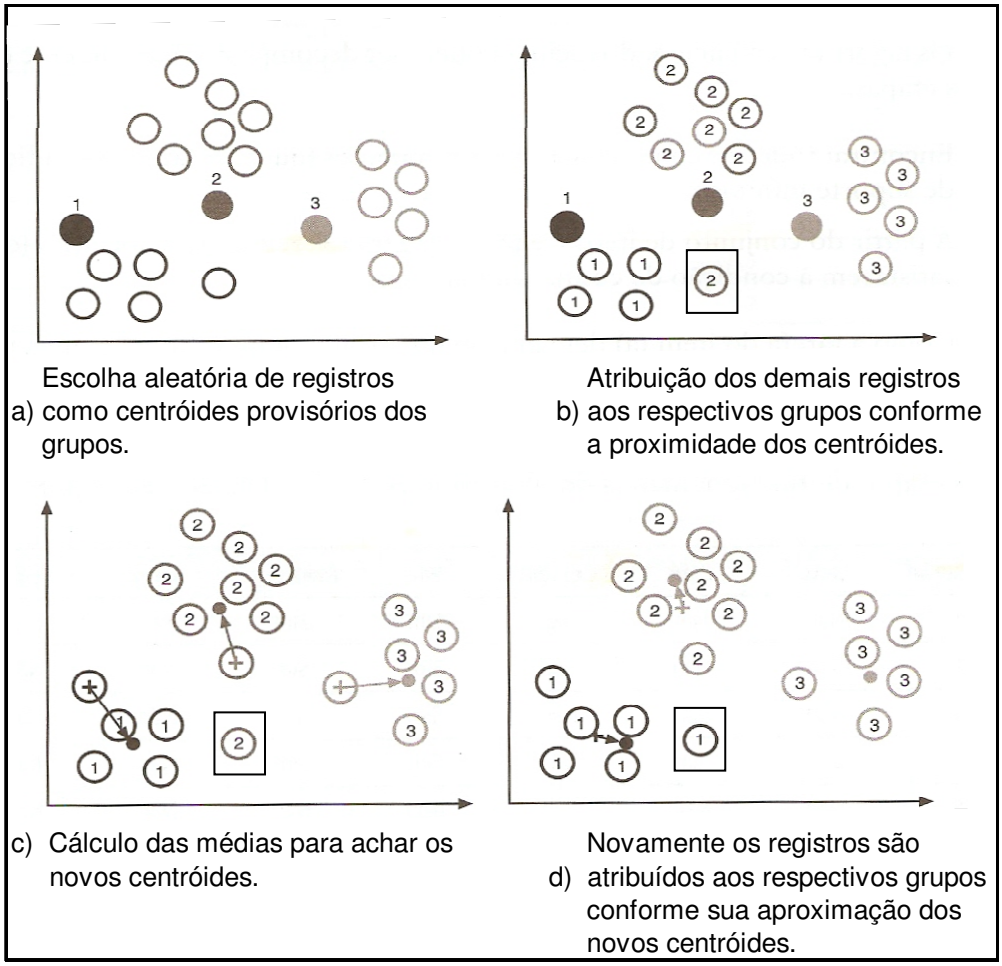


Figura 10. Processamentos do algoritmo K-Means (Passos *et al.*, 2005, p. 105, adaptado).

Nota-se que o item destacado na figura 10, “b” inicialmente é alocado ao grupo 2, por estar naquele momento mais próximo ao centróide daquele grupo. Após o cálculo das médias das distâncias entre elementos dos grupos, os novos centróides são encontrados e o elemento que anteriormente estava mais próximo ao do centróide do grupo 2, agora está mais próximo ao do grupo 1 e, por isso, o elemento passa a fazer parte deste grupo.

Da mesma forma, pela compreensão da sistemática de processamento do método K-Means, pode-se abstrair que o algoritmo apresenta uma estrutura tal como apresentada na figura 11, da página 85.

Entradas:

$I = \{i_1, ..., i_k\}$ (conjunto das instâncias a serem agrupadas)

n (quantidade de grupos)

Saídas:

$C = \{c_1, ..., c_n\}$ (conjunto dos centróides dos grupos)

$m : I \rightarrow C$ (membro do grupo)

Execute **KMeans**

Defina C como os centróides iniciais (escolha aleatória de valores de I)

Para cada $i_j \in I$

$m(i_j) = \underset{k \in \{1..n\}}{\text{menor}}$ distância(i_j, c_k) (atribui elemento ao grupo)

Fim

Enquanto m estiver mudando

Para cada $j \in \{1..n\}$

Processe i_j como centróide de $\{i | m(i) = j\}$

Fim

Para cada $i_j \in P$

$m(i_j) = \underset{k \in \{1..n\}}{\text{menor}}$ distância(i_j, c_k) (atribui elemento ao grupo)

Fim

Fim

Retorne C

Fim

Figura 11. Pseudo-código do algoritmo K-Means (Kadous, 2002, p. 85, adaptado).

Quando os dados são numéricos, há várias medidas que podem ser utilizadas para o cálculo das distâncias entre os itens (Larose, 2005, p. 148; Berry *et al.*, 1997, p. 197). Alguns exemplos de funções para cálculos de distâncias entre variáveis numéricas são demonstrados por Pedrycz (2005, p. 3), conforme a tabela 2, da página 86:

Cálculo de distância	Fórmula
Distância euclidiana	$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Distância Hamming (soma das distâncias absolutas, "city block")	$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\lvert x_i - y_i\rvert$
Distância Tchebyschev	$d(\mathbf{x},\mathbf{y}) = \max_{i=1,2,\ldots,n}\lvert x_i - y_i\rvert$
Distância Minkowski	$d(\mathbf{x},\mathbf{y}) = \sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p},\ p > 0$
Distância Canberra	$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert}{x_i + y_i}$, x_i e y_i são positivos
(...)	(...)

Tabela 2. Exemplos de funções para cálculos de distância entre variáveis numéricas (Pedrycz, 2005, p. 3, adaptado).

Vale destacar que a forma básica e original do algoritmo K-Means utiliza a medida de distância euclidiana como parâmetro para alocação dos elementos nos grupos, conforme a proximidade do respectivo centróide, mas outras medidas podem ser utilizadas (Tang *et al.,* 2005, p. 191; Larose, 2005, p. 153).

Como acontece com muitos métodos de clusterização, o número de *k* grupos precisa ser previamente definido como parâmetro a ser utilizado como critério para realização dos agrupamentos. A definição de um número muito alto pode resultar em grupos muito granulados e muito próximos (Kogan *et al.*, 2006, p. 66).

Uma das dificuldades para a aplicação do algoritmo K-Means é decidir sobre a quantidade grupos que o algoritmo deve procurar, isto é, definir o "*k*". Se essa quantidade não for condizente com a divisão natural dos dados, os resultados podem não ser muito bons. O analista de *data mining* pode conhecer tão bem a base de dados a ponto de ser capaz de estimar a quantidade de grupos nele existentes. Entretanto, uma boa alternativa é aplicar o algoritmo com várias definições de *k* e verificar os resultados, de forma a identificar a situação em que os grupos estão mais coesos e melhor divididos (Larose, 2005, p. 157; Berry *et al.*, 2000, p. 108; Passos *et al.*, 2005, p. 103).

Entre algumas das limitações do algoritmo K-Means é que o mesmo funciona convenientemente apenas com dados numéricos (isto é, o seu uso faz mais

sentido sob uma perspectiva geométrica) e os seus resultados podem ser negativamente afetados por algum desvio muito grande presente entre os dados (Kogan *et al.*, 2006, p. 37-39).

Essa limitação é facilmente compreendida, já que o princípio para a realização dos agrupamentos se baseia no cálculo de distâncias.

Entretanto, a grande popularidade do algoritmo resultou em muitas propostas de extensões e modificações (Kogan *et al.*, 2006, p. 42) e por isso há muitas variações da forma básica do algoritmo K-Means (Witten *et al.,* 2005, p. 142; Passos *et al.*, 2005, p. 103). Entre as diferenças que se pode encontrar são (Berry *et al.*, 1997, p. 205; Passos *et al.*, 2005, p. 103):

- métodos alternativos para se escolher as primeiras sementes;
- métodos alternativos para se calcular os próximos centróides;
- métodos alternativos para se verificar a proximidade ou distância dos elementos em relação aos centróides.

6- PROPOSTA DE UMA METODOLOGIA PARA A APLICAÇÃO DO PROCESSO DE MINERAÇÃO DE DADOS COM O USO DO ALGORITMO APRIORI

Ao considerar o referencial teórico, conclui-se que todo o processo de mineração de dados requer a realização de vários procedimentos e deve ser executado de forma planejada e sob determinados critérios. Como já foi abordado, os procedimentos a serem adotados, especialmente para a preparação dos dados, dependem da técnica de mineração de dados escolhida, pois cada técnica pode requerer uma formatação ou um pré-processamento de dados específico.

Assim, neste trabalho, é proposta uma metodologia para a realização de todo o processo de mineração de dados para a descoberta de associações, com a utilização do algoritmo Apriori. O principal objetivo é oferecer um direcionamento para este processo de forma que os resultados sejam aprimorados.

A maior preocupação se refere aos procedimentos da etapa de pré-processamento. Nesse sentido, a metodologia proposta visa preparar os dados disponíveis para torná-los mais adequados ao processamento do algoritmo Apriori, sob a premissa de que uma melhor preparação dos dados pode aprimorar os resultados do processo de mineração de dados.

Dessa forma, propõe-se um direcionamento para realização dos procedimentos durante a preparação dos dados.

Em síntese, a metodologia busca avaliar o conjunto de dados em estudo através de um processo de “filtragem” e “tratamentos” para definir quais e como os dados serão submetidos à aplicação do algoritmo Apriori. A princípio, todo o conjunto é útil, mas há uma sequência de fases em que são realizadas análises e tratamentos dos dados com o objetivo de aumentar o desempenho do algoritmo na tarefa de encontrar associações. Na medida em que se percorre a sequência de fases, os dados também passam for um processo rigoroso de seleção segundo os princípios de busca do algoritmo Apriori, sendo que os dados não selecionados são retirados do processo, enquanto os escolhidos são trabalhados e tratados para que possam ser mais bem aproveitados durante o processo de *data mining*. A metodologia pode ser melhor visualizada através da figura 12, da página 90.

Os principais objetivos da metodologia são:

- selecionar apenas os dados que são ou poderão, após um tratamento, tornar-se compatíveis e apropriados para a aplicação do algoritmo Apriori. Dessa forma, evita-se processamentos desnecessários e, consequentemente perda de tempo;
- tratar os dados de forma que se tenha a máxima quantidade possível de itens com frequência acima do suporte mínimo definido para o processamento do algoritmo em cada atributo selecionado. Isso porque os itens com frequência abaixo desse suporte são descartados imediatamente pelo algoritmo. Nesse sentido, busca-se garantir que mais dados possam ser ao menos considerados pelo algoritmo, ainda que não contribuam com alguma associação.

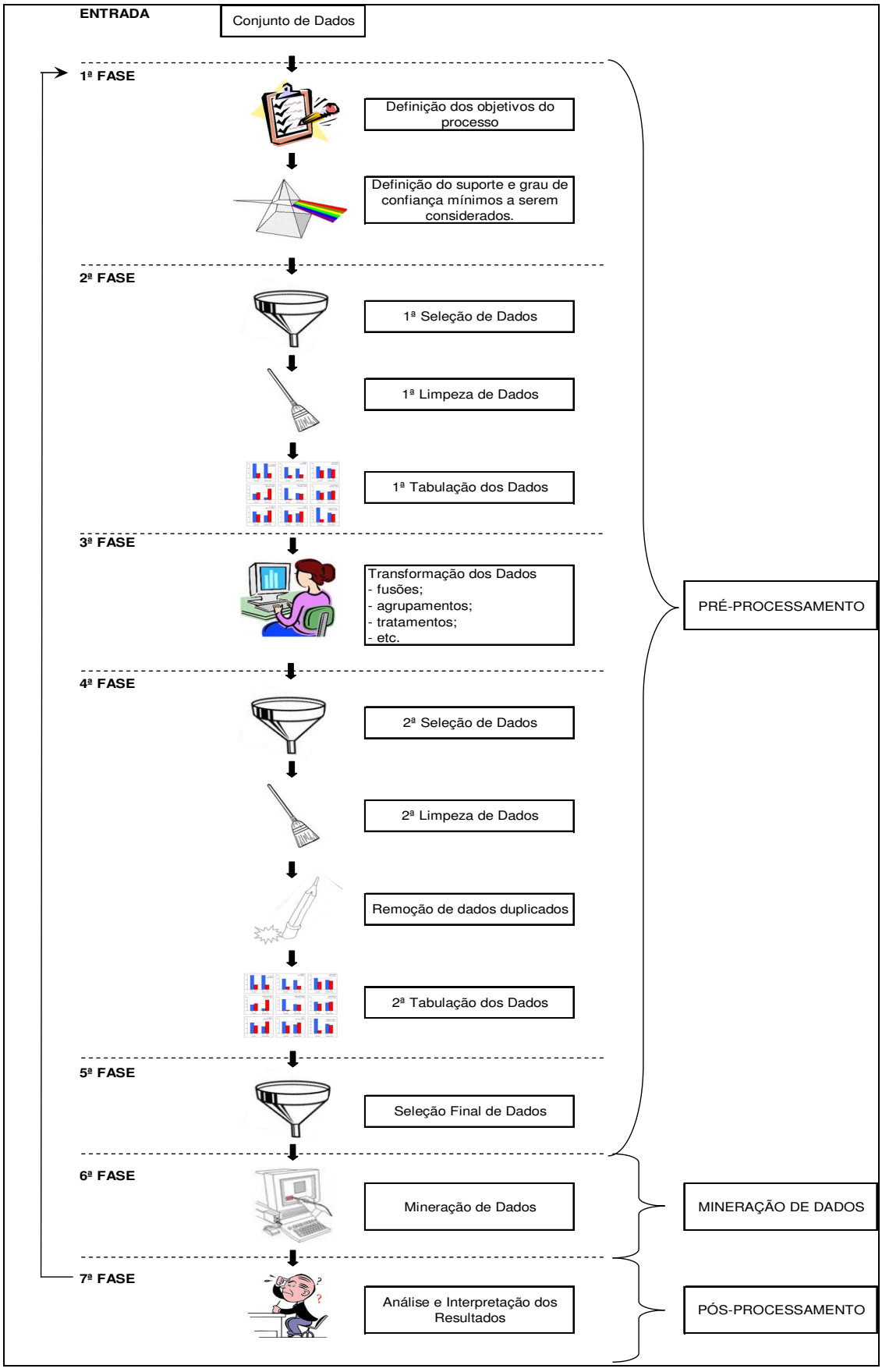


Figura 12. Metodologia proposta.

6.1- 1ª Fase

6.1.1- Definição dos objetivos do processo

A busca por associações precisa ser direcionada por um ou mais objetivos que devem ser definidos logo no início do processo. Esse procedimento é determinante para o processo, especialmente para a seleção de dados.

Uma base de dados pode apresentar grandes quantidades de dados de vários tipos, de diversas origens e relacionados a diferentes informações. Assim, a busca por associações precisa ter focos, que podem ser os tipos de associações a serem investigados, ou seja, que associações se espera encontrar.

Uma indústria, por exemplo, em determinado momento pode querer investigar possíveis associações existentes entre os dados de seus clientes e em outro, analisar os dados de produção, finanças ou vendas.

De certa forma, é uma delimitação e direcionamento sobre a abrangência da análise que se pretende fazer dos dados.

6.1.2- Definição do suporte e grau de confiança mínimos

O algoritmo Apriori processa os dados com base nos parâmetros de suporte e confiança mínimos que são definidos previamente à sua aplicação. Para essa definição é importante considerar os objetivos definidos para o processo de forma a avaliar os graus de suporte e confiança razoáveis para a busca por associações. Certos tipos de associações, por exemplo, só se mostram mais interessantes quando há um alto suporte e uma forte confiança, enquanto há casos que índices menores já despertam o interesse por uma análise mais profunda.

6.2- 2ª Fase

6.2.1- Primeira seleção de dados

Também com foco nos objetivos do processo, a base de dados é analisada e são selecionados os dados que serão preparados para ser submetidos ao processamento pelo algoritmo Apriori. Essa primeira seleção se refere à escolha dos atributos que serão analisados no contexto da mineração de dados.

Inicialmente, vale destacar que, conforme os objetivos estabelecidos, muitos dados já podem ser desconsiderados, se estiverem fora do contexto definido para a busca.

Dessa forma, alguns atributos, diante do contexto e dos objetivos da mineração de dados, não agregam qualquer valor para análise. Certos atributos podem existir na base de dados simplesmente para serem utilizados como suporte para algum controle do sistema informatizado ou talvez servir a outras finalidades alheias ao contexto em análise. É recomendável que a identificação desses atributos seja realizada com o auxílio de um especialista com grandes conhecimentos do contexto dos dados.

Por outro lado, deve ser considerado também o fato de que se trata de uma busca por associações e a técnica de mineração de dados a ser utilizada é o algoritmo Apriori. Assim, busca-se selecionar os dados que já estejam ou, após um tratamento, poderão estar adequados ao processamento do algoritmo.

Observa-se que o algoritmo Apriori operacionaliza sua busca por associações através da verificação de frequências em que os dados aparecem na base de dados, ou seja, é realizado um processo de contagem para verificar quantos registros da base de dados apresentam uma determinada característica no contexto de um atributo. Diante disso, atributos do tipo texto com características de preenchimento livre e sem qualquer padronização podem apresentar infinitas variações em uma base de dados. A não ser que haja uma mínima padronização no preenchimento dos dados, uma contagem apenas irá confirmar que cada item desse atributo tem uma frequência muito pequena. O atributo não deve ser descartado apenas se houver uma chance de ser trabalhado em um próximo momento com o

objetivo de agrupar os itens para reduzir as variações de seus dados e garantir uma maior padronização das suas informações.

Estes atributos, por sua própria natureza, geralmente não agregam valor durante a busca por associações, pois em muitos casos nem sequer oferecem boas condições para serem classificados ou tabulados. Como exemplos desse tipo de atributo, pode-se citar: nomes, contatos, documentação e alguns outros itens cadastrais de pessoas; descrições e determinadas características de objetos; redações que expressam observações, análises ou anotações; entre outros.

Assim, pelo fato de que o algoritmo Apriori descarta os dados com frequências mínimas, é muito provável que todos os dados desses atributos serão desconsiderados durante o seu processo de busca tão logo identifique essas baixas frequências. Mantê-los apenas acarretaria maior tempo de processamento durante o processo de mineração de dados. Dessa forma, assim que identificado, esse tipo de atributo é retirado do processo de mineração de dados, a não ser que se verifique alguma chance para que possa ser trabalhado em um segundo momento de forma a agrupar os seus dados para reduzir a quantidade de seus itens distintos e, consequentemente, obter-se grupos que apresentem maiores frequências.

É possível identificar, ainda, que determinados atributos descritivos já estejam representados ou ligados por outros mais codificados, como códigos de produtos, códigos de cidades, etc., geralmente utilizados como chaves para indexação do banco de dados. Nesse sentido, se dois atributos armazenam os mesmos dados apenas de formas diferentes, basta que um deles seja selecionado. Nesse caso, a preferência para a seleção recai sobre os atributos já codificados, pois os seus dados geralmente ocupam menos espaço e são mais rápidos de serem contados, o que melhora o desempenho do processamento do método Apriori.

Em qualquer processo de mineração de dados há momentos de tomada de decisão sobre a seleção de dados que serão utilizados no processo. Nessa metodologia, a seleção completa se concretiza ao se percorrer todas as suas etapas de seleção, como em um processo de eliminação. Assim, essa primeira seleção se trata apenas de um primeiro momento dedicado a decisões sobre a escolha de dados que, inicialmente, participarão do processo.

6.2.2- Primeira limpeza dos dados

Trata-se da manipulação dos dados de forma a reduzir o índice de ausência de informações e inconsistências na base de dados. Essa limpeza é realizada com o objetivo de retirar incoerências e falhas de preenchimento que possam ter ocorrido durante a entrada dos dados presentes no conjunto de registros e preencher os que possam estar nele ausentes.

Segundo Han *et al.* (2006, p. 48), dados ausentes em determinados atributos podem ser preenchidos através de inferências. Assim, nessa fase se busca realizar as correções, quando possível, diante da análise do contexto e da lógica das atividades que deram origem aos dados. Caso não seja possível corrigir os dados ou se dessa correção puder resultar riscos para integridade do conjunto de dados, esses dados tão somente são marcados e considerados como ausentes.

Nesse sentido, valores nulos, espaços em branco e dados inconsistentes, como datas inexistentes ou impossíveis e valores absurdos, entre outros que denotem uma situação impossível dentro do contexto analisado, se não puderem ser corrigidos, devem ser identificados como ausentes. Por outro lado, é comum existir atributos que apresentam dados ausentes ou com variações decorrentes de erros de digitação, como o uso de acentos e espaços (mais comum em atributos do tipo texto) e falta de algum dígito, que podem ser corrigidos através de inferências baseadas em análises do contexto da execução das atividades que originaram os dados.

Recursos de consultas automatizadas (*queries)* em bancos de dados podem ser utilizados para a identificação e análise dessas inconsistências, bem como a realização de possíveis correções.

Vale destacar que essa limpeza, embora seja apenas a primeira dessa metodologia, é uma atividade importante e normalmente é realizada em qualquer aplicação de um processo de mineração de dados.

6.2.3- Primeira tabulação dos dados

É interessante tabular os atributos selecionados após o primeiro processo de limpeza, isto é, realizar uma contagem dos itens distintos presentes em cada

atributo. Essa contagem permite uma visualização das frequências dos itens nos atributos de forma a possibilitar a avaliação das variações de categorias ou valores neles existentes. A representação dessas tabulações em gráficos auxilia a identificação dos atributos que apresentam maiores ou menores variações de categorias e até na verificação de possíveis polarizações ou concentração de dados em determinadas classes.

Essa tabulação permite conhecer melhor os dados que serão submetidos ao processo de mineração. Com os resultados da tabulação, é possível, inclusive, já prever que certos atributos teriam pouca ou nenhuma influência no processo durante a aplicação do algoritmo Apriori. Entretanto, após um processo de transformação de seus dados, esses atributos podem se tornar uma rica fonte para busca de associações.

6.3- 3ª Fase – Transformação de dados

Após a análise da tabulação dos dados, é possível verificar em determinadas bases de dados que certos atributos apresentam inúmeras variações de categorias (itens distintos); consequentemente apresentam baixas frequências. Existe até mesmo a possibilidade de maioria dessas categorias estar abaixo do suporte mínimo considerado para a aplicação do algoritmo Apriori e, nesse caso, nem seriam consideradas. Como exemplos, pode-se citar os atributos do tipo data, quantidades ou valores monetários. Dependendo da base de dados analisada, esses atributos podem apresentar inúmeras variações, pois são infinitas as combinações possíveis.

Para esses casos, pode-se buscar alguma alternativa para agrupar essas categorias, de forma a manter a integridade dos dados. Por exemplo, em um atributo do tipo data pode-se considerar apenas o mês, o trimestre ou ano em que o respectivo evento ocorreu. Pode-se, também, estabelecer intervalos ou níveis para abranger os valores de atributos sobre quantidades ou valores monetários, principalmente quando há muitos itens distintos no atributo. O tratamento desses atributos permite que se tenha menos categorias (itens), mas com frequências maiores, sem que haja perda da consistência dos dados.

A figura 13 apresenta dois exemplos sobre agrupamentos de dados relacionados a datas e valores numéricos em categorias.

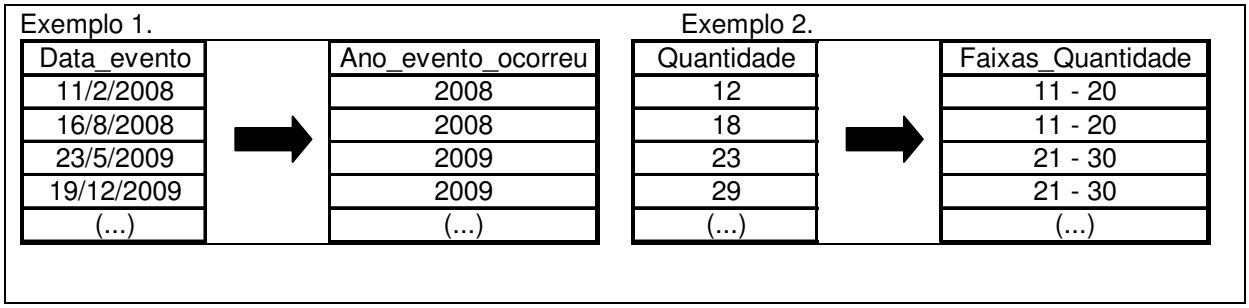


Figura 13. Exemplos de agrupamento de dados relacionados a datas e valores numéricos.

Outra situação possível é a polarização ou concentração de dados. Uma base de dados, por exemplo, pode ter atributos que apresentem uma ou duas categorias de dados com altas frequências, enquanto as demais possuem frequências menores até que o grau de suporte mínimo definido para a aplicação do método Apriori.

Para reduzir as concentrações ou polarizações de dados, as demais categorias com menor frequência podem ser agrupadas de forma a equilibrar as frequências dos itens do atributo. Esse tratamento, além de resultar em um equilíbrio maior entre as categorias do atributo, possibilita um melhor aproveitamento dos dados no processo de mineração, já que evita a participação de apenas algumas categorias durante o processamento dos dados.

A figura 14 demonstra um exemplo relacionado ao agrupamento de dados semelhantes para se obter um maior equilíbrio das frequências entre os itens dos atributos.

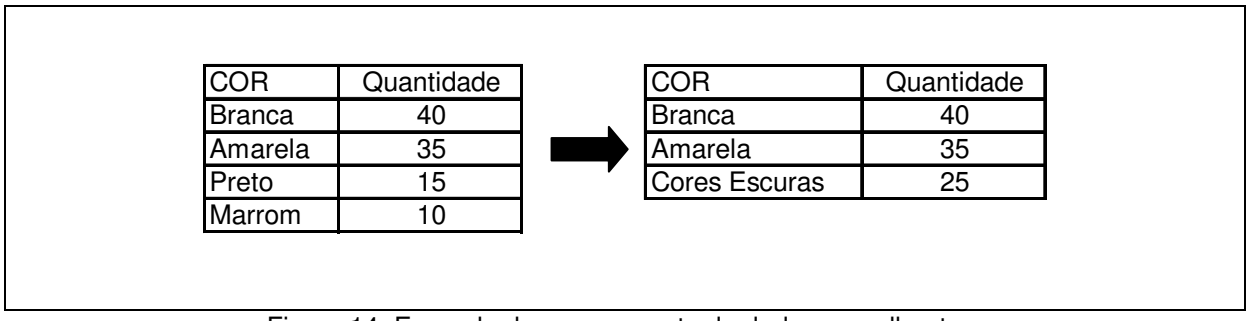


Figura 14. Exemplo de agrupamento de dados semelhantes.

Esses agrupamentos podem ser definidos por alguém que conheça profundamente o contexto da base de dados ou automaticamente, com a utilização

de técnicas automatizadas de clusterização capazes de encontrar um agrupamento natural dos dados.

Nesta metodologia proposta, a recomendação é a de que, primeiramente, haja uma tentativa para a realização da tarefa de clusterização como meio de agrupar os dados dos atributos com a aplicação de uma técnica automatizada de *data mining*. Sugere-se a utilização do algoritmo K-Means, que é uma metodologia muito eficiente na tarefa de clusterização.

Para a aplicação do K-Means, os dados que se pretende agrupar são selecionados e extraídos da base de dados e em seguida submetidos ao processamento da técnica. Esse processamento pode ser realizado com o uso de uma ferramenta (software) de mineração de dados que ofereça suporte para a aplicação do algoritmo K-Means. Na verdade, ocorre nesse momento uma etapa de mineração de dados que tem o objetivo principal de agrupar determinados dados para que sejam processados pelo algoritmo Apriori em seguida.

Caso não seja possível agrupar os dados com a utilização de técnicas automatizadas de clusterização, os itens de um determinado atributo podem ser agrupados com a orientação de um especialista no contexto dos dados em análise.

Alguns tratamentos também podem ser realizados para agregar maior valor ao conteúdo de um atributo e aprimorar ainda mais a busca por associações, de forma a aumentar as chances de descoberta de informações mais interessantes e valiosas. Em determinados casos, por exemplo, a data ou mesmo apenas o ano de nascimento não representam uma informação tão importante quanto à idade de uma pessoa. Com um simples cálculo, os dados de um atributo que representam datas de nascimento podem ser transformados em idades aproximadas de pessoas.

A seguir, a figura 15 mostra um exemplo de tratamento de um dado em busca de uma informação mais significativa.

