Aplicação de Mineração de Dados em uma Base Odontológica

Gustavo Guilherme de Alcenio Cunha, Carlos Rodrigo Dias

Bacharelado em Sistema de Informações - Faculdade Metodista Granbery (FMG)
Rua Batista de Oliveira, 1145 - 36010532 - Juiz de Fora - MG
(ggacunha@gmail.com, crdias@granbery.edu.br)

Resumo. Um dos problemas da atualidade é o que fazer com o grande volume de dados armazenados pelas organizações. A evolução da Tecnologia da Informação e o desenvolvimento de algoritmos de mineração de dados levaram estas organizações a investir na implantação de processos de KDD (Descoberta de Conhecimento em Base de Dados). O objetivo deste artigo é apresentar um estudo de caso, envolvendo um processo de KDD aplicado sobre a base de dados de uma clínica odontológica, seguindo o modelo de referência CRISP-DM. A essência deste trabalho está voltada para a análise dos aspectos mais relevantes, existentes nas etapas de um processo desta natureza.

Palavras Chave: dados, banco de dados, descoberta do conhecimento em base de dados (KDD), algoritmos de mineração de dados, modelo de referência CRISP-DM.

Abstract. *One of the problems in the present days is what to do with the great amount of data stored by the organizations. The evolution of Information Technology and the development of data mining algorithms force these organizations to invest on the development of KDD (Discovery of Knowledge in Databases) processes. The aim of this paper is to present an actual case involving the execution of the KDD process on the database of odontological issues, using the CRISP-DM reference model. The essence of this work is the analysis of the most important aspects to be considered about the steps of a KDD process.*

Key-Words: *data, data base, discovery of knowledge in databases, data mining algorithms, CRISP-DM reference model.*

1. INTRODUÇÃO

Nas últimas décadas, a evolução tecnológica contribuiu para o aumento da capacidade das organizações de gerar, captar, processar e armazenar informações em base de dados, acarretando um grande aumento no volume de dados armazenados. Com isso, percebeu-se a necessidade do desenvolvimento de novas técnicas e ferramentas computacionais que permitissem a utilização, de forma inteligente e, se possível automática, de informações obtidas a partir da análise de todo o volume de dados disponível em uma organização.

Como conseqüência disso, na área de Tecnologia da Informação surge uma oportunidade para o campo de pesquisa de extração de informações a partir de bases de dados, denominado Descoberta de Conhecimento em Base de Dados (KDD - *Knowledge Discorery in Databases).* Para Fayyad, Piatetsky e Smyth (1996), a descoberta de conhecimento em base de dados ocorre por meio de um processo dividido em várias etapas, orientado e executado por um analista humano conhecedor do negócio envolvido, que utiliza recursos computacionais para a busca de padrões potencialmente úteis e interpretáveis a partir das grandes bases de dados. As etapas desse processo são repetidas parcialmente ou integralmente, tantas vezes quantas se fizerem necessárias para se chegar a um resultado desejado, levando-se em conta a complexidade existente na sua execução. Para tanto, é necessária a utilização de uma metodologia eficiente e eficaz para o cumprimento desses requisitos operacionais, tornando o processo de KDD confiável e gerenciável. Em 1989, devido à grande quantidade de conceitos para a definição de Descoberta de Conhecimento em bases de dados, o termo KDD foi formalizado.

A prática da Odontologia produz uma grande quantidade de dados, e apenas disponibilizá-las da forma como estão armazenadas não é suficiente para que se tenha um bom aproveitamento das mesmas. É necessário conhecer as informações úteis e não triviais escondidas neste tipo de bancos de dados. Esses conhecimentos auxiliarão no desenvolvimento de estratégias que poderão contribuir para tomada de decisões, tanto no setor administrativo e financeiro quanto nos tratamentos clínicos de pacientes, além da expansão de conhecimentos acerca desta área da saúde.

Esse artigo apresenta a aplicação da metodologia CRISP-DM para a extração de conhecimento a partir de uma base de dados odontológica e está organizado da seguinte forma: na Seção 2 é descrita a metodologia de implantação CRISP-DM, na Seção 3 é descrita a base de dados a ser minerada e, finalmente, nas Seções 4, 5 e 6 é realizado o relato do caso de uso proposto, enumerando todos os passos da implantação, além das dificuldades, das descobertas e das conclusões obtidas.

2. METODOLOGIA CRISP-DM

Para a implantação do processo de KDD é necessária uma metodologia eficiente e eficaz para alcançar os objetivos, tornando o processo mais confiável e gerenciável. Atualmente, diversas metodologias podem ser aplicadas em projetos dessa natureza, a fim de aumentar suas probabilidades de sucesso.

A metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM) foi concebida em 1996 pelo consórcio composto por NCR Systems Engineering Copenhagen, DaimlerChrysler, SPSS e OHRA Verzekeringen en Bank Groep B.V, podendo ser descrita como um modelo envolvendo um processo hierárquico, sendo que este modelo compreende um conjunto de tarefas descritas em quatro níveis de abstração. Cada nível de abstração é composto por tarefas que vão da forma mais genérica até a forma mais específica (CHAPMAN *et al.*, 2000). O primeiro nível de abstração compreende as seis tarefas genéricas do ciclo de vida do processo de KDD, ou seja, a compreensão do negócio, compreensão dos dados, preparação dos dados, mineração de dados, avaliação e aplicação dos resultados. O segundo nível de abstração é descrito por tarefas menos genéricas, oriundas das tarefas do primeiro nível. As tarefas do segundo nível de abstração envolvem todas a possíveis situações de um processo de KDD. Como exemplo, a limpeza dos dados, dentro da fase de preparação dos dados. Já o terceiro nível é composto de tarefas específicas que descrevem as ações citadas no segundo nível de forma mais abrangente, para serem aplicadas em situações específicas, como, por exemplo, a transformação de um determinado dado numérico em categórico. No quarto e último nível é feita a instanciação dos processos, realizando os registros das ações, decisões e resultados do referido processo de KDD.

A fase de compreensão do negócio visa o entendimento dos objetivos e metas do KDD, avaliando os recursos, requisitos, suposições e limitações observadas, além dos custos com o processo e benefícios a serem alcançados. De posse dessas informações, obtém-se uma definição de problema de mineração de dados e desenvolve-se um projeto preliminar para alcançar os objetivos definidos. O objetivo da fase de compreensão dos dados está no entendimento dos dados, na qual especifica-se a coleta inicial de dados e os métodos utilizados para a coleta. A seguir faz-se a descrição dos dados coletados, devendo ser registrada a descrição individual de cada dado ou conjunto de dados. Executa-se também a identificação de problemas de qualidade com os dados. Assim, são executadas as atividades de familiarização com os dados da base, de forma a caracterizá-los pelas suas formas de representação.

A fase de preparação dos dados dispõe de atividades específicas que geram o banco de dados a ser minerado, a partir dos dados originais da base de dados operacional. Os dados da nova base de dados serão utilizados na ferramenta de mineração de dados. Para tanto, deve-se executar as tarefas de seleção de dados, de limpeza dos dados selecionados, construção de novos dados, formatação de dados existentes e, por último, da integração de novos dados a partir de fontes diferentes.

Na fase de mineração de dados são realizadas a seleção e a aplicação de técnicas de mineração de dados para extração de conhecimento. Os parâmetros dos algoritmos são definidos de forma a alcançar os objetivos do problema. Geralmente, em um mesmo problema de KDD, podem ser utilizadas inúmeras técnicas de mineração de dados, além da associação entre as mesmas. Normalmente, devido às características de cada técnica de mineração, o formato dos dados deve seguir exigências específicas. Dessa forma, a iteração entre as fases anteriores se faz necessária para o refinamento dos dados. Os resultados obtidos durante esta etapa podem ser combinados em um único relatório que descreverá como os modelos são construídos, testados e avaliados.

O objetivo da etapa de avaliação, proposta por CRISP-DM é a avaliação geral do processo de KDD implantado, pois se espera a detecção de algum aspecto relevante do negócio que não foi considerado adequadamente. Deve-se também rever todos os passos de cada fase, a fim de garantir um modelo de conhecimento eficiente, segundo uma perspectiva de análise de dados. Então, é realizada a avaliação final do modelos de conhecimento considerados relevantes na etapa anterior a esta. Caso os objetivos tenham sido alcançados, deve-se seguir para a última fase deste processo, para que sejam tomadas as decisões a partir do uso dos resultados obtidos.

