

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL  
FACULDADE DE INFORMÁTICA  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

**GESTÃO DE MÉTRICAS E INDICADORES DE DOENÇAS EM SAÚDE BUCAL  
SUPPORTADO POR UM AMBIENTE DE DESCOBERTA DE CONHECIMENTO  
EM BANCO DE DADOS**

**LUCIANO COSTA BLOMBERG**

Dissertação apresentada como requisito parcial à  
obtenção do grau de Mestre em Ciência da  
Computação na Pontifícia Universidade Católica do  
Rio Grande do Sul.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz

Porto Alegre  
2010

# **Livros Grátis**

<http://www.livrosgratis.com.br>

Milhares de livros grátis para download.

## Dados Internacionais de Catalogação na Publicação (CIP)

B653g Blomberg, Luciano Costa

Gestão de métricas e indicadores de doenças em saúde bucal suportado por um ambiente de descoberta de conhecimento em banco de dados / Luciano Costa Blomberg. – Porto Alegre, 2010.

107 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.

1. Informática. 2. Bancos de Dados. 3. Mineração de Dados (Informática). 4. Data Warehouse. I. Ruiz, Duncan Dubugras Alcoba. II. Título.

CDD 005.74

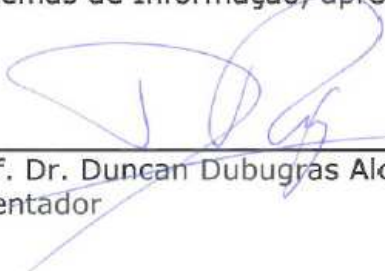
**Ficha Catalográfica elaborada pelo  
Setor de Tratamento da Informação da BC-PUCRS**




Pontifícia Universidade Católica do Rio Grande do Sul  
FACULDADE DE INFORMÁTICA  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

## TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

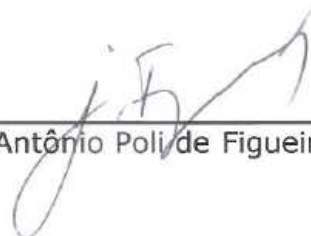
Dissertação intitulada "Gestão de Métricas e Indicadores de Doenças em Saúde Bucal Suportado por um Ambiente de Descoberta de Conhecimento em Banco de Dados", apresentada por Luciano Costa Blomberg, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Sistemas de Informação, aprovada em 16/03/10 pela Comissão Examinadora:

  
\_\_\_\_\_  
Prof. Dr. Duncan Dubugras Alcoba Ruiz -  
Orientador

PPGCC/PUCRS

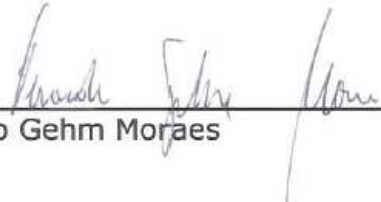
  
\_\_\_\_\_  
Profa. Dra. Ana Paula Terra Bacelo -

PPGCC/PUCRS

  
\_\_\_\_\_  
Prof. Dr. José Antônio Poli de Figueiredo -

PPGO/PUCRS

Homologada em...27/04/10..., conforme Ata No. 007... pela Comissão Coordenadora.

  
\_\_\_\_\_  
Prof. Dr. Fernando Gehm Moraes  
Coordenador.

**PUCRS**

**Campus Central**

Av. Ipiranga, 6681 - P32- sala 507 - CEP: 90619-900

Fone: (51) 3320-3611 - Fax (51) 3320-3621

E-mail: [ppgcc@pucrs.br](mailto:ppgcc@pucrs.br)

[www.pucrs.br/facin/pos](http://www.pucrs.br/facin/pos)

Dedico este trabalho aos meus pais pela educação exemplar e a minha esposa Anelise pelo apoio incondicional durante os momentos mais difíceis desta caminhada. Bela, este trabalho também é para você minha filha.

## **AGRADECIMENTOS**

É com extrema satisfação que deixo meus agradecimentos a todas as pessoas que acreditaram no meu trabalho e que foram essenciais para mais esta conquista.

Entre estas pessoas faço questão de destacar:

Minha esposa Anelise pela cumplicidade e apoio incondicional;

Meus pais e minha irmã por serem essenciais na minha formação humana;

A família de minha esposa pela confiança e incentivo;

O Prof. Duncan pela oportunidade de crescimento profissional e pelas preciosas orientações;

Os amigos do GPIN pelas dicas, trocas de conhecimento e pelos momentos “ahh poxa”;

O grande amigo Anderson Bestteti por ser um referencial de conduta profissional;

Os Professores e Pesquisadores da Faculdade de Odontologia, especialmente Figueiredo; Mota, Helenita e Fabiana por compartilharem seus conhecimentos;

Os meus amigos, pelo incentivo e compreensão nos momentos de ausência,

Para finalizar, agradeço ao PPGCC e CNPq por financiarem meus estudos.

# GESTÃO DE MÉTRICAS E INDICADORES DE DOENÇAS EM SAÚDE BUCAL SUPOSTADO POR UM AMBIENTE DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS

## RESUMO

Embora o último levantamento epidemiológico (BRASIL, 2006) revele uma sensível melhora na condição bucal da população brasileira nas últimas duas décadas, patologias bucais ainda demandam grandes investimentos financeiros por parte do governo federal. Este trabalho tem como objetivo desenvolver e documentar uma abordagem computacional (*Knowledge Discovery in Database*) capaz de gerenciar grandes volumes de dados e produzir modelos mais compreensíveis para o suporte à tomada de decisão, formulação de melhores políticas de saúde bucal, bem como a viabilização de novas atividades de ensino e pesquisa na área. Para tanto, analisamos uma amostra de dados referente às fichas odontológicas de 598 pacientes de baixa renda, atendidos junto a uma unidade da PUCRS vinculada ao SUS (Sistema Único de Saúde). Como principal resultado deste trabalho, identificamos oportunidades de *data mining* pela extração de modelos preditivos aplicados à análise de patologias periodontais, má-oclusão e indicadores de cárie dentária.

**Palavras-chave:** Banco de dados operacional, mineração de dados, data warehouse, indicadores, KDD.

# MANAGEMENT OF METRICS AND INDICATORS IN ORAL HEALTH DISEASES SUPPORTED BY AN ENVIRONMENT OF KNOWLEDGE DISCOVERY IN DATABASES

## ABSTRACT

Although the last epidemiological survey (BRAZIL, 2006) reveal a significant improvement in oral health status of the Brazilian population in the last two decades, oral diseases still require large financial investments by the federal government. This study aims to develop and document a computational approach (Knowledge Discovery in Database Process) capable of managing large volumes of data and producing more comprehensible models to support the decision making process and the formulation of improved oral health policies, as well as making new teaching and research activities in the area viable. Therefore, we analyze a sample of data concerning the dental records of 598 low-income patients, treated with one unit of PUCRS linked to the SUS (Unified Health System). The main result of this work, we identified opportunities for data mining for the extraction of predictive models applied to the analysis of periodontal diseases, malocclusion and indicators of dental caries.

**Keywords:** Database operational, data mining, data warehousing, indicators, KDD.



## LISTA DE FIGURAS

Figura 1 - Processo de formação do indicador de cárie dentária .....	21
Figura 2 - Etapas do processo de KDD (FAYAD; PIATETSKY; SMYTH, 1996) .....	25
Figura 3 – Fato, dimensão e medida (MACHADO, 2004).....	29
Figura 4 – Exemplo de análise de associações (TAN; STEINBACH; KUMAR, 2006)....	30
Figura 5 – Exemplo de árvore de decisão (MONTENEGRO; OLIVEIRA; CABRAL, 2008) 32	
Figura 6 – Processo de KDD em Saúde Bucal.....	35
Figura 7 – Projeto de construção de banco de dados (ELMASRI; NAVATHE, 2006) .....	36
Figura 8 – Modelagem lógica do BDSB .....	37
Figura 9 – Definições físicas do BDSB .....	38
Figura 10- Interface para população do BDSB .....	40
Figura 11 – Modelo analítico constelação de fatos.....	46
Figura 12 – Script SQL para transformação dos dados .....	47
Figura 13 – Gráfico de procedimentos odontológicos .....	49
Figura 14 – Gráfico de Tratamentos odontológicos.....	49
Figura 15 – Gráfico de Materiais Dentários.....	50
Figura 16 – Gráfico dos tratamentos x profissões.....	50
Figura 17 – Índice CPO-D por faixa-etária.....	51
Figura 18 – Primeiro Experimento – MP I .....	52
Figura 19 – Primeiro Experimento – MP II .....	53
Figura 20 – Segundo Experimento – MP I .....	56
Figura 21 – Segundo Experimento – MP II .....	57
Figura 22 – Segundo Experimento – MP III.....	58
Figura 23 – Segundo Experimento – MP IV.....	59
Figura 24 – Segundo Experimento – MP V .....	60

## LISTA DE QUADROS

Quadro 1 – Exemplo de Matriz de confusão para problema binário .....	31
Quadro 2 – Resultados da primeira etapa da apreensão de dados .....	40
Quadro 3 – Resultados da segunda etapa da apreensão de dados .....	41
Quadro 4 – Campos sem preenchimento (Anamnese).....	41
Quadro 5 – Campos sem preenchimento (Dados de saúde bucal).....	41
Quadro 6 – Campos sem preenchimento (Diagnóstico clínico) .....	41
Quadro 7 – Campos sem preenchimento (Plano de tratamento).....	42
Quadro 8 – Campos sem preenchimento (Odontograma).....	42
Quadro 9 – Campos sem registro temporal (Data de nascimento) .....	42
Quadro 10 – Campos sem registro temporal (Data de abertura da ficha).....	42
Quadro 11 – Nova amostra considerada .....	43
Quadro 12 – Primeira Questão de Negócio .....	46
Quadro 13 – Segunda Questão de Negócio .....	47
Quadro 14 – Oversampling .....	55

## **LISTA DE TABELAS**

Tabela 1 - Considerações gerais sobre o Projeto de Construção do BDSB.....	39
--	----

## LISTA DE SIGLAS

<b>BDSB</b>	Banco de Dados em Saúde Bucal
<b>CASE</b>	Computer Aided Software Engineering
<b>CERLAP</b>	Centro de Reabilitação de Fissuras Lábio-palatinas
<b>CEU</b>	Centro de Extensão Universitária
<b>CPO</b>	Cariados, Perdidos, Obturados
<b>CSV</b>	Comma-Separated Values
<b>DATASUS</b>	Departamento de Informática do SUS
<b>FO</b>	Faculdade de Odontologia
<b>IBGE</b>	Fundação Instituto Brasileiro de Geografia e Estatística
<b>J2SE</b>	Java 2 Standard Edition
<b>KDD</b>	Knowledge Discovery in Database
<b>MD</b>	Modelo Descritivo
<b>MP</b>	Modelo Preditivo
<b>OLAP</b>	On-line Analytical Processing
<b>OLTP</b>	On-line Transaction Processing
<b>OMS</b>	Organização Mundial da Saúde
<b>PPT</b>	Plano Preventivo-Terapêutico
<b>SGBD</b>	(Sistema Gerenciador de Banco de Dados)
<b>SQL</b>	Structured Query Language
<b>SUS</b>	Sistema Único de Saúde
<b>UML</b>	Unified Modeling Language

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b> .....	15
<b>1.1</b>	<b>DESCRIÇÃO DO CENÁRIO</b> .....	16
<b>1.2</b>	<b>CARACTERIZAÇÃO DO PROBLEMA</b> .....	17
<b>1.3</b>	<b>QUESTÃO DE PESQUISA</b> .....	18
<b>1.4</b>	<b>OBJETIVO GERAL</b> .....	18
<b>1.5</b>	<b>CARACTERIZAÇÃO DA CONTRIBUIÇÃO</b> .....	18
<b>1.6</b>	<b>CONSIDERAÇÕES FINAIS DO CAPÍTULO</b> .....	19
<b>2</b>	<b>REFERENCIAL TEÓRICO</b> .....	21
<b>2.1</b>	<b>MÉTRICAS E INDICADORES EM SAÚDE BUCAL</b> .....	21
<b>2.1.1</b>	<b>Principais Indicadores</b> .....	22
2.1.1.1	CPO-D .....	22
2.1.1.2	CEO-D=0 .....	22
2.1.1.3	IED (Índice de Estética Dentária) .....	23
2.1.1.4	Primeira Consulta Odontológica Programática .....	24
2.1.1.5	Ação Coletiva Escovação Dental Supervisionada .....	24
<b>2.2</b>	<b>PROCESSO DE KDD</b> .....	25
<b>2.2.1</b>	<b>Conceituação</b> .....	25
<b>2.2.2</b>	<b>Etapas do Processo de KDD</b> .....	25
<b>2.3</b>	<b>DATA WAREHOUSING</b> .....	27
<b>2.3.1</b>	<b>Modelo Analítico</b> .....	28
<b>2.4</b>	<b>DATA MINING</b> .....	29
<b>2.4.1</b>	<b>Tarefas Descritivas</b> .....	29
<b>2.4.2</b>	<b>Tarefas preditivas</b> .....	31
<b>2.5</b>	<b>CONSIDERAÇÕES FINAIS DO CAPÍTULO</b> .....	33
<b>3</b>	<b>DESENVOLVENDO UM PROCESSO DE KDD EM SAÚDE BUCAL</b> .....	35
<b>3.1</b>	<b>DESENVOLVIMENTO DE UM AMBIENTE OLTP</b> .....	35
<b>3.1.1</b>	<b>Identificação de Requisitos</b> .....	36
<b>3.1.2</b>	<b>Modelagem Conceitual</b> .....	36
<b>3.1.3</b>	<b>Modelagem Lógica</b> .....	37
<b>3.1.4</b>	<b>Implementação Física do BDSB</b> .....	38
<b>3.1.5</b>	<b>Implementação de Interface para População do BDSB</b> .....	39
<b>3.1.6</b>	<b>Apreensão dos dados</b> .....	40
<b>3.2</b>	<b>DESENVOLVIMENTO DE UM AMBIENTE OLAP</b> .....	43
<b>3.2.1</b>	<b>Identificando os requisitos de negócio</b> .....	43
<b>3.2.2</b>	<b>Construindo um Modelo de Análise de Dados</b> .....	45
<b>3.2.3</b>	<b>Extraindo os dados do BDSB</b> .....	47
<b>3.2.4</b>	<b>Transformação dos dados</b> .....	47
<b>3.2.5</b>	<b>Carga dos Dados</b> .....	48
<b>3.2.6</b>	<b>Geração do Cubo de Dados</b> .....	48
<b>3.2.7</b>	<b>Explorando o Cubo para Atender as Questões de Negócio</b> .....	48
<b>3.3</b>	<b>DATA MINING</b> .....	51
<b>3.3.1</b>	<b>Extração de modelos preditivos</b> .....	51
3.3.1.1	Primeiro Experimento .....	52
3.3.1.2	Segundo Experimento .....	54
<b>3.4</b>	<b>CONSIDERAÇÕES FINAIS DO CAPÍTULO</b> .....	60

<b>4</b>	<b>TRABALHOS RELACIONADOS.....</b>	<b>63</b>
<b>4.1</b>	<b>CONSIDERAÇÕES FINAIS DO CAPÍTULO.....</b>	<b>65</b>
<b>5</b>	<b>CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS .....</b>	<b>67</b>
<b>5.1</b>	<b>PUBLICAÇÕES.....</b>	<b>68</b>
	<b>REFERÊNCIAS .....</b>	<b>69</b>
	<b>APÊNDICE A – MODELO CONCEITUAL BDSB.....</b>	<b>73</b>
	<b>APÊNDICE B – ENTREVISTA / ÁREA BÁSICA .....</b>	<b>75</b>
	<b>APÊNDICE C – ENTREVISTA / ÁREA DIAGNÓSTICOS POR IMAGENS .....</b>	<b>77</b>
	<b>APÊNDICE D – ENTREVISTA / ÁREA ENDODONTIA.....</b>	<b>79</b>
	<b>APÊNDICE E – ENTREVISTA /ÁREA DE ENSAIOS MECÂNICOS .....</b>	<b>81</b>
	<b>APÊNDICE F – ENTREVISTA / ÁREA ESTOMATOLOGIA .....</b>	<b>83</b>
	<b>APÊNDICE G – ENTREVISTA / ÁREA ODONTOLOGIA EM SAÚDE COLETIVA .....</b>	<b>85</b>
	<b>APÊNDICE H – ENTREVISTA / ÁREA DE TRAUMA DENTO-ALVEOLAR .....</b>	<b>87</b>
	<b>APÊNDICE I - PRIMEIRO EXPERIMENTO – MP I.....</b>	<b>89</b>
	<b>APÊNDICE J - PRIMEIRO EXPERIMENTO – MP II.....</b>	<b>91</b>
	<b>APÊNDICE K – SEGUNDO EXPERIMENTO – MP I.....</b>	<b>93</b>
	<b>APÊNDICE L - SEGUNDO EXPERIMENTO – MP II .....</b>	<b>95</b>
	<b>APÊNDICE M - SEGUNDO EXPERIMENTO – MP III.....</b>	<b>97</b>
	<b>APÊNDICE N - SEGUNDO EXPERIMENTO - MP IV .....</b>	<b>99</b>
	<b>APÊNDICE O - SEGUNDO EXPERIMENTO - MP V.....</b>	<b>101</b>
	<b>APÊNDICE P - TERCEIRO EXPERIMENTO - MD I.....</b>	<b>103</b>
	<b>ANEXO A – SOLICITAÇÃO FACULDADE DE INFORMÁTICA .....</b>	<b>105</b>
	<b>ANEXO B – SOLICITAÇÃO FACULDADE DE ODONTOLOGIA .....</b>	<b>107</b>



# 1 INTRODUÇÃO

Apesar da significativa melhora dos índices de cárie na população infantil nas últimas duas décadas, agravos bucais ainda constituem-se como um problema de saúde pública diretamente relacionado às condições de vida da população. A cárie dentária é a mais comum doença crônica na infância, cinco vezes mais comum do que a asma e sete vezes mais comum do que a febre do feno (MONTENEGRO; OLIVEIRA; CABRAL, 2008). Segundo dados do último levantamento epidemiológico em saúde bucal (BRASIL, 2006), a doença periodontal mostrou-se alta em todas as faixas etárias com menos de 22% da população adulta e menos de 8% dos idosos apresentando gengivas saudáveis.

Considerando a incidência de má oclusão na população brasileira, o levantamento concluiu que os dados de prevalência na idade de 5 anos revelam problemas oclusais moderados ou severos em 14,5% da população nessa idade, variando de um mínimo de 5,6% na região Norte a um máximo de 19,4% na região Sul. Outrossim, observou-se também um aumento na mortalidade ocasionada por câncer de boca nos anos de 1979 a 1998, de 1,32 a 1,82 por 100.000 habitantes, sendo este aumento principalmente observado no sexo masculino (BRASIL, 2002).

Dada a gravidade deste quadro epidemiológico, é de extrema importância que os gestores públicos passem a adotar medidas que possibilitem a investigação dos fatores causadores destas doenças, mas principalmente, ações de planejamento, gestão e avaliação de políticas voltadas à melhoria da saúde bucal. Neste sentido, o uso de indicadores tem se mostrado como um valioso instrumento para gestão e avaliação de saúde em todos os níveis, permitindo a identificação de grupos humanos com maiores necessidades de saúde, estratificando o risco epidemiológico e identificando as áreas críticas (RIPSA, 2008).

Na área de odontologia, sua aplicação tem trazido grandes benefícios, como por exemplo: a) a identificação de novos grupos de risco, b) a definição da taxa de incidência de determinadas patologias (ex: cárie, má oclusão, doenças periodontais), c) conhecimentos sobre a perda dentária, bem como, uma importante base de conhecimento para tomada de ações preventivas em saúde bucal.

Durante as últimas duas décadas, uma série de estudos foram realizados com o objetivo de identificar o perfil odontológico da população brasileira. Em um cenário mais abrangente, os levantamentos epidemiológicos realizados pelo Ministério da Saúde em 1986, 1996 e 2003 têm servido como fonte de informação para novas pesquisas na área, especialmente àquelas realizadas



com o intuito de explorar a associação entre indicadores sociais e os de saúde bucal. Dentre as pesquisas mais relevantes, destacamos duas linhas de abordagens. Na primeira linha, estão os estudos descritivos e preditivos de cunho puramente estatístico (correlação, regressão linear e logística) (BALDANI; NARVAI; ANTUNES, 2004), (FERNANDES; PERES, 2005), (TAGLIAFERRO et al., 2006), (CELESTE; NADANOVSKY; LEON, 2007). Na segunda linha, estão os trabalhos que se utilizam de técnicas automatizadas de *data mining* (árvores de decisão, regras de associação, redes neurais.) para extração de modelos descritivos e preditivos em saúde bucal (GANSKY, 2003), (CUNHA; DIAS, 2007), (MONTENEGRO; OLIVEIRA; CABRAL, 2008), (TAMAKI et al., 2009).

Observando especificamente esta segunda linha de abordagem, constatamos que os trabalhos desenvolvidos na área ainda estão muito focados em uma visão parcial do processo, ou seja, uma exploração comparativa de técnicas de *data mining*, deixando de lado etapas de extrema importância (preparação dos dados, interpretação e validação dos resultados) para obtenção de melhores resultados em um processo de KDD (Knowledge Discovery in Database).

Destacada a importância do indicador como instrumento de diagnóstico e planejamento, o objetivo deste trabalho é desenvolver e documentar um processo de KDD completo, cujo resultado permita auxiliar os gestores em saúde bucal na tomada de decisão e formulação de políticas de saúde melhor direcionadas a partir do uso de indicadores. Para tanto, desenvolvemos um ambiente de *data warehousing* para o armazenamento de prontuários odontológicos, aplicando-se sob este repositório, técnicas preditivas de *data mining* para extração de modelos aplicados à análise de patologias periodontais, má-oclusão e indicadores de cárie dentária.

Nas próximas três seções, descreveremos o cenário encontrado junto à instituição parceira, os principais problemas encontrados, bem como o foco da contribuição deste trabalho.

## **1.1 DESCRIÇÃO DO CENÁRIO**

A FO (Faculdade de Odontologia) da PUCRS conta com mais de 50 anos de experiência nos campos do ensino, pesquisa e extensão. Juntamente à graduação, o Programa de Pós-Graduação *Stricto Sensu* com áreas de concentração em Cirurgia e Traumatologia Bucocomaxilofacial, Dentística Restauradora, Endodontia, Estomatologia Clínica, Materiais Dentários, Ortodontia e Ortopedia Facial, e Prótese Dentária somam esforços para desenvolvimento da Odontologia no Brasil. Com esta ampla área de atuação, um grande volume de dados clínicos, exames laboratoriais, exames de

imagem (radiografia, tomografia e ecografia) e fichas auxiliares são gerados diariamente, tornando sua análise manual um processo exaustivo.

No intuito de viabilizar o resgate de informações relevantes à pesquisa, como métricas e indicadores em saúde bucal, foi desenvolvido em parceria com o Programa de Pós-Graduação em Odontologia um ambiente para armazenamento, organização e recuperação de dados provenientes de prontuários odontológicos.

Mediante a diversidade de áreas envolvidas na parceria e o curto período de tempo para apreensão dos dados, optamos por selecionar uma única área (Odonto Social – CEU Vila Fátima) como unidade de estudo, a qual julgamos como de maior relevância social. Esta opção justifica-se à medida que o CEU Vila Fátima desenvolve um serviço de assistência social e odontológica pelo SUS (Sistema Único de Saúde) a uma população de baixa renda, com mais de 8 mil habitantes, situada na Zona Leste da capital gaúcha.

## **1.2 CARACTERIZAÇÃO DO PROBLEMA**

Como se pode observar, este é um trabalho laborioso que demandou um grande esforço para sua implementação. Entre os fatores complicadores, observados durante sua execução, destacamos:

- O grande volume de prontuários em diferentes áreas: estima-se, segundo informações da Faculdade de Odontologia da PUCRS que entre as unidades de atendimento odontológico analisadas exista aproximadamente 50 mil prontuários de pacientes.
- A inexistência de uma base de dados constituída: este fato gerou uma série de prejuízos, tanto operacionais, quanto para o andamento da pesquisa. Um dos maiores prejuízos operacionais referiu-se à falta de integração dos dados, ou seja, novos prontuários eram gerados em diferentes unidades de atendimento, fato este, que vem impossibilitando a compreensão integral do histórico odontológico do paciente. Sob o ponto de vista da pesquisa, outro grande prejuízo é gerado, à medida que os dados estavam sujeitos a descartes periódicos, desperdiçando assim, uma preciosa fonte de dados para a pesquisa.
- Falta de padronização no preenchimento dos prontuários: Pudemos observar na amostra de dados fornecida pelas unidades de atendimento odontológico que

não existia uma padronização para o preenchimento dos campos do prontuário, ou seja, um mesmo campo era preenchido ora com valor inteiro, ora com valor textual.

- Dados faltantes: a falta de preenchimento de determinados campos pode ser vista como uma contingência, ou até mesmo como uma decorrência da falta de padronização no preenchimento dos prontuários. Na análise dos prontuários, identificamos que muitas informações deixaram de ser registradas no seu campo destino para serem registradas como descrições textuais em outros campos.

Dadas as características típicas de um cenário de saúde bucal, bem como a constante necessidade de se fundamentar decisões em informações seguras e confiáveis, entendemos como problemática a falta de uma melhor documentação na literatura que auxilie o gestor de saúde bucal no tratamento de tais características. Deste modo, apresentamos na próxima seção uma significativa contribuição computacional para a aplicação de um processo de descoberta de conhecimento em saúde bucal.

### **1.3 QUESTÃO DE PESQUISA**

Mediante a necessidade de viabilizar novas atividades de ensino e pesquisa na área de saúde bucal, bem como, melhorar o embasamento no planejamento de ações e políticas na área, como podem contribuir os modelos preditivos para gestão de métricas e indicadores relacionados a patologias bucais?

### **1.4 OBJETIVO GERAL**

O objetivo geral desta pesquisa é desenvolver e documentar um processo de descoberta de conhecimento, voltado à gestão de métricas e indicadores em saúde bucal. Como resultado, espera-se obter modelos preditivos que viabilizem novas atividades de ensino e pesquisa e que forneçam informações confiáveis para o planejamento de políticas e ações em saúde bucal.