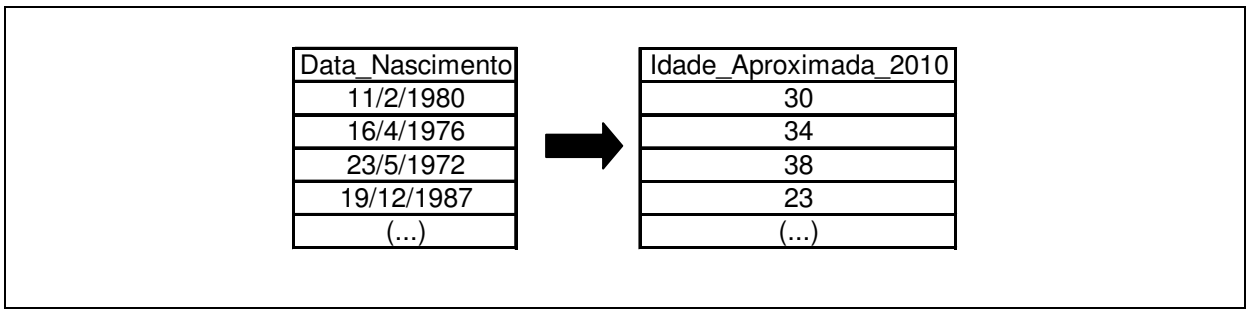


Figura 15. Exemplo de tratamento de dados em busca de uma informação mais significativa.

Há, ainda, situações em que, no mesmo conjunto de dados, dois ou mais atributos estão relacionados a uma mesma informação ou contexto. Eles podem ser processados em conjunto ou até fundidos, sem trazer prejuízos à análise. Duas datas, mesmo sendo diferentes, podem estar abrangidas por um período tão próximo que qualquer delas pode ser utilizada em uma análise. Esse tratamento pode, então, ajudar no preenchimento de dados ausentes, já que é possível processar um dado de outro atributo que, por inferência, retrata a mesma situação.

Na figura 16, pode-se observar um exemplo sobre a possibilidade de fusão de dados referentes a datas de ocorrência de dois eventos pela proximidade em que ocorrem.

Cliente	Data_Cadastro	Data_1ª_Visita
A	11/10/2007	21/10/2007
B	12/3/2008	-
C	-	1/9/2008
D	4/1/2009	-
E	-	13/10/2009

Ano_aproximado_ingresso
2007
2008
2008
2009
2009

Figura 16. Exemplo de fusão de dados com informações de datas.

Vale destacar que um atributo com dados resultantes de algum tratamento pode ser avaliado novamente sob a perspectiva de terem seus dados agrupados, caso apresente, ainda, um grande número de itens.

As transformações devem ser analisadas e realizadas cuidadosamente para que não haja o comprometimento da integridade dos dados ou, até mesmo, o surgimento de um efeito contrário ao esperado, ou seja, geração de resultados tendenciosos.

6.4- 4ª Fase

6.4.1- Segunda seleção de dados

Pelo fato de novos atributos terem sido criados através da transformação dos dados de outros, não é razoável submeter todos esses dados ao processo de mineração de dados. Os atributos que serviram de base para as transformações já

contribuíram com seus dados para a realização do processo e podem ser desconsiderados.

O principal objetivo dessa seleção é evitar redundâncias de dados, já que além de acarretar perda no desempenho de processamento do algoritmo Apriori, é provável o surgimento de associações óbvias. Nesse sentido, a preferência pela seleção recai sobre os atributos com dados transformados e mais bem trabalhados. Na verdade, o que ocorre é apenas uma substituição de atributos por outros.

Um exemplo sobre a seleção de um atributo que melhor representa a informação desejada pode ser visualizado na figura 17, a seguir.

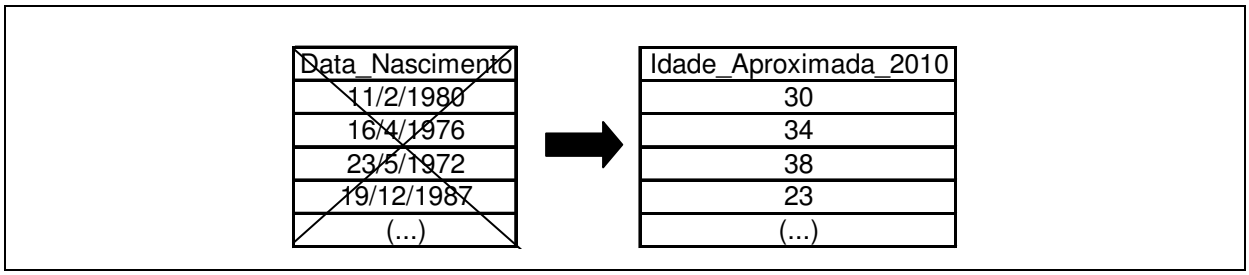


Figura 17. Exemplo de seleção dos dados mais bem trabalhados.

Deve-se evitar manter dois atributos relacionados praticamente à mesma informação.

6.4.2- Segunda limpeza dos dados

Após as transformações dos dados de vários atributos, praticamente um novo conjunto de dados é gerado. Durante a realização dessas transformações nas etapas anteriores, podem ter ocorrido erros ou falhas que resultaram no surgimento de novas inconsistências entre os dados transformados. Podem acontecer situações em que algumas incoerências apenas apareçam após as transformações realizadas. Por esses motivos, esse novo conjunto de dados precisa também ser submetido a um processo de limpeza. Essa limpeza busca corrigir eventuais inconsistências de dados surgidas, principalmente, nos atributos que passaram ou resultaram de um processo de transformação.

6.4.3- Remoção de dados duplicados

Por uma série de fatores como erros de inserção de dados ou mesmo falhas na substituição ou migração de um sistema de informação para outro, é possível encontrar na base de dados registros em que todos os seus atributos são iguais. Ao avaliar todo o contexto das atividades que originaram os dados é possível avaliar se estes registros são realmente distintos e coincidentemente iguais ou se há uma repetição de um mesmo registro.

A repetição de um mesmo registro por falhas ou sujeira na base de dados pode resultar na geração de falsas regras de associação após a aplicação de um processo de mineração de dados. Isso porque leva um algoritmo a interpretar erroneamente essas “coincidências” como sendo padrões de dados.

A retirada dos registros duplicados do processo é importante para evitar a geração de associações tendenciosas, mesmo com o risco de se remover registros apenas coincidentes.

Na figura 18, a seguir, pode-se visualizar um exemplo de um possível caso de registro duplicado, o qual deve ser removido caso haja a confirmação de que se trata de um mesmo fato.

Cliente	Idade	Quantidade	Total Compra
A	23	30	R$ 2.200,00
B	31	22	R$ 1.850,00
~~A~~	~~23~~	~~30~~	~~R$ 2.200,00~~
D	44	100	R$ 3.125,00
E	56	120	R$ 5.312,00
A	23	112	R$ 4.320,00

Figura 18. Exemplo de remoção de dados duplicados.

Registros duplicados podem ser facilmente removidos da base de dados com o uso de *queries*.

6.4.4- Segunda tabulação dos dados

O objetivo neste momento é realizar uma nova contagem dos dados de cada atributo do conjunto de dados, pois alguns atributos passaram por transformações, outros novos surgiram e, ainda, houve a retirada de registros duplicados. Também por meio de gráficos, busca-se visualizar novamente as frequências desse novo conjunto de dados gerado após as transformações e exclusões realizadas.

6.5- 5ª Fase – Seleção final de dados

Mesmo com as transformações e exclusões realizadas, ainda podem existir atributos que apresentam muitos dados ausentes. É também possível que haja outros que não ofereçam condições para apresentarem associações pelo fato de possuírem, ainda, muitas categorias com frequências abaixo do suporte mínimo definido para a aplicação do algoritmo Apriori. Nesse caso, esses atributos podem ser desconsiderados nessa última seleção.

Dessa forma, os atributos remanescentes, juntamente com seus respectivos dados, formam o conjunto de dados selecionados para participar da fase de mineração de dados, com o uso do algoritmo Apriori.

6.6- 6ª Fase – Mineração de dados

É nessa fase que os dados, previamente preparados, são efetivamente analisados pelo método de busca de associações sistematizado pelo algoritmo Apriori.

Para a aplicação do algoritmo, é preciso escolher e utilizar uma ferramenta (software) de mineração de dados, entre as disponíveis atualmente, que ofereça suporte para o uso da técnica Apriori.

Com a utilização da ferramenta, os dados são submetidos ao processamento do algoritmo Apriori, configurado com os parâmetros mínimos de suporte e confiança definidos no início do processo.

6.7- 7ª Fase – Análise e interpretação dos resultados

Após a aplicação do algoritmo Apriori para análise automática dos dados, as regras sobre padrões de associação existentes que atingiram os requisitos mínimos de suporte e confiança são identificadas. Essas regras de associação geradas devem ser analisadas para extração de informações sobre o contexto desses padrões de associações que poderão ser investigados mais profundamente.

Essa etapa é necessária porque as regras de associação identificadas pelo algoritmo Apriori são apresentadas com a representação do tipo “item A -> item B”. É possível, ainda, haver várias regras semelhantes ou complementares, o que é razoável, por se tratar de um processamento de forma automatizada.

Assim, cada representação de uma regra é analisada diante do contexto das demais em busca de uma interpretação mais coerente em relação aos objetivos do processo de mineração de dados.

Diante dos resultados alcançados o processo poderá ser novamente realizado com novas definições de objetivos ou parâmetros de suporte e confiança.

7- UM ESTUDO DE CASO DA APLICAÇÃO DA METODOLOGIA PROPOSTA: BASE DE DADOS DA CMAC JUAREZ BARBOSA

7.1- Contexto do estudo de caso

Pela Constituição Federal de 1988, até então vigente, “a saúde é direito de todos” e cabe ao Estado garantir o acesso universal e igualitário às ações e serviços de saúde aos cidadãos (Brasil, Constituição da República Federativa do Brasil, 1988, arts. 194 a 200). Os Poderes Públicos devem instituir e organizar um conjunto de ações integradas destinadas a assegurar os direitos da população relativos à saúde. Essas ações e serviços públicos constituem o Sistema Único de Saúde (SUS), que deve ser financiado com recursos financeiros provenientes do orçamento da União, dos Estados, do Distrito Federal e dos Municípios. Na verdade, tratam-se de recursos arrecadados da própria sociedade, de forma direta e indireta.

Pela Constituição, entre outras atribuições correlatas, compete ao SUS:

- controlar e fiscalizar procedimentos e produtos da área de saúde;
- participar da produção de medicamentos, equipamentos, imunobiológicos, hemoderivados, entre outros insumos, bem como fiscalizar a produção de produtos tóxicos e radiotativos;
- coordenar a formação de recursos humanos na área da saúde;
- formular políticas de saneamento básico;
- fomentar o desenvolvimento científico e tecnológico na área de saúde;
- promover ações de vigilância sanitária e epidemiológica.

O Governo do Estado de Goiás, por exemplo, que integra o SUS atualmente, principalmente por meio das ações da Secretaria de Estado da Saúde e suas unidades descentralizadas, tem investido na área de saúde anualmente, em média, mais de 12% de sua receita líquida, além de aplicar recursos públicos provenientes de repasses do Governo Federal e de outros entes públicos através de convênios e outros tipos de pactuações. A tabela 3 e a figura 19 mostram um histórico da aplicação de recursos na área da saúde pelo Governo de Goiás.

Em R$ 1.000

Período	Recursos Próprios Aplicados		Recursos de Outras Fontes de Receita	TOTAL
	Valor Aplicado	% da Receita Líquida		
2004	441.653	12,22%	418.617	860.270
2005	461.747	12,05%	448.640	910.387
2006	519.552	12,04%	493.547	1.013.099
2007	629.316	12,88%	612.613	1.241.929
2008	700.682	12,46%	664.307	1.364.989
2009	656.445	12,24%	791.514	1.447.959

Tabela 3. Aplicação de recursos financeiros na área da saúde pelo Governo do Estado de Goiás. Fonte: Gabinete do Controle Interno da Secretaria da Fazenda do Estado de Goiás[4].

[4] Relatório Resumido de Execução Orçamentária e Financeira – Demonstrativo da Receita Líquida de Impostos e das Despesas Próprias com Saúde:

- 6º Bimestre de 2004. Disponível em: http://www.controleinterno.goias.gov.br/site/relatorios/gestao_fiscal/2004/bim6/Demonstrativo%20dos%20Gastos%20em%20Saude.pdf. Acesso em: 11 de novembro de 2009.
- 6º Bimestre de 2005. Disponível em: http://www.controleinterno.goias.gov.br/site/relatorios/gestao_fiscal/2005/bim6/Gastos%20em%20Saude.pdf. Acesso em: 11 de novembro de 2009.
- 6º Bimestre de 2006. Disponível em: http://www.controleinterno.goias.gov.br/site/relatorios/gestao_fiscal/2006/bim6/vinculacoes-saude-2006.pdf. Acesso em: 11 de novembro de 2009.
- 6º Bimestre de 2007. Disponível em: http://www.controleinterno.goias.gov.br/site/relatorios/gestao_fiscal/2007/bim6/VINC.%20SA_DE.pdf. Acesso em: 11 de novembro de 2009.
- 6º Bimestre de 2008. Disponível em: http://www.controleinterno.goias.gov.br/site/relatorios/gestao_fiscal/2008/bim6/VinculacoesSaude.pdf. Acesso em: 11 de novembro de 2009.
- 6º Bimestre de 2009. Disponível em: http://www.controleinterno.goias.gov.br/site/relatorios/gestao_fiscal/2009/bim6/SAUDE.pdf. Acesso em: 25 de abril de 2010.

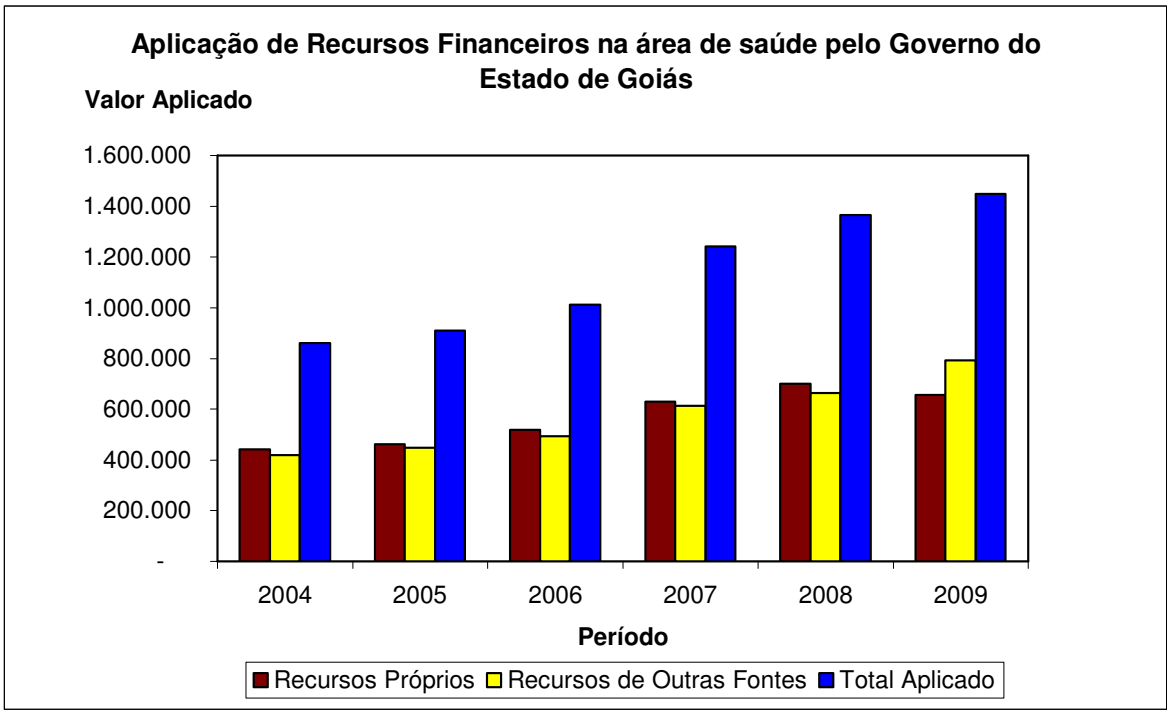


Figura 19. Aplicação de recursos financeiros na área da saúde pelo Governo do Estado de Goiás.
Fonte: Gabinete do Controle Interno da Secretaria da Fazenda do Estado de Goiás.

Observa-se que se trata de investimentos anuais com valores muito expressivos. Em 2009 os investimentos na área de saúde superaram o valor de R$ 1,4 bilhão. Para uma aplicação adequada de todo esse volume de recursos em ações, serviços e políticas públicas de saúde, para execução de atividades preventivas e assistenciais, torna-se imprescindível uma análise cuidadosa de muitas informações, sob diversas perspectivas, de forma a propiciar uma avaliação completa das necessidades da população, conforme o seu perfil e suas demandas.

Uma atividade muito importante de assistência à saúde desenvolvida pelo SUS é o fornecimento gratuito de medicamentos excepcionais para o tratamento de determinadas patologias. Estes medicamentos são extremamente necessários para melhorar a qualidade de vida e atenuar o sofrimento de muitas pessoas.

No Estado de Goiás, o fornecimento de medicamentos excepcionais é realizado pela CMAC Juarez Barbosa, unidade da Secretaria Estadual de Saúde, com base nas orientações da Portaria nº. 2.577/GM, de 27 de outubro de 2006, e suas alterações. Essa Portaria aprova o chamado Componente de Medicamentos de Dispensação Excepcional, como parte de uma política nacional de assistência farmacêutica.

A CMAC fornece, segundo informações coletadas no início do ano de 2010, 132 medicamentos diferentes. Os preços dos medicamentos, por unidade, variam de R$ 0,03 a R$ 3.537,47, com um valor médio de R$ 168,06. Vale destacar, que estes preços são menores que os praticados no mercado, já que o setor público geralmente adquire os medicamentos por um preço menor por conta do seu maior poder de compra e normas reguladoras que estabelecem os preços máximos para fornecimento a governos. A tabela 4 e a figura 20 mostram os investimentos realizados no programa de fornecimento de medicamentos excepcionais nos anos de 2008 e 2009.

Em R$ 1.000

Período	Recursos Próprios Aplicados	Recursos de Outras Fontes de Receita	TOTAL
2008	28.112	36.908	65.020
2009	4.696	47.527	52.223

Tabela 4. Aplicação de recursos financeiros com aquisições de medicamentos excepcionais.
Fonte: Sistema de Execução Orçamentária e Financeira (Siofi-Net) do Estado de Goiás.

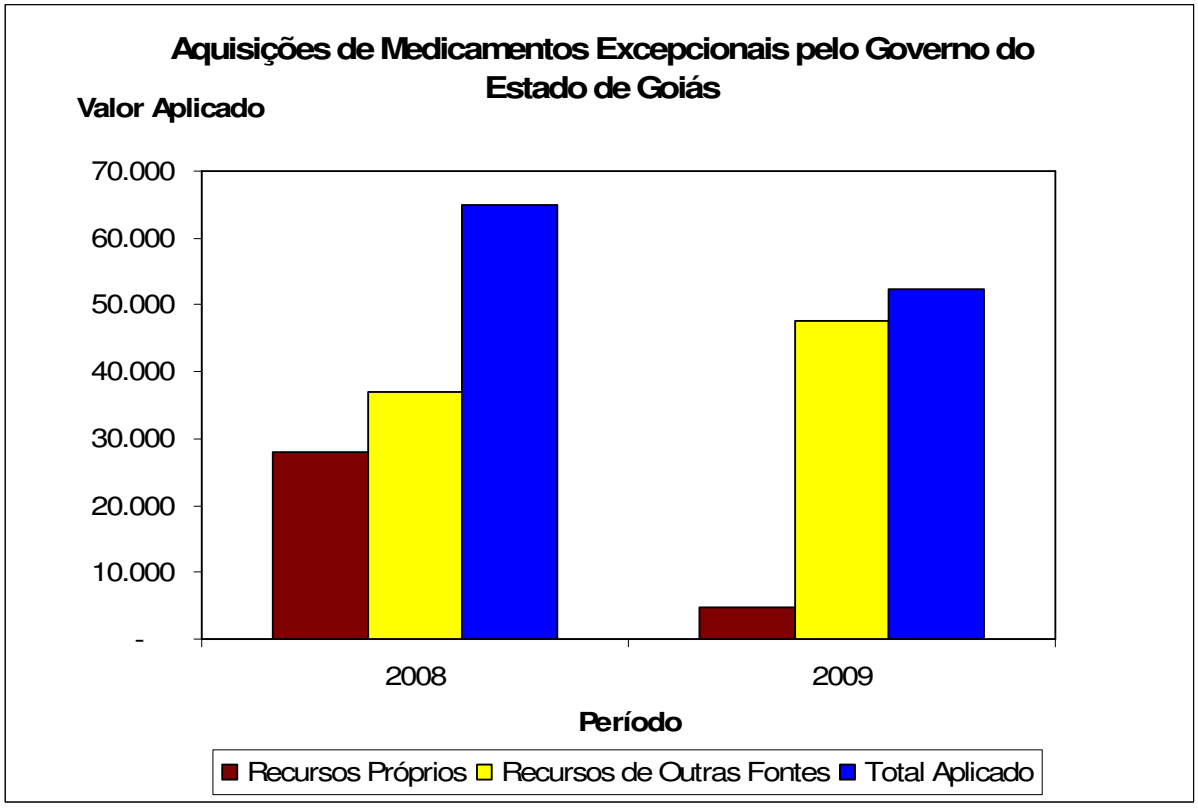


Figura 20. Aplicação de recursos financeiros com aquisições de medicamentos excepcionais.
Fonte: Sistema de Execução Orçamentária e Financeira (Siofi-Net) do Estado de Goiás.

O estudo de caso foi realizado com a aplicação da metodologia proposta neste trabalho na base de dados resultante das operações registradas através do sistema informatizado utilizado na CMAC Juarez Barbosa, que tem registrado eletronicamente, desde 1998, os atendimentos realizados pela unidade. Apenas com relação às dispensações, há mais de 700.000 registros presentes no banco de dados, além de dados relativos a outras operações da unidade.

A CMAC Juarez Barbosa é uma unidade da Secretaria Estadual de Saúde localizada, durante este estudo de caso, na Av. Tocantins c/ Rua 4, nº. 777, no Centro de Goiânia, Estado de Goiás.

São realizados na unidade, em média, cerca de 1.000 atendimentos por dia para fornecer medicamentos, realizar cadastramentos e registrar pedidos de pessoas portadoras de doenças como: hepatites “B” e “C”; esquizofrenia; esclerose múltipla; artritereumatóide; entre outras. Grande parte das pessoas atendidas apresenta doenças graves, sendo algumas consideradas raras, e uma possível falta de um medicamento pode interromper seu tratamento e causar sérios problemas à sua saúde.

Os critérios para o fornecimento de medicamentos excepcionais estão definidos na Portaria nº. 2.577/GM/2006. Conforme a Portaria, o fornecimento só deve ser realizado nos casos em que o tratamento de saúde exija uso de medicamento de alto valor unitário ou que, em caso de uso crônico ou prolongado, seja um tratamento de custo elevado para pessoas que se apresentarem nas seguintes condições:

- portadora de doença rara ou de baixa prevalência;
- portadora de doença prevalente, mas que necessariamente, ao se submeter a tratamento previsto no atendimento básico de saúde, apresentou intolerância, refratariedade ou evolução para quadro clínico de maior gravidade ou ficou evidente, pelo diagnóstico ou estabelecimento de conduta terapêutica, que a doença requer uma atenção especializada.

Resumidamente, os procedimentos adotados pela unidade para a dispensação de medicamentos excepcionais são:

1- inicialmente, é protocolado um processo formal para a dispensação do medicamento. O processo é protocolado apenas com a apresentação dos seguintes documentos:

a. formulário conhecido como LME (Laudo para Solicitação/Autorização de Medicamentos de Dispensação Excepcional e Estratégicos) preenchido em 2 vias. Neste formulário são preenchidas diversas informações sobre o paciente, a doença e o profissional de saúde ou hospital que atende o paciente;
b. receita médica relativa ao uso do medicamento, prescrito pelo seu princípio ativo, com a respectiva posologia, em 2 vias;
c. relatório médico com diagnóstico detalhado sobre o histórico, o atual quadro clínico, a doença do solicitante e a previsão do tempo de uso do medicamento;
d. cópias de exames complementares relativos à investigação diagnóstica;
e. cartão do SUS, documentos pessoais e comprovantes de endereço do solicitante;

2- o processo é encaminhado para análises e acompanhamento por uma equipe de assistência e serviço social;
3- em seguida, uma comissão médica também avalia o processo quanto aos diagnósticos e exames médicos anexados e manifesta pela autorização ou indeferimento da solicitação;
4- uma equipe do serviço farmacêutico também emite seu parecer sobre a autorização ou indeferimento da solicitação;
5- através do sistema informatizado é emitido um “cartão” ao paciente com validade de até 90 (noventa) dias, que dispensa a apresentação de novos documentos e diagnósticos médicos durante esse período e garante o fornecimento dos medicamentos nas quantidades autorizadas;
6- conforme a autorização da LME, são agendadas as datas para que, no período de validade do cartão, os pacientes possam retirar os medicamentos autorizados.

A figura 21 mostra o fluxo da solicitação de medicamentos excepcionais na CMAC Juarez Barbosa.

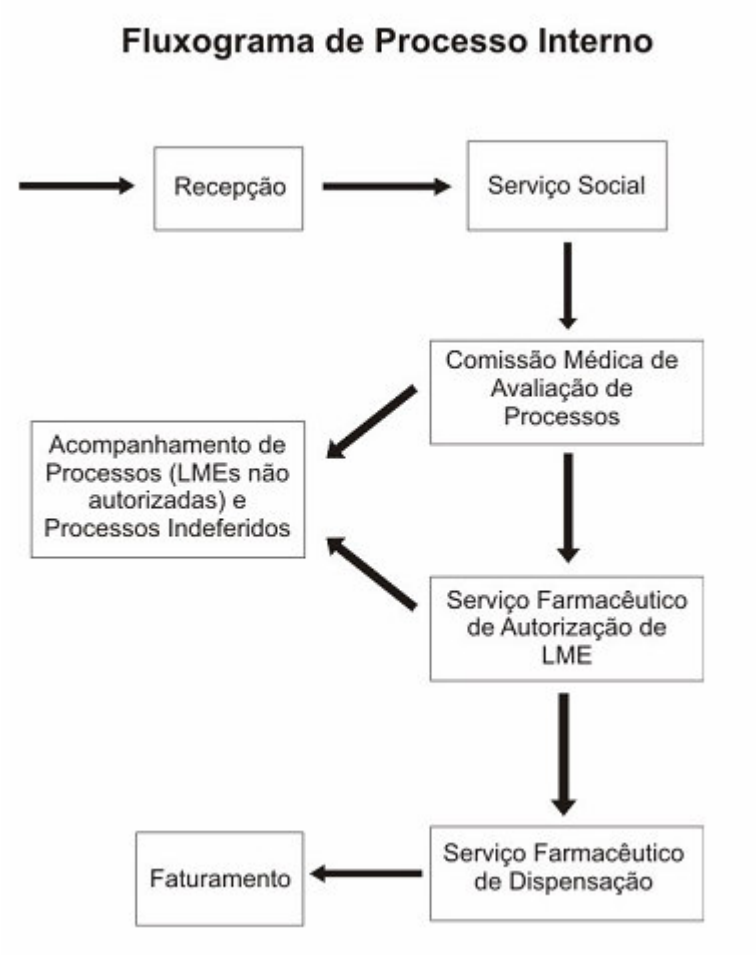


Figura 21. Fluxo do processo interno de dispensação de medicamentos excepcionais na unidade de saúde CMAC Juarez Barbosa, disponível em http://www.saude.go.gov.br/index.php?idEditoria=873, acesso em 04 de janeiro de 2010.

Há pacientes de diversas localidades, sendo que alguns, por apresentarem dificuldades de se locomover até o ponto de distribuição, muitas vezes autorizam parentes e amigos a receber o medicamento.

Por esses e outros motivos, é muito interessante realizar uma análise dos atendimentos realizados de forma a extrair o máximo de informações sobre as dispensações de medicamentos realizadas. As informações podem ser muito úteis para melhorar as políticas de atendimento a essas pessoas e reduzir os problemas referentes a abastecimentos, já que não pode haver falta de medicamentos e nem

aquisições em excessos que possam acarretar perdas pelo vencimento de prazos de validade. Com base em informações, pode-se até mesmo adotar ações preventivas de saúde com o objetivo de reduzir a demanda por este tipo de medicamento, responsável por uma despesa tão grande.

Entretanto, analisar, investigar e tirar conclusões dessa grande quantidade de registros sem o uso de uma ferramenta de tecnologia de informação passa a ser uma tarefa muito difícil, trabalhosa e demorada.

Nesse sentido, o uso de técnicas e ferramentas de mineração de dados pode ser uma alternativa. Vários autores mostram que a mineração de dados pode ser aplicada com sucesso em áreas relacionadas à proteção da saúde (Adriaans *et al.*, 1996, p. 82; Bigus, 1996, p. 20; Fayyad *et al.*, 1996, p. 23; Kantardzic *et al.*, 2005, p. 11; Elmasri *et al.*, 2005, p. 640; Han *et al.*, 2006, p. 654).

