UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

ESCOLA DE INFORMÁTICA APLICADA

APLICAÇÃO DE KDD NOS DADOS DOS SISTEMAS SIM E SINASC EM BUSCA DE PADRÕES DESCRITIVOS DE ÓBITO INFANTIL NO MUNICÍPIO DO RIO DE JANEIRO

CLAUDIO JESUS ROSA

Orientadora

KATE CERQUEIRA REVOREDO

RIO DE JANEIRO, RJ – BRASIL

JULHO DE 2015

APLICAÇÃO DE KDD NOS DADOS DOS SISTEMAS SIM E SINASC EM BUSCA DE PADRÕES DESCRITIVOS DE ÓBITO INFANTIL NO MUNICÍPIO DO RIO DE JANEIRO

CLAUDIO JESUS ROSA

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada por:

_____

KATE CERQUEIRA REVOREDO

_____

FERNANDA ARAÚJO BAIÃO

_____

FLÁVIA MARIA SANTORO

RIO DE JANEIRO, RJ – BRASIL.
JULHO DE 2015

Agradecimentos

Primeiramente, agradeço ao Eterno pela saúde e por permitir que coisas maravilhosas aconteçam na minha vida.

Agradeço aos meus pais que são a razão de tudo que eu sou e de tudo o que eu conquistei até hoje. Agradeço antecipadamente por aquilo que eu venha a me tornar e tudo que ainda tenho para conquistar (a estrada ainda é longa!), sem eles seria impossível.

Agradeço à professora Kate pela confiança e oportunidade de desempenhar a função de monitor junto à disciplina de Estruturas Discretas (ajudando e entendendo os colegas e fazendo novas amizades) e por ter aceitado ser a orientadora deste trabalho.

Aos meus amigos pelo respeito e pela confiança que sempre depositam em mim e por me fazerem acreditar que posso conquistar tudo o que eu quiser.

Aos amigos que fiz na UNIRIO, grandes companheiros desta difícil jornada que é cursar e concluir uma graduação, certamente um grande passo na vida de cada um de nós.

Aos professores da UNIRIO pelo conhecimento e pela experiência passada e a todos os funcionários que nos ajudam quando precisamos de algo.

Muito Obrigado!

RESUMO

A diminuição da taxa de mortalidade infantil é uma das grandes prioridades da saúde. Neste contexto, são utilizados sistemas de informação em saúde para a coleta, o armazenamento, o processamento e a utilização de informações para o planejamento de programas específicos que possam atender às necessidades deste segmento da população. O SIM registra os dados sobre ocorrência de óbito ocorridos em território nacional e o SINASC registra os dados sobre os nascimentos. É possível relacionar os registros de óbito com os respectivos registros de nascimento dos pacientes e assim investigar as características de nascimento que estão associadas ao óbito em menores de um ano de idade. Para agregar valor a essa investigação pode ser utilizado o processo de KDD que consiste em um processo que tem o potencial de descobrir conhecimentos ocultos em bases de dados e apresentar os resultados em formato de regras que demonstrem padrões de correlação de dados. Este trabalho demonstra a aplicação do processo de KDD em duas bases de dados reais, o SIM e o SINASC, em busca de padrões descritivos de ocorrência de óbito em crianças de até um ano de idade no município do Rio de Janeiro utilizando o algoritmo *Apriori* e a ferramenta WEKA.

PALAVRAS-CHAVE: MORTALIDADE INFANTIL, SIM, SINASC, KDD, MINERAÇÃO DE DADOS.

ABSTRACT

The decrease in infant mortality rate is a top priority of health. In this context, health information systems are used for the collection, storage, processing and use of information for planning specific programs that can meet the needs of this segment of the population. The SIM records data on death occurrence occurred in the country and SINASC records data on births. You can list the death records with their birth records of patients and thus to investigate the characteristics of birth that are associated with death in children under one year of age. To add value to this research can be used the KDD process consisting of a process that has the potential to discover hidden knowledge in databases and present the results in the shape of rules showing data correlation patterns. This work demonstrates the application of the KDD process in two real databases, SIM and SINASC, looking for descriptive patterns of occurrence of death in children under one year of age in the city of Rio de Janeiro using the Apriori algorithm and the WEKA tool.

Keywords: Infant Mortality, SIM, SINASC, KDD, Data Mining.

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

Os sistemas de informação desempenham papel fundamental no apoio às organizações tanto para avaliar e melhorar seus processos internos quanto para responder de forma eficiente as exigências do ambiente em que se encontram.

Os sistemas de informação em saúde possuem o objetivo de coletar dados sobre a saúde da população e possibilitam avaliar a eficiência dos serviços oferecidos, propor as melhorias necessárias para oferecer tratamentos específicos e o planejamento de programas de saúde para atender as necessidades da população.

A Constituição Federal do Brasil estabelece que a saúde da população é um dever do Estado e um direito de todos e a sua garantia se dá através de políticas que visem à redução do risco de doenças e de outros agravos, assim como, o acesso universal e igualitário às ações e serviços para a sua promoção, proteção e recuperação (CONSTITUIÇÃO FEDERAL, 1988).

Além de oferecer o serviço necessário e do atendimento indiscriminado à população, cada pessoa deve ser tratada individualmente. Não é suficiente o tratamento universal é preciso considerar as especificidades de cada paciente (MINISTÉRIO DA SAÚDE, 2004).

O Ministério da Saúde, através do Sistema Único de Saúde (SUS), tem por função propor políticas públicas de saúde para atender as necessidades específicas de diversos segmentos da população, especialmente, idosos, mulheres e crianças.