Na fase de implantação desenvolve-se um plano de aplicação e um plano de monitoramento e manutenção para o plano de aplicação, o qual deverá ser executado, devendo ser registrada em relatório a avaliação da aplicação dos resultados. Por último, é revisado todo o processo de KDD, registrando-se, na documentação própria, a experiência adquirida.

3. BASE DE DADOS ODONTOLÓGICA

A Clínica Odontológica utilizada neste trabalho existe há aproximadamente dez anos, sendo especializada em tratamentos nas áreas de Dentística, Periodontia, Implantodontia, Ortodontia, Exodontia e Cirurgia Oral Menor. Há três anos a clínica adquiriu um sistema

informatizado de controle odontológico que realiza o gerenciamento administrativo e operacional das atividades realizadas. Esse sistema encontra-se atualizado na sexta versão, com uma interface gráfica baseada em um odontograma gráfico que registra as intervenções realizadas nos dentes e suas faces, simulando a figura de uma arcada dentária humana. O sistema utiliza como repositório um banco de dados relacional Microsoft Access 2000, que será referenciado neste trabalho como BDO60.

De acordo com estudos realizados junto à base de dados BDO60 e ao material bibliográfico do sistema odontológico, ficou definido que o processo de KDD será voltado para tratar o perfil dos pacientes a partir das informações armazenadas. Assim, considerar-se-á como informação relevante aquela cujo conteúdo se referir às características pessoais e clínicas dos pacientes e dos procedimentos aplicados em seus tratamentos. Por essa razão, exclui-se um estudo voltado para área comercial ou administrativa da clínica.

Foram identificadas 81 (oitenta e uma) tabelas de dados. Desse total, vale ressaltar que 10 (dez) tabelas continham registros com informações relevantes para o processo de KDD. Essas informações estão focadas diretamente no tratamento do paciente. As dez tabelas com informações relevantes para o processo de KDD estão relacionadas a seguir, com as respectivas descrições:

a) tabela *PESSOAL*: armazena as informações pessoais dos pacientes, com 709 (setessentos e nove) pacientes cadastrados;
b) tabela *ANAMNESE_QUEST*: armazena os nomes dos principais questionários aplicados nos pacientes relativos ao seu estado de saúde. A tabela possui 3 (três) tipos de questionários cadastrados;
c) tabela *ANAMNESE_PERG*: armazena as perguntas dos questionários feitos ao paciente em relação ao seu estado de saúde. Possui 71 (setenta e uma) perguntas cadastradas;
d) tabela *ANAMNESE_RESP*: armazena as respostas do paciente relativas aos questionários sobre o seu estado de saúde. O questionário é respondido antes de iniciado o tratamento, e possui 608 (seiscentas e oito) respostas cadastradas;
e) tabela *TRATAMENTO*: armazena informações de controle do tratamento que o paciente se submeterá. De forma geral, um tratamento pode envolver vários procedimentos ou intervenções odontológicas. A tabela possui 1.514 (um mil, quinhentos e quatorze) tratamentos cadastrados;
f) tabela *ARCADA*: armazena a marcação das características e anomalias encontradas, além das intervenções já presentes na arcada dentária do paciente, também denominada "condição observada". Os registros dessa tabela são carregados no início do

tratamento. Atualmente, a tabela possui 48.448 (quarenta e oito mil, quatrocentas e quarenta e oito) arcadas cadastradas;

g) tabela *INTERVENÇÃO*: armazena informações de controle das intervenções ou procedimentos, que são realizadas em um tratamento odontológico, possuindo 5.537 (cinco mil, quinhentas e trinta e sete) intervenções cadastradas;
h) tabela *HISTÓRICO*: conhecido como prontuário do paciente, armazena dados clínicos dos pacientes, procedimentos, intervenções na boca e nos dentes, além de outras informações que o especialista julgar necessário. As intervenções cadastradas no odontograma, ao serem finalizadas, são automaticamente registradas nessa tabela, que é também denominada "diário" e possui 7.638 (sete mil, seiscentos e trinta e oito) históricos cadastrados;
i) tabela *DENTE*: armazena informações do dente que sofrerá uma determinada intervenção, dentro de um tratamento. A tabela possui 10.205 (dez mil, duzentos e cinco) registros;
j) tabela *FACE*: armazena o identificador de uma das cinco faces de um determinado dente que será submetido a uma determinada intervenção, dentro de um tratamento. A tabela contém 1.259 (mil, duzentas e cinqüenta e nove) faces cadastradas.

4. COMPREENSÃO DO NEGÓCIO E DOS DADOS

O objetivo do KDD considerado nesse processo é manter o bom desempenho profissional realizado na clínica odontológica. Esse desempenho é medido pela qualidade dos resultados obtidos nos tratamentos realizados nas bocas dos pacientes. Para tanto, além de capacitação profissional e experiência prática, é importante conhecer o perfil clínico dos pacientes e as tendências dos tratamentos bucais em suas especialidades.

O especialista do domínio da aplicação, neste caso o cirurgião dentista, por meio da implantação do processo de KDD, visa agregar conhecimentos pela identificação de padrões de ocorrências encontrados nos tratamentos realizados na clínica, a fim de poder medir e planejar os investimentos com especialização profissional. Considerar-se-á um bom resultado da aplicação do processo de KDD sobre a base de dados da clínica odontológica, a obtenção de qualquer resultado capaz de informar o perfil clínico dos pacientes e seus tratamentos, além de contribuir para predição de novos casos para tratamento.

Ao realizar a avaliação da situação, proposta pelo CRISP-DM, identificando os recursos disponíveis e os requisitos exigidos para a execução do KDD, constatou-se que

estavam envolvidos nesse processo quatro profissionais, sendo eles: o cirurgião dentista, pós-graduado em várias áreas da odontologia, e que detém o domínio da aplicação, sendo profundo conhecedor do negócio em que atua; a secretária, o profissional que auxilia diretamente o cirurgião dentista, tanto na parte operacional quanto na parte administrativa, sendo a operadora do sistema odontológico, tendo sido designada para contribuir administrativamente nesse processo; o especialista em KDD, graduando em Bacharelado de Sistemas de Informação, autor desse trabalho e responsável pela execução do KDD na clínica; e, por último, o professor e orientador do graduando, docente da Faculdade Metodista Granbery, especialista em processos de KDD.

Devido ao pouco tempo de informatização na clínica, a base de dados ainda não possui todas as informações dos pacientes no sistema, podendo citar como exemplo, a falta do campo da data de nascimento de alguns pacientes cadastrados. Assim, as fichas clínicas, manuscritas anteriormente à informatização, auxiliarão para completar os dados que faltam para o processo de KDD.

A base de dados não contém todos os tratamentos referentes à Ortodontia, uma vez que ainda não houve a implantação total dessa especialidade ao sistema. Portanto, essa especialidade odontológica não será abordada neste processo de KDD.

A ferramenta de KDD utilizada neste trabalho foi o sistema Weka, pois se trata de uma ferramenta de código aberto, flexível na sua manipulação, e que executa as tarefas de mineração de dados necessárias para o cumprimento deste trabalho. O Weka foi desenvolvido na linguagem de programação Java pelo curso de Ciências da Computação, na universidade de Waikato, Nova Zelândia. A interface do Weka se mostra dividida em quatro partes, as quais permitem que seus algoritmos sejam carregados diretamente. Essa ferramenta facilitou a migração dos dados da base de dados BDO60, pois suporta a abertura de arquivos com extensões *.ARF* e *.CSV*. O Weka permite a visualização dos dados de forma gráfica, utilizando histogramas.

A coleta inicial dos dados ocorreu por meio da criação de uma nova base de dados que somente recebeu as tabelas PESSOAL, TRATAMENTO, INTERVENÇÃO, e HISTÓRICO, contendo todos os registros e relacionamentos entre si. Inicialmente, foram considerados 15 (quinze) atributos, dispostos nessas tabelas.

Na tabela PESSOAL somente importarão os atributos sexo e a data de nascimento, pois deles poderão ser extraídos padrões que demonstrem as tendências do comportamento clínico. As tabelas TRATAMENTO e INTERVENÇÃO foram selecionadas pois o atributo *status*, em ambas tabelas, são responsáveis por definir se um tratamento ou

intervenção está aberto, finalizado ou interrompido. Somente serão considerados os tratamentos e intervenções que estiverem finalizados. A tabela HISTÓRICO foi selecionada, pois representa o prontuário clínico do qual serão extraídos os dados que identificam o paciente, o dente ou o grupo de dentes tratados, além das especialidades da odontologia que foram aplicadas no tratamento, por meio dos registros das intervenções ou procedimentos adotados. A data da intervenção registrada nessa tabela será útil para identificar a idade do paciente na época do tratamento.