7.2- Base de dados

Desde 1998, as atividades da CMAC Juarez Barbosa passaram a ser executadas com o auxílio de sistema informatizado. O sistema atual e sua base de dados, que armazena o histórico de dados disponíveis desde 1998, são gerenciados pela equipe de informática da Secretaria Estadual de Saúde. Esse banco de dados é relacional e é operacionalizado por um sistema gerenciador de banco de dados nos servidores da Secretaria.

Para realização deste estudo de caso, foi disponibilizada uma cópia completa de um *backup* dessa base de dados realizado no mês de janeiro do ano de 2010. Como acontece comumente, os dados fazem parte de uma base transacional que sofre constantes atualizações ao longo do tempo e, por isso, este é um procedimento muito importante, já que não é recomendável manipular e processar os dados diretamente no banco de dados que serve a um sistema de informação para que não haja interferência nas rotinas operacionais relacionadas a estes dados (Passos *et al.*, 2005, p. 26). Adriaans *et al.* (1996, p. 39) também ensinam que para facilitar todo o processo de *data mining* uma cópia dos dados deve ser extraída e armazenada em um banco de dados separado.

Observou-se, em uma primeira análise, que a base de dados sofreu mudanças ao longo do tempo. Foi informado pela equipe de informática que houve

trocas de sistemas de informação e que o atual está sendo aprimorado para se adequar às novas realidades da unidade. Novas funcionalidades estão sendo desenvolvidas. Isso explica a razão da base possuir algumas tabelas e campos que apresentam dados decorrentes apenas de operações mais recentes, enquanto outros componentes não estão atualizados pelo fato de terem sido substituídos por novas tabelas e campos criados.

O banco de dados apresenta mais de 80 tabelas. Estão armazenadas informações sobre atividades realizadas para registrar processos protocolados, autorizar solicitações, dispensar medicamentos, efetuar compras, prestar contas, entre outras relacionadas aos procedimentos da unidade. Também há tabelas com dados que apenas são cópias de segurança, as que são utilizadas temporariamente pelo sistema e as que nem sequer são utilizadas com frequência, mas permanecem na base para armazenar dados antigos para serem recuperados, se necessário.

7.3- Aplicação da metodologia

Para o estudo de caso sobre a verificação do desempenho do algoritmo Apriori na busca por possíveis associações existentes entre os dados do banco de dados da CMAC Juarez Barbosa, foi realizado um experimento com a aplicação completa da metodologia proposta neste trabalho, denominado experimento 1.

A escolha dos tipos de equipamentos e softwares utilizados para realização deste experimento não influenciaram diretamente nos resultados deste trabalho. Entretanto, vale apresentar esses recursos utilizados como forma de demonstrar que a realização de um processo de mineração de dados está mais acessível e sem maiores custos financeiros com aquisição de softwares e equipamentos. Há muitos softwares livres que podem ser utilizados nesse processo e computadores mais modestos também podem realizar todo o trabalho, embora com um menor desempenho no que se refere ao tempo de processamento. Os maiores investimentos em equipamentos e softwares talvez possam se justificar em casos específicos de manipulação de imensas bases de dados.

O experimento foi realizado com o uso de um computador do tipo *notebook* que apresentava as seguintes configurações:

- processador da marca Intel, modelo Core 2 Duo de 2.1 Ghz;

- memória física de 4 Megabytes;
- disco rígido de 320 Gigabytes;
- sistemas operacionais Windows XP Professional Service Pack 3 32 bits e Linux Ubuntu 9.04 64 bits em *dual boot*.

A ferramenta de *data mining* escolhida para realização do experimento foi o software WEKA (do acrônimo, *Waikato Environment for Knowledge Analysis*), versão 3.6.0, desenvolvido pela Universidade Waikato, situada na Nova Zelândia. A ferramenta foi escolhida, principalmente, pelas seguintes razões:

- é reconhecido e muito utilizado pelo meio acadêmico;
- possui uma interface gráfica bastante intuitiva, o que torna fácil a sua utilização;
- apresenta muitos recursos para análise e mineração de dados, inclusive métodos para visualização (através de tabelas e gráficos) e pré-processamento dos dados;
- é bem completo em relação a técnicas para realização de tarefas de *data mining* de associação, classificação e clusterização. Para cada tipo de tarefa há vários algoritmos diferentes;
- pode manipular dados com o acesso direto aos sistemas de banco de dados ou arquivos no formato texto, como os do tipo *.csv;
- é *open source* (código livre) e está disponível na Internet no endereço http://www.cs.waikato.ac.nz/ml/weka/;
- foi desenvolvido na linguagem Java, o que torna possível utilizá-lo em diferentes sistemas operacionais.

O software WEKA foi utilizado tanto para a aplicação do algoritmo Apriori como da técnica K-Means.

Os dados foram inicialmente manipulados, principalmente nas atividades da etapa de pré-processamento, com o uso do sistema gerenciador de banco de dados Mysql, versão 5.1.38, através das ferramentas com interface gráfica OpenOffice Portable Base 3.1.0 e HeidiSql 4.0 Portable, sob o ambiente operacional do sistema Microsoft Windows XP Professional. Vale destacar que a equipe de informática da Secretaria Estadual de Saúde adota o Mysql como servidor de banco de dados do sistema de informação utilizado pela CMAC Juarez Barbosa e faz sua manutenção com o auxílio do HeidiSql. Por esta razão, essas ferramentas também

foram adotadas para realização do experimento para facilitar o acesso e garantir maior integridade aos dados.

Apenas nos momentos de utilização do software WEKA para a aplicação e processamento das técnicas de mineração de dados o computador foi inicializado sob a plataforma Linux de 64 bits.

A utilização de sistemas operacionais de 64 bits possibilita o uso de uma maior quantidade de memória do computador pela *JVM*[5] (*Java Virtual Machine*) *JVM.* Esses sistemas garantem um maior espaço da memória do computador à *JVM*, que é o ambiente utilizado para carregar um sistema desenvolvido em Java.

Ao contrário, a maioria dos sistemas operacionais de 32 bits limita o espaço na memória do computador em menos de 2 Megabytes para o uso da *JVM*, o que impossibilita ao software WEKA carregar um volume de dados muito grande. O fato de ter sido desenvolvido em Java trouxe ao WEKA a desvantagem de apresentar menor desempenho de processamento e alto consumo de memória, se comparado com outros softwares desenvolvidos em outras linguagens.

Assim, com o uso do Linux, sob uma plataforma de 64 bits, a limitação de memória foi superada e foi possível ao software WEKA carregar e manipular a base de dados em estudo.

7.3.1- 1ª Fase

Definição dos Objetivos e da Entrada do Processo de data mining

A base de dados foi fornecida completa, portanto possui dados de diversas atividades realizadas pela CMAC Juarez Barbosa, além de outros referentes a *backups* e controles e configurações do sistema informatizado. No entanto, como já foi abordado (Elmasri *et al.*, 2005, p. 625; Adriaans *et al.*, 1996, p. 81; Westphal *et al.*, 1998, p. 25), é preciso definir um objetivo para o processo de

[5] "Máquina virtual Java (do inglês Java Virtual Machine - JVM) é um programa que carrega e executa os aplicativos Java, convertendo os bytecodes em código executável de máquina. A JVM é responsável pelo gerenciamento dos aplicativos, à medida que são executados. Graças à máquina virtual Java, os programas escritos em Java podem funcionar em qualquer plataforma de hardware e software que possua uma versão da JVM, tornando assim essas aplicações independentes da plataforma onde funcionam" (Wikipédia, 2010).

mineração de dados para que haja um foco e um direcionamento de todas as atividades do processo.

Após analisar a base de dados, percebeu-se que os dados relacionados à atividade de dispensação de medicamentos excepcionais oferecem melhores condições para a execução do processo de mineração de dados, principalmente porque há um histórico maior (desde 1998) e o seu preenchimento está mais adequado. Já quanto aos dados de outras atividades, há os casos daqueles que se referem apenas a um histórico mais recente e os demais que, geralmente, apresentam altas taxas de ausência, erros e perdas. Também vale destacar que a dispensação de medicamentos excepcionais é a atividade de maior importância para CMAC, já que é a principal atividade finalística da unidade.

Nesse sentido, com a anuência dos diretores da unidade, a atividade de dispensação de medicamentos excepcionais foi escolhida como base para a aplicação do processo de *data mining* neste estudo de caso.

Vale ressaltar dois dos principais momentos do processo de dispensação: a solicitação do medicamento pelo paciente e a efetiva entrega do medicamento. Dessa forma, algumas situações são possíveis:

1- a solicitação pode não ser autorizada ou ser atendida parcialmente;

2- por algum motivo a entrega pode não ocorrer ou a dispensação autorizada não ser realizada completamente, conforme a autorização;

3- várias entregas podem ser realizadas após a autorização, em momentos diferentes;

4- a dispensação pode ser efetivada completamente, conforme autorizado.

Assim, ficou estabelecido como objetivo desse processo, a busca por possíveis associações entre os dados com informações apenas sobre as solicitações de medicamentos, independentemente dos consequentes tratamentos e encaminhamentos posteriores. Dessa forma, não foram considerados nesse estudo, por exemplo, os dados referentes às autorizações, indeferimentos ou entregas.

Os principais dados das solicitações têm origem no registro das informações da solicitação do paciente encaminhada através de um formulário próprio utilizado pela CMAC, que o sistema informatizado trata como informações do Laudo para Solicitação de Medicamentos Excepcionais (LME), também conhecido, por assim ser identificado anteriormente, pela sigla SME.

A expectativa foi a de encontrar possíveis associações existentes entre os dados envolvidos e relacionados às informações sobre as solicitações de medicamentos, como os dos pacientes, dos medicamentos solicitados, das doenças diagnosticadas, entre outros.

Este é um cenário propício para a utilização do algoritmo Apriori para busca de associações durante a etapa de mineração de dados.

Definição do suporte e grau de confiança mínimos

Um dos requisitos para a utilização do algoritmo Apriori é a prévia definição dos graus de confiança e de suporte mínimos para servirem de parâmetros para as buscas de associações, ou seja, as regras que não atendam às condições mínimas pré-estabelecidas são desconsideradas. Apenas para possibilitar uma análise e comparação de resultados, foram considerados um suporte mínimo de 10% (dez por cento) e um grau de confiança mínimo de 80% (oitenta por cento) para a geração de possíveis regras de associação.

7.3.2- 2ª Fase

1ª Seleção de dados

Inicialmente, a base de dados foi analisada para verificar e extrair os dados relacionados ao contexto do objetivo estabelecido para o processo de mineração de dados que, no caso, se tratava do universo completo de informações referentes às solicitações de medicamentos.

Assim, foram identificadas as tabelas no banco de dados que armazenam os dados referentes às solicitações de medicamentos. A tabela que identifica detalhadamente cada medicamento solicitado é a sme_medicamentos, que está relacionada a outras do banco de dados que possuem as demais informações da solicitação, como os dados do paciente e informações adicionais dos medicamentos. Também com a orientação da equipe de informática da Secretaria de Estado da Saúde, os relacionamentos foram identificados, conforme demonstrado na figura 22, da página 116.

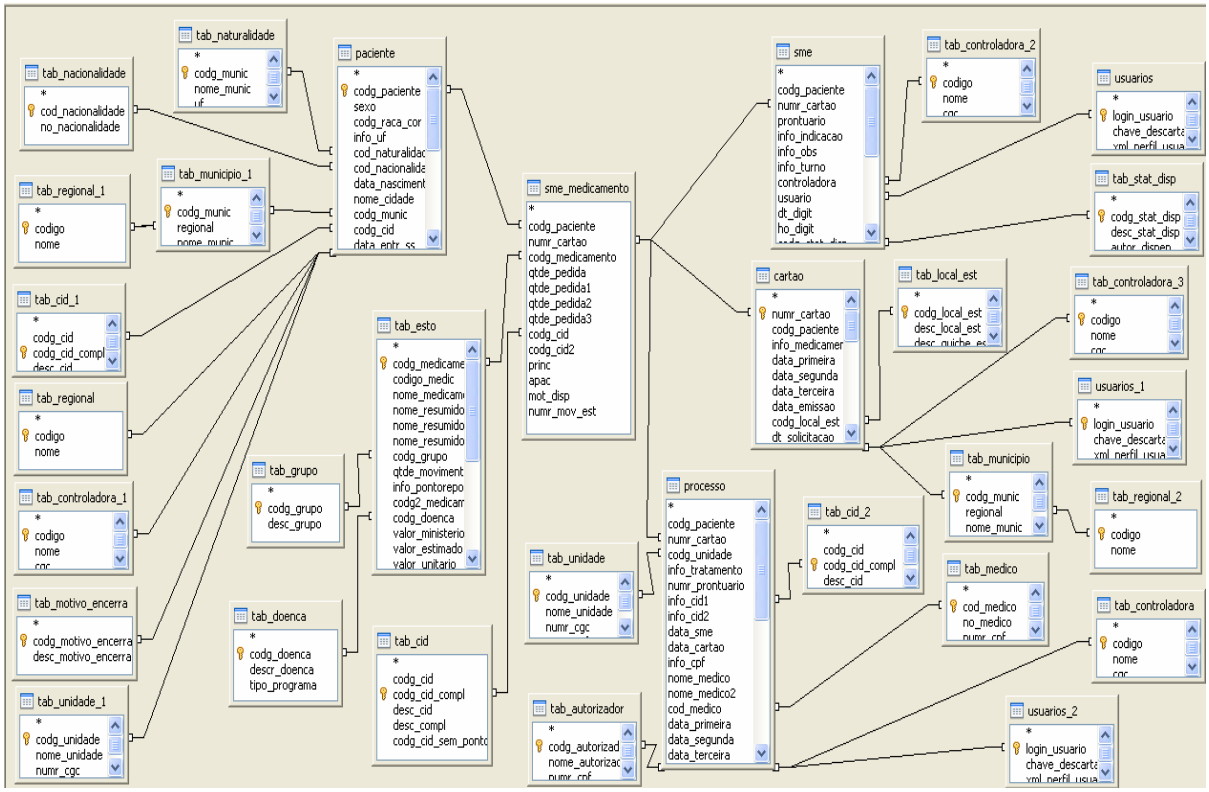


Figura 22. Relacionamentos da tabela sme_medicamentos.

Todas essas tabelas, com seus respectivos relacionamentos, formam um conjunto de registros de dados (as linhas) sobre as solicitações de medicamentos, organizados através de seus respectivos campos (as colunas). Para se obter um histórico completo de solicitações, foi feita, através de uma *query* (realização de uma consulta no banco de dados), a extração de todos os respectivos registros de dados presentes na base. Dessa operação, resultou um grande conjunto de registro de dados históricos sobre as solicitações de medicamentos formatado em uma única tabela em que as linhas são os registros (instâncias) das solicitações e as colunas são os seus respectivos atributos (características).

Conforme Witten *et al.* (2005, p. 52), o primeiro procedimento a ser realizado é trazer os dados disponíveis para uma mesma série de dados.

Bigus (1996, p. 44) ensina que os dados a serem utilizados no processo de *data mining* podem ser manipulados tanto em um banco de dados como em um arquivo de texto. Mesmo quando os dados já estão armazenados em um banco de dados, é comum exportá-los para um arquivo do tipo "*flat*" (um arquivo que apresenta os dados em apenas duas dimensões, a matriz instâncias x atributos, segundo Witten *et al.,* 2005, p. 45) para uso de algoritmos de mineração de dados.

Isso pode evitar perda de desempenho no processamento dos dados se a base de dados for muito grande.

Esse grande conjunto de dados passou a ser a referência para essa fase de seleção de dados. Suas principais características são:

- apresenta 746.914 registros e 255 atributos (incluindo as chaves do relacionamento que estão duplicadas);
- cada registro possui atributos relacionados a diferentes aspectos, como: do paciente (local e data de nascimento, tipo da doença, raça, cor, data de início do tratamento, etc.); da solicitação do medicamento (número do processo, data do pedido, data da autorização, autorizador, médico/hospital que atestou a doença, tipo de medicamento, quantidade solicitada, etc); do cartão fornecido (data de emissão e validade, entre outros);
- entre os 255 atributos, 63 apresentam mais de 70% dos dados ausentes e 133 apresentam uma taxa de ausência de até 10%;
- 67 atributos estão no formato texto, 116 são números, 25 são datas e 47 têm outros tipos de formato;
- 41 atributos estão duplicados por serem chaves para o relacionamento entre tabelas.

Em seguida, o conjunto de dados previamente extraído do banco de dados foi analisado para verificar os dados dos atributos que não ofereciam boas condições para serem classificados ou tabulados e, consequentemente, não apresentavam qualquer chance de constarem em regras de associação.

Dessa forma, foram retirados do processo os dados dos atributos do tipo texto com preenchimento livre e não padronizado, como: nomes de médico, de usuários, de unidades, de regionais e de municípios; descrições de medicamentos, de doenças e de tratamentos; observações gerais e específicas.

Vale ressaltar que dados de atributos que apenas se tratam de identificação de pessoas ou descrição de objetos raramente são selecionados para um processo de mineração de dados (Passos *et al.*, 2005, p. 31).

Também não foram considerados os atributos duplicados no conjunto de dados extraído, os relacionados a códigos identificadores ou documentos pessoais e os que apresentavam taxa de preenchimento de dados abaixo de 10%, pois se trata

de um nível de ausência de dados muito alto para o objetivo definido para o processo.

Após uma análise minuciosa dos demais atributos, juntamente com a equipe técnica da direção da entidade, foram selecionados os atributos que apresentam dados avaliados como mais importantes em uma solicitação de medicamentos e que despertam o maior interesse pelo descobrimento de possíveis associações.

Os atributos selecionados pelos critérios informados anteriormente para a participação do processo de *data mining* deste experimento são os seguintes, conforme demonstrado na tabela 5:

Identificação	Descrição do atributo
sme_medicamento-codg_medicamento	Código do medicamento solicitado.
sme_medicamento-qtde_pedida	Quantidade do medicamento solicitada.
sme-info_turno	Turno em que a solicitação foi registrada no sistema. M – Manhã e V – Vespertino.
processo-codg_unidade	Código do hospital ou médico cadastrado no sistema que recomendou a solicitação do medicamento.
processo-info_cid1	Código da doença do paciente registrado com a formalização do processo para a dispensação de medicamentos.
processo-data_sme	Data da solicitação da SME.
Cartão-guiche	Guichê que emitiu o cartão para início da dispensação.
Cartão-codg_munic	Código do município do endereço registrado no momento da emissão do cartão.
Cartão-peso	Peso do paciente.
Cartão-altura	Altura do paciente.
paciente-sexo	Sexo do Paciente. 1- Masculino e 2- Feminino
paciente-codg_raca_cor	Raça/Cor do Paciente. 01 - BRANCA; 02 - PRETA; 03 - PARDA; 04 - AMARELA; 05 - INDIGENA
paciente-cod_naturalidade	Código do município de naturalidade registrado no momento do cadastro do paciente.
paciente-data_nascimento	Data de nascimento do paciente.
paciente-data_entr_ss	Data da entrada do paciente no programa de dispensação.
tab_esto-codg_unidade	Unidade adotada para definir a quantidade do medicamento. Exemplos: comprimido, ampola, frasco, caixa, seringa, etc.

Tabela 5. Descrição dos atributos selecionados através da 1ª seleção.

1ª Limpeza dos dados

Várias *queries* foram executadas com o objetivo de identificar inconsistências presentes entre os dados selecionados. Foram identificadas situações absurdas e improváveis quanto ao preenchimento de dados, como anos improváveis anteriores a 1900 entre as datas de nascimento e de entrada do paciente, além das datas da solicitação. Situações impossíveis também ocorreram nas especificações de guichês e turnos, registros da raça ou cor do paciente, entre

outros atributos. Estas são falhas mais comumente decorrentes de erros de digitação.

Quando possível, as correções foram realizadas através de ajustes ou preenchimentos baseados em inferências sobre as situações encontradas e nas orientações da equipe de diretores da unidade. Nos demais casos, os dados simplesmente foram removidos e considerados como ausentes. Também foram removidos vários registros pelo fato de que boa parte de seus dados estavam ausentes, o que os tornavam inapropriados para participação de um processo de mineração de dados.

1ª Tabulação e visualização dos dados

Após a realização dos procedimentos das etapas anteriores, resultaram 746.355 registros com informações organizadas nos 16 atributos selecionados. Entre estes atributos foram verificadas as seguintes situações:

- somente 5 atributos possuíam um preenchimento abaixo de 60 % e 7 apresentavam mais de 90% dos dados preenchidos;
- a maior quantidade de itens com frequência acima de 10% do total de registros foi 3, encontrada em apenas 3 atributos;
- 4 atributos apresentavam uma concentração de mais de 75% da frequência dos dados preenchidos nos 3 itens com maior frequência;
- 5 atributos tinham diferença de mais de 20% entre o 1º e 3º item com maior frequência.

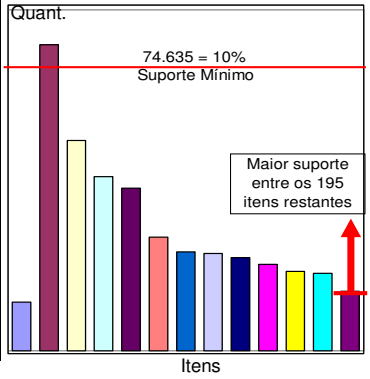
A tabela 6 e a figura 23 apresentam algumas das principais informações, inclusive através de gráficos, sobre os atributos selecionados:

Identificação	Taxa de ausência de dados	Número de itens distintos	Quant. de itens com frequência acima de 10%	Quant. de itens com frequência acima de 5%	Quant. de itens com frequência acima de 1%
sme_medicamento-codg_medicamento	1,7%	207	1	4	24
sme_medicamento-qtde_pedida	3,3%	427	2	4	15
sme-info_turno	44,3%	2	2	2	2
processo-codg_unidade	5,4%	6.242	0	2	9
processo-info_cid1	25,3%	905	1	4	15
processo-data_sme	2,1%	3.357	0	0	0
cartao-guiche	24,2%	41	3	4	6
cartao-codg_munic	56,8%	240	1	1	4
cartao-peso	68,6%	402	0	0	6
cartao-altura	69,4%	345	0	0	9
paciente-sexo	10,7%	2	2	2	2
paciente-codg_raca_cor	43,0%	5	2	2	3
paciente-cod_naturalidade	22,3%	1.656	1	1	7
paciente-data_nascimento	0,1%	27.518	0	0	0
paciente-data_entr_ss	0,9%	3.241	0	0	0
tab_esto-codg_unidade	1,8%	11	3	5	6

Tabela 6. Taxa de ausência de dados e quantidade de itens mais frequentes nos atributos selecionados nesta 2ª fase da metodologia.

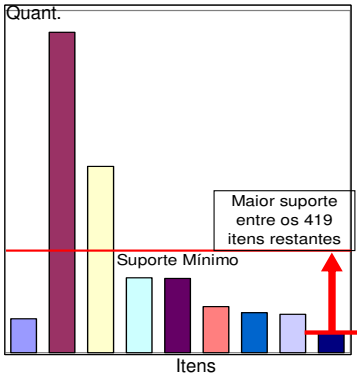
Atributo: sme_medicamento-codg_medicamento

	Item	Quant.	%
?	?	12.895	1,7%
1	14300	80.741	10,8%
2	10510	55.516	7,4%
3	16860	45.979	6,2%
4	16430	42.930	5,8%
5	36031020	29.999	4,0%
6	14570	26.160	3,5%
7	16350	25.694	3,4%
8	18210	24.643	3,3%
9	50610	22.819	3,1%
10	14900	21.027	2,8%
11	29680	20.395	2,7%
12	36381020	15.703	2,1%
13	10430	14.686	2,0%
14	16780	10.754	1,4%
15	15460	10.430	1,4%
16	18480	9.175	1,2%
17	36081120	9.134	1,2%
18	28440	8.602	1,2%
19	17670	8.573	1,1%
20	36381080	8.234	1,1%
(...)			



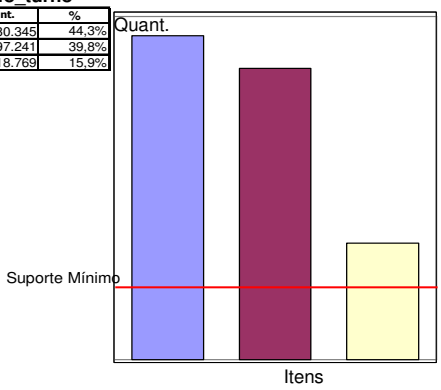
Atributo: sme_medic-qtde_pedida

	Item	Quant.	%
?	?	24.887	3,3%
1	90	234.251	31,4%
2	180	136.232	18,3%
3	36	54.758	7,3%
4	3	54.285	7,3%
5	270	33.756	4,5%
6	12	29.389	3,9%
7	360	28.094	3,8%
8	18	14.981	2,0%
9	6	14.822	2,0%
10	540	12.349	1,7%
11	27	10.011	1,3%
12	24	9.310	1,2%
13	135	8.580	1,1%
14	45	8.579	1,1%
15	60	7.996	1,1%
16	84	6.267	0,8%
17	30	6.045	0,8%
18	2	5.806	0,8%
19	1	5.441	0,7%
20	120	4.952	0,7%
(...)			



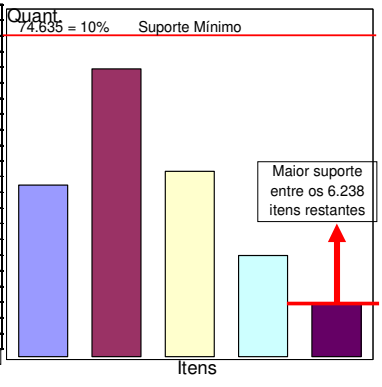
Atributo: sme-info_turno

	Item	Quant.	%
	?	330.345	44,3%
1	M	297.241	39,8%
2	V	118.769	15,9%



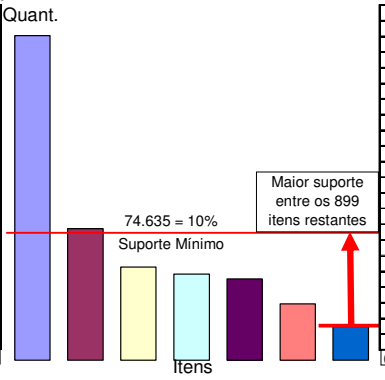
Atributo: processo-codg_unidade

	Item	Quant.	%
	?	39.940	5,4%
1	1082	67.053	9,0%
2	60	43.169	5,8%
3	72	23.630	3,2%
4	1724	12.505	1,7%
5	787	10.767	1,4%
6	52230	9.034	1,2%
7	28800	8.454	1,1%
8	1147	7.953	1,1%
9	1001752	7.669	1,0%
10	450	7.068	0,9%
11	1001932	7.039	0,9%
12	1000127	5.938	0,8%
13	1000337	5.804	0,8%
14	55468	5.557	0,7%
15	1520	5.459	0,7%
16	1001948	5.277	0,7%
17	1001615	4.523	0,6%
18	12211	4.358	0,6%
19	1000706	4.197	0,6%
20	1004795	3.999	0,5%
(...)			



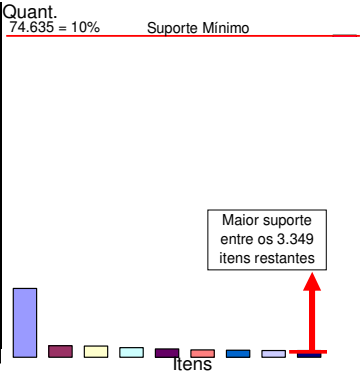
Atributo: processo-info_cid1

	Item	Quant.	%
	?	189.187	25,3%
1	N18.0	76.569	10,3%
2	Z94.0	54.415	7,3%
3	M81.0	50.208	6,7%
4	F20.0	47.310	6,3%
5	N18.9	32.701	4,4%
6	E78.0	19.547	2,6%
7	M81.8	19.247	2,6%
8	M80.0	16.199	2,2%
9	C61	14.572	2,0%
10	F20.5	10.381	1,4%
11	B18.2	10.337	1,4%
12	E23.0	10.274	1,4%
13	F29	10.006	1,3%
14	M81.9	8.766	1,2%
15	F20.9	8.729	1,2%
16	F20.8	6.939	0,9%
17	M32.1	6.473	0,9%
18	G20	6.330	0,8%
19	G30.0	5.695	0,8%
20	E22.8	5.642	0,8%
(...)			