Neste cenário, os cuidados relacionados à saúde da população infantil é uma das ações essenciais do Ministério da Saúde que busca desenvolver programas para oferecer um atendimento médico de melhor qualidade para as crianças (MINISTÉRIO DA SAÚDE, 2004).

Apesar do desenvolvimento de alguns programas que atendam às necessidades

específicas da população infantil, ainda há muito o que se fazer, principalmente, no que diz respeito ao oferecimento de tratamentos específicos que tenham por objetivo o combate às altas de taxas de mortalidade infantil, mesmo que estas tenham apresentado uma queda significativa nos últimos anos (MINISTÉRIO DA SAÚDE, 2004).

Há dois importantes sistemas de informação em saúde que o Ministério da Saúde implantou e que podem oferecer um subsídio inicial para esse processo de planejamento e construção de políticas de combate à mortalidade infantil: o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informações sobre Nascidos Vivos (SINASC).

O SIM foi desenvolvido pelo Ministério da Saúde no ano de 1975 para coletar dados sobre mortalidade no território nacional, o que possibilita a identificação do perfil epidemiológico da realidade brasileira.

O SINASC começou a ser implantando no ano de 1990 pelo Ministério da Saúde para criar um grande banco de dados sobre os nascimentos ocorridos em todo o território nacional possibilitando a criação do perfil epidemiológico dos nascidos vivos em todo o território nacional.

A integração entre esses dois sistemas possibilita realizar um estudo investigativo sobre as causas relacionadas à mortalidade infantil relacionando cada registro de óbito ocorrido em crianças de até um ano de idade com seus respectivos registros de nascimento e, assim, propor uma assistência especializada para faixa da população.

Neste contexto, pode ser utilizado o processo de Descoberta de Conhecimento em Bases de Dados – DCBD (*Knowledge Discovery in Databases* – KDD), (FAYAAD *et al*, 1996) que se constitui em um processo que tem o potencial de tornar explícito algum conhecimento que revele padrões de relacionamento de dados que dificilmente são identificados sem um processo automático.

A aplicação do processo de KDD pode ser utilizado para buscar os padrões relacionados às características dos nascidos vivos que vão a óbito antes de completar um ano de idade, o que pode servir de fundamento para o desenvolvimento de diversas ações de saúde para utilização no combate às causas da mortalidade nesta faixa etária.

1.1 Objetivo

Através da aplicação do processo de Descoberta de Conhecimento em Base de Dados (DCBD) busca-se extrair possíveis padrões de relacionamentos de dados na base de dados resultante da integração entre os registros coletados pelo Sistema de Informações sobre Mortalidade (SIM) e os registros coletados pelo Sistema de Informações sobre Nascidos Vivos (SINASC).

Este trabalho tem por objetivo a aplicação do processo de KDD para que seja possível descrever um comportamento padrão de relacionamento de dados que possam indicar um conhecimento sobre ocorrência de óbitos em crianças de até um ano de idade no Município do Rio de Janeiro através de dados disponibilizados pelo Departamento de Informática do SUS na internet.

1.2 Organização do texto

O presente trabalho está estruturado em capítulos e, além desta introdução, será desenvolvido da seguinte forma:

- Capítulo II: Principais Conceitos da Saúde – Onde são apresentados os principais conceitos relacionados à mortalidade infantil, à vigilância epidemiológica, os Sistemas de Informação em Saúde, o Sistema de Informações sobre Mortalidade e o Sistema de Informações sobre Nascidos Vivos.
- Capítulo III: Descoberta de Conhecimento em Base Dados – Onde são apresentados os conceitos do processo de Descoberta de Conhecimento em Base de Dados, suas etapas (Pré-processamento, Mineração de Dados e Pós-Processamento), a Ferramenta WEKA para a aplicação da mineração de dados, os conceitos de regras de associação e do algoritmo *Apriori*.
- Capítulo IV: Desenvolvimento – Nesta etapa são detalhadas as atividades realizadas neste trabalho de pesquisa que estão relacionadas à aplicação do processo de KDD na base de dados resultante da integração entre os dados do SIM e do SINASC.
- Capítulo V: Trabalhos Relacionados – Neste capítulo são citados trabalhos cujo escopo está relacionado à aplicação do processo de KDD em diversas áreas de saúde com objetivos específicos para cada caso.

- Capítulo VI: Conclusão – As considerações, com as contribuições, dificuldades encontradas e as sugestões de propostas de trabalhos futuros são apresentadas.

2. Principais Conceitos da Saúde

Neste Capítulo são apresentados os conceitos de Mortalidade Infantil, de vigilância epidemiológica, os Sistemas de Informação em Saúde, o SIM e o SINASC.

2.1. Mortalidade Infantil

O conceito de mortalidade infantil descreve o óbito que ocorre em crianças nascidas vivas desde o momento do nascimento até um ano de idade incompleto, ou seja, até 364 dias (PORTARIA MS n.º 72/2010).

A mortalidade infantil possui uma métrica conhecida como taxa de mortalidade infantil que consiste em um índice demográfico que mede a ocorrência de óbitos infantis em uma população em um determinado ano e local representada como a relação entre o número de óbitos para cada mil nascidos vivos.

A mortalidade infantil pode ser dividida em mortalidade neonatal e mortalidade pós-neonatal (ou infantil tardia). A mortalidade neonatal pode ser subdividida em neonatal precoce e neonatal tardia.

A taxa de mortalidade neonatal precoce é o número que expressa a relação de óbitos de crianças de 0 a 6 dias de vida completos por mil nascidos vivos, na população de um determinado espaço geográfico, no ano que está sendo considerado e possibilita uma estimativa do risco de uma criança nascida viva morrer durante a primeira semana de vida.