De um total de 709 pacientes cadastrados na tabela PESSOAL, inicialmente, foram selecionados para análise 378 pacientes, referenciando 7.638 procedimentos e intervenções nos prontuários (HISTÓRICO) desses pacientes. Foram selecionados todos os registros da tabela HISTÓRICO, estando nos seus dados o foco principal deste processo de KDD. Esses tratamentos ocorreram entre outubro de 2004 a maio de 2006, sendo esse o período estudado neste trabalho.

Posteriormente, os dados dessas tabelas foram convertidos para o formato **.***XLS*, da planilha Microsoft Excel. Esta conversão, além de facilitar a manipulação das informações, mediante a versatilidade da ferramenta Microsoft Excel, também auxiliará, ao final, para gerar o arquivo **.***CSV* utilizado pela ferramenta de mineração de dados Weka.

Ao fazer referência à qualidade dos dados, identificou-se inúmeras situações que deverão ser contornadas para a obtenção de êxito nas respostas esperadas. A seguir, serão explicitados os problemas e dificuldades observados nos dados com maior relevância para este processo de KDD:

a) a codificação utilizada para a identificação da primeira dentição é diferente dos dentes permanentes;
b) na tabela de HISTÓRICO, em um mesmo registro, no atributo "Região/Dente" tratado, pode ocorrer a identificação de um grupo de dentes que se dá por meio de separadores de intervalo. Os separadores são representados por "**-**" , "/" , "e" e "**,**". Também podem ser especificados todos os dentes com o texto "TODOS", bem como vir sem conteúdo algum;
c) na tabela de HISTÓRICO, a representação do atributo "Região/Dente" tratado obedece o sistema de identificação dentário da clínica, proposto na Figura 4.1. Onde o primeiro dígito significa o quadrante e o segundo, o número do dente por quadrante. Esta padronização é dada para a dentição permanente. Assim, em um intervalo entre os dentes 12 e 23, representados por 12/23 ou 12-23, seriam selecionados os dentes: 12, 11, 21, 22 e 23. Esta relação não ocorre entre os quadrantes superiores e os quadrantes inferiores;
d) na tabela de PESSOAL existem várias datas de nascimento em branco;

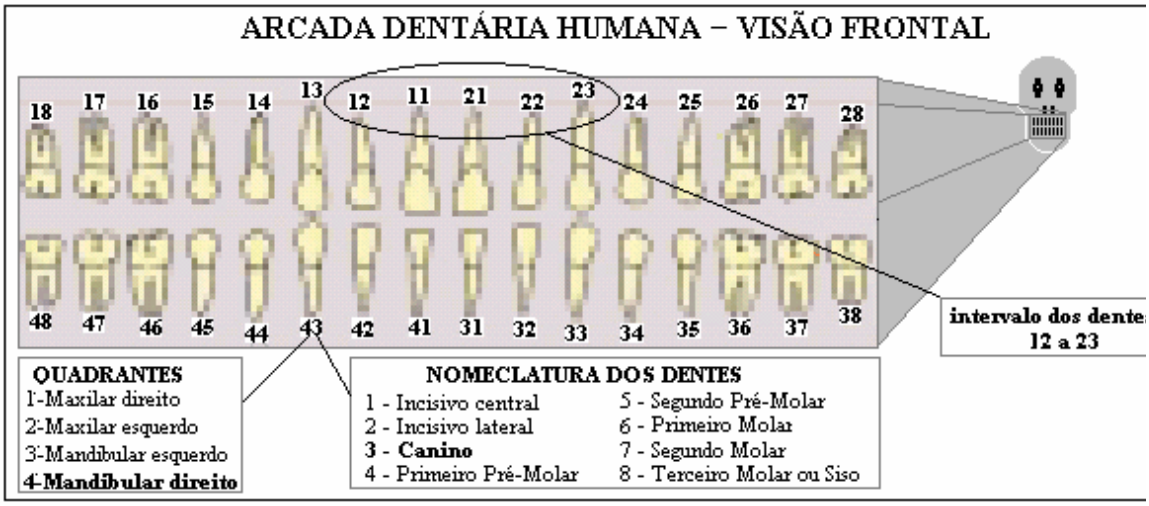

Figura 4.1: identificação técnica dos dentes de uma arcada dentária humana adulta.

e) há grande complexidade para a caracterização das especialidades aplicadas a um tratamento, registradas na tabela HISTÓRICO. Não há registro de qual especialidade atuou em um determinado tratamento. Essa informação deverá ser extraída dos dados referentes ao atributo "descrição da intervenção", devendo ser identificada por meio de análise em meio a inúmeros procedimentos auxiliares, a especialidade. A Tabela 4.1. mostra a relação de procedimentos e intervenções por especialidade e adotadas neste processo de KDD;

f) existem ruídos no atributo "data da intervenção", contendo valores irreais para o ano;

g) outra dificuldade encontrada está no fato de ser preciso identificar em grupos de registros de intervenções e procedimentos de um mesmo tratamento em um mesmo dente, quais registros devem ser preservados, pois representam o serviço principal que identifica a especialidade aplicada, e quais registros devem ser eliminados, pois representam um complemento do serviço executado pelo especialista do negócio;

h) se um tratamento aborda vários dentes, fazendo referência a um único dente por vez no atributo "Região/Dente", então ocorrerão tantos registros na tabela HISTORICO quanto o número de dentes tratado. Conforme a formatação do *dataset* a ser modelado, esses registros deverão ser suprimidos em um único registro que informará todos os dentes tratados da boca, dispostos em trinta e dois atributos;

i) foi observado que pacientes duplicados na tabela de PESSOAL geraram registros na tabela de HISTÓRICO como se fossem pacientes distintos;

j) observou-se que, no caso dos pacientes que já haviam realizado tratamentos anteriormente à informatização da clínica e, depois voltaram a realizar tratamentos, posteriormente à informatização, tiveram incluídos na tabela de HISTÓRICO, as intervenções manuscritas nas fichas de controle, referentes ao período anterior à informatização.

Tabela 4.1 : Procedimentos e Intervenções por especialidades adotadas no processo de KDD.

No	Procedimentos e Intervenções abordadas neste processo KDD	Especialidade
01	Restauração em resina composta	Dentística
02	Restauração em resina fotopolimerizável	Dentística
03	Restauração de superfície radicular	Dentística
04	Restauração em amálgama de prata	Dentística
05	Restauração de ionômero de vidro	Dentística
06	Colagem de fragmentos	Dentística
07	Enxerto ósseo	Periodontia
08	Enxerto gengival	Periodontia
09	Gengivectomia	Periodontia
10	Rizectomia	Periodontia
11	Frenectomia labial	Periodontia
12	Osteotomia e Osteoplastia	Cirurgia Oral Menor
13	Bridectomia	Cirurgia Oral Menor
14	Biopsia	Cirurgia Oral Menor
15	Alveoloplastia	Cirurgia Oral Menor
16	Exodontia de dente ou raiz	Cirurgia Oral Menor
17	Cirurgias de reconstrução	Implante e prótese
18	Implante ósteointegrável	Implante e prótese
19	Remoção bloco linha oblícua externa	Implante e prótese
20	Prótese fixa	Prótese
21	Prótese de porcelana removível	Prótese
22	Prótese total	Prótese
23	Prótese métalo cerâmica	Prótese
24	Núcleo métalo funfido	Prótese
25	Encaixes	Prótese
26	Coroa dentária	Prótese

5. PREPARAÇÃO E MINERAÇÃO DOS DADOS

A preparação dos dados foi realizada por intermédio da importação para planilha eletrônica, formatando o *dataset* a ser minerado para posterior aplicação de técnicas de mineração de dados para extração de conhecimento.

Identificou-se quais informações, armazenadas nas tabelas citadas na Seção 3, devem contribuir para efetivamente compor o *dataset*. Assim, considerar-se-á os seguintes atributos, com a respectiva justificativa para compor o *dataset* :

a) sexo: este atributo refere-se ao gênero da pessoa que gerou o fato, tendo a possibilidade de identificar quais especialidades odontológicas e quais dentes podem ser abordados devido à variação do gênero;
b) data de nascimento e data do tratamento: estes atributos foram selecionados, pois a partir deles serão calculados a idade do paciente na data do tratamento, além da faixa etária. Estes atributos podem ser importantes para avaliar o período da vida do paciente em que ocorreu o fato odontológico;

c) descrição da intervenção: este atributo armazena as informações relativas aos procedimentos adotados no tratamento, identificando-se a especialidade aplicada ao tratamento;
d) região/dente: o conteúdo deste atributo identificará os dentes envolvidos no tratamento, podendo ser classificados ou associados a outras características do paciente e dos tratamentos. Desse atributo serão originados os quatro atributos referentes aos quadrantes bucais.

