

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Programa de Pós-graduação em
Engenharia de Produção e Sistemas

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS EM PERIODONTIA

Dissertação de Mestrado

Hubert Chamone Gesser

Florianópolis

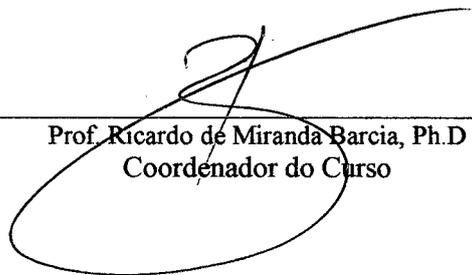
2001

Hubert Chamone Gesser

**UMA APLICAÇÃO DE MINERAÇÃO DE DADOS
EM PERIODONTIA**

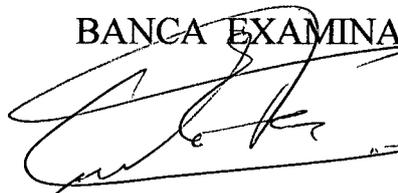
Esta dissertação foi julgada aprovada para a
obtenção do título de Mestre em Engenharia de
Produção no Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina

Florianópolis, 20 de julho de 2001.

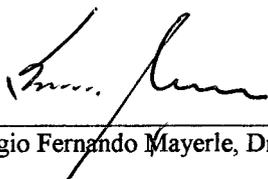


Prof. Ricardo de Miranda Barcia, Ph.D
Coordenador do Curso

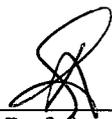
BANCA EXAMINADORA



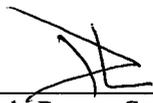
Prof. Carlos Ernani Fries, M.Eng.
Orientador



Prof. Sérgio Fernando Mayerle, Dr.



Prof. Aran Bey Tcholakian Morales, Dr.



Prof. Paulo Renato Corrêa Glavam, Dr.



Prof. Alexandre Leopoldo Gonçalves, M.Eng.

Habla Germano Cesar

INSTITUTO DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE FEDERAL DE SANTA CATARINA

Esta dissertação foi julgada aprovada para a
obtenção do título de Mestre em Engenharia de
Produção no Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina

Florianópolis, 20 de julho de 2001.

Prof. Raul de Almeida Júnior, PhD.
Coordenador do Curso

BANCA EXAMINADORA

Prof. Carlos Euzébio de Eng.
Orientador

Prof. Alan Rey Eschol. Matemática D.

Prof. Sérgio Fernando Alves, Dr.

Prof. Alexandre Roberto Gonçalves, M. Eng.

Prof. Carlos Roberto Gonçalves, M. Eng.

**UMA APLICAÇÃO DE MINERAÇÃO
DE DADOS EM PERIODONTIA**

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Programa de Pós-graduação em
Engenharia de Produção e Sistemas

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS EM PERIODONTIA

Hubert Chamone Gesser

Dissertação apresentada ao
Programa de Pós-graduação em
Engenharia de Produção e Sistemas da
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do título de mestre em
Engenharia de Produção.

Florianópolis

2001

Aos meus pais, Nilton e Márcia,
e à minha namorada Ivana.

AGRADECIMENTOS

À Universidade Federal de Santa Catarina.

Ao orientador Prof. Carlos Ernani Fries,
pelo acompanhamento pontual e competente.

Aos professores do Curso de Pós-graduação em
Engenharia de Produção e Sistemas.

À minha namorada, Ivana Corrêa de Oliveira,
pela paciência e prestatividade no desenrolar deste trabalho.

Aos meus pais e irmãos
por nunca deixar faltar amor, carinho e confiança.

Aos colegas Caroline, Alexandre e Fernando
pela colaboração e apoio.

A todos que direta ou indiretamente
contribuíram para a realização
desta pesquisa.

SUMÁRIO

1. INTRODUÇÃO	01
1.1 Objetivos e Hipóteses	02
1.2 Limitações do Estudo	03
1.3 Estrutura do Trabalho	04
2. O PERIODONTO E SUA EPIDEMIOLOGIA	05
2.1 A Doença Periodontal	06
2.1.1 A Epidemiologia das Doenças Periodontais	08
2.1.2 O Nível Socioeconômico como Fator de Influência na Doença Periodontal	13
2.1.3 O Fumo como Fator de Influência na Doença Periodontal	15
3. A ANÁLISE DE BANCO DE DADOS	19
3.1 A Análise Estatística de Dados pelo Teste de Associação Qui-Quadrado	19
3.2 A Análise de Dados através de Modelos de Inteligência Artificial	21
3.2.1 O Processo KDD – Knowledge Discovery in Databases	23
3.2.1.1 Regras de Associação	28
3.2.1.2 Classificação	30
3.2.1.3 Agrupamento (Clustering)	35
3.2.1.4 Regressão Linear	35
3.2.2 Ferramentas de Análise de Dados	36
4. METODOLOGIA EMPREGADA NO ESTUDO	38
4.1 População de Estudo	38
4.2 O Banco de Dados	38
4.2.1 A Composição do Banco de Dados	40
4.3 Plano de Análise do Banco de Dados	40
4.3.1 Análise Estatística	41
4.3.2 Análise pelo Processo KDD	41
5 APRESENTAÇÃO DOS RESULTADOS E DISCUSSÃO	43
5.1 Avaliação dos Ruídos	43
5.2 Distribuição de Frequência dos Dados Socioeconômicos	43
5.2.1 Escolaridade dos Pais e das Mães dos Alistandos	43
5.2.2 Escolaridade dos Alistandos	47
5.2.3 A Renda Familiar dos Alistandos	49
5.3 Distribuição de Frequência para o Tabagismo	51
5.4 Distribuição de Frequência dos Dados Clínicos	52
5.5 Análise Estatística	56
5.6 Análise pelo Processo KDD	64
5.7 Conclusões	68
6 CONCLUSÕES E RECOMENDAÇÕES	70
6.1 Conclusões	70
6.2 Recomendações	71
REFERÊNCIAS BIBLIOGRÁFICAS	72
APÊNDICES	
ANEXOS	

LISTA DE ILUSTRAÇÕES

Figura 1: As estruturas do periodonto	05
Figura 2: Códigos do índice CPI da Organização Mundial da Saúde	08
Figura 3: Distribuição de freqüência percentual de cálculo dentário, sangramento gengival e bolsas periodontais em recrutas de 20 anos de idade na Finlândia	09
Figura 4: Distribuição de freqüência percentual das condições de saúde periodontal em jovens brasileiros de 15 a 19 anos	10
Figura 5: Distribuição de freqüência percentual das condições de saúde periodontal em jovens da Região Sul do Brasil de 15 a 19 anos ...	11
Figura 6: Exemplo de tabela de contingência 2x2	19
Figura 7: Evolução da extração de conhecimento em base de dados	22
Figura 8: Etapas do processo de descoberta de conhecimento	25
Figura 9: Exemplo de Árvore de Decisão	33
Figura 10: Anos de estudo dos pais dos alistandos	46
Figura 11: Anos de estudo das mães dos alistandos	46
Figura 12: Anos de estudo dos alistandos.....	48
Figura 13: Comparação do grau de escolaridade dos alistandos com os pais e as mães	49
Figura 14: Freqüência percentual da renda familiar em salários mínimos e curva de tendência polinomial	51
Figura 15: Proporções de alistandos fumantes, não fumantes e ex-fumantes	52
Figura 16: Prevalências das condições periodontais	55
Figura 17: Distribuição percentual dos componentes do CPI	55

LISTA DE TABELAS

Tabela 1: Condições Periodontais observadas e mensuradas através do CPITN nas idades de 15-19 anos	12
Tabela 2: Distribuição de freqüência para a pergunta “Até que série seu pai estudou na escola?”	45
Tabela 3: Distribuição de freqüência para a pergunta “Até que série sua mãe estudou na escola?”	45
Tabela 4: Distribuição de freqüência para a pergunta “Até que série você estudou na escola?”	47
Tabela 5: Distribuição de freqüência da renda familiar dos alistandos	50
Tabela 6: Distribuição de freqüência do hábito de fumar dos alistandos em resposta a pergunta “Você fuma cigarro?”	51
Tabela 7: Distribuição de freqüência dos alistandos com e sem sangramento gengival	53
Tabela 8: Distribuição de freqüência do número de sextantes com sangramento gengival	53
Tabela 9: Distribuição de freqüência dos alistandos com e sem cálculo dentário..	53
Tabela 10: Distribuição de freqüência do número de sextantes com cálculo dentário	54
Tabela 11: Distribuição de freqüência dos componentes do CPI na amostra	54
Tabela 12: Sangramento Gengival segundo os anos de estudo do alistando	57
Tabela 13: Cálculo Dentário segundo os anos de estudo do alistando	58
Tabela 14: Bolsas Periodontais segundo os anos de estudo do alistando	58
Tabela 15: Sangramento Gengival segundo os anos de estudo do pai do alistando	58
Tabela 16: Cálculo Dentário segundo os anos de estudo do pai do alistando	59
Tabela 17: Bolsas Periodontais segundo os anos de estudo do pai do alistando	59
Tabela 18: Sangramento Gengival segundo os anos de estudo da mãe do alistando	59
Tabela 19: Cálculo Dentário segundo os anos de estudo da mãe do alistando	60
Tabela 20: Bolsas Periodontais segundo os anos de estudo da mãe do alistando	60
Tabela 21: Sangramento Gengival segundo a renda familiar	61
Tabela 22: Cálculo Dentário segundo a renda familiar	61
Tabela 23: Bolsas Periodontais segundo a renda familiar	62
Tabela 24: Sangramento Gengival segundo o hábito de fumar cigarro	62
Tabela 25: Cálculo Dentário segundo o hábito de fumar cigarro	63
Tabela 26: Bolsas Periodontais segundo o hábito de fumar cigarro	63
Tabela 27: Resumo das associações estatísticas entre os fatores socioeconômicos e o tabagismo com os dados clínicos da doença periodontal	64
Tabela 28: Regras de classificação para o atributo alvo Sangramento Gengival no alistando	65
Tabela 29: Regras de classificação para o atributo alvo Cálculo Dentário no alistando	66
Tabela 30: Regras de classificação para o atributo alvo Bolsas Periodontais no alistando	68

LISTA DE REDUÇÕES

CPI = Community Periodontal Index

(Índice Periodontal Comunitário)

CPITN = Community Periodontal Index of Treatment Needs

(Índice Periodontal Comunitário e Necessidades de Tratamento)

EUA = Estados Unidos da América

IBGE = Fundação Instituto Brasileiro de Geografia e Estatística

IgA = Imunoglobulinas tipo A

IgG = Imunoglobulinas tipo G

KDD = Knowledge Discovery in Databases

(Descoberta de Conhecimento em Banco de Dados)

LHS = Left Hand Side

(Lado Esquerdo da Regra)

NOHS = National Oral Health Search

(Pesquisa Nacional de Saúde Oral, África do Sul)

OLAP = On Line Analytical Process

OMS = Organização Mundial da Saúde

PNAD = Pesquisa Nacional por Amostra de Domicílios

RHS = Right Hand Side

(Lado Direito da Regra)

RESUMO

GESSER, Hubert Chamone, Uma Aplicação de Mineração de Dados em Periodontia. Florianópolis, 2001. 78f. Dissertação (Mestrado em Engenharia de Produção) Programa de Pós-graduação em Engenharia de Produção, UFSC, 2001.

O presente estudo tem como propósito conhecer a prevalência da doença periodontal e sua associação com questões socioeconômicas e com o hábito de fumar. Análises estatísticas e técnicas de regras de classificação pelo processo KDD (*Knowledge Discovery in Databases*) foram empregadas em um banco de dados dos alistandos do Exército em Florianópolis, Santa Catarina, Brasil, em 1999. O banco de dados contém informações obtidas através da aplicação de um questionário socioeconômico e realização um exame clínico para coleta de dados como: o sangramento gengival, a presença de cálculo e componentes do índice CPI (Community Periodontal Index). Os testes de associação estatística (Teste do Qui-quadrado e o Teste Exato de Fisher) mostram que as variáveis socioeconômicas como a escolaridade e a renda familiar, bem como o hábito de fumar cigarro, têm associação com as doenças periodontais. A mineração de dados do processo KDD mostra que os fatores socioeconômicos e o tabagismo estão presentes em muitas das regras de classificação geradas. Com a baixa ocorrência de problemas periodontais avançados nesta faixa etária (18 anos de idade), conclui-se que a doença periodontal mais severa não é um problema de saúde pública nesta amostra, apesar da mesma estar associada com questões socioeconômicas. Relata-se ainda que, devido à inexistência de outros estudos com esta população, faz-se necessária a confirmação destes achados com trabalhos posteriores.

Palavras-chave: Epidemiologia da Doença Periodontal, Data Mining, KDD.

ABSTRACT

GESSER, Hubert Chamone, Uma Aplicação de Mineração de Dados em Periodontia. Florianópolis, 2001. 78f. Dissertação (Mestrado em Engenharia de Produção) Programa de Pós-graduação em Engenharia de Produção, UFSC, 2001.

The present study has the purpose to know the prevalence of periodontal diseases and their association with socioeconomic level and smoking habits. Statistical analysis and technical classification rules of KDD (Knowledge Discovery in Databases) was used in database of army conscripts, in Florianópolis, Santa Catarina, Brazil, in 1999. The database contains information obtained through a questionnaire and a clinic exam was carried out to collect data, such as: gingival bleeding, calculus and CPI index components (Community Periodontal Index). Statistical association tests (Chi-square Test and Fisher's Exact Test) show that socioeconomic variables (education level and family income), as well as the smoking cigarette habit, have association with periodontal diseases. The Data Mining of KDD process shows that socioeconomic factors and smoking habits are present in many of the generated classification rules. The low occurrence of serious periodontal problems in this age group (18 years old) shows that periodontal diseases are not a problem of public health in this sample, despite their association with socioeconomic factors. The lack of similar studies in this kind of population is a flagrant evidence that additional studies are necessary to confirm these discoveries.

Keywords: Periodontal Diseases / Epidemiology, Data Mining, KDD.

1. INTRODUÇÃO

A palavra epidemiologia significa, originalmente, a ciência das epidemias. Na Medicina, bem como na Odontologia, os métodos epidemiológicos são utilizados para determinar a proporção de uma população que, em determinado tempo, é afetada por uma doença. O conhecimento deste percentual de indivíduos doentes é denominado prevalência da doença.

As doenças bucais vêm ganhando espaço ultimamente, como instrumento de pesquisa em levantamentos epidemiológicos de saúde bucal. Isto se deve ao fato dos altos índices de prevalência encontrados na população, caracterizando-as como sendo um problema de saúde pública, gerando graves conseqüências socioeconômicas.

A realização de estudos dentro de uma metodologia coerente com os princípios epidemiológicos, proporciona subsídios corretos ao bom planejamento e programação em saúde, através do conhecimento da real condição de saúde da população.

Relata-se na literatura que as doenças bucais mais comumente encontradas são a cárie dentária e as doenças relacionadas às gengivas, denominadas doenças periodontais, pois atingem a maioria das pessoas caracterizando assim, um quadro de alta prevalência.

A pouca existência de dados nacionais e locais a respeito do modo como as doenças periodontais se distribuem na população, em especial em adultos jovens, justifica a busca de informações através deste estudo realizado no banco de dados dos alistados do Exército Brasileiro em Florianópolis, onde através das análises estatísticas e das aplicações de técnicas de descoberta de conhecimento em banco de dados, entre elas o *Data Mining*, é possível identificar os grupos mais afetados.

Com as possibilidades modernas de processamento automático de dados, os resultados das pesquisas epidemiológicas vêm adquirindo uma importância crescente no

e) Observar quais variáveis são mais determinantes na presença de sangramento gengival, de cálculo dentário ou de bolsas periodontais, utilizando técnicas de mineração de dados, em particular a de regras de classificação.

A partir destes objetivos formam-se as seguintes hipóteses:

- a) Encontrar uma alta prevalência da doença periodontal na população estudada;
- b) Descobrir associação estatisticamente significativa entre o baixo nível socioeconômico e a doença periodontal;
- c) Descobrir associação estatisticamente significativa entre o tabagismo e a doença periodontal;
- d) Formular regras de classificação com base nas variáveis de significância estatística.

1.2 Limitações do Estudo

A elaboração do estudo apresentou as seguintes limitações:

- a) Como o banco de dados se refere aos alistados do Exército, todas as análises realizadas foram relacionadas apenas à indivíduos do sexo masculino. Todavia, como será mencionado posteriormente, referências bibliográficas sobre a doença periodontal, indicam que não existem diferenças significativas com relação ao sexo;
- b) O banco de dados é composto de uma amostra dos alistados pois, devido ao custo elevado de pessoal e de material, a realização de exames clínicos era inviável em todos os indivíduos;
- c) Há a dificuldade de comparação dos resultados encontrados, pois há uma carência muito grande de estudos de epidemiologia bucal em adultos jovens;

d) Os sistemas computacionais para uso em banco de dados têm custos de aquisição elevados. Desta forma, optou-se pelo uso da ferramenta CBA Data Mining II da National University of Singapore, por ser disponível na rede de computadores.

1.3 Estrutura do Trabalho

Com a finalidade de facilitar a compreensão do estudo este trabalho foi estruturado em outros 5 capítulos além deste.

O capítulo 2, O Periodonto e sua Epidemiologia, traz uma revisão bibliográfica da doença periodontal e sua epidemiologia, mostrando as definições do objeto do estudo.

No capítulo 3 relata-se o que a literatura apresenta sobre análise de banco de dados, enfatizando-se a análise estatística e a descoberta de conhecimento em banco de dados com as técnicas mais utilizadas. Neste capítulo é também apresentada a ferramenta de mineração de dados – CBA Data Mining II.

O capítulo 4 mostra a metodologia empregada, bem como apresenta a população de estudo, a formulação da amostra, as variáveis pertencentes ao banco de dados e o plano de análise dos dados.

No capítulo 5 estão demonstrados os resultados encontrados no estudo através das análises estatísticas e de mineração de dados propostas nos objetivos do trabalho.

As conclusões estão relatadas no capítulo 6, onde aparecem as afirmações pessoais do autor a respeito dos resultados encontrados no estudo, e que têm importância como contribuição científica ao conhecimento da doença periodontal, através do estudo de um banco de dados.

2. O PERIODONTO E SUA EPIDEMIOLOGIA

O periodonto (peri = em torno de, odontos = dente) compreende a gengiva, o ligamento periodontal, o cemento radicular e o osso alveolar. O periodonto é subdividido conforme sua morfologia em periodonto de proteção e periodonto de inserção. O de proteção é representado pela gengiva e o de inserção pelas demais estruturas [CAR 97]. A Figura 1 mostra as estruturas que compõem o periodonto [LIN 88].

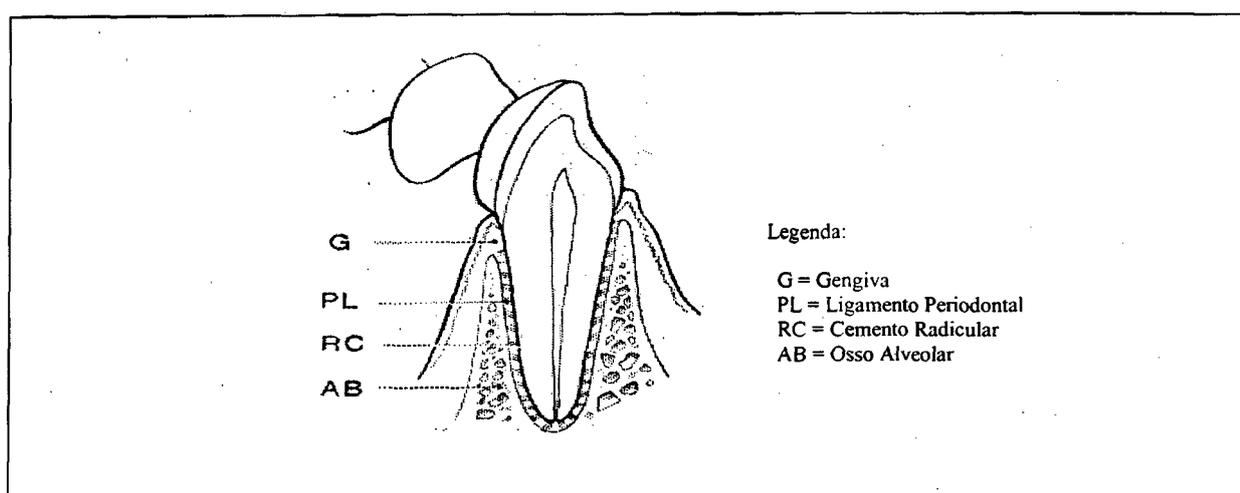


Figura 1: As estruturas do periodonto.

A gengiva é a parte da mucosa bucal que recobre os processos alveolares dos maxilares e circunda o colo dos dentes [EGE 95]. O ligamento periodontal é a estrutura de tecido conjuntivo que circunda a raiz do dente, onde se localiza o cemento e a liga ao osso alveolar. É contínuo ao tecido conjuntivo gengival e se comunica com os espaços medulares pelos canais vasculares no osso [CAR 97].

O periodonto, também chamado “aparelho de inserção” ou de “tecidos de sustentação dos dentes” sofre certas alterações com a idade e, por isso, está sujeito a alterações morfológicas e funcionais [CAR 97].

2.1 A Doença Periodontal

A doença periodontal e a cárie dental, representam as principais ameaças a saúde bucal. Enquanto a cárie dental estraga os dentes, as estruturas de suporte e de proteção do dente são afetadas pela doença periodontal [AMEa96] [AMEb96] [CAR 97] [AXE 98]. Afirma-se também que, a doença periodontal é a doença crônica mais prevalente que afeta a dentição humana [GUG 98].

Ao longo dos anos o termo doença periodontal tem tido diferentes significados e tem sido usado de forma ambígua. Ele é usado em um sentido geral, para abranger todas as doenças do periodonto, seja ele de suporte ou de proteção. O termo doença periodontal como sinônimo de periodontopatia vem caindo em desuso [CAR 97].

As doenças periodontais são tradicionalmente divididas em duas categorias principais: as doenças gengivais (gengivites) e as doenças que envolvem as estruturas de suporte do dente, como as periodontites [LIN 88].

O agente etiológico das doenças periodontais é a placa bacteriana. A placa bacteriana é formada por um conjunto de bactérias, células epiteliais descamadas, leucócitos polimorfonucleares, micoplasmas, leveduras e protozoários, que respondem pela doença periodontal [CAR 97].

O cálculo, que é a placa bacteriana mineralizada, atua apenas de forma indireta na patogenia da doença periodontal, uma vez que ele por si só não consegue induzir a doença. Sua remoção é indicada porque ele ajuda a reter mais placa e sua presença dificulta a higiene bucal [EGE 95] [CAR 97].

Inicialmente formulou-se a teoria da placa não específica, onde se acreditava que a doença periodontal resultava da elaboração de produtos nocivos por toda a flora da placa. Então se conclui que quanto maior o acúmulo de placa, maior era a possibilidade de superação

de nossas defesas imunológicas, e por conseguinte maior a chance de adquirir a doença periodontal [CAR 97].

Atualmente é mais aceita a teoria da placa específica, onde apenas certa placa é patogênica, e que sua patogenicidade depende da presença ou aumento do número de microorganismos específicos que produzem substâncias capazes de promover a doença periodontal [CAR 97].

O acúmulo de placa é o responsável pelo início da doença periodontal, denominada gengivite e caracterizada pelo sangramento gengival, mostrando então, uma relação causal entre a higiene bucal e a gengivite [LIN 88] [LAS 89] [REG 91] [EGE 95] [GEN 96] [CAR 97].

O quadro de gengivite pode evoluir para uma periodontite, onde se passa a observar a perda de inserção entre o periodonto e o dente, caracterizando a presença da bolsa periodontal [LIN 88] [LAS 89].

Não existem evidências de que a gengivite não tratada sempre progrida para uma periodontite. Porém, quando se desenvolve periodontite ela é freqüentemente precedida pela gengivite. A progressão das doenças periodontais apresenta-se de forma muito variada, podendo ter períodos de atividade de várias intensidades seguidos por intervalos de estagnação ou remissão, sendo esta evolução influenciada pelas nossas células de defesa e pelos tipos microbianos agressores envolvidos [REG 91] [GEN 96] [CAR 97].

Recentemente vários estudos epidemiológicos parecem mostrar que há um incremento nas perdas dentárias por problemas periodontais, tendência esta não observada com relação à cárie dentária [ONG 98]. Isto ocorre principalmente quando se analisa adultos em comunidades mais desenvolvidas [KRI 99].

2.1.1 A Epidemiologia das Doenças Periodontais

Em 1977, a Organização Mundial da Saúde (OMS), com o apoio da Federação Dentária Internacional, iniciou estudos sistemáticos acerca da epidemiologia das doenças periodontais e das necessidades de tratamento periodontal. Em 1983, foi adotado um índice denominado CPITN (*Community Periodontal Index of Treatment Needs*) [LAS 89].