### **1.5 CARACTERIZAÇÃO DA CONTRIBUIÇÃO**

A partir do trabalho realizado junto ao CEU Vila Fátima, caracterizamos nossa contribuição pelo desenvolvimento e documentação de um processo completo de KDD em saúde bucal, cujas características envolvam não apenas o tratamento de dados ausentes, mas também o tratamento de dados temporais e descritivos. Como resultado final, esperamos produzir modelos

preditivos aplicados à gestão de patologias periodontais, de má-oclusão e indicadores de cárie dentária. Desta forma, podemos visualizar a contribuição pelas seguintes etapas:

- A. Consolidação do BDSB (Banco de Dados em Saúde Bucal): Esforço empenhado para viabilizar o processo de descoberta de conhecimento, uma vez que não dispúnhamos de um repositório de dados para as áreas analisadas. Como resultado desta etapa, geramos um ambiente operacional para o armazenamento de prontuários odontológicos de 8 áreas vinculadas a FO.
- B. Construção de Data Warehouse Odontológico: A partir da consolidação do BDSB e escolha do CEU Vila Fátima como área de estudo, aplicamos o processo de ETC (Extração, Transformação, Carga) sob os dados do BDSB e construímos um novo ambiente para fins analíticos. Como resultado desta etapa, foram produzidos relatórios sumarizados para o atendimento das expectativas de negócio da instituição parceira.
- C. Exploração de modelos preditivos: Após a execução de um processo de *data warehousing*, utilizamos árvores de decisão (TAN; STEINBACH; KUMAR, 2006) para extração de modelos preditivos relacionados à gestão de patologias periodontais, de má-oclusão e indicadores de cárie dentária.

## 1.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo buscamos contextualizar o trabalho realizado junto a FO-PUCRS, bem como, apresentar as dificuldades encontradas para aplicação de um processo de KDD em um cenário atípico do convencionalmente encontrado na literatura, caracterizado por um grande volume de dados textuais, variantes no tempo e não armazenado digitalmente.

Como principal benefício desta pesquisa, esperamos fornecer ao gestor em saúde bucal um poderoso recurso computacional para apoio a tomada de decisão e formulação de políticas de saúde melhor direcionadas.

Para melhor compreensão, este trabalho está organizado da seguinte forma: no capítulo 1, contextualizamos o cenário de pesquisa e a contribuição deste trabalho. No capítulo 2, apresentamos o referencial teórico sobre o uso de indicadores, processo de KDD e as principais técnicas de *data mining*. No capítulo 3, apresentamos nossa solução de KDD em saúde bucal, bem como os modelos extraídos para gestão de patologias periodontais, de má oclusão e indicadores de cárie dentária. No capítulo 4 discutimos os trabalhos relacionados. Por fim, no capítulo 5 realizamos nossas considerações finais e perspectivas de trabalhos futuros.



## 2 REFERENCIAL TEÓRICO

### 2.1 MÉTRICAS E INDICADORES EM SAÚDE BUCAL

De um modo geral, tomadas de decisão não costumam ser procedimentos triviais, principalmente dentro da gestão pública, onde as decisões impactam diretamente na qualidade de vida da população. Ao longo dos últimos anos, o Governo Federal tem investido bilhões com a execução de políticas de saúde na área de saúde bucal. Entre as políticas mais efetivas, o “Brasil Sorridente” tem proporcionado ações de promoção, prevenção e recuperação da saúde bucal em todo território brasileiro. Como forma de diagnosticar a efetividade destas ações, indicadores tem sido amplamente utilizados.

Segundo documentos do Ministério da Saúde, indicadores são definidos como “[...] medidas-síntese que contêm informação relevante sobre determinados atributos e dimensões do estado de saúde, bem como do desempenho do sistema de saúde.” (RIPSA, 2008, p.13). Complementarmente, “[...] a construção de um indicador é um processo cuja complexidade pode variar desde a simples contagem direta de casos de determinada doença, até o cálculo de proporções, razões, taxas ou índices mais sofisticados [...]” (RIPSA, 2008).

A partir desta interpretação, bem como na obtida em (NOVELLO, 2006, p.13), ampliamos nosso entendimento sobre indicadores, definindo-o como uma variável utilizada na intenção de adicionar um alerta capaz de orientar decisões sobre determinado fenômeno de interesse. Em outra mão, o índice refere-se a uma combinação de variáveis que tenta sintetizar um conceito completo em um único valor, sendo muitas vezes base para formulação destes indicadores. Na figura 1 utilizamos o índice CPO-D (Dentes Cariados, Perdidos ou Obturados) para ilustrar o processo de formação de um indicador.

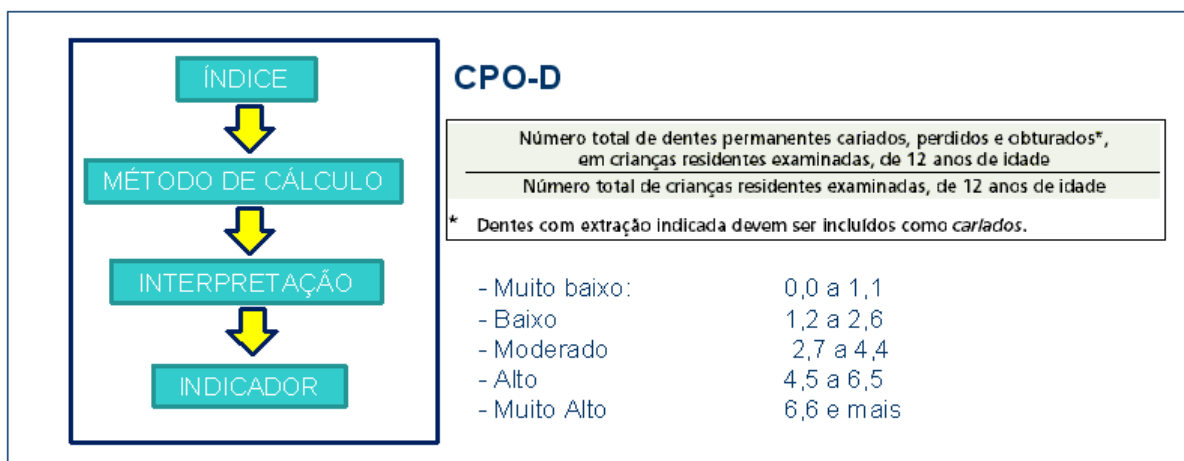


Figura 1 – Processo de formação do indicador de cárie dentária

Podemos ainda classificá-los em seis subconjuntos temáticos: a)demográficos, b)socioeconômicos, c)mortalidade, d)morbidade e fatores de risco, e)recursos e f)cobertura. Neste trabalho, caracterizamos o indicador por sua conceituação, método de cálculo, categorias de análise e fontes de dados. (RIPSA, 2008).

### 2.1.1 Principais Indicadores

Entre os principais índices associados a indicadores de morbidade e fatores de risco em saúde bucal estão:

#### 2.1.1.1 CPO-D

- a) conceituação:** número médio de dentes permanentes cariados, perdidos e obturados, aos 12 anos de idade, em determinado espaço geográfico, no ano considerado.
- b) método de cálculo:** número total de dentes permanentes cariados, perdidos e obturados, em crianças residentes examinadas, de 12 anos de idade, sobre o número total de crianças residentes examinadas, de 12 anos de idade.
- c) interpretação:** os valores do índice correspondem aos seguintes graus de severidade: muito baixo (0,0 a 1,1), baixo (1,2 a 2,6), moderado (2,7 a 4,4), alto (4,5 a 6,5) e muito alto (6,6 e mais).

#### 2.1.1.2 CEO-D=0

- a) conceituação:** Percentual de crianças de 5 a 6 anos de idade com índice CEO-D (número de dentes decíduos cariados, com extração indicada, perdidos devido à cárie ou obturados) igual a zero, em determinado espaço geográfico, no ano considerado.
- b) método de cálculo:** (número total de crianças residentes de 5 a 6 anos de idade examinadas com CEO-D = 0, sobre o número

total de crianças residentes de 5 a 6 anos de idade examinadas) x 100.

- c) interpretação:** a meta estabelecida pela OMS (Organização Mundial da Saúde) para o ano 2000 era de 50% das crianças livres de cárie (CEO-D = 0) aos 5 - 6 anos de idade.

### 2.1.1.3 IED (Índice de Estética Dentária)

- a) conceituação:** “[...] é um índice numérico que avalia as características oclusais, selecionadas de acordo com o potencial de causar incapacidade psicossocial.” (SANTOS et al., 2008, p.34).

- b) método de cálculo:** este índice pode ser calculado utilizando-se a seguinte equação de regressão:

$$(dentes\ ausentes\ visíveis\ x\ 6) + (apinhamento) + (espaçamento) + (diastema\ x\ 3) + (desalinhamento\ maxilar\ anterior) + (desalinhamento\ mandibular\ anterior) + (overjet\ maxilar\ anterior\ x\ 4) + (overjet\ mandibular\ anterior\ x\ 4) + (mordida\ aberta\ vertical\ anterior\ x\ 4) + (relação\ molar\ ântero-posterior\ x\ 3) + 13$$

- c) interpretação:** sem anormalidade ou má-oclusão leve (< 25), má-oclusão definida (26 a 30), má-oclusão severa (31 a 35), má oclusão muito severa ou incapacitante (> 36).

Poderíamos ainda destacar outros indicadores relacionados a patologias bucais, como câncer de boca, traumatismo dentário, doenças periodontais, fluorose dentária e edentulismo, porém, para condução deste trabalho, utilizaremos o indicador de cárie dentária, cujo conjunto de dados necessário para o cálculo do índice CPO-D é contemplado pela ficha odontológica do CEU Vila Fátima.

Entre os principais indicadores de cobertura relacionados à saúde bucal estão:



#### 2.1.1.4 Primeira Consulta Odontológica Programática

- a) conceituação:** é o percentual de pessoas que receberam uma primeira consulta odontológica programática, realizada com finalidade de diagnóstico e, necessariamente, elaboração de um PPT (Plano Preventivo-Terapêutico), para atender as necessidades detectadas.
- b) método de cálculo:** (número total de primeiras consultas odontológicas programáticas realizadas em um determinado local e período / pela população no mesmo local e período) x 100.
- c) interpretação:** a análise deste indicador está baseada no cumprimento de metas previamente definidas pelos órgãos de saúde.

#### 2.1.1.5 Ação Coletiva Escovação Dental Supervisionada

- a) conceituação:** é o percentual de pessoas que participaram da ação coletiva escovação dental supervisionada.
- b) método de cálculo:** média anual de pessoas participantes na ação coletiva Escovação Dental Supervisionada realizada em determinado local e período / pela população no mesmo local e período) x 100.
- c) interpretação:** a análise deste indicador está baseada no cumprimento de metas previamente definidas pelos órgãos de saúde.

Dada a importância destes indicadores como instrumento de planejamento, avaliação e diagnóstico em saúde, introduzimos o processo de KDD como um recurso computacional capaz de potencializar seus resultados. Conhecidamente, o processo de KDD tem sido amplamente explorado em diversas áreas de conhecimento, entre elas, a biologia (MACHADO, 2008), ciências médicas

(SAHAMA; CROLL, 2007), (STEINER et al., 2007), processo de desenvolvimento de software (BARROS et al., 2008), (FIGUEIRA; BECKER; RUIZ, 2007), workflow (GARCIA; RUIZ, 2005).

## 2.2 PROCESSO DE KDD

Esta seção foi elaborada com o objetivo de fornecer uma visão geral sobre o processo de descoberta de conhecimento em bases de dados.

### 2.2.1 Conceituação

O rápido avanço na coleta de dados e tecnologias de armazenamento tem permitido as organizações acumular grandes quantidades de dados ao longo das últimas décadas. No entanto, este grande volume de dados supera nossas habilidades de interpretá-los, criando-se assim a necessidade de se ter técnicas que permitam a sua automatização e análise de forma inteligente (TAN; STEINBACH; KUMAR, 2006).

Neste sentido, o processo de KDD tem se apresentado como um grande recurso para identificação de padrões válidos, novos, compreensíveis e potencialmente úteis (FAYAD; PIATETSKY; SMYTH, 1996).

### 2.2.2 Etapas do Processo de KDD

O processo de KDD é composto de cinco etapas, conforme ilustrado na figura 2.

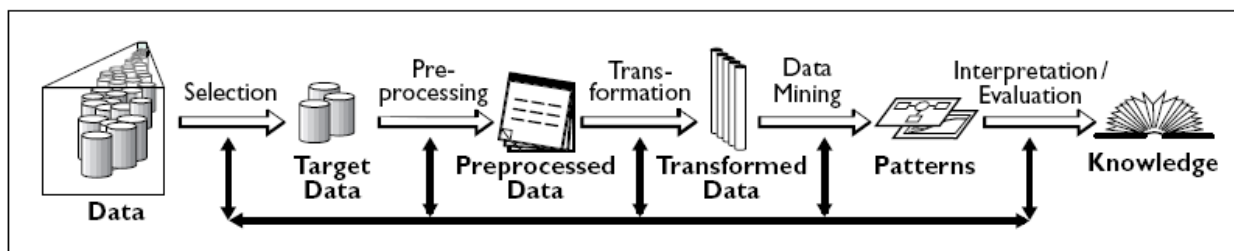


Figura 2 - Etapas do processo de KDD  
Fonte: Fayad, Piatetsky e Smyth (1996)

**a) Seleção dos dados:** consiste em selecionar um conjunto de dados ou focar sobre um subconjunto de variáveis em que se deseja realizar a descoberta;

**b) Pré-processamento e limpeza dos dados:** consiste na realização de operações básicas como exclusão de outliers, estratégias para tratamento de dados ausentes e valores desconhecidos. É considerada por muitos autores a etapa que mais consome tempo dentro do processo de KDD;

**c) Transformação dos dados:** consiste na realização de redução de dimensionalidade ou métodos para reduzir o número de variáveis;

**d) Data mining (mineração de dados):** pode ser vista em três partes:

- Escolha da função de mineração de dados: inclui a definição do propósito do modelo gerado pelo algoritmo de mineração de dados (ex. sumarização, classificação, regressão e agrupamento);
- Escolha do algoritmo de mineração de dados: consiste em selecionar os métodos a serem usados para a descoberta de padrões nos dados, tal como decidir que modelo e que parâmetros podem ser apropriados;
- *Data mining* (mineração de dados): consiste na busca de padrões de interesse em uma forma particular de representação ou conjunto de tais representações.

**e) Interpretação e avaliação dos resultados:** consiste em interpretar, avaliar e visualizar os padrões descobertos extraindo-se valores redundantes ou padrões irrelevantes.

A partir de Tan, Steinbach e Kumar (2006), complementamos o entendimento de algumas operações propostas por Fayad, Piatetsky e Smyth (1996) nas etapas de pré-processamento e transformação, conforme segue abaixo:

- **Agregação** – essa operação busca sumarizar o número de registros através da combinação de dois ou mais objetos em um único. Desta forma atributos quantitativos são geralmente agregados pegando-se uma soma ou média;
- **Amostragem** – abordagem utilizada para grandes volumes de dados, cujo objetivo é selecionar um subconjunto deste volume e reduzir consequentemente o tempo de processamento do algoritmo de mineração;
- **Redução de Dimensionalidade** – consiste na eliminação de atributos irrelevantes do conjunto de dados, permitindo assim, um melhor funcionamento dos algoritmos de mineração de dados e a geração de modelos mais compreensíveis;
- **Seleção de Atributos** – outra forma de redução de dimensionalidade, onde o objetivo é eliminar atributos que possam ser redundantes e irrelevantes ao objetivo da mineração;

- **Criação de Atributos** – consiste na criação de novos atributos a partir de outros existentes, de modo que informações importantes sejam capturadas em um conjunto de dados mais eficazmente;
- **Discretização** – a discretização consiste na transformação de atributos contínuos em atributos categóricos. É comumente aplicada em análise de classificação ou associação.

Nas próximas duas seções, buscaremos dar melhor compreensão sobre o desenvolvimento de um processo de *data warehousing* e o uso de técnicas de *data mining* para extração de modelos descritivos e preditivos.

## 2.3 DATA WAREHOUSING

Embora os sistemas OLTP (On-line Transaction Processing) cubram boa parte das operações do dia a dia (consultas, inserções, exclusões e atualizações), estes se mostram pouco adequados no suporte à tomada de decisão (HAN; KAMBER, 2001).

Em contrapartida, *data warehouses* são projetados precisamente para suportar extração, apresentação de dados e processamento eficientes, para finalidades analíticas e tomadas de decisão (HAN; KAMBER, 2001). Segundo Fayad, Piatetsky e Smyth (1996), sistemas de *data warehouse* podem ser definidos como um conjunto de dados baseado em assuntos, integrado, não volátil, e variável em relação ao tempo, e de apoio a decisões gerenciais. Já o termo *data warehousing* refere-se ao processo de construção e uso do *data warehouse*, compreendendo essencialmente as seguintes etapas:

- Extração é o primeiro passo no processo de obtenção dos dados para o *data warehouse*, cujo significado resume-se a ler e compreender a fonte de dados e copiar os dados necessários para o *data warehouse* através da *staging area* para manipulações posteriores. (KIMBALL; ROSS, 2002).
- Transformação: Assim que os dados são extraídos para a *staging area*, outras potenciais transformações podem ser realizadas, como por exemplo, a limpeza de dados (corrigindo erros ortográficos, resolvendo erros de domínio, resolver conflitos, lidar com elementos em falta, ou passar para formatos padronizados). (GARCIA; RUIZ, 2005). Durante o seguimento desta pesquisa, apresentaremos maiores detalhes sobre a execução desta etapa.

- Carga: Por fim, após os dados terem sido devidamente tratados na *staging area*, eles são carregados no *Data Warehouse*.

Uma popular abordagem para análise de data warehouses é chamada de Online Analytical Processing (OLAP) (FAYAD; PIATETSKY; SMYTH, 1996). Processamento analítico suporta operações OLAP incluindo operações de slice e dice, drill-down, roll-up.

Estas operações consistem em diferentes formas de visualização de dados históricos, apresentados de ambas formas detalhadas e resumidas, onde:

- **Slice:** redução de escopo, filtrar;
- **Dice:** é a mudança de perspectiva da visão.
- **Drill Down:** ocorre quando o usuário aumenta o nível de detalhe da informação, diminuindo o nível de granularidade;
- **Drill Up:** ocorre quando o usuário diminui o nível de detalhe da informação, aumentando o nível de granularidade;

*Data Warehouses* e ferramentas OLAP são comumente baseados em modelos analíticos (ou multidimensionais), permitindo que os dados sejam visualizados sob a forma de um cubo, ou seja, em múltiplas dimensões.

### 2.3.1 Modelo Analítico

Segundo Machado (2004), modelo analítico é um modelo de dados que apresenta um conjunto de medidas que descrevem aspectos comuns de negócios. Este modelo é utilizado especialmente para sumarizar e reestruturar dados e apresentá-los em visões que suportem a análise de valores destes dados.

Cada modelo analítico é formado por três elementos básicos:

- **Tabela Fato:** coleção de itens de dados, composta de dados e medidas de contexto;
- **Tabela Dimensão:** são os elementos que participam de um fato, ou as possíveis formas de visualizar um fato. Uma comum abordagem para identificação das dimensões é realizar as questões, onde, quando, quem e o que;
- **Medidas:** são os atributos numéricos que representam um fato.

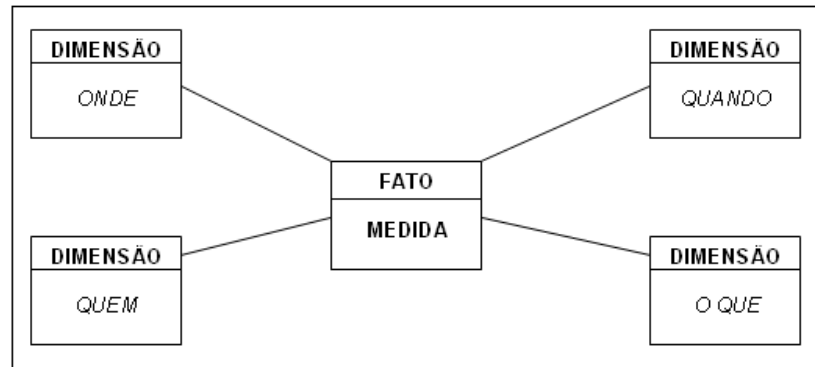


Figura 3 – Fato, dimensão e medida  
Fonte: Machado (2004)

Segundo Han e Kamber (2001), existem três tipos principais de modelo analítico:

**Modelo Estrela** – é o mais comum paradigma de modelagem, caracterizado, basicamente, por conter apenas uma tabela fato e diversas dimensões.

**Modelo Floco de Neve** – o modelo floco de neve é uma variação do modelo estrela, onde as tabelas dimensões são normalizadas;

**Modelo Constelação de Fatos** – neste tipo de modelo é possível ter mais de uma tabela fato de modo a compartilhar as dimensões existentes. Por esta razão também é vista como uma constelação de estrelas.

## 2.4 DATA MINING

Conforme descrevemos no início deste capítulo, data mining constitui uma importante etapa no processo de KDD, sendo tradicionalmente aplicada na busca de informações úteis em grandes repositórios de dados. Segundo Tan, Steinbach e Kumar (2006), podemos dividi-la em dois grandes tipos de tarefas: a) descritivas e b) preditivas.

### 2.4.1 Tarefas Descritivas:

Têm por objetivo derivar padrões que resumam os relacionamentos subjacentes nos dados. Seus principais propósitos são:

- *Análise de Associações*: utilizada para descobrir padrões que descrevem características fortemente associadas dentro dos dados. Os relacionamentos descobertos podem ser representados por uma técnica baseada na extração de regras de associação, cujo conceito é traduzido por uma expressão de implicação no formato  $X \rightarrow Y$ , onde X e Y são conjuntos

distintos dos itens (ex:  $X \cap Y = \emptyset$ ). A força de uma regra de associação pode ser medida pelo suporte e sua confiança (TAN; STEINBACH; KUMAR, 2006).

O suporte determina a frequência na qual uma regra é aplicável a um determinado conjunto de dados, enquanto que a confiança determina a frequência na qual os itens em Y aparecem em transações que contenham X. (TAN; STEINBACH; KUMAR, 2006). As definições formais destas métricas são:

$$\text{Suporte, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}; \quad \text{Confiança, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

Na figura 4 ilustramos um exemplo de análise de associações para os produtos de uma cesta básica.

TID	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Cola}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Cola}

Figura 4 – Exemplo de análise de associações  
Fonte: Tan, Steinbach e Kumar (2006)

{Fraldas  $\rightarrow$  Cerveja}

A regra acima sugere que há um relacionamento forte entre a venda de fraldas e a de cerveja porque muitos clientes que compram fraldas também compram cerveja. A partir deste tipo de regra, varejistas podem identificar novas oportunidades para vendas cruzadas dos seus produtos para os clientes (TAN; STEINBACH; KUMAR, 2009).

-*Análise de agrupamento*: procura dividir dados em grupos de itens que tem propriedades semelhantes. No comércio, por exemplo, o clustering pode ajudar a descobrir grupos distintos de clientes e caracterizar estes grupos com base no seu padrão de compra;

- *Detecção de anomalias*: é a tarefa de identificar observações cujas características são significativamente diferentes do resto dos dados. Tais observações são conhecidas como anomalias ou outliers. Aplicações de detecção de anomalias incluem: detecção de fraude, invasões de redes de computadores, padrões incomuns de doenças, e distúrbios no ecossistema.

## 2.4.2 Tarefas preditivas

Têm como objetivo prever o valor de um determinado atributo baseado nos valores de outros atributos. O atributo a ser previsto é comumente conhecido como a variável independente ou alvo, enquanto os atributos usados para fazer a previsão são conhecidos como variáveis independentes, explicativas ou classe.

Segundo Tan, Steinbach e Kumar (2006), as tarefas preditivas podem ser utilizadas para dois propósitos:

- *Classificação*: visa organizar objetos em uma entre diversas categorias pré-definidas, utilizando para tanto, técnicas implementadas sob algoritmos de aprendizagem. Este tipo de algoritmo utiliza o princípio da indução para aprender modelos de classificação a partir de um conjunto de registros com variável classe conhecida (conjunto de treinamento), e aplicá-los na dedução de outros registros, onde a variável classe é desconhecida (conjunto de teste).

A avaliação da qualidade de um modelo é baseada na contagem dos registros corretamente classificados no conjunto de teste. Estas contagens são tabuladas em uma tabela conhecida como matriz de confusão, onde o número total de previsões corretas é indicado pela diagonal principal da matriz.

		Classe Prevista	
		Classe = SIM	Classe = NÃO
Classe Real	Classe = SIM	4 ✓	1 ✗
	Classe = NÃO	2 ✗	3 ✓

Quadro 1 – Exemplo de Matriz de confusão para problema binário

No quadro 1 mostramos um exemplo de matriz de confusão, aplicada para resolução de um problema binário (sim, não) de classificação, como por exemplo, a incidência de uma determinada patologia bucal. Considerado 10 registros em um conjunto de teste, verificamos que destes, 4 foram corretamente classificados como SIM e 3 como NÃO (diagonal principal = 7). Na diagonal secundária verificamos os registros equivocadamente classificados, de modo que para nosso exemplo teríamos 1 registro classificado como NÃO (classe real=SIM) e 2 registros classificados como SIM (classe real=NÃO).

Adicionalmente, outras formas de avaliação podem se mostrar mais convenientes para avaliação de um modelo de classificação, como por exemplo, a adoção de métricas de



desempenho, onde a informação é resumida em um único número. Entre as métricas de desempenho mais frequentemente utilizadas estão:

- Precisão (Acurácia): Consiste na divisão do número de previsões corretas pelo número total de previsões, de modo que, para o exemplo ilustrado no quadro 12 teríamos uma precisão de 0,7 (7 / 10).
- Taxa de Erro: Consiste na divisão do número de previsões erradas pelo número total de previsões, de modo que, para o exemplo ilustrado no quadro 12 teríamos uma taxa de erro de 0,3 (3 / 10).

Entre as principais técnicas de classificação, destacamos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006), cuja estrutura é composta por uma hierarquia de nodos conectados por ramos. Segundo Han e Kamber (2001), cada nodo interno (ou não terminal) denota um teste lógico sobre um atributo (variável independente), no qual os ramos representam o resultado do teste, e cada nodo folha guarda um rótulo para classe (variável dependente).

Na figura 5 é apresentado um exemplo para aplicação de uma árvore de decisão (TAN; STEINBACH; KUMAR, 2006) extraído do estudo de Montenegro, Oliveira e Cabral (2008), onde o variável alvo considerado foi a incidência de cáries.