Os atributos "números do paciente", "número do tratamento" e "número da intervenção", que representavam as chaves primárias e estrangeiras, somente foram úteis para a criação da nova base de dados a ser minerada. O atributo "data do tratamento" foi útil para definir, junto à "data de nascimento", a idade e a faixa etária do paciente na época do tratamento. Já os atributos de "*status"* auxiliaram efetivamente para definir quais registros deveriam fazer parte do novo grupo de dados.

A execução da atividade de limpeza dos dados tem como objetivo realizar a verificação da consistência das informações, realizando a correção de possíveis erros, além de preencher ou eliminar valores desconhecidos ou redundantes. Esse objetivo visa garantir a qualidade dos modelos do conhecimento a serem extraídos. A seguir são apresentadas as ações realizadas nos dados, na ordem em que foram executadas:

a) eliminou-se 339 instâncias do *dataset*, cujo conteúdo do atributo *status* do procedimento indicava uma intervenção em aberto ou interrompida;
b) eliminou-se 758 instâncias do *dataset,* cujo conteúdo do atributo "Região/dente" estava ausente, não contribuindo para distinguir os dentes tratados nos pacientes;
c) eliminou-se 58 instâncias do *dataset*, no caso em que o conteúdo do atributo "Região/dente" referia-se à primeira dentição, não sendo abordada neste trabalho;
d) eliminou-se 191 instâncias do *dataset* por fazerem referências a especialidades de Ortodontia que não são abordadas neste trabalho;
e) eliminou-se 1.372 instâncias do *dataset*, por fazerem referências a consultas para orçamento ou verificação de histórico ou planejamento de tratamento ou modelo de estudo, ou qualquer outro procedimento clínico burocrático que seja comum a todas as especialidades odontológicas abordadas neste trabalho;
f) eliminou-se 164 instâncias do *dataset*, pois referiam-se a procedimentos clínicos comuns a todas as especialidades, como, por exemplo, a utilização de Raio X. Esses registros foram eliminados por não indicarem tratamentos efetivos ao dentes;

g) eliminou-se 403 instâncias do *dataset*, por não distinguirem os dentes uns dos outros. O atributo "Região/dente" referenciava todos os 32 dentes da boca, por meio de procedimentos do tipo: aplicação de flúor, limpeza geral, jato de bicarbonato;
h) eliminou-se 1.123 instâncias do *dataset*, pois os procedimentos representavam o complemento de um serviço executado, por exemplo: curativo, remoção de sutura, ajustes de provisório;
i) eliminou-se 2.514 instâncias do *dataset*, pois esses registros foram suprimidos em registros únicos.
j) inclusão de 225 valores de datas de nascimento no *dataset*, devido à ausência de conteúdo. Esses valores foram copiados das fichas clínicas manuscritas;
k) correção de 25 datas que faziam referência ao período do tratamento, por indicarem o ano fora dos padrões da escrita.

A tarefa utilizada na etapa de construção de dados consiste em gerar novos atributos a partir de atributos existentes. Essa operação foi realizada, pois novos atributos, além de expressarem relacionamentos conhecidos entre atributos existentes, podem reduzir o conjunto de dados, simplificando o processamento da mineração de dados. Nessa etapa, as ações realizadas foram:

a) criação de quatro atributos: primeiro, segundo, terceiro e quarto quadrantes tratados, indicando qual quadrante da arcada dentária humana foi tratado. Esses atributos são derivados do atributo "Região/Dente";
b) criação de trinta e dois atributos que identificam cada um dos dentes existentes na arcada dentária humana. Enquanto um único dente era referência em um registro da tabela HISTÓRICO, no *dataset*, em um único registro, serão identificados os trinta e dois dentes;
c) criação do atributo faixa etária que qualifica como "jovem" pacientes com idade até 21 anos incompletos, "adulto" aqueles com idade compreendida entre 21 e 65 anos incompletos e, por último, "idosos" os indivíduos com idade de 65 anos ou mais.

A formatação dos dados é a tarefa responsável pela forma da representação desses dados durante a etapa de mineração de dados. A formatação utilizada nessa seção ocorre para atender às necessidades específicas das técnicas de mineração de dados. Assim, o tipo de dados do atributo "sexo" na base de dados original é numérico, na qual o zero indica sexo masculino e o número um indica o sexo feminino. Após a formatação esse atributo passou a ser categórico, fazendo referência às letras ´M´ ou ´F´, de masculino ou de feminino.

Ao final das tarefas realizadas nas etapas anteriores que beneficiaram os dados, realizando o pré-processamento necessário para garantir uma melhor qualidade das informações deste *dataset*, os atributos que compõem a nova base de dados a ser trabalhada nas próximas seções são: sexo, idade, faixa etária, especialidade, primeiro quadrante bucal, segundo quadrante bucal, terceiro quadrante bucal, quarto quadrante bucal, dente 11, dente 12, dente 13, dente 14, dente 15, dente 16, dente 17, dente 18, dente 21, dente 22, dente 23, dente 24, dente 25, dente 26, dente 27, dente 28, dente 31, dente 32, dente 33, dente 34, dente 35, dente 36, dente 37, dente 38, dente 41, dente 42, dente 43, dente 44, dente 45, dente 46, dente 47, dente 48. A denominação dos dentes e dos quadrantes bucais encontram-se na Figura 4.1.

De um total de 7.638 registros, 6.922 foram eliminados devido às situações registradas no início dessa seção. Assim, a nova base de dados chega na etapa de mineração de dados com um total de 716 instâncias e 40 atributos.

A partir de agora serão descritas as técnicas de mineração de dados utilizadas nessa fase do processo de KDD. Para atingir o objetivo deste KDD, pretende-se realizar as tarefas de associação e classificação para a extração de padrões a partir dos dados armazenados no *dataset*. Para tanto, são necessários algoritmos específicos para a implementação dessas tarefas de mineração de dados. Decidiu-se por utilizar o algoritmo Apriori para a tarefa de Associação e o C4.5 para a Classificação, ambos suportados pela ferramenta Weka..

O algoritmo Apriori é o algoritmo clássico na categoria de algoritmos de extração de regras de associação em mineração de dados, seguindo o princípio da antimonotonicidade do suporte, ou seja, um conjunto de itens é freqüente se todos os seus subconjuntos forem freqüentes. Em contrapartida, para algum subconjunto não freqüente, seu conjunto também não será freqüente (Agrawal, 1993, *apud* Goldschmidt e Passos, 2005).

O algoritmo Apriori utiliza como medida de qualidade, parâmetros de suporte mínimo, confiança mínima e *lift* mínimo, sendo que esses valores devem ser informados pelo usuário. Outra exigência do Apriori é suportar somente atributos nominais no seu contexto, necessitando da discretização de atributos numéricos. Na tela de pré-processamento, do Weka, entre os atributos do *dataset*, pode-se selecionar aqueles que representarão um conjunto de dados que serão analisados pelo Apriori, criando a possibilidade de extração de novas informações. Assim, a rotina de obtenção de regras de associação é dada, praticamente, nas tarefas de encontrar todos os conjuntos de itens freqüentes, satisfazendo o suporte mínimo e, posteriormente, obedecendo a confiança mínima, identificando as referidas regras em função dos conjuntos de itens freqüentes

O algoritmo C4.5, segundo Goldschmidt e Passos (2005), é um método de mineração de dados baseado na indução de árvores de decisão. O C4.5 realiza uma abordagem recursiva de particionamento no *dataset,* sendo conhecido também como algoritmo de aprendizado supervisionado. Na ferramenta Weka, esse algoritmo é identificado com J48.

O algoritmo C4.5 possui características específicas como tratar valores ausentes como um valor em separado, tratar ruídos nos dados por meio de poda na árvore de decisão, além de suportar atributos numéricos no seu contexto. Assim como o algoritmo Apriori, o C4.5 também trabalha os atributos selecionados no pré-processamento do Weka. Devido à tarefa de Classificação, há a exigência de ser definida a classe que norteará a análise para a extração de regras.