A taxa de mortalidade neonatal tardia é o número que expressa a relação de óbitos de crianças de 7 a 27 dias de vida completos por mil nascidos vivos na população de um determinado espaço geográfico, no ano que está sendo considerado e possibilita uma estimativa do risco de uma criança nascida viva morrer entre a segunda a quarta semana de vida.

A mortalidade no período neonatal reflete de uma maneira geral, as condições socioeconômicas e de saúde da mãe, a inadequada assistência pré-natal, ao parto e ao recém-nascido. As informações relacionadas ao período neonatal permitem a análise para identificação de demandas por ações de serviços de saúde específicos, avaliação dos níveis de saúde e de desenvolvimento socioeconômico da população e fornecem subsídios para o processo de planejamento, gestão e avaliação de políticas e ações de saúde para a atenção pré-natal, ao parto e ao recém-nascido (MINISTÉRIO DA SAÚDE, 2009).

A taxa de mortalidade pós-neonatal ou infantil tardia é o número que expressa a relação de óbitos de crianças nascidas vivas na faixa etária entre 28 e 364 dias de vida completos por mil nascidos vivos na população de um determinado espaço geográfico, no ano que está sendo considerado e possibilita uma estimativa do risco de uma criança nascida viva morrer entre os 28 e 364 dias de vida.

A mortalidade no período pós-neonatal está geralmente associada às condições ambientais ao qual o recém-nascido é submetido bem com o acesso e a qualidade dos recursos disponíveis para a atenção à saúde materno-infantil. É fundamental o planejamento, a gestão e avaliação de políticas públicas, principalmente, no ambiente em que vive o recém-nascido, além de ações de saúde para a atenção ao período pré-natal, ao parto e para a proteção da saúde infantil (MINISTÉRIO DA SAÚDE, 2009).

A redução da taxa de mortalidade infantil é um desafio para o serviço público de saúde brasileiro e é um dos compromissos assumidos dentro das Metas do Desenvolvimento do Milênio da Organização das Nações Unidas (ONU). O Brasil vem apresentado diminuição na taxa de mortalidade infantil mas os níveis ainda são considerados elevados (MINISTÉRIO DA SAÚDE, 2009).

A tabela 1, apresenta a taxa de mortalidade infantil no município do Rio de Janeiro, entre os anos de 2008 e 2012.

Faixa Etária	2008	2009	2010	2011	2012
Neonatal Precoce	6,01	6,43	6,38	5,90	5,98
Neonatal Tardio	2,60	2,23	2,06	2,01	2,37
Pos Neonatal	4,98	4,97	4,62	5,06	4,70
Infantil	13,59	13,63	13,05	12,93	13,05

Tabela 1: Taxa de Mortalidade no Município do Rio de Janeiro[1].

1 Fonte: <http://www.rio.rj.gov.br/dlstatic/10112/1368636/4115727/mortalidadeinfantiltaxatabela_2012.pdf> - Acesso em: 11 de junho de 2015.

A tabela 2, demonstra um comparativo dos índices de mortalidade infantil entre os 10 países com menores índices de mortalidade e o Brasil, entre os anos de 2008 e 2012.

Países	2008	2009	2010	2011	2012
Singapura	2,3	2,31	2,32	2,32	2,65
Islândia	3,25	2,31	2,32	2,32	3,18
Japão	2,8	2,79	2,79	2,78	2,21
Suécia	2,75	2,75	2,74	2,74	2,74
Finlândia	3,5	3,47	3,45	3,43	3,4
Noruega	3,61	3,58	3,55	3,52	3,5
Luxemburgo	4,62	4,56	4,49	4,44	4,39
República Tcheca	3,83	3,79	3,76	3,73	3,7
França	3,36	3,33	3,31	3,29	3,37
Eslovênia	4,3	4,25	4,21	4,17	4,12
Brasil	23,33	22,58	21,86	21,17	20,5

Tabela 2: Comparativo de Taxa de Mortalidade Infantil Mundial[2].

A figura 1, apresenta os dados da tabela 1 de forma a evidenciar graficamente a diferença entre estes países e o Brasil.

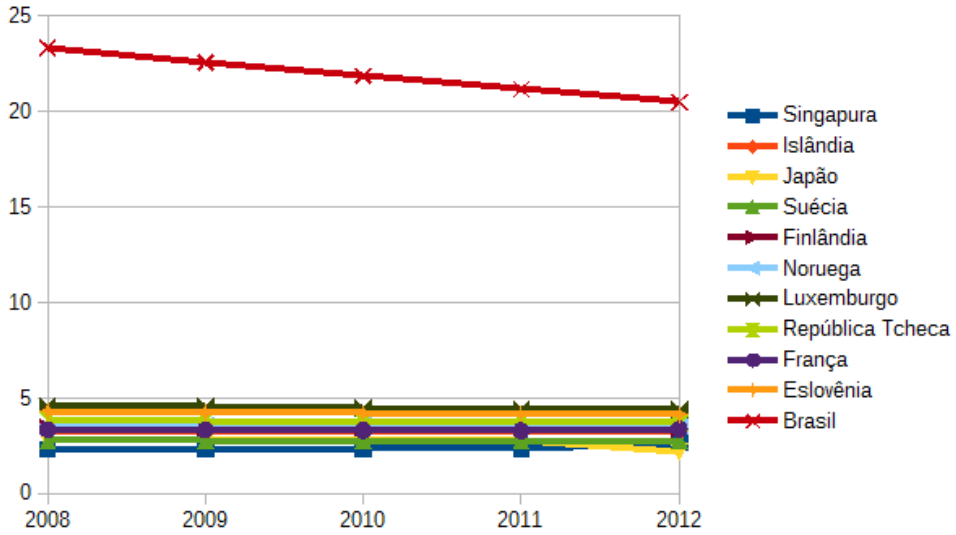


Figura 1: Comparativo da Taxa de Mortalidade Infantil.