O índice CPI (*Community Periodontal Index*), da Organização Mundial da Saúde (1997), derivado do antigo índice CPITN, serve para estabelecer a condição periodontal, baseado em 3 indicadores: no sangramento gengival, no cálculo e na bolsa periodontal. A boca é dividida em 6 sextantes e são analisados 6 dentes índices (16, 11, 26, 36, 31 e 46), onde são registradas as piores condições para cada dente examinado. Usa-se uma sonda especialmente desenhada para este fim, que tem uma esfera de 0,5mm na ponta e uma banda negra entre 3,5 e 5,5mm em sua parte ativa, além de anéis marcadores em 8,5mm e 11,5mm [OMS 97]. Os códigos e critérios do CPI podem ser vistos na Figura 2.

O código 0 do índice CPI se refere ao periodonto sadio, ou seja, aquele que não apresenta sangramento gengival à sondagem. O código 1 é usado quando se tem sangramento e o 2 quando se nota a presença de cálculo dentário. Os códigos 3 e 4 são referentes às bolsas periodontais rasas (até 5 mm) e profundas (com mais de 6 mm) respectivamente [OMS 97].

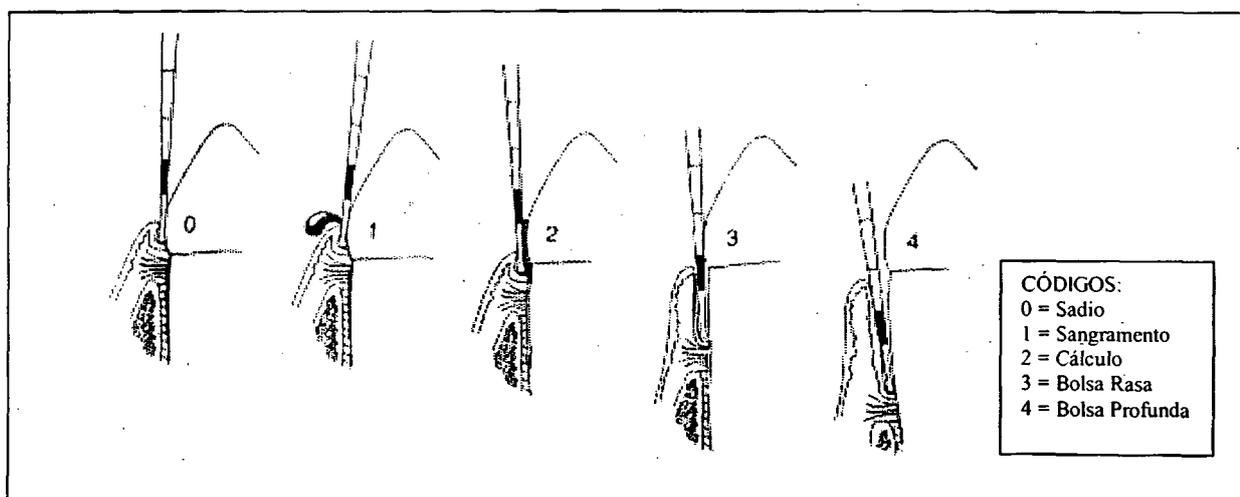


Figura 2: Códigos do índice CPI da Organização Mundial da Saúde [OMS 97].

planejamento e na administração da assistência em saúde. É neste contexto que se insere este trabalho, onde procura-se dar uma contribuição ao conhecimento das questões relacionadas ao periodonto, e por conseguinte à saúde bucal da coletividade.

1.1 Objetivos e Hipóteses

Este estudo tem como objetivo geral conhecer, a partir da análise de um banco de dados, a prevalência e severidade da doença periodontal nos alistados do Exército Brasileiro residentes em Florianópolis no ano de 1999.

Os objetivos específicos são:

- a) Conhecer o índice CPI (*Community Periodontal Index* – Índice Comunitário Periodontal) da Organização Mundial de Saúde (OMS) e seus componentes na população estudada. Este índice serve como um referencial para demonstrar a severidade da doença periodontal, pois segmenta as fases evolutivas da doença em níveis de menor ou maior complexidade;
- b) Conhecer a prevalência de sangramento gengival e de cálculo dentário na população estudada, mostrando desta forma a distribuição dos principais fatores biológicos agressores ao periodonto e favoráveis à instalação da doença periodontal;
- c) Verificar a existência de associação entre doença periodontal e o nível socioeconômico do indivíduo, através do nível de escolaridade dos pais, do alistando e da renda familiar. Assim, pode-se visualizar a influência de fatores socioeconômicos como determinantes na ocorrência da doença periodontal;
- d) Verificar a existência de associação entre doença periodontal e o tabagismo, permitindo visualizar a influência do hábito de fumar cigarro sobre a ocorrência da doença periodontal;

Alguns estudos têm sido feitos com adultos jovens aplicando-se os códigos e critérios da OMS. Dentre eles pode-se destacar:

Estudos epidemiológicos feitos entre 1982 e 1992 sobre a doença periodontal em Bangladesh na faixa etária entre 12 e 19 anos, mostraram a prevalência de cálculo entre 82 e 95% da população sendo de apenas 0 a 9% livres de sangramento gengival e cálculo. Em Bangladesh, nesta faixa etária, se encontra uma das mais altas prevalências de cálculo do mundo [HEL 96].

Estudos epidemiológicos sobre condições periodontais foram realizados nos anos de 1976, 1981, 1986 e 1991 em recrutas, com idade média de 20 anos (Figura 3), das Forças de Defesas Finlandesas. As prevalências de cálculo supragengival foram respectivamente (32,2%; 28,3%; 42% e 31,7%), de cálculo subgengival (35,9%; 22,4%; 15,4% e 24,8%), de sangramento gengival (73,4%; 69,1%; 68,5% e 74,9%) além de bolsas periodontais de 4mm ou mais (11%; 6,9%, 3,9% e 1,6%) [ANK 97].

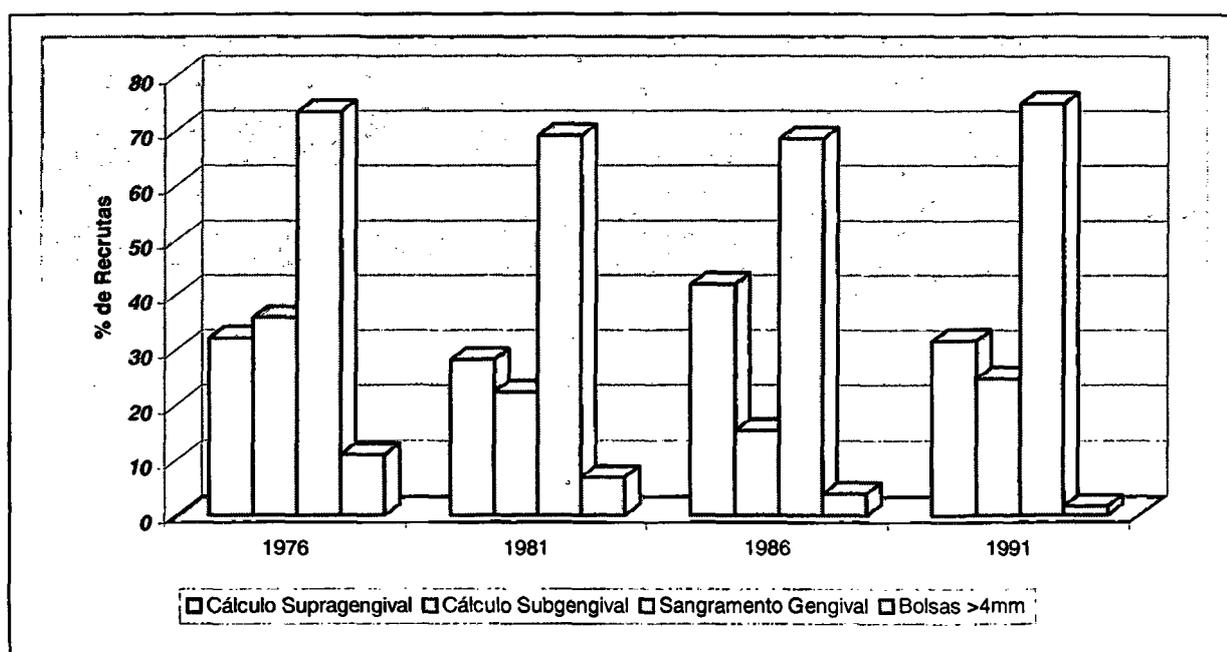


Figura 3: Distribuição de frequência percentual de cálculo dentário, sangramento gengival e bolsas periodontais em recrutas de 20 anos de idade na Finlândia.

Em jovens americanos de escolas públicas e privadas, durante os anos de 1986 e 1987, na faixa etária de 13 a 20 anos, a presença de inflamação gengival e do cálculo subgengival, mostraram-se associadas com a periodontite precoce [ALB 96].

Em Ukonga na Tanzânia, em 1988, na faixa etária de 15 a 19 anos encontrou-se, através do índice CPITN, a prevalência da doença periodontal de 97%, sendo o cálculo prevalente em 84,8% e as bolsas rasas atingindo 7,6% destes jovens [MUN 96].

Em 1986 foi realizado o Levantamento Epidemiológico em Saúde Bucal no Brasil, Zona Urbana (Figura 4). Na faixa etária de 15-19 anos foram examinados 4780 jovens, onde encontram-se os seguintes valores dos componentes do índice CPITN: sadios (28,8%); com doença periodontal (69,5%); com cálculo ou bolsas rasas (49,6%) e com bolsas profundas (0,7%). O número médio de sextantes com doença periodontal é de 2,72, com cálculo ou bolsas rasas é de 1,51 e com bolsas profundas é de 0,01 [MIN 88].

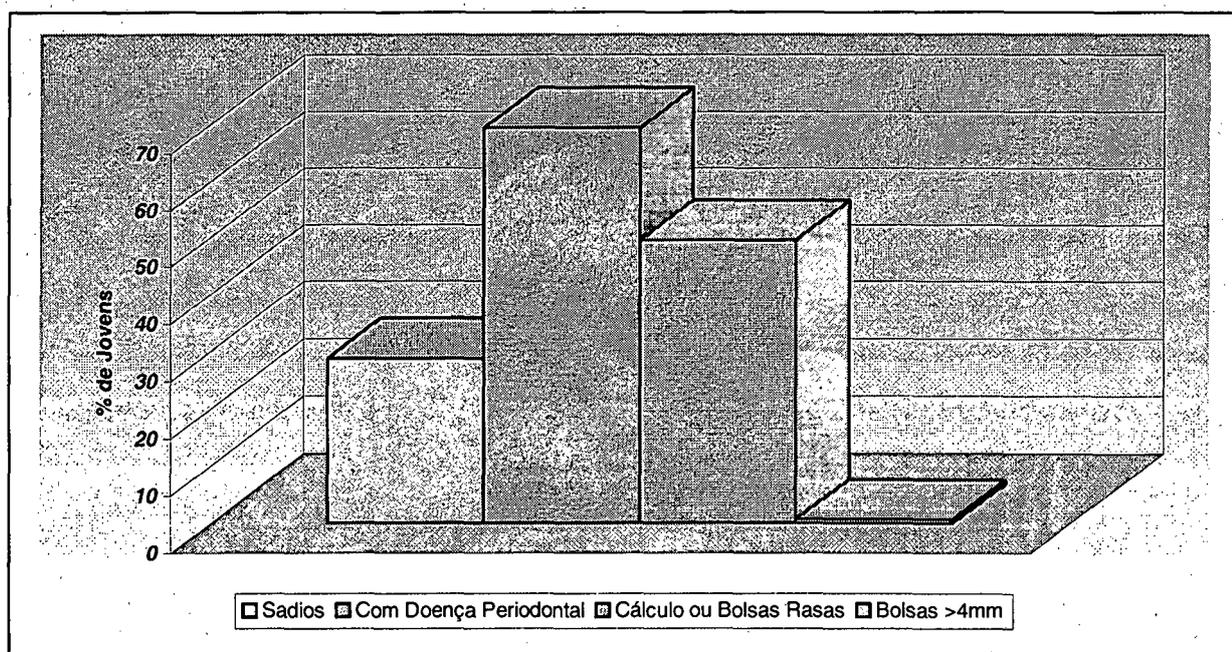


Figura 4: Distribuição de frequência percentual das condições de saúde periodontal em jovens brasileiros de 15 a 19 anos.

Os dados da Região Sul, nesta mesma faixa etária (Figura 5), mostram os seguintes valores: sadios (13,0%); com doença periodontal (86,3%); com cálculo ou bolsas rasas (60,5%) e com bolsas profundas (0,1%). O número médio de sextantes com doença periodontal é de 4,13, com cálculo ou bolsas rasas é de 1,95 e com bolsas profundas é de 0,001 [MIN 88].

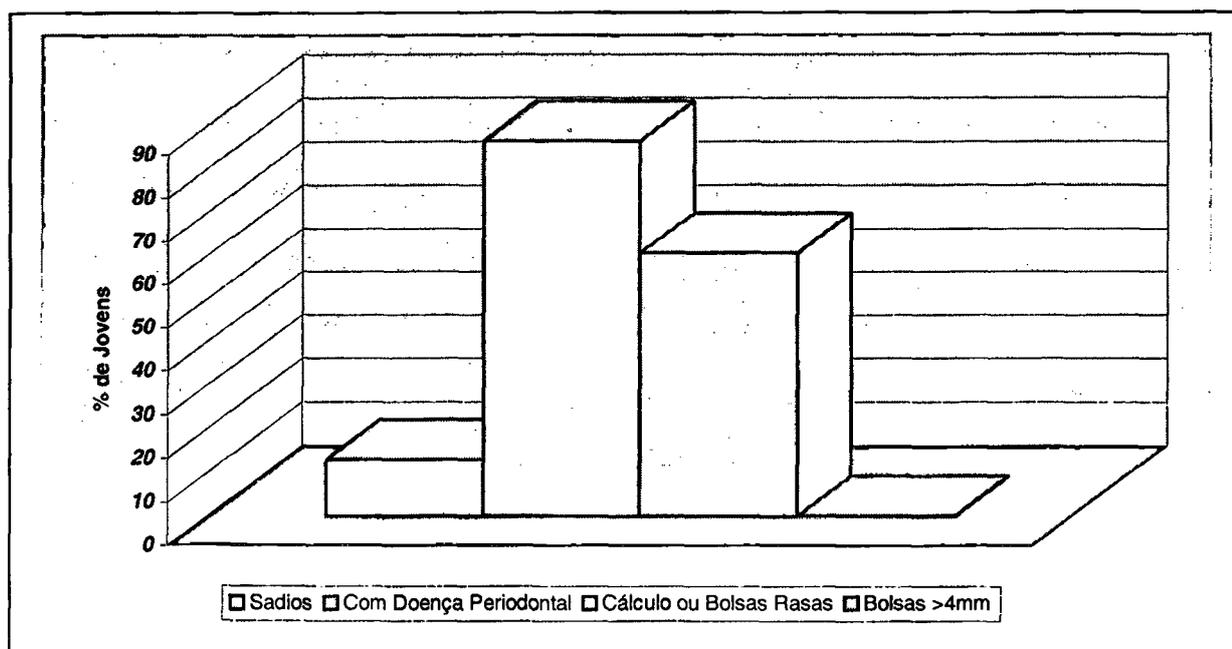


Figura 5: Distribuição de frequência percentual das condições de saúde periodontal em jovens da Região Sul do Brasil de 15 a 19 anos.

A Tabela 1 mostra um quadro global da condição periodontal, em jovens, em diversos países do mundo. É interessante observar a baixa ocorrência de bolsas periodontais e a alta prevalência de gengivite na maioria dos países.

Condições Periodontais observadas e mensuradas através do CPITN nas idades de 15-19 anos.
Adaptado do banco de dados da Organização Mundial da Saúde, 1994.

País	Ano	Idade	Amostra	% de pessoas com o maior escore CPITN					Número médio de sextantes				
				0 Sem doença	1 Sangramento	2 Cálculo	3 Bolsa 4-5mm	4 Bolsa +6mm	0 Sem doença	1 ou 2 ou 3 ou 4 Com Doença	2 ou 3 ou 4 Cálculo ou Bolsa	3 ou 4 Bolsa +4	
Suíça	1990	17-19	263	51	27	19	2	1	4,7	1,3	0,3	0,1	
	1991/92	15-19	189	7	11	11	62	9	1,2	4,8	3,0	1,9	
Suécia	1988	15	60	0	2	58	38	2	0,9	5,1	3,8	0,9	
	1990	18	95	5	5	66	22	1	1,7	5,1	3,8	0,9	
Suíça (de Janeiro)	1988	15-19	350	4	12	29	53	3	1,8	4,2	2,6	1,1	
	1986	15-19	4780	30	20	42	7	1	3,1	2,7	1,5	0,1	
Suíça (Urbana)	1989	15-19	187	17	36	42	5	0	2,5	3,5	1,1	0,1	
	1991	15	120	10	60	27	3	0	3,4	2,6	0,5	0,0	
Suíça (Rurais)	1992	15-19	155	3	1	90	6	0	2,0	4,0	3,9	0,1	
	1986	15-19	358	17	13	33	32	5	2,6	3,4	2,4	0,9	
Suíça (Rurais)	1983	15	99	0	36	54	9	0	1,7	4,3	1,2	0,2	
	1990	15-19	70	34	40	26	0	0	4,1	1,9	0,3	0,0	
Suíça (Rurais)	1992/93	15-19	956	45	3	51	1	0	4,4	1,6	1,4	0,0	
	1991	15	202	9	16	72	2	1	1,6	4,4	2,7	0,0	
Suíça (Rurais)	1990	15-19	107	6	11	79	5	0	1,2	4,8	3,2	0,1	
	1984	15	393	0	9	20	34	37	1,0	5,0	3,4	1,7	
Suíça (Rurais)	1988	15	100	11	4	76	9	0	2,7	3,1	2,5	0,2	
	1982	15	70	1	3	85	11	0	1,3	4,7	4,0	0,2	
Suíça (Rurais)	1989	15-19	63	3	2	52	41	2	1,8	4,2	3,1	0,9	
	1990	15-19	1154	1	3	46	42	8	1,1	4,9	4,4	1,1	
Suíça (Rurais)	1991	15-19	561	26	20	52	2	0	3,6	2,4	1,5	0,0	
	1989	15	705	21	16	63	0	0	3,8	2,2	1,3	0,0	
Suíça (Rurais)	1984	15	1898	1	52	40	6	2	1,1	4,9	1,2	0,2	
	1990/91	15-19	96	12	36	49	3	0	2,9	3,1	0,6	0,0	
Suíça (Rurais)	1991	15	123	1	10	80	9	1	1,1	4,9	3,1	0,1	
	1989	15	135	0	3	44	44	9	0,2	5,8	4,7	1,2	
Suíça (Rurais)	1987	15-19	62	1	8	26	55	10	1,6	4,3	2,9	1,3	
	1993	15-19	279	0	1	38	58	3	0,2	5,8	4,7	1,7	

WHO Oral Health Country Profile Programme, Division of Noncommunicable Diseases / Oral Health, School of Dentistry, Niigata University, Japan.
www.whocollab.odont.lu.se

Em 1994, estudou-se a condição periodontal de trabalhadores de 18 a 64 anos de idade em Araraquara, no estado de São Paulo. Uma amostra de 528 pessoas foi examinada, mostrando que o cálculo era mais freqüentemente encontrado na faixa etária de 18 a 34 anos e que somente após os 35 anos passava-se a observar mais bolsas rasas [DIN 94].

2.1.2 O Nível Socioeconômico como Fator de Influência na Doença Periodontal

Um dos fatores que influem na prevalência das doenças periodontais é o nível socioeconômico, pois muitos levantamentos mostram que melhores condições periodontais ocorrem à medida que aumentam os anos de educação formal e a renda da população [AHL 96]. Alguns estudos que evidenciam esta afirmação são relatados a seguir.

No levantamento epidemiológico feito em 1986 no Brasil, quando se analisou a distribuição da doença periodontal em função do estrato de renda familiar, tanto a nível nacional, quanto a nível da Região Sul, observa-se claramente que as piores condições periodontais são encontradas nos níveis mais baixos de renda [MIN 88].

Em 1990 são feitas explicações teóricas a respeito das desigualdades sociais em saúde oral na Dinamarca baseadas em dados de estudos longitudinais. Afirma-se que em adultos as desigualdades em saúde oral são marcantes quando compara-se indivíduos de baixo e alto nível social [PET 90].

Relata-se em 1994, na África do Sul, ao se analisar 3763 adultos entre 20 e 64 anos de idade, durante a Pesquisa Nacional de Saúde Bucal (NOHS), que existem diferenças no quadro de saúde bucal de pessoas com distintos padrões educacionais e níveis de renda [GIL 94]. Conclusões semelhantes são obtidas em um estudo feito em um grupo de mulheres grávidas em uma clínica de obstetrícia na Nigéria, quando se relacionou a associação entre a prevalência da doença periodontal e o baixo nível socioeconômico ($p < 0,05$)¹ [OGU 96].

1) p = probabilidade de significância, que será discutida posteriormente.

Estudou-se, na Tanzânia, 1764 pessoas (827 homens e 937 mulheres) entre 3 e 84 anos de idade. Encontrou-se gengivite em 81% das pessoas na faixa etária de 10 a 18 anos de idade, e bolsas periodontais com mais de 6 mm em apenas 4% dos indivíduos. Concluiu-se, que existe associação estatística e significativa entre o nível educacional e a prevalência da gengivite ($p < 0,01$) e com bolsa periodontal ($p < 0,001$) [MUN 95].

Em 1996 quando se analisou a influência de fatores socioeconômicos entre 118 homens diabéticos e 115 não diabéticos, no Veterans Affairs Medical Center no Kentucky (EUA), demonstrou-se que à medida que temos menores níveis de renda e padrões educacionais mais baixos, encontramos condições periodontais piores para os dois grupos ($p < 0,05$) [BRI 96].

Ao se estudar em 1996 a condição periodontal de 2400 jovens chilenos encontrou-se os códigos 3 (bolsas rasas) e 4 (bolsas profundas) do CPITN em 10,49% dos examinados. Ocorreu uma significativa diferença, quanto ao sexo, na prevalência de bolsas periodontais nos homens (9%) comparado as mulheres (12,16%). Também observou-se diferenças quanto ao nível socioeconômico (determinado em função da renda familiar), sendo a prevalência das bolsas periodontais maior nos indivíduos de baixo nível socioeconômico (15,87%), do que nos de médio (9,22%) e alto nível socioeconômico (8,93%) [LOP 96].

A prevalência da doença periodontal vem mostrando uma correlação inversa com o nível socioeconômico. E o emprego de princípios de qualidade na metodologia dos estudos técnicos é uma condição essencial para que se possa entender melhor as diferenças de morbidade entre os diferentes níveis sociais [MIC 96].

Demonstrou-se que, em idosos, fatores sociais e culturais influem na saúde oral da população no sul da Austrália, Ontário e na Carolina do Norte. Pois este estudo mostrou que os indivíduos com mais dentes perdidos, fragmentos radiculares retidos, cáries de raiz e bolsas periodontais, pertenciam aos níveis sociais mais baixos [SLA 96]. Estes achados ratificam as

afirmações que citam como as variáveis de maior influência no índice CPITN pela ordem como sendo: a idade, a classe socioeconômica, a escolaridade e o sexo [CAM 92].

Em 1997, estudou-se em Helsinki, Finlândia, 172 adolescentes, mostrando como um dos resultados, que o fator socioeconômico tem significância estatística com o sangramento gengival [KAL 97].

Em 1998, em um estudo com 3763 adultos com idade variando entre 20 e 64 anos, na África do Sul, pode-se mostrar que existe influência dos fatores socioeconômicos na distribuição da doença periodontal, através da análise retrospectiva dos dados da Pesquisa Nacional de Saúde Oral (NOHS, 1988/89). Este estudo relata ainda que se têm piores condições periodontais quando se têm menores níveis de educação e de renda [GUG 98].

Em países desenvolvidos, as populações com nível socioeconômico mais baixo não conseguem ter uma dieta adequada e têm uma higiene oral mais deficiente, contribuindo desta forma para uma pior condição periodontal. Em países subdesenvolvidos, a maior prevalência de problemas periodontais talvez possa ser explicada pelos baixos níveis de educação, associados com baixa renda e nutrição deficiente, atuando em conjunto [LAS 89].

Tomando-se como referência os estudos citados, conclui-se que a doença periodontal pode ter seu curso evolutivo agravado por questões relacionadas com o baixo nível socioeconômico. Mostrando assim uma possível associação entre a prevalência e a severidade da doença periodontal com o baixo nível socioeconômico.

2.1.3 O Fumo como um Fator de Influência na Doença Periodontal

Assim como os fatores de natureza socioeconômica, outros fatores ambientais, como o hábito de fumar, podem ter significativa relação com as doenças periodontais.

Diversos estudos demonstram que o tabagismo também pode influenciar a saúde periodontal. Assim suspeita-se que o fumante tenha um risco maior em desenvolver doença periodontal [PRE 86] [HAFa92] [HAFb92] [HAFc92] [BRI 96] [AMEb96] [TAA 96] [DNS 97] [PAG 97] [AXE 98] [MON 98] [REN 98].

Entende-se por fumante, como sendo aquele indivíduo que tem disposição duradoura e repetição freqüente do ato de fumar cigarro. Já em 1947, suspeitava-se que os fumantes tinham maior acúmulo de cálculo do que os não fumantes [PID 47].

