UNIVERSIDADE FEDERAL DE SANTA CATARINA
Programa de Pós-graduação em
Engenharia de Produção e Sistemas

Aplicação de Data Mining na Busca de um Modelo de Prevenção da Mortalidade Infantil

Dissertação de Mestrado

Ivana Corrêa de Oliveira

Florianópolis
2001

UNIVERSIDADE FEDERAL DE SANTA CATARINA
Programa de Pós-graduação em
Engenharia de Produção e Sistemas

Aplicação de Data Mining na Busca de um Modelo de Prevenção da Mortalidade Infantil

Ivana Corrêa de Oliveira

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

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE REDUÇÕES

Abreviaturas

DM - Data Mining

DN - Declaração de Nascimento

DO - Declaração de Óbito

MI - Mortalidade Infantil

MS - Ministério da Saúde

NV - Nascido Vivo

Siglas

CBA - Classification-Based on Association

CID - Código Internacional de Doenças

CENEPI - Centro Nacional de Epidemiologia

dBASE - Data Base

IBGE - Instituto Brasileiro de Geografia e Estatística

IBM - Internacional Business Machines Corporation

KDD - Knowledge Discovery in Databases

PNAD - Pesquisa Nacional por Amostra a Domicílio

SIM - Sistema de Informações sobre Mortalidade

SINASC - Sistema de Informações sobre Nascidos Vivos

LISTA DE TABELAS

RESUMO

Os sistemas de informações são instrumentos vitais para estabelecer as políticas voltadas à resolução de problemas de saúde. Nesse contexto, apresentam-se o Sistema de Informações sobre Nascidos Vivos - SINASC e o Sistema de Informações sobre mortalidade – SIM, que são utilizados como fonte de pesquisas e avaliação epidemiológica e permitem o conhecimento de importantes indicadores de saúde, dentre esses, o coeficiente de mortalidade infantil. O trabalho aplicou técnicas estatísticas, através do Teste do Qui-Quadrado e Data Mining do processo KDD (*Knowledge discovery in databases*), partindo-se da base de dados do SINASC, no ano de 1996, do município de Florianópolis e da ocorrência ou não de óbito no primeiro ano de vida. O objetivo é detectar as variáveis associadas à essas mortes e gerar regras de classificação que visam traçar o perfil dos recém-nascidos em risco de óbito no primeiro ano de vida. Os resultados revelam a associação estatística das variáveis socioeconômicas e biológicas com óbito; as regras de classificação permitem traçar o perfil dos recém-nascidos que devem receber assistência eficaz e auxiliar o processo de tomada de decisão, contribuindo para a redução da mortalidade infantil. Ao final do estudo, sugere-se novos trabalhos que poderão nortear ações de planejamento de saúde, contribuindo para a implantação de um modelo de prevenção.

Palavras-chave: Epidemiologia, Mortalidade Infantil, SINASC, KDD, Data Mining.

ABSTRACT

Information systems are important instruments to establish policies dealing with health issues and its solutions. In this context, the utilization of a record linkage information systems (*Sistema de Informações sobre Nascidos Vivos - SINASC e o Sistema de Informações sobre mortalidade – SIM*) on birth and mortality serve as a key source of research and epidemiological evaluation that provide knowledge on important health indicators, among others, the rate of infant mortality. The present work applied statistical techniques through the Chi-square Test and Data Mining of KDD (knowledge discovery in databases), in a database of live births for the year 1996, in the city of Florianópolis, Brazil, and the occurrence, or not, of deaths before the age of one year. The objective is to detect variables associated with deaths and to generate rules of classification that can draw the profile of recent live-births at risk of dying before reaching one year of age. The results of this analysis show the statistical association between death with some socioeconomic and biological variables. The classification rules allows for drawing a profile of the life births that should receive an effective assistance, and aid the decision-making process for reducing infant mortality. The study also suggested that new work could point to planning of health care policies through the elaboration of preventive models.

Key-words: Infant Mortality/ Epidemiology, SINASC, KDD, Data Mining.

CAPÍTULO 1

INTRODUÇÃO

Busca-se assegurar através de um sistema de informações sobre saúde, a permanente avaliação da situação de saúde da população e das ações executadas. Assim, é essencial concebê-lo como um instrumento para o processo de tomada de decisões, seja em questões técnicas ou em políticas a serem implementadas. É um processo gerador de conhecimento que descreve uma realidade social e, por meio deste conhecimento, procura-se não apenas

informações sobre saúde, mas a própria reorganização dos serviços de saúde (MINISTÉRIO DA SAÚDE, 1995).

O Sistema de Informações sobre Nascidos Vivos (SINASC) mantém, em âmbito nacional, um banco de dados importante sobre os recém-nascidos, através da coleta de dados sobre o nascimento e o parto e da geração de dados populacionais, caracteriza epidemiologicamente os nascimentos e disponibiliza essas informações a todos os níveis do sistema de

saúde (MINISTÉRIO DA SAÚDE, 1999; MINISTÉRIO DA SAÚDE, 1995).

A mortalidade infantil é considerada um dos indicadores mais importantes na avaliação do risco do ser humano morrer antes de completar um ano de vida, no reconhecimento do desempenho dos serviços de saúde e ainda, nas relações com as condições de vida e saúde da mulher.

A utilização conjunta dos dados sobre nascimentos e óbitos conduz a estudos epidemiológicos que buscam aprofundar o

conhecimento sobre as variáveis biológicas e socioeconômicas mais associadas à mortalidade infantil.

A tecnologia para a exploração e análise de dados conhecida por Data Mining tem por objetivo descobrir padrões significativos e gerar regras que descrevem o comportamento de uma base de dados, fundamentando a tomada de decisões com base no conhecimento (BERRY & LINOFF, 1997; DILLY, 1995).

Ao usar Data Mining procura-se identificar os recém-nascidos com maior risco de morrer antes de completarem um ano de idade, analisando o conjunto de variáveis a que estão expostos. Com esse conhecimento pode-se avaliar a assistência pré-natal, o parto e o primeiro ano de vida, possibilitando traçar um modelo de prevenção para a redução da mortalidade neonatal.

É neste contexto que se insere a presente dissertação. O trabalho consiste na utilização de técnicas de Data Mining para identificar e extrair conhecimento, tendo como fonte de dados o Sistema de Informações sobre Nascidos Vivos e ocorrência ou não óbito no primeiro ano de vida.

1.1 Objetivos do Trabalho

1.1.1 Objetivo Geral

Detectar, a partir da utilização de técnicas de Inteligência Artificial, em dados sobre nascidos vivos, as variáveis que estão associadas à morte de menores de um ano de idade, ocorridos no município de Florianópolis no ano de 1996.

1.1.2 Objetivos Específicos

- Verificar a existência da associação entre óbito de crianças menores de um ano com os dados clínicos e epidemiológicos do nascimento e os dados sociais da mãe, através do Teste estatístico do Qui-Quadrado;

- Aplicar técnicas de Data Mining na base de dados do Sistema de Informações sobre Nascidos Vivos, para descobrir as associações com as variáveis mais influentes na morte da população estudada;

- Gerar regras de classificação para os grupos em risco de óbito, a fim de auxiliar na elaboração de um modelo de prevenção.

1.2 Importância

Com o processo de descoberta de conhecimento, busca-se identificar as variáveis mais influentes na mortalidade infantil, e assim, definir medidas que possam auxiliar na elaboração de um modelo de prevenção e desencadear políticas de saúde para atender gestantes e recém-nascidos.

As variáveis coletadas na Declaração de Nascidos Vivos podem direcionar as formas de ação do setor de saúde para reduzir o número de mortes no primeiro ano de vida.

O estudo chama a atenção dos administradores de saúde para intervirem no planejamento estratégico, principalmente referindo-se à alimentação, nutrição, educação, condições de trabalho e de moradia.

O monitoramento das variáveis que influem nas condições de saúde da gestante e do neonato colabora para o estabelecimento de um quadro favorável, refletindo-se na melhoria dos indicadores sociais.

1.3 Limitações do Trabalho

A qualidade dos dados foi uma das limitações encontradas neste trabalho. Isto devido aos erros e ao mal preenchimento de algumas variáveis da declaração de nascidos vivos que dificultaram a interpretação dos resultados;

Os dados inconsistentes denotam a falta de comprometimento dos responsáveis com o correto preenchimento das declarações de nascidos vivos e com a alimentação do banco de dados, por conseqüência prejudicam a interpretação dos dados. Se não fosse assim, o sistema poderia trazer resultados mais precisos e uma melhor qualidade das informações.

A falta de integração entre as bases de dados de nascimento e de mortalidade impossibilitou que o estudo se estendesse a outros anos e a outras esferas.

1.4 Estrutura do Trabalho

Este trabalho é constituído por cinco capítulos. Esses capítulos objetivam introduzir e demonstrar a viabilidade da aplicação de Data Mining na relação de nascimentos e mortes em menores de um ano de idade.

A estrutura do trabalho apresenta-se da seguinte maneira:

- Capítulo 1: introdução, objetivos, importância e limitações do trabalho.
- Capítulo 2: uma pequena introdução dos sistemas de informações sobre saúde, SINASC e SIM, bem como sua relação com a mortalidade infantil.
- Capítulo 3: conceituação de Data Mining, os métodos, técnicas, ferramentas estudadas e aplicadas no estudo, e ainda, exemplos de aplicações na área da saúde.
- Capítulo 4: conceituação teórica do modelo para a aplicação de Data Mining na base de dados.
- Capítulo 5: aplicação de testes estatísticos e de Data Mining, demonstrando os resultados encontrados.
- Capítulo 6: conclusões do trabalho, recomendações e projetos futuros.
- Bibliografia referenciada e consultada, além do glossário e dos anexos.

CAPÍTULO 2

SISTEMAS DE INFORMAÇÕES SOBRE SAÚDE

Os sistemas de informações sobre saúde são instrumentos úteis e de importância vital para estabelecer as políticas voltadas à resolução de problemas de saúde e suas implicações socioeconômicas.

Os principais objetivos de um sistema de informação em saúde são: o de avaliar e apoiar o planejamento, a tomada de decisões e as ações em todos os níveis, político-estratégico, gerencial e operacional e o de apoiar o desenvolvimento científico e tecnológico do setor saúde. A informação constitui-se em insumo estratégico para a formulação de políticas e processos de planejamento, de decisão e de atuação nas diversas instâncias da organização e gerência dos serviços de saúde (MINISTÉRIO DA SAÚDE, 1995).

Neste contexto, apresenta-se o Sistema de Informações sobre Nascidos Vivos e o Sistema de Informações sobre Mortalidade.

2.1 O Sistema de Informações Sobre Nascidos Vivos

2.1.1 Histórico

A contagem de indivíduos foi sempre uma preocupação dos povos, desde os tempos mais remotos. Antes da era Cristã,

na Grécia, Roma e nos antigos povos do Oriente registravam-se apenas alguns fatos vitais com finalidades militares ou tributárias. A verdadeira origem dos dados populacionais é representada pelos livros de registros eclesiásticos feitos pela igreja. A primeira quantificação de pessoas que se tem conhecimento está registrada na Bíblia, conhecida como Censo do povo Hebreu (MELLO JORGE et al., 1993).

O primeiro registro instituído, não mais pela igreja, mas pelo Estado, ocorreu entre os Íncas no Peru. Por não conhecerem

formas escritas para registrar seus nascimentos e mortes, usavam cordões coloridos com nós chamados quipus, mediante os quais tinham o controle das pessoas que nasciam e morriam (LAURENTI et al., 1987).

Ao longo dos séculos, o registro dos eventos vitais foi ganhou amplitude e legalidade, passou a ser obrigatório em todos os países.

No Brasil, um Decreto datado de 7 de Março de 1888 foi o primeiro ato a

regulamentar os registros das pessoas naturais, ou sejam, nascimentos, casamentos e óbitos.

O Código Civil Brasileiro, de 1916, determinou em seu artigo 12 que: "*deverão ser inscritos em registros públicos os nascimentos, casamentos e óbitos"*, cabendo àunião legislar sobre eles. Outros Decretos foram instituídos a partir deste nos anos de 1939, 1969 e 1973, para regulamentar a prática e a obrigatoriedade desses registros, culminando com a Lei n.º 8069 de 1990 que ficou conhecida como a

Lei dos Registros Públicos do Brasil (Anexo I).

Até recentemente, os métodos adotados para quantificar os nascidos vivos eram manuais, dificultavam a execução e comprometiam a qualidade. Além disso, não havia um padrão operacional uniforme, coordenada e integrada com os diversos órgãos envolvidos, sejam hospitais, maternidades, cartórios e instituições que elaboram as estatísticas de saúde. As informações utilizadas apoiavam-se em

levantamentos aleatórios e bases comparativas (LAURENTI et al., 1987).

A inserção do Instituto Brasileiro de Geografia e Estatística (IBGE), em 1984, assumindo a responsabilidade de processar, analisar e divulgar as informações sobre nascidos vivos não foi suficiente para solucionar as falhas e imprecisões dessas informações. Esse procedimento gerava o dado relacionado apenas com os casos registrados, não alcançando ou alterando a situação dos sub-registros de nascimentos. A defasagem de tempo entre a coleta e a

publicação dos dados era outra limitação, visto que os cartórios encaminhavam trimestralmente seus dados e que o IBGE levava de 3 a 4 anos para disponibilizá-los, além disso, os dados estatísticos não registravam as variáveis essenciais para buscar qualquer tipo de prevenção.

Frente àsituação exposta, o Ministério da Saúde - MS decidiu investir em um sistema que permitisse a análise estatística e possibilitasse executar ações básicas de saúde.

2.1.2 Identificação

O MS começou a implantar, em 1990, sob sua coordenação, o Sistema de Informações sobre Nascidos Vivos – SINASC (MINISTÉRIO DA SAÚDE, 1999).

O objetivo desse sistema era registrar os nascidos vivos, a partir de um documento básico gerado nos hospitais, outras instituições de saúde que realizam partos e nos cartórios do Registro Civil, para os partos ocorridos no domicílio. Considerava-se essas instituições como os locais que poderiam prover as mais completas e corretas informações relativas aos nascimentos, visto que cerca de 80% dos partos no Brasil são hospitalares, segundo dados do IBGE em 1990.

MELLO JORGE citada por FURQUIM (1993) afirma que a implantação do SINASC permitiu a obtenção de dados mais detalhados e fidedignos sobre nascidos vivos do que os existentes anteriormente. As instituições de saúde passaram a utilizar esses dados como fonte de pesquisas, avaliação epidemiológica e administrativa da assistência materno-infantil prestada em suas áreas.

Um estudo do Ministério da Saúde apontou que de três crianças nascidas no Brasil, uma não possui certidão de nascimento. Assim, a cada ano, um milhão de crianças incrementam as estatísticas de brasileiros sem registro civil, ferindo seus direitos fundamentais, visto que o acesso aos serviços sociais básicos depende da comprovação da

existência civil. O sub-registro, além de ferir os princípios de cidadania, implica desrespeito à Convenção dos Direitos da Criança.

Em dezembro de 1997 foi aprovada a Lei nº 9534, Lei de Registro Civil, que determina a gratuidade do registro civil a todos os brasileiros. Segundo a lei, a certidão de nascimento e de óbito é gratuita. A gratuidade do registro civil é um dos caminhos para melhorar o planejamento e a execução de políticas públicas, principalmente na área da saúde e a declaração de nascido vivo passa a ser um instrumento para cumprir esse direito.

2.1.3 A Declaração de Nascidos Vivos

A fonte de dados do SINASC é o formulário de Declaração de Nascidos Vivos - DN (Anexo II), de uso padronizado em todo o país, emitido pelos estabelecimentos de saúde públicos ou privados, que prestam assistência ao parto, ou pelos cartórios, em caso de partos domiciliares.

A DN é preenchida em três vias, uma é entregue à mãe com o encaminhamento ao cartório para a emissão da certidão de nascimento. As outras duas são recolhidas, analisadas, processadas e utilizadas pelos órgãos de saúde pública, ou sejam, secretarias municipais e centros de saúde, para adotarem medidas que visem a acompanhar a mãe e o recém-nascido, subsidiadas nas informações geradas pelo sistema (SANTA CATARINA, 1999).

A DN é composta de sete blocos com as seguintes informações:

Bloco I - Número

Classifica numericamente a DN. Este número é pré-impresso e destina-se a identificar o evento.

Bloco II - Cartório

Refere-se a informar dados sobre o Cartório do Registro Civil onde o nascimento foi registrado. É composto pelo número e data do registro, município de ocorrência e código da Unidade da Federação.

Bloco III - Local de Ocorrência

Este bloco é relativo ao local onde ocorreu o parto, indica se a criança nasceu ou não em hospital e ainda registra o endereço completo do local.

Bloco IV - Mãe

Refere-se a identificar a mãe e a sua história reprodutiva. Contém o nome da mãe, idade, grau de instrução, filhos nascidos vivos, mortos ou abortados e o endereço completo da sua residência.

Bloco V - Gestação e Parto

Contém as características da gestação e do parto. É composto pela duração da gestação, tipo de gravidez, tipo de parto e número de consultas pré-natais.

Bloco VI - Recém-Nascido

Destina-se a anotar as características do recém-nascido. Possui os dados completos sobre a data do nascimento, sexo, peso ao nascer e índice de apgar no 1º e no 5º minuto.

Bloco VII - Responsável pelo Preenchimento

Refere-se a identificar o responsável pelo preenchimento da DN.