Conhecidos os principais conceitos relacionados à mortalidade infantil é necessário situar em que contexto dentro da estrutura do SUS a mortalidade infantil está inserida e a quem compete o conjunto de ações.

2 Fonte: <http://www.indexmundi.com/g/g.aspx?v=29&c=br&l=pt> - Acesso em 11 de junho de 2015.

2.2. Vigilância Epidemiológica

A vigilância epidemiológica é definida pela Lei n.º 8.080, no seu artigo 6º, inciso XI, parágrafo 2º, como “um conjunto de ações que proporcionam o conhecimento, a detecção ou prevenção de qualquer mudança nos fatores determinantes e condicionantes de saúde individual ou coletiva, com a finalidade de recomendar e adotar as medidas de prevenção e controle das doenças ou agravos.”

Dentre suas funções, a vigilância epidemiológica necessita coletar dados para análise e recomendação de medidas de prevenção e controle apropriadas para a promoção das ações de prevenção indicadas, com posterior avaliação da eficácia e eficiência das medidas adotadas e divulgação das informações pertinentes (MINISTÉRIO DA SAÚDE, 2009).

A Lei Orgânica da Saúde (MINISTÉRIO DA SAÚDE, 1990), estabelece no seu artigo n.º 18, inciso IV, alínea a, dentre outras atribuições, a competência da gestão municipal para a execução de serviços de vigilância epidemiológica e a atuação dos estados na coordenação e atuação complementar às atividades dos municípios.

A vigilância de óbitos está contida no conceito de vigilância epidemiológica que compreende o conhecimento das causas determinantes de óbitos maternos, óbitos infantis (objeto deste estudo), óbitos fetais e óbitos com causa mal definida (Portal da Saúde, SUS).

Neste sentido, o aumento na capacidade de captação de notificação de registros de nascimento e de registros de óbito, bem como a qualidade na definição da causa de óbito, é fundamental no apoio às atividades das Secretarias Municipais de Saúde no que diz respeito à vigilância de óbitos.

Assim, é necessário algum mecanismo eficiente de coleta de dados que facilite o trabalho de análise da situação epidemiológica da população e, neste caso, os sistemas de informação em saúde desempenham papel fundamental.

2.3. Sistemas de Informação em Saúde

Um Sistema de Informação em Saúde (SIS) é um software que permite a coleta, o processamento, a análise e a transmissão da informação necessária para subsidiar o planejamento, a organização, a operação e a avaliação dos serviços de saúde.

A informação em saúde serve tanto de fundamento para o planejamento de ações

para atender as necessidades comuns de um conjunto de pessoas, como por exemplo, programas e políticas de saúde específicos para a saúde da mulher, quanto para o planejamento de ações para atender as necessidades individuais de um paciente, como por exemplo, um profissional de saúde pode se basear nos registros históricos dos atendimentos a que foi submetido um paciente para um melhor acompanhamento e para ter condições de propor um tratamento específico.

É importante que os profissionais responsáveis pela alimentação dos dados nos SIS o realizem de maneira correta e completa para que os dados coletados reflitam de maneira fidedigna a situação de saúde da população e possibilitem a criação de serviços de saúde efetivos.

No Brasil, existem diversos SIS cada um com algum propósito específico, criados de forma independente sem uma preocupação inicial no compartilhamento de informações entre eles.

Com relação a este fato, há esforços do Ministério da Saúde para que seja possível que os SIS conversem entre si de alguma forma, os quais podem ser citados: o projeto Arquitetura SOA–SUS[3] e o Sistema Cartão Nacional de Saúde[4].

Para efeito de conhecimento podemos citar os seguintes SIS utilizados pelos órgãos do SUS:

- o SINAN (Sistema de Informações de Agravos de Notificação) tem por objetivo a coleta de dados referentes à notificação e investigação de casos de doenças e agravos que constam na lista nacional de doenças de notificação compulsória;
- o SIAB (Sistema de Informação da Atenção Básica) tem por objetivo o acompanhamento das ações e dos resultados das atividades realizadas pelas equipes do Programa Saúde da Família;
- o SIH/SUS (Sistema de Informações Hospitalares do SUS) tem por objetivo coletar dados referentes aos atendimentos realizados nas internações hospitalares financiados pelo SUS;
- o SIA/SUS (Sistema de Informações Ambulatoriais do SUS) tem por objetivo coletar dados referentes aos atendimentos ambulatoriais realizados pelas unidades de saúde financiados pelos SUS;
- o SIM (Sistema de Informações sobre Mortalidade) tem por objetivo a criação

3 Projeto de Interoperabilidade SOA-SUS, informações: <http://datasus.saude.gov.br/interoperabilidade> - Acesso em: 10 de junho de 2015.

4 Identificação dos usuários do SUS, informações: <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt0940_28_04_2011.html> - Acesso em: 10 de junho de 2015.

de um banco de dados nacional referente aos óbitos ocorridos no território nacional;

- o SINASC (Sistema de Informações sobre Nascidos Vivos) tem por objetivo a criação de um banco de dados nacional referente aos nascimentos ocorridos em território nacional.

Considerando o escopo definido para este trabalho, serão utilizadas as bases de dados de dois importantes sistemas de informação epidemiológicos: o SIM e o SINASC.

2.4. Sistema de Informações sobre Mortalidade

O Sistema de Informações sobre Mortalidade (SIM) é um sistema de informações epidemiológicas de abrangência nacional desenvolvido pelo Ministério da Saúde (MS) em 1975 composto por um conjunto de ações relacionada à coleta, codificação, processamento de dados, fluxo, consolidação, avaliação e divulgação de informações sobre os óbitos ocorridos no país.