Em 1986, analisou-se a redução das profundidades de bolsas periodontais, após tratamento não cirúrgico, em pacientes fumantes e não fumantes. A redução das bolsas mostrou-se similar em áreas posteriores, porém em áreas anteriores esta redução pareceu ser menor nos fumantes [PRE 86].

Alguns anos depois realizou-se um estudo com jovens fumantes, onde encontrou-se um significativo incremento na prevalência da recessão gengival localizada e da perda de inserção [ROB 90].

Em 1994 e em 1995 demonstrou-se que a existência de relação entre o fumo e a condição periodontal é verdadeira, pois mostra-se associação entre a prevalência de doença periodontal moderada e severa com o número de cigarros fumados por dia e o número de anos em que o paciente tem o hábito de fumar [GRO 94] [GRO 95].

Relata-se que o tabagismo incrementa o risco de infecção subgengival com patógenos periodontais, destacando-se entre eles: *Actinobacillus actinomycetemcomitans*, *Bacteroides forsythus* e *Porphyromonas gingivalis* [AXE 98] [DNS 97].

Segundo a Academia Americana de Periodontologia (1996), as substâncias tóxicas do cigarro podem trazer efeitos maléficos ao periodonto, seja afetando diretamente as células do periodonto (fibroblastos) ou alterando a resposta imunológica. A nicotina, por exemplo, é uma das mais de 2000 substâncias potencialmente tóxicas do cigarro, onde observa-

se que em baixas concentrações pode estimular a quimiotaxia dos neutrófilos, mas em altas concentrações pode prejudicar a fagocitose. Há também relatos, que afirmam que, nos fumantes ocorre uma diminuição dos níveis de anticorpos salivares (IgA) e séricos (IgG) para *P. intermedia* e *F. nucleatum*, além de possuírem uma redução no número de linfócitos [AMEb96].

Há relatos de que o fumo pode ser um dos mais significantes fatores de risco no desenvolvimento da doença periodontal [PAG 97]. Tem sido documentado que os componentes do fumo podem induzir ou exacerbar várias formas de doenças periodontais por dano local direto aos tecidos periodontais e/ou pela alteração de nossa resposta imunológica, que prejudicaria a neutralização da infecção e facilitaria a destruição dos tecidos do periodonto [AMEb96].

Em 1997, comparou-se 997 indivíduos jordanianos (45% fumantes e 55% não fumantes) entre 20 e 60 anos de idade. Foi encontrado um baixo nível de higiene oral tanto em fumantes como em não fumantes, porém com escores de placa e cálculo maiores em fumantes, sem no entanto encontrar diferenças significantes entre os dois grupos no que se refere a sangramento gengival e profundidade de bolsas periodontais [TAA 96].

No Charles Clifford Dental Hospital (Inglaterra), em 1997, comparou-se um grupo de 20 fumantes com outro de 20 não fumantes com similar nível de doença periodontal (Periodontite de Adulto Moderada para Severa) e verificou-se que há uma significante variação de temperatura em sítios subgengivais entre fumantes e não fumantes. Os fumantes apresentavam os sítios subgengivais 0,4 graus Celsius mais quentes, tanto para os sítios saudáveis como para os sítios doentes [DNS 97]. A importância do estudo é entendida uma vez que a elevação da temperatura do sítio subgengival acima de 35,5 graus Celsius pode ser um indicativo de futura perda de inserção, e que também induz a maiores proporções de patógenos

periodontais como: *Prevotella intermedia*, *Peptostreptococcus micros*, *Prophiromonas gingivalis* e *Actinobacillus actinomycescomitans* [HAFa92] [HAFB92] [HAFc92].

Em 1998 mostrou-se que os fumantes parecem não responder tão bem quanto os não fumantes nas formas de terapia periodontal não cirúrgica [REN 98].

Em um estudo feito em Varmland (Suécia) em 1998, comparando as proporções das necessidades de tratamento, medidas pelo índice CPITN (Índice Comunitário Periodontal e Necessidades de Tratamento da Organização Mundial da Saúde [OMS 97]), em fumantes e em não fumantes de 35, 50, 65 e 75 anos de idade, encontrou-se que as necessidades de tratamento eram maiores nos fumantes em todos os grupos etários [AXE 98].

Através destes estudos, afirma-se que a doença periodontal pode ter seu curso evolutivo agravado por questões relacionadas com o hábito de fumar. Mostrando assim uma possível associação entre a prevalência e a severidade da doença periodontal com o tabagismo.

Como a severidade da doença periodontal está relacionada à duração e a quantidade de cigarros fumados, observa-se que os dentistas devem adotar um novo papel, o de instruir a população e engajar-se em campanhas anti-tabagismo, exercendo assim a função de um verdadeiro promotor de saúde. Uma vez que o fumo está também associado com a incidência e a prevalência de uma variedade de doenças sistêmicas, como o câncer, o baixo peso ao nascer, as doenças pulmonares, as gastrointestinais e as cardiovasculares.

3. A ANÁLISE DE BANCO DE DADOS

As primeiras tentativas de se extrair conhecimento útil a partir de uma base de dados surgiram em meados dos anos 60, através da análise exploratória de dados – atividade desenvolvida no campo da Estatística. Atualmente, técnicas de Inteligência Artificial têm sido incorporadas com muito sucesso na análise de banco de dados. [DAT 00].

3.1 A Análise Estatística de Dados pelo Teste de Associação Qui-Quadrado

O Teste de Associação Qui-Quadrado é muito utilizado em pesquisas sociais e de saúde para verificar possíveis associações entre variáveis categóricas. Assim sendo, é possível determinar se uma variável qualitativa influi no comportamento de outra [SOA 91].

Para se executar o teste Qui-quadrado, inicialmente se elabora uma tabela de contingência 2x2 (Figura 6), onde são colocados os valores das variáveis em análise que serão utilizados no cálculo desta estatística. A expressão 3.1 utilizada para o cálculo do teste, é composta pelo tamanho da amostra (n) e pelas variáveis (a,b,c,d) oriundas da tabela de contingência [BAR 99].

	Sexo Masculino	Sexo Feminino
Fumantes	92 (a)	38 (b)
Não Fumantes	108 (c)	62 (d)

Figura 6: Exemplo de tabela de contingência 2x2.

$$X^2 = \frac{n \cdot (a \cdot d - b \cdot c)^2}{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)} \quad (3.1)$$

Numa tabela 2x2 a distribuição dos possíveis valores da estatística qui-quadrado, quando calculada com a correção de continuidade de Yates, aproxima-se mais da distribuição qui-quadrado com 1 grau de liberdade do que quando calculada sem esta correção. A estatística do teste com a correção de continuidade de Yates passa a ser calculada pela expressão 3.2 [BAR 99].

$$X^2 = \frac{n \cdot \left(|a \cdot d - b \cdot c| - \frac{n}{2} \right)^2}{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)} \quad (3.2)$$

O valor de χ^2 fornece, através da tabela da distribuição Qui-quadrado o valor da probabilidade de significância (p) [SOA 91] [BAR 99].

A significância é uma medida estatística que é estabelecida pelo analista, de forma arbitrária, mas que seja suficientemente precisa para concluir à respeito de hipóteses pré estabelecidas sobre inferências populacionais a partir de amostras desta população. Comumente, para os estudos em saúde, adota-se um nível de significância de 5% (0,05) e calcula-se, através deste teste, o valor da probabilidade de significância (p) [DAW 93].

Quando se tem a probabilidade de significância maior do que o nível de significância ($p > 0,05$), rejeita-se a hipótese de existir associação entre as variáveis estudadas. Porém quando o valor de p é menor que o nível de significância ($p < 0,05$) aceita-se a hipótese de existir associação estatística entre as variáveis. Quanto menor for o valor de p, mais significativo é a associação estatística [DAW 93] [SOA 91] [BAR 99].

Em algumas situações, o Teste do Qui-quadrado não deve ser empregado, sendo substituído pelo Teste Exato de Fisher. Adota-se este procedimento quando se têm, na tabela de contingência, alguma das frequências esperadas (total da linha multiplicado pelo total da coluna dividido pelo total geral) menor que 5 [DAW 93].

A equação 3.3 fornece o valor da probabilidade de significância através do Teste Exato de Fisher [DAW 93].

$$p = \frac{(a+b)! \cdot (c+d)! \cdot (a+c)! \cdot (b+d)!}{a! \cdot b! \cdot c! \cdot d! \cdot n!} \quad (3.3)$$

3.2 A Análise de Dados através de Modelos de Inteligência Artificial

Atualmente, têm sido desenvolvidas técnicas que compõem o campo do KDD (*Knowledge Discovery in Databases* – Descoberta de Conhecimento em Banco de Dados), em resposta à necessidade de se tratar volumes crescentes de dados. A disponibilidade de ferramentas mais poderosas proporcionou as condições para o desenvolvimento de uma enorme quantidade de sistemas de bancos de dados. Estes sistemas, criados para atender a necessidades específicas, provocaram um acúmulo de dados sem precedentes, levando à suspeita de que os mesmos poderiam conter informações muito além das que eles normalmente provêm [PRA 97].

A evolução no campo do KDD, nos últimos 30 anos, propiciada pelo avanço da tecnologia, permitiu a extração em bases de dados, de informações cada vez mais refinadas ao longo deste período. Esta evolução possibilita atualmente a extração de informações tão refinadas, ao ponto de auxiliarem no processo de tomada de decisões [DWB 01].

A Figura 7 ilustra esta evolução, onde podemos ver que até os anos 70 as informações eram oriundas basicamente da estatística. Na década de 80 surgiram as idéias de

Data Warehouse e as ferramentas OLAP (*On Line Analytical Process*), e somente na década de 90 surgem as técnicas de mineração de dados [DAT 00].

Nas últimas duas décadas se tem presenciado um crescente volume de informações sendo armazenadas por meios eletrônicos. A análise dos bancos de dados gerados podem trazer informações úteis e previamente desconhecidas [DWB 01].

Atualmente vem crescendo o número de publicações, principalmente em revistas de informática e negócios, acerca de métodos de análises de banco de dados eletrônicos, destacando-se o *Data Mining* [QUB 01].

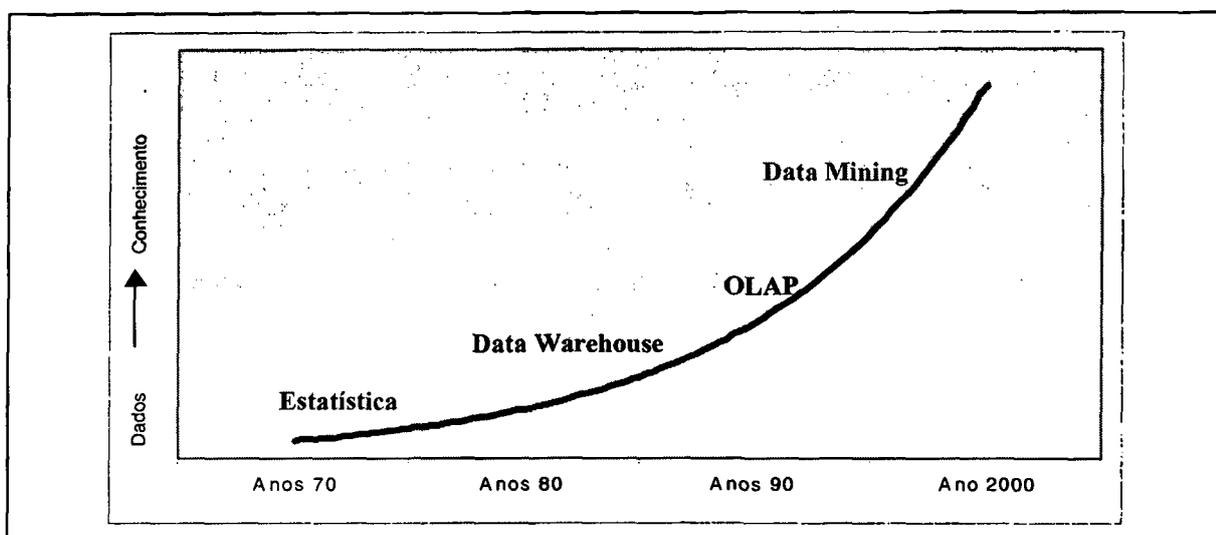


Figura 7: Evolução da extração do conhecimento em base de dados [DAT 00].

Os bancos de dados começaram a se formar para atender exigências mercadológicas de empresas comerciais [BER 97]. É possível aplicar-se as mesmas técnicas de análise para que sejam obtidas informações importantes no planejamento das políticas de saúde e no auxílio do estudo da epidemiologia das doenças.

3.2.1 O Processo KDD – Knowledge Discovery in Databases

A descoberta de conhecimento em base de dados (KDD) é uma tecnologia que utiliza técnicas de mineração de dados para extrair conhecimento a partir de uma base de dados. Estas técnicas, desenvolvidas principalmente na área da Inteligência Artificial, são diversas e sua aplicabilidade está diretamente vinculada às características da base de dados e aos resultados que se espera obter [BER 97].

Está se vivendo a era da informação. A evolução da tecnologia, propiciada pela informática, permitiu no decorrer das últimas décadas a formação de grandes bases de dados nas mais diversas áreas da sociedade [DAT 00].

Já há algum tempo, constatou-se que as informações armazenadas nestas bases de dados poderiam conter um conhecimento implícito agregado a elas. Impulsionada pela idéia de descobrir conhecimento sobre estes dados e poder utilizá-los em processos de tomada de decisões, surgiu uma nova área de pesquisa conhecida como Descoberta de Conhecimento em Base de Dados [QUB 01].

A tecnologia de Descoberta de Conhecimento em Base de Dados desponta como uma solução para um melhor aproveitamento das informações contidas nas bases de dados, otimizando os processos de tomada de decisões nos mais diversos campos do conhecimento humano [DAT 00].

A descoberta de conhecimento em base de dados é uma área de pesquisa que tem crescido e atraído esforços de pesquisadores, baseada na disseminação da tecnologia de bancos de dados e na suposição de que as grandes bases de dados atualmente existentes, podem ser uma fonte de conhecimento útil, que não está explicitamente representado. O objetivo da área é desenvolver e validar técnicas, metodologias e ferramentas capazes de extrair o conhecimento implícito em bases de dados e representar este conhecimento de uma forma acessível para o usuário [FEL 96].

Segundo [FAY 96], o KDD é um processo não trivial de identificar padrões válidos, não conhecidos, potencialmente úteis e interpretáveis em uma base de dados. Trata-se de uma tarefa intensiva de busca de conhecimento implícito que consiste em interações entre um usuário e uma grande base de dados, possivelmente apoiada por alguma ferramenta.

Assim, o processo de KDD visa automatizar a extração de conhecimento, a partir da identificação de padrões e relacionamentos implícitos nos dados armazenados. Esta descoberta de conhecimento resulta em informação de alto nível que pode ser gerada a partir de contextos diferentes e, por isso, empregada nas mais diversas aplicações, como por exemplo: no gerenciamento de informação, no processamento de consultas, nos processos de tomada de decisão, no controle de processos, entre outros [BER 97].

Inicialmente, foram determinados vários nomes à noção de achar padrões úteis em “dados crus”, como *Data Mining*, Extração de Conhecimento, Descoberta de Informação, Mineração de Dados e Processamento do Padrão de Dados. Apenas em 1989, o termo KDD - Descoberta de Conhecimento em Base de Dados, foi utilizado para se referir a todo o processo de procura do conhecimento em dados, enfatizando a aplicação de alto nível de técnicas de *Data Mining* [FAY 96].

Segundo [PRA 97], os resultados já alcançados pelas pesquisas em KDD são consequência da convergência de interesses de diversas áreas que antes trabalhavam isoladamente, muitas vezes sobre temas afins, sem capitalizarem os resultados umas das outras. Áreas como Estatística, Redes Neurais, Aprendizado de Máquina, Reconhecimento de Padrões e Bancos de Dados passaram a somar esforços no cenário de KDD. Esta união de interesses tem permitido grandes avanços na área, com a geração de ferramentas que realmente auxiliam na resolução de problemas com o uso de bancos de dados.

Considerando os benefícios que a Engenharia proporciona ao processo de desenvolvimento de *software*, surge a idéia de padronizar o processo de KDD, estruturando-o

em etapas bem definidas. A necessidade desta padronização advém da possibilidade dos resultados finais de um processo de KDD não serem os esperados, constituindo-se de informações não úteis ou que não expressem a realidade.

O processo ilustrado na Figura 8 segue a padronização proposta por [FAY 96].

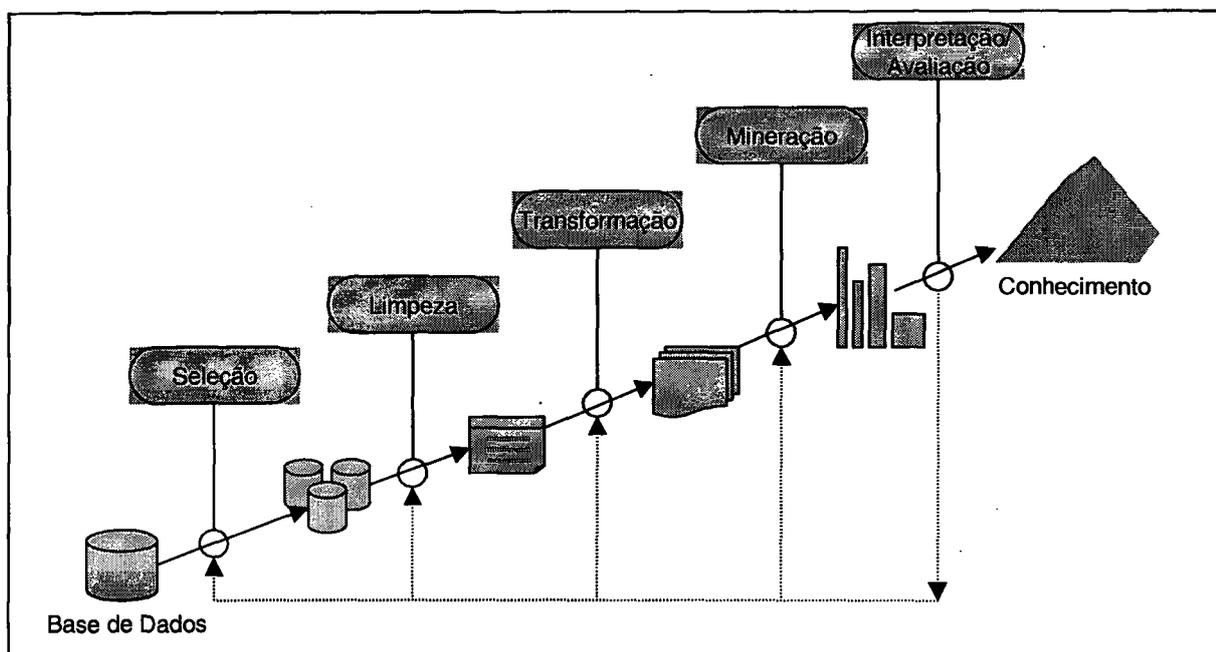


Figura 8: Etapas do processo de descoberta de conhecimento [FAY 96]

A seguir, descreve-se as etapas para chegar ao conhecimento:

- *Seleção* – é a extração dos dados necessários à aplicação. Pode ser necessário integrar e compatibilizar as bases de dados [BER 97]. É nesta etapa que se define o problema e os resultados esperados. Nesta fase inicial, é necessário conhecer o domínio sobre o qual aplicar-se-á o processo e traçar as metas que o usuário final pretende alcançar com a implantação deste processo. Trata-se de uma fase que se caracteriza pela intensa interação entre analista e usuário especialista, uma vez que, para traçar as metas, deve-se ter bem definido que tipo de conhecimento o usuário quer extrair da base de dados e em que contexto ele quer empregar

esse conhecimento. É importante salientar que estas especificações constituirão o ponto de partida no processo de descoberta de conhecimento [FAY 96];

- *Limpeza* – As informações consideradas desnecessárias são removidas e se adotam estratégias para manusear dados perdidos ou inconsistentes [QUB 01]. Se os erros não forem descobertos neste estágio, poderão contribuir para a baixa qualidade dos resultados. Esta etapa caracteriza-se por aplicar operações de remoção de ruídos sobre os dados. Ruídos são aquelas informações que não foram completadas ou que continham erros na entrada de dados. Também é necessário preocupar-se com uma forma de tratamento para informações inconsistentes [BER 97];

- *Transformação* – objetiva desenvolver um modelo sólido de dados de maneira que estes possam ser utilizados por algum algoritmo de extração de conhecimento. As transformações são ditadas pelo tipo de operação e técnica a ser utilizada. São conversões de um tipo de dados para outro, definição de novos atributos, etc. Esta etapa está diretamente relacionada ao tipo de descoberta de conhecimento que será aplicado, uma vez que encontrar características úteis para representar os dados depende dos objetivos da descoberta. Utiliza-se redução ou métodos de transformação para reduzir o número efetivo de variáveis consideradas ou para encontrar representações invariáveis de dados [QUB 01];

- *Mineração* – é o núcleo do processo, também chamado de *Data Mining*. Este estágio é relacionado com a extração de padrões dos dados. Os dados transformados são minerados usando-se uma ou mais técnicas de forma a tentar extrair o tipo de informação desejada. Enquanto se está minerando um conjunto de dados, pode ser necessário acessar dados adicionais e/ou executar transformações adicionais nos dados originalmente selecionados. Em alguns casos é necessário aplicar mais de uma técnica de *Data Mining*, criando um modelo para a aplicação de uma outra técnica neste modelo. Seleciona-se a(s) técnica(s) de mineração de dados (árvores de decisão, redes neurais, algoritmos genéticos, agentes, entre outros) para

ser utilizada na descoberta de padrões nos dados. Isto inclui decidir que modelos e parâmetros podem ser mais apropriados (por exemplo, modelos para dados categóricos são diferentes de modelos para dados de valores reais). Além disso, a(s) técnica(s) de mineração deve(m) ser compatível(is) com as metas do processo de KDD. É a etapa de descoberta de padrões de interesse em uma forma de representação particular ou um conjunto de tais representações - regras de classificação ou árvores, associações, agrupamento e tantos outros [DWB 00] [QUB 01];

- *Interpretação e avaliação* – verifica a validação do conhecimento extraído da base de dados. Os padrões identificados são interpretados em conhecimento, que poderão ser utilizados para dar apoio às tomadas de decisões. O objetivo da interpretação dos resultados é filtrar as informações que serão apresentadas aos tomadores de decisão. Se os resultados não forem satisfatórios, pode ser necessário repetir o passo do *Data Mining* ou retomar a qualquer um dos estágios anteriores. Somente após a avaliação dos resultados é que se encontra conhecimento. É a etapa de avaliar os conhecimentos extraídos no processo e distinguir o que constitui e o que não constitui conhecimento útil. Nesta etapa, dependendo do conhecimento obtido, surge a necessidade de retornar a alguma etapa precedente para iterações adicionais [FAY 96] [DAT 00] [QUB 01];

Após o cumprimento das etapas do Processo KDD, atinge-se a Descoberta de Conhecimento que se caracteriza por incorporar o conhecimento ao sistema, ou simplesmente documentá-lo e enviá-lo às partes interessadas [BER 97].

Os tipos de descoberta de conhecimento mais utilizados são as regras de associação, a classificação, o agrupamento (clustering) e a regressão linear. Estas técnicas serão descritas sumariamente a seguir.

3.2.1.1 Regras de Associação

A mineração de regras de associação em bancos de dados relacionais ou transacionais tem recentemente atraído grande atenção da comunidade de banco de dados [CHE 96].

Uma regra de associação é uma expressão do tipo $X \rightarrow Y$, onde X e Y são conjuntos de atributos, significando que nas linhas do banco de dados onde os atributos em X têm valores verdadeiros, os atributos em Y também possuem valor verdadeiro [HOL 95].

São muitas as áreas de aplicação para regras de associação, dentre as quais pode-se destacar análise de produtos comprados em conjunto, análise de matrículas de cursos universitários, descoberta de conjunto de produtos freqüentemente encomendados de um fabricante. Uma área especial de aplicação é a análise do comportamento de clientes no setor de vendas a varejo, a qual é chamada *análise de compras (basket analysis)*: quais itens o cliente freqüentemente compra junto em um supermercado. Por exemplo, numa base de dados, é possível encontrar uma regra de associação do tipo “Se um cliente compra (uma marca de) leite, este mesmo cliente usualmente compra (uma marca de) pão na mesma transação” [BER 97].

As associações se aplicam bem a estes domínios, pois permitem encontrar uma descrição compacta para um subconjunto de dados através da descoberta de relações funcionais entre variáveis [BER 97].

Segundo [AGR 96], um problema potencial, que usuários experientes podem enfrentar ao aplicar regras de associação em problemas reais, é que muitas regras não interessantes ou redundantes podem ser geradas juntamente com regras interessantes.

Além disso, a mineração de regras de associação pode requerer examinar repetidamente as transações do banco de dados para encontrar padrões de associação diferentes, o que pode ocasionar um processamento imenso. Assim, uma meta essencial na

mineração de regras de associação é a otimização da performance do sistema através da aplicação de métodos e algoritmos eficientes [AGR 96].