2.1.3.1 Descrição dos Dados

Os dados da DN representam as características da gestação e do parto através das variáveis de descrição da mãe e do recém-nascido. Algumas de caráter informativo e outras aferidas após o parto.

Os dados referentes à **gestação** e ao **parto** são:

a) *Duração da Gestação*: define em semanas o tempo gestacional. A partir desse número é possível identificar a presença de prematuridade, ou seja, gestações abaixo de 36 semanas.

b) *Pré-Natal*: demonstra o número de consultas realizadas durante o acompanhamento da gravidez. O pré-natal é uma ação fundamental para qualificar a atenção dada à saúde da mulher e da criança.

c) *Tipo de Parto*: refere-se ao tipo de procedimento adotado: parto normal, cesárea, fórceps ou outro.

d) *Tipo de Gravidez*: define o tipo de gravidez, única ou gemelar.

As variáveis de **descrição da mãe** são:

a) *Idade*: informa o número de anos completos da mãe no momento do parto. O interesse nessa variável é identificar as mães com idade de risco, ou sejam, as

com idade superior a trinta e cinco anos e as adolescentes, na faixa etária abaixo de dezenove anos.

b) *Escolaridade*: apresenta o nível de instrução da mãe, expressando sua situação socioeconômica e o contexto familiar onde está inserido o recém-nascido.

c) *Filhos Tidos*: indica se a mãe já teve filhos antes da gestação informada, caracteriza a história gestacional da mãe. É composta pelo conjunto das variáveis filhos tidos vivos, mortos e abortados.

As informações sobre o **recém-nascido** são:

a) *Sexo*: é uma variável padrão.

b) *Apgar no 1º e no 5º minuto*: representam as condições vitais do recém-nascido. Estes índices são aferidos na sala de parto no 1º e no 5º minuto de vida, recebendo uma pontuação que, somada, configura um valor que varia de zero a dez.

c) *Peso*: é uma variável importante para indicar problemas relacionados à desnutrição e a prematuridade.

2.1.4 Processo de Implantação do Sistema

A implantação do SINASC nos Estados se deu de forma gradativa e sob a responsabilidade das Secretarias Estaduais de Saúde, de acordo as suas condições estruturais e operacionais.

Em Santa Catarina, o sistema informatizado foi instalado em 1994 na Secretaria Estadual de Saúde e nas 18 Coordenações Regionais de Saúde, concentrando-se nessas a digitação dos dados oriundos dos municípios, uma vez que o formulário DN já estava previamente implantado nos hospitais e cartórios.

No âmbito estadual, o município de Florianópolis foi o primeiro a ter o sistema instalado, em meados de 1995, na sua Secretaria Municipal de Saúde.

Gradualmente, o sistema computacional foi implantado nos municípios que apresentavam condições técnicas para operacionalizá-lo e os dados cadastrados por local de residência da mãe.

A partir de 1996, o Serviço de Informática da Diretoria de Vigilância Epidemiológica da Secretaria de Estado da Saúde passou a gerenciar o SINASC. Começou-se então, um trabalho visando resgatar os registros não alimentados no sistema nos anos anteriores e consolidar uma base de dados estadual que, através da implementação de processos de filtragem de erros, pudesse proporcionar a qualidade e a integridade do banco de dados (SANTA CATARINA, 1999).

Os dados relativos às Secretarias Municipais de Saúde são repassados no papel ou em meio magnético, à respectiva Coordenação Regional de Saúde a qual encaminha, mensalmente, os dados à Diretoria de Vigilância Epidemiológica que analisa e consolida o banco de dados estadual, repassando-o mensalmente ao Centro Nacional de Epidemiologia do Ministério da Saúde, completando, assim, o fluxo de informação (Anexo III).

No estado de Santa Catarina, até Agosto de 2001, o sistema informatizado encontrava-se instalado em 103 municípios e em 03 hospitais, segundo dados da Secretaria Estadual da Saúde. A implantação do SINASC nos Municípios tem sido de forma positiva e gradual, de acordo com as suas condições técnicas e operacionais, apesar de ainda apresentar algumas dificuldades, principalmente para cumprir o fluxo em tempo satisfatório e obter o preenchimento correto da DN.

2.2 Mortalidade Infantil

A mortalidade ao longo da história da saúde pública tem sido estudada por muitos autores. Desde o trabalho de Willian Farr, em 1885, intitulado "Estatísticas Vitais", no qual inclui o estudo de uma série de aspectos sociais e da saúde, como por exemplo, a relação entre taxas de nascimento e morte, possibilidades de prolongamento da vida humana, relação entre saneamento e mortalidade.

Dentre os indicadores, o de mortalidade infantil - MI, mostra-se como um dos mais sensíveis para visualizar as nuances sociais (LAURENTI, 1997).

Estudos sobre a mortalidade, publicados pelo IBGE, demonstram que a defesa de mais investimentos governamentais em saúde pública e a ênfase na importância dos fatores econômicos e sociais como condicionantes da MI começaram novamente a ganhar a atenção dos estudiosos do assunto, utilizando o coeficiente de MI como indicador social.

É um tema de interesse científico e político utilizado historicamente como um bom indicador para avaliar as condições de saúde e de vida das populações. Permite subsidiar os processos de planejamento, gestão e avaliação de políticas e ações de saúde voltadas à

atenção pré-natal, ao parto e à criança, possibilita proceder a análise comparativa de situações de saúde em diferentes períodos, lugares e condições socioeconômicas (IBGE, 1999a, 2001).

2.2.1 Taxa de Mortalidade Infantil

As taxas de mortalidade permitem comparar a ocorrência de morte em diferentes populações, ou mortes por diferentes doenças na mesma população, ou ainda, mortes no mesmo período de tempo. O numerador na taxa de mortalidade é o número de pessoas que morrem durante um dado período de tempo, e o denominador é o número de pessoas com risco de morrer durante o período (DAWSON, 1993). Especificamente, a taxa de MI é o número de óbitos de menores de um ano, expresso por mil nascidos vivos, em determinado local e período, visando estimar o risco de uma criança morrer durante o primeiro ano de vida (MINISTÉRIO DA SAÚDE, 1997).

Segundo o IBGE (1999a), a taxa de MI é um dos indicadores mais sensíveis para medir o nível de desenvolvimento de um País. É geralmente classificada em *alta* (50 ou mais), *média* (20-49) e *baixa* (menos de 20), em função da proximidade ou distância dos valores já alcançados em sociedades mais desenvolvidas e que pode variar com o tempo (MINISTÉRIO DA SAÚDE, 1997).

2.2.2 Fatores de Risco

Pesquisas realizadas durante a década de 70 e início da de 80 procuram discutir as questões vinculadas ao problema da redução do ritmo de queda da MI nos países menos desenvolvidos. Preocupam-se com o caráter estratificado dessas sociedades, com destaque para a distribuição desigual de renda, acesso diferenciado aos recursos de saúde, saneamento, educação e outros componentes do padrão de vida das populações, culminando com avaliações sobre os diferentes impactos desses fatores nos níveis de mortalidade, entre os distintos estratos sociais (IBGE, 1999b).

A concentração dos recursos em determinadas áreas e grupos sociais tem sido um sério obstáculo a que se consigam maiores avanços na redução dos níveis da MI na maioria dos países do Terceiro Mundo. No Brasil, em particular, o modelo de desenvolvimento que vem vigorando ao longo dos anos, tem sido altamente excludente e concentrador de renda, recursos e serviços em determinadas regiões e estratos sociais. A partir da década de 70 o Brasil vem tomando algumas medidas de ações compensatórias, tais como: saneamento básico, programa de saúde materno-infantil, campanhas de vacinação e a ampliação da oferta de serviços médico-hospitalares. A partir dos anos 80, os programas de aleitamento materno e reidratação oral colaboraram, consideravelmente para a continuidade da redução dos níveis de MI (IBGE, 1999b).

A Figura 1 exibe a evolução da MI durante o período de 1930 a 1990, para o Brasil e Regiões (IBGE, 1999b). Esses valores estão detalhados no Anexo IV.

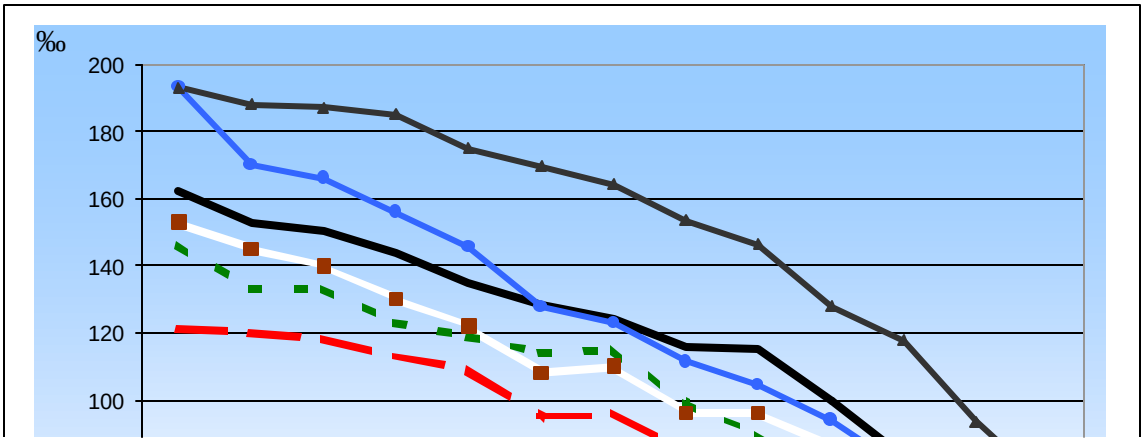


Figura 1: Taxa de mortalidade infantil, segundo IBGE, para as Regiões e Brasil - 1930/1990.

A inovação em tecnologias médicas talvez tenham impedido, de certa forma, impactos mais profundos da crise econômica sobre a MI, mas, ao mesmo tempo, impôs seus limites quando dissociados de políticas públicas mais gerais, através da educação da população, da melhor redistribuição dos recursos hospitalares, subsídios para alimentação, expansão dos sistemas de água potável, entre outros.

O aumento de renda geral não garantem, por si só, melhorias das condições de vida e saúde das populações, especialmente, quando ela é repartida de forma extremamente desigual entre os distintos estratos sociais, como vem ocorrendo no País (IBGE, 1999b).

A educação tem sido a variável chave na obtenção de quedas consistentes na MI, em todos os países, devido à maior percepção por parte das mulheres mais instruídas no

cuidado com seus filhos, possibilitando, desta forma, maior acesso aos serviços básicos de saúde (IBGE, 1999b, IBGE, 2001). A MI relacionada à mães sem instrução ou com pouca instrução chega a ser dez vezes superior às mais instruídas. Mesmo na situação em que as mães têm um nível educacional mínimo (quatro anos), a sobremortalidade infantil do grupo é 4,7 vezes superior à de crianças de mães com mais de 12 anos de instrução (IBGE, 1999b).

Solucionar os problemas socioeconômicos é de fundamental importância para garantir sustentabilidade à manutenção favorável do declínio da MI (IBGE, 1999b).

2.2.3 Tendências

Embora ao longo dos últimos anos, os índices de MI tenham diminuído, a taxa média do País está ainda entre as mais altas da América Latina. As taxas de mortalidade de crianças menores de 1 ano, calculadas pelo Ministério da Saúde, revelam a estreita ligação entre o estado de saúde da população, o acesso aos serviços e a qualidade do atendimento médico, condicionados pela situação socioeconômica da população (IBGE, 1999a).

Durante a década de 90, foram relevantes as transformações que ocorreram nos padrões de saúde da população brasileira. A mortalidade infantil vem mantendo a tendência histórica de queda (IBGE, 2001), passando, no período de 1992 a 1999, de 43‰ para 34,6‰, ou seja, um decréscimo de aproximadamente 20% e, neste cenário, perdem importância as causas relacionadas predominantemente às enfermidades infecciosas e parasitárias e às doenças respiratórias, passando a ser predominante as afecções perinatais, relacionadas a problemas congênitos e, também daqueles derivados da oferta de serviços de saúde de qualidade, especialmente no atendimento pré-natal (IBGE, 2001).

Entretanto, os diferenciais entre as regiões continuam elevados. Na Região Sul, que apresenta o menor índice para o mesmo período, a taxa de mortalidade infantil passou de 27,4‰ para 20,4‰, enquanto que no Nordeste, que apresenta o maior índice, passou de 65,2,3‰ para 53,0‰, conforme mostra a Figura 2.

Entre 1992 e 1999, o declínio da MI foi razoavelmente expressivo no grupo de mães com menor instrução, todavia, seu valor ainda é mais de 3 vezes superior ao observado no grupo de crianças com mães mais instruídas (IBGE,2001).

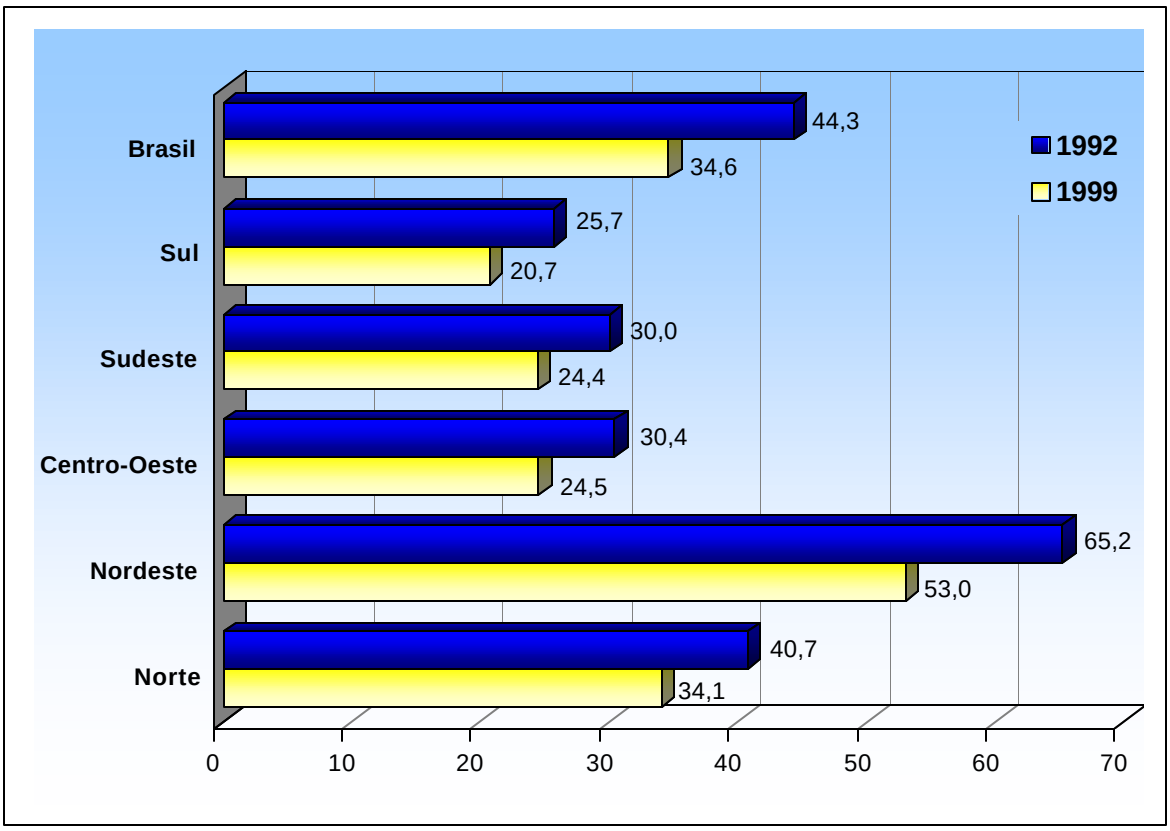


Figura 2: Taxas de mortalidade infantil, segundo dados do IBGE, para o Brasil e por regiões - 1992/1999.

É importante ressaltar que o valor da MI no Brasil para 1999 (34,6‰) está próximo da meta estipulada pela Organização das Nações Unidas pela Criança para o ano 2000 (IBGE, 2001), que é de 33‰. A Tabela 1 apresenta as taxas de MI por regiões e Unidades da Federação.

As estatísticas de óbitos mais recentes, produzidas pelo Ministério da Saúde, através do SIM, informam que, para o ano de 1998, dos mais de 71 mil óbitos de menores de 1 ano, aproximadamente 50% se referiam a causas perinatais (óbitos de crianças com menos de 7

dias de vida). Parte dessas mortes, tem como causa básica a ausência de atendimento pré-natal durante a gravidez, algumas vezes pela falta de oferta desse serviço, outros por falta de acesso a ele (IBGE, 2001).

A mortalidade de crianças menores de 1 ano relacionada às doenças infecto-contagiosas e parasitárias (11‰ em 1998), é um indicativo importante das precárias condições do saneamento básico, uma vez que cerca de 50% dos domicílios não dispõem de rede de esgoto adequado, expondo as crianças a contatos com dejetos.

De modo geral, as crianças pobres são as que mais sofrem pela ausência de saneamento adequado, visto que suas famílias carecem não apenas dos meios necessários para conseguir as instalações básicas, mas também de informações sobre a maneira de minimizar os efeitos nocivos das condições insalubres em que vivem (IBGE, 2001).

Assim, a MI no país ainda é elevada, principalmente quando comparada com a de países mais desenvolvidos, ou mesmo a países da América Latina, a exemplo da Costa Rica, Chile e Uruguai que apresentam patamares em torno de 10 mortes por mil nascimentos (IBGE, 2001).