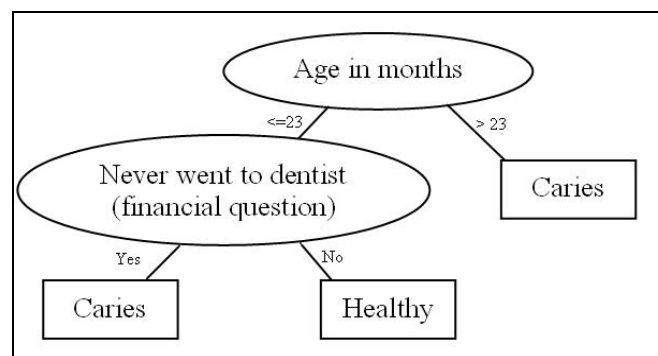


Figura 5 – Exemplo de árvore de decisão  
Fonte: Montenegro, Oliveira e Cabral (2008).

Conforme podemos observar na árvore gerada, pela interpretação do modelo podemos constatar uma provável relação entre idade e idas ao dentista com a incidência de cárie dentária. É importante destacar a importância deste tipo de informação, uma vez que sinalizam a necessidade de programas de saúde mais acessíveis a população de baixa renda.

- *Regressão*: Prevêem uma ou mais variáveis contínuas, como lucro ou perda, com base na análise de outros atributos do conjunto de dados. Por exemplo, prever as vendas do

próximo ano. Entre suas principais técnicas estão: Árvores de regressão, árvores modelo e aplicações estatísticas baseadas em regressão linear e logística.

## 2.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo apresentamos nossa revisão bibliográfica, dividindo-a em dois grandes tópicos. No primeiro, abordamos o uso de indicadores em saúde bucal como instrumentos de diagnóstico e planejamento, discutindo ainda sobre os principais tipos de indicadores e seus processos de formação.

No segundo tópico apresentado, abordamos o processo de KDD inicialmente definido por Fayad, Piatetsky e Smyth (1996) e adaptado por Han e Kamber (2001) com a inclusão da etapa de *data warehousing*. Discutimos também, a importância da preparação de dados e as principais tarefas de data mining apontadas por Tan, Steinbach e Kumar (2006). Para finalizar, apresentamos um exemplo prático de árvores de decisão (TAN; STEINBACH; KUMAR, 2006) aplicadas a um cenário de saúde bucal.

Com base no conhecimento obtido nesta fundamentação teórica, bem como na percepção do cenário de pesquisa, entendemos que uma solução ideal de KDD para gestão em Saúde Bucal deva contemplar necessariamente:

- **Robustez de ambiente:** Aplicado à gestão como um todo, uma solução ideal de KDD deve garantir suporte a manipulação de grandes conjuntos de dados. Neste caso, o desenvolvimento de ambientes de data warehousing mostra-se adequado a este tipo de necessidade, uma vez que tipicamente é projetado para fins analíticos.
- **Compreensibilidade dos modelos:** Embora a acurácia preditiva seja eminentemente almejada em grande parte dos projetos de KDD, é importante garantir que os modelos gerados neste processo sejam também compreensíveis ao entendimento outros profissionais. Desta forma, garantimos aos gestores de saúde melhores condições para entendimento do problema investigado, e conseqüentemente políticas de saúde melhores embasadas.
- **Tratamento de dados não convencionais:** Uma solução ideal de KDD na área de saúde bucal, igualmente deve ser adequada ao tratamento de dados não convencionais, como imagens, descrições textuais e dados temporais, tipicamente registrados ou anexados junto aos prontuários.

A partir das características apontadas, direcionamos a solução de KDD proposta para este trabalho.



### 3 DESENVOLVENDO UM PROCESSO DE KDD EM SAÚDE BUCAL

Embora a literatura de saúde bucal disponibilize uma série de trabalhos relacionados à exploração de modelos preditivos para as mais diversas patologias em saúde, boa parte destes trabalhos está eminentemente orientada a comparação de técnicas de data mining, desprezando aspectos importantes como a preparação dos dados, validação e interpretação dos resultados obtidos. Diferentemente destes trabalhos, a solução desenvolvida neste estudo visa dar o entendimento completo a este laborioso processo, fornecendo ao gestor em saúde bucal uma abordagem sistemática para apoio a tomada de decisão e formulação de políticas de saúde melhores direcionadas. Na figura 7, apresentamos o processo de KDD em saúde bucal proposto para este trabalho, dividindo-o em três grandes etapas:

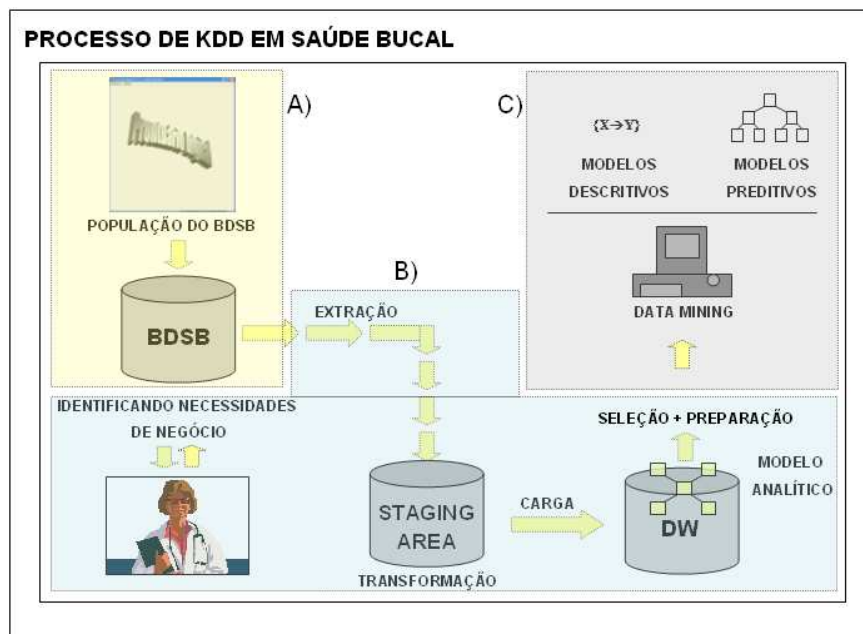


Figura 6 – Processo de KDD em Saúde Bucal

#### 3.1 DESENVOLVIMENTO DE UM AMBIENTE OLTP

Diferentemente de outros trabalhos realizados na área, o cenário encontrado junto à instituição parceira exigiu um grande esforço para viabilização de um processo de descoberta de conhecimento, uma vez que toda e qualquer informação era exclusivamente disponibilizada em prontuários de papel.

Desta forma, antes mesmo de se iniciar o processo de KDD, foi necessário criar um novo ambiente de armazenamento, organização e recuperação destes dados, o qual denominamos de BDSB (Banco de Dados em Saúde Bucal) (Etapa A).

Para tanto, o método adotado para a construção do BDSB foi baseado nas fases apontadas por Elmasri e Navathe (2006), conforme a ilustração simplificada da figura 7.

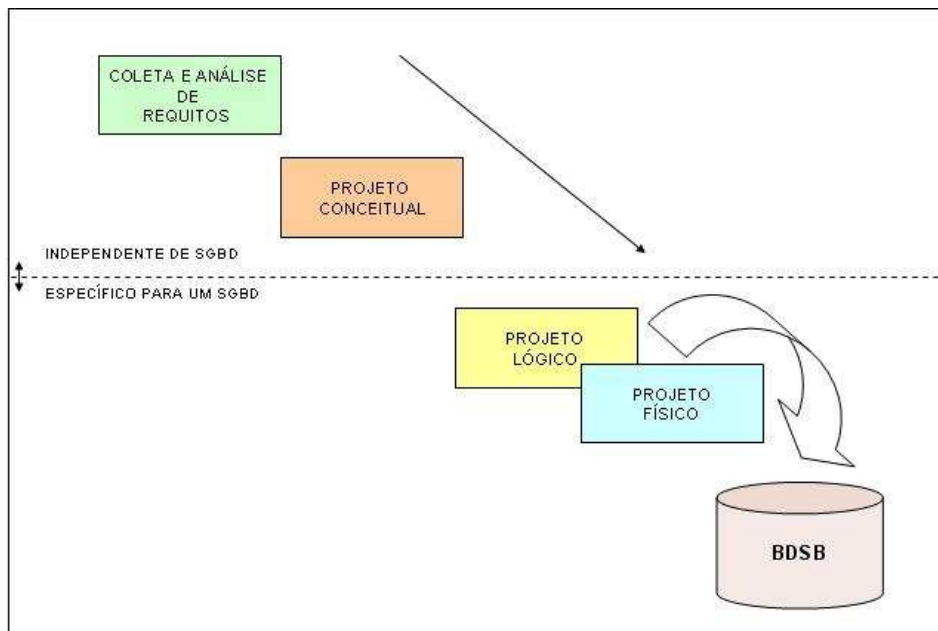


Figura 7 – Projeto de construção de banco de dados  
Fonte: Elmasri e Navathe (2006)

### 3.1.1 Identificação de Requisitos

A fase de coleta e análise de requisitos implicou na realização de entrevistas semi-estruturadas durante o período de três meses junto aos professores representantes das áreas de atendimento odontológico vinculadas à FO-PUCRS. Composto estas áreas estavam: o Centro de Extensão Universitária Vila Fátima (Unidade de Saúde SUS), o Hospital São Lucas (Serviço de Estomatologia), o CERLAP (Centro de Reabilitação de Fissuras Lábio-palatinas), Endodontia e a Área de Trauma Dento-Alveolar.

Nas entrevistas realizadas, foram coletadas informações referentes ao processo de armazenamento dos dados, abrangendo desde a abertura do prontuário até o tipo de dado gerado (textos, números, seqüências históricas e imagens). Ao final de cada entrevista, foi produzida uma ata (Apêndices B, C, D, E, F, G, H) contendo todas as informações coletadas e um diagrama de classes em notação UML (Unified Modeling Language) (RUMBAUGH; JACOBSON; BOOCH, 1999), onde descrevemos as principais entidades envolvidas no processo (consultas, anamneses, exames clínicos e complementares, diagnósticos e tratamentos).

### 3.1.2 Modelagem Conceitual

Depois de finalizada esta primeira fase, buscamos analisar a documentação gerada para cada uma das áreas, produzindo ao final desta atividade, um modelo abstrato de alto nível,

denominado modelo conceitual. O modelo conceitual é uma descrição concisa dos requisitos de dados (ELMASRI; NAVATHE, 2006), independente do SGBD (Sistema Gerenciador de Banco de Dados), e que inclui descrições detalhadas dos tipos de classes, atributos, associações e restrições.

Para o processo de modelagem conceitual optou-se pelo uso de um diagrama de classes em notação UML, conforme já utilizado na fase de coleta e análise de requisitos. Uma vez elaborado o modelo conceitual, foi validado junto aos professores envolvidos no estudo, executando-se ajustes na nomenclatura das classes e atributos do diagrama. O diagrama final, cujo objetivo foi integrar todas as áreas, tem 87 classes, 239 atributos e 97 associações, conforme ilustrado no Apêndice A.

### 3.1.3 Modelagem Lógica

Realizados os ajustes finais no modelo conceitual, partimos para o desenvolvimento do modelo lógico de dados, implementando-o no SGBD relacional Oracle pela linguagem SQL. Em SQL (Structured Query Language), convenções de nomenclatura como classe, atributo e associação são substituídas por tabelas, colunas e restrições de integridade.

No modelo lógico, foram definidas as estruturas de dados e seus detalhes de implementação, tais como, otimização do número de tabelas e a implementação de suas chaves primárias e estrangeiras. A Figura 8 ilustra a edição do modelo lógico construído para este estudo, através da ferramenta CASE (Computer Aided Software Engineering) DBDesigner 4 (Free Software Foundation, Boston, MA, USA).

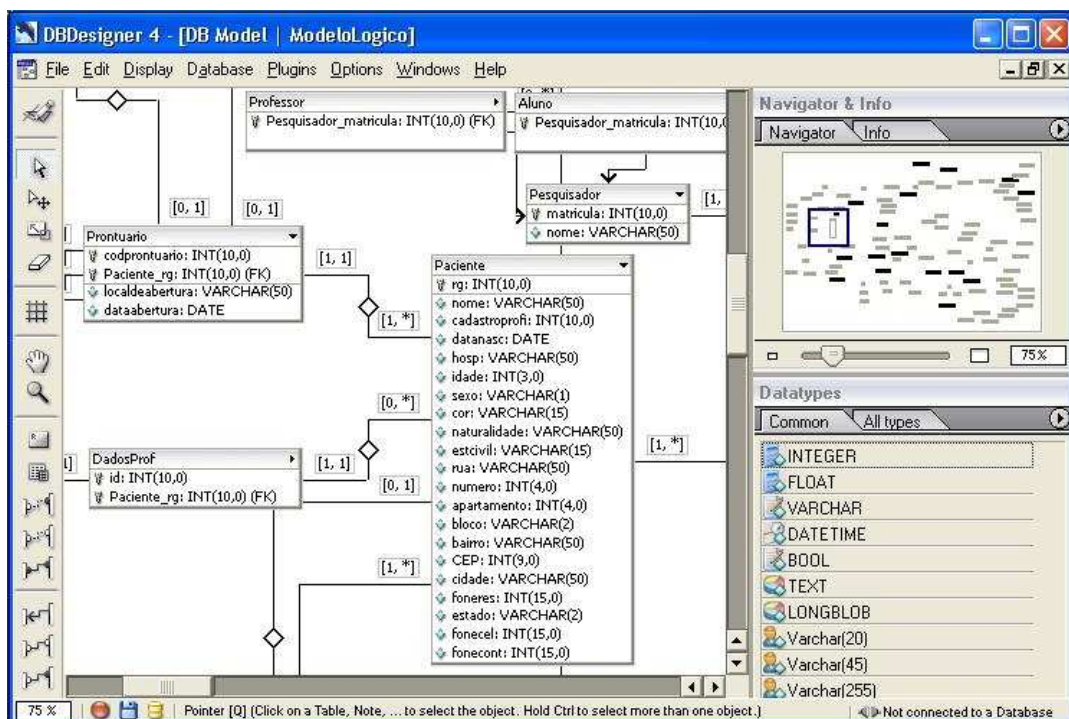


Figura 8 – Modelagem lógica do BDSB

De maneira complementar, outras modificações podem ser verificadas neste modelo em relação ao conceitual, como por exemplo, a inclusão da coluna Paciente\_rg como chave primária na tabela Prontuário. Tal modificação deve-se ao tipo de relacionamento (1 - \*, notação de cardinalidade Merise), entre as tabelas Prontuário e Paciente, ou seja, um registro da tabela Paciente pode ser vinculado a muitos (\*) registros da tabela Prontuário (no mínimo um), enquanto um registro da tabela Prontuário deve ser vinculado a um único registro da tabela Paciente (no mínimo um).

### 3.1.4 Implementação Física do BDSB

Ao final da modelagem lógica, utilizamos o módulo de exportação da ferramenta DBDesigner 4 para geração de um script na linguagem SQL com as especificidades do SGBD Oracle10.

Embora, boa parte dos autores de banco de dados inclua a modelagem física como fase independente no projeto de construção de banco de dados, muitas ferramentas CASE acabam tratando aspectos físicos, como estruturas de armazenamento e caminhos de acesso (índices), dentro da própria fase de modelagem lógica dos dados. Por este motivo, este trabalho adotou inicialmente a definição padrão da ferramenta CASE DBDesigner 4 para a composição do esquema físico do BDSB, em Oracle. Após a geração do script SQL, foi utilizada a ferramenta Oracle SQL Developer (Oracle Corporation, Redwood Shores, CA, USA) para sua importação e geração do BDSB, conforme ilustrado parcialmente na figura 9.

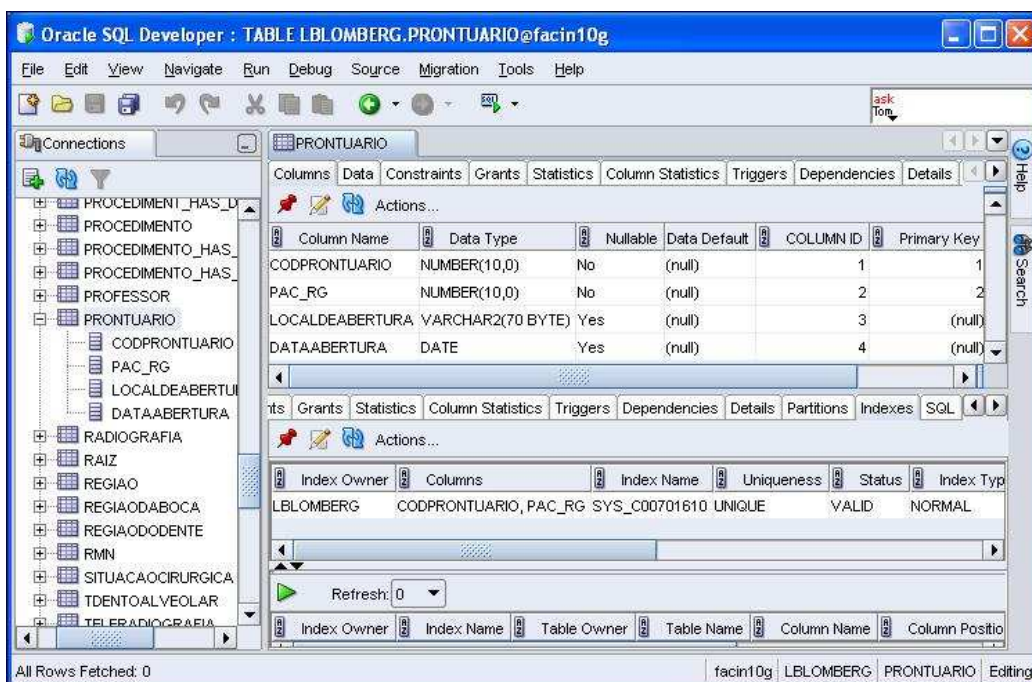


Figura 9 – Definições físicas do BDSB

Nesta mesma ilustração, podemos constatar alguns detalhes referentes à estrutura física dos dados da tabela prontuário como, por exemplo, a definição de chaves primárias, obrigatoriedade no preenchimento, e tipos de dados específicos do SGBD relacional Oracle 10g.

É importante destacar que trabalhos nesta área não comuns na literatura de saúde bucal. Os mais próximos encontrados foram os desenvolvidos por Taylor (2000; 2001; 2004), no entanto estes concentravam-se na estrutura lógica dos dados, e não sobre o projeto completo de construção de bancos de dados.

Tabela 1 - Considerações gerais sobre o Projeto de Construção do BDSB.

	<b>Taylor</b>	<b>Nossa Abordagem</b>
Proveniência dos dados	Dados provenientes de questionários clínicos armazenados em arquivos ASCII.	Dados provenientes de prontuários odontológicos.
Áreas de Atendimento	Clínica.	Clínica, Estomatologia, Trauma Dento-Alveolar, Fissuras Lábio Palatinas e Saúde Coletiva.
Fases do projeto de banco de dados implementadas	Lógica	Análise e coleta de requisitos, conceitual, lógica e física.*

\* Segundo Elmasri e Navathe (2006)

Construído o BDSB, selecionamos o CEU Vila Fátima como foco deste estudo. Esta escolha justificou-se pela relevância social dos dados lá gerados, conforme detalhado no primeiro capítulo. Na próxima seção, descrevemos o processo de apreensão de dados realizado junto a esta unidade.

### 3.1.5 Implementação de Interface para População do BDSB

Embora operacional, esta foi uma atividade de grande importância para viabilização da pesquisa. A partir da base operacional criada em Blomberg et al. (2009), foi desenvolvida uma aplicação em J2SE (Java 2 Standard Edition) cliente/servidor para o processo de população do BDSB. Na figura 10, ilustramos a interface da aplicação desenvolvida.



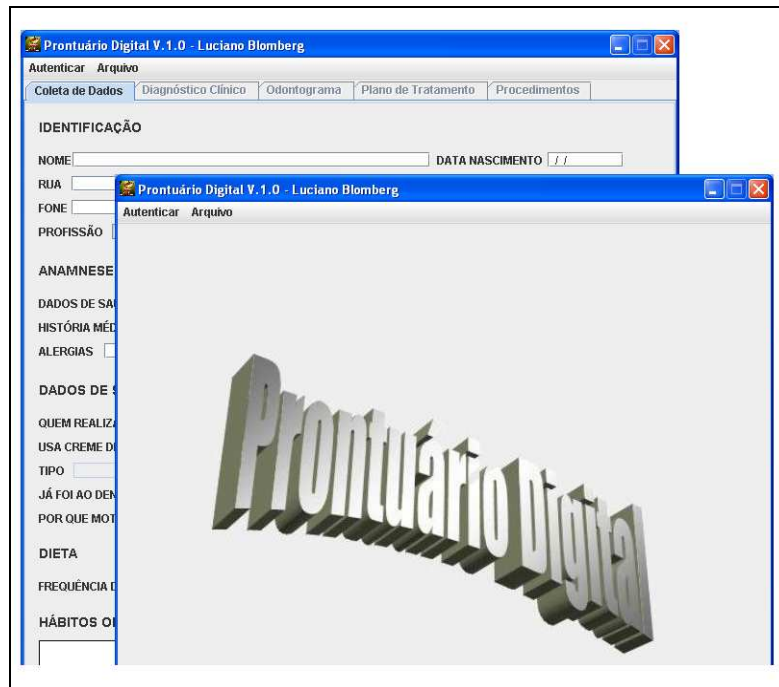


Figura 10 – Interface para a população do BDSB

A interface implementada contemplava 90% dos campos do prontuário odontológico utilizado no CEU Vila Fátima, diferenciando-se apenas pela ausência dos campos de índice de sangramento gengival.

### 3.1.6 Apreensão dos dados

Uma vez desenvolvida a aplicação, foi realizado um levantamento do volume de pastas familiares a serem apreendidas junto ao CEU Vila Fátima, onde se constatou o volume de 2089 pastas familiares, organizadas sequencialmente pela ordem de criação. Estabelecido um protocolo para retirada e entrega das fichas extraídas das pastas, foi iniciada a atividade de digitação, envolvendo inicialmente 4 bolsistas do curso de odontologia da FO-PUCRS durante o período de 8 semanas (11/5/2009 a 01/07/2009). Para tanto, foi necessário um esforço para a montagem de um ambiente de apreensão, solicitando junto às Faculdades de Informática e Odontologia autorização para deslocamento de equipamentos (Anexos A, B). Da mesma forma, foi necessário solicitar a autorização da Direção do CEU Vila Fátima para retirada / transporte de fichas odontológicas para o campus central da PUCRS.

Como resultado desta primeira etapa de apreensão de dados apresentamos:

<b>Volume total de pastas familiares: 2089</b>				
Pastas Familiares verificadas: 200 (9,57% do total)				
Período de apreensão: 8 semanas				
Total de				
Pacientes	Fichas	Consultas	Diagnósticos Clínicos	Procedim.
199	216	670	54	1044

Quadro 2 – Resultados da primeira etapa da apreensão de dados

Em um segundo momento (16/9/2009 a 23/10/2009), passamos a contar com a colaboração de 5 estagiários de odontologia da FO-PUCRS financiados pelo projeto PRÓ-SAÚDE do governo federal. Seguindo o mesmo protocolo de retirada e entrega de fichas odontológicas adotado anteriormente, dividimos a equipe em dois grupos conforme o grau de conhecimento do estagiário (considerado pelo semestre cursado). O primeiro ficou encarregado de lançar os dados de identificação, anamnese e diagnóstico clínico do paciente, enquanto o segundo grupo digitava o odontograma, plano de tratamento e procedimentos odontológicos realizados. Estes procedimentos foram divididos nas seguintes categorias: a)Primeira Consulta, b)acolhimento, c)procedimento preventivo, d)tratamento restaurador, e)tratamento cirúrgico, f)tratamento periodontal, g)tratamento protético, h)tratamento ortodôntico, i)tratamento endodôntico, j)consulta de manutenção, k)intercorrência, l)urgência, exame realizado, m)prescrição de medicamento, n)encaminhamento, o)alta clínica e p)exame solicitado. Para a definição destas categorias, contamos com o apoio de duas especialistas em saúde bucal vinculadas a instituição parceira.

Ao final deste processo, chegamos a uma amostra de 675 pastas familiares verificadas (31,32% do total), volume que entendemos adequado aos propósitos almejados neste trabalho. No quadro 2, apresentamos maiores detalhes sobre o volume apreendido.

<b>Volume total de pastas familiares: 2089</b>				
Pastas Familiares verificadas: 675 (31,32% do total)				
Período de apreensão: 5,4 semanas				
Total de				
Pacientes	Fichas	Consultas	Diagnósticos Clínicos	Procedim.
598	642	2031	168	3103

Quadro 3 – Resultados da segunda etapa da apreensão de dados

Com o término da etapa de apreensão de dados, pudemos observar algumas peculiaridades referentes à baixa qualidade dos dados registrados nas fichas odontológicas. Entre elas destacamos:

- Campos sem preenchimento, conforme ilustrado nos quadros 3, 4, 5, 6 e 7:

<b>Anamnese</b>	<b>642</b>	<b>100%</b>
Preenchidos	343	53.4%
Em branco	299	46.6%

Quadro 4 – Campos sem preenchimento (Anamnese)

<b>Dados de Saúde Bucal</b>	<b>642</b>	<b>100%</b>
Preenchidos	427	66.5%
Em branco	215	33.5%

Quadro 5 – Campos sem preenchimento (Dados de saúde bucal)

<b>Diagnóstico Clínico</b>	<b>642</b>	<b>100%</b>
Preenchidos	168	26.2%
Em branco	474	73.8%

Quadro 6 – Campos sem preenchimento (Diagnóstico clínico)

<b>Plano de Tratamento</b>	<b>642</b>	<b>100%</b>
Preenchidos	227	35.6%
Em branco	415	64.4%

Quadro 7 – Campos sem preenchimento (Plano de tratamento)

<b>Odontograma</b>	<b>642</b>	<b>100%</b>
Preenchidos	235	36.6%
Em branco	407	63.4%

Quadro 8 – Campos sem preenchimento (Odontograma)

Do mesmo modo, constatamos que a ausência de informação nos campos acima ilustrados, limitava não apenas a compreensão do histórico odontológico do paciente, mas também o cruzamento de indicadores bucais com os fatores epidemiológicos capturados nos campos de anamnese e de saúde bucal. Outrossim, grande parte dos dentes hígidos era omitida nos odontogramas, criando uma certa confusão entre o estado hígido e o estado não informado. Para se ter uma dimensão deste prejuízo, a amostra considerada para geração do indicador de cárie dentária foi dois terços menor do que a amostra original. Isto se deve ao baixo percentual de preenchimento do odontograma, bem como, a imprecisão do registro de eventos temporais.