No contexto de mineração de dados, as regras de associação, também denominadas regras de produção, servem tipicamente para representar com que frequência os padrões ocorrem nos dados. Sua principal função é a de caracterizar os dados, representando as regularidades encontradas. Um exemplo clássico seria analisar os registros de compras de um supermercado e observar entre três produtos qual a tendência dos mesmos serem comprados em conjunto. Esta informação pode ser o ponto de partida para uma campanha promocional que melhore os resultados das vendas [BER 97].

As regras de associação extraem o conhecimento através de uma ou mais regras, representadas genericamente na forma *Se (X) então (Y)*, onde X e Y são conjuntos de itens e $X \cap Y = \varnothing$ (X intersecção com Y = vazio). Assim, um exemplo de regra que poderia ser gerada a partir de uma base de dados é: “*Se salário = alto e crédito = baixo então empréstimo = Negado*” [BER 97].

As regras de associação representam os elementos ou itens com comportamento conjunto nos eventos descritos na base de dados. Portanto, o objetivo é achar a intersecção entre os elementos sendo analisados [DWB 01].

Segundo [BER 97] esta técnica, ao ser aplicada em uma base de dados, pode gerar uma grande quantidade de associações entre os dados. No entanto, muitas destas são de pouca importância, e caracterizam regras inúteis. Como forma de limitar a quantidade de regras que serão geradas, são definidos, antes de iniciar o processo de mineração, dois atributos que descrevem a qualidade destas regras:

- *suporte* (ou frequência): calculado em função do número de registros em que aparecem todos os itens analisados e pelo total de transações pesquisadas na base de dados. Assim, se

a análise está sendo feita para os itens X e Y, a expressão matemática que define o valor do suporte é definida por:

$$\text{Suporte} = \frac{\text{N}^\circ \text{ Transações com itens X e Y}}{\text{Total de Transações}}$$

- *fator de confiança*: calculado em função do número de registros em que aparecem todos os itens que estão sendo analisados e pelo total de transações de um determinado item. Considerando que os itens sendo analisados sejam X e Y, a expressão matemática que define o valor do fator de confiança é definida por:

$$\text{Fator de Confiança} = \frac{\text{N}^\circ \text{ Transações com itens X e Y}}{\text{N}^\circ \text{ Transações com X}}$$

Considera-se itens freqüentes numa base de dados os que têm suporte acima do limite estabelecido. E, o processo de extração inicia pela identificação destes itens [BER 97].

3.2.1.2 Classificação

A classificação de dados, também referenciada como regras de classificação, classifica um conjunto de dados em classes previamente estabelecidas, baseando-se nos valores de certos atributos [DAT 00].

Os recursos disponibilizados, atualmente, pelos bancos de dados, possibilitam acessar registros individualmente e informações sobre estes. Com a classificação, é possível, através da mineração de dados, a geração de uma descrição resumida de toda uma classe de registros, de acordo com as características comuns destes [DWB 01].

Para a geração de regras de classificação é necessário que se selecione um atributo alvo, também denominado variável dependente, e atributos preditores ou variáveis

independentes que serão utilizados no estudo. O atributo alvo é definido pela combinação dos atributos preditores [BER 97].

Uma regra de classificação, para clientes de um determinado banco, pode ser exemplificada da seguinte forma: “SE Salário > US\$2000 e Idade > 40 anos ENTÃO Risco de crédito = Baixo”. Os atributos preditores “Salário > US\$2000” e “Idade > 40 anos” formam o lado esquerdo da regra (LHS – *Left Hand Side*) e o atributo alvo “Risco de crédito = Baixo” corresponde ao lado direito da regra (RHS – *Right Hand Side*). Assim descreve-se que quando LHS é verdadeiro RHS também será [QUB 01].

Existem 3 tipos de categorias para as regras de classificação: as regras exatas, as fortes e as probabilísticas. As regras exatas são aquelas em que um determinado LHS resulta sempre no mesmo atributo alvo (RHS), ou seja a confiança é de 100%. Nas regras fortes o LHS gera um RHS, com um nível de confiança alto, porém inferior a 100%. As regras probabilísticas são aquelas em que se relaciona a condição de probabilidade $P(\text{RHS}/\text{LHS})$ para a probabilidade $P(\text{RHS})$ [QUB 01].

O objetivo da classificação é analisar os dados de forma a descobrir relacionamentos entre os atributos preditores e o atributo alvo, utilizando registros cuja classe é conhecida. Desta forma pode-se usar o relacionamento descoberto para prever a classe (o valor do atributo alvo) de um registro com classe desconhecida [BER 97].

Como a distribuição de frequência das classes dos atributos não é homogênea, as mais raras são as mais difíceis de serem previstas. Esta é a principal limitação do método de classificação [QUB 01].

Como uma forma de se medir o nível de informação de um atributo, se utiliza o conceito da entropia da teoria da informação. Quanto menor é o valor da entropia, menor será a incerteza e mais utilidade tem o atributo no processo de classificação [QUI 93] [WIT 00].

A informação contida em uma distribuição de probabilidade que é denominada entropia, é definida através da seguinte função [QUI 93]:

$$I(P) = - [p_1 * \log_2(p_1) + \dots + p_n * \log_2(p_n)]$$

Para um conjunto de T registros particionados em k classes, a informação necessária para a identificação da classe de um elemento de T é [QUI 93]:

$$\text{Info}(T) = I(P)$$

Se os T registros são particionados sobre a base dos valores do atributo X em conjuntos t_1, t_2, \dots, t_n , então a informação necessária para identificar a classe de um elemento T é [QUI 93]: $\text{Info}(X,T) = (t_i/T) * \text{Info}(t_i)$

A diferença entre a informação necessária para a identificação de um elemento de T antes e depois que o valor de X tenha sido considerado é denominado ganho de informação do atributo X, sendo expresso da seguinte forma [QUI 93]:

$$\text{Ganho}(X,T) = \text{Info}(T) - \text{Info}(X,T)$$

Ao se conhecer o ganho dos atributos, pode-se partir para a elaboração das árvores de decisão, haja vista que os atributos de maior ganho, são os que contém maior informação [QUI 93] [WIT 00].

A árvore de decisão representa graficamente, através de um determinado algoritmo, as regras de classificação facilitando sua interpretação de um conjunto de dados pré selecionados. Na árvore de decisão estão presentes os nós, os galhos e as folhas. Os nós são rotulados pelos nomes dos atributos, os galhos são os valores possíveis de cada atributo e as folhas são os valores das classes. Os registros são classificados seguindo um caminho para baixo na árvore, ficando a raiz no alto e as folhas embaixo [BER 97].

A Figura 9 mostra um exemplo de árvore de decisão, para um caso de concessão de crédito para clientes de uma instituição financeira. Os atributos apresentam informações

sobre a idade e o salário dos clientes e as classes são definidas como: Bom, para aqueles com crédito aprovado e Ruim para os clientes de risco [AGR 96].

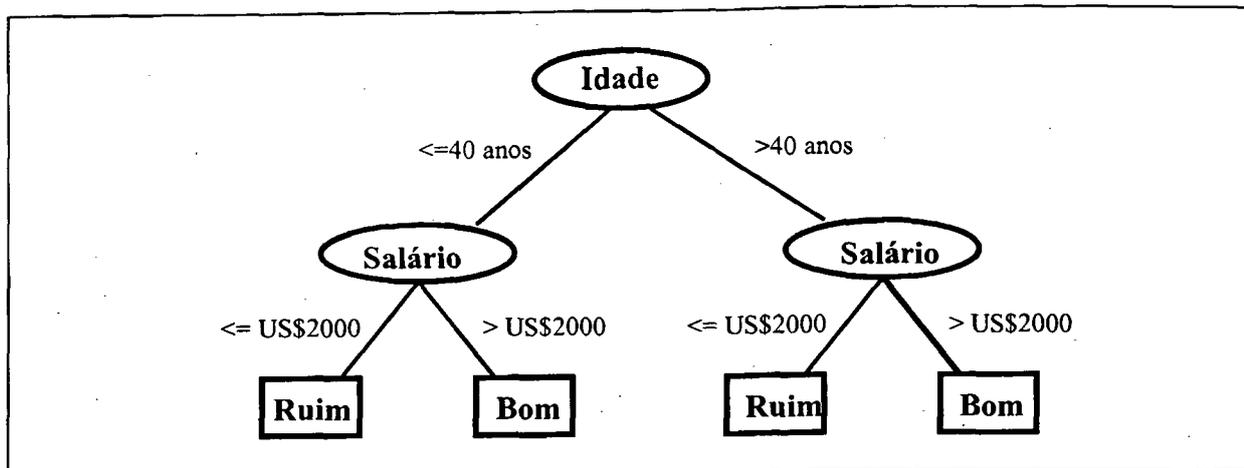


Figura 9: Exemplo de Árvore de Decisão [AGR 96].

As principais vantagens da árvore de decisão são a de que o modelo é bastante compreensível e há facilidade de expressão das regras como uma instrução lógica, aplicadas diretamente em novos registros [BER 97].

Os algoritmos mais conhecidos para se construir árvores de decisão são: CART, CHAID (*Chi-Squared Automatic Interaction Detection*), ID3 (*Iterative Dichotomiser 3*) e o C4.5 [BER 97].

Em 1993 o pesquisador australiano J. Ross Quinlan desenvolveu a partir do ID3 o algoritmo C4.5, atualmente encontrado em vários produtos comerciais. O algoritmo cria um conjunto de regras, as agrupa e elimina aquelas que não contribuem para a precisão do conhecimento a ser extraído. Assim se forma um pequeno conjunto de regras de fácil entendimento, obtidas pela combinação das regras que levam à mesma classificação [BER 97].

O algoritmo C4.5 emprega o conceito de poda (*pruning*) em árvores de decisão, onde parte da mesma é substituída por uma folha, com a finalidade de simplificar as regras de decisão [BER 97] [QUI 93] [WIT 00].

O pruning é empregado quando o valor esperado do erro da sub-árvore é maior que o erro da folha que fica em seu lugar. Para o cálculos desses erros é utilizado o número de exemplos do nó (N), o número de classes (k) e o número de exemplos de N pertencentes à classe C (n). A classe C é aquela com o maior número de elementos [QUI 93].

O erro esperado do nó e erro da sub-árvore podem ser expressos por [QUI 93]:

$$\text{Erro Esperado (nó)} = (N - n + k - 1) / (N + k)$$

$$\text{Erro Sub-árvore} = \text{Soma} (P_i * \text{Erro Esperado (nó } i \text{)}).$$

O algoritmo C4.5 pode ser descrito da seguinte forma [WIT 00]:

For each attribute,

For each value of that attribute, make a rule as follows:

Count how often each class appears

Find the most frequent class

Make the rule assign that class to this attribute-value.

Calculate the error rate of rules.

Choose the rules with the smallest error rate.

For each class C

Initialize E to the instance set

While E contains instances in class C

Create a rule R with an empty left-hand side that predicts class C

Until R is perfect (or there are no more attribute to use) do

For each attribute A not mentioned in R, and each value v,

Consider adding the condition A=v to the LHS of R

Select A and v to maximize the accuracy p/t

(break ties by choosing the condition with the largest p)

Add A=v to R

Remove the instances covered by R from E.

A técnica de classificação pode ser aplicada em muitas áreas do conhecimento humano. Como exemplos da utilização de regras de classificação têm-se:

- *aplicação em clínicas médicas ou hospitais:* utilização de dados históricos, para obter informações do tipo “paciente que já tenha feito este tipo de cirurgia” [DAT 00];

- *aplicação no marketing de vendas*: o vendedor de carros pode classificar seus clientes de acordo com suas preferências para carros. Deste modo, conhecidas suas preferências, o vendedor pode utilizar tal classificação para catalogar novos modelos de carros que podem ser direcionados para aqueles clientes com as características identificadas, como forma de maximizar as oportunidades de negócio [CHE 96].

3.2.1.3 Agrupamento (*Clustering*)

Este tipo de descoberta de conhecimento segmenta o banco de dados em diferentes grupos. A meta é encontrar grupos que diferem um do outro tão bem quanto as similaridades entre seus membros. O banco de dados é dividido em conjuntos de registros, onde registros com conteúdo similar são colocados num mesmo grupo e os diferentes agrupados como um possível outro grupo. Ao contrário da classificação, não se conhece quais são os grupos quando se inicia, ou em torno de quais atributos os dados se agruparão. Consequentemente, um analista de negócios deve ter conhecimento para interpretar o significado dos grupos [QUB 01].

Assim, como o principal objetivo do agrupamento é identificar um conjunto finito de categorias ou agrupamentos para descrever dados. Um exemplo da aplicação deste tipo de descoberta é em companhias de seguro para agrupar clientes de acordo com renda, idade, tipos de apólices compradas e outras características [QUB 01].

3.2.1.4 Regressão Linear

A regressão linear é um tipo de descoberta que utiliza uma função de aprendizado, que projeta um valor real específico, baseado numa variável de precisão associada a um item de dado [KEN 98].

Pode-se utilizar a regressão linear em casos onde a saída esperada pode ser ilimitada ou pelo menos com muitos valores possíveis. Devido a estas características, são muitas as aplicações de regressão linear. Como exemplo, pode-se citar: predição da quantidade de biomassa presente em uma dada floresta, através de medidas capturadas por sensoriamento remoto; estimativa da probabilidade de morte de um paciente dado um conjunto de diagnósticos [KEN 98].

3.2.2 Ferramentas de Análise de Dados

Várias ferramentas de análises de dados estão disponíveis comercialmente. Algumas apresentam limitações de uso quanto ao número de registros, outras são muito complexas e têm alto custo.

Como exemplos de ferramentas têm-se: Intelligent Miner da IBM, Oracle Data Mining Suite da Oracle Darwin, SPSS Neural Connection, IBM Neural Network Utility, IDIS da Information Discovery, Knowledge Seeker da Angoss Software, SAS System, IBM Parallel Visual Explorer, Aira Data Mining, See 5.0 e o CBA Data Mining II.

CBA é uma ferramenta de mineração de dados desenvolvida pela *School of Computing* da *National University of Singapore*. Esta ferramenta não tem limitações quanto ao número máximo de registros e está disponível de forma gratuita na rede de computadores.

É uma ferramenta útil para se descobrir padrões de relacionamentos entre os dados, haja vista que estão presentes dois módulos de análise: o de associação e o de classificação.

O módulo de associação é baseado no algoritmo *Apriori*, onde se analisam os valores do suporte e da confiança dos dados.

No módulo de classificação ocorre uma unificação com as regras de associação, onde as regras geradas não são derivadas apenas da aplicação do algoritmo C4.5, formando-se

então, regras de classificação mais precisas. O sistema é capaz de descobrir todas as regras possíveis, porém como muitas são redundantes, ou não adicionam conhecimento, estas são removidas através de seu nível de significância medido pelo Teste Qui-Quadrado.

Em virtude de sua praticidade, facilidade de acesso e por ser gratuita, foi selecionada como ferramenta de uso no presente estudo.

4. METODOLOGIA EMPREGADA NO ESTUDO

Um conjunto de processos e técnicas foi utilizado neste trabalho para ultrapassar a subjetividade do autor e se alcançar os objetivos do estudo. Inicialmente, determinou-se a população de estudo, selecionou-se, no banco de dados, as variáveis interessantes para o estudo e estabeleceu-se um plano de análise dos dados.

As informações necessárias à realização da pesquisa foram coletadas em um banco de dados sobre a situação periodontal dos alistados do Exército residentes em Florianópolis em 1999.

4.1 População de Estudo

A população de estudo foi composta exclusivamente, por jovens do sexo masculino, com 18 anos completados em 1999 e residentes na cidade de Florianópolis. Sendo portanto excluídos aqueles que prestaram alistamento militar nesta cidade, mas que nela não residem, como os alistados das cidades próximas. Foram ainda excluídos os refratários, ou seja, os alistados que tinham idade superior a 18 anos, e que portanto estavam se apresentando fora da idade correta do alistamento militar.

4.2 O Banco de Dados

O banco de dados consiste de uma amostra referente ao número de pessoas que necessitavam ser estudadas para responder os objetivos do estudo da prevalência da doença periodontal.

O cálculo estatístico amostral, foi obtido de um universo contendo apenas alistados residentes em Florianópolis. Como um dos princípios básicos para a determinação

do tamanho da amostra, é o de que a mesma deva ter precisão para estimar e poder demonstrar diferenças significativas utilizou-se a expressão matemática 4.1, proposta por [SOA 91]:

$$n = \frac{N \cdot z^2 \cdot p \cdot (1 - p)}{d^2 \cdot (N - 1) + z^2 \cdot p \cdot (1 - p)} \quad (4.1)$$

Onde foram obedecidos os seguintes parâmetros:

N = Tamanho da população (3452 alistandos);

z = Nível de confiança expresso em desvio padrão (95%) = 1,96;

d = Precisão ou erro mínimo permitido (4% ou 0,04);

p = Prevalência da doença periodontal na população (86% ou 0,86)

O parâmetro utilizado para se considerar a prevalência da doença periodontal ($p=0,86$) foi extraído do Levantamento Epidemiológico em Saúde Bucal feito no Brasil, Zona Urbana, em 1986, pelo Ministério da Saúde do Brasil, para a faixa etária entre 15 e 19 anos, na Região Sul do Brasil [MIN 88]. Como esta é a única fonte de dados sobre a distribuição da doença periodontal no Brasil, adotou-se a mesma para comparações com os resultados do presente estudo.

O tamanho da amostra encontrado após a aplicação da fórmula estatística foi de 267 indivíduos. Foram acrescentados mais 10% da amostra (27 indivíduos), para estimar eventuais perdas durante os exames clínicos. A amostra foi aumentada para 300 indivíduos (arredondando-se 294 para 300), sendo este o número de registros do banco de dados.

4.2.1 A Composição do Banco de Dados

O banco de dados é composto por dados clínicos relacionados às doenças periodontais e dados não clínicos como o tabagismo (dados comportamentais), a escolaridade e a renda (dados socioeconômicos).

Os dados clínicos apresentam as seguintes variáveis (atributos): presença de sangramento gengival à sondagem clínica, presença de cálculo (tártaro), presença de bolsas periodontais. Todos estes atributos originam classes dicotômicas, informando sobre a ocorrência ou não do atributo.

Os dados comportamentais apresentam o atributo hábito de fumar cigarro com as seguintes classes: fumante (para os que tinham o hábito), ex-fumante (para os que pararam de fumar a menos de 1 ano) e não fumantes (para os que nunca fumaram ou que deixaram de fumar a mais de 1 ano).

Os dados socioeconômicos continham os atributos relacionados à escolaridade do alistando, do pai do alistando e da mãe do alistando, além da renda familiar. Todos estes atributos originam classes dicotômicas, sendo a escolaridade dividida em indivíduos com até 8 anos de estudo ou com mais de 8 anos de estudo. A renda familiar contém a classe dos que recebem até 5 salários mínimos e a dos que recebem além de 5 salários mínimos.

Os dados socioeconômicos e comportamentais foram extraídos através de um questionário aplicado nos alistandos (Apêndice 1).

4.3 Plano de Análise dos Dados

Realizou-se dois tipos de análises sobre banco de dados. Uma análise estatística e outra utilizando a técnica de mineração de dados que gera regras de classificação pelo processo de descoberta de conhecimento em base de dados (KDD).

4.3.1 Análise Estatística

Utilizou-se como ferramenta de trabalho o Software SPSS for *Windows Release 8.0.0 Standard Version*, onde se analisou os seguintes aspectos:

- a prevalência da doença periodontal na população estudada;
- as proporções de indivíduos com sextantes sadios, com sangramento, com cálculo, com bolsas periodontais rasas (entre 3,5 e 5,5mm), com bolsas periodontais profundas (maiores de 5,5mm) e nulos para a população estudada;
- as proporções de fumantes, ex-fumantes e não fumantes na população estudada;
- o teste de associação qui-quadrado entre as variáveis clínicas (sangramento gengival, cálculo e bolsas periodontais) e as variáveis socioeconômicas (grau de escolaridade e renda familiar);
- o teste de associação qui-quadrado entre as variáveis clínicas (sangramento gengival, cálculo e bolsas periodontais) e o hábito de fumar cigarro.

4.3.2 Análise pelo Processo KDD

A análise pelo processo KDD foi feita, utilizando-se como ferramenta de trabalho o software CBA Data Mining II. Por meio desta análise geram-se regras de classificação em mineração de dados para os atributos alvos: sangramento gengival, cálculo dentário e bolsas periodontais.

Adotou-se 10% como suporte mínimo para a seleção das regras de classificação. Desta forma, limita-se a quantidade de regras geradas pelo software, facilitando a análise das mesmas.

Para o funcionamento do software é necessário que os dados sejam preparados em dois arquivos distintos. O primeiro arquivo deve ser do tipo *.data, onde estarão todos os dados da matriz de dados. O segundo arquivo deverá ser do tipo *.names, onde estarão

presentes a relação dos atributos e suas possíveis classes. A conversão dos dados pode ser facilmente conseguida utilizando-se ferramentas como Dbase ou Microsoft Excel, além do Microsoft WordPad que armazena os dois arquivos citados.

Como se pretende classificar as três variáveis clínicas, sangramento, cálculo e bolsas, gerou-se 3 arquivos NAMES (com a descrição das variáveis e suas respectivas categorias, bem como a categoria alvo da análise) e três arquivos DATA (com os dados propriamente ditos, no formato texto) presentes nos Apêndices 2, 3, 4, 5, 6 e 7.

5. APRESENTAÇÃO DOS RESULTADOS E DISCUSSÃO

Esta análise de banco de dados inicia-se com apresentação dos dados inconsistentes, em seguida são apresentadas as distribuições de frequência das variáveis, a análise estatística e a análise pelo processo KDD.

5.1 Avaliação dos Ruídos

O banco de dados apresentou poucos dados inconsistentes (ruídos), sendo estes inferiores a 10%. Observou-se estes ruídos, em 13 alistandos que apresentavam aparelho fixo em pelo menos uma arcada, fato este que inviabilizou a coleta de dados para o exame clínico (4,33%). O questionário socioeconômico apresentou, para a pergunta “Até que série seu pai estudou na escola?”, 25 respostas “Não sabe informar”, sendo esta a questão com maior grau de desconhecimento por parte dos alistandos (8,3%).

5.2 Distribuição de Frequência dos Dados Socioeconômicos

Os dados socioeconômicos são aqueles referentes aos níveis de renda familiar e de escolaridade dos alistandos e de seus pais.

5.2.1 Escolaridade dos Pais e das Mães dos Alistandos

As Tabelas 2 e 3 mostram os graus de escolaridade dos pais e das mães dos alistandos. Observa-se que os pais têm escolaridade ligeiramente superior às mães, principalmente se observarmos o primeiro quartil, onde 25% das mães só estudaram até a 4ª série do I grau, enquanto que 25% dos pais estudaram até a 5ª série do I Grau e também no percentual de pessoas que concluíram o ensino superior (15,4% das mães e 21,8% dos pais).

A escolaridade dos pais e das mães dos alistandos apresentaram níveis muito superiores aos nacionais, porém esta é uma peculiaridade do município de Florianópolis, conforme dados da Contagem da População de 1996 do IBGE e que podem ser visualizados no Anexo 2 [IBG 96]. A moda estatística, segundo o IBGE, para os homens residentes em Florianópolis na faixa etária de 40 a 59 anos de idade é ter 15 ou mais anos de estudo, e nesta pesquisa observou-se que os pais dos alistandos têm como moda o curso superior completo.

A maioria das mulheres de 20 a 49 anos de idade e residentes em Florianópolis tem entre 11 e 14 anos de estudo [IBG 96]. Na presente pesquisa a moda estatística das mães dos alistandos é a 4ª série do I Grau (51 ocorrências), porém com um valor muito próximo da 3ª série do II Grau (49 ocorrências) e do curso superior completo (44 ocorrências).

O padrão educacional brasileiro é segmentado em I grau (primário e ginásio), II grau e III grau (ensino superior). Os indivíduos de 1 a 4 anos de estudo correspondem aqueles que estudaram no máximo até a 4ª série do I grau, ou seja, não foram além do primário. Aqueles que têm de 5 a 8 anos de estudo alcançaram o ginásio. Os pais e mães que têm de 9 a 11 anos de estudo alcançaram o II grau e aqueles com 12 ou mais anos de estudo atingiram o ensino superior.

As Figuras 10 e 11 facilitam a interpretação das Tabelas 2 e 3, pois os níveis de escolaridade dos pais e das mães dos alistandos são agrupados conforme o padrão educacional brasileiro. Existe um maior percentual de mães nos níveis mais baixos de estudo, primário e ginásio. Os pais mostram-se mais freqüentes no II grau e no ensino superior.

Tabela 2: Distribuição de frequência para a pergunta “Até que série seu pai estudou na escola?”.