Apesar dessas taxas ainda serem elevadas, é possível vislumbrar, num curto período de tempo, a tendência a redução dos níveis de MI na maioria dos países latino-americanos (IBGE, 1999b).

Além dessas causas, é importante frisar que o envolvimento de tecnologias, sejam elas médicas ou não, pode ser um fator essencial para que as taxas de mortalidade infantil sejam reduzidas a patamares inferiores aos dos países desenvolvidos.

Tabela 1: Taxas de mortalidade infantil, segundo as Regiões do Brasil e Unidades da Federação – 1999.

Grandes Regiões e Unidades de Federação	Taxas de mortalidade infantil (‰)
Brasil	**34,6**
Norte	**34,1**
Rondônia	31,6
Acre	44,2
Amazonas	31,8
Roraima	38,3
Pará	34,6
Amapá	31,7
Tocantins	33,0
Nordeste	**53,0**
Maranhão	54,2
Piauí	45,3
Ceará	52,4
Rio Grande do Norte	48,7
Paraíba	60,3
Pernambuco	58,2
Alagoas	66,1
Sergipe	45,5
Bahia	45,4
Sudeste	**24,4**
Minas Gerais	26,3
Espírito Santo	26,0
Rio de Janeiro	24,4
São Paulo	21,9
Sul	**20,7**
Paraná	24,3
Santa Catarina	22,2
Rio Grande do Sul	18,4
Centro-Oeste	**24,5**
Mato Grosso do Sul	24,4
Mato Grosso	27,5
Goiás	25,0
Distrito Federal	22,6

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílios -PNAD, 1999.

2.3 O Sistema de Informações Sobre Mortalidade

O SIM foi criado pelo Ministério da Saúde em 1975 para obter dados sistematizados sobre a mortalidade que, de modo íntegro e abrangente pudesse embasar os diversos níveis gerenciais em suas ações de saúde e foi o primeiro sistema em informações de estatísticas vitais desenvolvido em microcomputadores no país (MINISTÉRIO DA SAÚDE, 1999).

Esse sistema visa produzir estatísticas de mortalidade e construir os principais indicadores de saúde, permitindo, assim, a realização de estudos não apenas do ponto de vista estatístico epidemiológico, mas também sócio-demográfico (MINISTÉRIO DA SAÚDE, 1999a). O documento padrão do SIM é a Declaração de Óbito - DO, padronizada em todo o território nacional (Anexo V).

As causas mortes são preenchidas pelo médico e, posteriormente recebem um código segundo a Classificação Internacional de Doença - CID, no "Modelo Internacional de Certificado Médico da Causa de Morte", utilizado em todos os países e recomendado pela Assembléia Mundial de Saúde, em 1948 (MINISTÉRIO DA SAÚDE, 1999a).

A Lei dos Registros Públicos (BRASIL, 1976), determina no seu Artigo 77, que nenhum sepultamento será feito sem Certidão Oficial de Registro expedida no lugar do falecimento e, quando se tratar do óbito de crianças com menos de um ano, o oficial verificará se ela foi devidamente registrada, em caso contrário, o documento será previamente feito.

2.4 O SINASC e a Mortalidade Infantil

Os cálculos da taxa de MI podem ser feitos de duas maneiras: diretamente com os dados obtidos através do registro civil de óbito ou indiretamente, apoiando-se em estimativas (Ministério da Saúde, 1997). Optando pela forma direta, as taxas são calculadas utilizando-se o SIM e o SINASC.

O SINASC, além de proporcionar o conhecimento do número de nascimentos, permite também definir o perfil epidemiológico dos partos, possibilita obter coeficientes de MI específicos, por meio da análise conjunta das variáveis constantes na DN e na DO. Segundo FURQUIM (1993), a DN viabilizou as informações populacionais importantes para analisar e interpretar os dados sobre a mortalidade infantil.

Estudos sobre a MI podem ser realizados através da técnica de *linkage*, que relaciona as informações de nascimento com as de óbito. Esse procedimento permite identificar o mesmo indivíduo nos dois bancos de dados, partindo-se de um *estudo de coorte* de nascidos vivos e encontrando-se às DOs relativas aos óbitos infantis das respectivas DNs (FURQUIM, 1993).

Existem alguns fatores que limitam a qualidade da taxa de mortalidade infantil, dentre eles, pode-se citar a subnotificação de nascidos vivos, e, por exclusão a de óbitos, sendo necessário ajustar as taxas usando-se estimativas (MINISTÉRIO DA SAÚDE, 1997).

Vale salientar que o SIM é um sistema consolidado e amplamente conhecido, o SINASC é um sistema ainda em construção, cujo aprimoramento é fundamental para estabelecer as ações e o relacionamento entre os dois sistemas.

2.5 Dados da Mortalidade Infantil em Santa Catarina e em Florianópolis

A Figura 3 apresenta os coeficientes de mortalidade infantil no Brasil, Santa Catarina e Florianópolis, de 1989 a 1998. Nesse contexto, observa-se que os índices têm decrescido ao longo dos anos. Para o período citado, a redução ficou na ordem de 22,09% para Florianópolis e 22,28% para Santa Catarina.

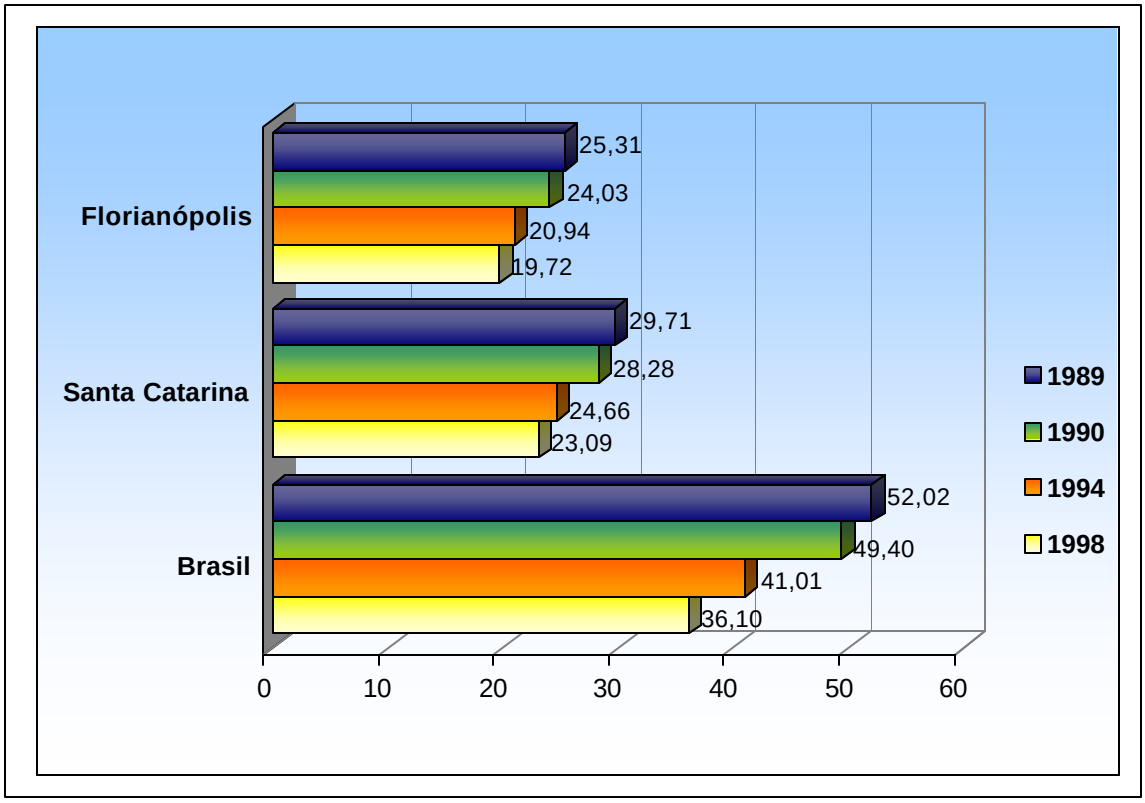


Figura 3: Taxas de Mortalidade Infantil em Santa Catarina e Florianópolis - 1989/1998, segundo dados do Ministério da Saúde.

Os dados analisados levaram em consideração o número de nascidos vivos em Santa Catarina para o ano de 1996, que segundo o IBGE, foi de 86.069, por residência da mãe.

CAPÍTULO 3

DATA MINING

As duas últimas décadas têm demonstrado um crescente aumento no número de informações e de dados armazenados em meio eletrônico e também que as organizações, em suas operações diárias, geram e coletam grandes volumes de dados, porém não são capazes de aplicá-los plenamente, pois as informações úteis estão implícitas e são de difícil compreensão (DILLY, 1995).

Para se manterem competitivas no mercado, as organizações precisam identificar as informações importantes e utilizá-las no processo de tomada de decisões (IBM, 1996). Para tanto, necessitam de técnicas de análises de dados automatizadas que as ajudem encontrá-las (LUBEL, 1998). Neste contexto, está o processo de descoberta de conhecimento, no qual Data Mining se apresenta como a principal etapa do processo.

3.1 O PROCESSO DE DESCOBERTA DE CONHECIMENTO

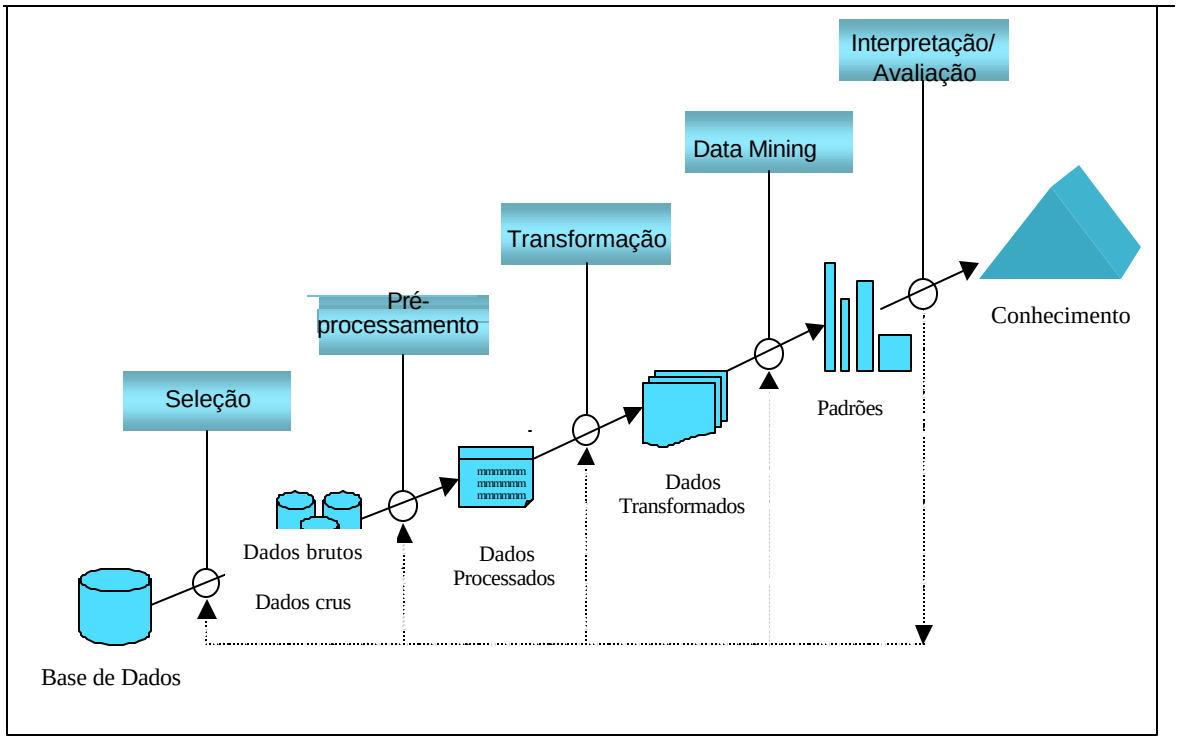
Obter conhecimentos em base de dados é uma área de pesquisa crescente que atrai esforços de pesquisadores. Fundamenta-se no fato de que as grandes bases de dados podem ser uma fonte de conhecimento útil, porém não explicitamente representado, e cujo objetivo é desenvolver e validar técnicas, metodologias e ferramentas capazes de extrair o conhecimento implícito nesses dados e representá-lo de forma acessível aos usuários (FELDENS, 1996).

Inicialmente, foram designados vários nomes à noção de achar padrões úteis em dados brutos, tais como *Data Mining*, Extração de Conhecimento,

> Descoberta de Informação, Mineração de Dados e Processamento do Padrão de Dados. Apenas em 1989, o termo "*Descoberta de Conhecimento em Base de Dados*" (KDD – *Knowledge Discovery in Databases*) foi utilizado para se referir ao processo total de procurar conhecimentos em dados, com a aplicação de técnicas de *Data Mining* (FAYYAD et al., 1996).

O desenvolvimento de sistemas de KDD está relacionado com diversos domínios de aplicações: marketing, análises corporativas, astronomia, medicina, biologia, entre outros. Existem diversas tarefas de KDD que são, principalmente, dependentes do domínio da aplicação e do interesse do usuário; cada tarefa de KDD extrai um tipo diferente de conhecimento do banco de dados e pode requerer um algoritmo diferente para cada tarefa.

Transformar os dados em informações que possam auxiliar à tomada de decisões é um processo complexo (IBM, 1996) e pode ser organizado em cinco passos, conforme ilustra a Figura 4.



O primeiro passo no processo de KDD é entender o domínio da aplicação, identificar o problema e definir os objetivos a serem atingidos. O processo inicia com os dados brutos e finaliza com a extração de conhecimento, como resultado das seguintes etapas:

1. *Seleção* - é a extração dos dados visando à aplicação. Nesta etapa pode ser necessário integrar e compatibilizar as bases de dados (DILLY, 1995).

2. *Pré-Processamento* - As informações consideradas desnecessárias são removidas. Adotam-se estratégias para manusear dados perdidos ou inconsistentes (DILLY, 1995; GONÇALVES, 2000). Se os erros não forem descobertos neste estágio, poderão contribuir para a obtenção de resultados de baixa qualidade (GILMAN apud LUBEL, 1998).

3. *Transformação* - consiste em desenvolver um modelo sólido de dados de maneira que possam ser utilizados por um algoritmo de extração de conhecimento. As transformações são ditadas pela operação e técnica a ser adotada. São conversões de um tipo de dados para outro, definição de novos atributos, etc. (GONÇALVES, 2000; IBM, 1996).

4. *Data Mining* - é o núcleo do processo. Aplicam-se algoritmos para extrair padrões dos dados ou gerar regras que descrevam o comportamento da base de dados (BERRY & LINOFF, 1997; DILLY, 1995). Para isto, utiliza-se uma ou mais técnicas para se extrair o tipo de informação desejada. Durante esse procedimento, pode ser necessário acessar dados adicionais e/ou executar outras transformações nos dados originalmente selecionados (IBM, 1996).

5. *Interpretação e avaliação* - significa validar o conhecimento extraído da base de dados, identificar padrões e interpretá-los, transformando-os em conhecimentos que possam apoiar as decisões (DILLY, 1995). O objetivo de interpretar os resultados é filtrar as informações que serão apresentadas aos tomadores de decisão.

Se os resultados não forem satisfatórios, faz-se necessário repetir a etapa de Data Mining ou retomar a qualquer um dos estágios anteriores. Somente após a avaliação e validação dos resultados é que se encontra conhecimento.

3.2 Conceituação de Data Mining

KDD refere-se ao processo completo de descoberta de conhecimento, enquanto que Data Mining é uma de suas etapas voltada a aplicar algoritmos específicos e a produzir padrões sobre uma base de dados (FAYYAD et al., 1996).

Data Mining - DM, ou Mineração de Dados, descende fundamentalmente da estatística clássica, da Inteligência Artificial e da *machine learning*. A primeira modalidade, a estatística clássica, é base da maioria das tecnologias a partir das quais DM foi construído. A segunda é a Inteligência Artificial, a qual tenta imitar a maneira do homem pensar sobre a resolução dos problemas estatísticos. A terceira, denominada *machine learning*, pode ser descrita como a união da estatística e da Inteligência Artificial (BUSINESS MINER, 1997). Estas três técnicas são usadas conjuntamente para estudar os dados e encontrar neles tendências e padrões, conforme mostra a Figura 5.

DM é uma tecnologia com grande potencial para auxiliar as organizações a extrair as informações mais importantes provenientes dos seus bancos de dados, predizendo padrões e comportamentos futuros,

respondendo a questões que tomariam muito tempo para serem resolvidas, o que possibilita as melhores decisões de negócio apoiadas em conhecimento. Para LUBEL (1998), DM é um recurso em ascensão que se tornará obrigatório aos mercados competitivos.