- Entre as imprecisões temporais destacamos aquelas referentes às datas de nascimento e procedimentos odontológicos, conforme ilustrado nos quadros 8 e 9:

<b>Pacientes</b>	<b>598</b>	<b>100%</b>
Com data de nascimento	583	97.5%
Sem data de nascimento	15	2.5%

Quadro 9 – Campos sem registro temporal (Data nascimento)

<b>Fichas Odontológicas</b>	<b>642</b>	<b>100%</b>
Com data de abertura	542	84.4%
Sem data de abertura	100	15.6%

Quadro 10 – Campos sem registro temporal (Data de abertura da ficha)

Como pudemos observar no capítulo 2, a precisão do registro temporal é de extrema importância para a gestão de indicadores, uma vez que boa parte destes utilizam-se da consideração de faixas-etárias e períodos de tempo.

- Encontramos também, erros de nomenclatura, como por exemplo o emprego do termo “restauração” para aplicação de materiais provisórios, quando o correto seria registrar um “selamento”. Em outras oportunidades, encontramos também a utilização do termo “conserto” em detrimento da restauração do dente.
- Outro tipo comum de erro foi o preenchimento em local indevido, como por exemplo, o registro do plano de tratamento (ex: exodontia) no odontograma e o registro de diagnóstico clínico dentro dos procedimentos odontológicos realizados

- Da mesma forma, campos de duplo entendimento dificultaram a compreensão do que era registrado, como por exemplo, a questão aplicada para identificar se o paciente já tinha ido alguma vez ao dentista e o motivo desta consulta. Frequentemente foram os casos onde os pacientes alegaram nunca ter ido ao dentista, mas mesmo assim informaram dor como motivo da ida. Fato que nos leva a questionar: dor é a razão do não comparecimento ou a razão do atendimento atual?

Mediante a baixa qualidade dos dados apreendidos, optamos por reduzir o tamanho da amostra, considerando apenas os pacientes com odontogramas preenchidos e sem imprecisões temporais nas datas de nascimento e procedimentos odontológicos. Deste modo, trabalhando com os dados mais homogêneos, acreditamos obter melhores resultados na execução da etapa de *data mining*. No quadro 10, ilustramos a nova amostra considerada para este trabalho.

	AMOSTRA ORIGINAL	IMPRECISÕES TEMPORAIS / AUSÊNCIA DE INFORMAÇÃO	AMOSTRA CONSIDERADA
PACIENTES	598	125	473
CONSULTAS	2031	425	1606
PROCEDIMENTOS	3103	579	2524
ODONTOGRAMAS	642	468	174

Quadro 11 – Nova amostra considerada

Com o término do processo de apreensão de dados, finalizamos a primeira etapa (A) de nossa solução. Adicionalmente, a partir desta discussão, entendemos ter clarificado as dificuldades encontradas, conforme apresentado no Capítulo 1.2.

## 3.2 DESENVOLVIMENTO DE UM AMBIENTE OLAP

### 3.2.1 Identificando os requisitos de negócio

A partir da amostra apreendida, iniciamos a segunda etapa (B) de nossa solução, implementando um processo de *data warehousing*. Deste modo, começamos definindo as questões de negócio junto ao especialista de saúde bucal, utilizando para tanto uma entrevista semi-estruturada.

As questões de negócio identificadas seguem abaixo.

#### Primeira Questão de Negócio

**Necessidade de negócio:** identificar na população atendida o tratamento realizado em função da condição dental.

**Informação gerada:** percentual dos principais tratamentos realizados considerados por faixa-etária, sexo e tempo entre consultas.

**Importância da informação:** identificar intervenções ao longo do tempo que resultem na manutenção ou perda dentária.

**Considerações do Entrevistador:** este tipo de tarefa é plenamente realizável dentro da perspectiva de OLAP. Para tanto relacionaremos os tipos de tratamento realizados com a condição dentária, aspectos sociais e epidemiológicos do paciente.

### Segunda Questão de Negócio

**Necessidade de negócio:** identificar os indicadores de cárie dentária, bem como o indicador de cobertura da primeira consulta programática para a população da Vila Fátima.

**Informação gerada:** índice CPO-D, CEO-D e de cobertura da primeira consulta programática.

**Importância da informação:** na tomada de ações de planejamento, gestão e avaliação de políticas voltadas à melhoria da saúde bucal.

**Considerações do Entrevistador:** tais indicadores podem ser perfeitamente gerados a partir dos dados presentes no odontograma e na seção de registro de procedimentos odontológicos no entanto, para este trabalho temos como foco os indicadores relacionados a patologias bucais.

### Terceira Questão de Negócio

**Necessidade de negócio:** identificar na população atendida, a saúde ou não saúde bucal, onde são definidos como saudáveis os pacientes sem nenhuma patologia bucal ou indicação de tratamento.

**Informação gerada:** percentual de pacientes saudáveis e não-saudáveis considerados por faixa etária, sexo, grupo familiar, tempo, ida ao dentista, hábitos alimentares e hábitos de higiene.

**Importância da informação:** na definição e planejamento das ações no nível de serviço; melhorar acesso preventivo de acordo com o perfil do paciente; qualificação da atenção; melhor custo benefício, ou seja, o custo em manter o paciente saudável é menor do que tratar a doença.

**Considerações do Entrevistador:** Dentro da perspectiva OLAP, este tipo de consulta poderia ser realizada pela verificação do percentual de pacientes com algum procedimento de tratamento definido. No entanto, o baixo percentual de altas clínicas registradas nos impossibilitou de ter

uma melhor compreensão real do estado de saúde do paciente, uma vez que este é variável no tempo.

#### **Quarta Questão de Negócio**

**Necessidade de negócio:** identificar principais doenças em saúde bucal.

**Informação gerada:** percentual de incidência das principais doenças em saúde bucal consideradas por faixa etária, sexo, hábitos alimentares, dados de saúde geral e vínculos familiares.

**Importância da informação:** na identificação de associações de doenças que incidam paralelamente, potencializando assim, a ação de conduta de vários profissionais.

**Considerações do Entrevistador:** mediante a forma textual de uma parte dos registros de diagnóstico clínico nas fichas odontológicas do CEU Vila Fátima, avaliamos que este tipo de análise requereria um tratamento prévio para categorização destes diagnósticos. Desta forma, optamos neste momento por analisar apenas a patologia de cárie dentária, cujos dados estão melhores documentados no prontuário odontológico.

Com base nos requisitos identificados na entrevista, foi construído um modelo analítico do tipo constelação de fatos para resolução das duas primeiras questões de negócio.

#### **3.2.2 Construindo um Modelo de Análise de Dados**

Embora o modelo estrela seja mais comumente utilizado, optamos por adotar o modelo constelação de fatos pela conveniência no aproveitamento das dimensões e pela visão mais abrangente da resolução das questões de negócio apontadas pelo especialista. Na figura 11, ilustramos o modelo analítico gerado.

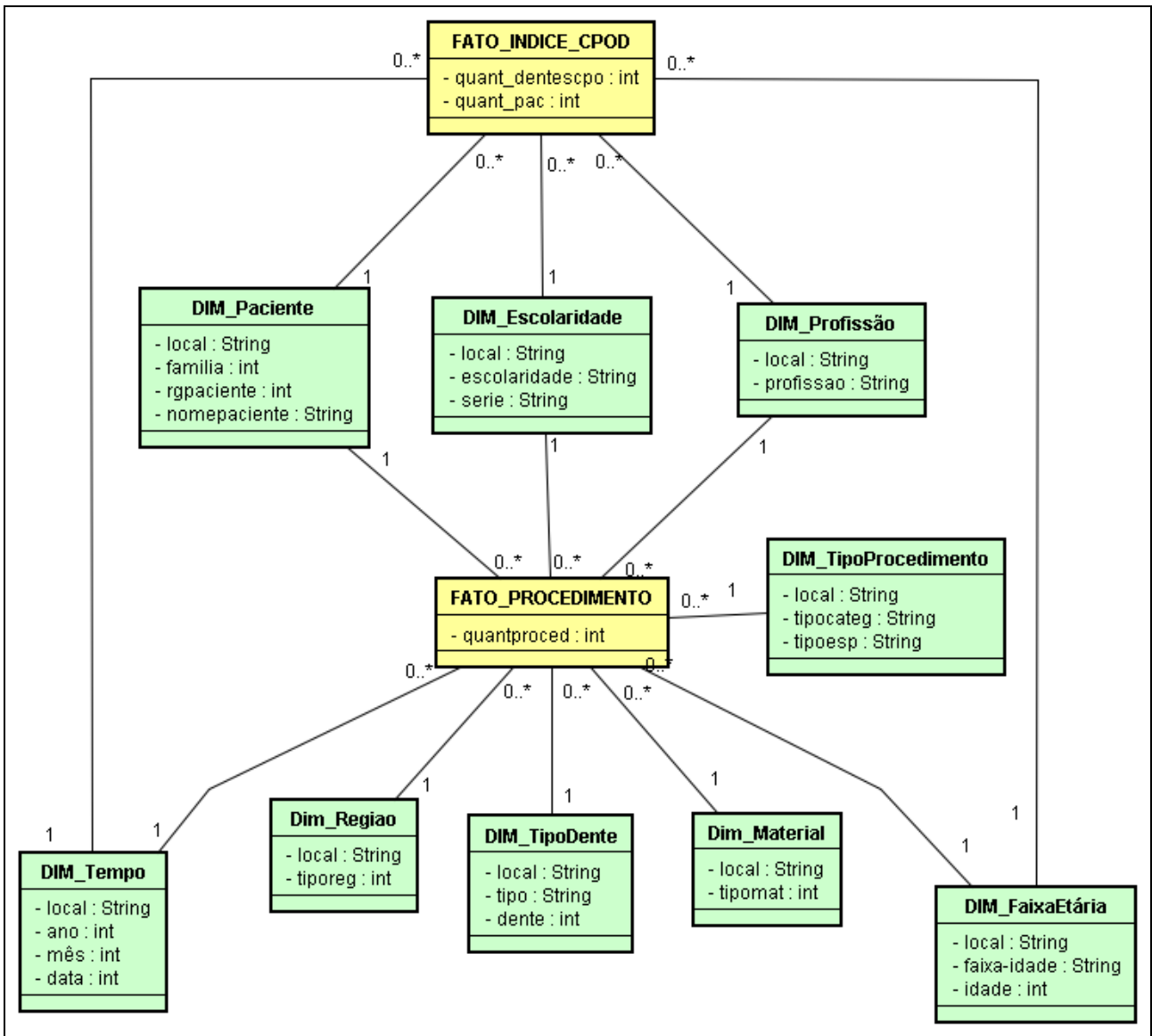


Figura 11 – Modelo analítico constelação de fatos

Para melhor compreensão do modelo produzido, ilustramos nos quadros 11 e 12 os critérios utilizados para composição das tabelas fato e dimensões modeladas.

1. Primeira Questão de Negócio

FATO	MEDIDA	PROPÓSITO	DIMENSÃO	PROPÓSITO
Fato_Procedimento	QuantProced	Quantificar o número de procedimentos mediante diferentes perspectivas.	Dim_Paciente	Definir características Sociais
			Dim_Escolaridade	
			Dim_Profissão	
			Dim_Tempo	
			Dim_FaixaEtaria	
			Dim_TipoProcedimento	Definir características odontológicas
			Dim_TipoDente	
			Dim_TipoRegião	
			Dim_TipoMaterial	

Quadro 12 – Primeira Questão de Negócio

## 2. Segunda Questão de Negócio

FATO	MEDIDA	PROPÓSITO	DIMENSÃO	PROPÓSITO
Fato_Índice_CPOD	Quant_dentesCPO	Quantificar o número de dentes CPO.	Dim_Paciente	Definir características Sociais
			Dim_Escolaridade	
			Dim_Profissão	
	Quant_Pac	Quantificar o número de pacientes com dentes CPO.	Dim_Tempo	Definir características sociais e perspectivas de ano e idade para formulação do índice CPO-D
		Dim_FaixaEtaria		

Quadro 13 – Segunda Questão de Negócio

Conforme relatamos nas considerações da entrevista, a resolução das demais questões de negócio ainda dependem de um melhor tratamento do texto lançado no prontuário odontológico. Para este tipo de problema, acreditamos que técnicas mais avançadas de PLN (Processamento da Linguagem Natural) possam agregar maior qualidade no reconhecimento automático destes textos. Desta forma, guiamos nossa modelagem pelo atendimento das duas primeiras questões de negócio.

### 3.2.3 Extraíndo os dados do BDSB

Definidas as questões de negócio, demos continuidade ao processo de *data warehousing* realizando para tanto a extração dos dados do BDSB. Para esta finalidade utilizamos a ferramenta SQL Developer, pela qual obtivemos um novo script SQL com os dados de 26 tabelas inerentes ao cenário do CEU Vila Fátima. Posteriormente, estes dados foram transferidos e tratados em uma staging area, conforme descrevemos na próxima seção.

### 3.2.4 Transformação dos dados

Na etapa de transformação compreendida pelo processo de *data*

```
--TABLE ESCOLARIDADE
--Campo série

UPDATE ESCOLARIDADE SET serie='Jardim' WHERE serie='jardim';
UPDATE ESCOLARIDADE SET serie='Analfabeto' WHERE serie='analfabeto';
UPDATE ESCOLARIDADE SET serie='1 s' WHERE serie='1S' or serie='1s';
UPDATE ESCOLARIDADE SET serie='2 s' WHERE serie='2s';
UPDATE ESCOLARIDADE SET serie='3 s' WHERE serie='3s';
UPDATE ESCOLARIDADE SET serie='3 a' WHERE serie='3A';
UPDATE ESCOLARIDADE SET serie='4 s' WHERE serie='4 s';
UPDATE ESCOLARIDADE SET serie='5 s' WHERE serie='5s';
UPDATE ESCOLARIDADE SET serie='8 s' WHERE serie='8 s';
UPDATE ESCOLARIDADE SET serie='Não Informado' WHERE NVL(serie,1)='1';

CREATE TABLE DIM_ESCOLARIDADE
AS
(SELECT DECODE (es.serie, 'Não Informado' , 'Não Informado',
                '1 s', 'Ensino Fundamental',
                '2 s', 'Ensino Fundamental',
                '3 s', 'Ensino Fundamental',
                '4 s', 'Ensino Fundamental',
                '5 s', 'Ensino Fundamental',
                '6 s', 'Ensino Fundamental',
                '7 s', 'Ensino Fundamental',
                '8 s', 'Ensino Fundamental',
                '1 a', 'Ensino Médio',
                '2 a', 'Ensino Médio',
                '3 a', 'Ensino Médio',
                'Jardim', 'Pré-Escola',
                'Analfabeto', 'Analfabeto') AS escolaridade,
es.serie,
pt.localdeabertura

FROM
paciente pa, prontuario pt, escolaridade es
WHERE
pa.ig=pt.pac_ig and pt.codprontuario=es.pro_codpr);
```

Figura 12 – Script SQL para transformação dos dados



*warehousing*, começamos a realizar as primeiras atividades de preparação de dados. Para tanto, utilizamos novamente scripts SQL, no entanto, dirigidos a correção de erros de digitação, tratamento de valores ausentes, categorização de valores não padronizados e criação de dimensões com valores agregados. Na figura 12, apresentamos parte do script utilizado para tratamento dos dados da tabela “Escolaridade”.

Adicionalmente, mediante a grande quantidade de erros ortográficos encontrados nas tabelas com campos descritivos (ex: *HabitosOrais*), tivemos a necessidade de adotar práticas de *stemming* dentro de nossos scripts de preparação de dados, ou seja, orientamos o reconhecimento dos hábitos orais pela busca dos radicais das palavras orientados por assunto. Segundo Orenge (2001), *stemming* pode ser definido como o processo de combinação de diferentes formas de uma palavra em uma representação comum, a raiz.

### **3.2.5 Carga dos Dados**

Após tratarmos os dados da base operacional em uma *staging area* e gerarmos nosso modelo multidimensional, extraímos um novo script SQL para a carga no *Data Warehouse*.

### **3.2.6 Geração do Cubo de Dados**

A partir dos dados carregados no *Data Warehouse*, utilizamos a ferramenta OLAP da Oracle AWM (Analytical Workspace Manager) para gerenciamento de objetos multidimensionais e geração do cubo de dados. A escolha desta ferramenta foi motivada por sua facilidade de uso e acesso livre, sendo inclusive, adotada institucionalmente pela Faculdade de Informática da PUCRS nas disciplinas de graduação e pós-graduação.

### **3.2.7 Explorando o Cubo para Atender as Questões de Negócio**

Na busca de uma melhor compreensão para a resolução da primeira questão de negócio, começamos gerando um relatório inicial com todos tipos de procedimentos odontológicos realizados no CEU Vila Fátima. Constatamos então, uma predominância dos procedimentos de tratamento (43.11%), seguidos pelos procedimentos preventivos (23.53%) e procedimentos de manutenção (7.68%) conforme ilustrado na figura 13.

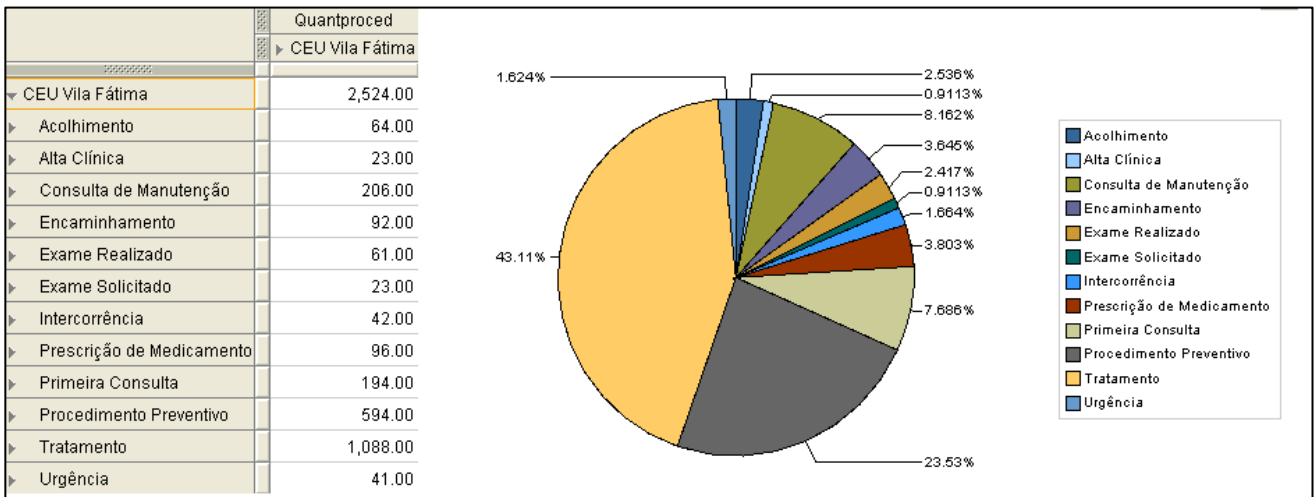


Figura 13 – Gráfico de procedimentos odontológicos

Analisando especificamente os procedimentos de tratamento realizados, produzimos novos relatórios onde verificamos os tipos de tratamento mais frequentes e seus relacionamentos com outros fatores sociais e odontológicos. Nas figuras 14 e 15, podemos observar que do total de 1088 procedimentos de tratamentos, 574 eram restauradores, e destes 240 utilizaram predominantemente os seguintes materiais restauradores: a) Ionômero (240), b) IRM (49) e c) Cimpat (4).

Considerando que Ionômero (para dentes permanentes), IRM e Cimpat são materiais restauradores provisórios, chegamos à conclusão que mais de um terço (33.44%) destes tratamentos são provisórios. Esta é uma informação que nos chamou a atenção para uma possível prática de protelação, sendo inclusive corroborada por especialistas em saúde bucal. Em outras palavras, podemos dizer que um mesmo paciente chegou a realizar dois ou mais procedimentos com materiais provisórios até que fosse feito um definitivo. Neste ponto, questionamos: por que não resolver logo o problema? Este é um fato que certamente merece a atenção de seus gestores, uma vez que gera um custo desnecessário à instituição.

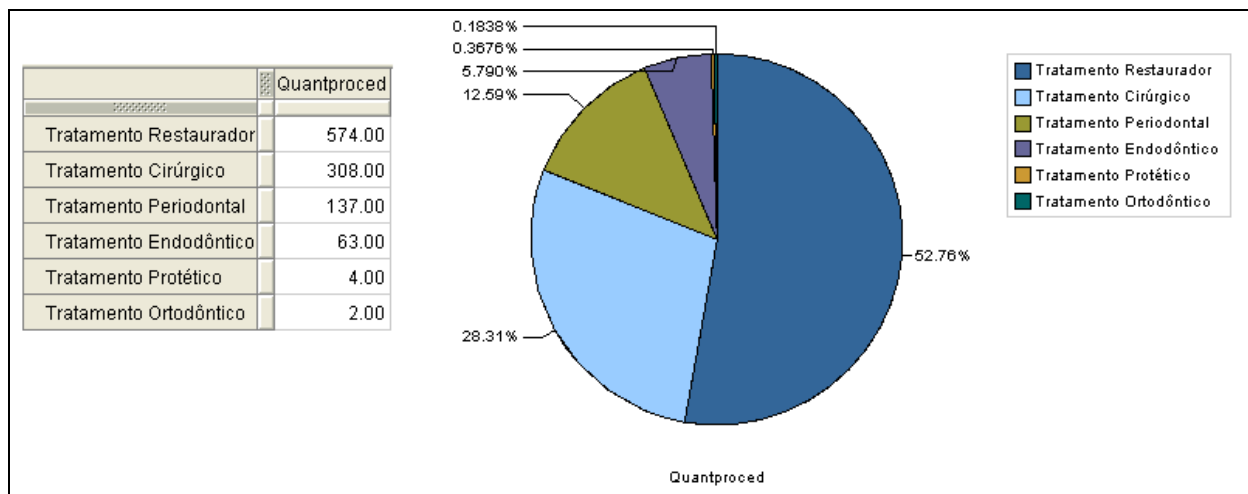


Figura 14 – Gráfico de Tratamentos odontológicos

		Quantproced			
		CEU Vila Fátima	Decíduo	Invalído	Permanente
Tratamento Restaurador	Ionômero	240.00	101.00	8.00	131.00
	Resina	171.00	8.00	4.00	159.00
	Não Informado	88.00	23.00	2.00	63.00
	IRM	49.00	14.00	2.00	33.00
	Amálgama	21.00	2.00		19.00
	Cimpat	4.00	3.00		1.00
	Óxido de Zinco	1.00			1.00

Figura 15 – Gráfico de Materiais dentários

Adicionalmente, a ferramenta OLAP AWM nos permitiu elaborar outros relatórios, relacionando os procedimentos de tratamentos realizados com fatores sociais, conforme almejado inicialmente. No relatório ilustrado na figura 16, mostramos os tratamentos realizados em função dos três tipos de profissões (ocupações) mais frequentes (Estudante, do lar e serviços gerais).

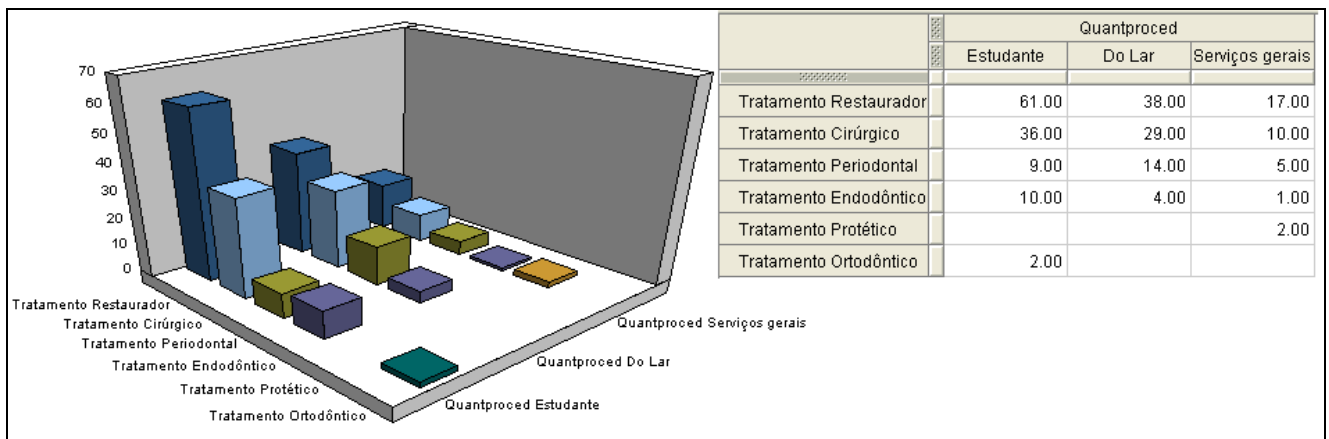


Figura 16 – Gráfico dos tratamentos x profissões

Para a resolução da segunda questão de negócio, utilizamos as dimensões DIM\_Paciente, DIM\_Tempo e DIM\_FaixaEtaria, cuja combinação com a medida CPO-D, forneceu-nos um indicador moderado (4,34) de cárie dentária. No entanto, este indicador reflete exclusivamente a média dos anos, e não um conjunto fechado, como por exemplo, a análise do índice para um ano específico, embora contemplado pelo modelo. Esta limitação deve-se a forma sequencial de lançamento das pastas familiares no BDSB, cuja organização física, não nos permitiu lançá-las por anos fechados.

De qualquer modo, temos neste momento, uma visão global do estado de saúde dos pacientes, pela qual pudemos observar tendências relacionadas ao comportamento e evolução da patologia dentária em diferentes grupos (familiar, etário, escolaridade, profissão). Na figura 17 ilustramos o índice CPO-D, visto por faixa-etária, onde constatamos uma maior severidade da

doença nos grupos 30-39, 40-49 e 50-59 anos. Através de outros relatórios gerados pela ferramenta AWM, verificamos também que os grupos com maior número de atendimentos eram formados por estudantes, donas de casa e auxiliares de serviços gerais cuja, escolaridade predominante era o ensino fundamental.