Resposta	Frequência Absoluta	Frequência Percentual	Freq. Percentual (Respostas Válidas)	Freq. % Acumulada (Respostas Válidas)
Não estudou	9	3,0	3,3	3,3
1ª série I Grau	7	2,3	2,5	5,8
2ª série I Grau	2	0,7	0,7	6,5
3ª série I Grau	12	4,0	4,4	10,9
4ª série I Grau	34	11,3	12,4	23,3
5ª série I Grau	20	6,7	7,3	30,5
6ª série I Grau	10	3,3	3,6	34,2
7ª série I Grau	9	3,0	3,3	37,5
8ª série I Grau	37	12,3	13,5	50,9
1ª série II Grau	9	3,0	3,3	54,2
2ª série II Grau	11	3,7	4,0	58,2
3ª série II Grau	47	15,7	17,1	75,3
Superior Incompleto	8	2,7	2,9	78,2
Superior Completo	60	20,0	21,8	100,0
Total	275	91,7	100,0	-
Não souberam informar	25	8,3	-	-
Total	300	100,0	-	-

Observações:

Moda = Superior completo; Mediana = 8ª série I Grau; 1º Quartil = até a 5ª série I Grau;

2º Quartil = até a 8ª série I Grau; 3º Quartil = até a 2ª série II Grau

Tabela 3: Distribuição de frequência para a pergunta “Até que série sua mãe estudou na escola?”.

Resposta	Frequência Absoluta	Frequência Percentual	Freq. Percentual (Respostas Válidas)	Freq. % Acumulada (Respostas Válidas)
Não estudou	15	5,0	5,3	5,3
1ª série I Grau	5	1,7	1,8	7,0
2ª série I Grau	4	1,3	1,4	8,4
3ª série I Grau	7	2,3	2,5	10,9
4ª série I Grau	51	17,0	17,9	28,8
5ª série I Grau	23	7,7	8,1	36,8
6ª série I Grau	13	4,3	4,6	41,4
7ª série I Grau	10	3,3	3,5	44,9
8ª série I Grau	38	12,7	13,3	58,2
1ª série II Grau	8	2,7	2,8	61,1
2ª série II Grau	7	2,3	2,5	63,5
3ª série II Grau	49	16,3	17,2	80,7
Superior Incompleto	11	3,7	3,9	84,6
Superior Completo	44	14,7	15,4	100,0
Total	285	95,0	100,0	-
Não souberam informar	15	5,0	-	-
Total	300	100,0	-	-

Observações:

Moda = 4ª série I Grau; Mediana = 8ª série I Grau; 1º Quartil = até a 4ª série I Grau;

2º Quartil = até a 8ª série I Grau; 3º Quartil = até a 3ª série II Grau

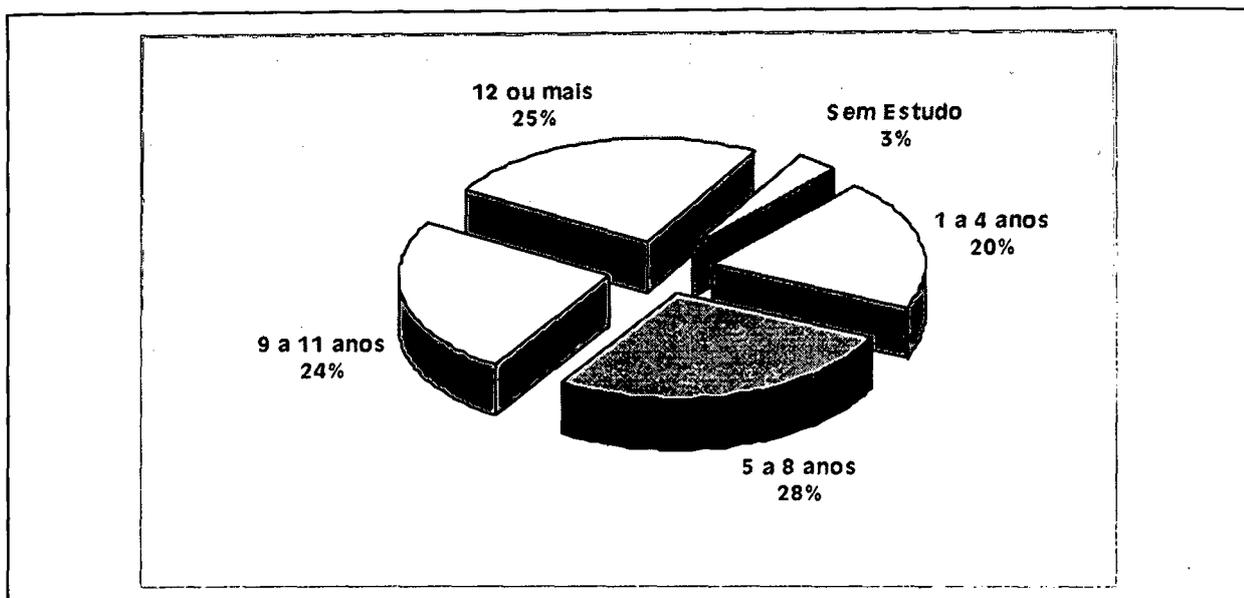


Figura 10: Anos de estudo dos pais dos alistandos

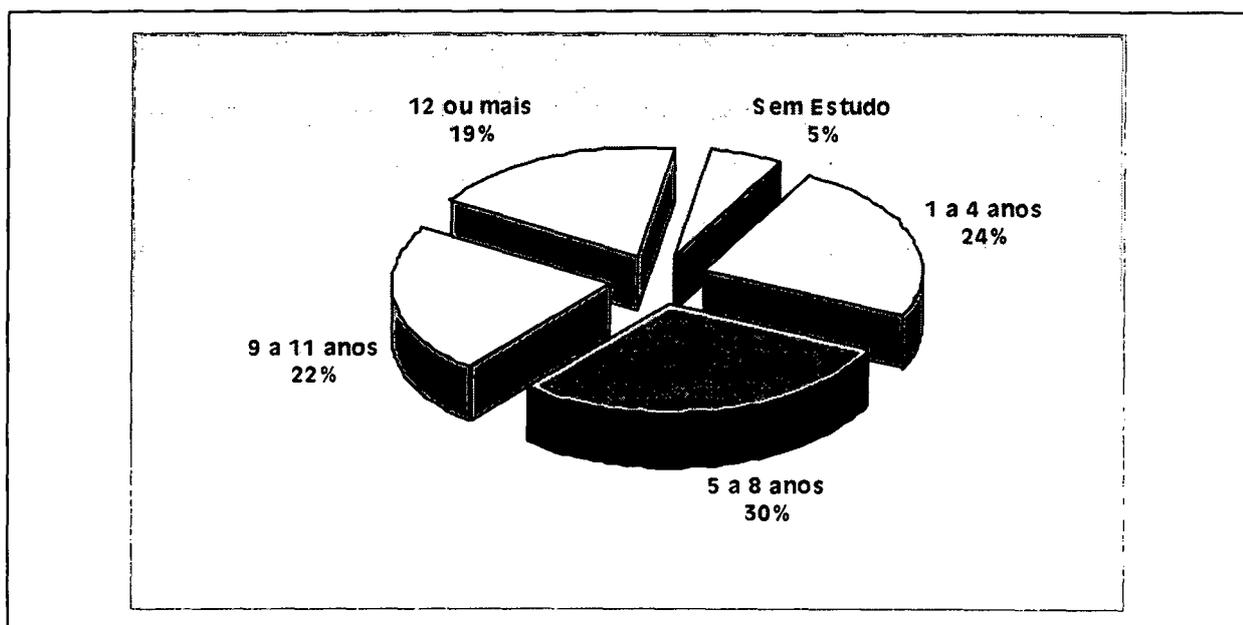


Figura 11: Anos de estudo das mães dos alistandos.

5.2.2 Escolaridade dos Alistandos

A Tabela 4 mostra a distribuição de freqüência do grau de escolaridade dos alistandos do Exército em Florianópolis. Na amostra analisada não se encontrou jovens sem estudo e a maioria já tinha cursado até o ginásio ou o 2º grau, conforme pode ser melhor visualizado na Figura 16. Este incremento no nível de instrução dos alistandos em relação aos seus pais pode ser justificado pelos dados da PNAD de 1998 (Pesquisa Nacional por Amostra de Domicílios) que conclui que vem subindo o nível de instrução no Brasil, pois em 1993 apenas 14,4% da população tinha pelo menos 11 anos de estudo (2º grau completo) e em 1998 passou-se para 18% [IBG 98]. Ressalta-se ainda que, mais de 50% dos jovens entre 15 e 19 anos de idade e residentes em Florianópolis tem acima de 8 anos de estudo [IBG 96].

Tabela 4: Distribuição de freqüência para a pergunta “Até que série você estudou na escola?”

Resposta	Freqüência Absoluta	Freqüência Percentual	Freq. Percentual (Respostas Válidas)	Freq. % Acumulada (Respostas Válidas)
2ª série I Grau	1	0,3	0,3	0,3
3ª série I Grau	5	1,7	1,7	2,1
4ª série I Grau	16	5,3	5,5	7,6
5ª série I Grau	28	9,3	9,7	17,2
6ª série I Grau	28	9,3	9,7	26,9
7ª série I Grau	22	7,3	7,6	34,5
8ª série I Grau	27	9,0	9,3	43,8
1ª série II Grau	25	8,3	8,6	52,4
2ª série II Grau	40	13,3	13,8	66,2
3ª série II Grau	73	24,3	25,2	91,4
Superior Incompleto	25	8,3	8,6	100,0
Total	290	96,7	100,0	-
Não souberam informar	10	3,3	-	-
Total	300	100,0	-	-

Observações:

Moda = 3ª série II Grau; Mediana = 1ª série II Grau; 1º Quartil = até a 6ª série I Grau;

2º Quartil = até a 1ª série II Grau; 3º Quartil = até a 3ª série II Grau

A Figura 12 mostra a segmentação dos níveis de escolaridade dos alistandos, conforme o padrão educacional brasileiro. É interessante observar que os jovens apresentam maior grau de escolaridade que os respectivos pais, conforme pode ser melhor visualizado na Figura 13.

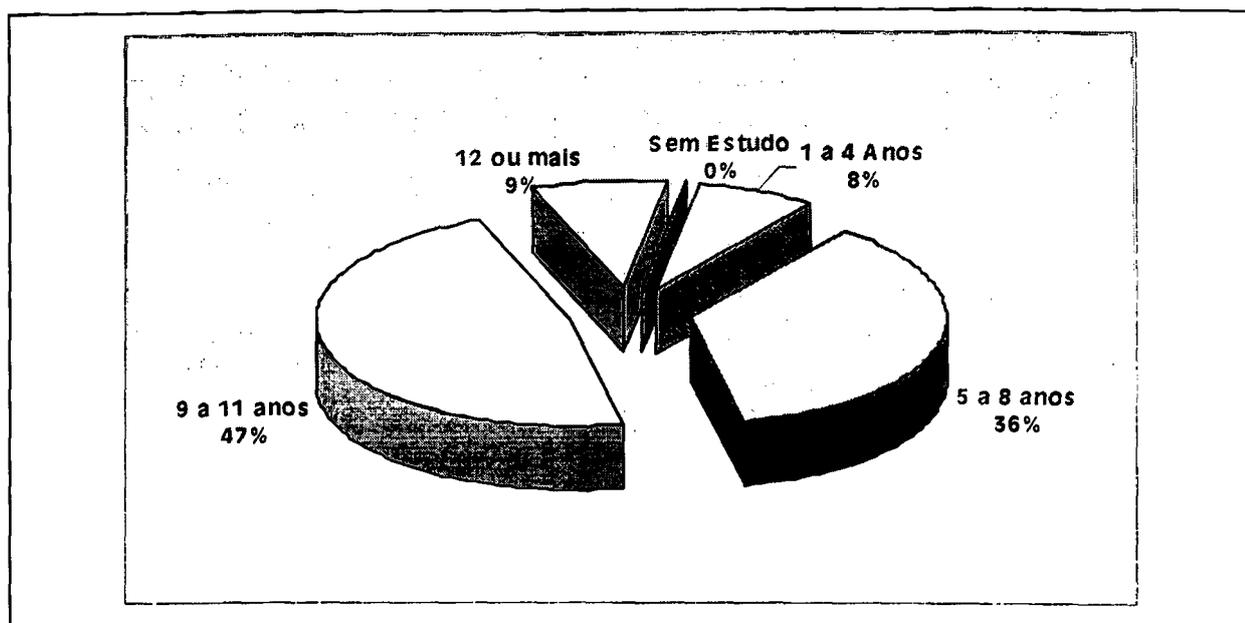


Figura 12: Anos de estudo dos alistandos.

Em Santa Catarina, no ano de 1998, para um contingente populacional de 5.044.996, a população com mais de 15 anos de estudo era de 165.345 (3,28%), com menos de um ano de estudo ou sem instrução era de 363.267 (7,20%), e com 10 anos de instrução era de 131.926 (2,61%) [IBG 98].

Em comparação com anos anteriores os dados educacionais vêm apresentando tendências positivas, mais lentas do que o ideal, porém significativas. Como esta tendência vem se confirmando no decorrer dos anos, justifica-se o fato dos alistandos terem mais acesso à educação do que seus pais.

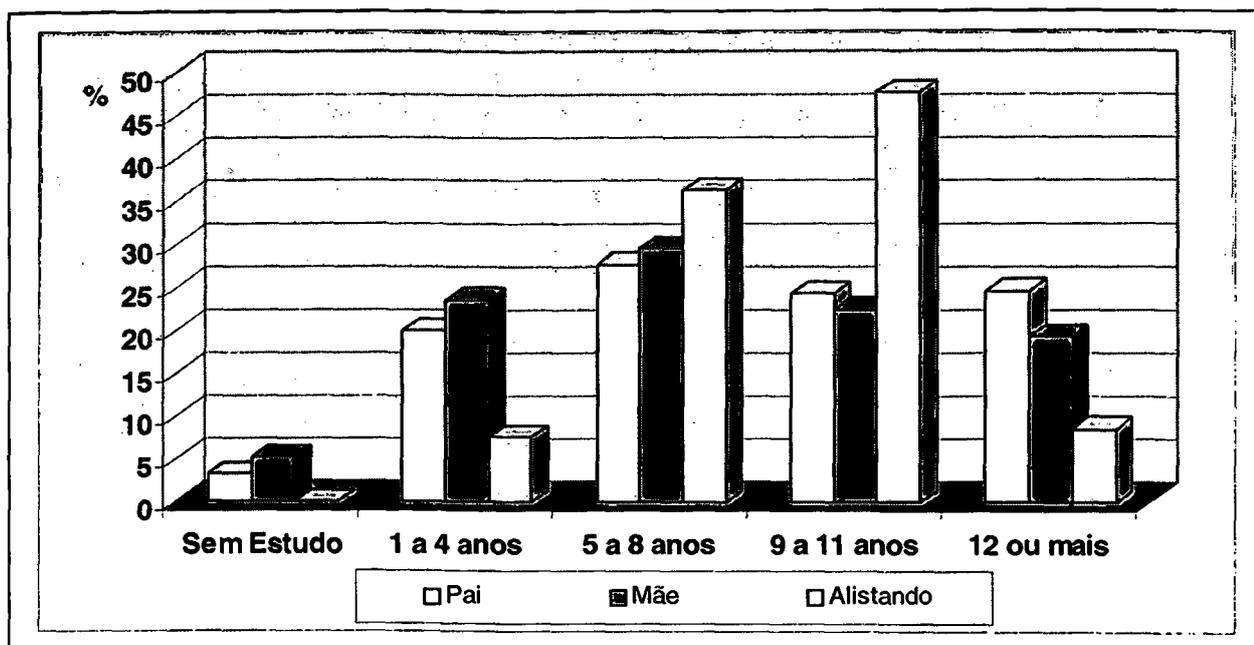


Figura 13: Comparação do grau de escolaridade dos alistas com os pais e a mães.

5.2.3 A Renda Familiar dos Alistados

A Tabela 5 mostra a distribuição da renda familiar nos domicílios dos alistas expressas em salários mínimos correspondente a R\$130,00, ou US\$71,00 (câmbio comercial) em setembro de 1999. É importante destacar que a renda familiar corresponde à somatória dos rendimentos de todas as pessoas que trabalham e moram no domicílio. A Figura 14 mostra, através da curva de tendência polinomial, que a medida em que se eleva a renda familiar se tem uma redução na quantidade de famílias envolvidas.

Os níveis de renda encontrados em Florianópolis, são bem superiores aos seus correspondentes a nível nacional, onde se mostra que 11% das famílias brasileiras ganham até 1 salário mínimo, pouco mais de 2% dos trabalhadores têm renda superior a 20 salários mínimos, e a renda mensal média está em torno de 3,9 salários mínimos [IBG 98].

Tabela 5: Distribuição de frequência da renda familiar dos alistandos.

Respostas em Salários Mínimos	Frequência Absoluta	Frequência Percentual	Freq. Percentual (Respostas Válidas)	Freq. % Acumulada (Respostas Válidas)
01	7	2,3	2,4	2,4
02	14	4,7	4,9	7,3
03	30	10,0	10,5	17,8
04	27	9,0	9,4	27,2
05	17	5,7	5,9	33,1
06	23	7,7	8,0	41,1
07	16	5,3	5,6	46,7
08	16	5,3	5,6	52,3
09	13	4,3	4,5	56,8
10	12	4,0	4,2	61,0
11	15	5,0	5,2	66,2
12	4	1,3	1,4	67,6
13	5	1,7	1,7	69,3
14	5	1,7	1,7	71,1
15	17	5,7	5,9	77,0
16	3	1,0	1,0	78,0
17	1	0,3	0,3	78,4
18	8	2,7	2,8	81,2
19	2	0,7	0,7	81,9
20	2	0,7	0,7	82,6
21	3	1,0	1,0	83,6
22	11	3,7	3,8	87,5
23	1	0,3	0,3	87,8
24	3	1,0	1,0	88,9
25	1	0,3	0,3	89,2
26	1	0,3	0,3	89,5
27	1	0,3	0,3	89,9
28	1	0,3	0,3	90,2
29	5	1,7	1,7	92,0
30	2	0,7	0,7	92,7
31	1	0,3	0,3	93,0
33	1	0,3	0,3	93,4
36	1	0,3	0,3	93,7
37	8	2,7	2,8	96,5
39	1	0,3	0,3	96,9
40 ou Mais	9	3,0	3,1	100,0
Total	287	95,7	100,0	-
Não Informado	13	4,3	-	-
Total	300	100,0	-	-

Observações:

Moda = 3 salários mínimos; Média = 12,1672 salários mínimos; Mediana = 8 salários mínimos;

1º Quartil = até 4 salários mínimos; 2º Quartil = até 8 salários mínimos; 3º Quartil = até 15 salários mínimos;

Desvio padrão = 12,2757 salários mínimos

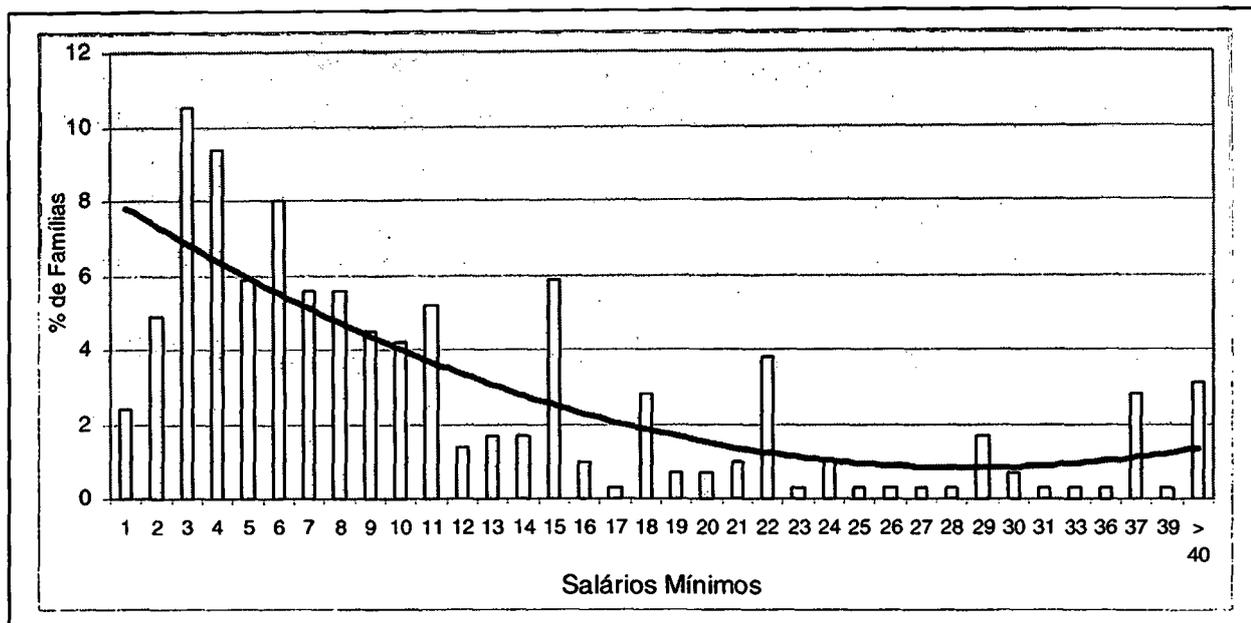


Figura 14: Frequência Percentual da Renda Familiar em Salários Mínimos e Curva de Tendência Polinomial.

5.3 Distribuição de Frequência para o Tabagismo

A distribuição de frequência do hábito de fumar cigarro pode ser vista na Tabela 6. Observa-se claramente que a maioria das pessoas na faixa etária estudada não tem o hábito de fumar cigarro, pois se somando os não fumantes e os ex-fumantes atinge-se um percentual de 81%. Como já citado, os ex-fumantes são todos aqueles que pararam de fumar a menos de 1 ano da realização dos exames clínicos bucais. As proporções de fumantes, ex-fumantes e não fumantes pode ser vista na Figura 15.

Tabela 6: Distribuição de frequência do hábito de fumar dos alistados em resposta a pergunta "Você fuma cigarro?".

	Frequência Absoluta	Frequência Percentual	Freq. Percentual dos examinados	Freq. % Acumulada dos examinados
Fumante	55	19,0	19,0	19,0
Ex-fumante	20	6,9	6,9	25,9
Não Fuma	215	74,1	74,1	100,0
Total	290	100,0	100,0	-

Observações:

Moda = Não Fuma

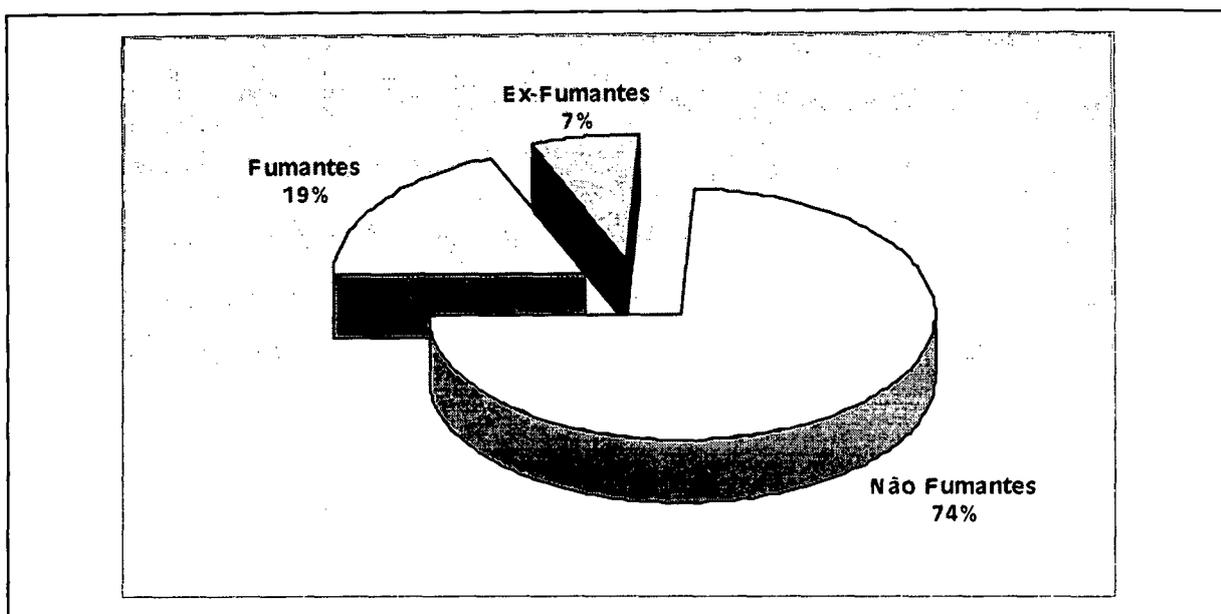


Figura 15: Proporções de alistados fumantes, não fumantes e ex-fumantes.

5.4 Distribuição de Frequência dos Dados Clínicos

Os dados clínicos coletados foram referentes ao sangramento gengival, à presença de tártaro e aos valores dos componentes do índice CPI. Na amostra de 300 alistados foram realizados 290 exames clínicos, pois alguns jovens apresentavam condições que inviabilizavam o referido exame, como o uso de aparelho ortodôntico fixo.

A ocorrência do sangramento gengival durante o exame clínico caracteriza a presença da doença periodontal. A Tabela 7 mostra a prevalência da doença (85,9%) na população de estudo, haja vista que nestas pessoas ocorre sangramento gengival. Os sadios (14,1%) são aqueles que não apresentam sangramento. A Tabela 8 mostra como é a distribuição do número de sextantes com sangramento gengival. Nota-se que quase a metade dos alistados tem mais de 2 sextantes afetados pela doença periodontal.