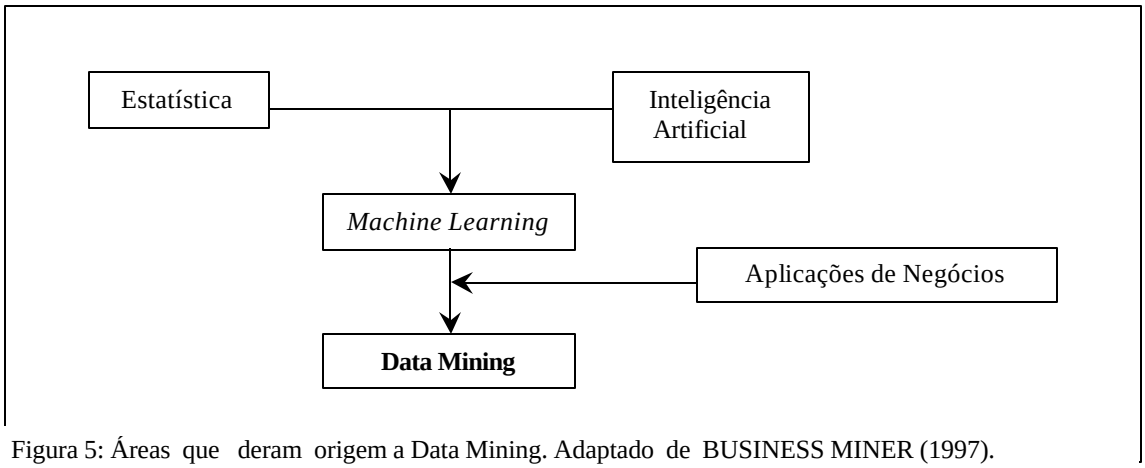


Figura 5: Áreas que deram origem a Data Mining. Adaptado de BUSINESS MINER (1997).

Na prática, os objetivos de DM são: predição e descrição. A *predição* envolve a utilização de algumas variáveis (atributos) da base de dados para predizer valores desconhecidos ou futuros de outras variáveis de interesse. A *descrição* procura por padrões que descrevem os dados interpretáveis (FAYYAD et al., 1996).

A seguir, são citadas algumas definições encontradas na literatura, que contextualizam DM:

a) Segundo BERRY & LINOFF (1997), DM é a exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados para descobrir padrões interessantes e regras. Envolve a transformação dos dados em informação, a informação em ação e a ação em valor.

b) DM é uma ferramenta que combina descobrimento com análise, utilizando um modelo para descobrir padrões passados que predizem comportamentos futuros. (LUBEL, 1998).

c) THE GARTNER GROUP (apud BERRY & LINOFF, 1997) definem DM como o processo de descobrir correlações significantes, padrões e tendências, através de filtragem de grandes quantidades de dados, pelo uso de tecnologias de reconhecimento desses padrões, bem como de técnicas estatísticas e matemáticas.

d) DM se refere ao uso de uma variedade de técnicas para identificar informações valiosas que podem ser usadas em áreas de apoio à decisão, predição e estimativa. Os dados geralmente são volumosos, mas de baixo valor para uso direto, pois, são as informações escondidas nestes dados que são úteis (CLEMENTINE USER GUIDE citado por DILLY, 1995).

e) Com o uso de softwares de DM é possível extrair informações importantes e em lugares inesperados, à medida que se extraem padrões aparentemente incompreensíveis ou tão óbvios que ninguém ainda os tenha notado (DILLY, 1995).

3.3 Métodos de Data Mining

Os métodos de DM podem ser classificados pela função que executam ou de acordo com a classe de aplicação em que podem ser usados (DILLY, 1995). Cada classe de aplicação tem como base um conjunto de algoritmos a serem utilizados na extração de relações relevantes de uma base de dados, diferindo uma das outras quanto aos tipos de problemas que o algoritmo será capaz de resolver.

Nesta sessão será apresentada uma breve introdução aos principais métodos de DM: Associação, Classificação, Clusterização e Previsão de Séries Seqüenciais/Temporais. A ênfase será no método de classificação empregado neste estudo.

3.3.1 Classificação

Essa modalidade é também conhecida como regras de classificação, indução supervisionada, aprendizado supervisionado ou processo direto.

Para classificar é necessário selecionar um atributo alvo, chamado variável dependente ou classe, cujo valor é usado para elaborar regras de classificação e as variáveis independentes ou atributos preditores (GROTH, 1998).

A classificação utiliza dados sobre o passado para encontrar padrões significantes de forma a induzir regras sobre o futuro, isto é, regras que predizem o valor do atributo alvo, através da combinação dos valores dos atributos preditores (BERRY & LINOFF, 1997; BISPO, 1998).

O processo inicia-se com um conjunto de treinamento e com os registros pré-classificados espera-se associar cada inclusão a um código de classe, fundamentado nos valores dos atributos preditores. O sistema deve inferir regras para classificar e encontrar a descrição da classe. Ao final do processo, tem-se um modelo da base de dados capaz de classificar um número maior de registros.

A precisão do resultado da classificação é medida pela **taxa de erro** que é o percentual de registros classificados incorretamente (BERRY & LINOFF, 1997).

Em geral, os algoritmos de classificação incluem as técnicas de árvore de decisão ou redes neuronais.

As aplicações para a classificação incluem análises de aprovação de crédito, definição de diagnóstico médico e efetividade de tratamento, determinação de alvos de campanhas de marketing, localização de lojas, etc. (AGRAWAL et al., 1996).

3.3.1.1 Metodologia para a Classificação

Na classificação de dados, o aprendizado vem de exemplos. O objetivo é analisar os dados e desenvolver um modelo ou descrição para uma classe. O modelo e a classe geram regras de classificação que serão aplicadas em dados futuros, cuja classe é desconhecida, ou então, para se entender melhor uma determinada classe (ALI et al., 1997).

Para desenvolver o sistema de extração de conhecimento, emprega-se uma metodologia. A Figura 6 apresenta a metodologia recomendada por BERRY & LINOFF, 1997.

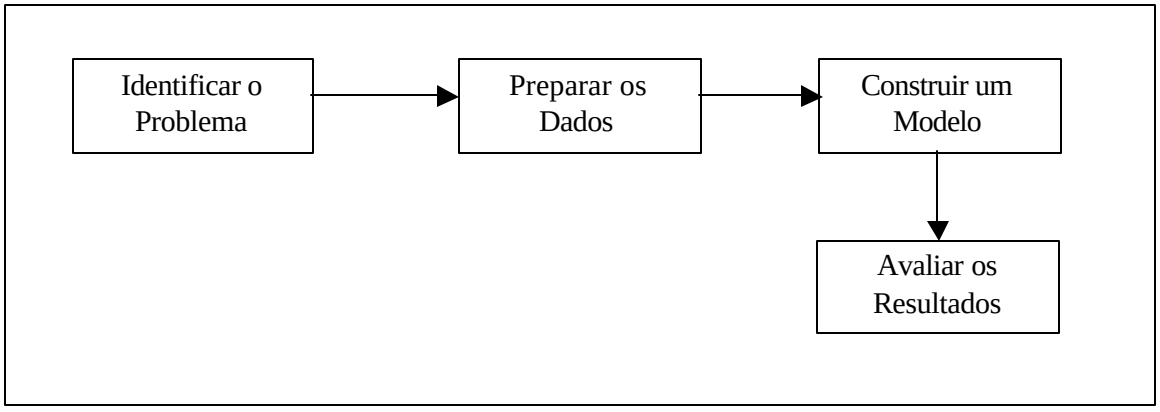


Figura 6: Metodologia para a extração de conhecimento. Adaptado de BERRY & LINOFF, 1997.

3.3.1.1.1 Identificação do Problema

Nesta etapa, procura-se identificar as características dos problemas e a as áreas dentro da organização onde a análise de dados pode prover valor, transformando-se em informações úteis.

Um especialista que conhece o negócio identifica as oportunidades de aplicar DM e planeja como deverá medir seus resultados para a organização, define o problema e as metas a serem atingidas. Nessa fase preocupa-se com critérios de desempenho, gargalos do domínio da aplicação e a interoperabilidade com o usuário final (GONÇALVES, 2000).

Segundo FAYYAD (1996), a busca por informações relevantes é realizada em três etapas: na primeira, decide-se se a respeito do processo, o qual pode ser de classificação, agrupamento ou sumarização. Na segunda, escolhe-se um dos métodos de busca de padrões e por último, a técnica a ser utilizada. Essa escolha depende do negócio, da aplicação e da qualidade dos dados disponíveis.

3.3.1.2 Preparação dos Dados

Nessa etapa planeja-se todas as atividades para se chegar ao ponto final de carga dos dados no ambiente de DM. A preparação dos dados vai de acordo com o algoritmo escolhido. Dependendo dessa escolha, os dados serão formatados de maneiras diferentes.

O primeiro requisito para que a classificação seja bem sucedida é possuir dados de qualidade. Isto implica limpeza e validação dos campos, tornando-os úteis ao processo, pois o modelo vai mostrar os acontecimentos passados. O cuidado na definição da classe alvo é outro requisito para o sucesso do método.

Para se construir o banco de dados para o DM, é preciso definir os grupos de dados e entender cada atributo. Esses grupos podem ser encontrados na organização ou serem provenientes de fontes externas (BARBIERI, 2001).

Deve-se tratar com dados provenientes de diferentes arquiteturas de computadores, múltiplas formas de representar a mesma coisa, dados no formato de texto, dados incompletos ou nulos. Em muitos casos, pode ser necessário criar novas variáveis derivadas das existentes.

Os dados são selecionados e passam por um processo de filtragem. Os atributos devem conter valores corretos e o conjunto selecionado possuir somente dados relevantes. Definem-se também as normas para o tratamento dos atributos, tais como, campos inválidos ou não preenchidos, atribuição de valores estatísticos, junção de variáveis, transformações e modificações na representação dos dados, como por exemplo, converter valores contínuos para discretos, datas para valores numéricos, etc.

Quando se está acostumado a usar técnicas puras de estatística, deve-se ter outro pensamento para começar a usar técnicas de DM. Ao invés de escolher cuidadosamente as variáveis independentes que se acredita serem importantes, as ferramentas de DM por si determinam as variáveis essenciais. Muitas vezes, uma variável é eliminada por não apresentar significância estatística, quando tem valor preditivo combinada com outras variáveis (BERRY & LINOFF, 1997).

A preparação dos dados costuma consumir mais de 50% do tempo e recursos destinados ao projeto e é essencial para o sucesso da aplicação (BARBIERI, 2001; BERRY & LINOFF, 1997).

3.3.1.3 Construção do Modelo

Envolve a escolha e aplicação de técnicas de DM sobre os dados

selecionados. Técnicas diferentes podem ser aplicadas para o mesmo problema, e por vezes, exigem formatos de dados diferentes, o que sugere prováveis retornos à fase de preparação.

A construção do modelo varia de técnica para técnica. Para a classificação, o conjunto de treinamento é usado para gerar uma explicação da variável alvo em relação às variáveis independentes. Essa explicação pode ser na forma de uma árvore de decisão, de rede neuronal ou de outra modalidade de relação entre a

variável que se deseja classificar e as demais variáveis da base de dados.

A classificação utiliza as ocorrências passadas para construir um modelo futuro. Para isto, é necessário dados pré-classificados, oriundos de dados históricos ou de um outro processo de descoberta de conhecimento.

Os dados pré-classificados são divididos em três partes. A primeira divisão representa o conjunto de treinamento aplicado para gerar uma explicação da classe alvo em função das variáveis preditoras, para construir um modelo inicial. A segunda, refere-se a série de testes utilizados para ajustar o modelo inicial, tornando-o mais geral e menos relacionado às características do conjunto de treinamento. A terceira representa o conjunto de avaliação aplicado para medir a efetividade do modelo quando os dados são desconhecidos. Os dados podem ser divididos em vários arquivos de treinamento para testar e avaliar cada conjunto. E, para otimizar o modelo, eliminam-se as regras que dependem inteiramente do conjunto de treinamento.

3.3.1.4 Avaliação do Modelo

Nesse passo, o modelo construído deverá ser criteriosamente avaliado visando a sua aplicação no problema sugerido. Objetiva determinar se algum conhecimento adicional foi descoberto ou se as hipóteses existentes foram confirmadas.

Um especialista define se as regras selecionadas no estudo agregam valores úteis à predição. A medida dos resultados se refere especificamente ao valor

para o negócio e se esse resultado pode ser usado no futuro. Deve-se identificar as informações úteis, sua incorporação aos processos de negócio e, mais importante, quem usará essas informações (BERRY & LINOFF, 1997).

Para conferir a performance do modelo, aplica-se uma estimativa à coleção final de registros pré-classificados. A taxa de erro do conjunto de treinamento é um bom preditor da taxa de erro dos demais dados.

Após este passo, fecha-se o ciclo de DM. Novas hipóteses podem ser formuladas, reiniciando o processo.

3.3.2 Associação

A associação ou afinidade de grupos visa a combinar itens importantes, tal que, a presença de um item em uma determinada transação pressupõe a de outro na mesma transação. Isto foi inicialmente proposto por AGRAWAL, em 1993.

As aplicações de técnicas de associação têm seu uso mais difundido na área de marketing, em que se pretende descobrir as associações existentes entre os produtos vendidos. A tecnologia possibilitou às organizações coletar e armazenar grandes quantidades de dados, como é o caso da tecnologia de código de barras sobre os dados de vendas (AGRAWAL, 1993). As grandes redes varejistas estudam as compras dos clientes para descobrir quais as vendas são normalmente realizadas ao mesmo tempo, chamando isso de *market basket analysis*. Essa análise pode determinar, por exemplo, os produtos que devem estar expostos juntos, objetivando incrementar as vendas (BUSINESS MINER, 1997).

A regra de associação é uma expressão representada na forma X => Y (X implica em Y) , em que X e Y são conjuntos de itens da base de dados; X é o antecedente da regra (lado esquerdo) e Y é o conseqüente da regra (lado direito) e pode envolver qualquer número de itens em cada lado da regra (DILLY, 1995). O significado desta regra é que as transações da base que contêm X tendem a conter Y. Um exemplo prático é

afirmar que "30% dos registros que contêm X também contêm Y; 2% dos registros contêm ambos" (AGRAWAL et al., 1997; AGRAWAL et al., 1993).

A regra de associação possui dois parâmetros básicos: o suporte e a confiança. Estes parâmetros limitam a quantidade de regras que serão extraídas e descrevem a qualidade delas.

Considerando que os conjuntos de itens X e Y estão sendo analisados, o **suporte** é definido como a fração de registros que satisfaz a união dos itens no conseqüente (Y) e no antecedente (X), correspondendo à significância estatística da regra (AGRAWAL et al., 1993).

A **confiança** é expressa pelo percentual de registros que satisfaz o antecedente (X) e o conseqüente (Y), medindo a força da regra ou sua precisão (AGRAWAL et al., 1993). No exemplo anteriormente citado, 30% é o fator de confiança e 2% é o suporte da regra.

BERRY & LINOFF (1997) definem a confiança como a freqüência com que o relacionamento mantém-se verdadeiro na amostra de treinamento e o suporte como a freqüência com que a combinação acontece. Assim, uma associação pode se manter 100% do tempo e ter a mais alta confiança, porém pode ser de pouca utilidade se a combinação ocorrer raramente.

Para AGRAWAL et al. (1997), o problema das regras de associação é encontrar todas as que possuem o suporte e a confiança acima de um determinado valor mínimo, pois, na prática os usuários normalmente estão interessados somente num subconjunto de associações.

Um dos algoritmos mais referenciados para este método é o *Apriori*, nas diversas variações, tais como, o *AprioriTid*, DHP e *Partition*.

3.3.3 Clusterização

É um exemplo de aprendizado não supervisionado ou indireto, cujo objetivo é agrupar tipos similares de dados ou identificar exceções (GROTH, 1998). O sistema tem que descobrir suas próprias classes, isto é, agrupar os dados e descobrir subconjuntos de objetos relacionados ao conjunto de treinamento, encontrando descrições de cada um destes subconjuntos (DILLY, 1995).

Um cluster pode ser definido como um conjunto de objetos agrupados pela similaridade ou proximidade e a clusterização como “a tarefa de segmentar uma população heterogênea em um número de subgrupos (ou clusters) mais homogêneos possíveis, de acordo com alguma medida” (BERRY & LINOFF, 1997; DILLY, 1995). Quando o processo é bem sucedido, os objetos do cluster têm alta homogeneidade interna e alta heterogeneidade externa. Um exemplo disso é a geração de clusters de sintomas de pacientes, que podem indicar diferentes doenças baseadas nas suas características.

Na clusterização, diferentemente da classificação, não há classes pré-definidas. Na classificação, a população é subdivida e associa cada registro a uma classe pré-definida, com base no modelo desenvolvido através de treinamento e exemplos pré-classificados. A clusterização é mais geral e freqüentemente realizada como primeira etapa de outros métodos de DM ou de modelagem. Assim, aplica-se o modo direto para reconhecer relações nos dados e o indireto para explicar estas relações (BERRY & LINOFF, 1997).

É aplicada em atividades de marketing para identificar os segmentos de mercado, para encontrar estruturas significantes nos dados e na descoberta de fraudes ou dados incorretos (GROTH, 1998).

3.3.4 Padrões Seqüenciais/Temporais

Este método procura eventos ou compras que ocorrem seqüencialmente em um período de tempo, determinando tendências (DILLY, 1995).

Uma aplicação típica é a venda por mala direta, que agrega os dados sobre os produtos adquiridos em cada compra. A descoberta de seqüência irá analisar este conjunto e detectar padrões de produtos comprados durante um determinado tempo. Pode ser útil também para identificar os itens que precedem a compra de um determinado produto (DILLY, 1995; IBM, 1996).

3.4 Técnicas de Data Mining

Para cada método de DM existe uma variedade de técnicas de extração de conhecimento. Selecionam-se as técnicas apropriadas de acordo com a característica dos dados e o objetivo a ser atingido.