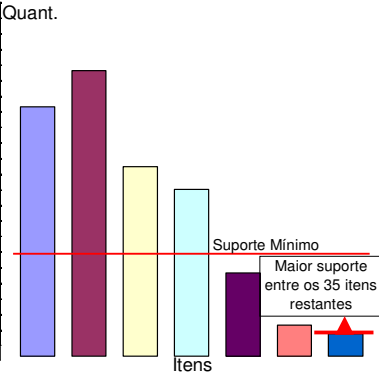
Atributo: processo-data_sme

	Item	Quant.	%
	?	15.943	2,1%
1	1/6/2006 00:00	2.628	0,4%
2	2/1/2008 00:00	2.610	0,3%
3	1/9/2006 00:00	2.252	0,3%
4	3/7/2006 00:00	1.884	0,3%
5	2/5/2006 00:00	1.703	0,2%
6	2/10/2006 00:00	1.598	0,2%
7	1/2/2008 00:00	1.523	0,2%
8	1/4/2008 00:00	1.505	0,2%
9	1/11/2006 00:00	1.494	0,2%
10	3/4/2006 00:00	1.476	0,2%
11	3/9/2007 00:00	1.469	0,2%
12	2/4/2007 00:00	1.418	0,2%
13	1/11/2007 00:00	1.417	0,2%
14	2/5/2007 00:00	1.397	0,2%
15	1/6/2007 00:00	1.381	0,2%
16	2/7/2007 00:00	1.377	0,2%
17	1/10/2007 00:00	1.366	0,2%
18	1/12/2006 00:00	1.315	0,2%
19	1/3/2007 00:00	1.292	0,2%
20	15/1/2008 00:00	1.251	0,2%
(...)			



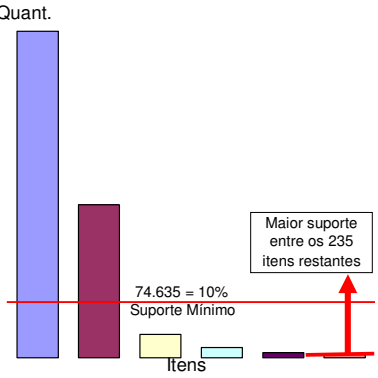
Atributo: cartao-guiche

	Item	Quant.	%
	?	180.943	24,2%
1	1	207.419	27,8%
2	2	137.563	18,4%
3	3	121.027	16,2%
4	4	60.351	8,1%
5	5	22.588	3,0%
6	6	16.341	2,2%
7	7	18	0,0%
8	20	11	0,0%
9	30	10	0,0%
10	8	9	0,0%
11	9	9	0,0%
12	12	7	0,0%
13	34	6	0,0%
14	24	5	0,0%
15	13	4	0,0%
16	29	4	0,0%
17	49	4	0,0%
18	10	3	0,0%
19	46	3	0,0%
20	84	3	0,0%
(...)			



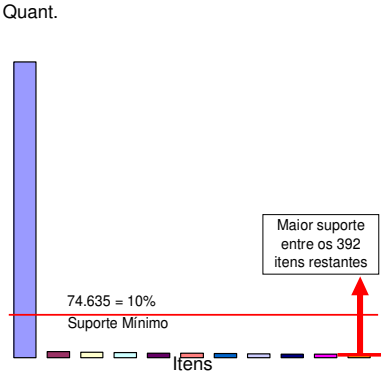
Atributo: cartao-codg_munic

	Item	Quant.	%
	?	423.766	56,8%
1	520870	198.726	26,6%
2	520140	30.421	4,1%
3	520110	13.493	1,8%
4	522140	7.226	1,0%
5	522045	4.483	0,6%
6	521880	4.417	0,6%
7	521000	2.827	0,4%
8	521150	2.230	0,3%
9	521190	1.821	0,2%
10	520860	1.777	0,2%
11	520510	1.740	0,2%
12	521930	1.680	0,2%
13	521450	1.533	0,2%
14	521250	1.418	0,2%
15	520450	1.322	0,2%
16	520880	1.243	0,2%
17	522010	1.199	0,2%
18	520330	1.194	0,2%
19	520130	1.172	0,2%
20	522185	1.162	0,2%
(...)			



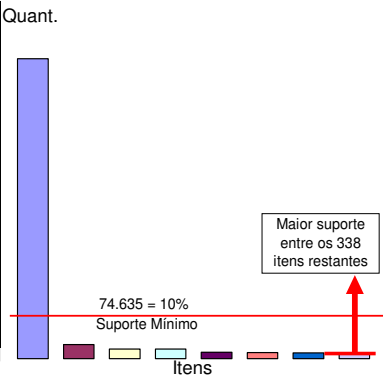
Atributo: cartao-peso

	Item	Quant.	%
	?	511.993	68,6%
1	60	11.090	1,5%
2	70	10.161	1,4%
3	65	9.251	1,2%
4	68	7.836	1,0%
5	62	7.785	1,0%
6	58	7.109	1,0%
7	80	6.938	0,9%
8	55	6.257	0,8%
9	63	6.090	0,8%
10	64	6.020	0,8%
11	75	5.797	0,8%
12	72	5.774	0,8%
13	56	5.551	0,7%
14	50	5.522	0,7%
15	52	5.292	0,7%
16	54	4.854	0,7%
17	67	4.736	0,6%
18	66	4.733	0,6%
19	57	4.647	0,6%
20	78	4.607	0,6%
(...)			



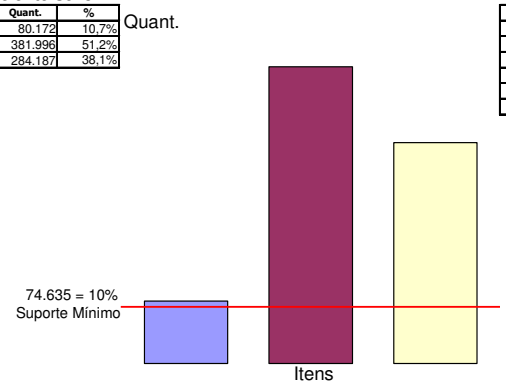
Atributo: cartao-altura

	Item	Quant.	%
	?	517.685	69,4%
1	160	24.378	3,3%
2	165	16.510	2,2%
3	170	16.437	2,2%
4	150	11.384	1,5%
5	155	10.339	1,4%
6	168	10.091	1,4%
7	162	8.764	1,2%
8	175	7.975	1,1%
9	158	7.776	1,0%
10	163	6.632	0,9%
11	172	6.352	0,9%
12	156	5.428	0,7%
13	167	5.247	0,7%
14	180	5.196	0,7%
15	152	5.085	0,7%
16	164	4.858	0,7%
17	153	4.503	0,6%
18	157	4.471	0,6%
19	169	4.005	0,5%
20	178	3.965	0,5%
(...)			



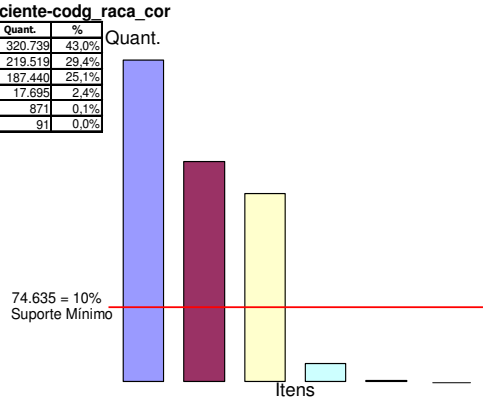
Atributo: paciente-sexo

	Item	Quant.	%
	?	80.172	10,7%
1	2	381.996	51,2%
2	1	284.187	38,1%



Atributo: paciente-codg_raca_cor

	Item	Quant.	%
	?	320.739	43,0%
1	3	219.519	29,4%
2	1	187.440	25,1%
3	2	17.695	2,4%
4	4	871	0,1%
5	5	91	0,0%



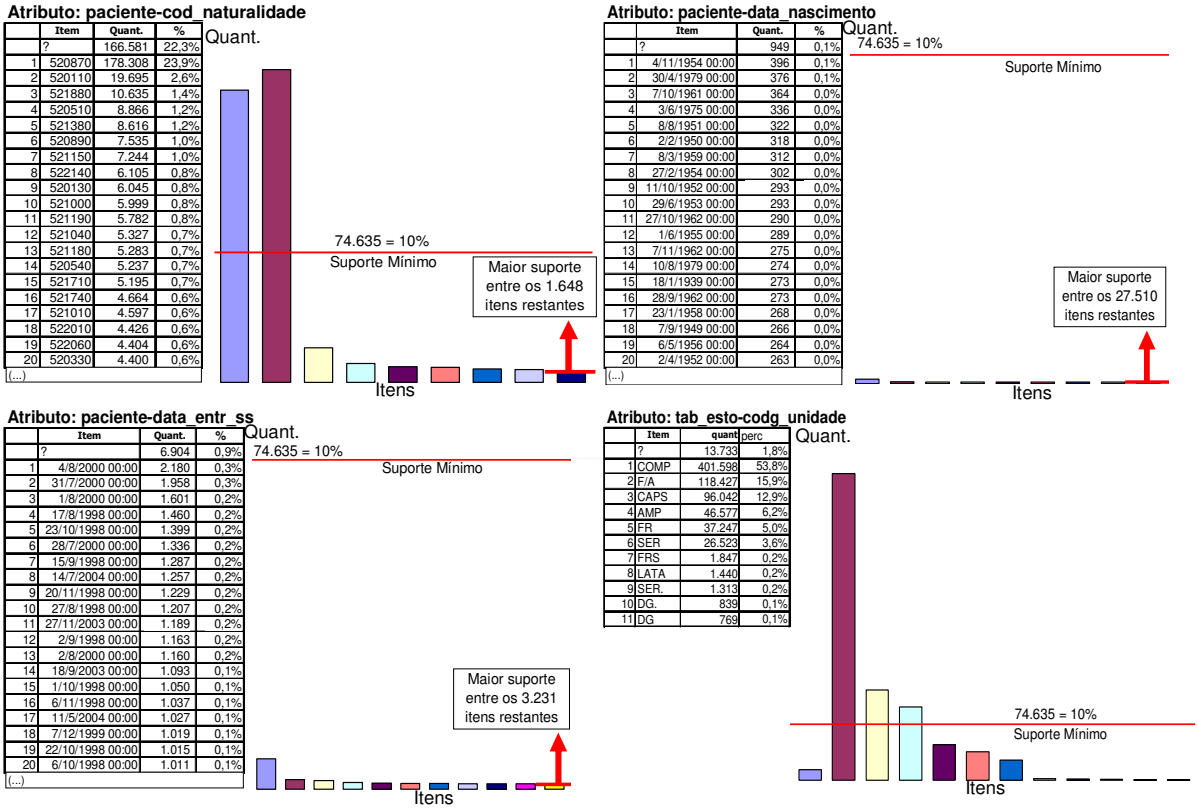


Figura 23. Gráficos com as frequências dos principais itens de cada atributo selecionado nesta 2ª fase da metodologia, com a discriminação dos 20 itens com maiores frequências.

Como forma de encontrar respostas para o principal questionamento deste trabalho em relação à influência dos procedimentos de pré-processamento nos resultados do processamento do algoritmo Apriori, neste momento os dados poderiam ser submetidos imediatamente ao processamento do algoritmo Apriori, tal como estão apresentados para que os resultados sejam comparados com os do processamento a ser realizado após a aplicação completa desta metodologia proposta.

Assim, para fins de comparação, no tópico 7.4, apresentado na página 150, esses resultados estão demonstrados.

7.3.3- 3ª Fase (transformação de dados)

Fusão de dados de atributos

Foi observado que determinados atributos selecionados para o processo de mineração de dados possuíam dados que retratavam realidades ou situações

muito próximas a de outros atributos da base de dados não selecionados. Verificou-se uma boa oportunidade para reduzir a ausência de dados de alguns atributos selecionados com a tentativa de se buscar os dados ausentes em outros da base. Sabia-se que o preenchimento de dados ausentes através dessa inferência precisava ser avaliado cuidadosamente para não comprometer a integridade das informações referentes ao atributo selecionado.

Com a orientação da equipe técnica da CMAC foi possível reduzir bem a taxa de ausência de dados de determinados atributos, como os relacionados à doença que fundamenta a solicitação do medicamento, ao período aproximado da solicitação e ao município de residência do solicitante, conforme demonstrado a seguir.

No caso da doença associada à solicitação do medicamento, observou-se que, além do código da CID (Classificação Estatística Internacional de Doenças e Problemas Relacionados com Saúde) registrado na base de dados com a efetiva geração de um processo formal para uma futura dispensação, há um código da CID relacionado à própria solicitação da LME. Verificou-se que em mais de 90% dos casos o código da CID do processo é o mesmo da LME. Assim, nota-se que quando os processos são formalizados há uma etapa de confirmação do código da CID apontado na LME e em poucos casos ocorre a sua mudança. Entretanto, observou-se que havia vários registros em que o código da CID relacionado à formalização do processo não foi informado, o que pode ter ocorrido por não ter sido preenchido em decorrência de falhas de entrada de dados, mas são grandes as chances de ser o mesmo da respectiva LME.

Considerando o fato de que dentro do contexto da dispensação de um medicamento a informação sobre a doença a ela relacionada é muito importante, houve o esforço para se obter essa informação no maior número possível de solicitações realizadas. Nesse sentido, foi realizada a fusão dos dados referentes ao código da CID do processo com os da própria LME, criando-se um atributo denominado Tr_CID. Através de uma *query,* buscou-se os dados do código da CID do processo e, na sua ausência, o código da CID da LME passou a ser utilizado. Veja-se que as informações sobre o código da CID do processo estavam ausentes em mais de 25% das solicitações, mas com o novo atributo a ausência de informações sobre a doença do paciente não passou de 2%.

Quanto ao município de residência do paciente, a situação é bem parecida. Entre os dados da emissão do cartão que dá direito ao paciente de receber os medicamentos autorizados há uma informação sobre o código do município de sua residência, mas também existe uma referência ao seu endereço no seu próprio cadastro. Verificou-se que em quase 90% das solicitações que resultaram na emissão de cartões o município informado no cartão do paciente é o mesmo do seu cadastro. É também um tipo de informação muito importante para o objetivo desse processo de mineração de dados, mas as informações do município registradas nos cartões estão ausentes em mais de 50% das solicitações. Da mesma forma, um novo atributo foi criado com a denominação Tr_Munic_Endereco com a fusão das informações do cartão e do cadastro do paciente. Buscou-se as informações do cadastro quando não disponíveis entre as do cartão. O novo atributo com informações do município de endereço do paciente apresentou uma taxa de ausência de dados de cerca de 8% apenas.

Com relação à data exata da solicitação do medicamento, pode-se deduzir que é a informada na LME, conforme registrada na base de dados. O que ocorre é que um atributo do tipo data pode receber inúmeros itens diferentes. Por exemplo, várias solicitações podem ocorrer em diferentes datas. Por isso, é muito raro uma data precisa constar em algum tipo de associação. Assim, muitas vezes não é necessário trabalhar com datas precisas durante um processo de mineração de dados, mas com períodos maiores como um ano, trimestre, mês, etc. Por exemplo, em vez de se trabalhar com a informação de que uma solicitação ocorreu em 01/12/2009, poder-se-ia optar por tratar o evento como ocorrido no ano de 2009 ou no 4º trimestre, ou seja, dentro de um período mais aproximado.

Por essa análise, observou-se que o processo de mineração de dados apresentaria melhores resultados, com mais chances para descoberta de associações, se fosse considerado o período aproximado da solicitação. Foi verificado, inclusive, após a avaliação de alguns outros atributos, que há informações sobre eventos que ocorreram em datas muito próximas a da solicitação e que poderiam ser aproveitadas para se conhecer a época da solicitação nos registros em que a informação está ausente, caso houvesse a opção por se trabalhar com um período aproximado.

Entre as informações do processo, há dados sobre a data do início de validade do cartão disponibilizado para a dispensação, a data da LME e a data de

previsão para a primeira dispensação. Há, inclusive, as informações sobre a data de emissão do cartão. Uma comparação foi realizada e mostrou que todas essas datas mencionadas são muito próximas, a grande maioria com diferenças de menos de 30 dias.

Dessa forma, a opção de se considerar um período maior e mais aproximado, em vez de exato, para tratar a época da solicitação poderia reduzir a taxa de ausência de dados sobre essa informação de 2,1% para quase 0%, com a fusão dos atributos. Nesse caso, os atributos deveriam passar por um processo de tratamento para que os períodos sejam ajustados, o que deveria ocorrer na etapa seguinte.

Tratamento de dados

Pelo contexto deste estudo de caso, observou-se também que os dados de alguns atributos poderiam ser mais bem trabalhados para representar informações mais significativas. Cada atributo selecionado, inclusive os que resultaram das fusões realizadas na etapa anterior, foi analisado sob a perspectiva da possibilidade de seus dados sofrerem alguma modificação para apresentar uma informação com maior significado.

Em decorrência das conclusões dessa análise, alguns atributos foram substituídos ou deles outros mais foram criados por meio de *queries* que realizaram cálculos e processamentos dos seus dados.

Cada processamento realizado está demonstrado a seguir.

Com relação a alguns atributos do tipo data, optou-se por trabalhar com informações aproximadas. Foi considerado apenas o ano da ocorrência do evento. No caso específico da data de solicitação, também foi trabalhado o trimestre em que o evento ocorreu.

Assim, foram criados os seguintes atributos, pelo processamento das informações da data de solicitação presentes na base de dados:

- Tr_ano_solicitacao: época aproximada em que a solicitação ocorreu, calculada pela fusão das datas de início de validade e de emissão do cartão, data da LME e a data prevista da primeira dispensação;

- Tr_trim_solicitacao: também pela fusão realizada no atributo anterior, extraiu-se, aproximadamente, o trimestre em que a solicitação ocorreu.

Os dados dos atributos sobre a data de nascimento e a data de entrada do paciente no programa de dispensação de medicamentos foram trabalhadas para apresentarem informações mais significativas para o contexto e objetivos deste processo de mineração de dados, como a idade e o tempo do paciente no programa nos diferentes momentos de solicitação de medicamentos. Das *queries* utilizadas para tratar esses dados resultaram os seguintes atributos:

- Tr_Idade_Entr_Prog: idade aproximada do paciente quando ingressou no programa de dispensação, pelo cálculo do ano de sua entrada subtraído pelo ano de seu nascimento;
- Tr_Idade_Disp: idade aproximada do paciente quando solicitou a dispensação do medicamento, pelo cálculo do ano de sua solicitação subtraído pelo ano de seu nascimento;
- Tr_Tempo_Prog_Disp: tempo aproximado do paciente no programa quando solicitou a dispensação do medicamento, pelo cálculo do ano de sua solicitação subtraído pelo ano de seu ingresso.

Há entre os dados informações sobre peso e altura do paciente, o que traz uma oportunidade para calcular o seu Índice de Massa Corporal (IMC) e verificar se o indivíduo está no peso ideal. Segundo a Wikipédia, a enciclopédia livre (2011),

> O índice de massa corporal (IMC) é uma medida internacional usada para calcular se uma pessoa está no peso ideal. Ele foi desenvolvido pelo polímata Lambert Quételet no fim do século XIX. Trata-se de um método fácil e rápido para a avaliação do nível de gordura de cada pessoa, ou seja, é um preditor internacional de obesidade adotado pela Organização Mundial da Saúde (OMS).
> O IMC é determinado pela divisão da massa do indivíduo pelo quadrado de sua altura, onde a massa está em quilogramas e a altura está em metros:
> $$IMC = \frac{massa}{(altura \times altura)}.$$

Conforme indicado na enciclopédia, o resultado deve ser analisado conforme a seguinte tabela 7, que indica o grau de obesidade do indivíduo:

IMC	CLASSIFICAÇÃO
< 18,5	Magreza
18,5 – 24,9	Saudável
25,0 – 29,9	Sobrepeso
30,0 – 34,9	Obesidade Grau I
35,0 – 39,9	Obesidade Grau II (severa)
≥ 40,0	Obesidade Grau III (mórbida)

Tabela 7. Grau de obesidade do indivíduo (Wikipédia, a enciclopédia livre, 2011).

Na própria fonte há informações sobre as limitações desse índice, especialmente com relações a crianças, adolescentes e idosos. No entanto, apenas para fins de busca de associações nesse processo de mineração de dados, o índice de massa corporal do paciente em cada registro de solicitação de medicamento foi calculado e armazenado no atributo IMC-Paciente. Em seguida, cada registro foi classificado conforme a classe do respectivo IMC através do atributo Tr_Classe_IMC.

Os locais de nascimento e de naturalidade do solicitante também são informações que necessariamente devem ser consideradas neste contexto de análise sobre a dispensação de medicamentos excepcionais, dada a sua importância para um estudo sobre a logística de distribuição.

Ao considerar o fato de que, pela política do SUS, cada Estado da federação e o DF são responsáveis pela dispensação desses medicamentos aos seus respectivos cidadãos, buscou-se identificar na base de dados os requerentes com naturalidade e residência fora do Estado de Goiás. Além disso, numa tentativa de melhor visualização dessas informações, foi realizado um trabalho de identificação dos residentes e com naturalidade em Goiás com relação às microregiões e mesoregiões definidas oficialmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE) nos seus diversos estudos sobre as regiões do Brasil.

Em seus estudos, o IBGE divide o estado de Goiás em 18 microregiões e 5 mesoregiões, conforme apresentado na tabela 8, da página 128:

MICROREGIÃO		MESOREGIÃO
1	SÃO MIGUEL DO ARAGUAIA	NOROESTE GOIANO
2	RIO VERMELHO	
3	ARAGARÇAS	
4	PORANGATU	NORTE GOIANO
5	CHAPADA DOS VEADEIROS	
6	CERES	CENTRO GOIANO
7	ANÁPOLIS	
8	IPORÁ	
9	ANICUNS	
10	GOIANIA	
11	VÃO DO PARANÃ	LESTE GOIANO
12	ENTORNO DE BRASILIA	
13	SUDOESTE DE GOIÁS	SUL GOIANO
14	VALE DO RIO DOS BOIS	
15	MEIA PONTE	
16	PIRES DO RIO	
17	CATALÃO	
18	QUIRINÓPOLIS	

Tabela 8. Tabela de microregiões e mesoregiões definidas pelo IBGE.
FONTE: IBGE. Divisão Territorial Brasileira. Disponível em ftp://geoftp.ibge.gov.br/Organizacao/Divisao_Territorial/2009/DTB_05_05_2009.zip. Acesso em 25 de abril de 2010.

Assim, novos atributos foram construídos para armazenar essas informações para cada solicitação:

- Tr_Micro_Reg_Res: microregião de Goiás onde o solicitante reside ou se reside em outro Estado da Federação;
- Tr_Meso_Reg_Res: mesoregião de Goiás onde o solicitante reside ou se reside em outro Estado da Federação;
- Tr_Micro_Reg_Nat: microregião de Goiás da naturalidade do solicitante ou se tem origem em outro Estado da Federação;
- Tr_Meso_Reg_Nat: mesoregião de Goiás da naturalidade do solicitante ou se tem origem em outro Estado da Federação.

Ainda, por conta da esperada concentração de solicitações de pacientes de regiões próximas a capital e com o objetivo de enriquecer a análise de dados, as informações sobre os municípios de nascimento (naturalidade) e de residência do paciente foram tratadas no sentido de adicionar uma informação complementar.

Quando se tratava de um município do Estado de Goiás foi identificada a distância, em quilômetros (km), desse município em relação à capital do Estado. Assim, foram criados dois atributos para identificar essa relação:

- Tr_Dist_Munic_Res: distância aproximada, em km, entre o município de residência do paciente em Goiás e a capital

(Goiânia). Os demais municípios foram tratados apenas como localizados em outros Estados, sem considerar qualquer distância como referência;

- Tr_Dist_Munic_Nat: distância aproximada, em km, entre o município de nascimento (naturalidade) do paciente em Goiás e a capital (Goiânia). Da mesma forma, as distâncias dos demais municípios localizados em outros estados não foram consideradas.

Agrupamento de dados

Para melhor aproveitamento dos dados e com o objetivo de aumentar as chances de geração de um maior número de associações pelo algoritmo Apriori, itens de alguns atributos foram agrupados de forma a reduzir a quantidade de itens presentes em um mesmo atributo. Isso, porque se observou que determinados atributos possuíam muitos itens e, consequentemente, grande parte com uma frequência muito baixa. Nesse sentido, essa grande parte de itens seria desconsiderada na análise do algoritmo Apriori.

Vale destacar, que após as modificações realizadas e, consequentemente, com os novos atributos criados, já houve um agrupamento natural. A quantidade de itens de alguns atributos sofreu uma redução, principalmente os do tipo data, já que as inúmeras datas foram reduzidas, por exemplo, para poucos anos. Ainda assim, os atributos sobre a época da solicitação, a idade e o período de tempo do paciente no programa foram analisados para se verificar a possibilidade de um agrupamento mais efetivo, como em intervalos ou grupos.

A distribuição dos dados de cada atributo foi analisada no sentido de verificar a possibilidade de serem agrupados de forma que cada grupo formado apresentasse uma frequência maior que o suporte mínimo definido para mineração de dados. Houve a tentativa de processar os dados de tal forma que o máximo possível de itens (ou de grupos formados) distintos de cada novo atributo que fosse criado tivesse uma frequência acima do suporte mínimo avaliado no processo.

Em síntese, a manipulação se deu da seguinte forma:

- para o atributo avaliado, um novo atributo foi criado para receber a classificação de seus dados nos respectivos grupos;

- após analisar o item do atributo, em cada registro do conjunto de dados foi registrada sua respectiva classificação no novo atributo criado.

Cada atributo selecionado e tratado foi analisado para se avaliar a possibilidade da realização da tarefa de clusterização, com o uso do algoritmo K-Means no seu formato original, disponível na ferramenta WEKA. O objetivo foi o de encontrar um agrupamento mais natural possível para os dados desses atributos. A utilização deste algoritmo requer que os dados sejam numéricos e apresentem alguma relação de "distância".

Para a aplicação do algoritmo K-Means, os dados de cada atributo foram exportados para um arquivo diferente, do tipo *.csv. Cada arquivo foi aberto na ferramenta WEKA e o algoritmo foi aplicado várias vezes com diferentes definições de quantidades de grupos (o *K)* para a realização dos agrupamentos. Os melhores resultados foram encontrados com a definição de até 6 (seis) grupos. Para garantir uma busca profunda foi definido que o número de iterações poderia chegar até o máximo de 500 (quinhentas). Assim, foram identificados os centróides de cada grupo encontrado. Consequentemente, os itens foram agrupados conforme a sua maior proximidade do respectivo centróide.

Os procedimentos realizados nos agrupamentos estão relatados a seguir.

A identificação de alguma classificação, grupo ou tipo de medicamento seria uma informação interessante para a análise de possíveis associações entre as variáveis de solicitação de medicamentos, já que poucos itens do atributo "sme_medicamento-codg_medicamento" têm uma frequência maior que o suporte mínimo considerado essas informações poderiam não ser aproveitadas pelo método Apriori.

A tentativa de um agrupamento dos itens do atributo "sme_medicamento-codg_medicamento" com a utilização do algoritmo K-Means não alcançaria o objetivo de se obter grupos de medicamentos, por se tratar de um atributo com dados que não estão formatados conforme os requisitos de análise do algoritmo. Os dados dos atributos são códigos que apenas identificam os medicamentos e, por isso, não há relação de distância entre os itens, o que impossibilita o método K-Means encontrar os centróides através de cálculos de médias.

Buscou-se, então, realizar um tratamento em alguns outros dados da base de dados relacionados à informação sobre os medicamentos solicitados de

forma a tornar possível a aplicação do algoritmo K-Means para se encontrar características semelhantes entre os medicamentos. Uma alternativa encontrada foi a de identificar o custo aproximado do medicamento em cada solicitação com o objetivo de se formar grupos de medicamentos por intervalos de valores de custo.

Foi verificado que cada medicamento cadastrado na base de dados possui informações sobre o custo atual do medicamento. Entre as informações do medicamento há dados sobre o seu efetivo custo unitário, o valor máximo de custo definido pelo Ministério da Saúde e um valor estimado para uma futura aquisição.

Ao avaliar o preenchimento dessas informações, verificou-se que esses valores são muito próximos. Assim, para melhorar o preenchimento dos dados, foi feita a fusão das informações sobre os custos dos medicamentos, com a prioridade para os dados do custo efetivo, em seguida o valor definido pelo Ministério da Saúde e, caso ambos estejam ausentes, foi considerado o custo estimado do medicamento.

Infelizmente, a base de dados não apresenta informações de valores na época da solicitação, mas apenas os atuais. Ainda assim, é uma referência interessante para se ter grupos de medicamentos com valores mais altos ou mais baixos.

Assim, após a identificação dos valores atuais dos medicamentos em cada solicitação, foi aplicado o algoritmo K-Means, com o uso da ferramenta WEKA, sobre esses dados. Foi possível encontrar 4 grupos de medicamentos, conforme o seu custo unitário atual (em R$). Foi criado o atributo Tr_Grupo_Medic que armazenou os grupos identificados como: “1-3”; “11-50”; “4-10”; e “>=51”.

Da mesma forma, o algoritmo K-Means foi aplicado no atributo “sme_medicamento-qtde_pedida”. Foi possível separar as solicitações em 4 grandes grupos, conforme a quantidade do medicamento solicitada: “01-15”; “16-80”; “81-100”; e “>=101”. O novo atributo criado para a identificação desses grupos foi Tr_Grupo_qtde_pedida.