Tabela 7: Distribuição de freqüência dos alistandos com e sem sangramento gengival.

	Freqüência Absoluta	Freqüência Percentual	Freq. Percentual dos examinados	Freq. % Acumulada dos examinados
Sem Sangramento	41	13,7	14,1	14,1
Com Sangramento	249	83,0	85,9	100,0
Total	290	96,7	100,0	-
Não Examinados	10	3,3	-	-
Total	300	100,0	-	-

Observação: Devido ao uso de aparelho ortodôntico fixo, 10 alistandos não foram examinados.

Tabela 8: Distribuição de freqüência do número de sextantes com sangramento gengival.

	Freqüência Absoluta	Freq. Percentual dos examinados	Freq. % Acumulada dos examinados
Nenhum sextante	41	14,1	14,1
1 sextante	51	17,6	31,7
2 sextantes	58	20,0	51,7
3 sextantes	44	15,2	66,9
4 sextantes	36	12,4	79,3
5 sextantes	39	13,4	92,8
6 sextantes	21	7,2	100,0
Total	290	100,0	-

O cálculo dentário mostrou-se bastante freqüente nos alistandos, apesar de sua prevalência apresentar valores elevados em populações de faixas etárias acima de 35 anos.

A Tabela 9 mostra a prevalência de cálculo na população estudada (50,3%) e a Tabela 10 mostra como é a distribuição do número de sextantes com cálculo. É interessante notar que entre os indivíduos que possuem cálculo, a maioria (43,1%) é afetada em somente 1 sextante.

Tabela 9: Distribuição de freqüência dos alistandos com e sem cálculo dentário.

	Freqüência Absoluta	Freqüência Percentual	Freq. Percentual dos examinados	Freq. % Acumulada dos examinados
Sem Cálculo	144	48,0	49,7	49,7
Com Cálculo	146	48,7	50,3	100,0
Total	290	96,7	100,0	-
Não Examinados	10	3,3	-	-
Total	300	100,0	-	-

Observação: Devido ao uso de aparelho ortodôntico fixo, 10 alistandos não foram examinados.

Tabela 10: Distribuição de frequência do número de sextantes com cálculo dentário.

	Frequência Absoluta	Freq. Percentual dos examinados	Freq. % Acumulada dos examinados
Nenhum sextante	145	50,0	50,0
1 sextante	125	43,1	93,1
2 sextantes	14	4,8	97,9
3 sextantes	5	1,7	99,7
4 sextantes	1	0,3	100,0
Total	290	100,0	-

As prevalências de sangramento gengival, cálculo dentário e bolsas periodontais podem ser visualizadas na Figura 16. O sangramento gengival (85,9%) é a condição periodontal mais freqüentemente encontrada e a presença de bolsas periodontais é muito baixa (7,6%).

A Tabela 11 mostra como os componentes do índice CPI foram distribuídos na população de estudo. É interessante ressaltar que quando se adota os valores 2, 3 ou 4 para um indivíduo pode-se estar subestimando a ocorrência do sangramento gengival. O mesmo ocorre em relação ao cálculo, quando se adota os valores correspondentes às bolsas periodontais (3 ou 4). Isto ocorre porque quando mais de 1 componente do índice CPI está presente no exame clínico, adota-se o de código maior. Assim sempre se registra a pior condição periodontal. A Figura 17 mostra graficamente a distribuição dos componentes do CPI na amostra.

Tabela 11: Distribuição de frequência dos componentes do CPI na amostra.

Código e Componente do CPI	Número de pessoas	Frequência Percentual
0 – Sem doença periodontal	37	12,8
1 – Com sangramento à sondagem	100	34,4
2 – Com cálculo	131	45,2
3 – Com bolsas rasas	21	7,3
4 – Com bolsas profundas	1	0,3
Total	290	100,0

Observação: Existiu ainda a ocorrência de um sextante nulo em um indivíduo.

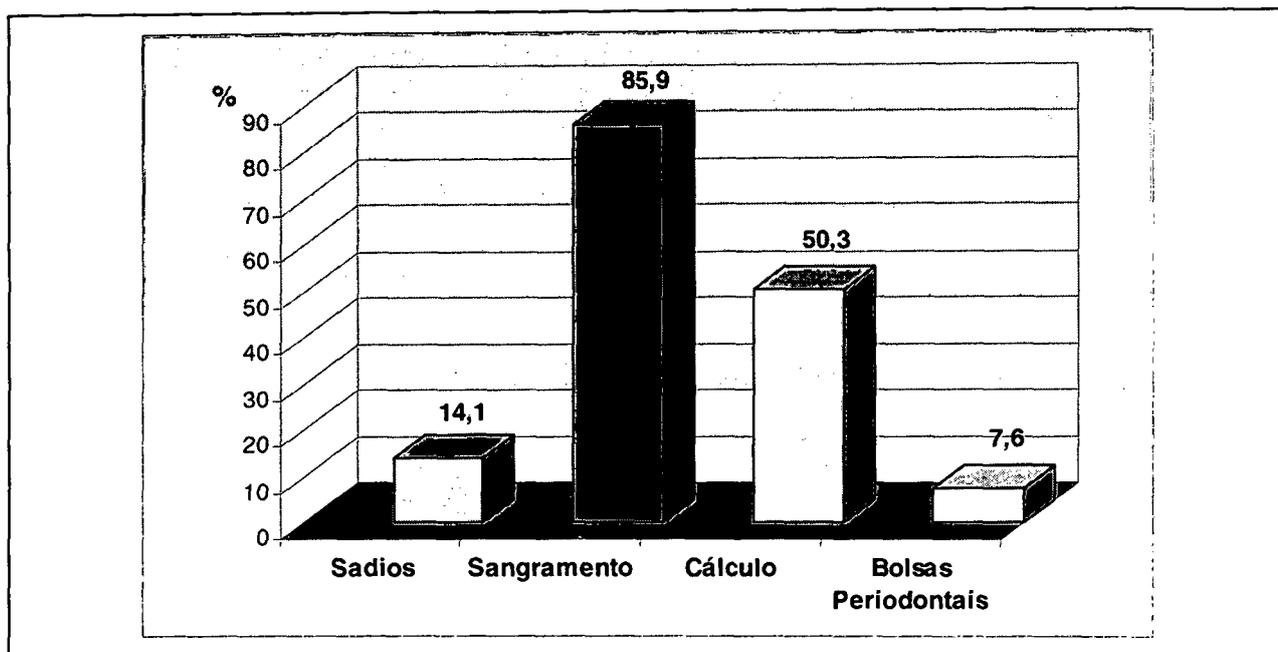


Figura 16: Prevalências das condições periodontais.

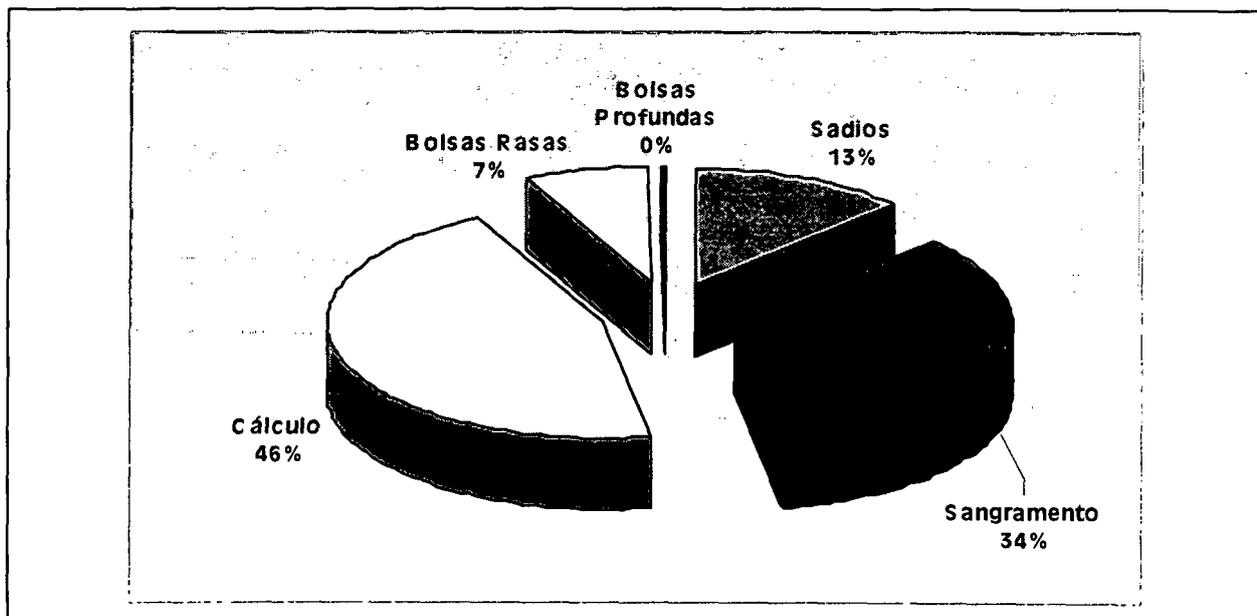


Figura 17: Distribuição percentual dos componentes do CPI.

Ao se comparar as prevalências de sangramento gengival, cálculo e bolsas periodontais (Figura 16) com as proporções dos componentes do índice CPI (Figura 17), observa-se que os valores são diferentes. Estas diferenças podem ser exemplificadas como no caso de se ter sangramento gengival e cálculo simultaneamente, sendo contabilizada nas duas prevalências. Para o índice CPI só é registrada a pior condição periodontal, no caso o cálculo, sendo a informação sobre sangramento gengival perdida.

Os componentes do índice CPI encontrados neste estudo nos alistados do Exército, mostraram-se semelhantes aos achados pelo Levantamento Epidemiológico Nacional de 1986, referentes aos jovens de 15 a 19 anos da Região Sul. Encontrou-se as mesmas proporções de sadios (cerca de 13%) e a baixa ocorrência de bolsas periodontais profundas em ambos os estudos (menos de 8%).

5.5 Análise Estatística

Além de ser influenciada por fatores socioeconômicos como a escolaridade e a renda familiar, a doença periodontal pode se relacionar com dados comportamentais como o tabagismo. Para verificar esta relação utiliza-se o Teste Estatístico Qui-quadrado (χ^2) para se confirmar a existência de associação entre a ocorrência da doença periodontal com o nível socioeconômico e com o hábito de fumar cigarro .

Para a confecção das tabelas de contingência 2 x 2 do referido teste estatístico, foi necessário definir arbitrariamente que a escolaridade e a renda sejam divididas em 2 grupos. Cada grupo deveria refletir um nível socioeconômico diferente.

A escolaridade foi dividida em um grupo contendo aqueles indivíduos que estudaram até no máximo a 8ª série do I grau, e em outro contendo os que estudaram além do I Grau.

A renda familiar foi dividida em um grupo contendo aqueles indivíduos pertencentes às famílias com renda de até 5 salários mínimos, e em outro contendo os indivíduos com renda superior a esta. Esta classificação foi utilizada no Levantamento Epidemiológico Nacional de 1986 e por isso foi adotado o mesmo procedimento.

Partiu-se então para a análise dos dados, fez-se o cruzamento das variáveis socioeconômicas (escolaridade e renda) e comportamentais (tabagismo) com os dados clínicos (sangramento gengival, cálculo e bolsa periodontal).

A Tabela 12 foi confeccionada com os dados de escolaridade do alistando e, presença ou ausência de sangramento gengival à sondagem. Mostra-se um valor de χ^2 (Qui-quadrado) de 7,303 (6,414 com a correção de Yates) e um valor de $p < 0,01$. Confirma-se assim, a hipótese de existir associação entre o grau de escolaridade do alistando e o sangramento gengival.

Tabela 12: Sangramento Gengival segundo os anos de estudo do alistando.

Anos de Estudo do Alistando	Com Sangramento	Sem Sangramento	TOTAL
Até a 8ª Série	117 (92,1%)	10 (7,9%)	127 (100,0%)
Além da 8ª série	132 (81,0%)	31 (19,0%)	163 (100,0%)
TOTAL	249 (85,9%)	41 (14,1%)	290 (100,0%)

** $p < 0,01$

A hipótese de existir associação entre o grau de escolaridade do alistando e a presença de cálculo dentário, também se confirmou. Pois, para os dados da tabela de contingência (Tabela 13), encontra-se um valor de χ^2 de 6,178 (5,604 com a aplicação da correção de Yates) e um valor do $p < 0,05$.

Tabela 13: Cálculo Dentário segundo os anos de estudo do alistando.

Anos de Estudo do Alistando	Com Cálculo	Sem Cálculo	TOTAL
Até a 8ª Série	74 (58,3%)	53 (41,7%)	127 (100,0%)
Além da 8ª série	71 (43,6%)	92 (56,4%)	163 (100,0%)
TOTAL	145 (50,0%)	145 (50,0%)	290 (100,0%)

* p<0,05

Ao se analisar a hipótese de existir associação entre o grau de escolaridade do alistando e a presença de bolsas periodontais, observa-se que a quantidade de bolsas na amostra é muito pequena, conforme pode ser visto na Tabela 14. Este fato, talvez, tenha sido o fator responsável pela não confirmação da hipótese de haver associação, pois o valor de χ^2 foi de 1,118 (0,695 com a correção de Yates) levando a um valor de p>0,05.

Tabela 14: Bolsas Periodontais segundo os anos de estudo do alistando.

Anos de Estudo do Alistando	Com Bolsas Period.	Sem Bolsas Period.	TOTAL
Até a 8ª Série	12 (9,4%)	115 (90,6%)	127 (100,0%)
Além da 8ª série	10 (6,1%)	153 (93,9%)	163 (100,0%)
TOTAL	22 (7,6%)	268 (92,4%)	290 (100,0%)

p = 0,29 (n.s.)

Verificou-se que a doença periodontal também está associada com o grau de escolaridade do pai do alistando. A escolaridade do pai e o sangramento gengival do alistando estão associados, pois com os dados obtidos na Tabela 15, aplicou-se o teste χ^2 e encontrou-se um valor de 16,167 (14,834 com a correção de Yates) resultando num p<0,01.

Tabela 15: Sangramento Gengival segundo os anos de estudo do pai do alistando.

Anos de Estudo do Pai	Com Sangramento	Sem Sangramento	TOTAL
Até a 8ª Série	131 (93,6%)	9 (6,4%)	140 (100,0%)
Além da 8ª série	103 (76,3%)	32 (23,7%)	135 (100,0%)
TOTAL	234 (85,1%)	41 (14,9%)	275 (100,0%)

** p<0,01

Ao se comparar a escolaridade do pai do alistando com a presença de cálculo dentário (Tabela 16) também se encontrou associação estatística, pois pelo Teste χ^2 tem-se um valor de 4,451 (3,957 com a correção de Yates) e um $p < 0,05$.

Tabela 16: Cálculo Dentário segundo os anos de estudo do pai do alistando.

Anos de Estudo do Pai	Com Cálculo	Sem Cálculo	TOTAL
Até a 8ª Série	79 (56,4%)	61 (43,6%)	140 (100,0%)
Além da 8ª série	59 (43,7%)	76 (56,3%)	135 (100,0%)
TOTAL	138 (50,2%)	137 (49,8%)	275 (100,0%)

* $p < 0,05$

A escolaridade do pai também se revelou associada com a presença de bolsas periodontais, mesmo com a baixa ocorrência das mesmas na amostra (Tabela 17), pois o valor de χ^2 é de 5,009 (4,023 com a correção de Yates) e o $p < 0,05$.

Tabela 17: Bolsas Periodontais segundo os anos de estudo do pai do alistando.

Anos de Estudo do Pai	Com Bolsas Period.	Sem Bolsas Period.	TOTAL
Até a 8ª Série	15 (10,7%)	125 (89,3%)	140 (100,0%)
Além da 8ª série	5 (3,7%)	130 (96,3%)	135 (100,0%)
TOTAL	20 (7,3%)	255 (92,7%)	275 (100,0%)

* $p < 0,05$

Como já citado, a escolaridade da mãe também foi objeto de análise neste estudo. Na Tabela 18 a escolaridade da mãe foi testada pelo método do χ^2 com o sangramento gengival do alistando, resultando no valor de 13,868 (12,623 com a correção de Yates) e em um $p < 0,01$, que confirma a hipótese de existir associação estatística.

Tabela 18: Sangramento Gengival segundo os anos de estudo da mãe do alistando.

Anos de Estudo da Mãe	Com Sangramento	Sem Sangramento	TOTAL
Até a 8ª Série	153 (92,2%)	13 (7,8%)	166 (100,0%)
Além da 8ª Série	91 (76,5%)	28 (23,5%)	119 (100,0%)
TOTAL	244 (85,6%)	41 (14,4%)	285 (100,0%)

** $p < 0,01$

Na tabela 19, encontram-se os valores correspondentes a escolaridade da mãe e a presença ou não de cálculo dentário no alistando. Ao se aplicar o Teste χ^2 obteve-se um valor de 9,321 (8,602 com a correção de Yates) resultando em um $p < 0,01$, que confirma a hipótese de haver associação.

Tabela 19: Cálculo Dentário segundo os anos de estudo da mãe do alistando.

Anos de Estudo da Mãe	Com Cálculo	Sem Cálculo	TOTAL
Até a 8ª Série	96 (57,8%)	70 (42,2%)	166 (100,0%)
Além da 8ª Série	47 (39,5%)	72 (60,5%)	119 (100,0%)
TOTAL	143 (50,2%)	142 (49,8%)	285 (100,0%)

** $p < 0,01$

Ao comparar-se a escolaridade da mãe com a presença de bolsas periodontais nos alistandos (Tabela 20), não se conseguiu comprovar a existência de associação estatística. Isto, talvez, devido a baixa ocorrência de bolsas periodontais na amostra. O valor de χ^2 encontrado foi de 3,549 (2,752 com a correção de Yates) e o $p > 0,05$.

Tabela 20: Bolsas Periodontais segundo os anos de estudo da mãe do alistando.

Anos de Estudo da Mãe	Com Bolsas Period.	Sem Bolsas Period.	TOTAL
Até a 8ª Série	17 (10,2%)	149 (89,8%)	166 (100,0%)
Além da 8ª Série	5 (4,2%)	114 (95,8%)	119 (100,0%)
TOTAL	22 (7,7%)	263 (92,3%)	285 (100,0%)

$p = 0,06$ (n.s.)

Um outro fator socioeconômico que merece análise é a renda familiar, pois muitas suspeitas existem entre a possível associação entre a prevalência da doença periodontal e a renda familiar.

A renda familiar foi comparada com o sangramento gengival do alistando, conforme os dados constantes na Tabela 21. Inicialmente, observa-se que existem somente 3

alistandos que têm renda familiar menor que 5 salários mínimos e ausência de sangramento gengival em todos os sextantes avaliados. Como isto gera um frequência esperada baixa, o Teste χ^2 não deve ser utilizado, sendo então substituído pelo Teste Exato de Fisher, que é mais preciso para valores muito pequenos. Com a aplicação do referido teste obteve-se um $p < 0,01$, que vem a confirmar a hipótese de existir associação estatística.

Tabela 21: Sangramento Gengival segundo a renda familiar.

Renda Familiar	Com Sangramento	Sem Sangramento	TOTAL
Até 5 Salários	75 (96,2%)	3 (3,8%)	78 (100,0%)
Além de 5 Salários	172 (82,3%)	37 (17,7%)	209 (100,0%)
TOTAL	247 (86,1%)	40 (13,9%)	287 (100,0%)

** $p < 0,01$

A Tabela 22 mostra os valores correspondentes as variáveis renda familiar e cálculo dentário. Com a aplicação do Teste χ^2 encontra-se um valor de 0,245 (0,131 com a correção de Yates), resultando em um $p > 0,05$, que vem a rejeitar a hipótese de existir associação estatística entre o cálculo dentário e a renda familiar.

Tabela 22: Cálculo Dentário segundo a renda familiar.

Renda Familiar	Com Cálculo	Sem Cálculo	TOTAL
Até 5 Salários	41 (52,6%)	37 (47,4%)	78 (100,0%)
Além de 5 Salários	103 (49,3%)	106 (50,7%)	209 (100,0%)
TOTAL	144 (50,2%)	143 (49,8%)	287 (100,0%)

$p = 0,621$ (n.s.)

Quando se analisa, na Tabela 23, a possibilidade da existência de associação estatística entre a renda familiar e a ocorrência de bolsas periodontais, voltamos a ter o já relatado problema da baixa ocorrência desta patologia periodontal. Após a aplicação do Teste χ^2 encontrou-se um valor de 0,022 (0,000 com a correção de Yates) resultando em um $p > 0,05$,

que vem a rejeitar a hipótese de haver associação. Porém esta conclusão deve ser observada com ressalvas.

Tabela 23: Bolsas Periodontais segundo a renda familiar.

Renda Familiar	Com Bolsas Period.	Sem Bolsas Period.	TOTAL
Até 5 Salários	6 (7,7%)	72 (92,3%)	78 (100,0%)
Além de 5 Salários	15 (7,2%)	194 (92,9%)	209 (100,0%)
TOTAL	21 (7,3%)	266 (92,7%)	287 (100,0%)

p = 0,881 (n.s.)

Comparou-se o hábito de fumar cigarro com o sangramento gengival do alistando, conforme os dados constantes na Tabela 24. Observa-se que quando o hábito de fumar está presente ocorre um incremento na ocorrência do sangramento gengival. Como os fumantes sem sangramento gengival geram uma frequência esperada baixa não é utilizado o teste χ^2 . Com a aplicação do Teste Exato de Fisher encontramos um valor de p igual a 0,007, sendo portanto menor que 0,01 e confirmando a hipótese de existir associação estatística.

Tabela 24: Sangramento Gengival segundo o hábito de fumar cigarro.

Hábito de Fumar	Com Sangramento	Sem Sangramento	TOTAL
Não	196 (83,4%)	39 (16,6%)	235 (100,0%)
Sim	53 (96,4%)	2 (3,6%)	55 (100,0%)
TOTAL	249 (85,9%)	41 (14,1%)	290 (100,0%)

**p<0,01

Ao analisar-se a Tabela 25 sobre a possibilidade da existência de associação estatística entre o hábito de fumar e a ocorrência de cálculo dentário, verifica-se que as proporções de alistandos com cálculo aumentam à medida que há a ocorrência do tabagismo. Após a aplicação do Teste χ^2 , encontra-se um valor de 0,000 (0,000 com a correção de Yates) resultando em um p<0,01 que vem confirmar a hipótese de haver associação.

Tabela 25: Cálculo dentário segundo o hábito de fumar cigarro.

Hábito de Fumar	Com Cálculo	Sem Cálculo	TOTAL
Não	105 (44,7%)	130 (55,3%)	235 (100,0%)
Sim	40 (72,7%)	15 (27,3%)	55 (100,0%)
TOTAL	145 (50,0%)	145 (50,0%)	290 (100,0%)

**p<0,01

Comparou-se o hábito de fumar cigarro com a ocorrência de bolsas periodontais nos alistandos, conforme os dados constantes na Tabela 26. Não foi possível observar diferenças significativas entre as proporções expressas na tabela de contingência. Com a aplicação do Teste Exato de Fisher, em virtude da baixa frequência esperada de fumantes com bolsa periodontal, encontra-se um valor de p igual a 0,09, sendo portanto maior que 0,05 e rejeitando a hipótese de existir associação estatística.

Tabela 26: Bolsas Periodontais segundo o hábito de fumar cigarro.

Hábito de Fumar	Com Bolsas Period.	Sem Bolsas Period.	TOTAL
Não	21 (8,9%)	214 (91,1%)	235 (100,0%)
Sim	1 (1,8%)	54 (98,2%)	55 (100,0%)
TOTAL	22 (7,6%)	268 (92,4%)	290 (100,0%)

P=0,09 (n.s.)

Parece-nos lógico afirmar, após nossas análises estatísticas, que os fatores socioeconômicos, podem influir na doença periodontal. Isto porque, conforme pode ser visualizado na Tabela 27, o grau de escolaridade do alistando pode ser associado com o sangramento gengival e a presença de cálculo; a escolaridade do pai pode ser associada com sangramento gengival, cálculo dentário e bolsas periodontais; a escolaridade da mãe pode ser associada com o sangramento gengival e o cálculo dentário; e a renda familiar pode ser associada com o sangramento gengival.

Tabela 27: Resumo das associações estatísticas entre os fatores socioeconômicos e o tabagismo com os dados clínicos da doença periodontal

Variável	Sangramento Gengival	Cálculo Dentário	Bolsa Periodontal
Escolaridade do Alistando	Significante (p<0,01)	Significante (p<0,05)	Não Significante (p=0,29)
Escolaridade do Pai	Significante (p<0,01)	Significante (p<0,05)	Significante (p<0,05)
Escolaridade da Mãe	Significante (p<0,01)	Significante (p<0,01)	Não Significante (p=0,06)
Renda Familiar	Significante (p<0,01)	Não Significante (p=0,62)	Não Significante (p=0,88)
Hábito de Fumar	Significante (p<0,01)	Significante (p<0,01)	Não Significante (p=0,09)

Pode-se afirmar também, que tanto o hábito de fumar quanto a escolaridade parecem ser mais determinantes no processo saúde-doença periodontal do que a renda familiar. Isto porque o hábito de fumar e a escolaridade mostraram-se sempre associadas ao sangramento gengival e ao cálculo, enquanto que a renda familiar só mostrou-se associada com o sangramento gengival.