Em alguns casos, pode ser necessário aplicar várias técnicas para se obter o melhor resultado (IBM, 1996). Entre estas técnicas, pode-se citar: algoritmos genéticos, redes neuronais, estatística multivariada, árvores de decisão, regras de associação e lógica difusa. Contudo, neste trabalho serão abordadas somente as técnicas referentes à Árvore de Decisão e Indução de Regras.

3.4.1 Árvore de Decisão

A árvore de decisão é uma ferramenta completa e bastante conhecida para classificar dados e apresentar os resultados sob a forma de regras (BERRY & LINOFF, 1997). A maioria das árvores de decisão executa a classificação em duas fases: construção da árvore e *prunning* (AGRAWAL et al., 1996):

1. Construção da árvore: a árvore vai se ramificando através de sucessivas divisões dos dados com base nos valores dos atributos. O processo é repetido recursivamente até que todos os registros pertençam a uma classe.

2. *Prunning* (poda): remove as ramificações que não tem valor significativo para criar o modelo de classificação, selecionando a sub-árvore que contém a menor taxa de erro estimada. 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, sendo desenhada com a raiz no topo e as folhas embaixo.

Um registro entra na árvore pelo nó raiz. Na raiz, é aplicado um teste para determinar o próximo nó onde o registro irá se posicionar. Há diferentes algoritmos para escolher o teste inicial, mas o objetivo é sempre o mesmo: escolher aquele que melhor descreve a classe alvo. O processo é repetido até que o registro chegue a uma folha, assim, todos os registros que terminam na mesma folha são classificados da mesma forma. Há somente um caminho da raiz até cada folha e este caminho é a expressão da regra usada para classificar os registros (BERRY & LINOFF, 1997).

A Figura 7, adaptada de AGRAWAL et al. (1996), descreve a situação para conceder créditos a clientes. Os atributos contêm informações sobre a idade e o salário; as classes são definidas como: *Bom*, para

designar os clientes que tem crédito aprovado e *Ruim* para os que apresentam risco na concessão de crédito. Diferentes folhas podem ter a mesma classificação, mas cada uma foi classificada por uma razão.

Idade	Salário	Classe
30	65	Bom
23	15	Ruim
40	75	Bom
55	40	Ruim
55	100	Bom
45	60	Bom

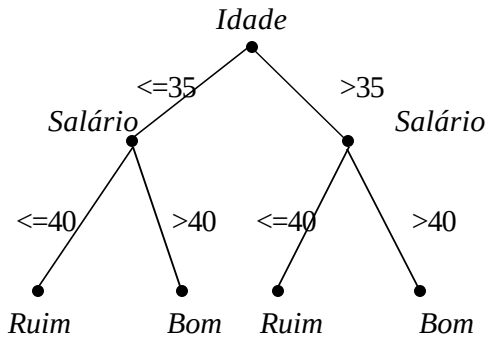


Figura 7: Exemplo de uma estrutura de árvore de decisão. Adaptado de AGRAWAL et al., 1996.

A árvore de decisão é vista como uma série de questões. A resposta da primeira questão determina qual a próxima questão a ser formulada. Se as questões são bem escolhidas, um caminho curto é suficiente para classificar com precisão um registro de entrada.

As principais vantagens da árvore de decisão, segundo BERRY & LINOFF (1997) são: a) facilitar compreender o modelo; b) permitir identificar os atributos chaves no processo; c) expressar facilmente as regras como instruções lógicas aplicadas diretamente aos novos registros. Acrescentam ainda como vantagens o fato das árvores de decisão serem relativamente mais rápidas quando comparadas às redes neuronais e, muitas vezes, obterem melhor precisão quando comparadas à outras técnicas de classificação (AGRAWAL et al.,1996) .

Há vários algoritmos para construir árvores de decisão. Os mais conhecidos são CART, CHAID (*Chi-Squared Automatic Interaction Detection*), ID3 (*Iterative Dichotomiser 3*) e C4.5.

3.4.1.1 Algoritmo C4.5

É uma evolução do algoritmo ID3 e um dos mais recentes algoritmos de árvore de decisão disponível. Foi desenvolvido pelo pesquisador Australiano J. Ross Quinlan em 1993 e está disponível em vários produtos comerciais. O algoritmo transforma a árvore de decisão em um conjunto de regras ordenadas

pela sua importância, permitindo ao usuário identificar, de imediato, os fatores que mais direcionam seus negócios (BERRY & LINOFF, 1997; BUSINESS OBJECTS, 1997).

O algoritmo produz uma árvore com um número variado de folhas por nó e assume os valores das categorias como um divisor, comportando-se diferentemente de algoritmos que produzem uma árvore binária, como o CART. O *prunning* é executado examinando a taxa de erro de cada folha, que somadas formam a taxa de erro da árvore.

Uma vez criado um conjunto de regras, o algoritmo agrupa as regras geradas para cada classe e elimina as que não contribuem para a precisão do conhecimento a ser extraído. O resultado final é um pequeno conjunto de regras de fácil entendimento, obtidas pela combinação das regras que levam à mesma classificação (BERRY & LINOFF, 1997).

3.4.2 Indução de Regras

O sistema de DM tem que inferir um modelo da base de dados, isto é, tem que definir as classes pelos atributos que denotam essa classe. Quando as classes são definidas o sistema deve inferir as regras que regem a classificação, ou seja, encontrar a descrição delas (DILLY, 1995).

Alguns algoritmos e índices executam esse processo, nesse contexto pode-se citar o Gini, o C4.5 e o CHAID. A maior parte do processo é realizado pelo computador e uma pequena parte pelo usuário (BUSINESS MINER, 1997).

> A tradução das regras para um modelo útil é feita pelo usuário, ou por uma interface de árvore de decisão. As regras são facilmente interpretadas, dada a sua modularidade, ou seja, pode ser entendida sem a necessidade de se referir a outras regras (DILLY, 1995).

3.5 Ferramentas de Data Mining

Atualmente existe uma grande variedade de produtos comerciais para DM. As ferramentas são apresentadas em pacotes comerciais e/ou encontradas na Internet.

Pesquisou-se alguns softwares que implementam o método de classificação e que podem ser aplicadas aos dados deste estudo.

3.5.1 A Ferramenta CBA

O CBA - *Classification-Based on Association* é uma ferramenta de DM desenvolvida pela School of Computing da National University of Singapore e foi apresentada na 4ª Conferência Internacional de Descoberta de Conhecimento e Data Mining, no ano de 1998 em Nova York, Estados Unidos (CBA, 1998), com o trabalho intitulado "*Integrating Classification and Association Rule Mining*" - Integrando Classificação e Regras de Associação de Mineração (LIU et al., 1998).

O CBA implementa a técnica de classificação baseada em associações, cujo objetivo é gerar subconjuntos de regras de associações, em que fica restrito ao lado direito das regras, o atributo alvo da classificação. Além de produzir regras de classificação, o CBA também pode ser aplicado para extrair regras normais de associação e categorização de textos (CBA, 1998). Estas funções podem ser observadas na Figura 8, onde se visualiza a tela inicial do sistema.

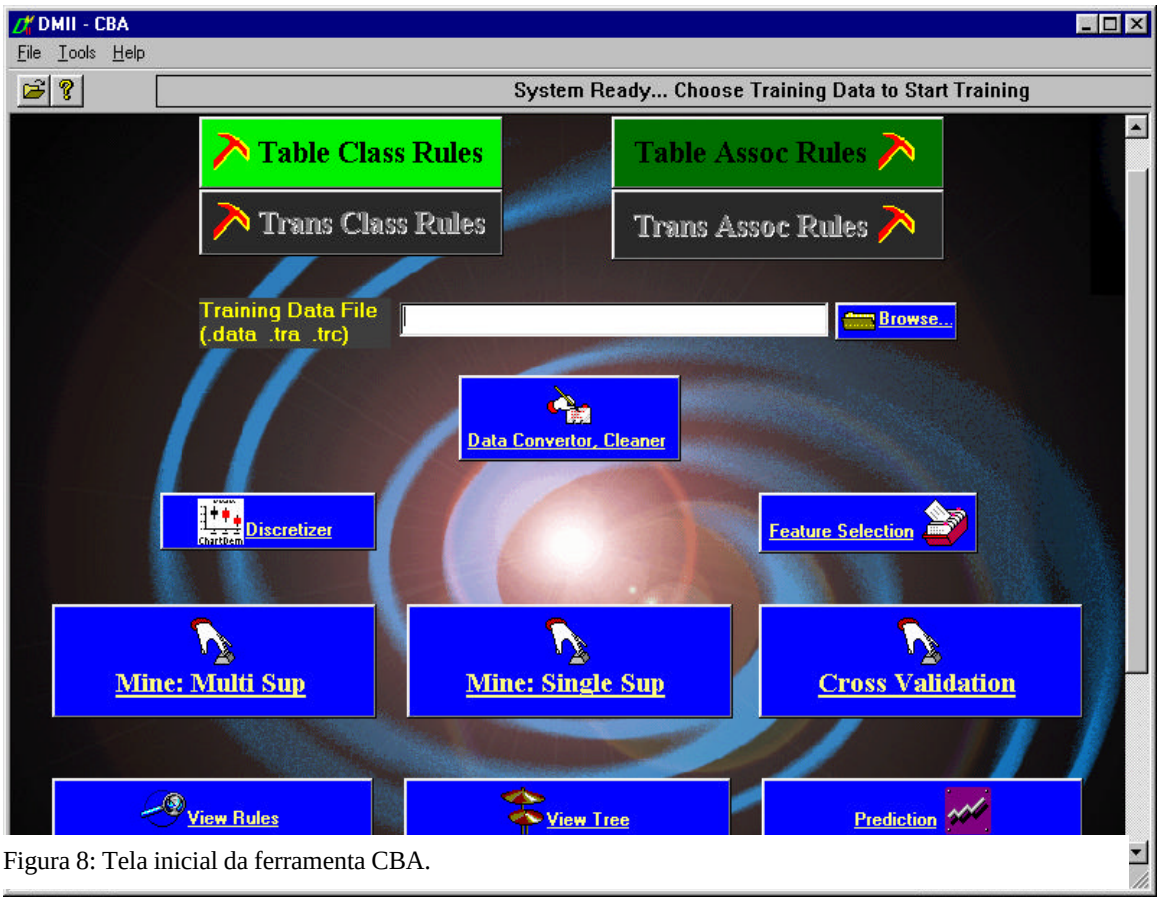


Figura 8: Tela inicial da ferramenta CBA.

Para LIU et. al (1999), o CBA apresenta as seguintes vantagens em relação aos métodos tradicionais de classificação:

- Unifica duas técnicas de DM: constrói a classificação através de um subconjunto de regras de associação, e, desta forma, produz regras mais precisas que os algoritmos puros;

- Abrange maior número de regras: é capaz de descobrir todas as regras que existem na base de dados, ao contrário dos sistemas de classificação tradicionais que produzem um conjunto pequeno de regras;

- Remove e sumariza as associações descobertas: em regras de associação muitas são redundantes ou não há ganho de conhecimento. O CBA elimina estas regras através da sua significância estatística, medida pelo Teste do Qui-Quadrado. O conjunto final contém um número consideravelmente reduzido de regras, pois as sem significância são removidas.

O CBA apresenta na sua arquitetura interna os algoritmos C4.5 e o Apriori. O C4.5 executa a classificação e o *prunning*, através da taxa de erro (CBA, 1998) e o algoritmo Apriori é aplicado para a geração de regras de associação (AGRAWAL, 1993).

A versão do software pode ser obtida pela Internet, sem custos. Não há restrições quanto ao número de registros para a execução do programa e nem em relação à quantidade de variáveis utilizadas.

3.5.1.1 Modelo de Dados na Ferramenta CBA

O programa utiliza dois arquivos: o arquivo de dados e o arquivo de definições. O primeiro, apresenta a extensão **.DATA*. Nesse, os dados selecionados para a aplicação devem estar limpos e compostos somente por valores discretos, para tanto, conta com as funções "*clean*" e "*discretize*r", respectivamente; o segundo, com a extensão **.NAMES*, contém o nome dos atributos preditorese com seus valores e a indicação da classe alvo. Ambos os arquivos são definidos no formato texto e devem possuir o mesmo nome.

As regras de classificação do CBA têm como base às associações entre as categorias das variáveis e cada regra gera um suporte e uma confiança expressos em valores percentuais e em valores absolutos. O *prunning* é executado examinando a taxa de erro de cada folha, que somadas formam a taxa de erro da árvore.

As opções de configuração podem ser alteradas, tais como: definir os valores mínimos do suporte e segurança, número total de regras a ser gerado, opção de executar ou não o *prunning*, etc.

As regras são apresentadas respeitando os valores mínimos de suporte e de confiança definidos pelo usuário e classificadas de acordo com o fator de confiança, que indica se a regra é forte, média ou fraca.

3.5.2 Outras Ferramentas de Data Mining

Os vendedores de produtos estão selecionando mercados verticalizados, atendo-se no tipo de serviço a ser oferecido: detecção de fraudes, gerência de telecomunicações, controle de manufaturamento, etc. Existe uma grande variedade de ferramentas de DM disponíveis no mercado. O Quadro 1 apresenta uma seleção de produtos que implementam a técnica de classificação.

Quadro 1 - Algumas Ferramentas Comerciais de Data Mining

Produto	**Fabricante**	**Características**
Intelligent Miner	IBM Corporation	É um conjunto de produtos para DM, inclui os métodos de classificação, associação, séries temporais/seqüenciais e clusterização. A IBM vende um conjunto diferente de algoritmos para resolver problemas separadamente. Trabalha em conjunto com banco de dados DB2, mas suporta outras fontes de dados (BISPO, 1998).
Clementine	Integral Solutions Ltd.	Utiliza árvore de decisão e rede neuronal para construir modelos de descoberta e executar predições (GROTH, 1998).
Scenario	Cognos Software	É uma solução que se integra a outras ferramentas da Cognos, o Powerplay e o Inpromptu. Baseia-se na técnica de árvore de decisão CHAID (GROTH, 1998).
SPSS CHAID	SPSS Inc.	Executa árvore de decisão para analisar e gerar modelos de predição, com diagramas de árvore de fácil entendimento. O SPSS também tem produtos para rede neuronal que provê modelagem e predição, séries temporais e clusterização (GROTH, 1998).
KnowledgeSeeker	Angoss International Limited	Utiliza técnicas de árvore de decisão (CHAID e CART) para construir modelos preditivos. Apresenta graficamente a árvore de decisão através de uma interface agradável, mostra a formação automática de todos os relacionamentos significantes. Os recursos iterativos permitem que os usuários explorem os dados, dividindo-os em nós selecionados na árvore ou forçando uma divisão particular que possa ser interessante (BISPO, 1998; GROTH, 1998).
See5	Rulequest Research	Executa a classificação, expressando-a como árvore de decisão ou como um conjunto de regras. Aceita qualquer número de atributos. A versão é disponibilizada gratuitamente na Internet, porém opera com o limite de 200 registros (RULEQUEST, 2000).

3.6 Aplicações de Data Mining

DM está sendo aplicado em uma variedade de áreas, entre elas, pode-se citar: vendas e marketing, bancos, saúde, telecomunicações, seguros, análise de vendas para promoções, análise de perfil e comportamento de consumidores, detecção de fraudes, etc. (IBM, 1996).

Como o presente trabalho está inserido no contexto da saúde, nas próximas seções apresenta-se alguns casos de aplicações nesta área.

3.6.1 Aplicações de Data Mining na Área da Saúde

BALLENGER (1999) cita algumas vantagens obtidas com o uso de DM em aplicações na saúde:

- Reduzir custos: respondendo a perguntas tais como, quais tratamentos são mais eficientes e, dentre esses, quais são os mais baratos;
- Melhorar a qualidade no atendimento, através da avaliação dos médicos e dos recursos empregados;
- Detectar fraudes: verificando se os médicos estão solicitando um número excessivo de exames e quais exames estão sendo solicitados por diagnóstico.

3.6.1.1 Extração de Conhecimento em Prontuários Médicos

O trabalho foi desenvolvido pela Universidade Católica de Pelotas e aplicado na Clínica Olivé Leite, em Pelotas - RS. O objetivo foi descobrir informações e padrões implícitos em textos da área médica, tais como prontuários, laudos, formulários de internações, entrevistas, etc.

Após a preparação dos dados para DM, utilizou-se o método de clusterização para fazer o agrupamento dos termos a fim de definir contextos. Posteriormente, aplicou-se a técnica de conjuntos difusos para determinar o grau de participação dos termos mais importantes em cada contexto.

As pesquisas em prontuários permitem aos médicos descobrir, por exemplo, quais pacientes tomaram determinados remédios, a que tratamento foram submetidos certos pacientes, o que tinham em comum aqueles que obtiveram alta, etc. (BAGATINI et al., 199-).

3.6.1.2 Aplicação de Data Mining para Estabelecer Padrões nos Tratamentos Clínicos

O projeto de DM foi implementado no Hospital da Flórida - Miami, com a finalidade de estabelecer padrões nas práticas clínicas. O software analisou os dados de tratamentos indicados por médicos e as despesas geradas .

Através deste sistema foi indicado, para cada doença, o tratamento que apresentou maior sucesso. Assim, pôde-se guiar a padronização de tratamento para diagnósticos específicos, diminuir a permanência dos pacientes no hospital e melhorar a qualidade dos serviços oferecidos (IBM, 2000).