A metodologia de testes utilizada para avaliar os resultados dar-se-á por meio de análise e interpretação das respostas obtidas. É importante observar que a partir de uma primeira análise serão separados conjuntos de informações considerados relevantes, para serem comparados com novos conjuntos, obtidos por meio de novas execuções. Para tanto, serão utilizadas a revisão dos parâmetros utilizados e a seleção de novos atributos para o *dataset*.

Na ferramenta Weka executou-se o algoritmo Apriori sobre o *dataset* definido previamente. A definição dos atributos foi realizada, a fim de identificar informações relevantes em uma determinada segmentação de dados, por meio de generalização ou não desses dados. A seguir, são apresentados os subconjuntos de atributos que foram selecionados para serem analisados pelo Apriori :

1º. Grupo: sexo, idade, especialidade e os trinta e dois dentes.;
2º. Grupo: dentes 11,12,13,14,15,16,17,18;
3º. Grupo: dentes : 21,22,23,24,25,26,27,28;
4º. Grupo: dentes : 31,32,33,34,35,36,37,38;
5º. Grupo: dentes : 41,42,43,44,45,46,47,48;
6º. Grupo: sexo, idade, especialidade e os quatro quadrantes bucais.

Para cada subconjunto de dados foram definidos e testados valores para o parâmetro suporte mínimo que variaram entre 90% e 10%, com intervalos de 5% entre um experimento e outro. Em todos os experimentos foi utilizado o valor 1,0 como o limite inferior para a medida de interesse *lift* e 1500 como o número máximo de regras a serem extraídas. O parâmetro que limita o número de regras extraídas foi definido com um valor suficientemente grande para armazenar todas as possibilidades de extração gerada pelo Weka.

O atributo "idade" é originalmente do tipo numérico e, conforme as exigências do algoritmo Apriori houve a necessidade de discretizar seus dados. Foram realizados vários experimentos com criação de intervalos entre as idades, na tentativa de identificar aquele que melhor produzisse informações. O filtro utilizado pelo Weka foi o *weka.filters.unsupervised.attribute.Discretize*. A discretização que melhor produziu informações foi aquela que gerou três intervalos de valores: de 8 a 37,33 anos, de 37,33 a 66,66 anos e de 66,66 a 96 anos.

No primeiro subconjunto de dados são utilizados os atributos "sexo", "idade", "especialidade" e os trinta e dois dentes de uma boca, de um indivíduo qualquer que se tratou na clínica, e são mostradas, a seguir, algumas regras de associação extraídas :

- 22, 11 => 21 (*suporte* 10,19%, *confiança* 92% *e lift* 5,75);
- 11, 13 => 12 (*suporte* 11,17%, *confiança* 95% *e lift* 5,46);
- 22, 11 => 12 (*suporte* 10%, *confiança* 91% *e lift* 5,22);
- 22 => 21, 11 (*suporte* 10,19%, *confiança* 65% *e lift* 5,36);
- *PROTESE=> IDADE entre* 37.33 *e* 66.66 anos (*sup.* 12,43%, *confiança* 75% *e lift* 1,30);
- *IDADE entre* 37.33 *e* 66.66 *anos => PROTESE* (*sup.* 12,43%, *confiança* 22 % *e lift* 1,30);
- *DENTISTICA => IDADE* < 37.33 *anos* (*suporte* 17,73%, *confiança* 41% *e lift* 1,29);
- *IDADE* < 37,33 *anos => DENTISTICA* (*suporte* 17,73%, *confiança* 55% *e lift* 1,29).

No segundo subconjunto de dados são utilizados somente os dentes do primeiro quadrante bucal, maxilar direito: 11-incisivo central, 12-incisivo lateral , 13-canino, 14-primeiro pré molar, 15-segundo pré molar, 16-primeiro molar, 17-segundo molar e 18-terceiro molar:

- 11, 13 ==> 12 (*suporte* 11,17%, *confiança* 95% *e lift* 5,46);
- 12 ==> 11, 15 (*suporte* 10,19%, *confiança* 58% *e lift* 5,16);
- 13, 15 ==> 14 (*suporte* 10,33%, *confiança* 95% *e lift* 4,89);
- 14 ==> 13, 15 (*suporte* 10,33%, *confiança* 53% *e lift* 4,89);
- 13 ==> 11, 12 (*suporte* 11,17%, *confiança* 67% *e lift* 4,68);
- 12, 13 ==> 11 (*suporte* 11,17%, *confiança* 92% *e lift* 4,67);
- 11 ==> 12, 13 (*suporte* 11,17%, *confiança* 57% *e lift* 4,67).

No terceiro subconjunto de dados são utilizados somente os dentes do segundo quadrante bucal, maxilar esquerdo: 21-incisivo central, 22-incisivo lateral, 23-canino, 24-primeiro pré molar, 25-segundo pré molar, 26-primeiro molar, 27-segundo molar, 28-terceiro molar ou siso:

- 21 ==> 22 (*suporte* 11,31%, *confiança* 70% *e lift* 4,50);
- 22 ==> 23 (*suporte* 10,89%, *confiança* 70% *e lift* 4,34);
- 21 ==> 23 (*suporte* 10%, *confiança* 63% *e lift* 3,90);
- 23 ==> 24 (*suporte* 10%, *confiança* 63% *e lift* 3,45);
- 24 ==> 25 (*suporte* 12,710%, *confiança* 70% *e lift* 3,66);
- 26 ==> 27 (*suporte* 11,87%, *confiança* 59% *e lift* 3,43).

No quarto subconjunto de dados são utilizados somente os dentes do terceiro quadrante bucal, mandibular esquerdo: 31-incisivo central, 32-incisivo lateral, 33-canino, 34-primeiro pré-molar, 35-segundo pré-molar, 36-primeiro-molar, 37-segundo-molar, 38-terceiro-molar ou siso:

- 35 ==> 34 (*suporte* 10,47%, *confiança* 60% *e lift* 3,98);
- 34 ==> 35 (*suporte* 10,47%, *confiança* 70% *e lift* 3,98);
- 37 ==> 36 (*suporte* 10,75%, *confiança* 62% *e lift* 3,55).

No quinto subconjunto de dados são utilizados os dentes do quarto quadrante bucal, mandibular direito: 1-incisivo central, 42-incisivo lateral, 43-canino, 44-primeiro pré molar, 45-segundo pré molar, 46-primeiro molar, 47-segundo molar, 48-terceiro molar ou siso:

- 44 ==> 45 (*suporte* 10,61%, *confiança* 67% *e lift* 3,92);
- 45 ==> 47 (*suporte* 10%, *confiança* 59% *e lift* 3,04);
- 46 ==> 47 (*suporte* 12,43%, *confiança* 52% *e lift* 2,70).

No sexto subconjunto de dados são utilizados os atributos "sexo", "idade", "especialidade" e os quatro quadrantes bucais, extraindo-se as seguintes regras:

- 2o.*Quadrante* ==> 1o.*Quadrante* (*suporte* 30,72%, *confiança* 65% *e lift* 1,25);
- 1o.*Quadrante* ==> 2o.*Quadrante* (*suporte* 30,72%, *confiança* 59% *e lift* 1,25);
- 3o.*Quadrante* ==> *FEMININO* (*suporte* 31,42% *confiança* 67% *e lift* 1,10);
- 4o.*Quadrante* ==> *FEMININO* (*suporte* 31,14% *confiança* 62% *e lift* 1,03).

A obtenção de regras de classificação sobre a base de dados odontológica se deu pela aplicação do algoritmo C4.5, implementado pelo classificador do *Weka*: *weka.classifiers.trees.J4*8. O algoritmo apresentou os resultados por meio de árvores de decisão, para análise posterior.

Foram utilizados todos os atributos do *dataset* original, sendo testados vários percentuais para confiança mínima, na utilização de poda. Selecionou-se para fazerem parte da classe classificadora, em momentos diferentes, os atributos: "sexo", "faixa etária" e "especialidade".

O atributo "idade" não precisa ser discretizado para a execução desse algoritmo, uma vez que o mesmo trabalha com atributos do tipo numérico. Contudo, por ser do tipo numérico, esse atributo não pode ser selecionado como atributo classe, sendo uma exigência do C4.5. O atributo "faixa etária", juntamente com a "idade" podem resultar alguma informação relevante. A seguir, são mostrados os resultados mais relevantes para cada classe escolhida.

Com 70% de confiança mínima para poda e definindo-se como classe o atributo "sexo", foi gerada uma árvore de decisão e, para simplificar a análise dos modelos obtidos, a Figura 5.1 mostra uma operação de transformação de modelo, mudando a representação do modelo para a figura de uma árvore de decisão.