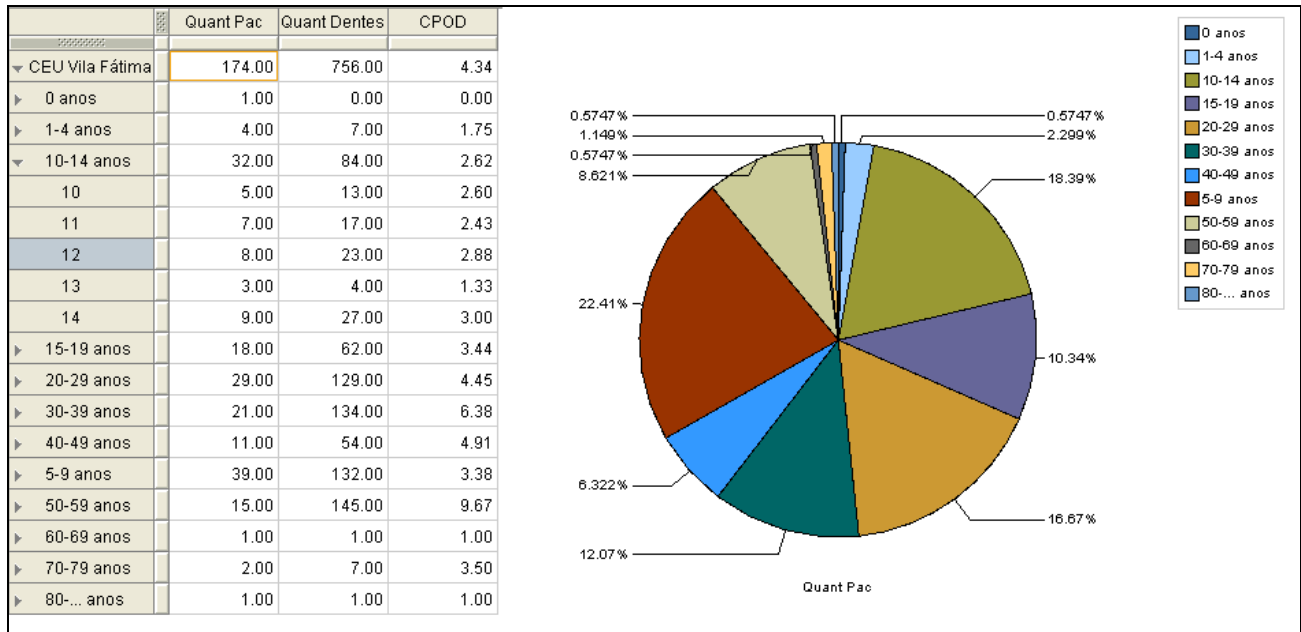


Figura 17 – Índice CPO-D por faixa-etária

### 3.3 DATA MINING

A partir da criação do *data warehouse*, começamos a terceira etapa de nossa solução (C), pela qual utilizamos os dados carregados do *data warehouse* e como base para criação de uma nova tabela de dados (ExpMiner01). Sua população era composta inicialmente por 174 registros provenientes da união das tabelas dimensões com as seguintes tabelas da base operacional: a)Anamnese, b)DadosSaudeBucal, c)Diagnósticos e d)HábitosOrais (Apêndice A). É importante ressaltar, que embora tivéssemos originalmente 598 pacientes, apenas 174 destes tinham alguma informação registrada em seu odontograma. Deste modo, executamos a etapa de *data mining* sob este conjunto de dados, sendo estes, diretamente relacionados a produção do índice de cárie dentária.

Na próxima seção, descrevemos as atividades implementadas para preparação dos dados, bem como as técnicas de *data mining* utilizadas para extração de modelos preditivos aplicados a análise de patologias periodontais, má-oclusão e indicadores de cárie dentária.

#### 3.3.1 Extração de modelos preditivos

Para extração dos modelos preditivos, optamos pelo uso de árvores de decisão (TAN; STEINBACH; KUMAR, 2006), uma vez que este tipo de técnica tem se mostrado mais adequada a

extração de modelos compreensíveis. Deste modo, iniciamos a etapa de *data mining* adotando a ferramenta *freeware* Weka 3.7.0, pela qual realizamos três experimentos.

### 3.3.1.1 Primeiro Experimento

Para realização do primeiro experimento, utilizamos a ferramenta SQL Developer para exportar os dados da tabela ExpMiner01 e gerar um novo arquivo (ExpMiner01.CSV) de dados reconhecível pela ferramenta Weka 3.7.0. Este arquivo foi submetido ao algoritmo J48 para inicialmente prever os seguintes atributos classe: incidência de cárie dentária (sim, não) e indicador de cárie dentária (muito baixo, baixo, moderado, alto, muito alto). Para melhor compreensão deste experimento, optamos por relatar os modelos obtidos e posteriormente sua validação.

#### a) Modelos Obtidos

Na análise de incidência de cárie dentária, fizemos uma série de tentativas para obtenção de árvores significativas, reduzindo manualmente a dimensionalidade de 105 para 11 atributos. O critério utilizado para escolha destes atributos foi baseado nos seguintes fatores:

- Conhecimento de domínio;
- Relevância das árvores geradas, de forma a eliminar árvores semanticamente ruins;
- Eliminação de atributos com baixa frequência.

Como melhor resultado obtido, geramos uma árvore de 11 nodos e com alto índice de instâncias corretamente classificadas (91.95%), conforme detalhado no Apêndice I. Na figura 18 apresentamos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada.

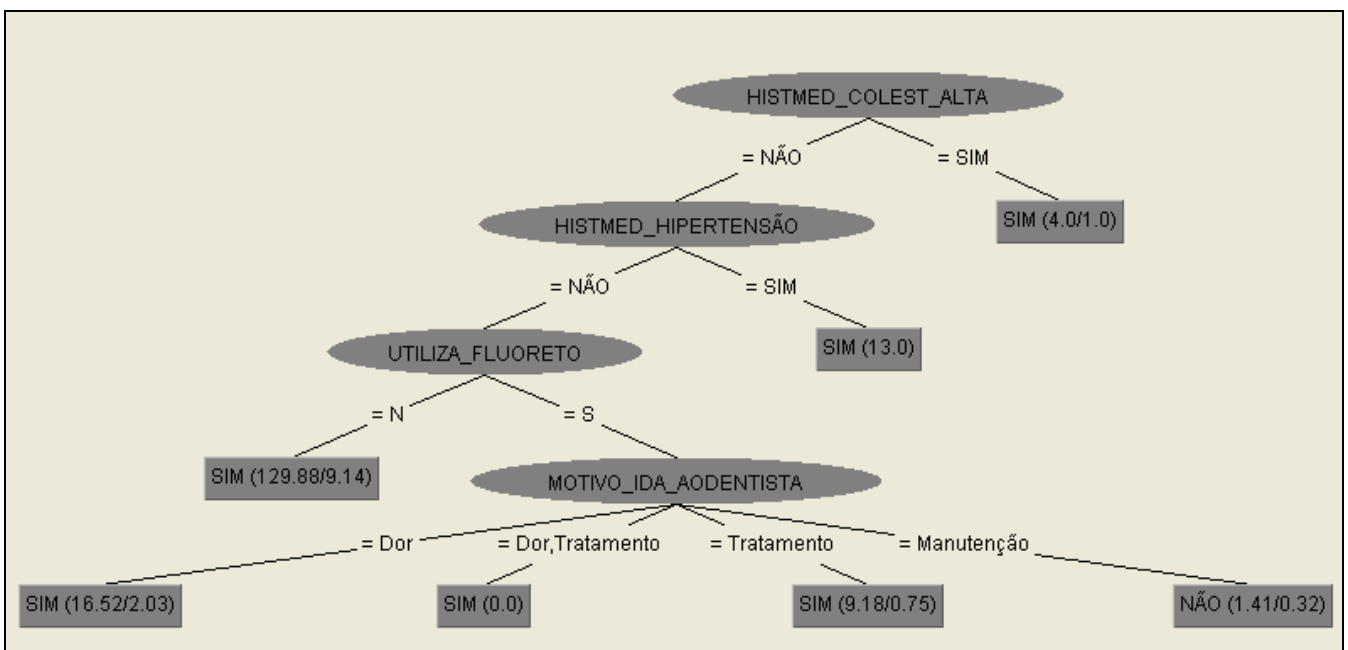


Figura 18 – Primeiro Experimento – MP I

Para análise do indicador de cárie dentária, utilizamos o mesmo arquivo de dados CSV e o mesmo critério de eliminação manual dos atributos, considerando, no entanto, 12 campos. Diferentemente do modelo anterior, tivemos como melhor resultado uma árvore de 13 nodos e com baixo índice de instâncias corretamente classificadas (28.16%), conforme detalhado no Apêndice J. Na figura 19 apresentamos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada.

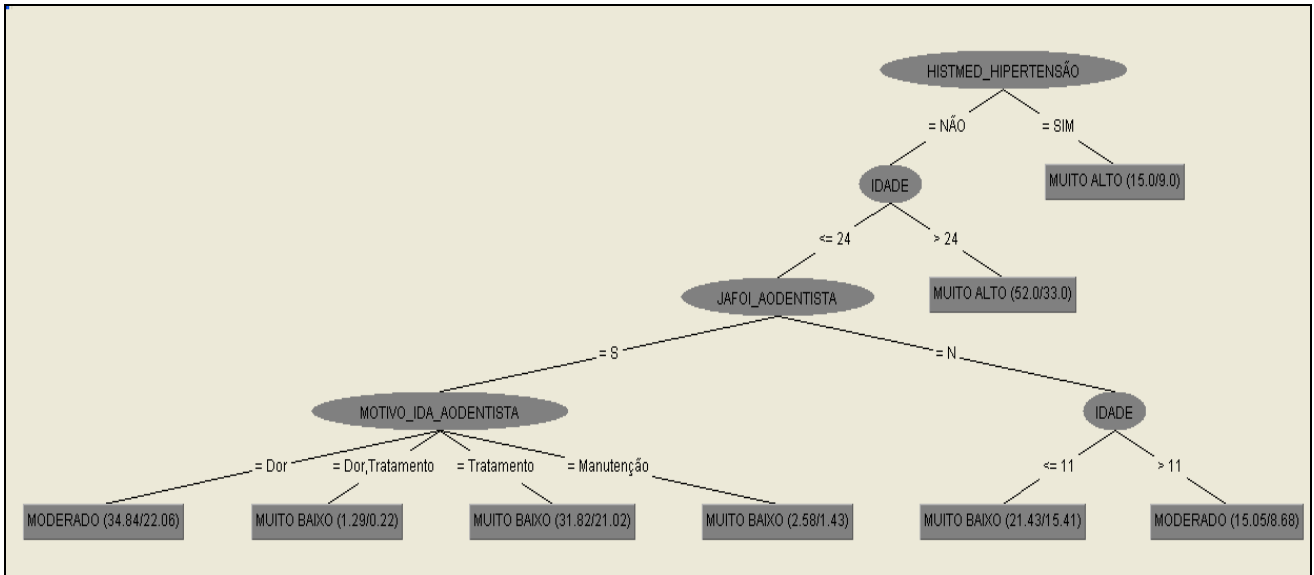


Figura 19 – Primeiro Experimento – MP II

## b) Validação dos Modelos

Neste momento, a forma utilizada para mensurar a qualidade dos modelos gerados foi a interpretação da matriz de confusão e a avaliação do especialista em saúde bucal.

- Incidência de Cárie Dentária – Embora tivéssemos um alto índice de registros corretamente classificados, avaliamos o modelo gerado como sendo de baixa qualidade, uma vez que:
  - a) Mostra-se exclusivamente direcionado a classificação de casos de incidência de cárie dentária, não obtendo nenhum acerto para casos de não incidência (conforme matriz de confusão). Este é um problema provavelmente ocasionado pela falta de balanceamento da amostra, sendo muito comum em tarefas de mineração de dados. Nestes casos, o algoritmo tende a interpretar a classe menos freqüente como ruído.
  - b) É semanticamente ruim, pois sabendo-se que a origem da cárie dentária provém de uma ação bacteriana, modelos que expressem

fatores relacionados ao histórico médico do paciente (Alergias e outras doenças), tendem a não trazer novas contribuições.

- c) O conjunto de campos explorados nos prontuários do CEU Vila Fátima pouco nos favoreceu para geração de resultados diferentes daqueles já consagrados pela literatura. Desta forma, para os próximos experimentos, deixaremos de considerar incidência de cárie, para focar exclusivamente na extração de modelos que possam estabelecer alguma diferença no perfil do paciente mediante diferentes níveis de severidade da doença.

Na árvore gerada pelo algoritmo evidenciamos uma possível relação entre motivos prévios de idas ao dentista com a incidência de futuras cáries. Em outras palavras, pacientes que já foram ao dentista por motivos de manutenção estariam mais propensos a não ter cárie, ao contrário daqueles que foram por motivos de dor ou tratamento. Este tipo de constatação ratifica a importância de ações preventivas para o controle da doença. Curiosamente, constatamos também uma relação da utilização de fluoreto com a incidência de cárie, mas atribuímos este tipo de ocorrência à subjetividade da pergunta durante a anamnese, ou seja, a autenticidade da resposta depende exclusivamente do paciente, embora a realidade possa ser evidenciada em um exame clínico.

- Indicador de Cárie Dentária – Diferentemente do modelo anterior, neste tivemos um baixo índice de registros corretamente classificados, provavelmente ocasionados pelo grande número de categorias do atributo classe (muito baixo, baixo, moderado, alto e muito alto). Adicionalmente, identificamos os mesmos problemas de balanceamento e de qualidade semântica do modelo anterior, fato que nos motivou a adotar novas práticas de preparação de dados para busca de melhores modelos.

Analisando a árvore gerada, observamos a relação sugerida entre pessoas com idade superior a 24 anos e indicadores de cárie muito altos, no entanto, desconsideramos este tipo de previsão uma vez que seu índice de acerto é inferior a 36%.

### 3.3.1.2 Segundo Experimento

Neste segundo experimento, adotamos algumas modificações na forma de preparação dos dados, gerando como resultado um novo arquivo de dados (ExpMiner02.CSV), conforme feito anteriormente para (ExpMiner01.CSV).

As modificações implementadas neste arquivo consistiram na execução de práticas que conhecidamente tornam o conjunto de dados mais adequado à aplicação de algoritmos baseados em árvore de decisão (TAN; STEINBACH; KUMAR, 2006), como por exemplo:

- Discretização e Binarização de atributos: Neste momento, discretizamos apenas o campo *frequência\_escovação*, transformando-o em um novo campo categórico denominado *escova\_maisde1x(sim,não)*. Para os campos *Motivo\_Ida\_Aodontista* e *Frequência\_Consumo\_Açucar* optamos por binarizar os valores possíveis, diminuindo de 3 para 2 em ambos casos. Outra importante mudança na preparação dos dados foi binarização do atributo classe Case (Indicador de Cárie), de modo a considerar apenas dois valores para o indicador: baixo ou alto. Com esta mudança esperamos aumentar a acurácia preditiva dos modelos obtidos no primeiro experimento.
- Oversampling (TAN; STEINBACH; KUMAR, 2006): técnica de amostragem utilizada para replicação de classes pouco frequentes, de forma a equilibrá-las com as classes mais frequentes. A partir desta técnica, utilizamos o arquivo base ExpMiner02.CSV e o derivamos em novos arquivos CSV, inclusive para predição de outras patologias não consideradas até este momento.

ATRIBUTO CLASSE	CLASSES	FREQUÊNCIA ORIGINAL	FREQUÊNCIA EQUILIBRADA	ARQUIVO
Indicador de Cárie	Muito Baixo	47	47	ExpMiner02_Indicador.CSV
	Baixo	43	47	
	Moderado	34	47	
	Alto	30	47	
	Muito Alto	20	47	
Indicador de Cárie Binarizado	Baixo	43	94	ExpMiner02_IndicadorBin.CSV
	Alto	30	94	
Oclusão	Normais		34	ExpMiner02_Oclusao.CSV
	Alterações		34	
DoençaPeriodontal	S		34	ExpMiner02_Periodont.CSV
	N		34	

Quadro 14 - OverSampling

- Criação de atributos agregados: A partir dos atributos já existentes, referentes ao estado dos elementos dentários do paciente (decíduos e permanentes), criamos outros novos, agregando-os em *arcada\_inferior*, *arcada\_superior*, *primeiro\_quadrante*, *segundo\_quadrante*, *terceiro\_quadrante*, *quarto\_quadrante*;
- Seleção de atributos: Embora fosse possível continuar selecionando os atributos mais relevantes pelo uso do conhecimento de domínio adquirido, adotamos para



este experimento uma abordagem sistemática baseada na aplicação do filtro ChiSquaredAttributeEval disponível na ferramenta Weka 3.7.0. Este filtro avalia individualmente cada atributo pelo cálculo de seu qui-quadrado relacionado ao atributo classe. Assim, descartamos os atributos menos relevantes.

A partir da inclusão de novas técnicas de preparação de dados, bem como através *feedback* fornecido pelo especialista na primeira fase de experimentos, conseguimos obter melhores modelos, ainda que abaixo de nossas expectativas. Esse fato deve-se a baixa qualidade dos dados, o que de certa forma pode ser encarado como um fator limitador para obtenção de melhores resultados. Abaixo, fazemos uma breve análise sobre os modelos obtidos para as seguintes patologias bucais:

#### A1) Modelo Obtido - Indicador de Cárie convencional

No modelo detalhado no Apêndice K, podemos constatar uma melhora significativa na acurácia preditiva (42.97%), especialmente para os casos de predição dos indicadores Muito Baixo (80%) e Muito Alto (78%). Na figura 20 apresentamos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada.

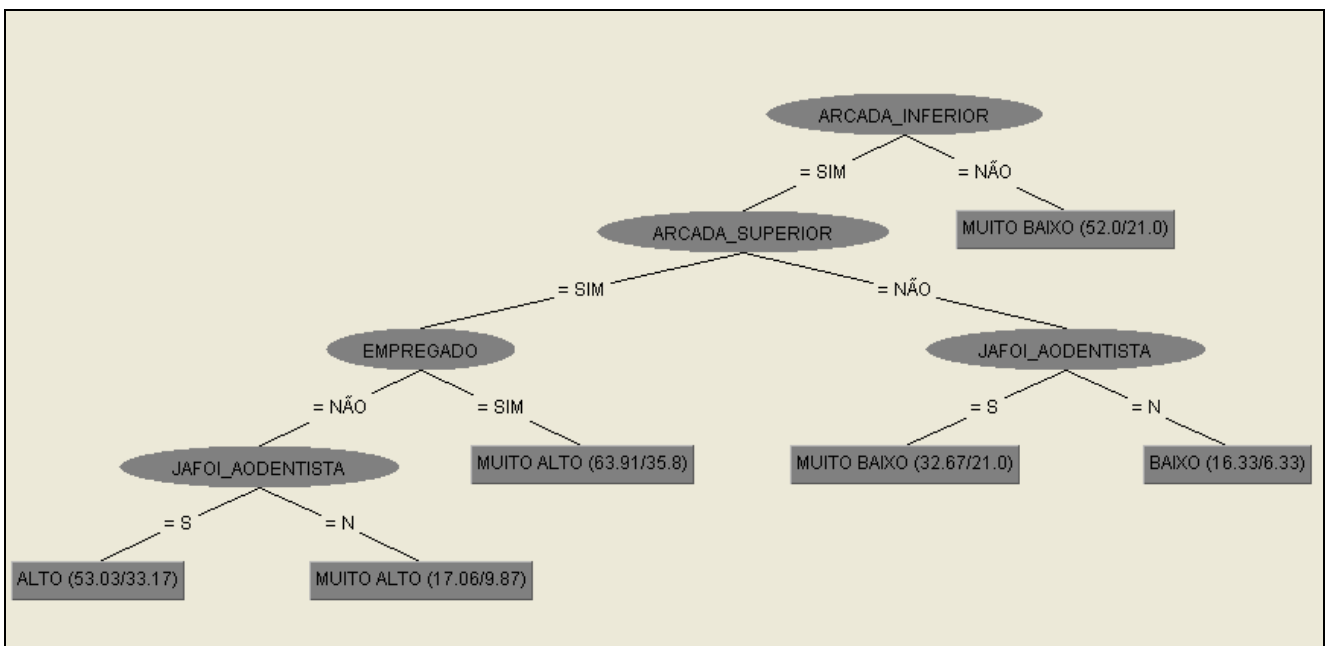


Figura 20 – Segundo Experimento – MP I

#### B1) Validação do modelo

Entre as evidências apontadas pelo modelo, destacamos:

- Relação da incidência de cáries na arcada inferior com indicadores de cárie muito baixos;
- Relação da ocupação profissional com indicadores de cárie muito altos, o que pode ser explicado pelo menor grau de instrução e tempo em casa dos que trabalham fora em relação às donas de casa, estudantes e desempregados.
- Em contrapartida, relações de idas prévias ao dentista desta vez não se mostram relevantes ou até mesmo lógicas, provavelmente resultado de distorções ocasionadas pela baixa qualidade dos dados.

## A2) Modelo Obtido - Indicador de Cárie Binarizado

Com a binarização do indicador de cárie dentária (Detalhes no Apêndice L), obtivemos uma sensível melhora na acurácia preditiva (50.53%), sendo 42% para altos indicadores e 58% para baixos indicadores. Na figura 21 ilustramos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada.

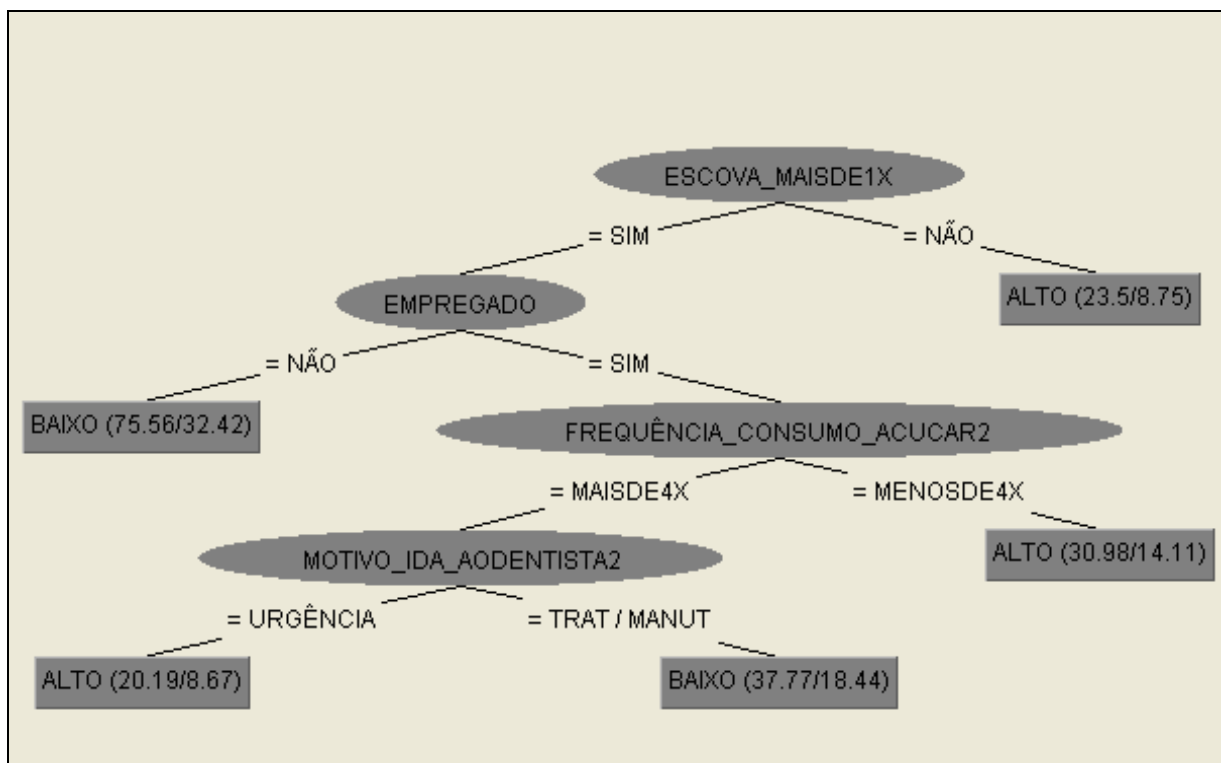


Figura 21 – Segundo Experimento – MP II

## B2) Validação do Modelo - Indicador de Cárie Binarizado

Conforme apontado pelo modelo anterior, verificou-se agora, novas evidências para a relação de ocupação profissional com baixos indicadores de cárie dentária. Da mesma forma, a adoção de medidas preventivas, como baixa frequência de escovações diárias foram determinantes para a incidência de altos indicadores de cárie dentária.

Em um segundo modelo gerado (Detalhes no Apêndice M) foi sugerido pelo modelo que pessoas (crianças neste caso) que não contam com alguma ajuda na escovação dentária, estão menos propensas a ter altos indicadores de cárie, o que vem a valorizar ainda mais a atuação dos pais na prevenção desta patologia. Em contrapartida, o modelo apontou que pessoas sem ocupação profissional (como donas de casa e estudantes) que não escovam mais de uma vez por dia os dentes, estão mais propensas a apresentar altos indicadores de cárie. Na figura 22 ilustramos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada.

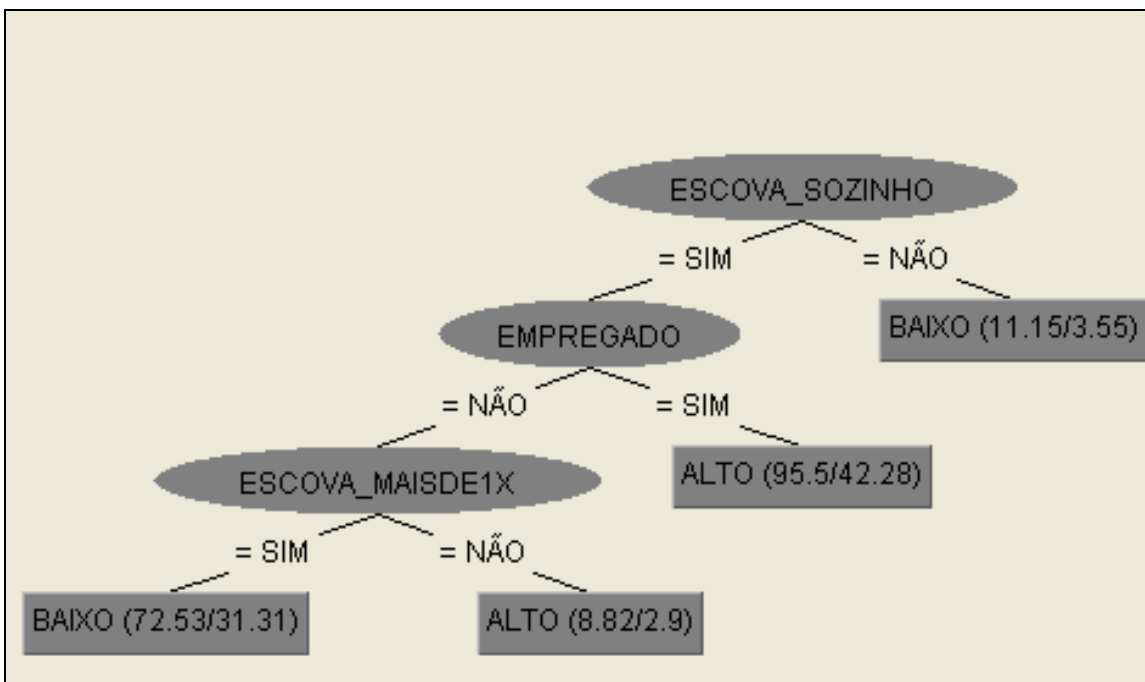


Figura 22 – Segundo Experimento – MP III

## A3) Modelo Obtido - Incidência de Má-oclusão

Quando analisada a incidência de má oclusão (Detalhes no Apêndice N), chegamos a uma acurácia preditiva de 54%, obtendo melhores resultados nos casos de classificação de oclusões normais (79%). Na figura 23 ilustramos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada.