3.6.1.3 Classificação de Cromossomos Humanos Usando Rede Neuronal Artificial

A análise de cromossomos é um procedimento fundamental para detectar anormalidades genéticas, no diagnóstico do câncer e no diagnóstico pré-natal de doenças citogenéticas. Essa análise, normalmente, é realizada através dos modelos convencionais da Teoria de Reconhecimento de Padrões, fundamentados nos princípios da estatística e da probabilidade. Neste trabalho aplicou-se, confrontando estes métodos, uma rede neuronal artificial com uma função de base radial para classificar cromossomos humanos em 24 classes, com o objetivo de diminuir a taxa de erro na classificação destes cromossomos. Com isso, obtém-se maior precisão e rapidez nos testes laboratoriais, tornando-os economicamente mais acessíveis.

Este trabalho foi desenvolvido no Programa de Pós-Graduação da Engenharia de Produção da Universidade Federal de Santa Catarina (TODESCO, 1995).

3.6.1.4 Data Mining na Indústria Farmacêutica

A empresa farmacêutica americana Merck-Medco utilizou DM, através do método de associação, para descobrir vínculos entre as enfermidades e os tratamentos realizados e definir os remédios mais efetivos para cada paciente, reduzindo, desta forma, o custo de cada tratamento (BISPO,1998; LUBEL, 1998).

3.6.1.5 Descoberta de Conhecimento em uma Base de Dados na Área Biomédica

Um problema que sempre existiu na biologia molecular é a predição da estrutura secundária de uma proteína a partir da sua estrutura primária. Utilizou-se DM para gerar regras que fornecessem uma descrição geral dos dados da base. Isso foi implementado através de um algoritmo de indução de regras que prediz múltiplas estruturas secundárias a partir dos dados sobre a estrutura primária das proteínas.

O estudo foi realizado no Departamento de Computação da Universidade Brunel, Inglaterra (ALNAHI & ALSHAWI, 1993).

3.6.1.6 Identificação de Padrões no Controle de Infecção Hospitalar

A análise foi realizada nos dados coletados durante um ano sobre uma determinada infecção hospitalar ocorrida nos pacientes do Hospital Birmingham da Universidade do Alabama. Através de regras de associação, foram identificados padrões das infecções e resistência antimicrobiana à bactéria *pseudomonas auriginosa*, o que permitiu traçar um programa efetivo de vigilância e controle (BROSSETE et al., 1998).

CAPÍTULO 4

CONCEPÇÃO DO MODELO PARA OS NASCIDOS VIVOS

Para gerar o modelo e extrair conhecimentos nos dados sobre a coorte de nascidos vivos, empregou-se a metodologia referenciada por BERRY & LINOFF (1997). A Figura 9 apresenta o esquema aplicado no estudo.

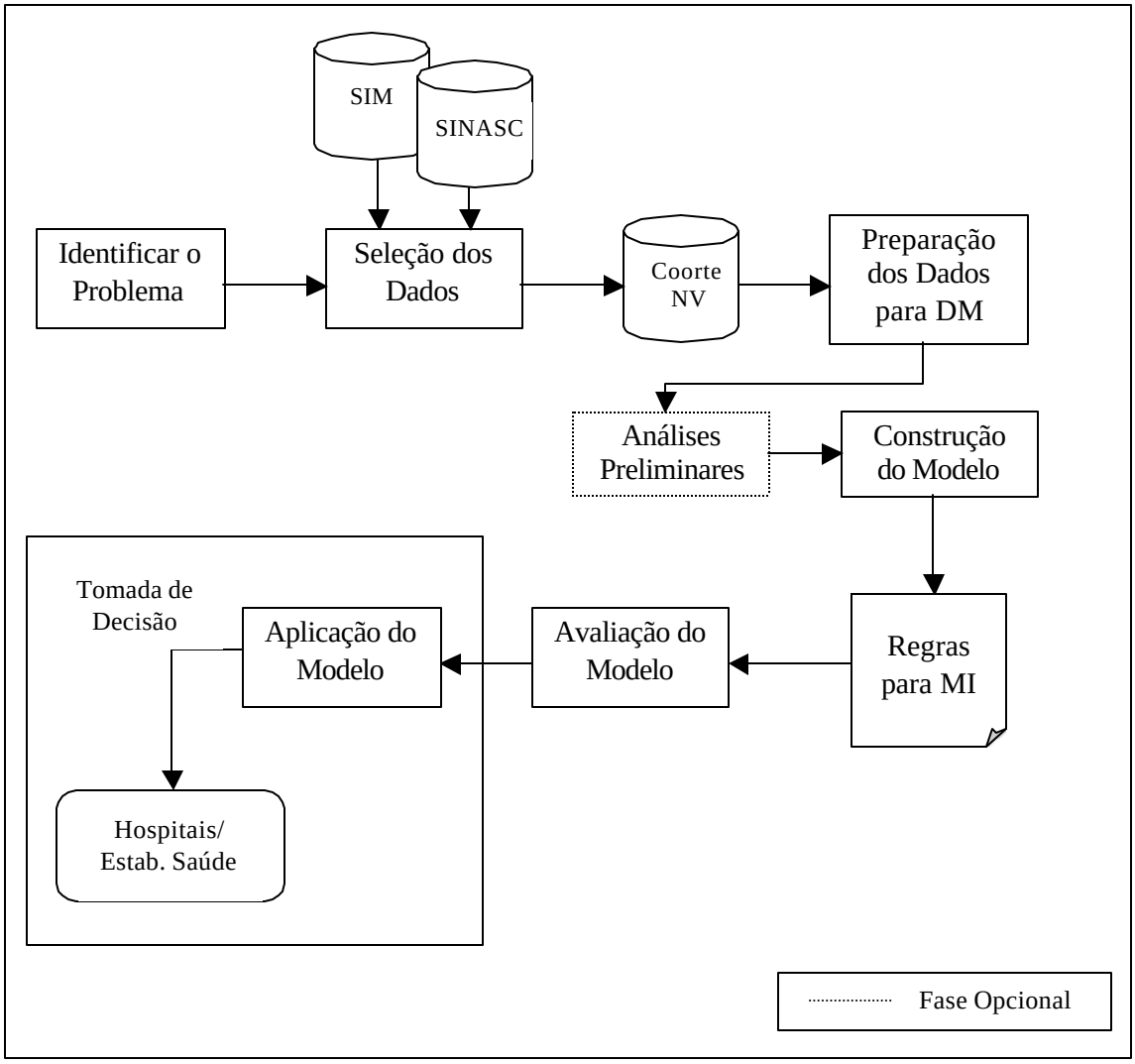


Figura 9: Modelo do Processo KDD para a coorte de nascidos vivos. Adaptado de BERRY & LINOFF, 1997.

4.1 Problema a ser tratado

O objetivo desta pesquisa é identificar os recém-nascidos com maior risco de morrer antes de completar um ano, analisando o conjunto de variáveis a que estão expostos, tendo como fonte de dados os sistemas SINASC e SIM. Para isso, aplica-se a técnica de classificação do processo de KDD, cuja variável alvo é a ocorrência do óbito no primeiro ano de vida, proveniente do arquivo de mortalidade.

Com a aplicação de técnicas de DM é possível ampliar o conhecimento sobre a mortalidade infantil, produzir regras que definam a ocorrência do óbito no primeiro ano de vida.

Com o modelo gerado na classificação, é possível elaborar um programa de prevenção para identificar, na gestação e parto, os recém-nascidos com risco de óbito futuro.

4.2 Base de Dados

O SINASC e o SIM apresentam-se em banco de dados dBASE. Apesar do SIM conter no seu documento de entrada o número da declaração de nascimento, para os casos de óbito fetal ou óbitos em menor que um ano, esta variável não é corretamente preenchida. Desta forma, não existe um mecanismo automatizado para relacionar as informações dos dois sistemas.

A ausência de um identificador universal, comum às bases de dados dos diferentes sistemas de saúde dificulta a integração das informações, pois é necessário apoiar-se em identificadores secundários fundamentados em variáveis comuns, nem sempre preenchidos corretamente, além de apresentar duplicidade de informações (JORGE MELLO, 1993). A localização dos dados nos dois sistemas, faz-se através de variáveis comuns, tais como:

idade da mãe, data de nascimento da criança, município de residência, endereço da mãe, etc.

Parte-se, então, de uma coorte de nascidos vivos e identifica-se, para cada declaração de óbito de menores de um ano, sua respectiva declaração de nascimento, relacionando os dados dos dois arquivos.

4.3 Seleção dos Dados

Com a integração dos arquivos de nascimento e de mortalidade, excluem-se as variáveis fora da área de pesquisa, por serem usadas com finalidades operacionais ou que não se aplicam a menores de um ano.

O conjunto de variáveis selecionadas permitirá extrair uma variedade de informações e revelar a situação do recém-nascido em relação à mortes no primeiro ano de vida, constituindo-se em um passo importante para se obter bons resultados na geração das regras .

4.4 Preparação dos Dados

Com os dados do SINASC e do SIM, relevantes ao estudo, procede-se, então, ao processamento, tornando-os limpos e consistentes.

A falta de comprometimento dos responsáveis pelo preenchimento dos dados da DN e da DO, acarreta dados incorretos ou, muitas vezes, não preenchidos. Faz-se necessário, então, uma avaliação desses dados e a respectiva correção.

É necessário também modificar à representação dos dados, ou seja, converter os valores contínuos para discretos, datas para valores numéricos, etc.

Define-se, então, a categorização das variáveis, seguindo-se a padronização estabelecida pelas entidades de saúde pública, em grupos que sejam mais representativos ao estudo.

4.5 Análises Preliminares

As análises preliminares compreendem as aplicações de testes estatísticos, como a distribuição de freqüência e o Teste do Qui-Quadrado.

Através da distribuição de freqüência pode-se observar a proporção dos indivíduos em cada categoria e o grau de preenchimento de cada variável.

O Teste do Qui-Quadrado (Anexo VI) é bastante utilizado em pesquisas sociais e de saúde para verificar possíveis associações entre variáveis categóricas. Desta forma, é possível determinar se uma variável qualitativa influi no comportamento de outra (SOARES, 1991). Através do teste, verifica-se a existência da associação entre o óbito de menores de um ano com os dados clínicos e epidemiológicos do nascimento e os dados sociais da mãe.

4.6 Construção do Modelo

Para aplicar DM a um conjunto inicial de variáveis, é preciso selecionar uma ferramenta que implemente o método de classificação e converter os dados para o formato padrão do software.

Pode ser necessário redefinir o conjunto de variáveis aplicadas ou até mesmo, preparar uma nova categorização das variáveis. O que se procura é o conjunto de variáveis preditoras que melhor define a ocorrência do óbito.

As regras de classificação são apresentadas de acordo com os parâmetros definidos pelo usuário, tais como: valores mínimos de suporte e de confiança, o número de regras, *prunning*, etc. As regras que apresentaram os maiores percentuais de confiança são selecionadas, e correspondem as regras mais fortes. Essas regras descrevem o óbito e geram um modelo da base de dados capaz de classificar outros registros.

4.7 Avaliação do Modelo

Nesta etapa, um especialista define, apoiado em conhecimentos técnicos, se as regras geradas para os óbitos agregaram valores úteis àpredição.

A elaboração de um modelo capaz de predizer o perfil dos que são mais afetados pelos fatores associados inerentes à mortalidade infantil é de grande importância

para a elaborar políticas públicas. Desta forma, o princípio da eqüidade pode ser melhor aplicado, haja vista que os mais necessitados podem ser identificados e receberem a atenção necessária.

Do ponto de vista populacional, a implantação deste modelo, tende a trazer consideráveis melhorias nos índices de mortalidade infantil, refletindo, assim, nos indicadores de qualidade de vida.

As regras geradas indicam as variáveis que estão mais associadas àmortalidade infantil. O analista seleciona as regras

relevantes, com base no fator de confiança e um especialista define, as regras que geram valores úteis a predição. Pretende-se assim, traçar o perfil dos recém-nascidos que devem receber uma assistência diferenciada, tanto em ações médicas quanto sociais.

CAPÍTULO 5

APLICAÇÃO

O trabalho propõe aplicar técnicas estatísticas e de DM nos dados do SINASC do município de Florianópolis, no segundo ano de implantação, juntamente com as informações sobre mortalidade infantil. O objetivo é identificar o conjunto de variáveis que levam a ocorrência do óbito e possibilite a geração de um modelo para auxiliar na sua prevenção.

Para isso, segue as etapas definidas no processo de descoberta de conhecimento (KDD), no qual DM está inserido.

5.1 A Base de Dados

O relacionamento (*linkage*) entre os arquivos de nascimentos e de mortalidade foi feito de forma manual, na Secretaria de Estado da Saúde de Santa Catarina, constituindo-se em novo arquivo contendo os registros dos dois sistemas. Deve-se observar que esse processo não é objeto do presente estudo.

A base de dados utilizada foi a do SINASC, versão 5.0, que se encontra em um banco de dados dBASE, referente aos nascidos vivos no ano de 1996, cujas mães são residentes em Florianópolis.

Foram localizados no SIM, os óbitos ocorridos de 1º de Janeiro de 1996 a 31 de Dezembro de 1997, através do nome da mãe, o endereço completo e o hospital de ocorrência (para os casos de óbito hospitalar). O novo arquivo, chamado "Coorte de

Nascidos Vivos (*coorte_nv*)", mantido em dBASE converter-se-á, posteriormente, para o formato padrão da ferramenta de DM aplicada no estudo.

O número de nascidos vivos em Florianópolis no ano de 1996 foi de 5.337, destes, foram localizados 81 óbitos de menores de um ano de vida, do total de 104. A localização dos registros depende do correto preenchimento da variável nome da mãe, tendo-se para este período uma perda de 23 registros.

5.2 Seleção dos Dados

As variáveis selecionadas do SINASC foram: *Sexo*, *Peso*, *Apgar no 1º* e no *5º minuto*, *Idade*, *Duração da Gestação*, *Escolaridade*, *Pré-Natal*, *Tipo de Parto*, *Tipo de Gravidez* e *Filhos Tidos*.

No SIM escolheu-se apenas duas variáveis que são interessantes ao estudo:

- *Óbito*: define a ocorrência ou não de óbito no primeiro ano de vida. É a classe alvo do estudo.

- *Data do Óbito*: informa a data da ocorrência do óbito.

A Tabela 2 descreve cada variável e seus valores categorizados.

5.3 Preparação dos Dados

As variáveis categorizadas e inicialmente aplicadas no estudo estão na Tabela 2. Utilizou-se a abreviatura "Categ." para distinguir as variáveis categorizadas das variáveis originais do banco de dados e "Classe" para definir a classe alvo do estudo.

Tabela 2: Conjunto de variáveis categorizadas.

Variável	Descrição	Valores
Categ_Sexo	Sexo da criança	Masculino Feminino
Categ_Peso	Peso da criança ao nascer	Baixo Peso = menor ou igual a 2.500 g Sobrepeso = maior que 2.500 g
Categ_Apgar1	Apgar no 1º minuto de vida	Baixo = de 1 a 3 Médio = de 4 a 7 Alto = de 8 a 10
Categ_Apgar5	Apgar no 5º minuto de vida	Baixo = de 1 a 3 Médio = de 4 a 7 Alto = de 8 a 10
Categ_Idade	Idade da mãe	Idade de risco = 19 anos ou menos ou idade superior a 35 anos Normal = Faixa etária de 20 a 35 anos
Categ_Gestação	Duração da gestação	Baixa = até 27 semanas Intermediária = de 28 a 36 semanas Alta = Mais que 36 semanas
Categ_Escolaridade	Escolaridade da mãe	Baixa = nenhuma instrução, 1º grau incompleto Intermediária = 1º grau completo ou 2º grau Alta = curso superior
Categ_PréNatal	Número de consultas pré-natal	Nenhuma consulta 1 ou mais consultas
Categ_Parto	**Tipo de parto**	Normal Outro = cesárea ou fórceps
Categ_Gravidez	Tipo de gravidez	Única Múltipla
Categ_Filhos	A mãe teve filhos anteriormente.	Sim = já teve filhos Não = sem filhos
Classe_Óbito	Ocorrência ou não de óbito	Vivo Óbito
Categ_IdadeÓbito	Idade da criança quando ocorreu o óbito	28 dias ou menos Mais que 28 dias

5.4 Análises Preliminares

As análises preliminares compreenderam a distribuição de freqüência e a aplicação do teste de associação estatística do Qui-Quadrado.

O software utilizado nesta etapa foi o Epi-Info versão 6.2.

5.4.1 Distribuição de Freqüência

Através da distribuição de freqüência pode-se observar a proporção dos indivíduos em cada categoria, conforme mostra a Tabela 3.

Dos 5.337 recém-nascidos, 81 morreram com menos de um ano de vida, representando uma freqüência de 1,5%. Desta forma, a ausência de óbito representa 98,5% dos casos.

Na variável pré-natal é marcante o número de casos ignorados ou não informados (34,5%), ocasionando uma perda de precisão ao avaliar os dados. Observa-se também que algumas variáveis concentram os dados em uma de suas categorias, tais como apgar no 1º minuto alto (83,6%), apgar no 5º minuto alto (80,8%), ausência de filhos tidos anteriormente (86,7%), sobrepeso (91,8%) e duração da gestação (86,9%).

O número de mortes de crianças com 28 dias ou menos, representa 71,6% dos óbitos.