Após o tratamento dos dados referentes ao período aproximado da solicitação, foi possível agrupar as solicitações por trimestre e por ano. Foi criado o atributo Tr_Grupo_Ano_Soli para armazenar os 7 grupos encontrados após a aplicação do algoritmo K-Means para separar as solicitações por ano: “<=2001”; “2002-2003”; “2004-2005”; “2006”; “2007”; “2008”; e “>=2009”. Evidentemente, a identificação do trimestre aproximado das solicitações, anteriormente, praticamente

já formou os grupos de solicitações realizadas nos 4 trimestres, os quais foram registrados no atributo criado Tr_Trim_Solicitacao.

O algoritmo K-Means também foi utilizado para agrupar os itens dos atributos referentes às idades e tempo aproximado dos pacientes no programa. Vale destacar que esses atributos foram tratados na etapa anterior com o processamento de dados de atributos com informações sobre datas de nascimento e entrada no programa juntamente com o período aproximado da solicitação. Os novos atributos criados são:

- Tr_Grupo_Idade_Entrada_Prog: que identifica os grupos de pacientes que ingressaram no programa através de suas idades aproximadas. Foram identificados 6 grupos: “00-19 anos”; “20-33 anos”; “34-44 anos”; “45-56 anos”; “57-69 anos”; e “>=70 anos;
- Tr_Grupo_Idade_Disp: que identifica os grupos de pacientes através de suas idades aproximadas na época da solicitação de medicamentos. Foram identificados 6 grupos: “00-25 anos”; “26-40 anos”; “41-50 anos”; “51-60 anos”; “61-70 anos”; “>=71 anos”;
- Tr_Grupo_Tempo_Prog_Disp: que identifica os grupos de pacientes através do tempo aproximado neste programa assistencial na época da solicitação de medicamentos. Foram identificados 7 grupos: “0 anos”; “1 anos”; “2 anos”; “3 anos”; “04-05 anos”; “06-10 anos”; “>=11 anos”.

No mesmo sentido, embora com taxas de preenchimento muito baixas, os itens dos atributos com informações sobre peso e altura do paciente também foram agrupados com o uso do algoritmo K-Means. Foram criados os novos atributos:

- Tr_Grupo_Peso: com 3 grupos (“01-60”; “61-90”; e “>=91”), em kg;
- Tr_Grupo_Altura: com 4 grupos (“01-140”; “141-160”; “161-190”; e “>=191”), em cm.

Quanto ao agrupamento dos itens dos atributos referentes ao município de naturalidade e de residência do paciente solicitante, avaliou-se, inicialmente, as microregiões e mesoregiões oficialmente definidas nos trabalhos realizados pelo IBGE, no entanto, como há uma concentração muito forte dos dados de ambos os atributos nas regiões próximas a capital, apenas algumas poucas microregiões e mesoregiões teriam grandes frequências em detrimento das demais. Por outro lado,

ao aplicar o algoritmo K-Means nos dados sobre as distâncias (em km) dos municípios em relação a capital de Goiás, foi possível encontrar um agrupamento mais interessante. Os municípios foram agrupados e as informações armazenadas nos seguintes atributos:

- Tr_Grupo_Munic_Res: que identifica os grupos de municípios de residência dos pacientes em Goiás com distâncias muito próximas em relação a capital. Foram encontrados 3 grupos (“001-050 km”; “051-170 km”; e “>=171 km”) com o K-Means e identificados os pacientes residentes na própria capital (“CAP”) e os de fora do Estado de Goiás (“FE”);
- Tr_Grupo_Munic_Nat: que identifica os grupos de municípios de naturalidade dos pacientes em Goiás com distâncias muito próximas em relação a capital. Também foram encontrados 3 grupos (“001-100 km”; “101-200 km”; e “>=201 km”) com o K-Means e identificados os pacientes com naturalidade na própria capital (“CAP”) e os de fora do Estado de Goiás (“FE”);

Quanto ao código da CID, vale ressaltar que se trata de uma identificação de uma doença, conforme a Classificação Internacional de Doenças e Problemas Relacionados à Saúde (CID), utilizada pela Organização Mundial de Saúde (OMS). Ao consultar a versão 10 dessa classificação, no sítio do Departamento de Informática do SUS (DATASUS) na Internet, observou-se o agrupamento das doenças em 22 categorias, conforme tabela 9, da página 134:

Item	Categorias de Doenças	Códigos da CID
1	Doenças infecciosas e parasitárias	A00 até B99
2	Neoplasias - (Tumores)	C00 até D48
3	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários	D50 até D89
4	Doenças endócrinas, nutricionais e metabólicas	E00 até E90
5	Transtornos mentais e comportamentais	F00 até F99
6	Doenças do sistema nervoso	G00 até G99
7	Doenças do olho e anexos	H00 até H59
8	Doenças do ouvido e da apófise mastóide	H60 até H95
9	Doenças do aparelho circulatório	I00 até I99
10	Doenças do aparelho respiratório	J00 até J99
11	Doenças do aparelho digestivo	K00 até K93
12	Doenças da pele e do tecido subcutâneo	L00 até L99
13	Doenças do sistema osteomuscular e do tecido conjuntivo	M00 até M99
14	Doenças do aparelho geniturinário	N00 até N99
15	Gravidez, parto e puerpério	O00 até O99
16	Algumas afecções originadas no período perinatal	P00 até P96
17	Malformações congênitas, deformidades e anomalias cromossômicas	Q00 até Q99
18	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	R00 até R99
19	Lesões, envenenamento e algumas outras consequências de causas externas	S00 até T98
20	Causas externas de morbidade e de mortalidade	V01 até Y98
21	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	Z00 até Z99
22	Códigos para propósitos especiais	U00 até U99

Tabela 9. Lista de categorias de três caracteres da CID vigente.
Fonte: CID-10 – DATASUL, disponível em http://www.datasus.gov.br/cid10/v2008/cid10.htm, acesso em 05 de março de 2010.

Assim, uma alternativa encontrada para reduzir o número de itens do atributo, de forma que apresentem frequências maiores, foi a de utilizar esse agrupamento. A categoria da doença informada em cada solicitação foi identificada e armazenada no atributo Tr_Grupo_CID.

Da mesma forma, a classificação dos registros de solicitação pelo IMC dos respectivos pacientes, já realizada e armazenada no atributo Tr_Classe_IMC anteriormente, se baseou nos critérios adotados também pela OMS, conforme já demonstrado na tabela 7, o que já possibilitou um agrupamento dos registros.

Por apresentarem apenas 2 (dois) itens, com ambos acima do suporte considerado, não houve a necessidade de realização de agrupamentos dos atributos “sme-info_turno” e “paciente-sexo”.

Por opção, não foi adotado nenhum critério para um agrupamento específico dos atributos “cartao-guiche”, “paciente-codg_raca_cor” e “tab_esto-codg_unidade”, por já se tratarem de variáveis que já definem categorias pré-determinadas e informações diretas e peculiares. Apenas para aumentar o

desempenho de processamento, os itens dos atributos com menores frequências foram agrupados com identificação de “outros casos”. Nesse sentido, os novos atributos criados são:

- Tr_Grupo_Guiche: com 4 itens possíveis (Guichê 1, 2, 3 ou “O” - outros);
- Tr_Grupo_Raca_Cor: com 3 itens possíveis (1-Branca, 3-Parda ou “O” - outras);
- Tr_Grupo_Un_Med: com 4 itens possíveis (“COMP” - comprimidos, “F/A” – fracos ou ampolas, “CAPS” – cápsulas e “O” - outras).

O atributo “processo-codg_unidade” se trata de dados de códigos identificadores de unidades hospitalares e médicos que atestam a necessidade do uso do medicamento. Da mesma forma, o atributo tem grande importância na análise das solicitações de medicamentos, mas apresenta mais de 6.000 médicos ou hospitais diferentes entre os registros de solicitações. Os dados adicionais de identificação dessas unidades e médicos na base de dados são apenas nome, CNS ou CRM e CNPJ ou CPF. A tentativa de se atribuir alguma outra característica a cada unidade para se tornar uma referência para um agrupamento teria de ser feita manualmente e, pelo volume de unidades diferentes e poucos dados de identificação, o trabalho seria difícil, demorado e com muitas chances de erro. Por exemplo, pode-se imaginar a dificuldade de identificar a especialidade de um médico ou hospital (que até pode realizar atendimentos em diversas especialidades) sem que haja um ponto de partida. Por esses motivos, o atributo não foi objeto de procedimentos para realização de um agrupamento.

7.3.4- 4ª Fase

2ª Seleção de dados

Como vários novos atributos foram criados, para retratar a mesma informação de outros de maneira mais bem trabalhada, foi feita uma análise para verificar possíveis redundâncias de informações entre os atributos. Assim, quando

diagnosticada a redundância, optou-se por manter os atributos que tiveram origem durante as transformações.

Assim, foram selecionados nesta fase os dados dos seguintes atributos, conforme demonstrado na tabela 10, a seguir:

Atributos Originais	Atributos Trabalhados	Atributos Selecionados
sme_medicamento-codg_medicamento	Tr_Custo_Medic; Tr_Grupo_Medic	Tr_Grupo_Medic
sme_medicamento-qtde_pedida	Tr_Grupo_qtde_pedida	Tr_Grupo_qtde_pedida
sme-info_turno		sme-info_turno
processo-codg_unidade		processo-codg_unidade
processo-info_cid1	Tr_CID; Tr_Grupo_CID	Tr_Grupo_CID
processo-data_sme	Tr_ano_solicitacao; Tr_trim_solicitacao; Tr_Grupo_Ano_Soli;	Tr_Grupo_Ano_Soli Tr_Trim_Solicitacao
cartao-guiche	Tr_Grupo_Guiche;	Tr_Grupo_Guiche
cartao-codg_munic	Tr_Munic_Endereco; Tr_Micro_Reg_Res; Tr_Meso_Reg_Res; Tr_Dist_Munic_Res; Tr_Grupo_Munic_Res	Tr_Grupo_Munic_Res
cartao-peso	Tr_Grupo_Peso; IMC-Paciente; Tr_Classe_IMC	Tr_Grupo_Peso; Tr_Classe_IMC;
cartao-altura	Tr_Grupo_Altura;	Tr_Grupo_Altura
paciente-sexo		paciente-sexo
paciente-codg_raca_cor	Tr_Grupo_Raca_Cor;	Tr_Grupo_Raca_Cor
paciente-cod_naturalidade	Tr_Micro_Reg_Nat; Tr_Meso_Reg_Nat; Tr_Dist_Munic_Nat; Tr_Grupo_Munic_Nat.	Tr_Grupo_Munic_Nat
paciente-data_nascimento	Tr_Idade_Entr_Prog; Tr_Idade_Disp; Tr_Grupo_Idade_Entrada_Prog; Tr_Grupo_Idade_Disp;	Tr_Grupo_Idade_Entrada_Prog; Tr_Grupo_Idade_Disp.
paciente-data_entr_ss	Tr_Tempo_Prog_Disp; Tr_Grupo_Tempo_Prog_Disp;	Tr_Grupo_Tempo_Prog_Disp
tab_esto-codg_unidade	Tr_Grupo_Un_Med.	Tr_Grupo_Un_Med

Tabela 10. Demonstração dos atributos selecionados nesta 2ª seleção prevista na metodologia.

2ª limpeza dos dados

Os dados foram novamente analisados, principalmente os dos atributos resultantes das manipulações realizadas na fase anterior, com o objetivo de verificar se há, ainda, alguma inconsistência.

As principais adequações foram realizadas nos atributos resultantes de fusões e tratamento de dados, como: custo do medicamento, código da CID e ano

de solicitação. Buscou-se identificar e retirar possíveis incoerências surgidas com a fusão ou tratamento dos dados.

Da mesma forma, foram removidos vários registros pelo fato de que boa parte de seus dados estavam ausentes ou retirados anteriormente por conta das inconsistências.

Remoção de registros duplicados

Uma *query* foi realizada para identificar e remover os registros duplicados, oriundos, provavelmente, das mudanças realizadas na base de dados em razão das substituições do sistema de informações da unidade de saúde.

2ª tabulação dos dados

Foi realizada uma nova contagem de itens, com o uso de *queries*, para cada atributo remanescente. O objetivo foi o de visualizar as frequências dos seus itens e avaliar melhor os resultados das transformações e manipulações realizadas até o momento. As frequências foram representadas em gráficos de barras para facilitar a análise. As principais informações podem ser verificadas através das tabelas 11 e 12, das páginas 138-140, e figura 24, das páginas 141-143.

7.3.5- 5ª Fase – Seleção final de dados

Após a visualização dos dados, percebeu-se que apenas dois atributos, Tr_Grupo_CID e processo-codg_unidade, apresentavam a maioria de seus itens com baixas frequências, o que significa a influência desses itens no processo de busca por associações seria mínima.

Quanto ao atributo processo-codg_unidade, vale destacar que a quantidade de médicos e hospitais que prescreveram os medicamentos excepcionais é muito grande, por isso as solicitações estão bastante distribuídas. Devido à dificuldade e a complexidade de uma tentativa de clusterização desses dados e ao fato de que esses dados não oferecem opções para definição de critérios de agrupamento, decidiu-se não realizar um tratamento do atributo. Nesse sentido, o

atributo foi retirado do processo, já que pouco contribuiria com possíveis associações.

Já com relação ao atributo Tr_Grupo_CID, há vários itens com frequências maiores que o suporte mínimo definido neste experimento; todavia a maioria dos itens apresenta uma baixa frequência. Dessa forma, foi decidido manter o atributo no processo para verificar se os itens com maiores frequências poderiam estar associados a algum dado de outro atributo.

Visualização dos dados

Do processo de pré-processamento de dados desse experimento, em que os critérios e procedimentos foram aplicados conforme a metodologia proposta, resultaram-se 746.445 registros com informações organizadas nos 18 atributos selecionados. Entre esses atributos foram verificadas as seguintes situações:

- apenas 5 atributos possuíam uma taxa de ausência de dados acima de 30% e 9 atributos apresentam mais de 90% dos seus dados preenchidos;
- 11 atributos tinham todos os itens com freqüência superior ao suporte considerado no experimento e outros 4 possuíam apenas um sem alcançar o nível desse suporte.

Por fim, os atributos selecionados para terem seus dados submetidos ao processamento do algoritmo Apriori, na etapa seguinte, foram os listados juntamente com as relativas informações nas tabelas 11 e 12, das páginas 138-140, e figura 24, das páginas 141-143:

Identificação	Descrição do atributo	Dados do atributo
Tr_Grupo_Medic	Enquadramento da solicitação em um dos grupos definidos pelo custo unitário atual do medicamento solicitado (em R$).	Grupos: • 1-3 • 4-10 • 11-50 • >=51
Tr_Grupo_qtde_pedida	Enquadramento da solicitação em um dos grupos definidos por intervalos de quantidades dos medicamentos solicitados.	Grupos: • 01-15 • 16-80 • 81-100 • >=101
sme-info_turno	Turno em que a solicitação foi registrada no sistema.	M – Manhã; V – Vespertino.
Tr_Grupo_CID	Enquadramento da solicitação em um dos grupos definidos pelo enquadramento do código da CID em categorias de doenças.	Categorias identificadas pelos códigos de 1 a 22 e NI (Não identificada).

Tr_Grupo_Guiche	Enquadramento da solicitação conforme a identificação dos guichês que realizaram os atendimentos.	1- Guichê 1 2- Guichê 2 3- Guichê 3 O- Outros guichês
Tr_Grupo_Peso	Enquadramento da solicitação em um dos grupos definidos por intervalos de pesos (em kg) dos pacientes solicitantes.	Grupos: • 01-60 • 61-90 • >=91
Tr_Classe_IMC	Enquadramento da solicitação em uma das classes adotadas pela OMS, conforme o IMC do respectivo paciente solicitante.	1 – Magreza; 2 – Saudável; 3 – Sobrepeso; 4 – Obesidade grau I 5 – Obesidade grau II (severa) 6 – Obesidade grau III (mórbida)
Tr_Grupo_Altura	Enquadramento da solicitação em um dos grupos definidos por intervalos de alturas (em cm) dos pacientes solicitantes.	Grupos: • 01-140 • 141-160 • 161-190 • >=191
paciente-sexo	Sexo do paciente.	1- Masculino; 2- Feminino
Tr_Grupo_Raca_Cor	Enquadramento da solicitação conforme a raça/cor dos pacientes solicitantes.	1- Branca; 3- Parda; O- Outras
Tr_Grupo_Un_Med	Enquadramento da solicitação conforme a unidade de medida utilizada para atender à posologia do medicamento solicitada.	CAPS - Cápsulas COMP - Comprimidos F/A – Frasco ou Ampola O - Outras;
Tr_Grupo_Ano_Soli	Enquadramento da solicitação em um dos grupos definidos por intervalos dos anos em que as solicitações foram realizadas.	Grupos: • <=2001 • 2002-2003 • 2004-2005 • 2006 • 2007 • 2008 • >=2009
Tr_Trim_Solicitacao	Enquadramento da solicitação conforme o trimestre em que foi realizada.	1- 1º trimestre 2- 2º trimestre 3- 3º trimestre 4- 4º trimestre
Tr_Grupo_Munic_Res	Enquadramento da solicitação em um dos grupos definidos conforme intervalos das distâncias do município de endereço do paciente em relação a capital.	Grupos: • Cap - Capital; • 001-050 km; • 051-170 km; • >=171 km; • FE - Fora do Estado.
Tr_Grupo_Munic_Nat	Enquadramento da solicitação em um dos grupos definidos conforme intervalos das distâncias do município de naturalidade do paciente em relação a capital.	Grupos: • Cap - Capital; • 001-100 km; • 101-200 km; • >=201 km; • FE - Fora do Estado.
Tr_Grupo_Idade_Entrada_Prog	Enquadramento da solicitação em um dos grupos definidos por intervalos das idades aproximadas em que os pacientes ingressaram no programa de dispensação de medicamentos.	Grupos: • 00-19 anos; • 20-33 anos; • 34-44 anos; • 45-56 anos; • 57-69 anos; • >=70 anos.

Tr_Grupo_Idade_Disp	Enquadramento da solicitação em um dos grupos definidos por intervalos das idades aproximadas dos pacientes no momento da solicitação de medicamentos.	Grupos: • 00-25 anos; • 26-40 anos; • 41-50 anos; • 51-60 anos; • 61-70 anos; • >=71 anos;
Tr_Grupo_Tempo_Prog_Disp	Enquadramento da solicitação em um dos grupos definidos por intervalos dos tempos proximados em que os pacientes participam do programa de dispensação de medicamentos.	Grupos: • 0 anos; • 1 anos; • 2 anos; • 3 anos; • 04-05 anos; • 06-10 anos; • >=11 anos;

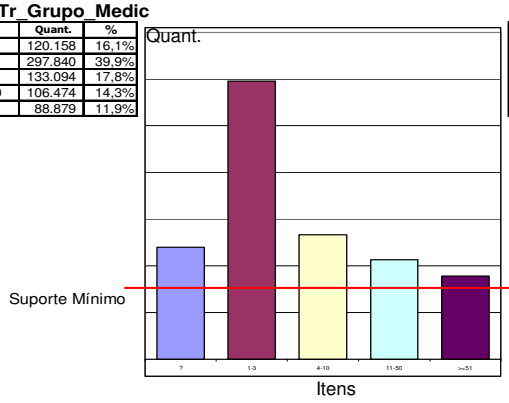
Tabela 11. Descrição dos atributos selecionados para o processo de mineração de dados.

Identificação	Taxa de ausência de dados	Número de itens distintos	Quant. de itens com frequência acima de 10%	Quant. de itens com frequência acima de 5%	Quant. de itens com frequência acima de 1%
Tr_Grupo_Medic	16,1%	4	4	4	4
Tr_Grupo_qtde_pedida	3,3%	4	4	4	4
sme-info_turno	44,3%	2	2	2	2
Tr_Grupo_CID	1,4%	19	3	6	11
Tr_Grupo_Guiche	24,2%	4	4	4	4
Tr_Grupo_Peso	68,6%	3	2	2	3
Tr_Classe_IMC	69,5%	6	1	2	4
Tr_Grupo_Altura	69,4%	4	2	2	3
paciente-sexo	10,7%	2	2	2	2
Tr_Grupo_Raca_Cor	43,0%	3	2	2	3
Tr_Grupo_Un_Med	1,8%	4	4	4	4
Tr_Grupo_Ano_Soli	0,0%	7	7	7	7
Tr_Trim_Solicitacao	0,0%	4	4	4	4
Tr_Grupo_Munic_Res	7,9%	5	4	4	4
Tr_Grupo_Munic_Nat	22,3%	5	5	5	5
Tr_Grupo_Idade_Entrada_Prog	1,1%	6	6	6	6
Tr_Grupo_Idade_Disp	0,1%	6	6	6	6
Tr_Grupo_Tempo_Prog_Disp	2,6%	7	6	6	6

Tabela 12. Taxa de ausência de dados e quantidade de itens mais frequentes nos atributos selecionados para processamento do algoritmo Apriori.

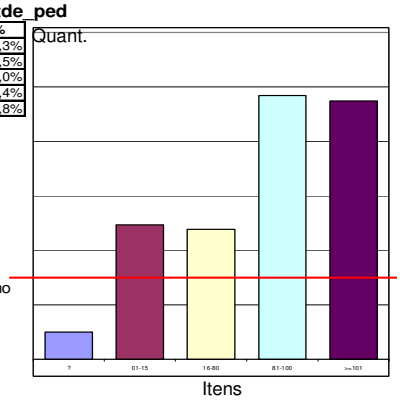
Atributo: Tr_Grupo_Medic

	Item	Quant.	%
	?	120.158	16,1%
1	1-3	297.840	39,9%
2	4-10	133.094	17,8%
3	11-50	106.474	14,3%
4	>=51	88.879	11,9%



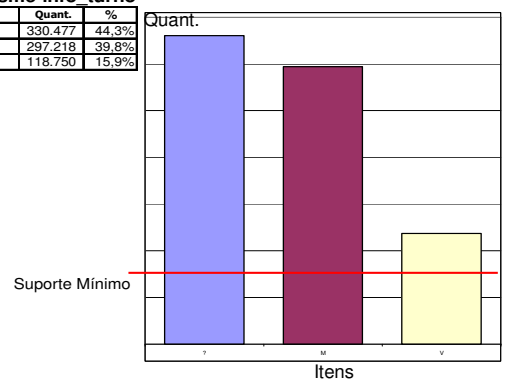
Atributo: Tr_Grupo_qtde_ped

	Item	Quant.	%
	?	24.965	3,3%
1	01-15	123.354	16,5%
2	16-80	119.227	16,0%
3	81-100	241.887	32,4%
4	>=101	237.012	31,8%



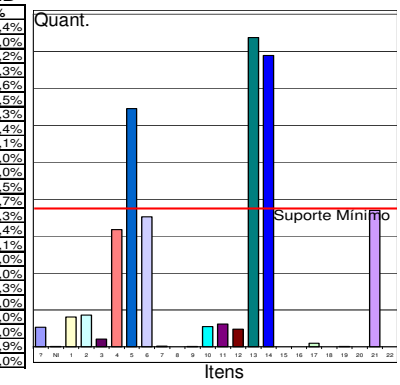
Atributo: sme-info_turno

	Item	Quant.	%
	?	330.477	44,3%
1	M	297.218	39,8%
2	V	118.750	15,9%



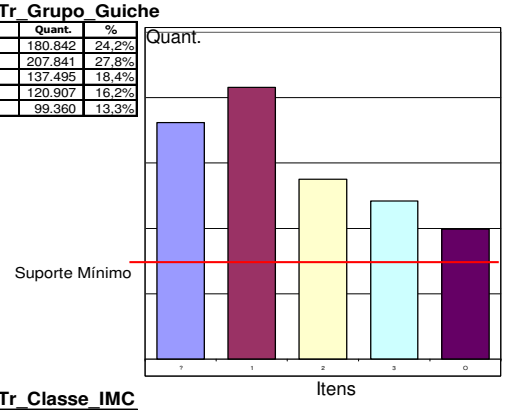
Atributo: Tr_Grupo_CID

	Item	Quant.	%
	?	10.569	1,4%
1	NI	197	0,0%
2	1	16.167	2,2%
3	2	17.251	2,3%
4	3	4.238	0,6%
5	4	63.487	8,5%
6	5	128.994	17,3%
7	6	70.494	9,4%
8	7	406	0,1%
9	8	-	0,0%
10	9	206	0,0%
11	10	11.051	1,5%
12	11	12.445	1,7%
13	12	9.455	1,3%
14	13	167.429	22,4%
15	14	157.707	21,1%
16	15	1	0,0%
17	16	-	0,0%
18	17	1.987	0,3%
19	18	48	0,0%
20	19	352	0,0%
21	20	-	0,0%
22	21	73.961	9,9%
23	22	-	0,0%



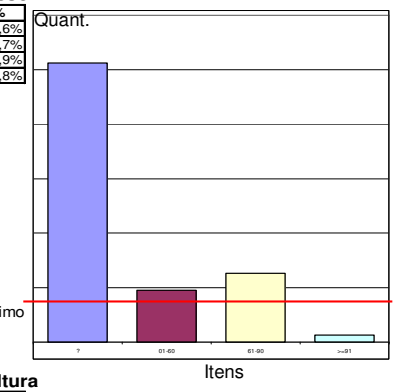
Atributo: Tr_Grupo_Guiche

	Item	Quant.	%
	?	180.842	24,2%
1	1	207.841	27,8%
2	2	137.495	18,4%
3	3	120.907	16,2%
4	O	99.360	13,3%



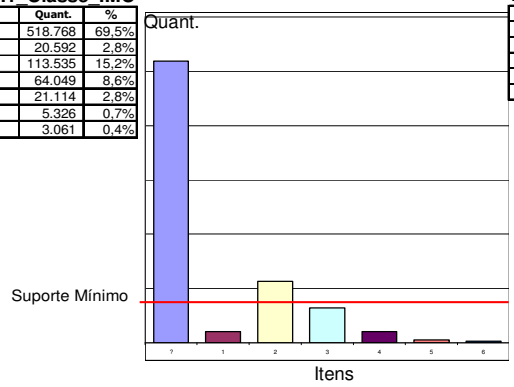
Atributo: Tr_Grupo_Peso

	Item	Quant.	%
	?	512.171	68,6%
1	01-60	95.032	12,7%
2	61-90	125.995	16,9%
3	>=91	13.247	1,8%



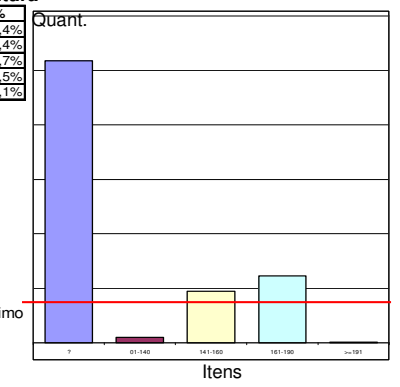
Atributo: Tr_Classe_IMC

	Item	Quant.	%
	?	518.768	69,5%
1	1	20.592	2,8%
2	2	113.535	15,2%
3	3	64.049	8,6%
4	4	21.114	2,8%
5	5	5.326	0,7%
6	6	3.061	0,4%



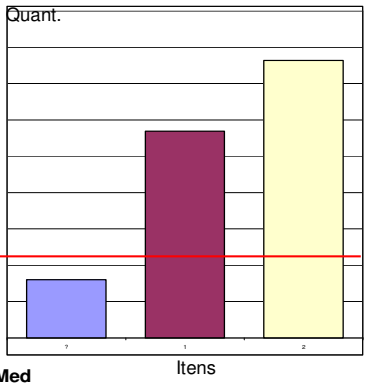
Atributo: Tr_Grupo_Altura

	Item	Quant.	%
	?	517.863	69,4%
1	01-140	10.134	1,4%
2	141-160	94.734	12,7%
3	161-190	122.791	16,5%
4	>=191	923	0,1%



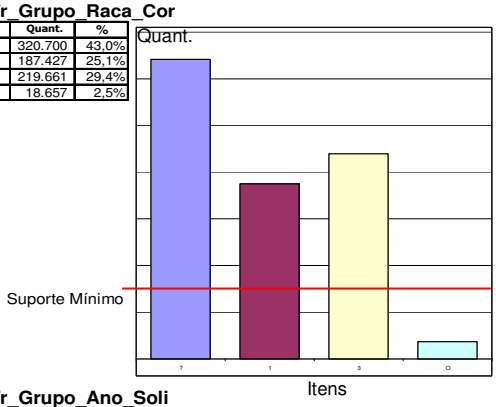
Atributo: paciente-sexo

	Item	Quant.	%
	?	80.118	10,7%
1	1	284.374	38,1%
2	2	381.953	51,2%



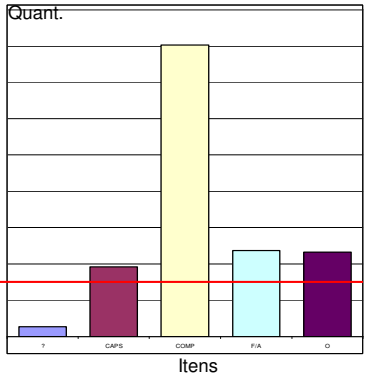
Atributo: Tr_Grupo_Raca_Cor

	Item	Quant.	%
	?	320.700	43,0%
1	1	187.427	25,1%
2	3	219.661	29,4%
3	O	18.657	2,5%