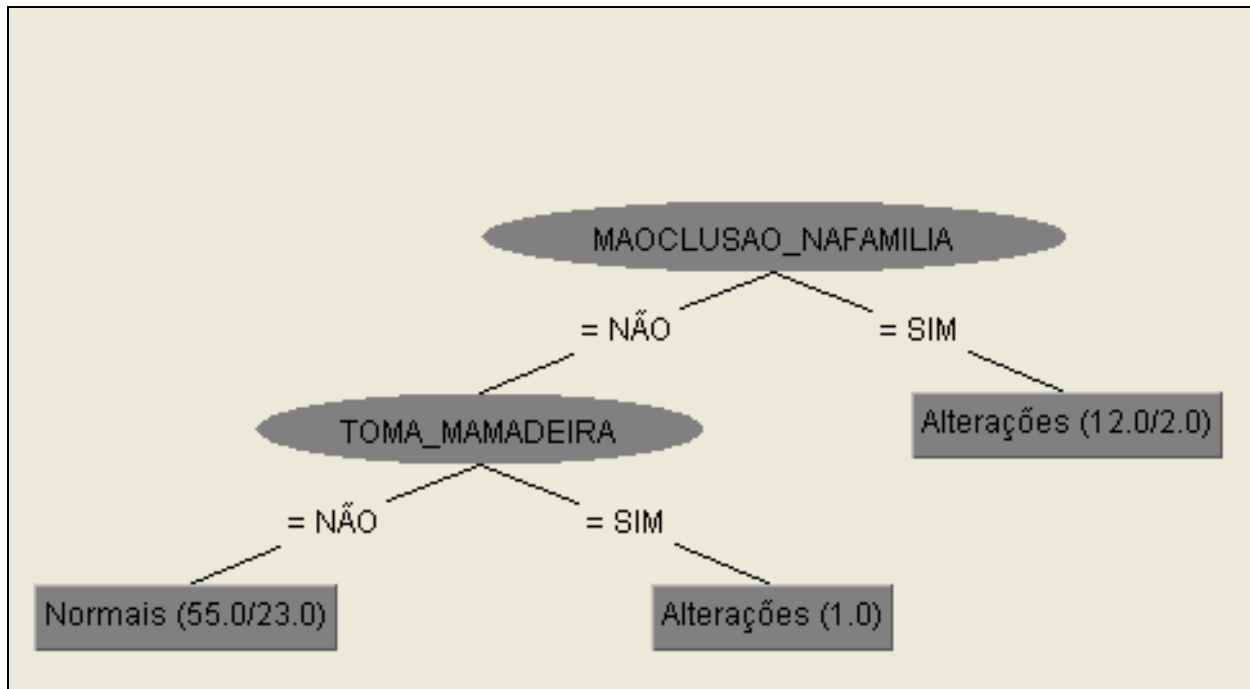


Figura 23 – Segundo Experimento – MP IV

### B3) Validação do Modelo - Incidência de Má-oclusão

Embora tenhamos constatado um baixo percentual de acerto para os casos de classificação de alterações oclusais (29.41%), informações obtidas junto a especialistas em saúde bucal nos permitem acreditar que casos anteriores de má oclusão na família são bons preditores de novas ocorrências, assim como o uso excessivo de mamadeiras durante a infância da criança.

### A4) Modelo Obtido - Incidência de Doença Periodontal

Considerando a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada para predição de doenças periodontais (Detalhes no Apêndice O), chegamos ao nosso melhor modelo preditivo neste segundo experimento, atingindo um total de 66.17% de instâncias corretamente classificadas, sendo 73.52% para casos de não incidência de doenças periodontais e 58.82% para casos de incidência. Na figura 24 ilustramos a árvore de decisão (TAN; STEINBACH; KUMAR, 2006) gerada.

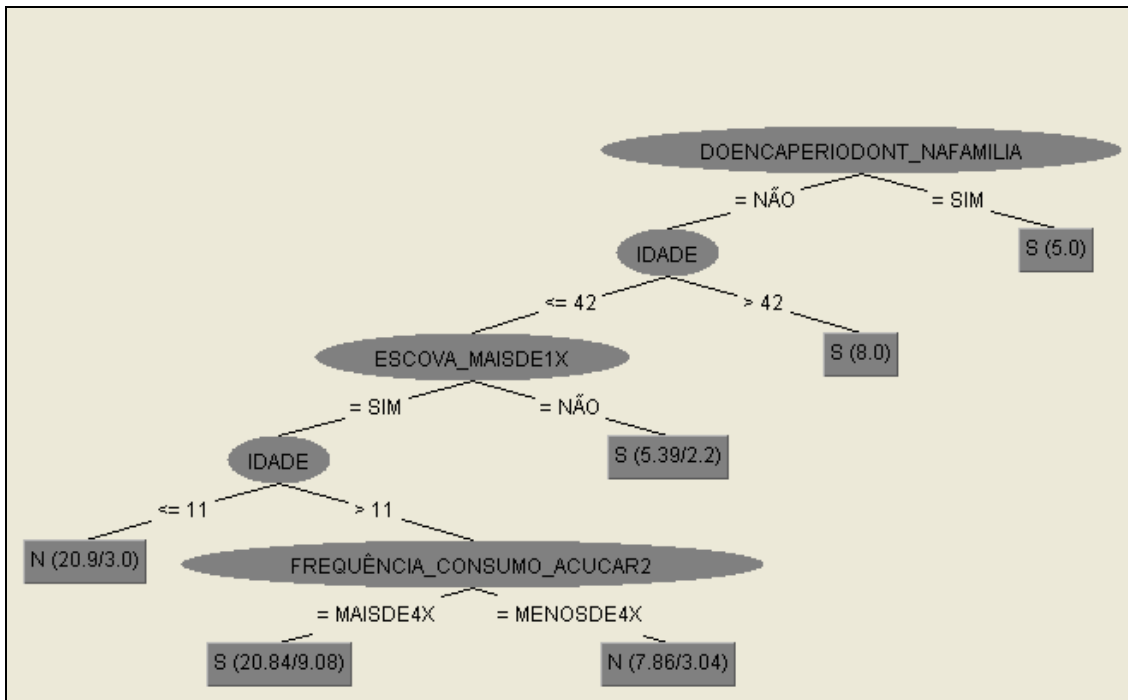


Figura 24 – Segundo Experimento – MP V

#### B4) Validação do Modelo – Incidência de Doença Periodontal

Da mesma forma que encontramos relação de fatores genéticos na incidência de má oclusão, verificamos novamente sua relevância para incidência de doenças periodontais, sendo mais frequente em pessoas entre 11 a 42 anos com baixa frequência de escovação dentária (menos de 2x por dia) e que consomem altas quantidades de açúcar (neste caso, mais de 4x diárias)

### 3.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Como se pode observar neste capítulo, realizamos todo um esforço para o desenvolvimento de uma solução de KDD em saúde bucal, desde a criação de um ambiente operacional até o desenvolvimento de um novo ambiente para fins analíticos. É importante destacar a importância deste tipo de ambiente, não apenas para análise histórica dos dados, mas também como uma etapa preliminar de preparação dos dados em que começamos a adequá-los a etapa de *data mining*. Igualmente, o desenvolvimento de modelos multidimensionais dentro de um processo de KDD nos permite visualizar novas oportunidades de mineração. No entanto para este trabalho, não chegamos a uma oportunidade específica, o que nos motivou a manter nosso objetivo original na busca por modelos preditivos relacionados a patologias periodontais, de má-oclusão e indicadores de cárie dentária. Como meio para atingirmos este objetivo, concentramos nossos experimentos na aplicação de árvores de decisão (TAN; STEINBACH; KUMAR, 2006), cuja técnica entendemos ser mais adequada a geração de modelos mais compreensíveis. Outrossim, também utilizamos abordagens

descritivas como análise de clustering (K-Means) e extração de regras de associação (Apriori), mas a baixa qualidade dos modelos gerados nos motivou a não ilustrá-los durante este trabalho.

No (Apêndice P) realizamos um terceiro experimento, no qual mostramos os resultados obtidos pela extração de regras de associação para identificação de variáveis relacionadas ao indicador de cárie dentária. Para tanto utilizamos o algoritmo APriori da ferramenta Weka 3.7.0.

Para facilitar a compreensão, reduzimos o número de regras geradas e selecionamos as mais convenientes para ilustrar sua aplicação. Embora o modelo aponte para um determinado grupo de dentes cariados como mais frequentemente associados a altos indicadores de cárie, outras situações mostram-se contraditórias, como por exemplo a associação de práticas preventivas com altos indicadores de cárie dentária. Novamente, atribuímos este tipo de ocorrência a baixa qualidade dos dados.

Com a discussão documentada neste capítulo, entendemos ter contribuído para identificação de situações previamente desconhecidas à instituição parceira, como por exemplo, o uso excessivo de materiais provisórios em tratamentos de restauração. Adicionalmente, geramos modelos preditivos para diferentes patologias bucais, identificando situações corroboradas pela literatura, bem como anomalias ocasionadas pela baixa qualidade dos dados registrados.

Da mesma forma, ressaltamos que não é objetivo deste trabalho atribuir a implementação técnica de um processo de KDD a gestores de saúde bucal, mas sim, dar maior visibilidade e conhecimento para seu acompanhamento e controle.



## 4 TRABALHOS RELACIONADOS

Durante as últimas duas décadas, uma série de estudos foram realizados com o objetivo de identificar o perfil odontológico da população brasileira. Em um cenário mais abrangente, os levantamentos epidemiológicos realizados pelo Ministério da Saúde em 1986, 1996 e 2003 têm servido como fonte de informação para novas pesquisas na área, especialmente às realizadas com o intuito de explorar a associação entre indicadores sociais e os de saúde bucal. Dentre as pesquisas mais relevantes, destacam-se duas linhas de abordagens. Na primeira linha, seguem os estudos descritivos e preditivos de cunho puramente estatístico (correlação, regressão linear e logística). Na segunda linha, seguem os trabalhos que se utilizam de técnicas automatizadas de *data mining* (árvores de decisão, regras de associação, redes neurais.) para descrição e predição em saúde bucal.

No que diz respeito a esta segunda linha de abordagem, observamos que os trabalhos desenvolvidos na área ainda estão muito focados em uma visão parcial do processo, ou seja, uma exploração comparativa de técnicas de data mining, deixando de lado etapas de extrema importância como a preparação dos dados e a interpretação de resultados.

Em (BALDANI; NARVAI; ANTUNES, 2004), modelos de regressão linear múltipla foram desenvolvidos para associação da variação do índice CPO-D com indicadores sociais e socioeconômicos de renda, habitação, escolaridade, oferta de serviços odontológicos e fluoretação das águas no estado do Paraná. Os resultados deste trabalho corroboram com a literatura, apontando a cárie dentária como uma patologia diretamente associada às populações de baixa renda.

Similarmente, o estudo de (FERNANDES; PERES, 2005) explorou a associação de indicadores em atenção básica de cobertura de primeira consulta programática, média anual de população entre zero a quatorze anos de idade coberta por procedimentos odontológicos coletivos, e proporção de exodontias na dentição permanente em relação ao total de procedimentos odontológicos individuais, com indicadores sociais, econômicos e de serviços municipais no estado de Santa Catarina. Para tanto, foi utilizado o teste estatístico de Spearman.

No estudo de (TAGLIAFERRO et al., 2006), foi aplicada regressão logística para a identificação de fatores de risco associados ao aumento da cárie dentária em crianças de 6 a 8 anos de idade no estado de São Paulo. Como principal resultado deste trabalho, constatou-se uma forte relação do aumento do índice de cárie com o baixo nível escolar da mãe. Da mesma forma, experiências prévias de cárie na dentição decídua foram considerados significantes preditores para futuras cáries.



Em (CELESTE; NADANOVSKY; LEON, 2007), foram utilizados dados de 4033 jovens gaúchos entre 15 e 19 anos e aplicada a técnica de regressão logística multinível para avaliar a associação entre atividades de prevenção da cárie dentária e a prevalência do índice CPO-D. Entre as conclusões obtidas, ratificou-se a importância dos procedimentos preventivos como um fator inibidor da cárie dentária.

Na segunda linha, seguem os estudos que utilizam técnicas automatizadas de data mining para tarefas de descrição e predição em saúde bucal. Entre as técnicas mais utilizadas destacaram-se aquelas com propósitos de regressão e classificação.

Em (GANSKY, 2003) foram analisados os dados de 466 crianças de até vinte e quatro meses de idade para predição do risco de cáries. Foram então utilizadas e comparadas as técnicas de regressão logística, árvores de regressão e classificação, e redes neurais. Embora tenha introduzido conceitos fundamentais do processo de KDD, este ainda é um trabalho eminentemente orientado a comparação da acurácia preditiva de diferentes técnicas de *data mining*.

No trabalho de (CUNHA; DIAS, 2007), foi apresentado um estudo de caso envolvendo a exploração de um banco de dados clínico gerenciado pelo banco de dados Microsoft Access 2000 (Microsoft Corporation, Redmond, WA, USA), onde foram aplicados algoritmos de mineração de dados (C4.5 e Apriori) (TAN; STEINBACH; KUMAR, 2006) visando à identificação preditiva e descritiva do perfil do paciente e os procedimentos aplicados em seu tratamento.

Em (MONTENEGRO; OLIVEIRA; CABRAL, 2008) foi realizado um estudo experimental com dados coletados em entrevistas com mães de 3864 crianças abaixo de 5 anos de idade. O estudo consistiu na avaliação e comparação das técnicas de árvores de decisão, redes neurais, KNN e Support Vector Machine aplicadas à predição de cárie dentária. Conforme já verificado em outros estudos, constatou-se uma provável influência das condições financeiras e das experiências anteriores de cárie com a incidência de novos casos da doença.

No trabalho de (TAMAKI et al., 2009) foi realizado um estudo para predição de cáries utilizando os dados de 560 crianças com idade entre 5 e 8 anos no Japão. Neste estudo foram utilizadas as técnicas de regressão logística, árvore de decisão e redes neurais. Como resultado deste estudo, verificou-se a relação do pH salivar e sexo do paciente com a incidência de novas cáries.

Mediante os trabalhos já realizados na área, entendemos que a predição de indicadores possa apresentar resultados mais satisfatórios, fornecendo ao gestor de saúde não apenas a

informação de incidência da patologia, mas principalmente a classificação do perfil do paciente conforme o grau de severidade da patologia analisada.

#### **4.1 CONSIDERAÇÕES FINAIS DO CAPÍTULO**

Apresentada a solução deste trabalho e feitas as devidas considerações sobre os trabalhos realizados na área, podemos concluir que embora exista na literatura uma diversidade de estudos sobre a predição de patologias bucais (principalmente cáries dentárias), ainda há uma carência de trabalhos que explorem o desenvolvimento completo de um processo de KDD, bem como a exploração de outras patologias tão importantes quanto a cárie dentária.

Dadas estas considerações, poderíamos ainda, estabelecer critérios para comparação deste trabalho com os apresentados neste capítulo, no entanto, julgamos inadequados, à medida que nenhum destes explora um processo completo de KDD em saúde bucal.



## 5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Motivado pela constante necessidade de se adotar melhores instrumentos de gestão em saúde, este trabalho buscou extrair modelos descritivos e preditivos para análise de patologias bucais. Para tanto, selecionamos na literatura, trabalhos relacionados ao uso de indicadores em diversos cenários de descoberta de conhecimento, permitindo-nos chegar as seguintes constatações:

- a) as abordagens sistemáticas para predição de métricas e indicadores em saúde bucal são pouco comuns na literatura, sendo mais frequente a análise de incidência de patologias;
- b) abordagens estatísticas que exploram a análise preditiva de indicadores tendem a demandar grande esforço, à medida que estão sustentadas na validação de hipóteses e na compreensão de resultados exclusivamente numéricos;
- c) os trabalhos estão excessivamente orientados a comparação de técnicas de data mining, desprezando etapas importantes no processo de descoberta de conhecimento, como avaliação e interpretação dos modelos gerados.

A partir destas constatações, o estudo realizado junto ao CEU Vila Fátima buscou explorar o uso de métricas e indicadores em saúde bucal juntamente com a análise de incidência de outras patologias bucais tipicamente negligenciadas na literatura.

Como principal contribuição desta pesquisa desenvolvemos e documentamos um processo completo de KDD adotando para tanto os seguintes critérios para sua avaliação:

- a) Robustez da solução: Através da construção de um ambiente de data warehousing, agregamos robustez à solução, viabilizando desta forma, a análise de grandes volumes de dados, requisito fundamental para cenários de gestão pública. Em contrapartida, este tipo de característica pode não ser vantajosa para análise de pequenos volumes de dados, à medida que demanda um considerável esforço para sua implementação.
- b) Compreensibilidade dos modelos: Embora a baixa qualidade dos dados tenha influenciado diretamente na qualidade dos modelos gerados, entendemos que a solução produzida é capaz de gerar modelos compreensíveis aos gestores de saúde, facilitando a interpretação dos resultados e fornecendo melhor embasamento para tomada de decisão. Uma solução futura para este tipo de

problema seria a adoção de sistemas informatizados para registro das informações dos prontuários odontológicos. Desta forma, atribuiríamos ao sistema o controle de preenchimento das informações, o que aumentaria significativamente a qualidade dos dados e dos modelos gerados. Neste trabalho, especificamos e implementamos um sistema (Prontuário Digital) para esta finalidade, o que foi de extrema importância para viabilização deste trabalho.

- c) Tratamento de dados não convencionais: Entre as características típicas deste tipo de cenário em saúde, destacamos a importância de se adotar técnicas para o tratamento de dados não convencionais, como descrições textuais e dados temporais. Para o tratamento de dados temporais, modelamos dimensões que nos permitiram observar os fatores relacionados em diferentes momentos de tempo. No entanto, para a extração de modelos preditivos de cárie dentária, adotamos um critério de seleção baseado na escolha do odontograma mais recente do paciente. Para o tratamento de dados descritivos adotamos práticas de preparação baseadas em *stemming*. No entanto, entendemos que novas frentes de atuação baseadas no processamento da linguagem natural possam agregar maior qualidade ao tratamento de textos não estruturados.

Como trabalhos futuros pretendemos garantir a continuidade do estudo junto ao CEU Vila Fátima, bem como explorar novas áreas de atuação junto a Faculdade de Odontologia, como por exemplo o reconhecimento de novos padrões a partir da análise de imagens médicas.

## 5.1 PUBLICAÇÕES

Entre os principais resultados obtidos nesta pesquisa, destacamos o artigo aceito e publicado na revista Odonco Ciência em 21/06/2009, o artigo aceito no ERBD 2010 em 09/03/2010 e o capítulo publicado no livro “Inovação, Universidade e Relação com a Sociedade: boas práticas na PUCRS”.

Este estudo recebeu o suporte do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) de acordo com o edital MCT/CNPq 27/2007.

## REFERÊNCIAS BIBLIOGRÁFICAS

- BALDANI, M. H.; NARVAI, P. C.; ANTUNES, J. L. F. *Cárie Dentária e Condições Sócio-econômicas no Estado do Paraná, Brasil*. Cad. Saúde Pública, Rio de Janeiro, vol.20, n1, 2004, pp. 143-152.
- BARROS, R C. ; BASGALUPP, M. P. ; TENORIO, N. ; RUIZ, D. D. ;BECKER, K. . *Issues on Estimating Software Metrics in a Large Software Operation*. In: 32nd Annual IEEE Software Engineering Workshop. Kassandra, Chalkidiki, Greece, 2008, pp. 152-160.
- BLOMBERG, L. C. ; MOTA, E. G. ; FIGUEIREDO, J. A. ; RUIZ, D. D. A. *Development of an oral health database for the management of clinical records*. Revista Odonto Ciência (PUCRS. Impresso), vol. 24, 2009, pp. 249-253.
- BRASIL, Ministério da Saúde - Instituto Nacional de Câncer. *Falando Sobre Câncer da Boca*. Rio de Janeiro: INCA, 2002, 52p.
- BRASIL, Ministério da Saúde. *A Política Nacional de Saúde Bucal do Brasil: Registro de uma Conquista Histórica*. José Felipe Riani Costa, Luciana de Deus Chagas, Rosa Maria Silvestre (Orgs.). Brasília: Organização Pan-Americana da Saúde, 2006, 39p.
- CELESTE, R. K.; NADANOVSKY, P.; LEON, A. P. *Associação entre procedimentos preventivos no serviço público de odontologia e a prevalência de cárie dentária*. Rev. Saúde Pública, São Paulo, vol. 41, n. 5, 2007, pp.830-838.
- CUNHA, G. G. D. ; DIAS, C. R. *Aplicação de Mineração de Dados em uma Base Odontológica*. Revista Eletrônica da Faculdade Metodista de Granbery. Minas Gerais, Ed. Jul/Dez, 2007, 25p.
- ELMASRI R, NAVATHE S. B. *Sistemas de banco de dados*. São Paulo: Pearson Education; 4.ed., 2006, 724p.
- FAYYAD, U.; PIATETSKY-S G.; SMYTH P. *The KDD Process for Extracting Useful Knowledge from Volumes of Data*. Communications of the ACM, N.Y, vol. 39, n11, 1996, pp. 27-34.
- FERNANDES, L. S.; PERES, M. A. *Associação entre Atenção Básica em Saúde Bucal e Indicadores Socioeconômicos Municipais*. Rev. Saúde Pública. Rio de Janeiro, vol. 39, n6, 2005, pp.930-936.
- FIGUEIRA, F. V.; BECKER, K.; RUIZ, D. D. *Mineração em métricas de software*. In: ESCOLA REGIONAL DE BANCO DE DADOS (ERBD'07). Caxias do Sul, SBC, 2007, p.204-213.
- GANSKY S. A. *Dental data mining: potential pitfalls and practical issues*. Adv Dent Res. Ed.17, 2003, pp.109-14.
- GARCIA, R. S.; RUIZ, D. D. *Pré-processamento de dados para descoberta de conhecimento em processos de workflow*. In: ESCOLA REGIONAL DE BANCO DE DADOS (ERBD'05), 2005. Porto Alegre. SBC, 2005, pp.204-213.
- HAN, J.; KAMBER, M. *Data Mining : concepts and techniques*. San Francisco (CA): Morgan Kaufmann, c2001. 550 p.

- KANER, C.; BOND, W. *Software engineering metrics: What do they measure and how do we know?* In 10th International Software Metrics Symposium (METRICS 2004). CS Press, Setembro 2004, 12p.
- KIMBALL, R.; ROSS, M.. *The data warehouse toolkit: the complete guide to dimensional modeling*. São Paulo: Campus, 2. ed., 2002, 436p.
- MACHADO, F. N. R. *Tecnologia e Projeto de Data Warehouse: uma Visão Multidimensional*. 1. ed. São Paulo: Érica, 2004, 318p.
- MARCONI, M. A.; LAKATOS, E. M. *Fundamentos de metodologia científica*. 5. ed. São Paulo : Atlas, 2003, 311 p
- MONTENEGRO, R. D.; OLIVEIRA, L. I.; CABRAL, G. G. *A Comparative Study of Machine Learning echniques or Caries Prediction*. In: 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), vol. 2, 2008., pp.477-481.
- NOVELLO, T. C. *Uma abordagem de Data Warehouse para análise de processos de desenvolvimento de software*. Porto Alegre, 2006. 153 p. Dissertação (Mestrado em Ciência da Computação) – PUCRS, Faculdade de Informática. Acessado em: 20/10/2009. Disponível em: [http://tede.pucrs.br/tde\\_busca/arquivo.php?codArquivo=1159](http://tede.pucrs.br/tde_busca/arquivo.php?codArquivo=1159)
- ORENGO, V.; HUYCK, C. *A Stemming Algorithmm for the Portuguese Language*. In: Eighth Symposium on String Processing and Information Retrieval (SPIRE'01), 2001, pp.186-193.
- RIPSA. *Indicadores Básicos para a Saúde no Brasil: conceitos e aplicações*. 2ed. Brasília: Organização Pan-America em Saúde, 2008, 350p. Disponível em: <http://www.opas.org.br/sistema/arquivos/matriz.pdf>. Acessado em: 22/out/2008.
- RUMBAUGH J, JACOBSON I, BOOCH G. *The unified modeling language reference manual*. Reading: Addison-Wesley, c1999, 550 p.
- SAHAMA, T. R.; CROLL, P. R. *A Data Warehouse Architecture for Clinical Data Warehousing*. In: ACM International Conference Proceeding Series. Ballarat, vol. 249, 2007, pp.227-232.
- SANTOS, P. C. F. ; MONTEIRO, A. L. B. ; ROCHA, R. G. ; JUNIOR, C. B. J. *Uma ferramenta alternativa para avaliação do índice dental estético*. Revista Clínica de Ortodontia Dental Press, vol. 7, 2008, pp. 34-39.
- STEINER, M. T. A.; SOMA, N. Y.; SHIMIZU, T.; NIEVOLA, J. C.; NETO, P. J. S. *Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados*. Gest. Prod. Vol. 11, n2, 2006, pp.11-25.
- TAGLIAFERRO, E. P. S. ; PEREIRA, A. C. ; MENEGHIM, M. C. ; AMBROSANO G. M. B. *Assessment of Dental Caries Predictors in a Seven-Year Longitudinal Study*. Journal Public Health Dent. 2006; n66, pp.169-73.
- TAMAKI Y; NOMURA Y; KATSUMURA S; OKADA A; YAMADA H; TSUGE S; KADOMA Y; HANADA N. *Construction of a Dental Caries Prediction Model by Data mining* Journal of Oral Science. Vol. 51, n1, 2009, pp. 61-68.

TAN, P.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. Boston: Addison-Wesley, c2006., 769 p.

TAYLOR, D.; NAGUIB, R. N. G.; AMIN, S.; JAMES, A.; BOULTON, S. *Design considerations for a relational database schema to hold dental epidemiological data*. Information Technology Applications in Biomedicine, 2000. pp.342-346.

TAYLOR, D.; NAGUIB, R. N. G.; AMIN, S.; JAMES, A.; BOULTON, S. *Storage, querying and visualisation of clinical dental records*. Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, 2001;4: pp. 3718-3720.

TAYLOR D, NAGUIB RNG, BOULTON S. *A dynamic clinical dental relational database*. Information Technology in Biomedicine, IEEE Transactions on 2004, vol. 8, n3., pp.298-305.