A fonte de alimentação do SIM é o formulário da Declaração de Óbito (DO). A DO é um documento padronizado pelo MS, com numeração sequencial única, de uso obrigatório em todo o território nacional. A figura 2, demonstra os principais conceitos captados pelo SIM.

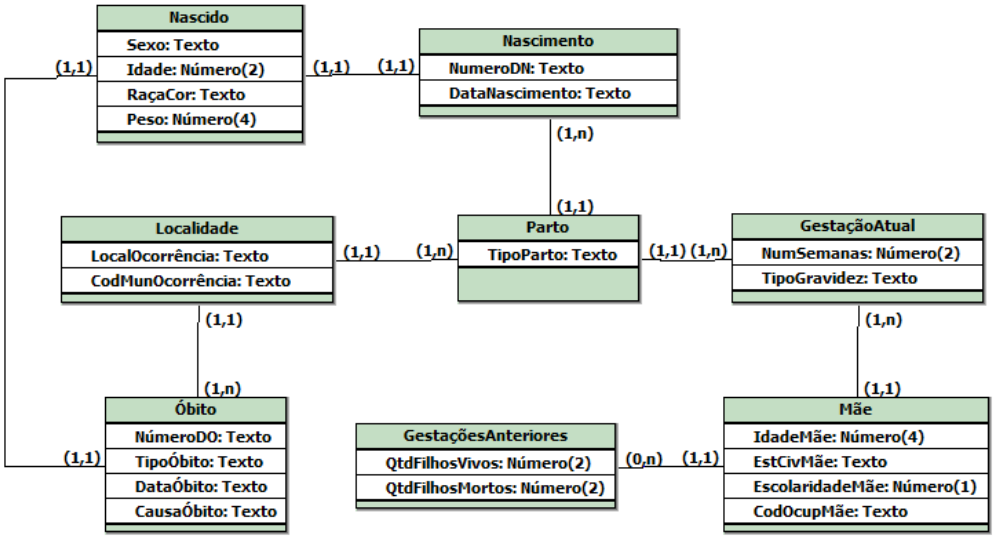


Figura 2: Modelo Lógico: Principais Conceitos captados pelo SIM

A padronização do modelo da DO permite a uniformização do modo pelo qual os dados sobre mortalidade são notificados e registrados no país. No período anterior à implantação do SIM existiam diferentes tipos de modelos de declaração de óbito que dificultavam o trabalho de consolidação de informações sobre as mortes ocorridas no território nacional.

A DO é fornecida em três vias autocopiativas às Secretarias Estaduais de Saúde cuja impressão, controle e distribuição está sob a responsabilidade do Ministério da Saúde. As Secretarias Estaduais de Saúde são responsáveis pela distribuição da DO às suas Secretarias Municipais de Saúde.

A distribuição das três vias da DO é feita da seguinte maneira: uma via é de responsabilidade da unidade de saúde para posterior envio para a respectiva secretaria municipal de saúde; uma segunda via é entregue ao responsável pelo paciente falecido para obtenção da Certidão de Óbito e uma terceira via é arquivada no prontuário do paciente falecido na respectiva unidade de saúde. A figura 3 demonstra este fluxo das três vias da DO.

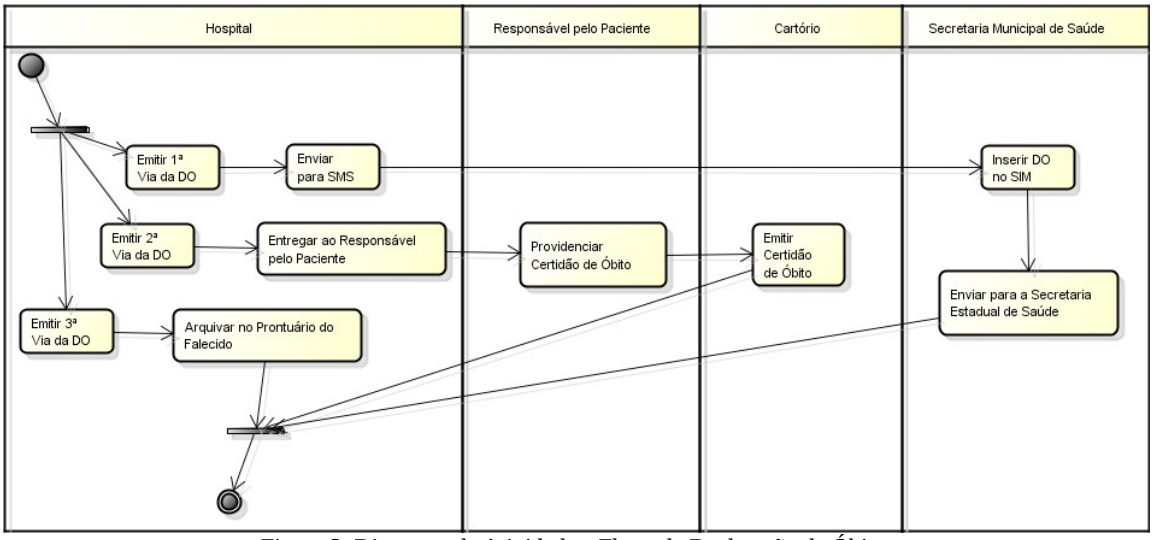


Figura 3: Diagrama de Atividades: Fluxo da Declaração de Óbito.

O preenchimento dos dados da DO é de responsabilidade do médico que atestou o óbito (RESOLUÇÃO Nº 1779/2005, CFM) excetuados os casos confirmados ou suspeitos de morte por causas externas, cuja responsabilidade é atribuída ao médico do Instituto Médico Legal (IML) ou equivalente (PORTARIA MS nº. 116, 2009).