5.6 Análise pelo Processo KDD

A descoberta de conhecimento em banco de dados, também chamado processo KDD, é aplicada para extrair informações adicionais dos dados. Nesta pesquisa, aplicou-se a técnica de geração de regras de classificação. Esta técnica se baseia em associações entre as variáveis do estudo e o atributo alvo, de forma que as associações mais freqüentes formulem regras de classificação que satisfaçam um suporte mínimo pré-determinado.

Neste estudo, adotou-se 10% como suporte mínimo, desta forma, muitas regras que não trariam conhecimento adicional foram excluídas da análise.

Através da análise pelo processo KDD obteve-se 3 conjuntos de regras de classificação organizadas de acordo com os atributos alvos: sangramento gengival, cálculo dentário e bolsas periodontais.

Para o atributo alvo sangramento gengival, o conjunto de regras esta representado na Tabela 28. O sangramento gengival apresenta duas possíveis classificações: presença de sangramento gengival (ComSG) ou ausência (SemSG). Ressalta-se que a ocorrência de sangramento gengival é de 85,9% na base de dados e sua ausência de 14,1%. As 8 regras geradas para a presença de sangramento gengival apresentam um grau de confiança entre 96,875% e 100,000%, superiores portanto aos 85,9%. Quando se verifica a ausência de sangramento gengival observa-se que as duas regras geradas têm confiança de 38,889% e 40,000%, sendo portanto maiores que os 14,1% da base de dados.

Tabela 28: Regras de classificação para o atributo alvo Sangramento Gengival no alistando

Seqüência	Sangramento Gengival	Regra	Suporte (%)	Confiança (%)
1	ComSG	Se o pai tem menos de 8 anos de estudo e o alistando tem cálculo, então há sangramento gengival.	27,241	100,000
2	ComSG	Se o alistando é fumante e tem cálculo, então há sangramento gengival.	13,793	100,000
3	ComSG	Se o alistando é fumante e o pai tem menos de 8 anos de estudo, então há sangramento gengival.	12,759	100,000
4	ComSG	Se o alistando é fumante e a mãe tem menos de 8 anos de estudo, então há sangramento gengival.	12,759	100,000
5	ComSG	Se o alistando tem menos de 8 anos de estudo e tem cálculo, então há sangramento gengival.	25,517	98,649
6	ComSG	Se o alistando tem cálculo e sua mãe menos de 8 anos de estudo, então há sangramento gengival.	33,448	97,938
7	ComSG	Se o alistando não fuma, mas tem cálculo então há sangramento gengival.	33,448	97,938
8	ComSG	Se a renda familiar for abaixo de 5 salários mínimos e o pai apresentar menos de 8 anos de estudo, então há sangramento gengival.	22,069	96,875
9	SemSG	Se o alistando não fuma, não tem cálculo e sua escolaridade e de seu pai forem superiores a 8 anos, então não há sangramento gengival.	18,966	40,000
10	SemSG	Se o alistando não tiver cálculo, a renda familiar for maior que 5 salários mínimos e a sua escolaridade e a de seu pai forem superiores a 8 anos, então não há sangramento gengival.	18,621	38,889

A Tabela 29 apresenta as 13 regras de classificação para o atributo alvo Cálculo Dentário. O cálculo dentário apresenta duas possíveis classificações: presença do mesmo (ComCalculo) ou sua ausência (SemCalculo). A presença de cálculo na base de dados é de 50,3% e sua ausência de 49,7%. As 7 regras geradas para a presença de cálculo dentário apresentam um grau de confiança entre 70,968% e 94,444%, superiores portanto aos 50,3%. Já para a ausência de cálculo dentário observa-se que as 6 regras geradas tem confiança variando entre 64,103% e 94,118%, sendo portanto maiores que os 49,7% da base de dados.

Tabela 29: Regras de classificação para o atributo alvo Cálculo Dentário no alistando

Seqüência	Cálculo Dentário	Regra	Suporte (%)	Confiança (%)
1	ComCalculo	Se o alistando é fumante, tem renda familiar superior a 5 salários mínimos e a escolaridade da mãe é inferior a 8 anos, então há Cálculo.	6,207	94,444
2	ComCalculo	Se o alistando é fumante e tem juntamente com sua mãe escolaridade inferior a 8 anos, então há cálculo.	11,379	84,848
3	ComCalculo	Se o alistando é fumante, apresenta sangramento gengival e tem renda familiar acima de 5 salários mínimos, então há cálculo.	10,000	82,759
4	ComCalculo	Se o alistando tem sangramento gengival, renda familiar superior a 5 salários mínimos, seu pai tem mais de 8 anos de estudo e sua mãe menos de 8 anos de estudo, então há cálculo.	6,897	75,000
5	ComCalculo	Se o alistando tem sangramento gengival, não apresenta bolsas periodontais, tem renda familiar abaixo de 5 salários mínimos, e seu pai e sua mãe têm menos de 8 anos de estudo, então há cálculo.	17,241	72,000
6	ComCalculo	Se o alistando tem sangramento gengival, renda familiar abaixo de 5 salários mínimos e juntamente com seu pai e sua mãe possui menos de 8 anos de estudo, então há cálculo.	15,862	71,739
7	ComCalculo	Se o alistando tem sangramento gengival, sua mãe menos de 8 anos de estudo e seu pai mais, então há cálculo.	10,690	70,968
8	SemCalculo	Se o alistando não tem sangramento gengival e não fuma, então não há cálculo.	11,724	94,118
9	SemCalculo	Se o alistando não tem sangramento gengival e sua mãe tem mais de 8 anos de estudo, então não há cálculo.	8,966	92,308
10	SemCalculo	Se o alistando não tiver sangramento gengival, então não há cálculo.	13,793	90,000

Segue

11	SemCalculo	Se o alistando não tem bolsas periodontais, não fuma e seu pai e sua mãe tem escolaridade superior a 8 anos, então não há cálculo.	26,552	68,831
12	SemCalculo	Se o alistando não tem bolsas periodontais e ele, juntamente com seu pai e sua mãe têm mais de 8 anos de estudo, então não há cálculo.	28,966	65,476
13	SemCalculo	Se o alistando não fuma, tem renda familiar acima de 5 salários mínimos e seu pai e sua mãe tem escolaridade inferior a 8 anos de estudo, então não há cálculo.	13,448	64,103

Na Tabela 30 pode-se observar as 12 regras de classificação para o atributo alvo Bolsas Periodontais. Este atributo apresenta duas possíveis classificações: presença de bolsas periodontais (ComBolsa) ou a ausência das mesmas (SemBolsa). É muito baixa a presença de bolsas periodontais na base de dados (7,6%) sendo que a maioria dos alistandos (92,4%) não apresenta este problema de saúde periodontal. Gerou-se 5 regras de classificação para a presença de bolsas periodontais com variações na confiança de 27,931% a 40,000%, superiores portanto aos 7,6%. Na análise da ausência de bolsas periodontais observa-se que as 7 regras geradas têm confiança superior aos 92,4% da base de dados, sendo que ocorrem variações entre 97,674% e 100,000%.

As regras referentes a ausência de Bolsas Periodontais apresentam-se com um maior nível de confiança do que as relativas a presença de Bolsas. Isto ocorre devido a maior ocorrência do atributo alvo SemBolsa no banco de dados, fato que justifica a elaboração de regras mais precisas e portanto mais importantes para a extração de conhecimento.

A análise pelo processo KDD gera regras de classificação utilizando as variáveis socioeconômicas e o hábito de fumar cigarro. Estes mesmos achados são encontrados na análise estatística. Assim, observa-se que, o algoritmo C4.5, utilizado na análise KDD, é eficiente na elaboração das regras de classificação e sua utilização em estudos epidemiológicos é válida.

Tabela 30: Regras de classificação para o atributo alvo Bolsas Periodontais no alistando

Seqüência	Cálculo Dentário	Regra	Suporte (%)	Confiança (%)
1	ComBolsa	Se o alistando apresenta sangramento gengival, não fuma, tem renda familiar superior a 5 salários mínimos e seu pai e sua mãe tem menos de 8 anos de estudo, então há bolsa periodontal.	5,172	40,000
2	ComBolsa	Se o alistando tem sangramento gengival, não fuma, não apresenta cálculo e tem junto com seu pai menos de 8 anos de estudo, então há bolsa periodontal.	6,897	25,000
3	ComBolsa	Se o alistando apresenta sangramento gengival, não fuma e tem juntamente com seu pai menos de 8 anos de estudo, então há bolsa periodontal.	14,828	20,930
4	ComBolsa	Se o alistando tem sangramento gengival, não tem cálculo, não fuma e sua mãe tem menos de 8 anos de estudo, então há bolsa periodontal.	15,517	20,000
5	ComBolsa	Se o alistando tem sangramento gengival, e ele juntamente de seu pai tem menos de 8 anos de estudo, então há bolsa periodontal.	27,931	13,580
6	SemBolsa	Se o alistando não tem cálculo e sua mãe tem mais de 8 anos de estudo, então não há bolsa periodontal.	24,828	100,000
7	SemBolsa	Se o alistando não tem cálculo, a renda familiar é superior a 5 salários mínimos e seu pai tem mais de 8 anos de estudo, então não há bolsa periodontal.	21,034	100,000
8	SemBolsa	Se o alistando apresenta sangramento gengival, não tem cálculo e seu pai tem mais de 8 anos de estudo, então não há bolsa periodontal.	16,552	100,000
9	SemBolsa	Se o alistando não apresenta sangramento gengival e tem renda familiar superior a 5 salários mínimos, então não há bolsa periodontal.	11,034	100,000
10	SemBolsa	Se o alistando é fumante, então não há bolsa periodontal.	18,966	98,182
11	SemBolsa	Se o alistando e seu pai têm mais de 8 anos de estudo e a renda familiar for superior a 5 salários mínimos, então não há bolsa periodontal.	31,724	97,826
12	SemBolsa	Se o alistando, seu pai e sua mãe tiverem mais de 8 anos de estudo, então não há bolsa periodontal.	29,655	97,674

5.7 Conclusão

A pesquisa evidenciou que a doença periodontal severa (nos estágios mais avançados), em jovens de 18 anos de idade, não é um problema de saúde pública na população estudada. Pode-se afirmar isto com base nos resultados obtidos, uma vez que a ocorrência de bolsas periodontais (7,6%) é baixa.

A prevalência do sangramento gengival foi de 85,9% e do cálculo dentário 50,3%, sendo estas as condições mais freqüentemente encontradas. As proporções dos componentes do CPI foram: CPI 0 (12,8%); CPI 1 (34,4%); CPI 2 (45,2%); CPI 3 (7,3%) e CPI 4 (0,3%) sendo semelhantes às do Levantamento Epidemiológico Nacional de 1986 para a Região Sul.

O tabagismo e as variáveis socioeconômicas têm importante papel no desenvolvimento da doença periodontal, havendo significativa associação estatística entre elas. A escolaridade foi a variável socioeconômica de maior grau de significância .

As regras de classificação geradas pela técnica de mineração de dados, mostram claramente que o perfil das pessoas que apresentam problemas gengivais, presença de cálculo e bolsas periodontais, apresentam também piores indicadores socioeconômicos associados.

Existe portanto uma forte relação entre as variáveis socioeconômicas e o tabagismo com a doença periodontal, sendo que estas variáveis apresentam-se como agravantes à doença.

6. CONCLUSÕES E RECOMENDAÇÕES

Este trabalho teve como objetivo a aplicação de técnicas estatísticas e de *Data Mining* para a descoberta de conhecimento em um banco de dados dos alistados do Exército em Florianópolis no ano de 1999.

Em virtude da carência de dados sobre a forma de distribuição das doenças periodontais na população, especialmente em adultos jovens, as ações coletivas de promoção de saúde bucal são inexistentes. Desta forma, esta doença não vem recebendo a atenção necessária no âmbito preventivo, haja vista que suas seqüelas podem levar, no decorrer dos anos, a perda dos dentes caracterizando um problema de saúde pública.

6.1 Conclusões

É bastante provável que as prevalências encontradas neste estudo possam ser inferidas na população, desta faixa etária, do município de Florianópolis. Isto porque segundo dados populacionais do IBGE (Instituto Brasileiro de Geografia e Estatística) grande parte desta população masculina está contida no universo deste estudo, uma vez que o alistamento no serviço militar é obrigatório a todos.

A descoberta de conhecimento em banco de dados pode ser um instrumento útil para o gerenciamento de ações em saúde, uma vez que, através dessas análises, pode-se conhecer melhor os grupos populacionais de risco e focalizar neles as ações preventivas mais necessárias.

O índice CPI, recomendado pela Organização Mundial da Saúde para aferição das enfermidades do Periodonto, aplicado de maneira isolada, esconde a ocorrência do sangramento gengival e/ou do cálculo em detrimento de uma condição periodontal mais grave,

fazendo-se necessário os registros destas condições em separado. Além disso, por considerar doentes os sextantes com cálculo sem sangramento e as bolsas inativas (sem sangramento), eleva-se a prevalência da doença para 87,2%. Portanto, sugere-se que ao utilizar-se o índice CPI, sejam registrados em separado as proporções de sangramento gengival e cálculo dentário.

Em virtude das associações estatísticas e das regras de classificação geradas entre o nível socioeconômico e a doença periodontal, pode-se afirmar que, com a melhoria da qualidade de vida da população obtêm-se melhores indicadores de saúde. Portanto, a luta pelas melhorias na qualidade de vida deve ser o objetivo principal de todo aquele que se envolve acerca das questões da Odontologia em Saúde Coletiva.

6.2 Recomendações

A falta de estudos semelhantes nesta população dificulta comparações, e mostra a necessidade dos mesmos, para orientar com mais eficácia a coordenação de serviços de saúde bucal no município, através dos recursos que a epidemiologia pode vir a oferecer.

Trabalhos futuros em banco de dados sobre grupos populacionais são importantes para que o conhecimento sobre as doenças ultrapasse as fronteiras da Biologia e alcancem as esferas socioeconômicas e comportamentais, que também estão presentes no processo saúde-doença.

Outros tipos de análise, além das regras de classificação, podem e devem ser empregadas para que este processo seja melhor compreendido. A Análise de Agrupamentos (*Cluster*), por exemplo, pode ser empregada para que se formem agrupamentos por similaridade interna e assim se estabeleçam os grupos mais homogêneos para receber uma determinada atenção em saúde.

REFERÊNCIAS BIBLIOGRÁFICAS

Capítulo 2 – O Periodonto e sua Epidemiologia

- [AHL 96] AHLBERG, J., TUONIMEN, R., MURTOMAA, H. Periodontal status among male industrial works in Southern Finland with or without access to subsidized dental care. *Acta Odontol. Scand.*, n.54, p.166-170, 1996.
- [ALB 96] ALBANDAR, J. M. et al. Gingival state and dental calculus in early-onset periodontitis. *J. Periodontol.*, n.10, p.953-958, Oct., 1996.
- [AMEa96] AMERICAN ACADEMY OF PERIODONTOLOGY, Position paper: epidemiology of periodontal diseases. *J. Periodontol.*, v.67, n.9, p.935-945, Sep., 1996.
- [AMEb96] AMERICAN ACADEMY OF PERIODONTOLOGY, Tobacco use and the periodontal patient. *J. Periodontol.*, n.1, p.51-56, Jan., 1996.
- [ANK 97] ANKKURINIEMI, O., AINAMO, J. Dental health and dental treatment needs among recruits of the Finnish Defence Forces, 1919-91. *Acta Odontol. Scand.*, n.55, p.192-197, 1997.
- [AUI 83] AUIL ATALA, E. La enfermedad periodontal afecta al 80 por ciento de jóvenes. *Esencia Odontol.*, n.13, 1983.
- [AXE98] AXELSSON, P. et al. Relationship between smoking and dental status in 35, 50, 65 and 75 year-old individuals. *J. Clin. Periodontol.*, n.25, p.297-305, Apr., 1998.
- [BND 93] BERNDSEN, M. et al. Compliance perceived by dutch periodontists and higienists. *J. Clin. Periodontol.*, v.20, n.9, p.668-672, Oct., 1993.
- [BIJ 95] BIJELLA, V. T. et al. Prevalência de cárie, doença periodontal e má-oclusão em crianças de zero a seis anos no município de Bauru, Estado de São Paulo. Determinação social do processo saúde doença. *Cecade News*, v.3, n.1, p.1-11, jan./abr., 1995.
- [BOU 91] BOURGEOIS, D. et al. Evaluation of the utilization of complex periodontal care in na industrialized country with a national health insurance. *Int. Dent. J.*, v.41, n.1, p.49-54, Jan., 1991.
- [MIN 88] MINISTÉRIO DA SAÚDE DO BRASIL. Secretaria Nacional de Programas Especiais de Saúde. Divisão Nacional de Saúde Bucal. Fundação de Serviços de Saúde Pública. **Levantamento Epidemiológico em Saúde Bucal: Brasil, zona urbana**, 1986. Brasília: Centro de Documentação do Ministério da Saúde, 1988. 137p.

- [BRI 96] BRIDGES, R. B. et al. Periodontal status of diabetic and non-diabetic men: effects of smoking, glycemic control, and socioeconomic factors. **J. Periodontol.**, v.67, n.11, p.1185-1192, Nov., 1996.
- [CAM 92] CAMPOS, A. Jr. Análise da influência de variáveis explanatórias conjunturais e individuais na situação de saúde periodontal medida pelo CPITN. **Periodontia**, v.1, n.1, p.28-36, abr./set., 1992.
- [CAR 97] CARRANZA, F. A., NEWMAN, M. G. **Periodontia Clínica**. 8.ed. Rio de Janeiro: Guanabara Koogan, 1997.
- [CHA 89] CHAMBRONE, L. A. Prevalência da doença periodontal no Brasil. Um alerta aos dentistas. **Rev. Inst. Ciênc. Saúde**, v.7, n.2, p.5-8, jul./dez., 1989.
- [DIN 94] DINI, E. L. et al. Periodontal conditions and treatment needs (CPITN) in a worker population in Araraquara, SP, Brazil. **Int. Dent. J.** v.44, n.4, p.309-311, Aug., 1994.
- [DNS 97] DINSDALE, C.R. et al. Subgingival temperature in smokers and non-smokers with periodontal disease. **J. Clin. Periodontol.**, v.24, n.10, p.761-766, Oct., 1997.
- [EGE 95] EGELBERG, J., BADERSTEIN, A. **Exame periodontal**. 1.ed. São Paulo: Livraria Santos, 1995.
- [GEL 99] GELSKEY, S. C. Cigarette smoking and periodontitis: methodology to assess the strength of evidence in support of a causal association. **Community Dent. Oral Epidemiol.**, n.27, p.16-24, 1999.
- [GEN 96] GENCO, R. J., COHEN, D. W., GOLDMAN, H. M. **Periodontia contemporânea**. 1.ed. São Paulo: Livraria Santos, 1996.
- [GIL 94] GILBERT, L. Social factors and self-assessed oral health in South Africa. **Community Dent. Oral Epidemiol.** v.22, n.1, p.47-51, Feb., 1994.
- [GRO 94] GROSSI S.G et al. Assessment of risk for periodontal disease. I. Risk indicators for attachment loss. **J. Periodontol.**, v.65, n.3, p.260-267, Mar., 1994.
- [GRO 95] GROSSI S.G. et al. Assessment of risk for periodontal disease. II. Risk indicators for alveolar bone loss. **J. Periodontol.**, n.66, p.23-29, 1995.
- [GUG 98] GUGUSHE, T. S. The influence of socio-economic variables on the prevalence of periodontal disease in South Africa. **SADJ**, v.53, n.2, p.41-46, Feb., 1998.
- [GUI 90] GUIMARÃES, M. M. **Estudo sócio epidemiológico da mortalidade dental**, Belo Horizonte, Minas Gerais, Brasil. Belo Horizonte. Tese de mestrado, Universidade Federal de Minas Gerais, 86p., 1990.

- [HAFa92] HAFFAJEE, A.D., SOCRANSKY, S.S., GOODSON J.M. Subgingival temperature (I). Relation to baseline clinical parameters. **J. Clin. Periodontol.**, n.19, p.401-408, 1992.
- [HAFb92] HAFFAJEE, A.D., SOCRANSKY, S.S., GOODSON J.M. Subgingival temperature (II). Relation to future periodontal attachment loss. **J. Clin. Periodontol.**, n.19, p.409-416, 1992.
- [HAFc92] HAFFAJEE, A.D., SOCRANSKY, S.S., GOODSON J.M. Subgingival temperature (III). Relation to microbial counts. **J. Clin. Periodontol.**, n.19, p.417-422, 1992.
- [HEL 96] HELDERMAN , v. P., KALAM JOARDER, M. A., AZIZA BEGUM, M. A. Prevalence and severity of periodontal diseases and dental caries in Bangladesh. **Int. Dent. J.**, n.46, p.76-81, 1996.
- [IBG 96] IBGE, Fundação Instituto Brasileiro de Geografia e Estatística. **Contagem da População de 1996 v.2 resultados relativos à Grande Florianópolis**
- [IBG 98] IBGE, Fundação Instituto Brasileiro de Geografia e Estatística. **Pesquisa Nacional por Amostra de Domicílios – PNAD.**
<http://www.ibge.gov.br/presidencia/noticias/scripts/novopress.asp?cod=pnad98>
- [KAL 97] KALLIO, P. et al. Determinants of self assessed gingival health among adolescents. **Acta Odontol. Scand.**, v.55, n.4, p.106-110, Apr., 1997.
- [KRI 99] KRIGER, L. et al. **ABOPREV: Promoção de saúde bucal.** 2.ed. São Paulo: Artes Médicas, 1999.
- [LAN 94] LANG, W. P. et al. The relation of preventive dental behaviors to periodontal health status. **J. Clin. Periodontol.**, v.21, n.3, p.194-198, Mar., 1994.
- [LAS 89] LASCALA, N. T., MOUSSALI, N. H. **Periodontia clínica II.** São Paulo: Artes Médicas, 1989.
- [LIN 88] LINDHE, J. **Tratado de periodontologia clínica.** Rio de Janeiro: Editora Guanabara, 1988.
- [LOP 96] LÓPES, N. J. et al. Periodontal conditions in 15-19 years-old Chileans. **Int. Dent. J.**, v.46, n.3, p.161-164, 1996.
- [MAC 99] MACHUCA, G. et al. The influence of general health and socio-cultural variables on the periodontal condition of pregnant women. **J. Periodontol.**, v.70, n.7, p.779-785, Jul., 1999.
- [MED 89] MEDEIROS, U. V. **Redução da prevalência de doença periodontal através de um programa de higienização supervisionada em saúde pública.** São Paulo. Tese de doutorado. Universidade de São Paulo, 157p., 1989.

- [MIC 96] MICHEELIS, W., BAUCH, J. Oral health of representative samples of germans examined in 1989 and 1992. **Community Dent. Oral Epidemiol.**, v.24, n.1, p.62-67, Feb., 1996.
- [MON 98] MONTEIRO DA SILVA, A. M. Psychosocial factors, dental plaque levels and smoking in periodontitis patients. **J. Clin. Periodontol.**, n.25, p.517-523, Jun., 1998
- [MUL 94] MULLALLY ,B. H., LINDEN, G. J. The periodontal status of irregular dental attenders. **J. Clin. Periodontol.**, v.21, n.8, p.544-548, Sep.,1994
- [MUN 95] MUNGHAMBA, E. G. et al. Risk factors for periodontal diseases in Ilala, Tanzania. **J. Clin. Periodontol.**, v.22, n.6, p.347-354, May, 1995.
- [MUN 96] MUNGHAMBA, E. G. S., MARKKANEN, H. A., KONKALA, E. Periodontal status and treatment needs in a rural area of Ukonga, Tanzania. **Int. Dent. J.**, n.46, p.156-160, 1996.
- [NOR 98] NORDERYD O. et al. Risk of severe periodontal disease in a swedish adult population. A cross-sectional study. **J. Clin. Periodontol.**, n.28, p.1022-1028, Dec., 1998.
- [NOV 99] NOVAES, A. B. Jr., et al. Supportive periodontal therapy in South America. A retrospective multi-practice study on compliance. **J. Clin. Periodontol.**, n.70, p.301-306, Mar., 1999.
- [OGU 96] OGUNBODEDE, E. O. et al. Socio-economic factors and dental health in na obstetric population. **West Afr J Med.**, v.15, n.3, p.158-162, Jul., 1996.
- [ONG 98] ONG, G. Periodontal diseases and tooth loss. **Int. Dent. J.**, v.48, n.3, suppl.1, p.233-238, Jun., 1998.
- [OMS 97] ORGANIZACIÓN MUNDIAL DE LA SALUD. **Encuestas de salud bucodental: Métodos básicos.** 4.ed. Genebra: OMS, 1997.
- [PAG 97] PAGE R. C. et al. Risk assessment for periodontal diseases. **Int. Dent. J.**, n.47, p.61-87, Apr., 1997.
- [PER 98] PERSON, R. E. Periodontal effects of a biobehavioral prevention program. **J. Clin. Periodontol.** n.25, p.322-329, 1998.
- [PET 90] PETERSEN, P. E. Social Inequalities in dental health. Towards a theoretical explanation. **Community Dent. Oral Epidemiol.**, v.18, n.3, p.153-158, Jun., 1990.
- [PID 47] PINDBORG, J.J. Tobacco and gingivitis. I. Statistical examiation of the significance of tobacco in the development of ulceromembranous gingivitis and in the formation of calculus. **J. Dent. Res.**, n.26, p.261-264, 1947.