## APÊNDICE B – ENTREVISTA / ÁREA BÁSICA

### PROJETO BANCO DE DADOS EM SAÚDE BUCAL ATA 01

Aos vinte e dois dias do mês de abril de dois mil e oito, reuniram-se às 11 horas, no Prédio 30 do Campus Central da PUCRS, os professores Eraldo e Maria Martha, da Faculdade de Odontologia, e o Mestrando Luciano, do Curso de Pós-Graduação em Ciência da Computação da FACIN. A entrevista foi aberta pelo Mestrando Luciano que apresentou o seguinte roteiro para sua condução:

**1) Identificação da Área e seus membros:** O Prof. Eraldo informou que ele e a Profa. Maria Martha compõem a Área definida como Básica – Diagnóstico Molecular, sendo ele o Coordenador. **2) Quais dados são manipulados e como estes dados são armazenados atualmente:** Os Professores informaram que atualmente não existe nenhum sistema para entrada / manipulação dos dados, assim como, nenhum banco de dados para armazenamento das informações obtidas. As únicas informações armazenadas atualmente encontram-se em papel e referem-se aos dados cadastrais e de acompanhamento do paciente. O Prof. Eraldo informou que no intuito de pesquisa, são realizadas coletas de DNA de pacientes voluntários onde posteriormente um pesquisador poderá solicitar exame de sequenciamento de determinados genes com a finalidade de se realizar estudos mais aprofundados sobre a relação gene x patologia. Nesta etapa da reunião, o Mestrando Luciano elaborou questionamentos para obter as entidades, atributos e seus respectivos relacionamentos, mas para melhor entendimento, tais aspectos serão descritos em novo documento separado, onde os dados serão representados através de uma linguagem formal caracterizada por diagramas de classe. **3) Quais são os objetivos visados com a modelagem do Banco de Dados:** Os professores explicaram que não existe uma tabela predefinida, que determine a correlação mencionada acima, pois este é um assunto em freqüente estudo e que ainda está em evolução. A idéia principal é criar uma estrutura de dados que possibilite resgatar mais eficientemente as correlações entre patologias bucais e os genes causadores, minimizando assim a necessidade de recorrer a estudos e levantamentos bibliográficos. **4) Material solicitado / recebido:** Buscando melhor caracterização das entidades de dados, o Mestrando Luciano solicitou que fosse disponibilizado um prontuário de um paciente (em branco ou rasurado). Os Professores Eraldo e Maria Martha comprometeram-se em deixar o material solicitado até sexta-feira (25-04-08), na Secretaria da Faculdade de Odontologia. **5) Material a ser gerado:** O mestrando Luciano informou aos professores que a partir desta reunião seriam gerados os seguintes documentos: Formulário de identificação de Área, Ata e diagrama de classes. **6) Considerações finais:** O Mestrando Luciano relatou que haverá a necessidade de novas interações com os professores para esclarecimento de eventuais dúvidas, visto as restrições de tempo e disponibilidade dentro do projeto.

Nada mais havendo a tratar no momento, encerrou-se a reunião às 12h20min, cuja Ata vai assinada por mim e pelos Professores da Área.

---

Eraldo L. B. Júnior  
Prof. Coordenador

---

Luciano Blomberg  
Mestrando

---

Maria Marta Campos  
Professora



## APÊNDICE C – ENTREVISTA / ÁREA DIAGNÓSTICOS POR IMAGENS

### PROJETO BANCO DE DADOS EM SAÚDE BUCAL ATA 01

Aos dois dias do mês de maio de dois mil e oito, reuniram-se às 12 horas e 10 minutos, no Prédio 30 do Campus Central da PUCRS, a Professora Helena, da Faculdade de Odontologia, e o Mestrando Luciano, do Curso de Pós-Graduação em Ciência da Computação da FACIN. A entrevista foi aberta pelo Mestrando Luciano que apresentou o seguinte roteiro para sua condução:

**1) Identificação da Área e seus membros:** A Profa. Helena informou que ela é a única integrante da Área, definida como de Diagnósticos por imagem em odontologia. **2) Como os dados são manipulados e como estes dados são armazenados atualmente:** A Profa. Helena relatou que não existe nenhum sistema para manipulação dos dados. Assim como em outras áreas, os principais dados manipulados referem-se aos dados pertinentes ao prontuário e exames complementares. Dentre estes exames destacam-se os de filmes (imagens) que são objetos alvo da Área. Atualmente estes prontuários ficam armazenados em dois lugares: no térreo da Faculdade de Odontologia, no Setor de Triagem, e no 5º andar, no CERLAP (Centro de reabilitação de Fissuras Lábio Palatinas). Este último utiliza o mesmo prontuário do primeiro, mas diferencia-se por registrar dados de pacientes que apresentem alguma fissura. Inicialmente trataremos apenas do primeiro caso. **3) Quais são os objetivos visados com a modelagem do Banco de Dados:** A Profa. Helena apresentou como objetivos almejados: recuperação ágil de prontuários e exames de imagem; acesso aos dados referentes ao histórico de solicitações destes exames. Em um momento futuro, deseja-se converter estes exames em arquivo digital (DAICON), assim como, viabilizar sua manipulação por algum software de edição. Foi apresentado também, o interesse em tratar questões referentes a custos, ou seja, melhorar o controle sobre o pagamento de exames e outros procedimentos, algo que atualmente apresenta falhas. **4) Material solicitado / recebido:** Buscando melhor caracterização das entidades de dados, o Mestrando Luciano solicitou que fosse disponibilizado um prontuário. A Professora comprometeu-se em disponibilizar o material solicitado e solicitou ao Mestrando Luciano que retirasse na sua sala na Faculdade de Odontologia. **5) Material a ser gerado:** O Mestrando Luciano informou às professoras que a partir desta reunião seriam gerados os seguintes documentos: Formulário de identificação de área, Ata e diagrama de classes. **6) Considerações finais:** Constatou-se na reunião, que a Área de diagnósticos por imagens em odontologia não gera requisições para outras áreas, mas sim serve como Área meio para as demais. O Mestrando Luciano relatou ainda, que haverá a necessidade de novas interações com os professores para esclarecimento de eventuais dúvidas, visto as restrições de tempo e disponibilidade dentro do projeto. Nada mais havendo a tratar no momento, encerrou-se a reunião às 13h10min, cuja Ata vai assinada por mim e pela Professora da Área.

---

Profa. Helena Oliveira  
Professora

---

Luciano Blomberg  
Mestrando



## APÊNDICE D – ENTREVISTA / ÁREA ENDODONTIA

### PROJETO BANCO DE DADOS EM SAÚDE BUCAL ATA 01

Aos vinte dias do mês de maio de dois mil e oito, reuniram-se às 10 horas e 50 minutos, no Prédio da Faculdade de Odontologia do Campus Central da PUCRS, os professores Figueiredo, Alexandre, Maristela e Patrícia, da Faculdade de Odontologia, e o Mestrando Luciano, do Curso de Pós-Graduação em Ciência da Computação da FACIN. A entrevista foi aberta pelo Mestrando Luciano que apresentou o seguinte roteiro para sua condução: **1) Identificação da Área e seus membros:** O Prof. Figueiredo informou que ele e os Professores Alexandre, Patrícia, Simone, Márcio e Maristela compõem a Área de Endodontia, sendo a última citada Coordenadora desta Área. **2) Quais dados são manipulados e como estes dados são armazenados atualmente:** Os Professores informaram que não existe nenhum sistema para manipulação e armazenamento dos dados. O gerenciamento dos dados é realizado da mesma forma que nas outras áreas, ou seja, o procedimento inicial para entrada de dados se dá através da criação do prontuário, neste caso no setor de triagem do prédio da Odontologia, e posteriormente novos dados são agregados a este prontuário, como dados de acompanhamento, exames complementares, diagnósticos, e planos de tratamento. Durante a reunião foram apresentados protótipos de fichas de acompanhamento do tratamento do paciente. O objetivo é incorporar tais dados na modelagem do banco de dados a ser criada. **3) Quais são os objetivos visados com a modelagem do Banco de Dados:** Identificou-se nesta reunião com a Área, três frentes de trabalho. A primeira refere-se ao controle dos dados históricos do paciente, algo que já está relativamente adiantado através do trabalho que está sendo feito com as demais Áreas. A segunda frente seria o interesse em criar um projeto piloto com a Área de Endodontia cujo objetivo visa estruturação a organização dos dados de ensino, permitindo aos professores da graduação e pós-graduação acesso ágil e completo a informações como Plano de Curso, Plano de Disciplina e Plano de aula. A última frente consiste na estruturação dos dados de pesquisa como, por exemplo, monografias, dissertações e teses, organizadas por tema, permitindo que alunos e professores possam utilizá-las como fonte para continuidade ou desenvolvimento de novas pesquisas. **4) Material solicitado / recebido:** Não foi solicitado nenhum material aos Professores da Área. Foi entregue pelo Professor Figueiredo, as fichas de acompanhamento do tratamento do paciente. **5) Material a ser gerado:** O Mestrando Luciano informou aos professores que a partir desta reunião seriam gerados os seguintes documentos: Formulário de identificação de área, Ata e diagrama de classes. **6) Considerações finais:** Neste primeiro momento será dada prioridade à finalização da modelagem em cima dos dados do paciente, visto que já se tem um trabalho mais adiantado neste sentido. Posteriormente, colocou-se como prioridade a modelagem dos dados de Ensino. Nada mais havendo a tratar no momento, encerrou-se a reunião às 11h50min, cuja Ata vai assinada por mim e pelos Professores da Área.

---

Maristela Borba  
Professora Coordenadora

---

Luciano Blomberg  
Mestrando

---

José A. P. de Figueiredo  
Professor

---

Alexandre Ghisi  
Professor

---

Patrícia Kopper  
Professora

---

Simone Luise  
Professora

---

Márcio Menin  
Professor





## APÊNDICE E – ENTREVISTA /ÁREA DE ENSAIOS MECÂNICOS

### PROJETO BANCO DE DADOS EM SAÚDE BUCAL ATA 01

Aos vinte sete dias do mês de maio de dois mil e oito, reuniram-se às 11 horas e 10 minutos, no Prédio 30 do Campus Central da PUCRS, o Prof. Eduardo Mota da Faculdade de Odontologia, e o Mestrando Luciano, do Curso de Pós-Graduação em Ciência da Computação da FACIN. A entrevista foi aberta pelo Mestrando Luciano que apresentou o seguinte roteiro para sua condução: **1) Identificação da Área e seus membros:** O Prof. Eduardo Mota é o responsável pela Área de Ensaaios Mecânicos. **2) Quais dados são manipulados e como estes dados são armazenados atualmente:** Atualmente há dois tipos de dados na área, cujo registro é mantido. A primeira categoria de dados refere-se aos dados provenientes de ensaios mecânicos gerenciados pelo software MTeste que aparentemente gera arquivos de relatórios em um formato proprietário. O Segundo tipo são dados de testes de microdureza que são armazenados dentro de planilhas do Microsoft Excel. Somente professores e alunos vinculados a projetos podem realizar tais procedimentos. **3) Quais são os objetivos visados com a modelagem do Banco de Dados:** Segundo o professor Eduardo, há o interesse de criar uma estrutura de armazenamento para estas informações de forma a proporcionar que novos pesquisadores possam resgatá-las mais agilmente, bem como disponibilizá-las como fonte de conhecimento prático para os profissionais da área clínica. **4) Material solicitado / recebido:** Foi solicitado ao professor que enviasse por e-mail uma imagem dos relatórios dos procedimentos acima relatados. A pedido da Profa. Helena, foi entregue no momento da reunião pelo Prof. Eduardo, os documentos solicitados a área de “Diagnóstico por imagem em Odontologia”. **5) Material a ser gerado:** O Mestrando Luciano informou aos professores que a partir desta reunião seriam gerados os seguintes documentos: Formulário de identificação de área, Ata e diagrama de classes. **6) Considerações finais:** Esta entrevista de área encerra a primeira etapa do processo de modelagem do banco de dados em saúde bucal. Nada mais havendo a tratar no momento, encerrou-se a reunião às 12h10min, cuja Ata vai assinada por mim e pelo Professor da Área.

---

Eduardo Mota  
Professor Coordenador

---

Luciano Blomberg  
Mestrando



## APÊNDICE F – ENTREVISTA / ÁREA ESTOMATOLOGIA

### PROJETO BANCO DE DADOS EM SAÚDE BUCAL ATA 01

Aos vinte e cinco dias do mês de abril de dois mil e oito, reuniram-se às 8 horas e 30 minutos, no Prédio 30 do Campus Central da PUCRS, as professoras Liliane e Karen, da Faculdade de Odontologia, e o Mestrando Luciano, do Curso de Pós-Graduação em Ciência da Computação da FACIN. A entrevista foi aberta pelo Mestrando Luciano que apresentou o seguinte plano para sua condução: **1) Identificação da Área e seus membros:** O Profa. Liliane informou que ela e as professoras Karen, Maria e Fernanda compõem a Área definida como Estomatologia, sendo a Profa. Liliane Coordenadora. **2) Quais são os dados manipulados e como estes dados são armazenados atualmente:** As Professoras informaram que atualmente não existe nenhum sistema para entrada / manipulação dos dados, assim como, nenhum banco de dados para armazenamento das informações obtidas, exceto um determinado sistema que gera impressões de alguns exames, armazenando-os por um curto período de tempo. No mais, as únicas informações armazenadas atualmente, encontram-se em papel e referem-se ao prontuário do paciente, que é composto por dados cadastrais, de acompanhamento e por diversos exames em anexo. Entre estes exames, foram citados os de sangue, imagem e anatomopatológicos. A Profa. Karen informou ainda, que este prontuário não é padronizado com as demais áreas. Nesta etapa da reunião, o Mestrando Luciano elaborou questionamentos para obter as entidades, atributos e seus respectivos relacionamentos, mas para melhor entendimento, tais aspectos serão descritos em novo documento separado, onde os dados serão representados através de uma linguagem formal caracterizada por diagramas de classe. **3) Quais são os objetivos visados com a modelagem do Banco de Dados:** As professoras relataram que os principais objetivos almejados com a criação de um banco de dados são caracterizados pela possibilidade de se resgatar mais agilmente os exames vinculados ao paciente, bem como, a capacidade de se identificar indicadores até então desconhecidos. **4) Material solicitado / recebido:** Buscando melhor caracterização das entidades de dados, o Mestrando Luciano solicitou que fossem disponibilizados alguns exames (rasurados). No momento da reunião, as Professoras Liliane e Karen forneceram um prontuário em branco para o Mestrando Luciano. A Profa. Liliane comprometeu-se ainda a enviar um documento compilando as principais informações necessárias nesse prontuário. **5) Material a ser gerado:** O Mestrando Luciano informou às professoras que a partir desta reunião seriam gerados os seguintes documentos: Formulário de identificação de área, Ata e diagrama de classes. **6) Considerações finais:** O Mestrando Luciano relatou que haverá a necessidade de novas interações com os professores para esclarecimento de eventuais dúvidas, visto as restrições de tempo e disponibilidade dentro do projeto. Nada mais havendo a tratar no momento, encerrou-se a reunião às 09h50min, cuja Ata vai assinada por mim e pelos Professores da Área.

---

Liliane Soares Yurgel  
Profa. Coordenadora

---

Luciano Blomberg  
Mestrando

---

Karen Cherubini  
Professora



## APÊNDICE G – ENTREVISTA / ÁREA ODONTOLOGIA EM SAÚDE COLETIVA

### PROJETO BANCO DE DADOS EM SAÚDE BUCAL ATA 01

Aos vinte e nove dias do mês de abril de dois mil e oito, reuniram-se às 11 horas e 30 minutos, no Prédio 30 do Campus Central da PUCRS, os professores Denis, Salete e Helenita, da Faculdade de Odontologia, e o Mestrando Luciano, do Curso de Pós-Graduação em Ciência da Computação da FACIN. A entrevista foi aberta pelo Mestrando Luciano que apresentou o seguinte roteiro para sua condução: **1) Identificação da Área e seus membros:** O Profa. Helenita informou que ela e os professores Denis e Salete compõem a Área definida como Odontologia em Saúde Coletiva, sendo a Prof. Denis Coordenador. **2) Quais dados são manipulados e como estes dados são armazenados atualmente:** Os Professores informaram que a entrada de dados se dá pelo preenchimento de um prontuário familiar que é composto por prontuários individuais de cada membro da família. Dentro desses prontuários são registrados dados cadastrais e de acompanhamento do paciente, podendo conter ainda uma diversidade de exames complementares. A Professora Helenita relatou que estes prontuários são os mesmos adotados pelas unidades de saúde de Porto Alegre. Foi informado ainda, que existe um sistema (EPIBUCO) recentemente implantado para manipulação dos dados, mas que não atende a todas necessidades almejadas. O Mestrando relatou que seria interessante em uma próxima oportunidade, conhecer o funcionamento de tal sistema. **3) Quais são os objetivos visados com a modelagem do Banco de Dados:** Os professores relataram que os principais objetivos almejados com a criação de um banco de dados estão relacionados à possibilidade de acompanhar o histórico de atendimentos do paciente, mesmo em outras áreas, a recuperação ágil de informações sobre pacientes e exames complementares, bem como a produção de indicadores que possam considerar, por exemplo, a relação de doenças em determinados grupos familiares com a sua situação socioeconômica e geográfica. **4) Material solicitado / recebido:** Buscando melhor caracterização das entidades de dados, o Mestrando Luciano solicitou que fosse disponibilizado um prontuário familiar (rasurado). Os Professores comprometeram-se em disponibilizar o material solicitado até sexta-feira (02-05), na Secretaria da Faculdade de Odontologia. No momento da reunião, as Professores da área de Odontologia em Saúde Coletiva forneceram um prontuário individual do paciente para o Mestrando Luciano. **5) Material a ser gerado:** O Mestrando Luciano informou às professoras que a partir desta reunião seriam gerados os seguintes documentos: Formulário de identificação de área, Ata e diagrama de classes. **6) Considerações finais:** Identifica-se nesta área uma maior complexidade de informações em relação às áreas Básica e de Estomatologia anteriormente vistas. O Mestrando Luciano relatou, que haverá a necessidade de novas interações com os professores para esclarecimento de eventuais dúvidas, visto as restrições de tempo e disponibilidade dentro do projeto. Cabe salientar que uma parte significativa dos dados deixou de ser modelado devido a tais restrições. Nada mais havendo a tratar no momento, encerrou-se a reunião às 12h10min, cuja Ata vai assinada por mim e pelos Professores da Área.

---

Denis Marcelo Dockhorn  
Prof. Coordenador

---

Luciano Blomberg  
Mestrando

---

Salete Maria Pretto  
Professora

---

Helenita Correia Eli  
Professora



## APÊNDICE H – ENTREVISTA / ÁREA DE TRAUMA DENTO-ALVEOLAR

### PROJETO BANCO DE DADOS EM SAÚDE BUCAL ATA 01

Aos seis dias do mês de maio de dois mil e oito, reuniram-se às 10 horas e 50 minutos, no Prédio 30 do Campus Central da PUCRS, os professores João e Maristela, da Faculdade de Odontologia, e o Mestrando Luciano, do Curso de Pós-Graduação em Ciência da Computação da FACIN. A entrevista foi aberta pelo Mestrando Luciano que apresentou o seguinte roteiro para sua condução:

**1) Identificação da Área e seus membros:** A Profa. Maristela informou que ela e o Prof. João compõem a Área definida como de Trauma Dento-Alveolar. **2) Quais dados são manipulados e como estes dados são armazenados atualmente:** Os Professores informaram que atualmente não existe nenhum sistema de entrada / manipulação dos dados, assim como nenhum banco de dados para armazenamento das informações obtidas. A forma de manipulação dos dados é a mesma adotada pela maioria das Áreas dentro da Odontologia, ou seja, existe um arquivo físico de prontuários dentro desta Faculdade, que contém dados de identificação, acompanhamento e exames anexos a este prontuário. Neste caso, tem-se um interesse em especial nos exames de imagem, que registram diversos tipos de trauma. O problema é que os campos do prontuário não atendem as necessidades da Área de Trauma-Alveolar, tanto que, os Professores João e Maristela elaboraram uma ficha de dados, visando suprir informações não contempladas pelo atual prontuário. Para conhecimento, esta ficha não está sendo adotada, pois está em fase de elaboração. O Mestrando Luciano relatou que, o processo de modelagem do banco de dados em saúde bucal, visa justamente atender este tipo de necessidade, não se fazendo necessário assim, a criação de mais um documento físico **3) Quais são os objetivos visados com a modelagem do Banco de Dados:** Os Professores informaram que um dos objetivos almejados com a criação do banco de dados é a possibilidade de obter mais agilmente, informações referentes ao histórico de consultas do paciente, seus exames complementares e principalmente informações que permitam o acompanhamento da evolução do trauma ao longo do tempo. Conforme a evolução do caso pode-se ter um novo diagnóstico e conseqüentemente um novo plano de tratamento. Deseja-se ainda criar um banco para armazenar fotografias clínicas do dente, algo que não existe atualmente. **4) Material solicitado / recebido:** Buscando melhor caracterização das entidades de dados, o Mestrando Luciano solicitou que fosse disponibilizada a ficha de dados que está sendo elaborada pela Área. A Profa. Maristela informou que a ficha seria disponibilizada tão logo fosse finalizada. Para esta Área dispensou-se o pedido de prontuário, já que este é o mesmo adotado pela Área de Diagnósticos por Imagens em Odontologia. **5) Material a ser gerado:** O Mestrando Luciano informou às professoras que a partir desta reunião seriam gerados os seguintes documentos: Formulário de identificação de área, Ata e diagrama de classes. **6) Considerações finais:** O Mestrando Luciano relatou que haverá a necessidade de novas interações com os professores para esclarecimento de eventuais dúvidas, visto as restrições de tempo e disponibilidade dentro do projeto.

Nada mais havendo a tratar no momento, encerrou-se a reunião às 12h, cuja Ata vai assinada por mim e pelos Professores da Área.

---

João Batista Blessmann Weber  
Professor

---

Luciano Blomberg  
Mestrando

---

Maristela Borba  
Professora





## APÊNDICE I - PRIMEIRO EXPERIMENTO – MP I

**Propósito de Mineração:** Classificação

**Técnica:** Árvore de Decisão – Predição de Incidência de Cárie Dentária

**Algoritmo:** J48

**Conjunto de dados:** EXPMINER01.CSV

**Instâncias:** 174

**Atributos Selecionados:** 11 (escolaridade, usacremedental, utiliza\_fluoreto, jafoi\_aodentista, motivo\_ida\_aodentista, frequência\_consumo\_acucar, alergias\_pó, histmed\_hipertensão, histmed\_cholesterol\_alta, histmed\_asma, carie).

**Atributo Classe:** Cárie

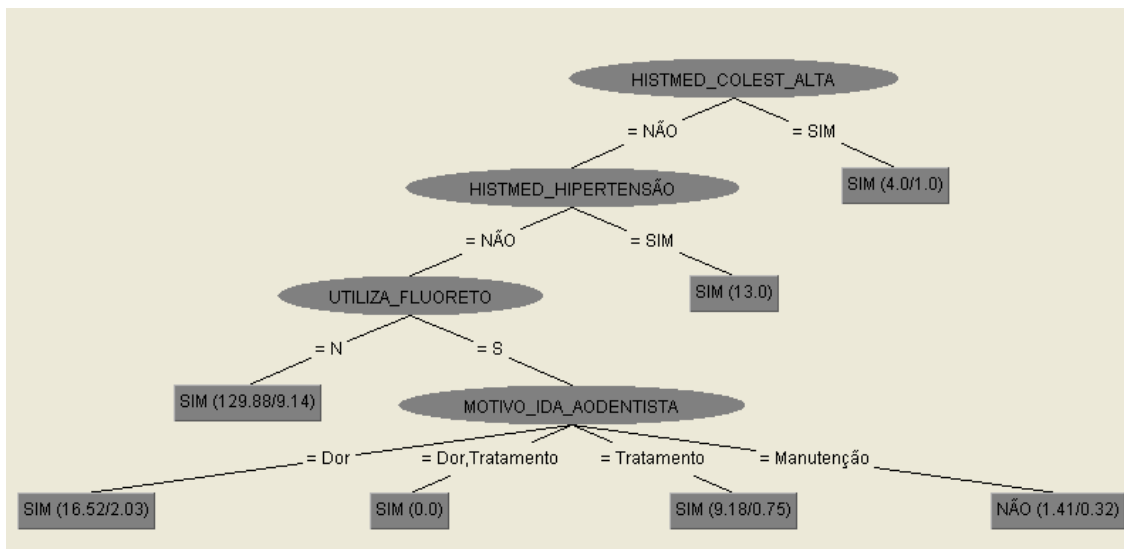
**Balanceamento:** Não

**Feature Selection:** Sim. Atributos selecionados manualmente.

**MinNumObj:** 2

**Unpruned:** true

### Árvore Gerada:



=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	160	91.954 %
Incorrectly Classified Instances	14	8.046 %
Kappa statistic	0	
Mean absolute error	0.1463	
Root mean squared error	0.2739	
Relative absolute error	95.9499 %	
Root relative squared error	100.5818 %	
Total Number of Instances	174	

=== Confusion Matrix ===

```

a  b  <-- classified as
160  0 |  a = SIM
14   0 |  b = NÃO
  
```

**Frequência do atributo classe**

**SIM: 160**

**NÃO: 14**



## APÊNDICE J - PRIMEIRO EXPERIMENTO – MP II

**Propósito de Mineração:** Classificação

**Técnica:** Árvore de Decisão – Predição de Indicador de Cárie Dentária

**Algoritmo:** J48

**Conjunto de dados:** EXPMINER01.CSV

**Instâncias:** 174

**Atributos Seleccionados:** 12 (idade, escolaridade, usacremedental, utiliza\_fluoreto, jafoi\_aodentista, motivo\_ida\_aodentista, frequência\_consumo\_acucar, alergias\_pó, histmed\_hipertensão, histmed\_coolest\_alta, histmed\_asma, case).