A gestão do SIM é de responsabilidade das Secretarias Municipais de Saúde seguindo as normas e diretrizes nacionais e estaduais. As Secretarias Municipais de Saúde tem como atribuições a coleta, o processamento, a consolidação e a avaliação dos

dados que alimentam o SIM recebidos das suas unidades de saúde e, posteriormente, os enviam para suas respectivas Secretarias Estaduais de Saúde conforme os fluxos e prazos estabelecidos (Portaria MS nº. 116, 2009).

2.5. Sistema de Informações sobre Nascidos Vivos

O Sistema de Informações sobre Nascidos Vivos (SINASC) é um sistema de informações epidemiológicas de abrangência nacional desenvolvido pelo Ministério da Saúde (MS) em 1990 composto por um conjunto de ações relacionadas à coleta, codificação, processamento de dados, fluxo, consolidação, avaliação e divulgação de informações sobre nascidos vivos ocorridos no país.

A fonte de alimentação do SINASC é o formulário da Declaração de Nascidos Vivos (DNV). A DNV é um documento padronizado pelo MS, com numeração sequencial única, de uso obrigatório em todo o território nacional que deve ser preenchido para todos os nascidos vivos independentemente do local onde tenha ocorrido o nascimento (unidades de saúde, vias públicas, domicílios, veículos de transporte, etc…). A figura 4 demonstra os principais conceitos captados pelo SINASC.

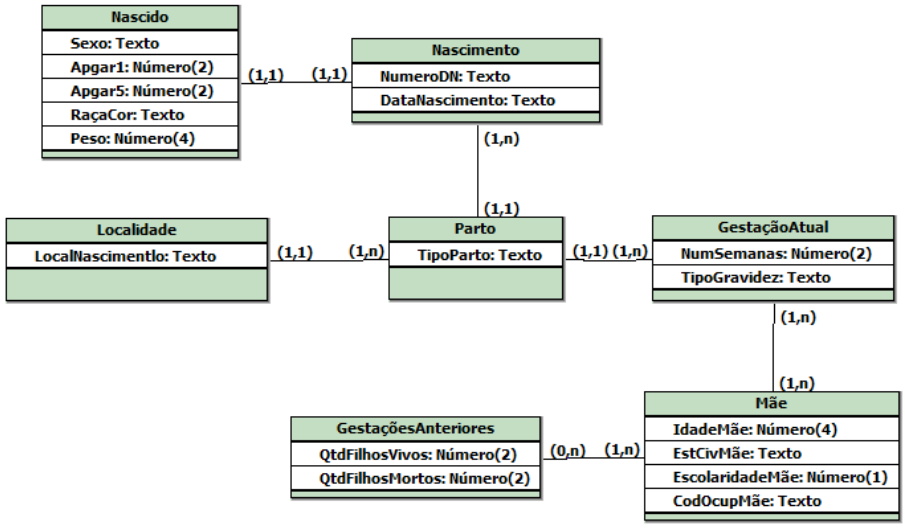


Figura 4: Modelo Lógico: Principais Conceitos captados pelo SINASC.

A DNV é fornecida em três vias autocopiativas cuja impressão, controle e distribuição está sob a responsabilidade da Secretaria de Vigilância Sanitária do Ministério da Saúde (SVS/MS). Cada Secretaria Estadual de Saúde é responsável pela distribuição da DNV às suas Secretarias Municipais de Saúde.

A distribuição das três vias da DNV é feita da seguinte maneira: a primeira via é

de responsabilidade da unidade de saúde que realizou o parto para posterior envio à Secretaria Municipal de Saúde; a segunda via é entregue à família para apresentação no Cartório de Registro Civil para emissão da Certidão de Nascimento (o Cartório arquiva essa via); a terceira via é arquivada no prontuário médico da gestante ou do recém-nascido na unidade de saúde que realizou o parto. A figura 3 demonstra o fluxo das três vias da DNV.

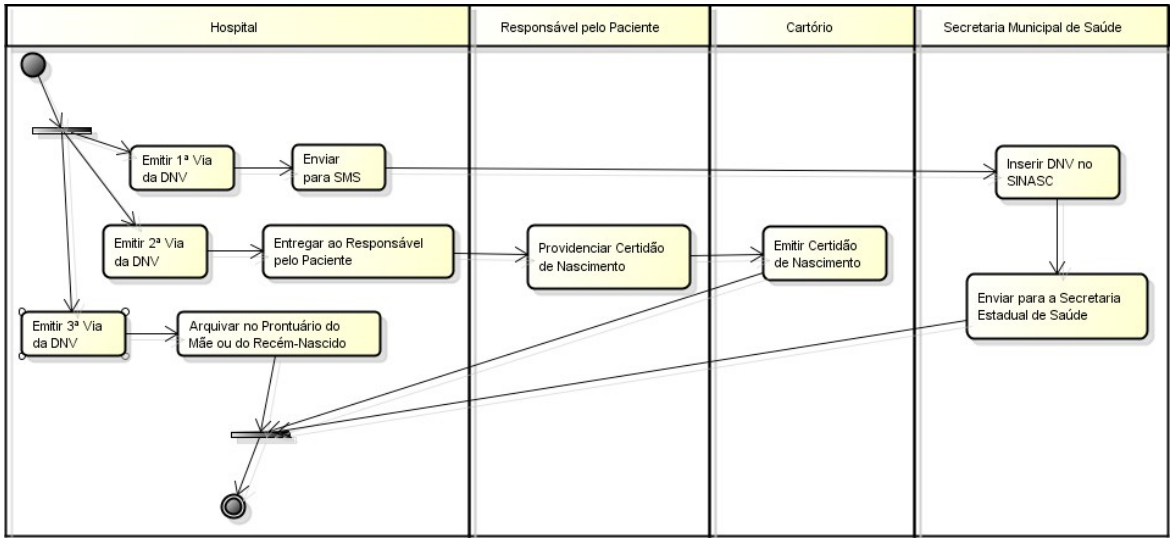


Figura 5: Diagrama de Atividades: Fluxo da Declaração de Nascido Vivo.