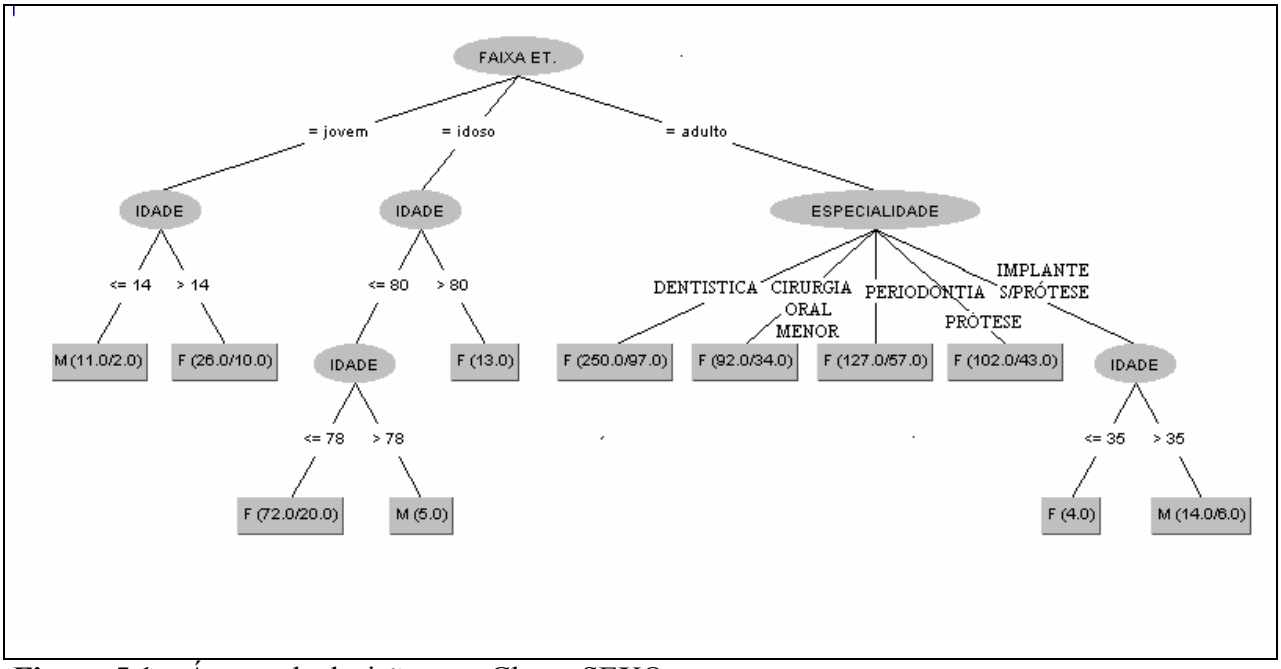

Figura 5.1 – Árvore de decisão para Classe SEXO.

O atributo "faixa etária", quando selecionado para ser o atributo classificador, apresentou resultados óbvios, sem acrescentar nada a este processo de KDD, gerando a seguinte árvore de decisão:

IDADE <= 64
| IDADE <= 20: jovem (37.0)
| IDADE > 20: adulto (589.0)
IDADE > 64: idoso (90.0)

Com 40% de confiança mínima para poda e classe ESPECIALIDADE, abstraiu-se uma árvore de decisão, na qual a Figura 5.2 mostra a representação simplificada do modelo.

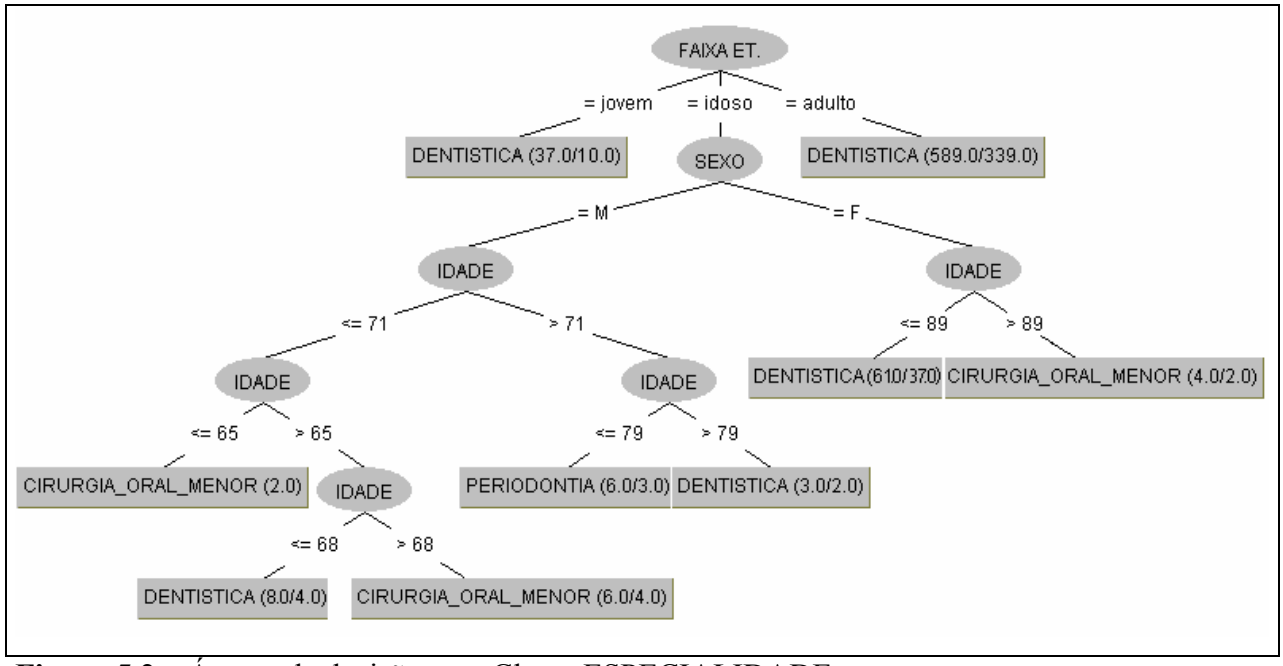

Figura 5.2 – Árvore de decisão para Classe ESPECIALIDADE.

A seguir será realizada a avaliação dos resultados obtidos na fase de mineração dos dados, detalhando as configurações e parâmetros utilizados nas execuções dos algoritmos de mineração de dados, além das principais características observadas durante a construção do modelo de conhecimento.

Em relação à extração de regras de associação da base odontológica, os resultados evidenciaram características descritivas relacionadas à realidade que envolve os pacientes, os tratamentos e procedimentos clínicos adotados da clínica odontológica.

Os primeiros resultados obtidos com a execução do Apriori demonstraram que todas as regras descobertas referenciavam somente os dentes não tratados, produzindo somente informações irrelevantes. Percebeu-se então que o conteúdo dos trinta e dois dentes tratados, quando retratasse a falta de abordagem, através da letra ´N´, deveriam ser alterados para ´?´, indicando que o conteúdo daquele atributo deve ser desconsiderado. Esse fato leva à iteração com a fase de pré-processamento, a qual prepara os dados a serem modelados.

Observou-se que, após a alteração dos dados citados, o número de regras extraídas foi inferior a quatrocentas incidências. Também em todos os grupos de atributos testados avaliou-se que não houve ocorrências de regras sem sentido algum, ou seja, todas tinham algum sentido, mesmo que fossem óbvios.

Foi experimentada a associação do atributo "faixa etária" com os demais atributos do *dataset* e não foi obtida nenhuma regra com relevância para este estudo uma vez que a

incidência de pacientes adultos representa mais de 80% da base. Dessa forma, esse atributo não foi utilizado nesta técnica de mineração de dados.

Na avaliação dos resultados da mineração de dados no primeiro subconjunto de dados, o valor de suporte mínimo configurado para 10% permitiu uma melhor produção de informações com variedade de combinações entre os atributos do grupo testado, possuindo, em alguns casos, confiança mínima superior a 90% e lift superior a 5,0, indicando interdependência entre esses itens. Foram extraídas 202 regras de associação para esse subconjunto de dados.

A maioria das regras extraídas a partir dos conjuntos de itens freqüentes, apontam para o relacionamento entre quase todos os dentes do primeiro e segundo quadrante bucal, com maior incidência de relacionamento para os dentes 11, 12, 13, 21, 22.