Tabela 3: Distribuição de freqüência das variáveis categorizadas.

Variável	**Categoria**	**Freqüência**	**Percentual**
Categ_Sexo	masculino	2.709	50,8
	feminino	2.586	48,5
	ignorado	42	0,8
Categ_Peso	baixo peso	399	7,5
	sobrepeso	4.898	91,8
	ignorado	40	0,7
Categ_Apgar1	baixo	130	2,4
	médio	612	11,5
	alto	4.462	83,6
	ignorado	133	2,5
Categ_Apgar5	baixo	59	1,1
	médio	123	2,3
	alto	4.312	80,8
	ignorado	843	15,8
Categ_Idade	risco	1.348	25,3
	normal	3.784	70,9
	ignorado	205	3,8
Categ_Gestação	baixa	23	0,4
	intermediária	245	4,6
	alta	4.635	86,9
	ignorado	433	8,1
Categ_Escolaridade	baixa	1.970	36,9
	intermediária	2.145	40,2
	alta	676	12,7
	ignorado	545	10,2
Categ_PréNatal	nenhuma	139	2,6
	1 ou mais	3.359	62,9
	ignorado	1.839	34,5
Categ_Parto	normal	2.825	52,9
	outro	2.403	45,0
	ignorado	109	2,0
Categ_Gravidez	única	5.067	94,9
	múltipla	107	2,0
	ignorado	163	3,1
Categ_Filhos	sim	712	13,8
	não	4.625	86,7
Classe_Óbito	óbito	81	1,5
	vivo	5.256	98,5
Categ_IdadeÓbito	mais que 28 dias	58	71,6
	28 dias ou menos	23	28,4
	vivo	-	-

A distribuição de freqüência dos dados do SINASC é visualizada graficamente nas Figuras 10 e 11.

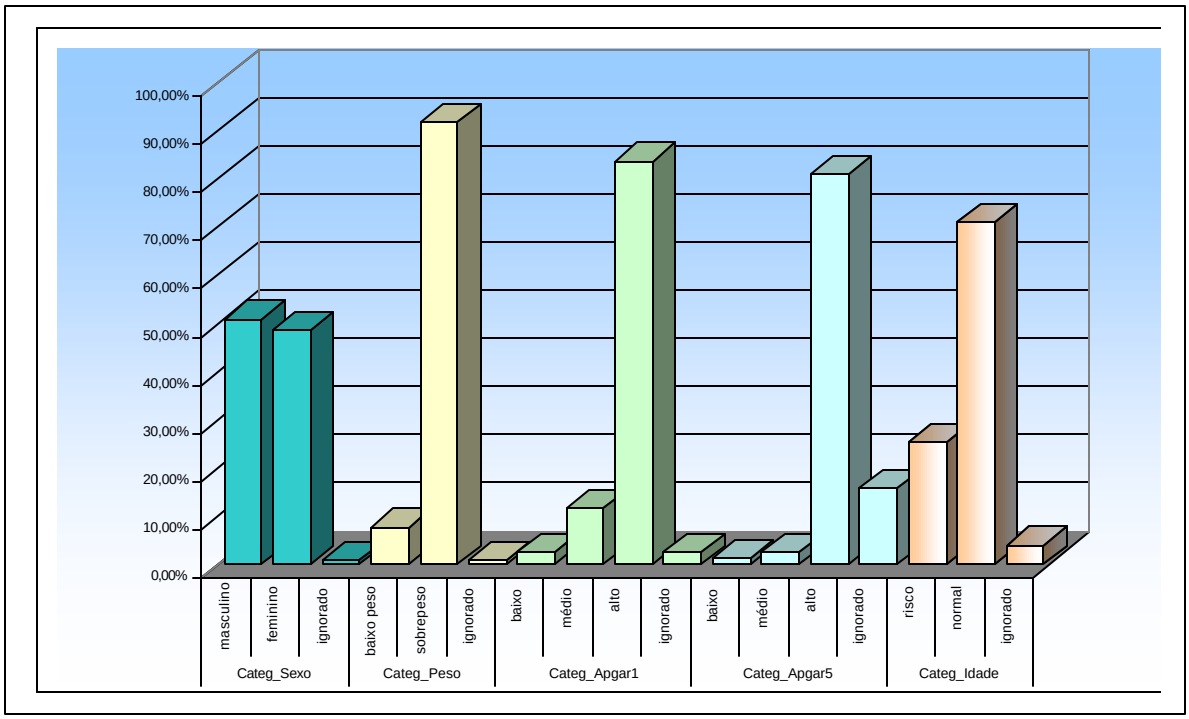


Figura 10: Distribuição de Freqüência Percentual das Variáveis: Sexo, Peso, Apgar1, Apgar5, Idade.

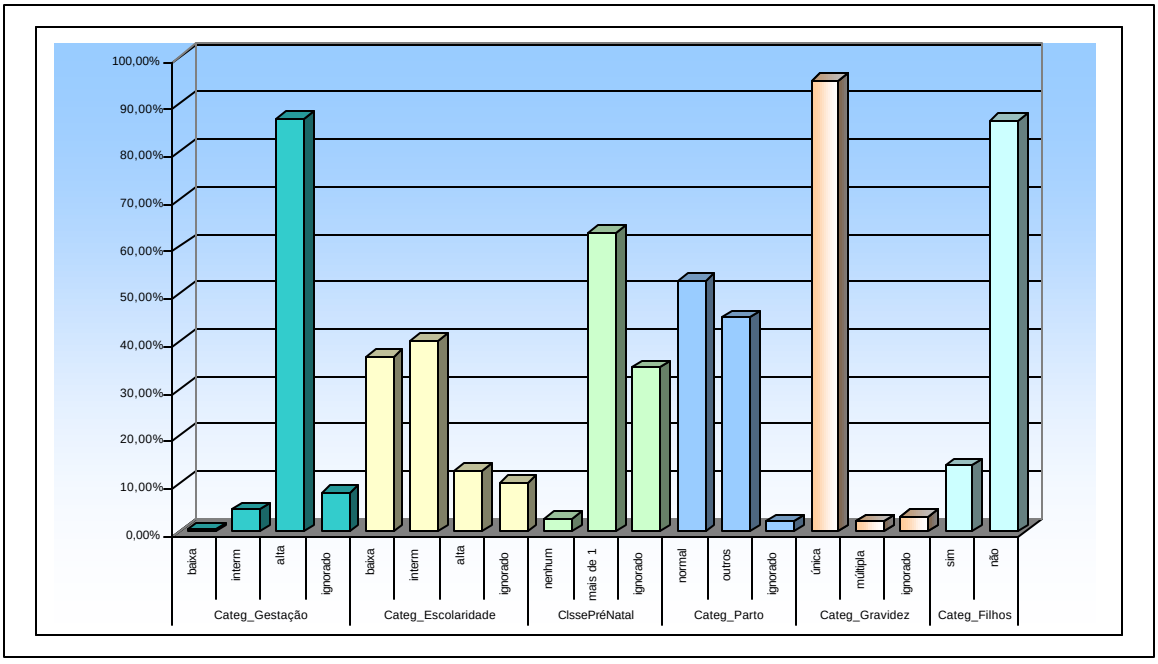


Figura 11: Distribuição de Freqüência Percentual das Variáveis: Gestação, Escolaridade, Pré-natal, Tipo de parto, Tipo de Gravidez, Filhos.

5.4.2 Aplicação do Teste do Qui-Quadrado

Considerando-se a ocorrência ou não do óbito como variável alvo (Classe_Óbito) e as variáveis categorizadas (Tabela 2), aplicou-se o Teste do Qui-Quadrado (Anexo VI) como medida de efeito associativo, com um nível de significância de 95%. O resultado pode ser visualizado na Tabela 4.

As variáveis que tem significância estatística (p<0,05) foram: peso, apgar no 1º e no 5º minuto, duração da gestação e tipo de gravidez. As variáveis que não apresentaram associação estatística com óbito foram: sexo, idade e escolaridade da mãe da mãe, pré-natal, tipo de parto e filhos tidos anteriormente.

Tabela 4: Estudo de associação da classe óbito com as demais variáveis.

Variável	**Teste Estatístico** (1)	**Nível de Significância** (2)
Categ_Sexo	Qui-Quadrado	não significante (p=0,355)
Categ_Peso	Exato de Fisher	muito significante (p=0,000)
Categ_Apgar1	Qui-Quadrado	muito significante (p=0,000)
Categ_Apgar5	Exato de Fisher	muito significante (p=0,000)
Categ_Idade	Qui-Quadrado	não significante (p=0,798)
Categ_Gestação	Exato de Fisher	muito significante (p=0,000)
Categ_Escolaridade	Qui-Quadrado	não significante (p=0,092)
Categ_PreNatal	Exato de Fisher	não significante (p=0,141)
Categ_Parto	Qui-Quadrado	não significante (p=0,774)
Categ_Gravidez	Exato de Fisher	muito significante (p=0,001)
Categ_Filhos	Qui-Quadrado	não significante (p=0,222)

(1) Utilizou-se o Teste Exato de Fisher para freqüência esperada menor que 5.
(2) Níveis de significância:
- não significante: p>0,05
- significante: p<0,05
- muito significante: p<0,01

5.5 Construção do Modelo

Devido ao alto custo dos sistemas de DM, optou-se pela ferramenta CBA, a qual é disponibilizada gratuitamente na rede de computadores e não limita esta aplicação.

5.5.1 Modelo de Dados na Ferramenta CBA

Com os dados relevantes à descoberta de conhecimento, procedeu-se a modelagem para o formato da ferramenta CBA. Assim, gerou-se os arquivos: *coorte_dn.names* (Anexo VII) e *coorte_dn.data* (Anexo VIII). O primeiro contendo a definição das variáveis com suas categorias e a indicação da classe alvo do estudo (óbito) e o segundo, com os dados no formato texto.

Com os arquivos formatados, selecionou-se o método de classificação e definiu-se os parâmetros iniciais para executá-lo, como ilustra a Figura 12. Inicia-se a fase de testes, selecionando o método de classificação. Ao final, vários modelos foram gerados, alterando-se, em cada um, o conjunto de variáveis categorizadas.

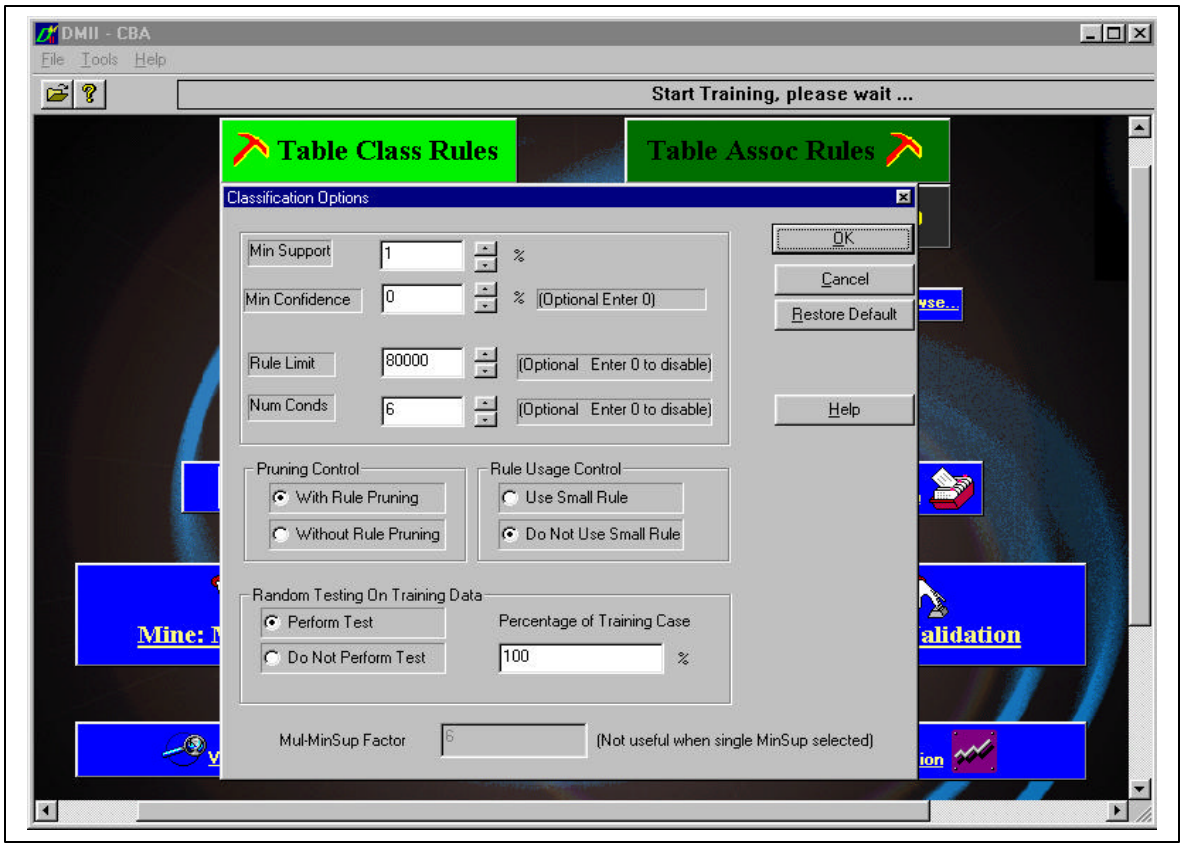


Figura 12: Definição dos parâmetros iniciais na ferramenta CBA.

Como parâmetros iniciais à execução da classificação foram mantidos os valores *default* do software, não se definindo limites mínimos para o suporte e a confiança, em virtude da freqüência do óbito ser muito baixa, 1,5 % dos registros, o que restringiria a formação de regras para o óbito.

5.5.2 Variáveis Relevantes

Algumas variáveis, apesar da significância estatística com o óbito, estavam presentes em regras de difícil interpretação ou que diminuíram consideravelmente a confiança da regra. Essas variáveis apresentaram um baixo grau de preenchimento e

ocasionaram desvios na análise (Tabela 3). Algumas variáveis que não apresentaram significância na associação estatística, quando combinadas com outras variáveis, apresentaram bons resultados na classificação, sendo então mantidas no estudo.

As variáveis: apgar no 1º minuto, apgar no 5º minuto, peso da criança, duração da gestação, idade da mãe e filhos tidos, estão presentes nas regras de classificação mais influentes para a descoberta de conhecimento e, por conseqüência, apresentaram os maiores valores de confiança.

5.5.3 Regras Geradas

O modelo final gerou onze regras de classificação (Anexo IX), algumas com alta similaridade, variando apenas a composição das variáveis. A taxa de erro foi de 1,16%, indicando o percentual de registros incorretos.

As regras selecionadas apresentaram alto grau de confiança, tanto para óbitos, quanto para vivos. Para óbitos, a confiança ficou entre 64,706% e 83,333%, sendo que o valor esperado era 1,5%. Para vivos, a

confiança ficou entre 99,643% e 99,711%, quando se esperava 98,5%.

O suporte das regras para óbito foi baixo, ficando entre 0,543% e 0,956%. Isso se deve ao baixo número de registros desta categoria, 81 casos em 5.337 (1,5% dos registros). Para vivos, os suportes ficaram altos, entre 64,849% e 78,640% (98,5% dos registros do banco de dados).

As regras que apresentaram os maiores percentuais de confiança foram selecionadas, sendo 3 regras para óbitos e 3 para vivos. Essas regras fornecem as melhores informações para o modelo em estudo e são mostrados na Tabela 5.

As regras que definem os **óbitos** no primeiro ano de vida são:

A primeira regra afirma que, se a idade da mãe é de risco, o peso da criança é baixo, o apgar no apgar no 1º minuto é baixo e o apgar no 5º minuto é médio, ocorre óbito. Neste caso, o suporte é de 0,112% (6 casos) e a confiança é de 83,333%.

A segunda mostra que, se o peso da criança é baixo e o apgar no 1º e no 5º minuto são baixos, então há óbito. Esta regra apresentou um suporte de 0,543% (29 registros em 5.337) e uma confiança de 82,759%.

A terceira afirma que, se o apgar no 1º minuto é baixo e o peso da criança é baixo, há óbito. Obteve um suporte de 0.956% (51 registros) e uma confiança de 64,706%.

Para confirmar os resultados, utilizou-se também o oposto da regra, caracterizando as crianças que não morreram no primeiro ano de vida.

As regras selecionadas em que **não ocorreram óbitos** são:

A quarta regra mostra que, se a mãe não teve filhos anteriormente, o peso da criança é alto e o apgar no 5º minuto é alto, não há óbito. Isto ocorre com um suporte de 64,849% (3.461 em 5.337 casos) e uma confiança de 99,711%.

A quinta regra afirma que, se o apgar no 5º minuto é alto e o peso da criança é alto, não ocorre óbito. Apresentou um suporte de 75,155% (4.011 casos) e uma confiança de 99,676%.

A sexta regra apresenta uma relação onde o apgar no 1º minuto é alto, o peso é alto e não há óbito. O suporte é de 78,640% (4.197 registros) e a confiança da regra é de 99,643%.

Tabela 5: Regras selecionadas para o modelo em estudo.