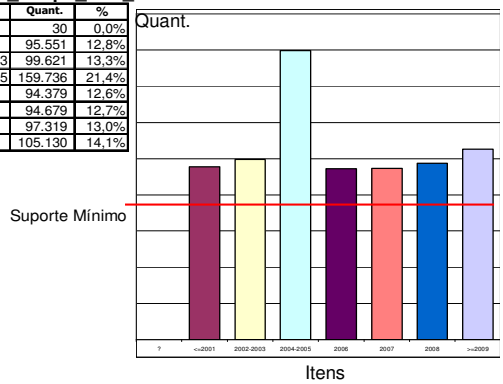
Atributo: Tr_Grupo_Un_Med

	Item	Quant.	%
	?	13.726	1,8%
1	CAPS	96.167	12,9%
2	COMP	401.574	53,8%
3	F/A	118.592	15,9%
4	O	116.386	15,6%



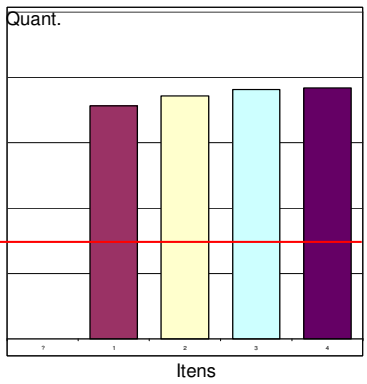
Atributo: Tr_Grupo_Ano_Soli

	Item	Quant.	%
	?	30	0,0%
1	<=2001	95.551	12,8%
2	2002-2003	99.621	13,3%
3	2004-2005	159.736	21,4%
4	2006	94.379	12,6%
5	2007	94.679	12,7%
6	2008	97.319	13,0%
7	>=2009	105.130	14,1%



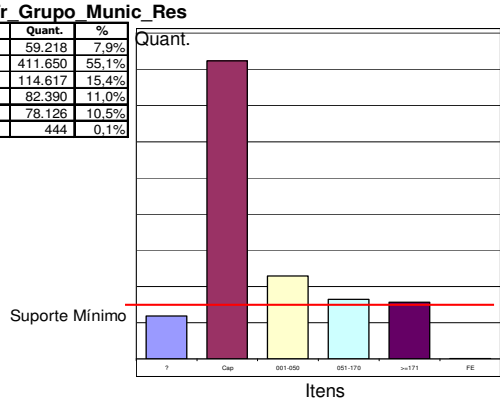
Atributo: Tr_Trim_Solicitacao

	Item	Quant.	%
	?	30	0,0%
1	1	178.100	23,9%
2	2	185.638	24,9%
3	3	190.844	25,6%
4	4	191.833	25,7%



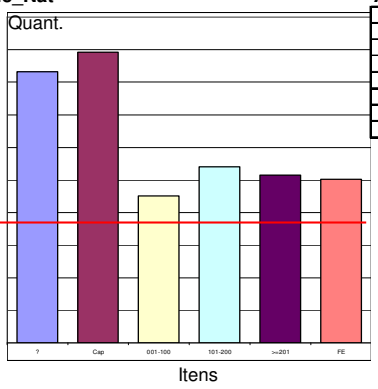
Atributo: Tr_Grupo_Munic_Res

	Item	Quant.	%
	?	59.218	7,9%
1	Cap	411.650	55,1%
2	001-050	114.617	15,4%
3	051-170	82.390	11,0%
4	>=171	78.126	10,5%
5	FE	444	0,1%



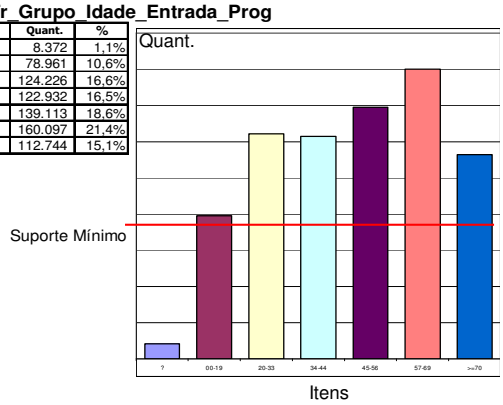
Atributo: Tr_Grupo_Munic_Nat

	Item	Quant.	%
	?	166.530	22,3%
1	Cap	178.379	23,9%
2	001-100	90.251	12,1%
3	101-200	107.948	14,5%
4	>=201	102.912	13,8%
5	FE	100.425	13,5%



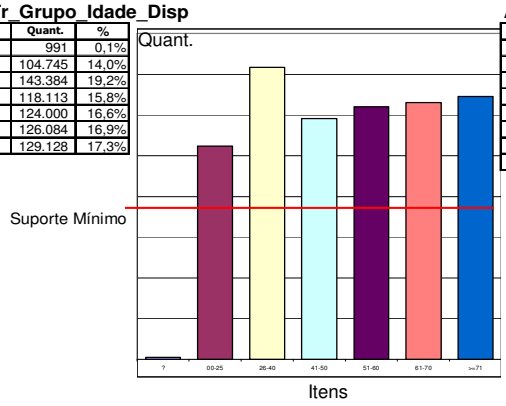
Atributo: Tr_Grupo_Idade_Entrada_Prog

	Item	Quant.	%
	?	8.372	1,1%
1	00-19	78.961	10,6%
2	20-33	124.226	16,6%
3	34-44	122.932	16,5%
4	45-56	139.113	18,6%
5	57-69	160.097	21,4%
6	>=70	112.744	15,1%



Atributo: Tr_Grupo_Idade_Disp

	Item	Quant.	%
	?	991	0,1%
1	00-25	104.745	14,0%
2	26-40	143.384	19,2%
3	41-50	118.113	15,8%
4	51-60	124.000	16,6%
5	61-70	126.084	16,9%
6	>=71	129.128	17,3%



Atributo: Tr_Grupo_Tempo_Prog_Disp

	Item	Quant.	%
	?	19.619	2,6%
1	0	157.088	21,0%
2	1	158.600	21,2%
3	2	111.352	14,9%
4	3	84.865	11,4%
5	04-05	113.902	15,3%
6	06-10	95.999	12,9%
7	>=11	5.020	0,7%

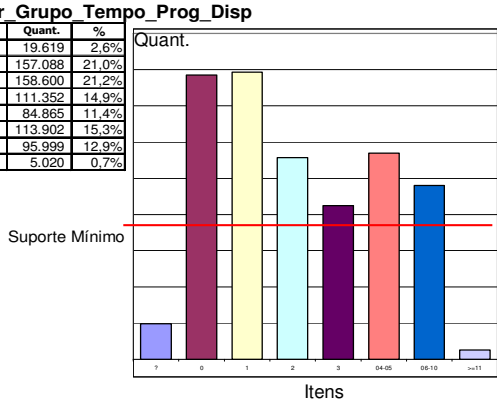


Figura 24. Gráficos com as frequências dos itens de cada atributo selecionado para a aplicação do algoritmo Apriori.

7.3.6- 6ª Fase – Mineração de dados

Nesta etapa foi aplicado o algoritmo Apriori, com o uso do software WEKA, sobre os dados pré-processados. Com o objetivo de aumentar o desempenho de processamento, antes da aplicação do algoritmo o conjunto de dados selecionados foi exportado para um arquivo texto do tipo *.csv. A ferramenta foi configurada para que o algoritmo tentasse encontrar até 200 regras de associação, com base nos suportes e o grau de confiança mínimos determinados.

O algoritmo foi aplicado com a definição de um grau de suporte mínimo de 10% e de um grau de confiança mínimo de 80% (oitenta por cento), conforme demonstrado na figura 25 a seguir.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 200 -T 0 -C 0.8 -D 0.05 -U 1.0 -
M 0.1 -S -1.0 -c -1
Relation:     Conjunto-Solicitacoes-Medicamentos-Transformados
Instances:    746445
Attributes:   18
              Tr_Grupo_Medic
              Tr_Grupo_qtde_pedida
              sme-info_turno
              Tr_Grupo_CID
              Tr_Grupo_Guiche
              Tr_Grupo_Peso
              Tr_Classe_IMC
              Tr_Grupo_Altura
              paciente-sexo
              Tr_Grupo_Raca_Cor
              Tr_Grupo_Un_Med
              Tr_Grupo_Ano_Soli
              Tr_Trim_Solicitacao
              Tr_Grupo_Munic_Res
```

```
               Tr_Grupo_Munic_Nat
               Tr_Grupo_Idade_Entrada_Prog
               Tr_Grupo_Idade_Disp
               Tr_Grupo_Tempo_Prog_Disp
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (74645 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 68

Size of set of large itemsets L(2): 115

Size of set of large itemsets L(3): 39

Size of set of large itemsets L(4): 14

Size of set of large itemsets L(5): 2

Best rules found:

  1. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
Tr_Grupo_Guiche=3.0 80978 ==> Tr_Grupo_Un_Med=COMP 80874    conf:(1)
  2. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0
81797 ==> Tr_Grupo_Un_Med=COMP 81654    conf:(1)
  3. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0 85466 ==>
Tr_Grupo_Un_Med=COMP 85270    conf:(1)
  4. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0
81951 ==> Tr_Grupo_Un_Med=COMP 81649    conf:(1)
  5. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 83813 ==>
Tr_Grupo_Un_Med=COMP 83457    conf:(1)
  6. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap
79288 ==> Tr_Grupo_Un_Med=COMP 78755    conf:(0.99)
  7. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
paciente-sexo=2.0 87493 ==> Tr_Grupo_Un_Med=COMP 86898    conf:(0.99)
  8. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0 94664
==> Tr_Grupo_Un_Med=COMP 93959    conf:(0.99)
  9. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
105197 ==> Tr_Grupo_Un_Med=COMP 104394    conf:(0.99)
 10. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 87075 ==> Tr_Grupo_Un_Med=COMP
86383    conf:(0.99)
 11. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 113920 ==>
Tr_Grupo_Un_Med=COMP 112939    conf:(0.99)
 12. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0
Tr_Grupo_Un_Med=COMP 81649 ==> Tr_Grupo_CID=13.0 80874    conf:(0.99)
 13. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0
Tr_Grupo_Un_Med=COMP 81654 ==> Tr_Grupo_Medic=1-3 80874    conf:(0.99)
 14. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0
81797 ==> Tr_Grupo_Medic=1-3 80978    conf:(0.99)
 15. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap 77799 ==>
Tr_Grupo_Un_Med=COMP 76954    conf:(0.99)
 16. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0
81797 ==> Tr_Grupo_Medic=1-3 Tr_Grupo_Un_Med=COMP 80874    conf:(0.99)
```

```
 17. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0
81951 ==> Tr_Grupo_CID=13.0 80978    conf:(0.99)
 18. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0 91816 ==>
Tr_Grupo_Un_Med=COMP 90695    conf:(0.99)
 19. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 86383 ==>
Tr_Grupo_CID=13.0 85270    conf:(0.99)
 20. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0
81951 ==> Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 80874    conf:(0.99)
 21. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 110624 ==> Tr_Grupo_Un_Med=COMP
109084    conf:(0.99)
 22. Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 86521 ==>
Tr_Grupo_Medic=1-3 85270    conf:(0.99)
 23. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 87075 ==> Tr_Grupo_CID=13.0
85466    conf:(0.98)
 24. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 87075 ==> Tr_Grupo_CID=13.0
Tr_Grupo_Un_Med=COMP 85270    conf:(0.98)
 25. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP
83457 ==> Tr_Grupo_CID=13.0 81654    conf:(0.98)
 26. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP
83457 ==> Tr_Grupo_Medic=1-3 81649    conf:(0.98)
 27. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 83813 ==>
Tr_Grupo_Medic=1-3 81951    conf:(0.98)
 28. Tr_Grupo_Idade_Entrada_Prog=00-19 78961 ==> Tr_Grupo_Idade_Disp=00-25
77078    conf:(0.98)
 29. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 83813 ==>
Tr_Grupo_CID=13.0 81797    conf:(0.98)
 30. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 83813 ==>
Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 81654    conf:(0.97)
 31. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 83813 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_Un_Med=COMP 81649    conf:(0.97)
 32. Tr_Grupo_CID=5.0 128994 ==> Tr_Grupo_Un_Med=COMP 125452    conf:(0.97)
 33. Tr_Grupo_Idade_Entrada_Prog=>=70 112744 ==> Tr_Grupo_Idade_Disp=>=71
109516    conf:(0.97)
 34. Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 89264 ==> Tr_Grupo_CID=13.0
86521    conf:(0.97)
 35. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP
83457 ==> Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 80874    conf:(0.97)
 36. Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 89264 ==> Tr_Grupo_Medic=1-3
86383    conf:(0.97)
 37. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 83813 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 80978    conf:(0.97)
 38. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Guiche=3.0 83813 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 80874
conf:(0.96)
 39. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 90695 ==> Tr_Grupo_qtde_pedida=81-100 86898
conf:(0.96)
 40. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 109084 ==>
Tr_Grupo_qtde_pedida=81-100 104394    conf:(0.96)
 41. Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 89264 ==> Tr_Grupo_Medic=1-3
Tr_Grupo_CID=13.0 85270    conf:(0.96)
 42. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0 91816 ==>
Tr_Grupo_qtde_pedida=81-100 87493    conf:(0.95)
 43. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 110624 ==>
Tr_Grupo_qtde_pedida=81-100 105197    conf:(0.95)
 44. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0
Tr_Grupo_Un_Med=COMP 85270 ==> Tr_Grupo_qtde_pedida=81-100 80874
conf:(0.95)
 45. Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 Tr_Grupo_Munic_Res=Cap
104851 ==> Tr_Grupo_Un_Med=COMP 99427    conf:(0.95)
```

```
 46. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0 85466 ==>
Tr_Grupo_qtde_pedida=81-100 80978    conf:(0.95)
 47. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0 91816 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP 86898    conf:(0.95)
 48. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0 85466 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP 80874    conf:(0.95)
 49. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 86383 ==>
Tr_Grupo_qtde_pedida=81-100 81649    conf:(0.95)
 50. Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 86521 ==>
Tr_Grupo_qtde_pedida=81-100 81654    conf:(0.94)
 51. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 110624 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP 104394    conf:(0.94)
 52. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 87075 ==>
Tr_Grupo_qtde_pedida=81-100 81951    conf:(0.94)
 53. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 87075 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP 81649    conf:(0.94)
 54. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 86383 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 80874    conf:(0.94)
 55. Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 89264 ==>
Tr_Grupo_qtde_pedida=81-100 83457    conf:(0.93)
 56. Tr_Grupo_CID=13.0 Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 86521 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 80874    conf:(0.93)
 57. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0
102787 ==> Tr_Grupo_Un_Med=COMP 95777    conf:(0.93)
 58. Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 156999 ==>
Tr_Grupo_Un_Med=COMP 146142    conf:(0.93)
 59. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 87075 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 80978    conf:(0.93)
 60. Tr_Grupo_Medic=1-3 Tr_Grupo_Guiche=3.0 87075 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 80874
conf:(0.93)
 61. Tr_Grupo_Guiche=3.0 paciente-sexo=2.0 94447 ==> Tr_Grupo_CID=13.0
87403    conf:(0.93)
 62. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 93959 ==> Tr_Grupo_Medic=1-3 86898    conf:(0.92)
 63. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP
112939 ==> Tr_Grupo_Medic=1-3 104394    conf:(0.92)
 64. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0 94664
==> Tr_Grupo_Medic=1-3 87493    conf:(0.92)
 65. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 113920 ==>
Tr_Grupo_Medic=1-3 105197    conf:(0.92)
 66. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Munic_Res=Cap 153856 ==>
Tr_Grupo_Un_Med=COMP 141680    conf:(0.92)
 67. Tr_Grupo_Idade_Disp=61-70 126084 ==> Tr_Grupo_Idade_Entrada_Prog=57-69
115940    conf:(0.92)
 68. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0 94664
==> Tr_Grupo_Medic=1-3 Tr_Grupo_Un_Med=COMP 86898    conf:(0.92)
 69. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 113920 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_Un_Med=COMP 104394    conf:(0.92)
 70. Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 89264 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 81654    conf:(0.91)
 71. Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 89264 ==> Tr_Grupo_Medic=1-3
Tr_Grupo_qtde_pedida=81-100 81649    conf:(0.91)
 72. Tr_Grupo_Guiche=3.0 120907 ==> Tr_Grupo_CID=13.0 110301    conf:(0.91)
 73. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Munic_Res=Cap
93108 ==> Tr_Grupo_Un_Med=COMP 84828    conf:(0.91)
 74. Tr_Grupo_qtde_pedida=81-100 sme-info_turno=M 106019 ==>
Tr_Grupo_Un_Med=COMP 96536    conf:(0.91)
 75. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 95777 ==> Tr_Grupo_CID=13.0 86898    conf:(0.91)
```

```
 76. Tr_Grupo_Guiche=3.0 Tr_Grupo_Un_Med=COMP 89264 ==> Tr_Grupo_Medic=1-3
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 80874    conf:(0.91)
 77. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP Tr_Grupo_Munic_Res=Cap 87883
==> Tr_Grupo_qtde_pedida=81-100 78755    conf:(0.9)
 78. Tr_Grupo_qtde_pedida=81-100 241887 ==> Tr_Grupo_Un_Med=COMP 215992
conf:(0.89)
 79. Tr_Grupo_Medic=4-10 133094 ==> Tr_Grupo_Un_Med=COMP 118611
conf:(0.89)
 80. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP Tr_Grupo_Munic_Res=Cap 87883
==> Tr_Grupo_Medic=1-3 76954    conf:(0.88)
 81. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 139067 ==>
Tr_Grupo_Un_Med=COMP 121447    conf:(0.87)
 82. Tr_Grupo_CID=13.0 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 107825 ==>
Tr_Grupo_qtde_pedida=81-100 93959    conf:(0.87)
 83. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 130370 ==>
Tr_Grupo_qtde_pedida=81-100 112939    conf:(0.87)
 84. Tr_Grupo_Ano_Soli=2004-2005 Tr_Grupo_Munic_Res=Cap 98498 ==> sme-
info_turno=M 85277    conf:(0.87)
 85. Tr_Grupo_Ano_Soli=2004-2005 159736 ==> sme-info_turno=M 138026
conf:(0.86)
 86. Tr_Grupo_Un_Med=COMP Tr_Grupo_Ano_Soli=2004-2005 92760 ==> sme-
info_turno=M 80105    conf:(0.86)
 87. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP
121447 ==> Tr_Grupo_CID=13.0 104394    conf:(0.86)
 88. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0
102787 ==> Tr_Grupo_CID=13.0 87493    conf:(0.85)
 89. Tr_Grupo_Idade_Disp=>=71 129128 ==> Tr_Grupo_Idade_Entrada_Prog=>=70
109516    conf:(0.85)
 90. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0
102787 ==> Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 86898    conf:(0.85)
 91. Tr_Grupo_CID=13.0 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 107825 ==>
Tr_Grupo_Medic=1-3 90695    conf:(0.84)
 92. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 130370 ==> Tr_Grupo_Medic=1-3
109084    conf:(0.84)
 93. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
Tr_Grupo_Un_Med=COMP 104394 ==> paciente-sexo=2.0 86898    conf:(0.83)
 94. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP
112939 ==> paciente-sexo=2.0 93959    conf:(0.83)
 95. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
105197 ==> paciente-sexo=2.0 87493    conf:(0.83)
 96. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 109084 ==>
paciente-sexo=2.0 90695    conf:(0.83)
 97. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 113920 ==> paciente-
sexo=2.0 94664    conf:(0.83)
 98. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 110624 ==> paciente-sexo=2.0
91816    conf:(0.83)
 99. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 130370 ==> paciente-sexo=2.0
107825    conf:(0.83)
100. Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap 107689 ==> paciente-sexo=2.0
89016    conf:(0.83)
101. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
105197 ==> paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 86898    conf:(0.83)
102. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 113920 ==> paciente-
sexo=2.0 Tr_Grupo_Un_Med=COMP 93959    conf:(0.82)
103. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 110624 ==> paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 90695    conf:(0.82)
104. Tr_Grupo_Altura=141-160 94734 ==> paciente-sexo=2.0 77429
conf:(0.82)
105. Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap 107689 ==>
Tr_Grupo_Un_Med=COMP 87883    conf:(0.82)
```

```
106. Tr_Grupo_CID=13.0 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 107825 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 86898    conf:(0.81)
107. Tr_Grupo_CID=13.0 167429 ==> paciente-sexo=2.0 134681    conf:(0.8)
108. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 130370 ==> Tr_Grupo_Medic=1-3
Tr_Grupo_qtde_pedida=81-100 104394    conf:(0.8)
109. Tr_Grupo_CID=13.0 paciente-sexo=2.0 134681 ==> Tr_Grupo_Un_Med=COMP
107825    conf:(0.8)
```

Figura 25. Relatório gerado pelo software WEKA após a execução do algoritmo APRIORI com o suporte de 10% e grau de confiança de 80%, durante a realização do experimento 1.

7.3.7- 7ª Fase – Análise e interpretação dos resultados

Da análise das regras de associação geradas na etapa de mineração de dados deste experimento para o suporte e grau de confiança mínimos determinados, observou-se que foram geradas 109 regras de associação com o envolvimento de 12 atributos:

- Tr_Grupo_Medic;
- Tr_Grupo_qtde_pedida;
- sme-info_turno;
- Tr_Grupo_CID;
- Tr_Grupo_Guiche;
- Tr_Grupo_Altura;
- paciente-sexo;
- Tr_Grupo_Un_Med;
- Tr_Grupo_Ano_Soli;
- Tr_Grupo_Munic_Res;
- Tr_Grupo_Idade_Entrada_Prog;
- Tr_Grupo_Idade_Disp.

Grande parte das regras de associação geradas envolvem solicitações de medicamentos em comprimidos. Das regras geradas, 81 envolvem a unidade de medida do medicamento em comprimidos.

Quase 90% das solicitações de medicamentos com custo entre R$ 4,00 e R$ 10,00 são por comprimidos. Os medicamentos com valores entre R$ 1,00 e R$ 3,00 também são, geralmente, solicitados em comprimidos, principalmente nas quantidades entre 81 e 100, sendo mais comuns nas solicitações do guichê 3, por mulheres.

As regras também mostram que as solicitações de quantidades entre 81 e 100 de medicamentos geralmente são para atender às indicações de posologias em comprimidos. Vale ressaltar, que a grande maioria se refere à solicitação da quantidade de 90. Por conta da sistemática de fornecimento de medicamentos a cada 90 dias, pressupõe-se que a solicitação se refere ao consumo de 1 comprimido por dia.

Verifica-se, também, um indício de correlação entre a categoria da CID nº. 13 (doenças do sistema osteomuscular e do tecido conjuntivo) com pacientes do sexo feminino, pois cerca de 80% das solicitações de medicamentos que apontam algum código dessa categoria da CID são de pessoas do sexo feminino. Ainda sobre as mulheres, há uma referência significativa de que mais de 80% das solicitações de pessoas com 1,41 a 1,60 metros de altura também foram realizadas por pacientes do sexo feminino.

Observa-se que quando as mulheres solicitam uma quantidade entre 81 a 100 de medicamentos, geralmente indicam os comprimidos, para um período de 3 meses, o que reforça o entendimento sobre o consumo de 1 comprimido por dia.

É possível deduzir, ainda, que as solicitações dos medicamentos para doenças da categoria da CID nº. 13 – a maioria são de pacientes do sexo feminino – foram direcionadas para o guichê 3 e, geralmente, esses medicamentos têm custo de cerca de R$ 1,00 a R$ 3,00.

Ainda quanto à categoria da CID, das solicitações que informaram a nº. 5 (transtornos mentais e comportamentais) 97% se tratavam de busca por medicamentos em comprimidos.

Algumas regras mostram informações bem interessantes com relação à idade em que as pessoas entraram no programa e a idade no momento da solicitação.

Uma regra mostra que 98% das solicitações das pessoas que ingressaram no programa com idade até 19 anos foram realizadas quando as mesmas tinham idade de aproximadamente 25 anos. Outra aponta que mais de 90% das solicitações das pessoas com idade entre 61 e 70 anos no momento da solicitação eram de pessoas que ingressaram no programa com idade entre 57 a 69 anos.

Também foi identificado o fato de que 85% das solicitações de pessoas com mais de 70 anos de idade no momento da solicitação eram de pessoas que ingressaram no programa já com 70 anos ou mais.

Por outro lado, vale destacar, até mesmo por se tratar de uma busca automatizada, que uma regra bem óbvia foi encontrada. Trata-se da informação de que as solicitações de pessoas que ingressaram com mais de 70 anos no programa foram realizadas quando as mesmas tinham mais de 70 anos no momento da solicitação.

Chamou a atenção o fato de que mais de 85% das solicitações realizadas entre os anos de 2004 e 2005 ocorreram no período da manhã. A princípio, deduz-se que o atendimento era mais direcionado para esse turno naquela época.

Em algumas regras apareceram referências a solicitações de pacientes residentes na capital, entretanto, após uma avaliação, foi observado que se tratam apenas de confirmações das associações apontadas anteriormente. O que ocorre é que pelo fato de haver uma grande demanda de medicamentos de pacientes residentes na capital, as associações anteriores também valem para esse grupo de pacientes específico.

Todas essas afirmações foram confirmadas junto à diretoria da unidade.

7.4- Avaliação do impacto do uso da metodologia nos resultados

Para avaliar o impacto do uso da metodologia proposta neste trabalho nos resultados da aplicação do algoritmo Apriori um outro experimento foi realizado, denominado experimento 2.

Como o principal foco da metodologia proposta é a etapa de pré-processamento dos dados, com o objetivo de proporcionar uma melhor preparação dos dados para torná-los mais adequados ao processamento do algoritmo Apriori, constatou-se uma oportunidade para verificar e comparar os resultados que seriam alcançados, caso alguns dos procedimentos da etapa de pré-processamento da metodologia deixassem de ser realizados.

Dessa forma, busca-se verificar e avaliar o principal questionamento deste trabalho, que é o de verificar a influência da etapa de pré-processamento nos resultados do algoritmo Apriori.

Assim, apenas para fins de comparação, o experimento 2 visa demonstrar que, se no momento da conclusão da 2ª fase da aplicação da metodologia proposta no experimento 1 fosse tomada a decisão de submeter os dados até então selecionados imediatamente ao processamento do algoritmo Apriori, os resultados que seriam gerados seriam os demonstrados conforme discriminado na figura 26 a seguir:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 200 -T 0 -C 0.8 -D 0.05 -U 1.0 -
M 0.1 -S -1.0 -c -1
Relation:     Exp1
Instances:    746355
Attributes:   16
              sme_medicamento-codg_medicamento
              sme_medicamento-qtde_pedida
              sme-info_turno
              processo-codg_unidade
              processo-info_cid1
              processo-data_sme
              cartao-guiche
              cartao-codg_munic
              cartao-peso
              cartao-altura
              paciente-sexo
              paciente-codg_raca_cor
              paciente-cod_naturalidade
              paciente-data_nascimento
              paciente-data_entr_ss
              tab_esto-codg_unidade
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (74636 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 18

Size of set of large itemsets L(2): 31

Size of set of large itemsets L(3): 7

Best rules found:

  1. sme_medicamento-codg_medicamento=14300.0 80741 ==> tab_esto-
codg_unidade=COMP 80741    conf:(1)
  2. sme_medicamento-codg_medicamento=14300.0 sme_medicamento-
qtde_pedida=90.0 77223 ==> tab_esto-codg_unidade=COMP 77223    conf:(1)
```

```
  3. sme_medicamento-qtde_pedida=90.0 cartao-guiche=3.0 77965 ==> tab_esto-
codg_unidade=COMP 77777    conf:(1)
  4. sme_medicamento-codg_medicamento=14300.0 80741 ==> sme_medicamento-
qtde_pedida=90.0 77223    conf:(0.96)
  5. sme_medicamento-codg_medicamento=14300.0 tab_esto-codg_unidade=COMP
80741 ==> sme_medicamento-qtde_pedida=90.0 77223    conf:(0.96)
  6. sme_medicamento-codg_medicamento=14300.0 80741 ==> sme_medicamento-
qtde_pedida=90.0 tab_esto-codg_unidade=COMP 77223    conf:(0.96)
  7. sme_medicamento-qtde_pedida=90.0 paciente-sexo=2.0 150585 ==>
tab_esto-codg_unidade=COMP 140645    conf:(0.93)
  8. sme_medicamento-qtde_pedida=90.0 sme-info_turno=M 101005 ==> tab_esto-
codg_unidade=COMP 92391    conf:(0.91)
  9. sme_medicamento-qtde_pedida=90.0 234251 ==> tab_esto-codg_unidade=COMP
209930    conf:(0.9)
 10. cartao-guiche=3.0 tab_esto-codg_unidade=COMP 89302 ==>
sme_medicamento-qtde_pedida=90.0 77777    conf:(0.87)
```

Figura 26. Relatório gerado pelo software WEKA, após a aplicação do algoritmo Apriori com definição do suporte mínimo de 10% e grau de confiança mínimo de 80%, durante a realização do experimento 2.