De acordo com a regra "22, 11 ==> 21 (*suporte* 10,19%, *confiança* 92% *e lift* 5,75)", em 92% das vezes que os dentes 22 e 11 são tratados, o dente 21 é tratado também. A regra "11, 13 ==> 12 (*suporte* 11,17%, *confiança* 95% *e lift* 5,46)" também demonstra uma correlação entre os dentes 11, 13 e o 12, tendo 95% das ocorrências, ou seja, em 84 instâncias contendo o dente 11 e 13, 80 incluíam o dente 12. Nas duas regras avaliadas neste parágrafo, o valor da medida de interesse *lift* ultrapassou os cinco pontos. Esse valor significa que, estatisticamente há uma possibilidade cinco vezes maior dos dentes situados no conseqüente das regras, quando tratados, serem tratados os dentes abordados no antecedente das referidas regras, em relação à probabilidade dos tratamentos ocorrerem juntos caso fossem estatisticamente independentes.

A discretização do atributo "idade" em três intervalos resultou em regras sem perder qualidade nas informações, sendo evidenciados na regra: "*IDADE*<37,33*anos*==>*DENTISTICA* (*suporte* 17,73%, *confiança* 55% *e lift* 1,29)", que mostra que em 55% das ocorrências um paciente com idade inferior aos trinta e sete anos e meio utiliza recursos da Dentística em seu tratamento. A regra "*IDADE entre* 37.33 *e* 66.66 *anos* ==> *PROTESE* (*suporte* 12,43%, *confiança* 22% *e lift* 1,30)", demonstra que 22% das ocorrências dos pacientes próximos à terceira idade estão adotando próteses dentárias. Nas duas regras que relacionaram o atributo "idade" e seus intervalos, observou-se que o valor da *lift* ficou entre 1,29 e 1,30, indicando pouca dependência na ocorrência do conseqüente, dado o antecedente.

Na avaliação dos resultados da mineração de dados no segundo subconjunto de dados, foram consideradas as regras de associação referentes aos dentes tratados do primeiro quadrante bucal, sendo obtidas 72 regras com um suporte mínimo no valor de 10%, confiança

mínima superior a 90% e *lift* superior a 2,10, indicando interdependência entre o conseqüente e o antecedente das regras.

Os grupos de itens mais freqüentes referem-se às regras de associação com trios de dentes dispostos entre o antecedente e o conseqüente da regra. Contudo, os dentes que se relacionam nessas regras seguem uma estrutura seqüencial para sua composição, sendo, na maioria das vezes, vizinhos uns dos outros. Os grupos de dentes evidenciados nessas regras são (11, 12 e 13), (13, 14 e 15), (14, 15 e 16) e (15, 16 e 17). Assim, a regra "13, 15 ==> 14 (*suporte* 10,33%, *confiança* 95% *e lift* 4,89)" demonstra esse fato. Em 95% das ocorrências em que os dentes 13 e 15 apareciam juntos, aparecia também o dente 14, ou seja, em 78 instâncias do *dataset,* contendo os dentes 13 e 15, 74 incluíam o dente 14. O *lift* dessa regra indica que há uma probabilidade 4,89 vezes maior dos dentes 13 e 15 estarem relacionados em um tratamento do dente 14, do que se as ocorrências desses dentes fossem estatisticamente independentes.

Na avaliação dos resultados da mineração de dados no terceiro subconjunto de dados, a extração de regras de associação referente aos dentes tratados do segundo quadrante bucal totalizou 20 regras com um suporte mínimo no valor de 10%, confiança mínima superior a 50% e *lift* superior a 2,0, indicando interdependência entre o conseqüente e o antecedente das regras.

Ao contrário das regras de associação obtidas para os dentes do primeiro quadrante bucal, as regras de associação extraídas do segundo quadrante bucal evidenciaram somente pares de dentes. No segundo quadrante bucal, a associação entre os dentes posteriores (24, 25, 26 e 27) é maior em relação aos dentes anteriores (21, 22 e 23).

A regra "21 ==> 22 (*suporte* 11,31%, *confiança* 70% *e lift* 4,50)" tem confiança de 70%, ou seja, em 70% das vezes que o dente 21 é tratado, o dente 22 é tratado também. O valor da *lift* mostra que há 4,50 vezes mais possibilidade do dente 21 estar relacionado ao dente 22, em um tratamento, do que se as ocorrências fossem estatisticamente independentes.

Na avaliação dos resultados da mineração de dados no quarto subconjunto de dados foram considerados os valores de 10% para suporte mínimo, confiança mínima superior a 50% e *lift* superior a 2,0.

As regras de associação extraídas pelo Apriori*,* referenciando os dentes do terceiro quadrante bucal, mostraram associações entre os dentes posteriores (34 e 35) e (36 e 37), totalizando apenas quatro regras. Os dentes 31, 32 e 33 podem não ter sido relacionados em regras devido ao baixo número de instâncias no *dataset*, devendo ser levado em consideração a quantidade de instâncias dos dentes 34, 35, 36 e 37.

A regra de associação "34 ==> 35 (*suporte* 10,47%, *confiança* 70% *e lift* 3,98)" demonstra esse fato. Nessa regra, em 70% das vezes que o dente 34 é tratado, o dente 35 também é tratado. O *lift* da regra avaliada ultrapassou os 3,50 pontos. Este valor significa que, estatisticamente há uma probabilidade três vezes e meia maior dos dentes situados no conseqüente das regras, quando tratados, serem tratados os dentes abordados no antecedente das referidas regras, em relação ao caso das ocorrências de tratamento serem independentes estatisticamente.

Na avaliação dos resultados da mineração de dados no quinto subconjunto de dados, identificou-se que não é diferente da avaliação feita nos modelos referentes aos dentes do terceiro quadrante bucal. Foram configurados os valores de 10% para suporte mínimo e 1,0 para *lift*, sendo extraídos somente seis regras que relacionam os pares de dentes (44 e 45), (45 e 47) e (46 e 47).

A regra "46 ==> 47 (*suporte* 12,43%, *confiança* 52% *e lift* 2,70)" mostra um desses relacionamentos, significando que em 52% das vezes que o dente 46 é tratado, o dente 47 também é tratado. O *lift* dessa regra mostra que há 2,70 vezes mais possibilidades de tratamento do dente 46 quando tratado o dente 47, do que se as ocorrências dos tratamento fossem estatisticamente independentes.

Nas regras de associação extraídas dos tratamentos no terceiro e quarto quadrante bucal, observou-se que não há incidência de regras referenciando os dentes anteriores, ocorrendo somente regras para os dente posteriores. Todavia, no primeiro e segundo quadrante bucal houve extração de regras relacionando tanto os dentes anteriores quanto os posteriores. Também não evidenciou-se em regras, os dentes de siso dos quatro quadrantes bucais: 18, 28, 38 e 48. Tal fato pode ser explicado pelo baixo número de instâncias no *dataset* que referenciavam os dentes citados

Na avaliação dos resultados da mineração de dados no sexto subconjunto de dados abordou-se os atributos "sexo", "idade", "especialidade", além dos quatro quadrantes bucais. Assim, o suporte mínimo que apresentou o melhor resultado na extração de regras de associação foi o valor de 30%. Para as medidas de interesse lift e confiança mínima foram considerados valores superiores a 1,12 pontos e 50% respectivamente.

Constatou-se a interdependência entre os dentes do primeiro e segundo quadrante bucal, por meio das regras "2o.*Quadrante* ==> 1o.*Quadrante* (*suporte* 30,72%, *confiança* 65% *e lift* 1,25)" e "1o.*Quadrante* ==> 2o.*Quadrante* (*suporte* 30,72%, *confiança* 59% *e lift* 1,25)", mostrando que mais de 59% das vezes que tratou-se algum dente do segundo quadrante bucal, tratou-se também um dente do primeiro quadrante, e vice versa. O *lift* dessas regras indica que

há uma possibilidade 1,25 vezes maior dos dentes do primeiro e segundo quadrante bucal estarem relacionados em um mesmo tratamento, em comparação à probabilidade esperada caso as ocorrências dos tratamentos fossem estatisticamente independentes entre si.

Em relação à execução da tarefa de classificação realizada com o intuito de predizer novas situações clínicas, no âmbito odontológico, os resultados obtidos evidenciam as características descritivas relacionadas a toda dinâmica que envolve os pacientes, tratamentos e procedimentos clínicos adotados dentro do período estudado pelo KDD.

O atributo "sexo", quando selecionado para ser o atributo classificador, apresentou resultados que apontam para a tendência de mulheres idosas tratarem seus dentes com maior freqüência em relação aos homens.