Nº	Óbito	**Regra**	**Suporte (%)**	**Confiança (%)**
1	Sim	Se a idade da mãe é de risco, o peso da criança é baixo, o apgar no 1º minuto é baixo e o apgar no 5º minuto é médio, então ocorre óbito.	0,112	83,333
2	Sim	Se o peso da criança é baixo, o apgar no 1º e no 5º minuto são baixos, então há óbito.	0,543	82,759
3	Sim	Se o apgar no 1º minuto é baixo e o peso da criança é baixo, então há óbito.	0,956	64,706
4	Não	Se a mãe não teve filhos anteriormente, o peso da criança é alto e o apgar no 5º minuto é alto, então não há óbito.	64,849	99,711
5	Não	Se o apgar no 5º minuto é alto e o peso da criança é alto, então não ocorre óbito.	75,155	99,676
6	Não	Se apgar no 1º minuto é alto e o peso da criança é alto, então não há óbito.	78,640	99,643

5.6 Avaliação do Modelo

As regras de classificação baseadas em associação comprovaram a existência de um forte relacionamento entre o óbito e as variáveis peso, apgar no 1º minuto e apgar no 5º minuto. Quando essas assumiram valores baixos, a ocorrência de óbito foi alta.

Confirmou-se também a associação do óbito com as mães com idade de risco (faixa etária abaixo dos 19 anos ou superior a 35). Esse dado associado ao baixo peso e apgar no 1º e 5º minuto com valores baixos, indicam gravidez de risco.

A data da morte é uma variável importante, pois dos 81 óbitos registrados, 58 ocorreram com menos de 28 dias, reforçando a teoria que as variáveis peso, apgar no 1º e 5º minuto com valores baixos, associadas à idade da mãe de risco, tem alta probabilidade de ocorrência de morte antes do primeiro mês de vida.

5.7 Considerações Finais

A pesquisa mostrou, por meio do Teste do Qui-Quadrado, as variáveis de maior associação estatística com o óbito, e, desta forma, agravantes às taxas de mortalidade infantil. As variáveis apgar no 1º minuto, apgar no 5º minuto e peso do recém-nascido se mostraram mais fortemente associadas e, assim, mais presentes na elaboração das regras de classificação.

As regras que definem o óbito e as que não ocorreram óbitos apresentaram similaridades opostas. Desta maneira, concluiu-se que, apesar do número reduzido de registros para os casos de óbitos, as regras foram bastante elucidativas.

CAPÍTULO 6

CONCLUSÕES E RECOMENDAÇÕES

Este estudo objetivou a aplicar Data Mining na produção de conhecimento, através de regras de classificação baseadas em associação com banco de dados do SINASC e a informação sobre óbito no primeiro ano de vida, no município de Florianópolis, no ano de 1996.

Em virtude da falta de dados precisos sobre a distribuição da mortalidade infantil, muitas das ações coletivas de promoção de saúde não são bem orientadas. Desta forma, esta questão não vem recebendo a atenção necessária no âmbito preventivo, caracterizando um problema de saúde pública.

São apresentadas nesta seção as conclusões e as recomendações de projetos futuros.

6.1 Conclusões

Este estudo confirma que a tecnologia de Data Mining é adequada para produzir conhecimento a partir de bases de dados da área da saúde. É um instrumento para avaliar o planejamento e as políticas de saúde pública.

A pesquisa apresenta a aplicação da técnica de classificação baseada em associações do processo de KDD (*Knowledge Discovery in Database*) para traçar o perfil dos recém-nascidos e identificar as variáveis biológicas e socioeconômicas que mais interferem na mortalidade infantil.

Os resultados revelam a associação estatística significante entre o óbito de menores de um ano e as variáveis peso ao nascer, os valores do apgar no 1º e no 5º minuto de vida, a duração da gestação e o tipo de gravidez

As seis regras de classificação selecionadas mostram que, em todas elas estão presentes ao menos uma das variáveis que apresentaram associação estatística.

DM, além de confirmar as hipóteses do teste de associação, serve também para e determinar o perfil dos nascidos vivos de maior risco.

Busca-se, com esse procedimento, alertar para as variáveis com alto grau de associação na ocorrência de óbito, e assim, monitorar os recém-nascidos que devem receber assistência efetiva e contribuir, desta forma, para reduzir a mortalidade infantil.

6.2 Recomendações para Futuros Trabalhos

As ferramentas de Data Mining utilizadas como instrumento de apoio àtomada de decisões em ações de saúde pública, não tem uma tradição histórica. Justifica-se então, a sua utilização em trabalhos futuros pela potencialidade do método.

A existência de poucos trabalhos semelhantes dificulta comparações. Sugere-se novos estudos para se confirmar os resultados desta pesquisa.

Outros métodos de Data Mining podem ser empregados nesta base de dados para auxiliar o processo de geração de conhecimento, como por exemplo, a análise de cluster, em que se formam agrupamentos que facilitam visualizar os grupos de risco.

Recomenda-se ainda que os registros das declarações de óbitos sejam codificados de

forma conjugada às declarações de nascidos vivos para integrar as bases de dados dos dois sistemas, SIM e SINASC, e assim, facilitar o cruzamento dos dados das duas bases com conseqüente ganho de informações que, se agregadas, permitem estudos com maior abrangência.

Com o modelo gerado na classificação é possível implantar um sistema de vigilância, em hospitais ou em outros estabelecimentos onde ocorrem os partos, para auxiliar na tomada de decisões e prevenir óbitos no primeiro ano de vida.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, Rakesh, IMICLINSKI, Tomasz, SWAMI, Arun. Mining association rules between sets of items in large databases. In: **Proccedings of the ACM SIGMOD Conferenc**e, Washington, D.C., May, 1993. Disponível na Internet: http://www.cs.bham.ac.uk/~anp/bibtex/kdd.bib.html Acessado em 20 de abril de 2001.

AGRAWAL, Rakesh, MEHTA, Manish, RISSANEN, Jorma. SLIQ: a fast scalable classifier for data mining. In: **Fifth In'tl Conference on Extending Database Technology**, Avignon, France, Mar, 1996.

AGRAWAL, Rakesh, SRIKANT, Ramakrish, VU, Quoc. Mining association rules with item constraints. In: **Future Generations Computer System**, Elsevier, Netherlands, v. 13, n. 2-3, Nov, 1997, p. 161-80.

ALI, Kamal, MANGANARIS, Stefanos, SKIRANT, Ramakrishnan. Partial classification using association rules. In: **Third International Conference on Knowledge Discovery and Data Mining**, Newport Beach, California, Aug, 1997.

ALNAHI, H., ALSHAWI, S. Knowledge discovery in biomedical databases - a machine induction approach. **Journal Computer Methods and Programs in Biomedicine**. v. 39. 1993. Issue 3-4. p. 343-349.

BAGATINI, Daniela, LOH, Stanley, GASTAL, Cláudio. **Correção ortográfica em prontuários médicos**. Universidade Católica de Pelotas, [199-] http://esin.ucpel.tche.br/napi/sidi/public.html 05 de Julho de 2000.

BALLENGER, Kitti. **Data mining presentation**. White Paper, 1999. http://www.ils.nuc.edu/DataMining/OurClassPage.htm 03 de Julho de 2000.

BARBETTA, Pedro A. **Estatística aplicada às ciências sociais**. Florianópolis: UFSC, 1994.

BARBIERI, Carlos. **BI - business inteligence:** modelagem & tecnologia. Axcel Books do Brasil, Rio de Janeiro, 2001, p. 177-212.

BEAGLEHOLE, R. et al. **Epidemiologia básica**. São Paulo, Santos, 1996. 176p.

BERRY, Michel J. A., LINOFF, Gordon. **Data mining techniques for marketing, sales, and customer support**. John Wiley & Sons, New York, 1997, 454 p.

BISPO, Carlos Alberto Ferreira. **Uma análise da nova geração de sistemas de apoio à decisão**. São Carlos, 1998. Dissertação (Mestrado em Engenharia de Produção) - Escola de Engenharia de São Carlos, Universidade de São Paulo.
http://cazarini.cnp.ecsc.sc.usp.br/Bispo/Di 05 de Julho de 2000.

BRASIL, Lei nº 8.069, de 13.07.90. Dispõe sobre o Estatuto da Criança e do Adolescente. **Diário Oficial**, n. 135, de 16.07.90, p. 13563 – 13577, Seção I.

BRASIL, Leis e Decretos. **Registros Públicos.** Lei 6015, de 31 de dezembro de 1973. São Paulo, Atlas, 1976.

BROSSETE, Stephen E. et al. Association rules and data minig in hospital infection control and public health surveillance. **Journal American Medical Information Asssocitation**, july, 1998, 5:4, p. 373-81.

BUSINESS OBJECTS. **Introducing Business Miner:** business mining for decision support insights. White Paper, 1997.

CAMPANHA NACIONAL DE REGISTRO CIVIL. Ministério da Saúde, 1997.
http://www.saude.rn.gov.br/campanha_registro_civil.htm 06 de Fevereiro de 2001.

CANUTO, Anne Magály de Paula, GOTTGTIOY, Márcia de Paiva Bastos. Data mining: geração de dados com qualidade para sistemas agropecuários. In: **AGROSOFT97**, 1997.
http://www.agrosoft.com/ag97/papers/c3a1100.html 06 de Julho de 2000

CBA Overview. White Paper, 1998.
http://www.csqueensu.ca/home/yu/cba.html 23 de Novembro de 2000.

DAWSON-SAUNDERS, Beth, TRAPP, Robert G. **Biostatistics basic & clinical**. Norwalk: Appleton & Lange, 1993.

DEAN, Andrew G., DEAN, Jeffrey A., SMITH, Richard C. et al. **Epi-Info, Version 6:** a word processing database, and statistics program for epidemiology on microcomputers. Centers of Disease Control and Prevention, Atlanta, Georgia, U.S.A., 1994.

DILLY, Ruth. **Data Mining -** an introduction. Parallel Computer Centre - Queen's University of Belfast. Dezembro, 1995. http://www.pcc.qub.ac.uk/tec/coursers/datamining/stu_notes/dm_book_2.html 10 de Julho de 2000.

DW Brasil. White Paper. http://www.dwbrasil.com.br/html/dmining.html 10 de Julho de 2000.

FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory, SMYTH, Padhraic. From data mining to knowledge discovery: an overview. In: **Advances in Knowledge Discovery and Data Mining**, AAAI Press / The MIT Press, MIT, Cambridge, Massachusets, and London, England, 1996, p.1-34.

FELDENS, Miguel A. **Descoberta de conhecimento aplicada à detecção de anomalias em base de dados**. Porto Alegre: PPGCC da UFRGS, 1996. (Trabalho Individual).

FREITAS, Alex A. Understanding the crucial differences between clasification and discovery of association rules - a position paper. In: **SIGKDD Explorations**, Jul, 2000, v. 2, Issue 1.

FURQUIM, Márcia, JORGE, Maria Helena de Melo. O uso da técnica de "linkage" de sistemas de informação em estudos de coorte sobre mortalidade neonatal. **Revista de Saúde Pública**, v. 30, n. 2, p. 141-7, 1993.

GONÇALVES, Alexandre Leopoldo. **Utilização de técnicas de mineração de dados na análise dos grupos de pesquisa no Brasil**. Florianópolis, 2000. Dissertação (Mestrado em Engenharia de Produção) - Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

GROTH, Robert. **Data mining:** a hands-on approach for business professionals. Prentice Hall, New Jersey, 1998.

HERRCHEN, Beate et al. Vital statistics linked birth/infant death and hospital discharge record linkage for epidemiological studies. **Computers and Biomedical Research**, n. 30, p. 290-305, 1997.

IBGE. **Criança e adolescelentes:** indicadores sociais, 1997. Rio de Janeiro, 1999. v.6, p. 30-41.

IBGE. **Evolução e perspectivas da mortalidade infantil no Brasil**. Rio de Janeiro, 1999. 45 p. (Estudos e Pesquisas. Informação Demográfica e Socioeconômica, n. 2).

IBGE. **Síntese dos indicadores sociais 2000**. Rio de Janeiro, 2001, p. 47-57. (Estudos e Pesquisas. Informação Demográfica e Socioeconômica, n. 5).

IBM. **IBM'S data mining technology**. White Paper, 1996.

IBM. **Healthcare software solutions:** Florida Hospital improves operations with DB2 Intelligent Miner, 2000.
http://www-4.ibm.com/software/data/bi 17 de Abril de 2001.

JORGE, Maria Helena de Melo et. al. **O Sistema de informação sobre nascidos vivos - SINASC**. São Paulo, Centro Brasileiro de Classificação de Doenças, 1992. (série Divulgação nº 7).

JORGE, Maria Helena de Melo et. al. Avaliação do Sistema de Informação sobre Nascidos Vivos e o uso de seus dados em epidemiologia e estatística de saúde. **Revista de Saúde Pública**, v. 27, 1993, supl.

LAURENTI, R. et al. **Estatísticas de saúde**. EPU, 2 ed., São Paulo, 1987. 186p.

LAURENTI, Ruy, BUCHALLA, Cássia Maria. Indicadores da Saúde Materna e Infantil: Implicações da Décima Revisão da Classificação Internacional de Doenças. **Revista Panamericana de Saúde Pública**, v. 1, n. 1, p. 18-22, jan. 1997.

LIU, Bing, HSU, Wynne, MA, Yiming. Integrating classification and association rule mining. In: **KDD-98**, New York, Aug, 1998.

LIU, Bing, HSU, Wynne, MA, Yiming, CHEN, Shu. Mining interesting knowledge using DM-II. In: **ACM SIGKDD - International Conference on Knowledge Discovery & Data Mining (KDD99)**, San Diego, CA, 1999.

LUBEL, Kenneth S. **Data mining:** a new way to find answers. White Paper. University of Maryland European Division, 1998.
http://faculty.ed.umuc.ed/~jmeinke/inss690/lubel.htm 07 de Novembro de 2000.

MALLACH, Efrem G. **Decision support and data warehouse systems**. Mc-Graw-Hill, USA, 2000.

MARCOPITO, L.F. et al. **Epidemiologia geral:** exercícios para discussão. São Paulo, Atheneu, 1996. 135p.

MINISTÉRIO DA SAÚDE. **Informe epidemiológico do SUS (Brasil)** - Ano 4. Brasília: CENEPI, 1995. 147 p. Sistemas de Informações de Saúde.

MINISTÉRIO DA SAÚDE. **Qualidade de indicadores básicos do IDB-97**: taxa de mortalidade infantil, 1997. http://www.datasus.gov.br 20 de Julho de 2001.

MINISTÉRIO DA SAÚDE. **Sistema de Informações sobre Mortalidade**: Manual de Procedimentos. 3ª Edição, Brasília, 1999.

MINISTÉRIO DA SAÚDE. **Sistema de Informações sobre Nascidos Vivos**: Manual de Procedimentos. Brasília, 1999.

PEREIRA, M.G. **Epidemiologia teoria e prática**. Rio de Janeiro: Guanabara Koogan, 1995. 583p.

QUINLAN, J. Ross. Improved use of continuous attributes in C4.5. **Journal of Artificial Intelligence Research**, v.4, 1996 p. 77-90.

ROTHMAN Kenneth J. **Epidemiologia moderna**. Madri: Díaz de Santos, 1987, p. 11-27

RULEQUEST RESEARCH. **See5:** an informal tutorial, 2000. http://rulequest.com/datamining/see5.0/see5-win.html 28 de Setembro de 2000.

SANTA CATARINA. SECRETARIA DE ESTADO DA SAÚDE. **SINASC - Sistema de Informações sobre Nascidos Vivos no estado de Santa Catarina,** 1ª Avaliação descritiva. Florianópolis: Ed. da UFSC, 1999. 64 p.

SOARES, José, FARIAS, A., CESAR, Cibel C. **Introdução à estatística**. Rio de Janeiro: LTC, 1991.

TODESCO, José L. **Reconhecimento de padrões usando rede neuronal artificial com uma função de base radial:** uma aplicação na classificação de cromossomos humanos. Florianópolis, 1995. Tese (Doutorado em Engenharia de Produção) - Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina.

WITTEN, Ian H., Frank, Eibe. **Data mining:** practical machine learning tools and techniquees with java implementations. Morgan Kaufmann Publishers, London, United Kingdom, 2000, p. 353-397.

GLOSSÁRIO

Epidemiologia - estudo das relações dos diversos fatores que determinam a freqüência e distribuição de um processo ou doença infecciosa numa comunidade. Ciência que estuda todos os possíveis fatores que, de alguma forma, contribuem para modificar a saúde de uma comunidade.

Mortalidade - percentual de mortes em uma comunidade em determinado período de tempo.

Natalidade - percentual de nascimentos de uma comunidade em determinado período de tempo.

Indicador de saúde - revela a situação de saúde de uma população, isto é, a quantificação de um aspecto da realidade.

Estudo de Coorte - na medicina, as pessoas nos estudos de coorte são selecionados por alguma característica (ou características) suspeitas de serem precursoras para um fator de risco de uma doença ou efeito na saúde, questionando o que irá acontecer nesse período.

Neonato - recém-nascido.

Perinatal - óbitos de crianças com menos de 7 dias de vida.

Puerpério - período que se segue ao parto até que os órgãos genitais e o estado geral da mulher retornem ànormalidade.

Nascimento Vivo - Nascimento vivo é a expulsão ou extração completa do corpo da mãe, independentemente da duração da gravidez, de um produto de concepção que, depois da separação, respire ou apresente qualquer outro sinal de vida, tal como batimentos do coração, pulsações do cordão umbilical ou movimentos efetivos dos músculos de contração voluntária, estando ou não cortado o cordão umbilical e estando ou não desprendida a placenta. Cada produto de um nascimento que reúna essas condições se considera como uma criança viva.