Nesse sentido, os dados de 16 atributos selecionados do conjunto de dados com informações sobre solicitações de medicamentos (conforme se observou nas tabelas 5 e 6 e figura 23) teriam sido submetidos ao processamento do algoritmo Apriori basicamente como registrados e extraídos da base de dados, ou seja, sem tratamentos ou transformações.

Da análise das regras de associação geradas na etapa de mineração de dados observou-se que foram geradas somente 10 regras de associação com o envolvimento de apenas 6 atributos:

- sme_medicamento-codg_medicamento;
- sme_medicamento-qtde_pedida;
- sme-info_turno;
- cartao-guiche;
- paciente-sexo;
- tab_esto-codg_unidade.

A principal associação encontrada se refere aos medicamentos distribuídos em comprimidos, que geralmente são solicitados na quantidade de 90.

A regra número 9 informa que das solicitações com indicação da quantidade 90 (que representa 30% do total das solicitações), 90% são por comprimidos.

Apenas como análise dessa associação específica, vale destacar que nos resultados da tabulação realizada, observou-se que mais de 50% do total de solicitações registradas na base de dados são de comprimidos, que é a maior

demanda, sendo que a segunda maior representa pouco mais de 15%. Nota-se que a demanda por medicamentos em comprimidos é bem maior. Isso pode ser influência do fato de que dos 220 medicamentos cadastrados na base de dados, utilizados ao longo do tempo para atender às solicitações registradas, 94 (43 %) são adquiridos para o atendimento às posologias indicadas em comprimidos, seguida de frascos ou ampolas (38 – 17 %) e cápsulas (35 – a 16 %). Já com relação à quantidade solicitada, verificou-se em que cerca de 30% de todas as solicitações há a indicação de 90 unidades.

Sabe-se, ainda, que os pacientes solicitam medicamentos para o seu tratamento durante 3 meses. Assim, percebe-se que há uma tendência de uma previsão para o uso de 1 comprimido por dia. Este fato foi confirmado junto a diretoria da unidade.

As demais regras encontradas são apenas partes da associação principal em determinadas situações específicas.

O medicamento com código 14300 somente apareceu entre as regras por ser o único a ter frequência acima do suporte mínimo de 10%, mas apenas confirma a associação principal. A associação com esse medicamento é esperada, já que sua distribuição se dá através de comprimidos.

Da mesma forma, as variáveis sexo do paciente, guichê que emitiu o cartão e o turno da solicitação possuem itens que concentram muitos dados e isso refletiu em coincidências de várias combinações desses itens com a associação principal encontrada, o que reforça essa associação. Isso porque, se mais de 90% das solicitações de medicamentos na quantidade de 90 se referem à posologia em comprimidos, então é natural que solicitações com indicação da quantidade 90 por determinado sexo, em determinado guichê ou em determinado turno, também sejam em comprimidos. Contudo, se essas variáveis também estão envolvidas entre as associações, vale a realização de um estudo mais profundo desses itens específicos de sexo, guichês e turno apontados entre as regras.

7.5- Experimento com dados mais recentes

Para aprofundar o estudo com mais comparações, optou-se por realizar um outro experimento, denominado experimento 3. A experiência foi realizada no sentido de verificar os resultados da aplicação do processo de mineração de dados, com o uso da metodologia proposta neste trabalho, apenas em uma parte da base de dados em que há as informações mais recentes sobre as solicitações de medicamentos.

Essa base de dados foi gerenciada por sistemas de informação que foram aprimorados no decorrer dos anos. Assim, pressupõe-se que os dados mais recentes tenham mais qualidade, especialmente no que se refere ao seu preenchimento. Observou-se, inclusive, que muitas informações de alguns atributos sobre a solicitação, como, raça e cor, sexo, peso, altura, naturalidade e residência do paciente, por exemplo, estão menos ausentes entre os dados mais recentes. Vale destacar que alguns dos outros atributos apresentam uma ausência maior neste período.

Foi constatado que houve uma profunda mudança no sistema de informação da unidade de saúde Juarez Barbosa a partir do ano de 2006. Novos recursos e funções foram adicionados ao sistema. Há uma queda substancial de ausência de dados, a partir desse período, entre os atributos trabalhados nos processos de mineração de dados realizados através dos experimentos anteriores. Por isso, vale uma investigação no sentido de verificar se, com essa queda de ausência de dados, outras associações podem ser descobertas.

A metodologia proposta sugere que o processo seja iterativo, ou seja, todo o processo de mineração de dados pode ser realizado novamente após a análise dos resultados já encontrados. Nesse sentido, para aprimorar os resultados da busca por associações, foi realizado este outro experimento.

Em síntese, todos os procedimentos adotados no experimento 1, em que a metodologia proposta foi completamente aplicada no conjunto de dados, foram realizados novamente. O processo foi realizado da mesma forma e nas mesmas condições, com a diferença apenas em relação ao fato de que os dados das

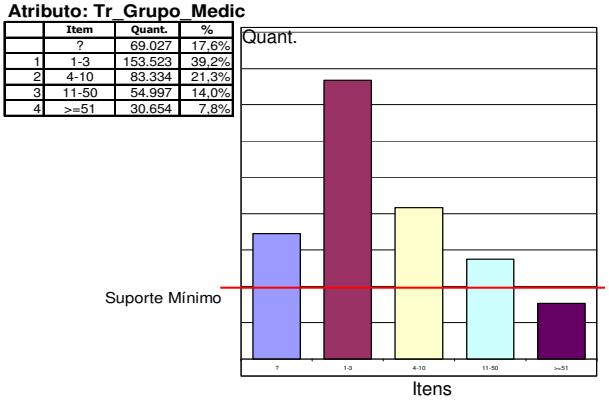
solicitações de medicamentos manipulados neste experimento se referem apenas ao período de solicitação à partir do ano de 2006.

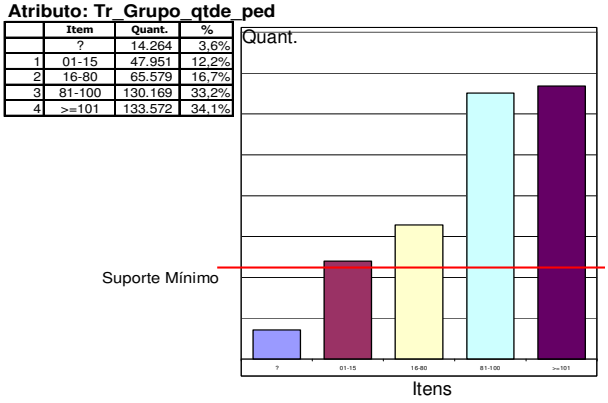
7.5.1- Visualização dos dados

Do processo de pré-processamento de dados desse experimento resultaram 391.535 registros com informações organizadas nos mesmos 18 atributos selecionados e considerados no experimento 1, conforme tabela 11. As principais informações dos atributos trabalhados neste experimento 3 podem ser visualizadas na tabela 13 e na figura 27 a seguir.

Identificação	Taxa de ausência de dados	Número de itens distintos	Quant. de itens com frequência acima de 10%	Quant. de itens com frequência acima de 5%	Quant. de itens com frequência acima de 1%
Tr_Grupo_Medic	17,6%	4	3	4	4
Tr_Grupo_qtde_pedida	3,6%	4	4	4	4
sme-info_turno	41,1%	2	2	2	2
Tr_Grupo_CID	0,6%	18	4	6	10
Tr_Grupo_Guiche	40,8%	4	3	4	4
Tr_Grupo_Peso	40,2%	3	2	2	3
Tr_Classe_IMC	41,9%	6	2	4	5
Tr_Grupo_Altura	41,6%	4	2	2	3
paciente-sexo	2,2%	2	2	2	2
Tr_Grupo_Raca_Cor	26,2%	3	2	2	3
Tr_Grupo_Un_Med	2,9%	4	4	4	4
Tr_Grupo_Ano_Soli	0,0%	4	4	4	4
Tr_Trim_Solicitacao	0,0%	4	4	4	4
Tr_Grupo_Munic_Res	0,0%	5	4	4	4
Tr_Grupo_Munic_Nat	8,5%	5	5	5	5
Tr_Grupo_Idade_Entrada_Prog	0,9%	6	6	6	6
Tr_Grupo_Idade_Disp	0,1%	6	6	6	6
Tr_Grupo_Tempo_Prog_Disp	0,9%	7	6	6	6

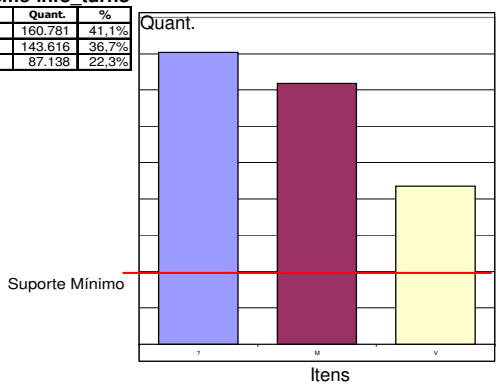
Tabela 13. Taxa de ausência de dados e quantidade de itens mais frequentes nos atributos selecionados para processamento do algoritmo Apriori no experimento 3.





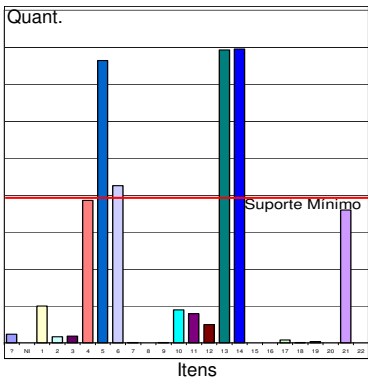
Atributo: sme-info_turno

	Item	Quant.	%
	?	160.781	41,1%
1	M	143.616	36,7%
2	V	87.138	22,3%



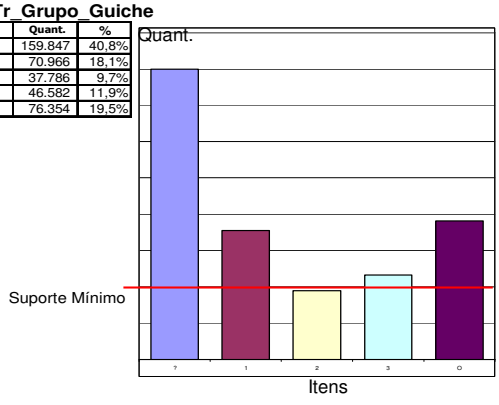
Atributo: Tr_Grupo_CID

	Item	Quant.	%
	?	2.329	0,6%
1	NI	5	0,0%
2	1	9.971	2,5%
3	2	1.627	0,4%
4	3	1.878	0,5%
5	4	38.536	9,8%
6	5	76.432	19,5%
7	6	42.580	10,9%
8	7	158	0,0%
9	8	-	0,0%
10	9	44	0,0%
11	10	8.953	2,3%
12	11	7.938	2,0%
13	12	4.997	1,3%
14	13	79.321	20,3%
15	14	79.567	20,3%
16	15	-	0,0%
17	16	-	0,0%
18	17	811	0,2%
19	18	47	0,0%
20	19	352	0,1%
21	20	-	0,0%
22	21	35.989	9,2%
23	22	-	0,0%



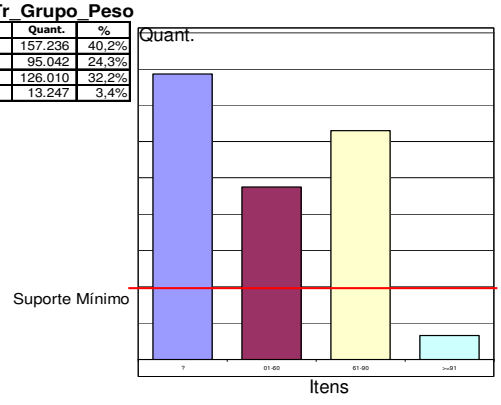
Atributo: Tr_Grupo_Guiche

	Item	Quant.	%
	?	159.847	40,8%
1	1	70.966	18,1%
2	2	37.786	9,7%
3	3	46.582	11,9%
4	O	76.354	19,5%



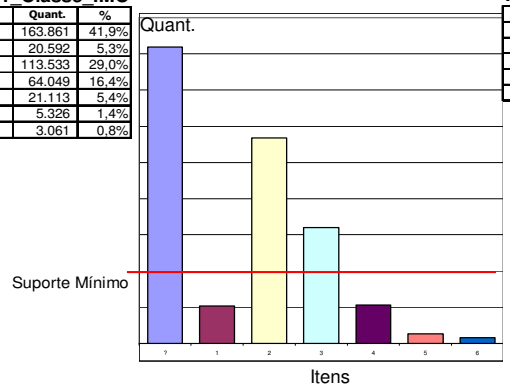
Atributo: Tr_Grupo_Peso

	Item	Quant.	%
	?	157.236	40,2%
1	01-60	95.042	24,3%
2	61-90	126.010	32,2%
3	>=91	13.247	3,4%



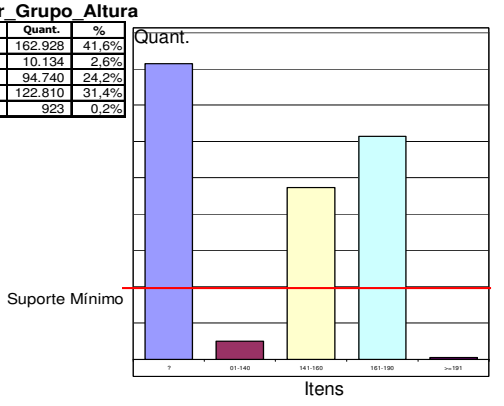
Atributo: Tr_Classe_IMC

	Item	Quant.	%
	?	163.861	41,9%
1	1	20.592	5,3%
2	2	113.533	29,0%
3	3	64.049	16,4%
4	4	21.113	5,4%
5	5	5.326	1,4%
6	6	3.061	0,8%



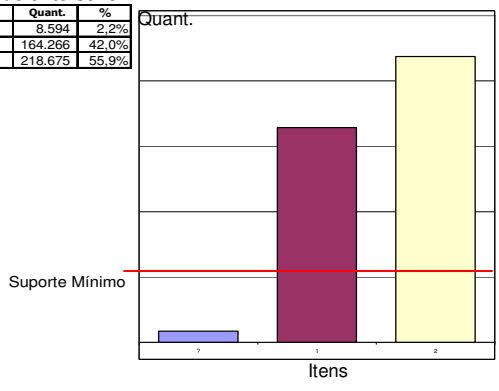
Atributo: Tr_Grupo_Altura

	Item	Quant.	%
	?	162.928	41,6%
1	01-140	10.134	2,6%
2	141-160	94.740	24,2%
3	161-190	122.810	31,4%
4	>=191	923	0,2%



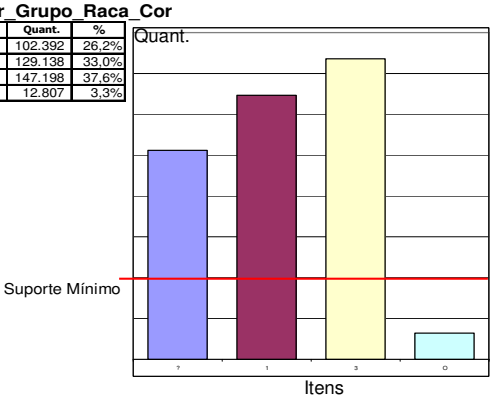
Atributo: paciente-sexo

	Item	Quant.	%
	?	8.594	2,2%
1	1	164.266	42,0%
2	2	218.675	55,9%



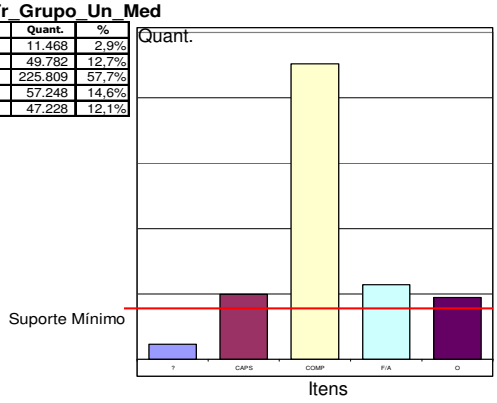
Atributo: Tr_Grupo_Raca_Cor

	Item	Quant.	%
	?	102.392	26,2%
1	1	129.138	33,0%
2	3	147.198	37,6%
3	O	12.807	3,3%



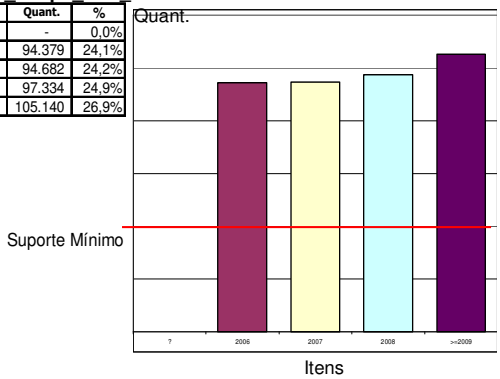
Atributo: Tr_Grupo_Un_Med

	Item	Quant.	%
	?	11.468	2,9%
1	CAPS	49.782	12,7%
2	COMP	225.809	57,7%
3	F/A	57.248	14,6%
4	O	47.228	12,1%



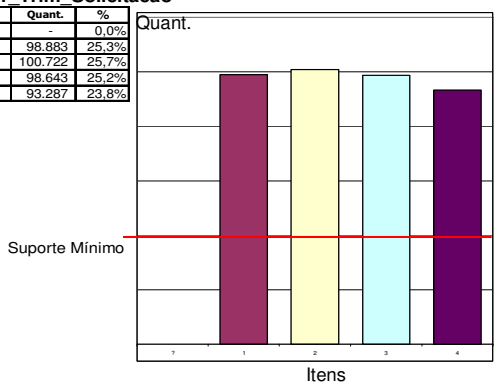
Atributo: Tr_Grupo_Ano_Soli

	Item	Quant.	%
	?	-	0,0%
1	2006	94.379	24,1%
2	2007	94.682	24,2%
3	2008	97.334	24,9%
4	>=2009	105.140	26,9%



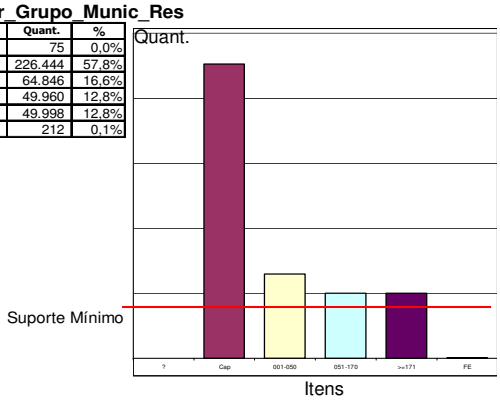
Atributo: Tr_Trim_Solicitacao

	Item	Quant.	%
	?	-	0,0%
1	1	98.883	25,3%
2	2	100.722	25,7%
3	3	98.643	25,2%
4	4	93.287	23,8%



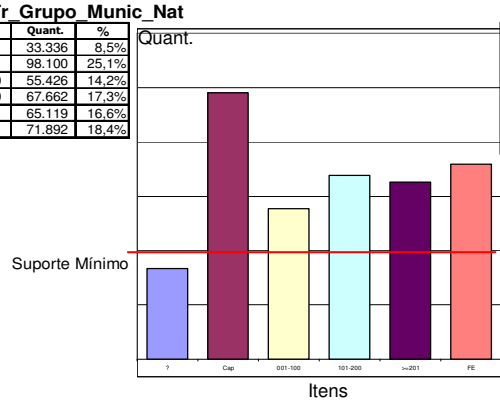
Atributo: Tr_Grupo_Munic_Res

	Item	Quant.	%
	?	75	0,0%
1	Cap	226.444	57,8%
2	001-050	64.846	16,6%
3	051-170	49.960	12,8%
4	>=171	49.998	12,8%
5	FE	212	0,1%



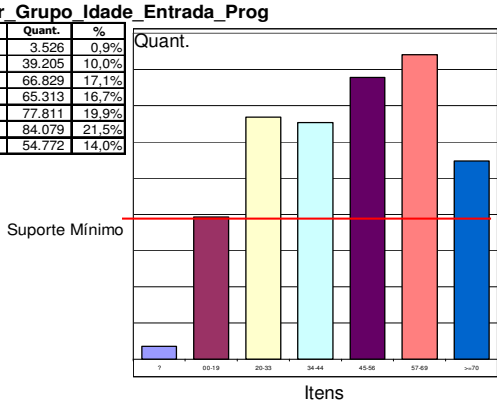
Atributo: Tr_Grupo_Munic_Nat

	Item	Quant.	%
	?	33.336	8,5%
1	Cap	98.100	25,1%
2	001-100	55.426	14,2%
3	101-200	67.662	17,3%
4	>=201	65.119	16,6%
5	FE	71.892	18,4%



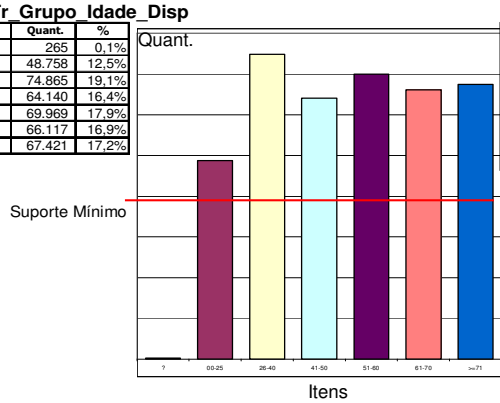
Atributo: Tr_Grupo_Idade_Entrada_Prog

	Item	Quant.	%
	?	3.526	0,9%
1	00-19	39.205	10,0%
2	20-33	66.829	17,1%
3	34-44	65.313	16,7%
4	45-56	77.811	19,9%
5	57-69	84.079	21,5%
6	>=70	54.772	14,0%



Atributo: Tr_Grupo_Idade_Disp

	Item	Quant.	%
	?	265	0,1%
1	00-25	48.758	12,5%
2	26-40	74.865	19,1%
3	41-50	64.140	16,4%
4	51-60	69.969	17,9%
5	61-70	66.117	16,9%
6	>=71	67.421	17,2%



Atributo: Tr_Grupo_Tempo_Prog_Disp

	Item	Quant.	%
	?	3.345	0,9%
1	0	62.397	15,9%
2	1	67.663	17,3%
3	2	54.611	13,9%
4	3	47.061	12,0%
5	04-05	72.256	18,5%
6	06-10	79.239	20,2%
7	>=11	4.963	1,3%

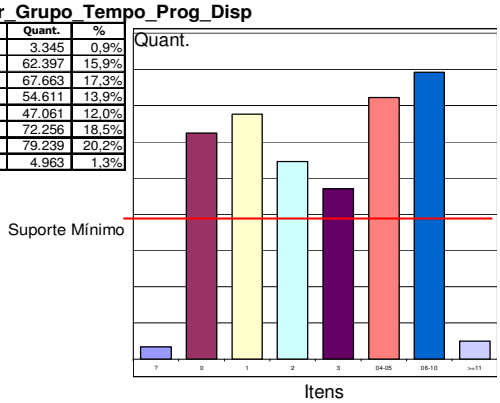


Figura 27. Gráficos com as frequências dos itens de cada atributo selecionado para a aplicação do algoritmo Apriori no experimento 3.

7.5.2- Mineração dos dados

Da mesma forma e nas mesmas condições de como foi realizado no experimento 1, o software WEKA foi configurado para que o algoritmo Apriori tentasse encontrar até 200 regras de associação, diante dos mesmos suportes e o grau de confiança mínimos determinados.

Assim, neste conjunto de dados pré-processados o algoritmo foi aplicado com a definição de um grau de suporte mínimo de 10% e de um grau de confiança mínimo de 80% (oitenta por cento), conforme demonstrado na figura 28 a seguir.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 200 -T 0 -C 0.8 -D 0.05 -U 1.0 -
M 0.1 -S -1.0 -c -1
Relation:     Conjunto-partir—ano-2006
Instances:    391535
Attributes:   18
              Tr_Grupo_Medic
              Tr_Grupo_qtde_pedida
              sme-info_turno
              Tr_Grupo_CID
              Tr_Grupo_Guiche
              Tr_Grupo_Peso
              Tr_Grupo_Altura
              paciente-sexo
              Tr_Grupo_Raca_Cor
              Tr_Grupo_Un_Med
              Tr_Grupo_Ano_Soli
              Tr_Trim_Solicitacao
              Tr_Grupo_Munic_Res
              Tr_Grupo_Munic_Nat
              Tr_Grupo_Idade_Entrada_Prog
              Tr_Grupo_Idade_Disp
              Tr_Grupo_Tempo_Prog_Disp
              Tr_Classe_IMC
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (39154 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 65

Size of set of large itemsets L(2): 204

Size of set of large itemsets L(3): 59
```

```
Size of set of large itemsets L(4): 12

Size of set of large itemsets L(5): 1

Best rules found:

  1. Tr_Grupo_Medic=4-10 Tr_Grupo_CID=5.0 44652 ==> Tr_Grupo_Un_Med=COMP
44646    conf:(1)
  2. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap
41540 ==> Tr_Grupo_Un_Med=COMP 41335    conf:(1)
  3. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
paciente-sexo=2.0 48361 ==> Tr_Grupo_Un_Med=COMP 48122    conf:(1)
  4. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 52879
==> Tr_Grupo_Un_Med=COMP 52601    conf:(0.99)
  5. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0 52882
==> Tr_Grupo_Un_Med=COMP 52589    conf:(0.99)
  6. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 58158 ==>
Tr_Grupo_Un_Med=COMP 57809    conf:(0.99)
  7. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap 41094 ==>
Tr_Grupo_Un_Med=COMP 40728    conf:(0.99)
  8. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0 51457 ==>
Tr_Grupo_Un_Med=COMP 50928    conf:(0.99)
  9. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 56374 ==> Tr_Grupo_Un_Med=COMP
55719    conf:(0.99)
 10. Tr_Grupo_qtde_pedida=>=101 Tr_Grupo_CID=5.0 39920 ==>
Tr_Grupo_Un_Med=COMP 39318    conf:(0.98)
 11. Tr_Grupo_Idade_Entrada_Prog=>=70 54772 ==> Tr_Grupo_Idade_Disp=>=71
53761    conf:(0.98)
 12. Tr_Grupo_CID=5.0 paciente-sexo=1.0 41055 ==> Tr_Grupo_Un_Med=COMP
39676    conf:(0.97)
 13. Tr_Grupo_CID=5.0 76432 ==> Tr_Grupo_Un_Med=COMP 73751    conf:(0.96)
 14. Tr_Grupo_CID=5.0 Tr_Grupo_Munic_Res=Cap 44465 ==> Tr_Grupo_Un_Med=COMP
42849    conf:(0.96)
 15. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Idade_Entrada_Prog=57-69 42043
==> Tr_Grupo_Un_Med=COMP 40392    conf:(0.96)
 16. Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 Tr_Grupo_Munic_Res=Cap
61596 ==> Tr_Grupo_Un_Med=COMP 58699    conf:(0.95)
 17. Tr_Grupo_CID=5.0 Tr_Grupo_Guiche=O 42158 ==> Tr_Grupo_Un_Med=COMP
40088    conf:(0.95)
 18. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 50928 ==> Tr_Grupo_qtde_pedida=81-100 48122
conf:(0.94)
 19. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 55719 ==>
Tr_Grupo_qtde_pedida=81-100 52601    conf:(0.94)
 20. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 56356
==> Tr_Grupo_Un_Med=COMP 53036    conf:(0.94)
 21. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0 51457 ==>
Tr_Grupo_qtde_pedida=81-100 48361    conf:(0.94)
 22. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 56374 ==>
Tr_Grupo_qtde_pedida=81-100 52879    conf:(0.94)
 23. Tr_Grupo_Medic=4-10 Tr_Grupo_Munic_Res=Cap 47948 ==>
Tr_Grupo_Un_Med=COMP 44963    conf:(0.94)
 24. Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 90919 ==>
Tr_Grupo_Un_Med=COMP 85136    conf:(0.94)
 25. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 paciente-sexo=2.0 51457 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP 48122    conf:(0.94)
 26. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 56374 ==>
Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP 52601    conf:(0.93)
 27. Tr_Grupo_Medic=4-10 83334 ==> Tr_Grupo_Un_Med=COMP 77538
conf:(0.93)
```