A coleta, o processamento, a consolidação e a avaliação dos dados que alimentam o SINASC é de responsabilidade das respectivas Secretarias Municipais de Saúde que os recebem dos hospitais notificantes e, posteriormente, as enviam as suas respectivas Secretarias Estaduais de Saúde conforme os fluxos e prazos estabelecidos (PORTARIA MS nº. 116, 2009).

Neste capítulo foi definido o conceito de mortalidade infantil, situando-o dentro de um conceito maior, a vigilância epidemiológica, referenciada como competência prioritária da esfera municipal. Depois, foi descrita a importância de um mecanismo de coleta de dados e explicados os conceitos de Sistemas de Informação em Saúde e, especificamente, o SIM e o SINASC.

No próximo capítulo, serão demonstrados os principais conceitos relacionados ao processo de descoberta de conhecimento em base de dados que será utilizado para a extração de padrões descritivos para, conforme o objetivo do trabalho, demonstrar a ocorrência de óbito em crianças de até um ano de idade no município do Rio de Janeiro.

3. A Descoberta de Conhecimento em Base de Dados

Neste Capítulo, são apresentados os conceitos do processo de Descoberta de Conhecimento em Base de Dados, as etapas de Pré-processamento, Mineração de Dados e Pós-Processamento, as regras de associação, o algoritmo *Apriori* e a Ferramenta WEKA.

3.1. O Processo de Descoberta de Conhecimento em Base de Dados

O processo de descoberta de conhecimento em base de dados (*Knowledge Discovey in Database – KDD*) é um importante recurso para auxiliar os gestores na tarefa de tomada de decisão e está relacionado à procura de conhecimento a partir de bases de dados.

O KDD é um processo que tem uma natureza interdisciplinar em que há uma interseção entre diversos campos de pesquisa, tais como, aprendizado de máquina, reconhecimento de padrões, banco de dados, estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados e computação de alto desempenho (FAYAAD *et al*, 1996).

Segundo (FAYYAD *et al*, 1996) “o KDD é um processo interativo e iterativo, não trivial, composto por várias etapas, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados”

O processo é interativo porque é necessário um elemento humano responsável por intervir e controlar as atividades do processo definindo os objetivos e avaliando os resultados. O processo é iterativo porque é possível realizar o refinamento sucessivo através da repetição de cada uma das etapas ou de todo o processo de KDD na busca de resultados satisfatórios.

O processo é não trivial porque é necessária a utilização de alguma técnica de busca ou inferência. O conceito previamente desconhecido significa que a informação deve ser nova para o sistema e para o usuário. O conceito de potencialmente útil significa que a informação deve possibilitar ao usuário algum ganho.

A figura 6, demonstra o processo de KDD através de uma representação composta de três principais etapas: a etapa de pré-processamento, a etapa de mineração de dados e a etapa de pós-processamento.

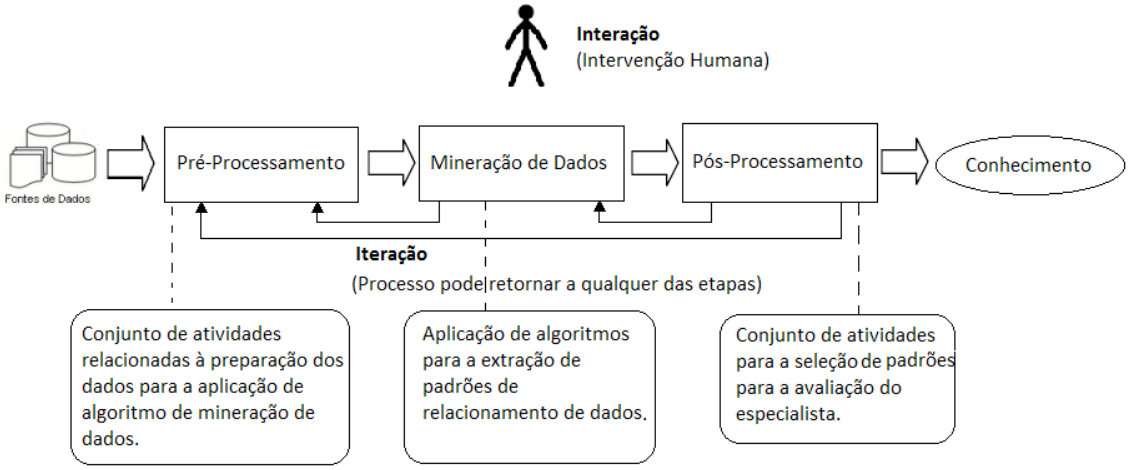


Figura 6: O Processo de KDD.

O pré-processamento é a primeira etapa do processo de KDD e consiste em um conjunto de atividades que preparam o conjunto de dados em formatos específicos para a aplicação da etapa de mineração de dados.

A mineração de dados é a etapa do processo de KDD onde ocorre a aplicação de algoritmos específicos para a extração de padrões a partir da base de dados preparada na etapa de pré-processamento.

No pós-processamento é possível a visualização, análise e seleção de regras para a interpretação dos resultados encontrados na aplicação da etapa de mineração de dados que podem ser representados através de gráficos, diagramas e relatórios demonstrativos.