**Atributo Classe:** Case (Indicador de Cárie)

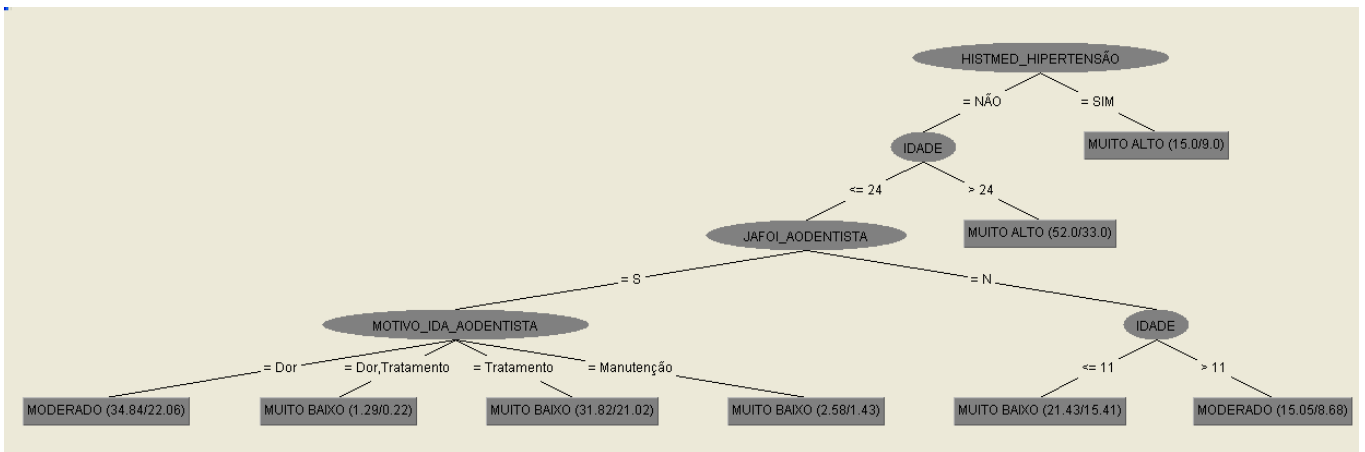
**Balanceamento:** Não

**Feature Selection:** Sim. Atributos seleccionados manualmente.

**MinNumObj:** 15

**Unpruned:** true

### Árvore Gerada:



### === Stratified cross-validation ===

#### === Summary ===

Correctly Classified Instances	49	28.1609 %
Incorrectly Classified Instances	125	71.8391 %
Kappa statistic	0.0541	
Mean absolute error	0.3121	
Root mean squared error	0.4031	
Relative absolute error	99.3348 %	
Root relative squared error	101.7175 %	
Total Number of Instances	174	

### === Confusion Matrix ===

a	b	c	d	e	<-- classified as
19	16	0	0	8	a = MODERADO
14	18	0	3	12	b = MUITO BAIXO
8	5	0	0	7	c = ALTO
14	10	0	0	6	d = BAIXO
10	12	0	0	12	e = MUITO ALTO

### Frequência do atributo classe

<b>MUITO BAIXO:</b>	<b>47</b>
<b>BAIXO:</b>	<b>43</b>
<b>MODERADO:</b>	<b>34</b>
<b>ALTO:</b>	<b>30</b>
<b>MUITO ALTO:</b>	<b>20</b>



## APÊNDICE K – SEGUNDO EXPERIMENTO – MP I

**Propósito de Mineração:** Classificação

**Técnica:** Árvore de Decisão – Predição de Indicador de Cárie Dentária Convencional

**Algoritmo:** J48

**Conjunto de dados:** EXPMINER02\_INDICADOR.CSV

**Instâncias:** 235

**Atributos Selecionados:** 15(escolaridade, empregado, usacremedental, escova\_sozinho, escova\_maisde1x, roe\_unhas, come\_doces, naousa\_fiodental, fuma, utiliza\_fluoreto, jafoi\_aodontista, frequência\_consumo\_acucar2, arcada\_superior, arcada\_inferior, case)

**Atributo Classe:** Case(Indicador\_Cárie)

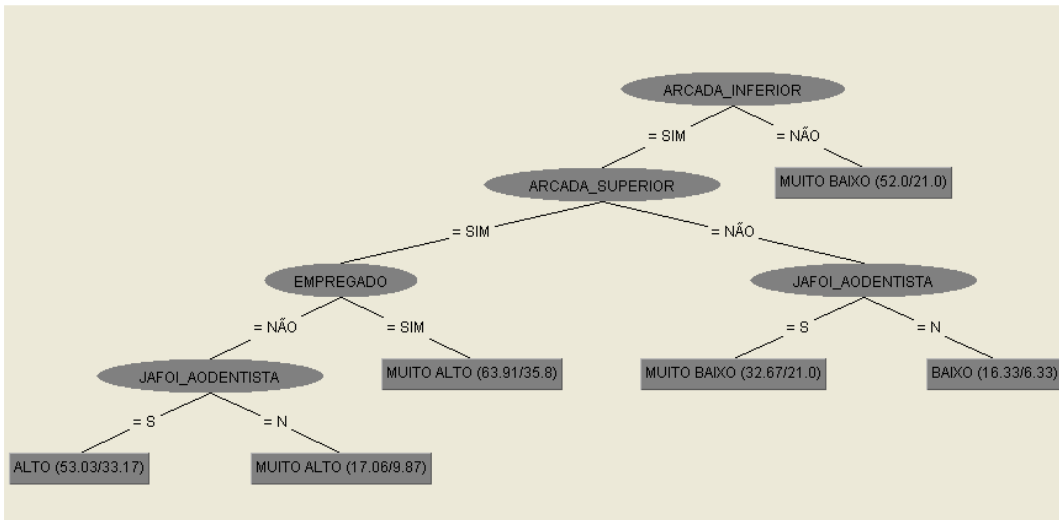
**Balanceamento:** Sim

**Feature Selection:** Sim. Filtro ChiSquaredAttributeEval da ferramenta Weka.

**MinNumObj:** 12

**Unpruned:** true

### Árvore Gerada:



### === Stratified cross-validation ===

#### === Summary ===

Correctly Classified Instances	101	42.9787 %
Incorrectly Classified Instances	134	57.0213 %
Kappa statistic	0.2872	
Mean absolute error	0.2713	
Root mean squared error	0.3701	
Relative absolute error	84.7651 %	
Root relative squared error	92.505 %	
Total Number of Instances	235	

### === Confusion Matrix ===

```

a b c d e <-- classified as
11 4 26 2 4 | a = ALTO
13 2 17 11 4 | b = MODERADO
8 1 37 1 0 | c = MUITO ALTO
0 0 0 38 9 | d = MUITO BAIXO
5 1 9 19 13 | e = BAIXO
  
```

### Frequência do atributo classe

```

MUITO BAIXO: 47
BAIXO: 47
MODERADO: 47
ALTO: 47
MUITO ALTO: 47
  
```



## APÊNDICE L - SEGUNDO EXPERIMENTO – MP II

**Propósito de Mineração:** Classificação

**Técnica:** Árvore de Decisão – Predição de Indicador de Cárie Dentária Binarizado

**Algoritmo:** J48

**Conjunto de dados:** EXPMINER02\_INDICADORBIN.CSV

**Instâncias:** 188

**Atributos Selecionados:** 14 (escolaridade, empregado, usacremedental, escova\_sozinho, escova\_maisde1x, come\_doces, toma\_mamadeira, usa\_fiodental, naousa\_fiodental, toma\_cafe, jafoi\_aodontista, motivo\_ida\_aodontista2, frequência\_consumo\_acucar2, case2)

**Atributo Classe:** Case2(Indicador\_Cárie)

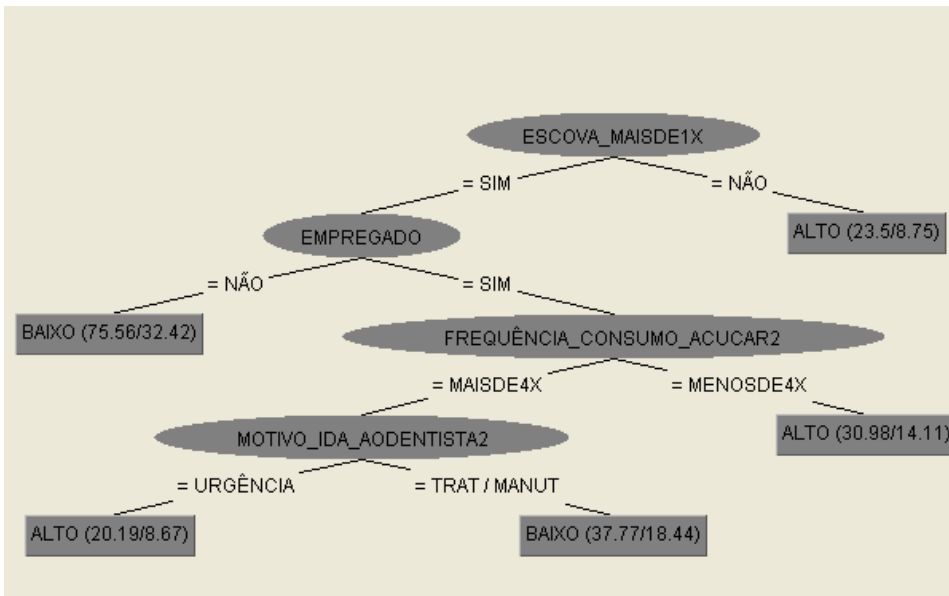
**Balanceamento:** Sim

**Feature Selection:** Sim. Filtro ChiSquaredAttributeEval da ferramenta Weka.

**MinNumObj:** 11

**Unpruned:** true

### Árvore Gerada:



### === Stratified cross-validation ===

#### === Summary ===

Correctly Classified Instances	95	50.5319 %
Incorrectly Classified Instances	93	49.4681 %
Kappa statistic	0.0106	
Mean absolute error	0.5018	
Root mean squared error	0.504	
Relative absolute error	100.3415 %	
Root relative squared error	100.7655 %	
Total Number of Instances	188	

### === Confusion Matrix ===

```

a b  <-- classified as
40 54 | a = ALTO
39 55 | b = BAIXO
  
```

### Frequência do atributo classe

**BAIXO:** 94  
**ALTO:** 94





## APÊNDICE M - SEGUNDO EXPERIMENTO – MP III

**Propósito de Mineração:** Classificação

**Técnica:** Árvore de Decisão — Predição de Indicador de Cárie Dentária Binarizado

**Algoritmo:** J48

**Conjunto de dados:** EXPMINER02\_INDICADORBIN.CSV

**Instâncias:** 188

**Atributos Selecionados:** 12 (escova\_maisde1x, escova\_sozinho, empregado, jafoi\_aodontista, frequência\_consumo\_acucar2, usacremedental, escolaridade, come\_doces, toma\_chimarrão, toma\_cafe, usa\_fiodental, case2)

**Atributo Classe:** Case2(Indicador\_Cárie)

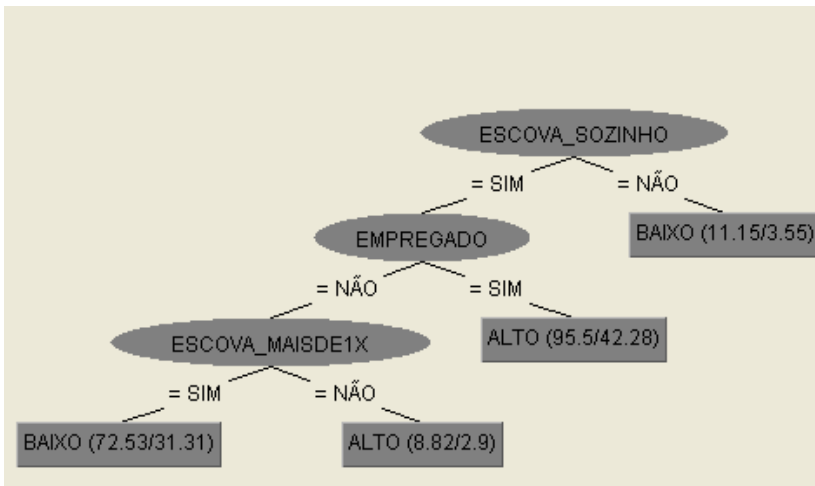
**Balanceamento:** Sim

**Feature Selection:** Sim. Filtro ChiSquaredAttributeEval da ferramenta Weka.

**MinNumObj:** 6

**Unpruned:** true

**Árvore Gerada:**



=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	98	52.1277 %
Incorrectly Classified Instances	90	47.8723 %
Kappa statistic	0.0426	
Mean absolute error	0.5005	
Root mean squared error	0.5034	
Relative absolute error	100.0813 %	
Root relative squared error	100.6499 %	

Total Number of Instances 188

=== Confusion Matrix ===

```

a  b  <-- classified as
43 51 | a = ALTO
39 55 | b = BAIXO
  
```

**Frequência do atributo classe**

**BAIXO: 94**  
**ALTO: 94**



## APÊNDICE N - SEGUNDO EXPERIMENTO - MP IV

**Propósito de Mineração:** Classificação

**Técnica:** Árvore de Decisão – Predição de Incidência de Má-Oclusão

**Algoritmo:** J48

**Conjunto de dados:** EXPMINER02\_OCLUSAO.CSV

**Instâncias:** 68

**Atributos Selecionados:** 10 (maoclusao\_nafamilia, toma\_mamadeira, morde\_lápis, naousa\_fiodental, usa\_fiodental, idade, tem\_bruxismo, chupa\_bico, toma\_chimarrão, oclusao)

**Atributo Classe:** Oclusão

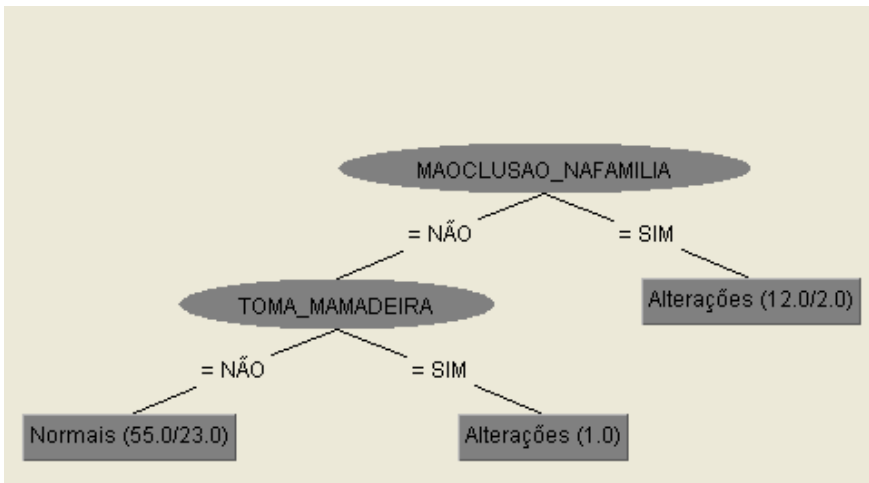
**Balanceamento:** Sim

**Feature Selection:** Sim. Filtro ChiSquaredAttributeEval da ferramenta Weka.

**MinNumObj:** 1

**Unpruned:** true

**Árvore Gerada:**



=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	37	54.4118 %
Incorrectly Classified Instances	31	45.5882 %
Kappa statistic	0.0882	
Mean absolute error	0.4691	
Root mean squared error	0.5221	
Relative absolute error	93.6352 %	
Root relative squared error	104.2091 %	
Total Number of Instances	68	

=== Confusion Matrix ===

```

a b  <-- classified as
10 24 | a = Alterações
 7 27 | b = Normais
  
```

**Frequência do atributo classe**

**ALTERAÇÕES: 34**  
**NORMAIS: 34**



## APÊNDICE O - SEGUNDO EXPERIMENTO - MP V

**Propósito de Mineração:** Classificação

**Técnica:** Árvore de Decisão – Predição de Incidência de Doenças Periodontais

**Algoritmo:** J48

**Conjunto de dados:** ExpMiner02\_Periodont.CSV

**Instâncias:** 68

**Atributos Selecionados:** 16 (idade, doencaperiodont\_nafamilia, escova\_sozinho, usacremedental, escova\_maisde1x, frequência\_consumo\_acucar2, chupa\_bico, tem\_bruxismo, toma\_mamadeira, escolaridade, roe\_unhas, morde\_lápis, chupa\_dedo, naousa\_fiodental, usa\_fiodental, atividade\_doenca\_periodontal)

**Atributo Classe:** Atividade\_doenca\_periodontal

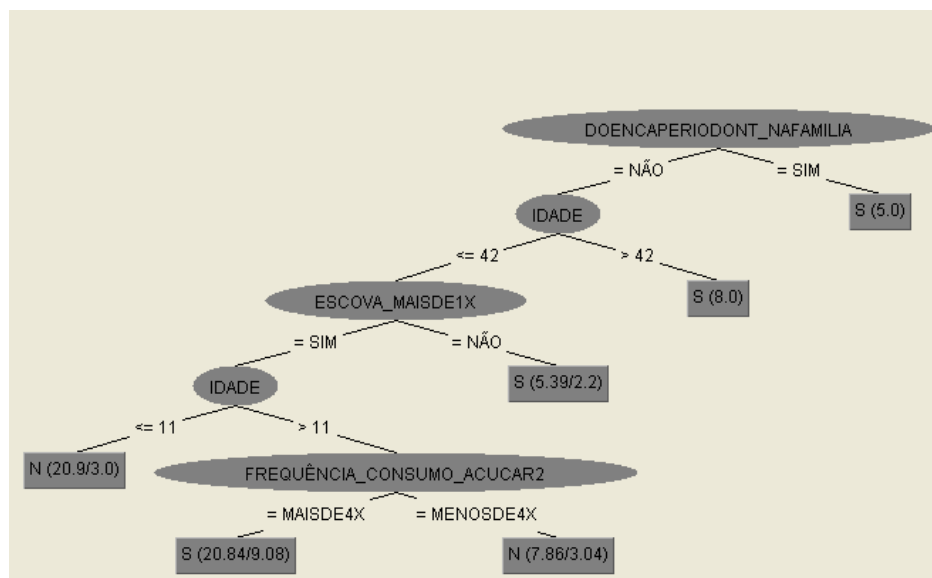
**Balanceamento:** Sim

**Feature Selection:** Sim. Filtro ChiSquaredAttributeEval da ferramenta Weka.

**MinNumObj:** 5

**Unpruned:** true

### Árvore Gerada:



==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	45	66.1765 %
Incorrectly Classified Instances	23	33.8235 %
Kappa statistic	0.3235	
Mean absolute error	0.4126	
Root mean squared error	0.4828	
Relative absolute error	82.3646 %	
Root relative squared error	96.3682 %	
Total Number of Instances	68	

==== Confusion Matrix ====

```

a b <-- classified as
25 9 | a = N
14 20 | b = S
  
```

**Frequência do atributo classe**

**S: 34**  
**N: 34**



## APÊNDICE P - TERCEIRO EXPERIMENTO - MD I

**Propósito de Mineração:** Análise Associativa

**Técnica:** Extração de Regras de Associação – Descrição de Características Associadas à Cárie Dentária

**Algoritmo:** APriori

**Conjunto de dados:** EXPMINER03

**Arquivo:** ..\experimentos\técnicas\j48\sibalanc\expminer03IndicadorCarie.arff

**Instâncias:** 235

**Atributos Selecionados:** 59 (faixaidade, escova\_maisde1x, escova\_sozinho, utiliza\_fluoreto, empregado, requência\_consumo\_acucar, motivo\_ida\_aodentista2, jafoi\_aodentista, frequência\_consumo\_acucar2, usacremedental, dentecpo17, dentecpo16, dentecpo18, come\_doces, toma\_cafe, usa\_fiodental, dentecpo83, dentecpo42, dentecpo31, dentecpo41, dentecpo85, dentecpo46, dentecpo45, dentecpo43, dentecpo44, dentecpo84, dentecpo75, dentecpo74, dentecpo35, dentecpo38, dentecpo36, dentecpo37, dentecpo71, dentecpo32, dentecpo34, dentecpo33, dentecpo73, dentecpo11, dentecpo12, dentecpo52, dentecpo21, dentecpo14, dentecpo15, dentecpo55, dentecpo53, dentecpo54, dentecpo13, dentecpo27, dentecpo65, dentecpo26, dentecpo47, dentecpo28, dentecpo48, dentecpo23, dentecpo22, dentecpo64, dentecpo25, dentecpo24, case2)

**Atributo Classe:** Case2 (Indicador de Cárie Dentária)

**Balanceamento:** Sim

**Feature Selection:** Sim. Filtro ChiSquaredAttributeEval da ferramenta Weka.

**Frequência do atributo classe**

**BAIXO: 94**

**ALTO: 94**

**Apriori**

=====

1. DENTECPO15=SIM 38 ==>	CASE2=ALTO 38	conf:(1)
2. DENTECPO26=SIM, DENTECPO47=SIM 33 ==>	CASE2=ALTO 33	conf:(1)
3. DENTECPO46=SIM, DENTECPO26=SIM 29 ==>	CASE2=ALTO 29	conf:(1)
4. DENTECPO15=SIM, DENTECPO26=SIM 29 ==>	CASE2=ALTO 29	conf:(1)
5. DENTECPO37=SIM, DENTECPO26=SIM 26 ==>	CASE2=ALTO 26	conf:(1)
6. DENTECPO17=SIM, DENTECPO26=SIM 25 ==>	CASE2=ALTO 25	conf:(1)
7. DENTECPO27=SIM, DENTECPO26=SIM 25 ==>	CASE2=ALTO 25	conf:(1)
8. DENTECPO26=SIM, DENTECPO25=SIM 25 ==>	CASE2=ALTO 25	conf:(1)
9. JAFOI_AODENTISTA=S, DENTECPO15=SIM 22 ==>	CASE2=ALTO 22	conf:(1)
10. ESCOVA_MAISDE1X=SIM, USACREMEDENTAL=S, DENTECPO26=SIM, DENTECPO47=SIM 22 ==>	CASE2=ALTO 22	conf:(1)
11. JAFOI_AODENTISTA=S, DENTECPO35=SIM 21 ==>	CASE2=ALTO 21	conf:(1)
12. DENTECPO16=SIM, DENTECPO46=SIM 21 ==>	CASE2=ALTO 21	conf:(1)
13. ESCOVA_MAISDE1X=SIM, ESCOVA_SOZINHO=SIM, USACREMEDENTAL=S, DENTECPO17=SIM 19 ==>	CASE2=ALTO 19	conf:(1)
14. ESCOVA_SOZINHO=SIM, USACREMEDENTAL=S, DENTECPO27=SIM, DENTECPO26=SIM 19 ==>	CASE2=ALTO 19	conf:(1)





**ANEXO A – SOLICITAÇÃO FACULDADE DE INFORMÁTICA**

Ao Prof. Dr. Avelino Francisco Zorzo  
M.D. Diretor da Faculdade de Informática da PUCRS

Senhor Diretor

Mediante um novo cenário, caracterizado pela oferta de uma sala junto à Faculdade de Odontologia para o processo de apreensão de dados (projeto BDSB), gostaríamos de retificar nosso pedido anterior, solicitando autorização para transferirmos temporariamente o microcomputador (w113613-4) e o monitor (113614-2), adquiridos via projeto FAPERGS, para a Faculdade de Odontologia. Adicionalmente precisaríamos transferir mais duas CPUs (w140584-4, w136697-0) e dois monitores (140585-1, 140586-9), pois embora a sala disponibilizada ofereça melhores condições de trabalho (espaço, segurança e suporte), essa ainda não conta com microcomputadores.

Desde já agradecemos pela atenção dispensada,

Luciano Costa Blomberg  
Mestrando.

Prof. Dr. Duncan Dubugras Alcoba Ruiz  
Orientador.

Prof. Dr. José Antônio Poli de Figueiredo  
Coordenador do Programa de Pós-Graduação em Odontologia



**ANEXO B – SOLICITAÇÃO FACULDADE DE ODONTOLOGIA**

Ao Prof. Dr. Prof. Marcos Túlio Mazzini Carvalho  
M.D. Diretor da Faculdade de Odontologia da PUCRS

Senhor Diretor

Solicitamos autorização para o envio dos microcomputadores (w113613-4, w140584-4, w136697-0) e dos monitores (113614-2, 140585-1, 140586-9), para a sala 507.03 da Faculdade de Odontologia, visando viabilizar o processo de apreensão de prontuários odontológicos junto ao CEU Vila Fátima. Destacamos que esta atividade está inserida dentro do projeto BDSB (Banco de Dados em Saúde Bucal) desenvolvido em parceria entre a FACIN e a FOPUCRS.

Durante o período de dois meses, o Mestrando em Ciência da Computação Luciano Costa Blomberg atuará nesta atividade juntamente com duas bolsistas vinculadas à FOPUCRS e supervisão da Profa. Dra. Helenita Correia Eli. A princípio, nosso intuito é que os microcomputadores permaneçam durante este mesmo período.

Desde já agradecemos pela atenção dispensada,

Luciano Costa Blomberg  
Mestrando.

Prof. Dr. Duncan Dubugras Alcoba Ruiz  
Orientador.

Prof. Dr. José Antônio Poli de Figueiredo  
Coordenador do Programa de Pós-Graduação em Odontologia

# Livros Grátis

( <http://www.livrosgratis.com.br> )

Milhares de Livros para Download:

[Baixar livros de Administração](#)

[Baixar livros de Agronomia](#)

[Baixar livros de Arquitetura](#)

[Baixar livros de Artes](#)

[Baixar livros de Astronomia](#)

[Baixar livros de Biologia Geral](#)

[Baixar livros de Ciência da Computação](#)

[Baixar livros de Ciência da Informação](#)

[Baixar livros de Ciência Política](#)

[Baixar livros de Ciências da Saúde](#)

[Baixar livros de Comunicação](#)

[Baixar livros do Conselho Nacional de Educação - CNE](#)

[Baixar livros de Defesa civil](#)

[Baixar livros de Direito](#)

[Baixar livros de Direitos humanos](#)

[Baixar livros de Economia](#)

[Baixar livros de Economia Doméstica](#)

[Baixar livros de Educação](#)

[Baixar livros de Educação - Trânsito](#)

[Baixar livros de Educação Física](#)

[Baixar livros de Engenharia Aeroespacial](#)

[Baixar livros de Farmácia](#)

[Baixar livros de Filosofia](#)

[Baixar livros de Física](#)

[Baixar livros de Geociências](#)

[Baixar livros de Geografia](#)

[Baixar livros de História](#)

[Baixar livros de Línguas](#)

[Baixar livros de Literatura](#)  
[Baixar livros de Literatura de Cordel](#)  
[Baixar livros de Literatura Infantil](#)  
[Baixar livros de Matemática](#)  
[Baixar livros de Medicina](#)  
[Baixar livros de Medicina Veterinária](#)  
[Baixar livros de Meio Ambiente](#)  
[Baixar livros de Meteorologia](#)  
[Baixar Monografias e TCC](#)  
[Baixar livros Multidisciplinar](#)  
[Baixar livros de Música](#)  
[Baixar livros de Psicologia](#)  
[Baixar livros de Química](#)  
[Baixar livros de Saúde Coletiva](#)  
[Baixar livros de Serviço Social](#)  
[Baixar livros de Sociologia](#)  
[Baixar livros de Teologia](#)  
[Baixar livros de Trabalho](#)  
[Baixar livros de Turismo](#)