```
 28. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Munic_Res=Cap 84919 ==>
Tr_Grupo_Un_Med=COMP 78931    conf:(0.93)
 29. Tr_Grupo_Altura=141-160 Tr_Classe_IMC=2.0 46365 ==> Tr_Grupo_Peso=01-
60 43059    conf:(0.93)
 30. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Munic_Res=Cap
47897 ==> Tr_Grupo_Un_Med=COMP 44365    conf:(0.93)
 31. Tr_Grupo_Peso=61-90 Tr_Classe_IMC=2.0 47667 ==> Tr_Grupo_Altura=161-
190 44138    conf:(0.93)
 32. Tr_Grupo_Medic=4-10 paciente-sexo=1.0 42843 ==> Tr_Grupo_Un_Med=COMP
39644    conf:(0.93)
 33. paciente-sexo=2.0 Tr_Grupo_Idade_Disp=61-70 43516 ==>
Tr_Grupo_Idade_Entrada_Prog=57-69 40058    conf:(0.92)
 34. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Raca_Cor=1.0 46407 ==>
Tr_Grupo_Un_Med=COMP 42692    conf:(0.92)
 35. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 52589 ==> Tr_Grupo_Medic=1-3 48122    conf:(0.92)
 36. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0
Tr_Grupo_Un_Med=COMP 52601 ==> paciente-sexo=2.0 48122    conf:(0.91)
 37. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 52879
==> paciente-sexo=2.0 48361    conf:(0.91)
 38. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0 52882
==> Tr_Grupo_Medic=1-3 48361    conf:(0.91)
 39. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 55719 ==>
paciente-sexo=2.0 50928    conf:(0.91)
 40. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 56374 ==> paciente-sexo=2.0 51457
conf:(0.91)
 41. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 52879
==> paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 48122    conf:(0.91)
 42. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 paciente-sexo=2.0 52882
==> Tr_Grupo_Medic=1-3 Tr_Grupo_Un_Med=COMP 48122    conf:(0.91)
 43. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP
57809 ==> Tr_Grupo_Medic=1-3 52601    conf:(0.91)
 44. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP
57809 ==> paciente-sexo=2.0 52589    conf:(0.91)
 45. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 58158 ==> paciente-
sexo=2.0 52882    conf:(0.91)
 46. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 58158 ==>
Tr_Grupo_Medic=1-3 52879    conf:(0.91)
 47. Tr_Grupo_qtde_pedida=81-100 sme-info_turno=M 48813 ==>
Tr_Grupo_Un_Med=COMP 44373    conf:(0.91)
 48. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP Tr_Grupo_Munic_Res=Cap 46350
==> paciente-sexo=2.0 42134    conf:(0.91)
 49. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 53036 ==> Tr_Grupo_CID=13.0 48122    conf:(0.91)
 50. Tr_Grupo_qtde_pedida=81-100 130169 ==> Tr_Grupo_Un_Med=COMP 117875
conf:(0.91)
 51. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 66469 ==> paciente-sexo=2.0
60145    conf:(0.9)
 52. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 58158 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_Un_Med=COMP 52601    conf:(0.9)
 53. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 58158 ==> paciente-
sexo=2.0 Tr_Grupo_Un_Med=COMP 52589    conf:(0.9)
 54. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 56374 ==> paciente-sexo=2.0
Tr_Grupo_Un_Med=COMP 50928    conf:(0.9)
 55. Tr_Grupo_Idade_Disp=61-70 66117 ==> Tr_Grupo_Idade_Entrada_Prog=57-69
59709    conf:(0.9)
 56. Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap 53718 ==> paciente-sexo=2.0
48121    conf:(0.9)
 57. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 68500 ==>
Tr_Grupo_Un_Med=COMP 61120    conf:(0.89)
```

```
 58. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP Tr_Grupo_Munic_Res=Cap 46350
==> Tr_Grupo_qtde_pedida=81-100 41335    conf:(0.89)
 59. Tr_Grupo_CID=13.0 79321 ==> paciente-sexo=2.0 70337    conf:(0.89)
 60. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Raca_Cor=3.0 44823 ==>
Tr_Grupo_Un_Med=COMP 39730    conf:(0.89)
 61. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP Tr_Grupo_Munic_Res=Cap 46350
==> Tr_Grupo_Medic=1-3 40728    conf:(0.88)
 62. Tr_Classe_IMC=3.0 64049 ==> Tr_Grupo_Peso=61-90 56278    conf:(0.88)
 63. Tr_Grupo_CID=13.0 paciente-sexo=2.0 Tr_Grupo_Munic_Res=Cap 48121 ==>
Tr_Grupo_Un_Med=COMP 42134    conf:(0.88)
 64. Tr_Grupo_CID=13.0 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 60145 ==>
Tr_Grupo_qtde_pedida=81-100 52589    conf:(0.87)
 65. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 66469 ==>
Tr_Grupo_qtde_pedida=81-100 57809    conf:(0.87)
 66. Tr_Grupo_Peso=61-90 paciente-sexo=1.0 65473 ==> Tr_Grupo_Altura=161-
190 56839    conf:(0.87)
 67. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP
61120 ==> paciente-sexo=2.0 53036    conf:(0.87)
 68. Tr_Grupo_Guiche=3.0 46582 ==> paciente-sexo=2.0 40346    conf:(0.87)
 69. Tr_Grupo_Altura=141-160 Tr_Grupo_Un_Med=COMP 56752 ==> paciente-
sexo=2.0 49147    conf:(0.87)
 70. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 55719 ==>
Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 48122    conf:(0.86)
 71. Tr_Grupo_CID=13.0 Tr_Grupo_Munic_Res=Cap 53718 ==>
Tr_Grupo_Un_Med=COMP 46350    conf:(0.86)
 72. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Un_Med=COMP
61120 ==> Tr_Grupo_CID=13.0 52601    conf:(0.86)
 73. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 56356
==> Tr_Grupo_CID=13.0 48361    conf:(0.86)
 74. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 56374 ==>
Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 48361    conf:(0.86)
 75. Tr_Grupo_CID=13.0 paciente-sexo=2.0 70337 ==> Tr_Grupo_Un_Med=COMP
60145    conf:(0.86)
 76. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 56356
==> Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 48122    conf:(0.85)
 77. Tr_Grupo_Medic=1-3 Tr_Grupo_CID=13.0 56374 ==>
Tr_Grupo_qtde_pedida=81-100 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 48122
conf:(0.85)
 78. Tr_Grupo_Guiche=3.0 46582 ==> Tr_Grupo_CID=13.0 39458    conf:(0.85)
 79. Tr_Grupo_CID=13.0 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 60145 ==>
Tr_Grupo_Medic=1-3 50928    conf:(0.85)
 80. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_Munic_Res=Cap
47897 ==> paciente-sexo=2.0 40406    conf:(0.84)
 81. Tr_Grupo_Altura=141-160 Tr_Grupo_Munic_Res=Cap 55089 ==> paciente-
sexo=2.0 46436    conf:(0.84)
 82. Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP 66469 ==> Tr_Grupo_Medic=1-3
55719    conf:(0.84)
 83. Tr_Grupo_CID=13.0 79321 ==> Tr_Grupo_Un_Med=COMP 66469    conf:(0.84)
 84. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 Tr_Grupo_Un_Med=COMP
57809 ==> Tr_Grupo_Medic=1-3 paciente-sexo=2.0 48122    conf:(0.83)
 85. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 58158 ==>
Tr_Grupo_Medic=1-3 paciente-sexo=2.0 48361    conf:(0.83)
 86. Tr_Grupo_qtde_pedida=81-100 Tr_Grupo_CID=13.0 58158 ==>
Tr_Grupo_Medic=1-3 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 48122
conf:(0.83)
 87. Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 68500 ==> paciente-
sexo=2.0 56356    conf:(0.82)
 88. Tr_Grupo_Altura=141-160 94740 ==> paciente-sexo=2.0 77430
conf:(0.82)
 89. Tr_Grupo_Idade_Entrada_Prog=20-33 66829 ==> Tr_Grupo_Idade_Disp=26-40
54487    conf:(0.82)
```

```
 90. Tr_Grupo_Peso=01-60 Tr_Grupo_Altura=141-160 55797 ==> paciente-
sexo=2.0 45428    conf:(0.81)
 91. paciente-sexo=1.0 Tr_Classe_IMC=2.0 49387 ==> Tr_Grupo_Altura=161-190
40021    conf:(0.81)
 92. Tr_Grupo_CID=13.0 paciente-sexo=2.0 Tr_Grupo_Un_Med=COMP 60145 ==>
Tr_Grupo_Medic=1-3 Tr_Grupo_qtde_pedida=81-100 48122    conf:(0.8)
```

Figura 28. Relatório gerado pelo software WEKA após a execução do algoritmo APRIORI com o suporte de 10% e grau de confiança de 80%, durante a realização do experimento 3.

7.5.3– Análise e interpretação dos resultados

Da análise das regras de associação geradas na etapa de mineração de dados deste experimento, observou-se que foram geradas 92 regras de associação com o envolvimento de 14 atributos:

- Tr_Grupo_Medic;
- Tr_Grupo_qtde_pedida;
- sme-info_turno;
- Tr_Grupo_CID;
- Tr_Grupo_Guiche;
- Tr_Grupo_Peso;
- Tr_Classe_IMC;
- Tr_Grupo_Altura;
- paciente-sexo;
- Tr_Grupo_Raca_Cor;
- Tr_Grupo_Un_Med;
- Tr_Grupo_Munic_Res;
- Tr_Grupo_Idade_Entrada_Prog;
- Tr_Grupo_Idade_Disp.

Muitas das informações encontradas no experimento 1 foram confirmadas, com apenas algumas exceções de regras que não apareceram durante a realização deste experimento, conforme demonstrado a seguir.

Como aconteceu no experimento 1, grande parte das regras de associação geradas envolvem medicamentos em comprimidos. Das 92 regras geradas, 64 envolvem a solicitação do medicamento em comprimido. As solicitações de quantidades entre 81 e 100 de medicamentos geralmente são para atender às indicações de posologias em comprimidos.

Esta associação ocorre, por exemplo, em mais de 90% dos casos de pacientes do sexo feminino, da capital do Estado e com doença da categoria da CID 13 (doenças do sistema osteomuscular e do tecido conjuntivo). Os pacientes dessa categoria de doença, na maioria mulheres, solicitam medicamentos com custo entre R$ 1,00 a 3,00. As solicitações desse quantitativo em comprimidos também são realizadas mais frequentemente por pacientes que entraram no programa com idade entre 57 a 69 anos, sendo uma grande parte por pessoas da cor branca e parda. Há uma frequência de mais de 90% das solicitações de comprimidos nessa quantidade no turno da manhã.

Dessas solicitações, chamou a atenção um indício muito forte de correlação entre a categoria da CID nº. 13 (doenças do sistema osteomuscular e do tecido conjuntivo) com pacientes do sexo feminino.

Quase 90% das solicitações direcionadas para o guichê 3 são de pacientes do sexo feminino e para as doenças da categoria da CID 13.

Também há uma referência significativa de que mais de 80% das solicitações de pessoas com 1,41 a 1,60 metros de altura e com peso de até 60 kg também foram realizadas por pacientes do sexo feminino, especialmente da capital do Estado.

Mais de 90% das solicitações de medicamentos com custo entre R$ 4,00 e R$ 10,00 são por comprimidos. A maior parte das solicitações é de pacientes da capital do Estado, do sexo masculino e com doença classificada na categoria da CID 5 (transtornos mentais e comportamentais).

Os pacientes com altura entre 1,41 e 1,60 metros e na classe de IMC 2 (saudável) possuem peso até 60 kg. Os pacientes da classe de IMC 3 (sobrepeso), geralmente possuem peso entre 61 a 90 kg.

Já os pacientes com peso entre 61 a 90 kg e na classe IMC 2 (saudável) possuem altura entre 1,61 a 1,90 metros, sendo a maioria do sexo masculino.

Os pacientes, especialmente do sexo feminino, que solicitaram medicamentos quando tinham idade entre 61 a 70 anos entraram no programa quando tinham idade entre 57 a 69 anos.

As pessoas que ingressaram no programa com idade entre 20 a 33 anos passaram a solicitar mais medicamentos quando tinham entre 26 e 40 anos.

Novamente a mesma regra de associação óbvia do experimento 1 apareceu, por conta da sistemática de uma busca automatizada, em que a maioria

das pessoas que ingressaram no programa com idade maior que 70 anos começaram a solicitar medicamentos quando tinham mais de 71 anos.

8- ANÁLISE DOS RESULTADOS DO ESTUDO DE CASO

Para o estudo de caso, três experimentos de aplicação do processo de mineração de dados foram realizados, sendo diferentes, apenas, quanto a alguns procedimentos realizados na etapa de pré-processamento de cada experimento.

Com o destaque para essa peculiaridade, os experimentos foram realizados sob as mesmas condições, as quais, resumidamente, podem ser destacadas:

- utilização da ferramenta WEKA;
- aplicação do algoritmo Apriori, com a definição dos parâmetros mínimos de suporte de 10% e de confiança de 80%;
- seleção inicial de dados dos mesmos atributos como entrada do processo de mineração de dados;
- mesmo ambiente computacional.

Com relação às diferenças dos procedimentos adotados na etapa de pré-processamento dos experimentos, vale destacar que:

- no experimento 2 os dados dos atributos selecionados foram submetidos apenas a um processo de limpeza, para ajustes de dados inconsistentes. Foram mantidos da forma como extraídos, inclusive quanto ao seu formato;
- nos experimentos 1 e 3 foram adotados completamente a metodologia proposta neste trabalho. Basicamente, além dos procedimentos adotados no experimento 2, os dados dos atributos selecionados foram tratados e transformados em consonância com os princípios de buscas de associações do algoritmo Apriori. Buscou-se, principalmente, agrupar os itens em grupos ou classes com o uso do algoritmo específico de clusterização K-Means, quando possível.

Nesse sentido, é possível realizar uma comparação entre os três experimentos realizados, sob o foco dos resultados alcançados, com o objetivo de analisar o desempenho do algoritmo Apriori em cada um deles.

No experimento 2, o algoritmo Apriori considerou apenas 6 dos 16 atributos submetidos ao processo de mineração de dados na geração de regras de

associação, enquanto no experimento 1, dos 18 atributos submetidos ao processo, 12 foram referenciados nas regras de associação encontradas, enquanto que no experimento 3, 14 dos mesmos 18 atributos foram envolvidos. Isso significa que nos experimentos 1 e 3 os dados dos atributos estavam em melhores condições de alcançar os requisitos mínimos estabelecidos no processo. Vale destacar que esse fator contribui para o aumento das chances de serem geradas maior quantidade de regras de associação e, consequentemente, descoberta de mais conhecimento.

Quanto ao número de regras geradas pelo algoritmo Apriori, foram geradas no experimento 2 apenas 10, enquanto no experimento 1, 109 e no experimento 3, 92. Embora seja um indicador meramente quantitativo, essa grande diferença já demonstra a dimensão do impacto dos procedimentos adicionais adotados nos experimentos 1 e 3 no desempenho do algoritmo Apriori.

Pode-se dizer que um dos critérios mais importantes na avaliação de um algoritmo de *data mining* é a sua capacidade de descoberta de conhecimentos. No caso do algoritmo Apriori, devem ser consideradas as conclusões possíveis que podem ser extraídas de uma análise das regras de associação geradas pelo método. Essa análise ocorre na etapa de pós-processamento.

Ao avaliar as conclusões relacionadas à interpretação dos resultados alcançados em ambos os experimentos, verifica-se que no experimento 2 apenas uma informação sobre possíveis relações entre os dados foi efetivamente extraída após a análise das regras geradas, enquanto que no experimento 1, observa-se pelo menos 13 informações bem interessantes sobre possíveis relações entre vários dados. Da mesma forma, no experimento 3, mais de 15 informações puderam ser extraídas.

Vale destacar que os experimentos 1 e 3 se diferem apenas em relação ao período das solicitações de medicamentos que formam o conjunto de dados, ou seja, no experimento 1 todos os dados disponíveis sobre as solicitações de medicamentos foram processados, enquanto no experimento 3, apenas as solicitações realizadas após o ano de 2006 foram analisadas pelo algoritmo Apriori.

Pelas melhorias realizadas no sistema de informação após esse período, esperava-se que os dados de solicitações após o ano de 2006 tivessem maior qualidade. Assim, ao comparar os experimentos 1 e 3 verifica-se que praticamente as mesmas regras foram confirmadas, sendo que apenas algumas do experimento 1 não apareceram no experimento 3.

No entanto, no experimento 3, mais atributos foram referenciados, especialmente os relacionados ao peso e o IMC do paciente, que passaram a apresentar uma menor ausência de dados no conjunto de dados analisado. Consequentemente, mais informações puderam ser extraídas em relação a esses atributos.

Chama a atenção o fato de que, ao analisar o contexto da dispensação de medicamentos excepcionais, alguns atributos da base de dados, que podem ser considerados muito importantes e essenciais para descobertas de conhecimentos interessantes para a área de saúde, ainda apresentam taxas de ausência de dados muito elevadas, como são os casos das informações sobre peso, altura, raça/cor e naturalidade do paciente. Vale destacar que há atributos com outras informações na base de dados que não foram selecionados nos processos experimentados neste trabalho por apresentarem taxas de ausência de dados ainda maiores, mas que também seriam informações muito impactantes nesse contexto, como: horários das ocorrências dos eventos; “status” das tarefas e análises realizadas; informações adicionais do paciente como quantidade de transplantes, se é gestante, se tem hemofilia ou se usa algum inibidor.

Assim como foi constatado pelo experimento 3, em que as taxas de ausência de dados diminuiu e mais conhecimentos foram descobertos, é muito provável que os resultados seriam ainda melhores e que haveria mais descobertas de conhecimento se essas taxas de ausência de dados fossem ainda menores. Vale ressaltar que foi observado que essas informações estão mais presentes e mais consistentes nos registros mais recentes, o que denota uma maior chance para descobertas de informações de forma automatizada. Isso indica que se o sistema informatizado e os procedimentos da unidade forem ainda mais aprimorados, com uma maior presença de dados e uma maior riqueza de detalhes das informações sobre as dispensações, melhores serão os resultados nas próximas tentativas de realização de processos de mineração de dados.

9- CONCLUSÕES E CONSIDERAÇÕES FINAIS

Pelo estudo de caso se pode concluir, inicialmente, que é possível extrair informações de uma base de dados, por meio de um processo de mineração de dados com o uso do algoritmo Apriori. Como foi relatado, é possível aplicar o processo com a utilização softwares livres, como a ferramenta de *data mining* WEKA e o banco de dados MySQL, entre outros. Assim, o uso de técnicas de mineração de dados está mais acessível.

Em casos específicos de manipulação de grandes bases de dados, possíveis limitações do ambiente computacional utilizado no processo podem influenciar na capacidade e tempo de processamento das técnicas de mineração de dados. Nesse sentido, um maior investimento em equipamentos pode ser necessário.

O algoritmo Apriori alcança resultados satisfatórios na tarefa de busca por associações; no entanto, não é recomendável que os atributos selecionados para o processo apresentem muitos itens distintos a ponto de terem baixas frequências, situação comumente encontrada em atributos do tipo data, variáveis numéricas e textos livres ou não padronizados.

Assim, os dados dos atributos nessa situação precisam ser mais bem trabalhados para evitar que eles sejam desconsiderados pelo método de busca de associações do algoritmo Apriori. Não se deve simplesmente escolher um grupo de atributos de uma base de dados, ainda que limpa, para que seus dados sejam submetidos à aplicação do algoritmo Apriori. A preparação dos dados deve ser direcionada e específica para essa aplicação.

O objetivo desse trabalho era identificar as principais limitações do algoritmo Apriori, com as respectivas orientações para que sejam superadas, e avaliar a influência de uma preparação dos dados disponíveis, diante da definição dos parâmetros e critérios mínimos para a realização da busca de associações pelo algoritmo, nos resultados encontrados. Outra questão importante que se buscava analisar é se seria viável o uso do algoritmo K-Means, que é um método que pode ser utilizado em tarefas de clusterização, na etapa de pré-processamento dos dados

e se a sua utilização contribuiria para o aprimoramento das buscas realizadas pelo algoritmo Apriori.

Pela comparação dos resultados alcançados nos três experimentos é possível concluir que a aplicação da metodologia proposta, que também sugere o uso do algoritmo K-Means na etapa de pré-processamento dos dados, melhorou o desempenho do algoritmo Apriori, ao se observar tanto aspectos quantitativos quanto qualitativos das regras de associação geradas.

Vale destacar que esse trabalho se limitou a avaliar as regras de associação geradas pelo algoritmo Apriori. O foco foi direcionado para a análise do desempenho do algoritmo com base na coerência e possibilidade para as regras serem entendidas, interpretadas e transformadas em conhecimento. Nesse sentido, não foi objeto desse trabalho a confirmação ou investigação mais profunda das regras para verificar sua aplicação prática, embora as regras tenham sido confirmadas e tidas como verdadeiras junto a técnicos especializados.

As limitações dessa dissertação não permitem afirmar se as conclusões do autor sobre a utilização da metodologia utilizada na base de dados da unidade de saúde estudada valem também para outras bases de dados similares.

Nesse sentido, como sugestão para trabalhos futuros, alguns aspectos relacionados à busca automatizada de associações, como os descritos a seguir, ainda podem ser estudados e explorados.

Embora o K-Means, na sua versão original, tenha identificado adequadamente os grupos nos atributos que foram submetidos ao método Apriori, conforme a metodologia sugerida, vale ressaltar que seu uso é limitado aos atributos com dados numéricos. Dessa forma, os dados precisam ser originalmente numéricos ou serem trabalhados para serem representados nesse formato. Também existem outros métodos alternativos capazes de realizar a tarefa de clusterização de dados não numéricos, inclusive, há métodos construídos com variações do próprio algoritmo K-Means que podem ser avaliados.

Sugere-se que a metodologia seja aplicada em outras bases de dados, com dados da área da saúde ou não, já que foi observado que a estudada apresenta deficiências de preenchimento de dados, o que possivelmente prejudicou os resultados encontrados. Por outro lado, pela análise dos dados dos registros mais recentes, há a expectativa de que o preenchimento dos dados da base em estudo esteja sendo aprimorado, o que motiva uma nova aplicação da metodologia em um

momento futuro para que os resultados sejam analisados. Também seria muito interessante avaliar os resultados decorrentes de experimentos em que sejam definidos diferentes níveis mínimos de suporte e de confiança.

Outros algoritmos foram construídos com base nos princípios do método Apriori e também podem ser testados juntamente com o uso da metodologia.

Vale destacar que as experiências, os resultados e as conclusões desse trabalho também foram abordados por Almeida *et al.* (2010) e discutidos no XVII Simpósio de Engenharia da Produção (SIMPEP), realizado pela Universidade Estadual Paulista (UNESP). Nele os autores demonstram que com a aplicação do processo de mineração de dados com o uso do algoritmo Apriori, após uma melhor preparação dos dados através da metodologia por eles proposta, obtem-se resultados satisfatórios (Almeida *et al.*, 2010).

REFERÊNCIAS

ADRIAANS, Pieter; ZANTINGE, Dolf. Data Mining. Harlow: Addison-Wesley, 1996.

ALMEIDA, Derciley C.; BRITO, Leonardo. C.; GUEDES, Leonardo. G. R.; CRUZ JUNIOR, Gelson. Proposta de Metodologia para Descoberta Automatizada de Associações em Bases de Dados, aplicando o processo de Mineração de Dados: Um Estudo de Caso em uma unidade pública de saúde.In: SIMPÓSIO DE ENGENHARIA DA PRODUÇÃO, 7., 2010, Bauru. Anais... Bauru: UNESP, 2010, Disponível em http://www.simpep.feb.unesp.br/abrir_arquivo_pdf.php?tipo=artigo&evento=5&art=940&cad=10557&opcao=com_id. Acesso em dezembro de 2010.

ANDRADE, Maria Margarida. Introdução à Metodologia do Trabalho Científico. 6. ed. São Paulo: Atlas, 2003.

BARROS, Aidil Jesus da Silveira. Fundamentos de Metodologia Científica. 2. ed. São Paulo: Pearson Makron Books, 2004.

BERRY, Michael J. A.; LINOFF, Gordon. Data Mining: Techniques for Marketing, Sales, and Customer Suport. New York: Willey, 1997.

BERRY, Michael J. A.; LINOFF, Gordon. Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: Willey, 2000.

BERSON, Alex; SMITH, Stephen; THEARLING, Kurt. Building Data Mining Applications for CRM. New York: McGraw-Hill, 2000.

BIGUS, Joseph P. Data Mining with neural networks: solving business problems – from application development to decision support. Crawfordsville: McGraw-Hill, 1996.

BRASIL, Constituição da República Federativa do Brasil: promulgada em 5 de outubro de 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituiçao.htm. Acesso em 11 de novembro de 2009.

BRASIL, Portaria Nº 2.577/GM, de 27 de outubro de 2006. Aprova o Componente de Medicamentos de Dispensação Excepcional. Disponível em: http://dtr2001.saude.gov.br/sas/PORTARIAS/Port2006/GM/GM-2577.htm. Acesso em 12 de novembro de 2009.

CABENA, Peter; HADJINIAN, Pablo; STADLER, Rolf; VERHESS, Jaap; ZANASI, Alessandro. Discovering Data Mining: from Concept to Implementation. Upper Saddle River: Prentice Hall, 1998.

CARVALHO, Maria Cecília M. (org.). Construindo o Saber: Metodologia Científica – Fundamentos e Técnicas. 18. ed. Campinas: Papirus, 2007.

ELMASRI, Ramez; NAVATHE, Shamkant B. Sistemas de Banco de Dados. São Paulo: Pearson Addison Wesley, 2005.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic; UTHURUSAMY, Ramasamy. Advances in Knowledge Discovery and Data Mining. Menlo Park: American Association for Artificial Intelligence, 1996.

GARBER, Rogério. Inteligência Competitiva de Mercado. São Paulo: Madras, 2001.

HAN, Jiawei; KAMBER, Micheline. Data Mining: Concepts and Techniques. 2. ed. San Francisco: Morgan Kaufmann Publishers, 2006.

KADOUS, Mohammed W. Temporal Classification: extending the classification paradigm to multivariate time series. Sydney, 2002. Thesis (Degree of Doctor of Philosophy) - The University of New South Wales, School of Computer Science and Engineering. Disponível em http://www.cse.unsw.edu.au/~waleed/phd/phd.pdf. Acesso em abril de 2010.

KANTARDZIC, Mehmed M.; ZURADA, Jozef. Next Generation of Data-Mining Applications. New Jersey: Wiley-Interscience, 2005.

KOCHE, José Carlos. Fundamentos de Metodologia Científica: Teoria da Ciência e Iniciação à Pesquisa. 20. ed. Petrópolis: Vozes, 2002.

KOGAN, Jacob; NICHOLAS, Charles; TEBOULLE Marc. Grouping Multidimensional Data: Recent Advances in Clustering. New York: Springer, 2006.

KUDYBA, Stephan; HOPTROFF, Richard. Data Mining and Business Intelligence: A guide to productivity. Hershey: Idea Group Publishing, 2001.

LAROSE, Daniel T. Discovering knowledge in data: an introduction to data mining. Hoboken: Wiley-Interscience, 2005.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. Fundamentos de Metodologia Científica. 6. ed. São Paulo: Atlas, 2007.

MARTINS, Gilberto de Andrade. Manual para Elaboração de Monografias e Dissertações. 3. ed. São Paulo: Atlas, 2002.

PASSOS, Emmanuel; GOLDSCHMIDT, Ronaldo. Data Mining: um guia prático. Rio de Janeiro: Elsevier, 2005.

PEDRYCZ, Witold. Knowledge-based clustering: from data to information granules. Hoboken: Wiley, 2005.

REZENDE, Solange O. (coord.). Sistemas Inteligentes: fundamentos e aplicações. Barueri: Manole, 2005.

RUDIO, Franz Victor. Introdução ao Projeto de Pesquisa Científica. 32. ed. Petrópolis: Vozes, 2004.

RUIZ, João Álvaro. Metodologia Científica: Guia para Eficiência nos Estudos. São Paulo: Atlas, 2002.

SEVERINO, Antônio Joaquim. Metodologia do Trabalho Científico. São Paulo: Cortez, 2007.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. Sistema de Banco de Dados. 5. ed. Rio de Janeiro: Elsevier, 2006.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introdução ao Data Mining: Mineração de Dados. Rio de Janeiro: Ciência Moderna, 2009.

TANG, Zhaohui; MACLENNAN, Jamie. Data Mining with SQL Server 2005. Indianapolis: Wiley, 2005.

WESTPHAL, Christopher; BLAXTON, Teresa. Data Mining Solutions: Methods and Tools for Solving Real-World Problems. New York: Wiley, 1998.

Wikipédia, a enciclopédia livre. Armazém de Dados. Disponível em http://pt.wikipedia.org/wiki/Armazém_de_dados. Acesso em fevereiro de 2011.

Wikipédia, a enciclopédia livre. Clustering. Disponível em http://pt.wikipedia.org/wiki/Clustering. Acesso em fevereiro de 2011.

Wikipédia, a enciclopédia livre. Índice de massa corporal. Disponível em http://pt.wikipedia.org/wiki/Índice_de_massa_corporal. Acesso em fevereiro de 2011.

Wikipédia, a enciclopédia livre. Máquina virtual Java. Disponível em http://pt.wikipedia.org/wiki/Máquina_virtual_Java. Acesso em abril de 2010.

Wikipedia, the free encyclopedia. Apriori Algorithm. Disponível em http://en.wikipedia.org/wiki/Apriori_algorithm. Acesso em abril de 2010.

WITTEN, Ian H.; EIBE, Frank. Data Mining: practical machine learning tools and techniques. 2. ed. San Francisco: Elsevier, 2005.