- [PRE 86] PREBER, H., BERGSTROM, J. The effect of non-surgical treatment on periodontal pockets in smokers and non-smokers. **J. Clin. Periodontol.**, n.13, p319-323, 1986.
- [RED 92] REDDY, J. The WHO oral health goals for the year 2000 in South Africa. **Int. Dent. J.**, v.42, n.3, p.150-156, Jun., 1992.
- [REG 91] REGEZI, J. A., SCIUBBA, J. J. **Patologia Bucal: correlações clinicopatológicas.** Rio de Janeiro: Guanabara Koogan, 1991.
- [REN 98] RENVERT, S. et al. The clinical and microbiological effects of non-cirurgical periodontal therapy in smokers and non-smokers. **J. Clin. Periodontol.**, v.24, n.3, p.153-157, Feb., 1998.
- [ROB 90] ROBERTSON, P. B. et al. Periodontal effects associated with the use of smokeless tobacco. **J. Periodontal.** N.61, p.438-443, 1990.
- [SLA 96] SLADE, G. D. et al. Variations in the social impact of oral conditions among older adults in South Australia, Ontario, and North Carolina. **J. Dent. Res.** v.75, n.7, p.1439-1450, Jul., 1996.
- [STE 96] STEELE, J. G. et al. Major clinical findings from a dental survey of elderly people in three different English communities. **Br. Dent. J.**, v.180, n.1, p.17-23, Jan., 1996.
- [TAA 96] TAANI, D. S. Dental health of 13-14-years-old Jordanian school children and its relationship with socio-economic status. **Int. J. Paediatr. Dent.** v.6, n.3, p.183-186, Sept., 1996.
- [VIE 91] VIEIRA, L. C. C. **Análise da influência de variáveis diversas na situação de saúde periodontal da população de Bauru, São Paulo.** Bauru, Tese de Doutorado, Universidade de São Paulo, 116p., 1991.

Capítulo 3 – A Análise de Banco de Dados

- [AGR 96] AGRAWAL, Rakesh. et al. **The Quest Data Mining System.** San Jose, California: IBM Almaden Research Center, 1996.
Disponível em: http://cora.jprc.com/Artificial_Intelligence/Data_Mining/index.html
- [ALM 99] ALMENDA, Camilo C. TEIXEIRA, Ivan R., PRUDÊNCIO, Ricardo B. C. Descoberta de Conhecimento em Banco de Dados com Redes Neurais. **In: XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 1999, Rio de Janeiro. Anais...** Rio de Janeiro: PUC-RJ, 1999. p. 555-563.

- [AVI 98] ÁVILA, Bráulio Coelho Data Mining. In: **VI ESCOLA DE INFORMÁTICA DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO**, 1998, Blumenau. *Anais...* Curitiba: PUC-PR, 1998. p. 87-106.
- [BAR 99] BARBETTA, Pedro A. **Estatística Aplicada às Ciências Sociais**. Florianópolis: UFSC, 1994.
- [BER 96] BERNDT, Donald J., CLIFFORD, James. *Finding Patterns in Time Series: A Dynamic Programming Approach*. In: **Advances in Knowledge Discovery and Data Mining**. Califórnia: American Association for Artificial Intelligence, 1996. 611 p.
- [BER 97] BERRY, Michael J. A., LINOFF, Gordon. **Data Mining techniques**: for marketing, sales and customer support. New York : John Wiley, 1997. 454 p.
- [BOJ 99] BOJARCZUK, Célia C., LOPES, Heitor S., FREITAS, Alex A. Programação Genética e Data Mining para Descoberta de Conhecimento em Medicina. In: **XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO**, 1999, Rio de Janeiro. *Anais...* Rio de Janeiro: PUC-RJ, 1999. p. 31-42.
- [CHE 96] CHEN, Ming-Syan, HAN, Jiawei, YU, Philip S. **Data Mining: An Overview from Database Perspective**, 1996.
Disponível em: http://cora.jpcc.com/Artificial_Intelligence/Data_Mining/index.html
- [DAT 00] DATA MINING **Data Mining**. Disponível em <http://www.datamining.com>
- [DAW 93] DAWSON-SAUNDERS, Beth, TRAPP, Robert G. **Biostatistics** Basic & clinical. Norwalk: Appleton & Lange; 1993.
- [DWB 01] DWBRASIL **Data Mining**. Disponível em: <http://www.dwbrasil.com.br/html/dmining.html>
- [EBE 98] EBECKEN, Nelson F. F. **Data Mining**. Boston: WIT Press/Computacional Mechanics Publications, 1998. 449 p.
- [EDE 96] EDELSTEIN, Herb. **Data Mining: Exploiting the Hidden Trends in Your Data**. DB2 On Line Magazine, 1996.
- [EXC 99] EXCLUSIVE ORE. **The DBMS Guide to Data Mining Solutions** written by Exclusive. Disponível em: <http://www.xore.com/>
- [FAY 96] FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory, SMITH, Padhraic. **From Data Mining to Knowledge Discovery: An Overview**. In: *Advances in Knowledge Discovery and Data Mining*. Califórnia: American Association for Artificial Intelligence, 1996. 611 p.
- [FEL 96] FELDENS, Miguel A. **Descoberta de Conhecimento Aplicada à Detecção de Anomalias em Base de Dados**. Porto Alegre: PPGCC da UFRGS, 1996. (Trabalho Individual).

- [FEL 99] FELDENS, Miguel A. et al. **Metodologias para Implementação da Inteligência do Negócio: desenvolvimento de sistema de informação para Database Marketing**, 1999.
- [GOT 98] GOTTGROY, Márcia P. B., RODRIGUES, Márcia J. N., SOUSA, Marcel T. G. *Data Mining Agents*. In: **Data Mining**. Boston: WIT Press/Computacional Mechanics Publications, 1998. 449 p.
- [HAL 98] HALMENSCHLAGER, Carine. **Aplicação de Data Mining na Descoberta de Conhecimento em Bases de Dados**. Santa Cruz do Sul: UNISC, 1998. 144 p. (Monografia, Graduação em Ciência da Computação)
- [HAN 92] HAN, Jiawei. et al. **Knowledge Discovery in Databases: An Attribute-Oriented Approach**, 1992. http://cora.jpcc.com/Artificial_Intelligence/Data_Mining/index.html
- [HOL 95] HOLSHEIMER, Marcel. et al. **A Perspective on Databases and Data Mining**, 1995. Disponível em: http://cora.jpcc.com/Artificial_Intelligence/Data_Mining/index.html
- [KEN 98] KENNEDY, Ruby L. et al. **Solving Data Mining problems through pattern recognition**. Upper Saddle River: Prentice Hall, 1995.
- [MOR 93] MORIK, Katharina et al. **Knowledge Acquisition and Machine Learning: theory, methods and applications**. London: Academic Press Limited, 1993. 305 p.
- [PIN 99] PINHEIRO, Carlos A. R. **Data Mining: Obtendo Vantagens com seu Data Warehouse**. *Developers Magazine*, Rio de Janeiro, Ano 3, n. 35, p. 38-40, jul. 1999.
- [PRA 97] PRADO, Hércules A. **Conceitos de Descoberta de Conhecimento em Banco de Dados**. Porto Alegre: PPGCC da UFRGS, 1997. (Trabalho Individual).
- [QUB 01] THE QUEEN'S UNIVERSITY OF BELFAST **Data Mining**. Disponível em: http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_4.html
- [QUI 93] QUINLAN, J. Ross **C4.5 Programs for machine learning**. San Mateo: Morgan Kaufmann, 1993. 302p.
- [SOA 91] SOARES, José., FARIAS, Alfredo A., CÉSAR, Cibele C. **Introdução à Estatística**. Rio de Janeiro: LTC, 1991.
- [WIT 00] WITTEN, Ian H., FRANK, Eibe. **Data Mining** Practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann, 2000. 371p.
- [WU 98] WU, Xindong, KOTAGIRI, Ramamohanarao, KORB, Kevin B. **Research and Development in Knowledge Discovery and Data Mining**. Second Pacific-Asia Conference, PAKDD-98 Proceedings... Berlin: Springer, 1998. 421p.

APÊNDICES

APÊNDICE 1 – Questionário Aplicado nos Alistados para a Composição do Banco de Dados

APÊNDICE 2 – Arquivo Sangramento.names (CBA Data Mining II)

APÊNDICE 3 – Arquivo Sangramento.data (CBA Data Mining II)

APÊNDICE 4 – Arquivo Calculo.names (CBA Data Mining II)

APÊNDICE 5 – Arquivo Calculo.data (CBA Data Mining II)

APÊNDICE 6 – Arquivo Bolsa.names (CBA Data Mining II)

APÊNDICE 7 – Arquivo Bolsa.data (CBA Data Mining II)

APÊNDICE 1 - Questionário Aplicado para a Composição do Banco de Dados

Número:

Questionário

ATENÇÃO: Este questionário não possui respostas certas ou erradas, os dados são de caráter sigiloso e seu nome não será associado com qualquer uma das respostas dadas.

1) Seu pai sabe ler e escrever?

- 1 Sim
- 2 Não
- 3 Só assina
- 9 Não sei informar

EDPAI _____

2) Sua mãe sabe ler e escrever?

- 1 Sim
- 2 Não
- 3 Só assina
- 9 Não sei informar

EDMÃE _____

3) Até que série estudou na escola?

Você: _____
Pai: _____
Mãe: _____

SÉRIEV _____

SÉRIEP _____

SÉRIEM _____

4) Quanto ganharam, em Reais, no último mês, as pessoas da casa em que você mora (sua família) que estão trabalhando ou aposentadas?

Pessoa 1: _____ Pessoa 4: _____
Pessoa 2: _____ Pessoa 5: _____
Pessoa 3: _____ Pessoa 6: _____

RENDAF _____

5) Você fuma cigarro?

- 1 Sim, sou fumante
- 2 Não, mas já fui fumante. Parei no ano de 19 _____.
- 3 Não fumo e nunca fui fumante.

FUMO _____

As questões 6 e 7 só devem ser respondidas por quem fuma ou já foi fumante!

6) Com que frequência você fuma ou fumava?

Número de cigarros fumados por dia _____
ou
Número de cigarros fumados por semana _____

- 8 Não se aplica

FUMOF _____

7) Durante quanto tempo você fumou, ou a quanto tempo fuma?

Número de anos fumados _____
ou
Número de meses fumados _____

- 8 Não se aplica

FUMOT _____

APÊNDICE 5 – Arquivo Calculo.data

| Universidade Federal de Santa Catarina – UFSC
| Centro Tecnológico - CTC
| Programa de Pós Graduação em Engenharia de Produção – PPGEP
| Arquivo Calculo.data CBA Data Mining II

| NSEXTSG, CPI, BOLSAS, SEXTSAD, SEXTALT, EDALIST, EDPAI, EDMAE, RENDAFAM, FUMO, SG, CALCULO

Nenhum, Dois, SemBolsa, 3ouMais, Um, Mais8anos, Mais8anos, Mais8anos, Maior5SM, Ex Fumante, SemSG, ComCalculo
Um, Um, SemBolsa, 3ouMais, Um, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemCalculo
Nenhum, Zero, SemBolsa, 3ouMais, Nenhum, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, SemSG, SemCalculo
Dois, Dois, SemBolsa, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, NaoInform, NaoFuma, ComSG, ComCalculo
Nenhum, 3ou4, ComBolsa, 3ouMais, Um, Mais8anos, Mais8anos, Ate8anos, NaoInform, NaoFuma, SemSG, SemCalculo
Nenhum, Zero, SemBolsa, 3ouMais, Nenhum, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, SemSG, SemCalculo
3ouMais, 3ou4, ComBolsa, Dois, 3ouMais, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
3ouMais, Um, SemBolsa, 3ouMais, 3ouMais, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemCalculo
Dois, Dois, SemBolsa, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
Nenhum, Zero, SemBolsa, 3ouMais, Nenhum, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, SemSG, SemCalculo
Um, Dois, SemBolsa, 3ouMais, Um, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
Dois, Dois, SemBolsa, 3ouMais, Dois, Mais8anos, Ate8anos, Ate8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
Dois, Um, SemBolsa, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemCalculo
Dois, Dois, SemBolsa, 3ouMais, Dois, Mais8anos, NaoInform, Ate8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
3ouMais, Um, SemBolsa, Dois, 3ouMais, Mais8anos, Mais8anos, Mais8anos, Ate5SM, NaoFuma, ComSG, SemCalculo
3ouMais, Dois, SemBolsa, 3ouMais, 3ouMais, Mais8anos, Ate8anos, Ate8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
3ouMais, Dois, SemBolsa, 3ouMais, 3ouMais, Mais8anos, Mais8anos, Ate8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
Dois, Dois, SemBolsa, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, ComCalculo
Dois, Dois, SemBolsa, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, ComCalculo

APÊNDICE 6 – Arquivo Bolsa.names

Universidade Federal de Santa Catarina – UFSC
Centro Tecnológico - CTC
Programa de Pós Graduação em Engenharia de Produção – PPGEP
Arquivo Bolsa.names CBA Data Mining II

ComBolsa, SemBolsa.

NSextSg: ignore. | Número de sextantes com sangramento. Classes: Nenhum,Um,Dois,3ouMais.
Calculo: ComCalculo, SemCalculo. | Ocorrência de cálculo dentário.
NSextCal: ignore. | Número de sextantes com cálculo: Nenhum,Um,Dois,3ouMais.
CPI: ignore. | Valor do índice CPI: Zero,Um,Dois,3ou4.
NSextSad: ignore. | Número de sextantes sadios: Nenhum,Um,Dois,3ouMais.
NSextAlt: ignore. | Número de sextantes doentes: Nenhum,Um,Dois,3ouMais.
EdAlist: Mais8anos, Ate8anos. | Nível de instrução do alistando.
EdPai: Mais8anos, Ate8anos, NaoInform. | Nível de instrução do pai do alistando.
EdMae: Mais8anos, Ate8anos, NaoInform. | Nível de instrução da mãe do alistando.
RendaFam: Maior5SM, Ate5SM, NaoInform. | Renda familiar do alistando.
Fumo: ExFumante, Fumante, NaoFuma. | Hábito de fumar do alistando.
SG: ComSG, SemSG. | Ocorrência de sangramento gengival.

APÊNDICE 7 – Arquivo Bolsa.data

| Universidade Federal de Santa Catarina – UFSC
| Centro Tecnológico - CTC
| Programa de Pós Graduação em Engenharia de Produção – PPGE
| Arquivo Bolsa.data CBA Data Mining II

| NSEXTSG, CALCULO, NSEXTCAL, CPI, SEXTSAD, SEXTALT, EDALIST, EDPAI, EDMAE, RENDAFAM, FUMO, SG, BOLSAS

Nenhum, ComCalculo, Um, Dois, 3ouMais, Um, Mais8anos, Mais8anos, Mais8anos, Maior5SM, ExFumante, SemSG, SemBolsa
Um, SemCalculo, Nenhum, Um, 3ouMais, Um, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
Nenhum, SemCalculo, Nenhum, Zero, 3ouMais, Nenhum, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, SemSG, SemBolsa
Dois, ComCalculo, Um, Dois, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, NaoInform, NaoFuma, ComSG, SemBolsa
Nenhum, SemCalculo, Nenhum, 3ou4, 3ouMais, Um, Mais8anos, Mais8anos, Ate8anos, NaoInform, NaoFuma, SemSG, ComBolsa
Nenhum, SemCalculo, Nenhum, Zero, 3ouMais, Nenhum, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, SemSG, SemBolsa
3ouMais, ComCalculo, Um, 3ou4, Dois, 3ouMais, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, ComBolsa
3ouMais, SemCalculo, Nenhum, Um, 3ouMais, 3ouMais, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
Dois, ComCalculo, Um, Dois, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
Nenhum, SemCalculo, Nenhum, Zero, 3ouMais, Nenhum, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, SemSG, SemBolsa
Um, ComCalculo, Um, Dois, 3ouMais, Um, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
Dois, ComCalculo, Um, Dois, 3ouMais, Dois, Mais8anos, Ate8anos, Ate8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
Dois, SemCalculo, Nenhum, Um, 3ouMais, Dois, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
Dois, ComCalculo, Um, Dois, 3ouMais, Dois, Mais8anos, NaoInform, Ate8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
3ouMais, SemCalculo, Nenhum, Um, Dois, 3ouMais, Mais8anos, Mais8anos, Mais8anos, Ate5SM, NaoFuma, ComSG, SemBolsa
3ouMais, ComCalculo, Um, Dois, 3ouMais, 3ouMais, Mais8anos, Ate8anos, Ate8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
3ouMais, ComCalculo, Um, Dois, 3ouMais, 3ouMais, Mais8anos, Mais8anos, Ate8anos, Maior5SM, NaoFuma, ComSG, SemBolsa
.
.
Um, SemCalculo, Nenhum, Um, 3ouMais, Um, Mais8anos, Mais8anos, Mais8anos, Maior5SM, NaoFuma, ComSG, SemBolsa

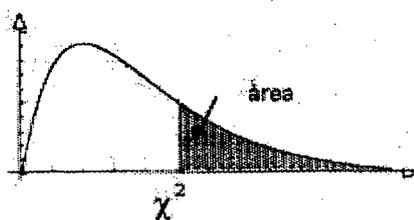
ANEXOS

ANEXO 1 – Tabela da Distribuição Qui-quadrado (χ^2)

ANEXO 2 – Tabela da Distribuição da Escolaridade em Florianópolis, por sexo e grupos de idade.

ANEXO 1 – Tabela de Distribuição Qui-quadrado (χ^2)

Distribuição
qui-quadrado



gl	área na cauda superior						
	0,250	0,100	0,050	0,025	0,010	0,005	0,001
1	1,32	2,71	3,84	5,02	6,63	7,88	10,8
2	2,77	4,61	5,99	7,38	9,21	10,6	13,8
3	4,11	6,25	7,81	9,35	11,3	12,8	16,3
4	5,39	7,78	9,49	11,1	13,3	14,9	18,5
5	6,63	9,24	11,1	12,8	15,1	16,7	20,5
6	7,84	10,6	12,6	14,4	16,8	18,5	22,5
7	9,04	12,0	14,1	16,0	18,5	20,3	24,3
8	10,2	13,4	15,5	17,5	20,1	22,0	26,1
9	11,4	14,7	16,9	19,0	21,7	23,6	27,9
10	12,5	16,0	18,3	20,5	23,2	25,2	29,6
11	13,7	17,3	19,7	21,9	24,7	26,8	31,3
12	14,8	18,5	21,0	23,3	26,2	28,3	32,9
13	16,0	19,8	22,4	24,7	27,7	29,8	34,5
14	17,1	21,1	23,7	26,1	29,1	31,3	36,1
15	18,2	22,3	25,0	27,5	30,6	32,8	37,7
16	19,4	23,5	26,3	28,8	32,0	34,3	39,3
17	20,5	24,8	27,6	30,2	33,4	35,7	40,8
18	21,6	26,0	28,9	31,5	34,8	37,2	42,3
19	22,7	27,2	30,1	32,9	36,2	38,6	43,8
20	23,8	28,4	31,4	34,2	37,6	40,0	45,3
21	24,9	29,6	32,7	35,5	38,9	41,4	46,8
22	26,0	30,8	33,9	36,8	40,3	42,8	48,3
23	27,1	32,0	35,2	38,1	41,6	44,2	49,7
24	28,2	33,2	36,4	39,4	42,8	45,6	51,2
25	29,3	34,4	37,7	40,6	44,3	46,9	52,6
26	30,4	35,6	38,9	41,9	45,6	48,3	54,1
27	31,5	36,7	40,1	43,2	47,0	49,6	55,5
28	32,6	37,9	41,3	44,5	48,3	51,0	56,9
29	33,7	39,1	42,6	45,7	49,6	52,3	58,3
30	34,8	40,3	43,8	47,0	50,9	53,7	59,7
40	45,6	51,8	55,8	59,3	63,7	66,8	73,4
50	56,3	63,2	67,5	71,4	76,2	79,5	86,7
60	67,0	74,4	79,1	83,3	88,4	92,0	99,6
70	77,6	85,5	90,5	96,0	100	104	112
80	88,1	96,6	102	107	112	116	125
90	98,6	108	113	118	124	128	137
100	109	118	124	130	136	140	149

ANEXO 2 – Escolaridade em Florianópolis por Sexo e Grupos de Idade

SANTA CATARINA - Florianópolis

SANTA CATARINA - Florianópolis
Tabela 6 - Pessoas residentes de 4 anos ou mais de idade, por anos de estudo,
segundo o sexo e os grupos de idade

Sexo e grupos de idade	Total	Anos de estudo						15 anos ou mais	Não determinado
		Sem instrução e menos de 1 ano	1 a 3 anos	4 a 7 anos	8 a 10 anos	11 a 14 anos			
Total	254 647	24 627	31 094	65 248	41 404	54 261	34 748	2 385	
4 a 6 anos	13 372	12 823	111	0	0	0	0	438	
7 a 9 anos	14 226	4 971	8 972	69	0	0	0	213	
10 a 14 anos	26 603	572	8 179	16 939	669	0	0	244	
15 a 19 anos	27 527	316	1 205	8 899	12 670	4 072	0	385	
20 a 24 anos	28 491	282	992	5 820	6 704	11 297	2 246	170	
25 a 29 anos	23 347	352	846	4 775	4 807	7 998	4 583	114	
30 a 34 anos	24 393	388	1 129	4 621	4 584	7 924	5 888	121	
35 a 39 anos	21 842	390	1 159	4 354	3 556	6 304	5 738	341	
40 a 44 anos	19 250	477	1 427	3 953	2 745	5 165	5 370	112	
45 a 49 anos	15 710	426	1 418	3 649	3 890	3 786	4 494	47	
50 a 54 anos	11 420	506	1 265	2 994	3 260	2 474	2 780	39	
55 a 59 anos	8 564	451	1 162	2 451	1 117	1 753	1 609	21	
60 a 64 anos	6 922	520	1 187	2 051	885	1 269	977	34	
65 anos ou mais	14 640	2 157	2 801	4 568	1 588	2 155	1 271	50	
Idade ignorada	381	36	43	105	45	64	32	58	
Homem	123 017	12 206	15 736	31 234	19 101	25 347	16 223	1 170	
4 a 6 anos	6 893	6 616	46	0	0	0	0	231	
7 a 9 anos	7 304	2 733	4 408	39	0	0	0	124	
10 a 14 anos	13 642	364	4 577	8 281	305	0	0	125	
15 a 19 anos	13 805	175	710	4 730	6 027	1 800	0	183	
20 a 24 anos	13 126	149	545	3 048	2 800	5 530	974	82	
25 a 29 anos	11 395	187	511	2 401	2 221	3 958	2 063	54	
30 a 34 anos	11 679	194	527	2 274	2 131	3 842	2 644	62	
35 a 39 anos	10 237	174	545	1 964	1 675	2 945	2 776	158	
40 a 44 anos	9 055	206	632	1 611	1 229	2 373	2 750	54	
45 a 49 anos	7 448	159	629	1 643	812	1 640	2 539	26	
50 a 54 anos	5 459	178	561	1 336	532	1 085	1 728	19	
55 a 59 anos	4 085	184	460	1 064	432	782	1 129	14	
60 a 64 anos	3 160	192	492	880	349	560	673	14	
65 anos ou mais	5 756	691	1 036	1 710	569	802	931	17	
Idade ignorada	177	14	17	55	19	30	16	26	
Mulher	131 630	12 421	16 258	34 014	22 303	28 914	18 525	1 195	
4 a 6 anos	6 479	6 207	65	0	0	0	0	207	
7 a 9 anos	6 921	2 238	4 564	30	0	0	0	69	
10 a 14 anos	12 961	218	3 602	8 658	361	0	0	118	
15 a 19 anos	13 922	141	495	4 189	8 643	2 272	0	202	
20 a 24 anos	13 385	113	447	2 774	2 904	5 767	1 272	68	
25 a 29 anos	11 952	165	435	2 374	2 380	4 038	2 800	90	
30 a 34 anos	12 718	172	602	2 347	2 433	4 062	3 024	58	
35 a 39 anos	11 605	216	614	2 390	1 881	3 359	2 962	183	
40 a 44 anos	10 195	271	795	2 142	1 516	2 793	2 620	58	
45 a 49 anos	8 262	287	787	2 006	1 078	2 148	1 855	21	
50 a 54 anos	5 991	330	664	1 656	828	1 369	1 057	20	
55 a 59 anos	4 479	267	662	1 387	665	971	480	7	
60 a 64 anos	3 762	328	695	1 171	538	708	304	20	
65 anos ou mais	8 844	1 466	1 765	2 858	1 029	1 353	340	33	
Idade ignorada	204	22	26	50	26	34	16	30	

Fonte - Fundação Instituto Brasileiro de Geografia e Estatística - IBGE