Observou-se que, o atributo "especialidade", quando selecionado para ser o atributo classificador, não extraiu regras de classificação que relacionassem a especialidade Implante e Prótese sobre Implante. Deve ser observado que o número de instâncias no *dataset* que se referem a esta especialidade, somente 23 instâncias, é relativamente pequena, em relação à quantidade de instâncias das demais especialidades, influenciando para este resultado. Já a especialidade Prótese possui uma quantidade superior de instâncias, 119, em relação à especialidade Cirurgia Oral Menor, 115. Contudo, extraiu-se regras referenciando Cirurgia Oral Menor, representando então um padrão de comportamento.

Outro detalhe observado, quando selecionado o atributo "especialidade' para ser o atributo classificador, é que nas faixas etárias jovem e adulto predominou a utilização da especialidade Dentística, enquanto o tratamento ao idoso dividiu-se entre Dentística, Cirurgia Oral Menor e Periodontia.

6. AVALIAÇÃO E IMPLANTAÇÃO

Durante a implantação do processo de KDD foram realizadas revisões dos objetivos várias vezes, uma vez que a falta de conhecimento do especialista do domínio da aplicação com os princípios do KDD e a falta de conhecimento do especialista de KDD com a área odontológica, associados às condições dos dados na base, levou-os a um estudo para adequação desses objetivos.

Na fase de avaliação dos modelos cogitou-se a possibilidade de um estudo voltado para a evolução dos materiais odontológicos utilizados nos tratamentos de Dentística e Prótese, levando-se em conta a classe social e a história bucal do paciente. O intuito seria identificar o perfil do paciente, na qualidade de consumidor desses produtos.

Os modelos obtidos na fase de mineração de dados foram apresentados ao especialista do domínio da aplicação para avaliação dos resultados quanto à qualidade e aplicabilidade desses para as tomadas de decisão.

Segundo o especialista do domínio, os resultados aumentaram o conhecimento sobre o perfil dos pacientes e das especialidades aplicadas nos tratamentos, descrevendo as características do cotidiano da clínica odontológica. Os modelos também confirmaram algumas tendências estudadas no meio odontológico, validando os resultados obtidos.

Para Mondelli, et al (1984), a questão da estética leva as pessoas a se preocuparem mais com seus dentes, uma vez que uma boca quando sorri deixa à mostra seus dentes e, qualquer alteração na aparência pode provocar implicações psicológicas sérias. Assim, a alta incidência de dentes frontais superiores em regras de associação seria confirmada por essa afirmativa.

As regras que referenciam associações entre dentes vizinhos ocorre, pois segundo Charbeneau et al (1978), as lesões dentárias provenientes de cárie são oriundas de colônias de bactérias e por esta razão, a probabilidade do dente vizinho ser lesionado também é elevado, ocasionando a cárie de contato.

Segundo Charbeneau et al (1978), os dentes superiores são mais suscetíveis ao ataque de cárie do que os dentes inferiores, observando que os dentes anteriores do terceiro e quarto quadrante bucal são os mais resistentes a essa doença. Dessa forma, a extração de regras de associação, referenciando dentes anteriores e posteriores do primeiro e segundo quadrante bucal, além dos dentes posteriores do terceiro e quarto quadrante bucal, estaria pausada nessa afirmação.

De acordo com Navarro e Côrtes (1995), a idade do paciente é fator importante de risco de cárie, mostrando em estudos recentes que o grau de ataque da cárie é similar em adultos e jovens. Contudo, em populações com faixas etárias próximas de 60 anos, a cárie é a principal causa da perda dos dentes. Assim, a extração de regras de associação, relacionando pacientes com idade inferior a 37 anos e meio a tratamento de Dentística é plenamente justificável, bem como a utilização de recursos de Prótese em pacientes entre 37 anos e meio e 66 anos e meio de idade.

Conforme Carranza (1985), o envelhecimento é o declínio da função natural, uma desintegração do controle equilibrado e da organização que caracteriza o jovem. No idoso, foram identificadas modificações gengivais e em outras áreas da mucosa bucal, redução na altura do osso alveolar, diminuição da vascularização, além do aumento nas fibras elásticas no ligamento periodontal. Esses fatores contribuem para os efeitos acumulativos das doenças

bucais. Esse fato também descreve uma tendência natural que na terceira idade a saúde bucal necessita de mais cuidados e que há a predominância de tratamentos na área de Dentística, Periodontia e de Cirurgia Oral Menor.

As regras extraídas na etapa de mineração de dados mostraram conhecimentos que, na maioria das vezes, são considerados triviais, mas somente com a utilização de métodos estatísticos podem disponibilizá-los ao especialista, confirmando a validade desses conhecimentos. Assim, considerou-se como modelo para estudo as regras que evidenciaram a especialidade Dentística, responsável pela estética bucal. Contudo, a questão de predizer novos casos não atingiu as expectativas esperadas, devido ao baixo número de instâncias na base de dados, considerando a inexistência de atributos que representem fatores importantes relacionados aos tratamentos realizados por outros especialistas.

Como próximo passo, o especialista do domínio irá rever a padronização dos dados que são inseridos no prontuário do paciente, além de terminar a atualização dos dados que ainda não foram inseridas na base de dados, perdurando um total de um a dois anos para o término dessa adequação. Como conseqüência, o volume de dados na base aumentará.

Uma seqüência deste trabalho é a proposição do mesmo estudo, para reavaliar os resultados deste trabalho, juntamente com um novo estudo que utilizaria a mesma base de dados construída, sendo inseridos novos atributos que avaliariam a evolução dos materiais odontológicos utilizados nos tratamentos de Dentística e Prótese, levando-se em conta a classe social e a história bucal do paciente. Com isso, o conhecimento obtido neste trabalho poderia ser comparado, verificando-se a qualidade das informações e, conseqüentemente, garantindo confiança nos resultados para aplicá-los em seguida.

A metodologia CRISP-DM garantiu uma condução segura da implantação do processo de KDD, indicando em todas as suas etapas o direcionamento burocrático das questões a serem respondidas e das ações a serem praticadas. A execução deste trabalho não seguiu totalmente a forma iterativa proposta pela metodologia CRISP-DM, entre as três primeiras etapas. Tão pouco se executou todas as atividades propostas, uma vez que se para o problema em questão definiu-se não serem necessárias. Os primeiros passos do processo de KDD tiveram foco nas análises na compreensão dos dados e avaliação da situação para posteriormente interagir com as demais etapas propostas. A fase mais complexa foi a compreensão do negócio, pois foi necessário definir um objetivo em meio a uma diversidade de opções e suas limitações.

7. CONSIDERAÇÕES FINAIS

A confecção deste trabalho deu-se pelo objetivo de ampliar, por meio de um estudo de caso, o entendimento e a compreensão dessa área da Tecnologia da Informação, denominada KDD (*Knowledge Discovery in Data Base*).

A descoberta de conhecimentos úteis em banco de dados é um processo seguro quando implantado por uma metodologia estruturada em tarefas bem definidas a serem executadas, considerando-se o elevado grau de complexidade que envolve sua execução. Uma das complexidades está em interpretar adequadamente o conhecimento extraído em meio a vários questionamentos referentes à escolha certa das informações selecionadas para compor o *dataset* a ser minerado. A outra complexidade é a capacitação dos especialistas envolvidos no processo, uma vez que é necessário ter fundamentação teórica das áreas abordadas no processo de KDD. Contudo, o sucesso dos resultados dependerá principalmente da disponibilidade, qualidade e do formato dos dados dispostos em suas bases de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

CARRANZA, F.; **Periodontia Clínica de Glickman.** 6.ed, Rio de Janeiro: Interamericana, 1985.

CHAPMAN, P.; et al. **CRISP-DM 1.0 Step-by-step data mining guide.** USA, 2000. Disponível em: <http://www.crisp-dm.org/download.htm>. Acesso em: 20 maio 2006.

CHARBENEAU, G.; et al. **Princípio e Prática de Dentística Operatória.** Rio de Janeiro: Guanabara, 1978.

FAYYAD, U. M.; PIATETSKY, G.; SMYTH, P. From Data Mining to Knowledge Discovery: Na Overview. Knowledge Discovery and Data Mining, Menlo Park: AAAI Press,1996.

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

MONDELLI, J.; et al. **Restaurações estéticas.** São Paulo**.** Rio de Janeiro: Sarvier, 1984.

NAVARRO, M.; CÔRTES, D. Avaliação e tratamento do paciente com relação ao risco de cárie. **Maxi-Odonto: Dentística**, Bauru, SP, v. 1, n. 4, p. 5-11, jul./ago. 1995