3.2. Pré Processamento

A etapa de pré-processamento de dados é composta por um conjunto de atividades necessárias para realizar a preparação do conjunto de dados utilizados para a aplicação dos algoritmos da etapa de mineração de dados.

São diversas as causas que demonstram a necessidade de realização das atividades da etapa de pré-processamento de dados:

- a numerosa quantidade de registros é diminuída devido a problemas na forma como os dados estão registrados (dados incorretos, dados faltantes, valores duplicados);
- os registros utilizados no trabalho possuem origem em mais de uma base de dados e a etapa de mineração de dados é aplicada sobre uma única base de dados, o que demanda um trabalho de integração (muitas vezes não é fácil de ser realizada devido a falta de campos comuns entre as fontes de dados);
- os dados necessários à realização do trabalho estão armazenados em formatos diferentes que muitas vezes não são suportados pelo algoritmo de mineração de dados específico;

A fase de pré-processamento de dados é semiautomática porque depende da capacidade de quem a conduz em identificar os problemas presentes nos dados e utilizar os métodos mais apropriados para solucionar cada um destes problemas.

3.2.1. Seleção de Dados

A seleção de dados refere-se à escolha do conjunto de dados necessários ao entendimento do problema que se busca resolver e sobre o qual o processo de descobrimento será aplicado contendo os registros e suas variáveis – também conhecidas como características ou atributos (FAYYAD *et al*, 1996).

A complexidade desse processo pode variar dependendo de quais fontes serão extraídos os conjuntos de dados e em quais formatos eles foram armazenados.

3.2.2. Limpeza de Dados

A limpeza de dados é a atividade que tem o objetivo aumentar a qualidade do conjunto de dados que será utilizado no processo de extração de conhecimento.

É neste momento que são definidas as estratégias para o tratamento de registros com dados faltantes, dados incorretamente informados, dados muito fora da faixa de valores ou dados informados com valores diferentes nas fontes de dados utilizadas.

Para a correção dos registros incompletos, assim conhecidos como valores faltantes ou *missing values*, algumas propostas são sugeridas, como por exemplo:

excluir os registros, calcular a média dos valores preenchidos e completar o campo, utilizar uma constante global, predizer um possível valor através da utilização de regressão, utilizar uma fonte secundária para preencher o campo.

Outro problema a ser tratado são os registros que possuem dados com valores extremos (*outliers*), atípicos ou cujas características o tornam muito diferente dos demais registros. O que normalmente ocorre é que esse tipo de registro é descartado da amostra. Dependendo do objetivo a ser alcançado com o processo de KDD e do conjunto de dados utilizado, esse tipo de registro pode ser interessante para descobrir comportamentos discrepantes, como por exemplo, na detecção de fraudes.

3.2.3. Integração de Dados

A etapa de mineração de dados é aplicada sobre uma base de dados única e geralmente muitos trabalhos utilizam registros que estão armazenados em fontes de dados distintas.

A integração de dados pode ser realizada através da combinação de um campo identificador comum que estão presentes em bases de dados diferentes, caso em que o processo pode ser realizado de forma automática, bem como a utilização de um conjunto de atributos secundários identificadores.

3.2.4. Transformação dos Dados

A transformação de dados consiste em representar os dados em formatos adequados para serem manipulados pelo algoritmo de aprendizado que será utilizado na etapa de mineração, dependendo dos objetivos que se quer alcançar com a aplicação do processo de KDD (FAYYAD *et al*, 1996).

3.3. Mineração de Dados

A mineração de dados (*Data Mining*) é uma das principais etapas no processo de KDD. É nesta etapa que o conhecimento pode ser extraído possibilitando a descoberta de padrões, relações ocultas e regras para predizer ou descrever e correlacionar dados através da aplicação de algoritmos de aprendizado de máquina (MITCHELL, 1997).

Segundo (BERRY *et al*, 1997), **“***Data Mining* é a exploração e análise, de forma

automática ou semiautomática, de grandes bases de dados com objetivo de descobrir padrões e regras. O objetivo do processo de mineração é fornecer as corporações informações que as possibilitem montar melhores estratégias de *marketing*, vendas, suporte, melhorando assim os seus negócios".

Segundo (FAYYAD *et al*, 1996), "Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados."

Os dois principais objetivos da mineração de dados é a predição e a descrição. A predição utiliza as variáveis existentes na base de dados para predizer valores futuros ou desconhecidos pelo usuário. A descrição é baseada na procura de padrões contidos em uma base de dados para que o usuário possa interpretá-los. A escolha do tipo de tarefa depende do tipo de problema que se quer resolver e dos objetivos que se quer alcançar com a aplicação da mineração de dados.

A seguir serão apresentados alguns tipos de padrões descritivos (regras de associação e agrupamento) e preditivo (classificação), antes é necessário diferenciar dois tipos de aprendizado: o aprendizado supervisionado e o aprendizado não supervisionado.

O aprendizado supervisionado consiste em classificar os registros a partir de exemplos que já estão previamente rotulados em uma classe conhecida. O aprendizado não supervisionado não utiliza uma classificação prévia e consiste na tentativa de agrupamento dos registros de alguma forma.

Uma descoberta de associação é um tipo de aprendizado não supervisionado que demonstra a relação entre um conjunto de itens nos registros de uma determinada base de dados e um outro conjunto distinto de itens nos mesmos registros. Uma regra de associação é representada por uma expressão condicional na forma $X \rightarrow Y$, onde X é a condição e Y é a consequência ou na forma **se** <condição> **então** <consequência> que significa que sempre que um conjunto de itens de dados estiver presente (X) implica na presença de um outro conjunto diferente de itens de dados nos mesmos registros (Y). Para este tipo de aprendizado pode ser utilizado o algoritmo *apriori* (AGRAWAL, 1994).

O agrupamento ou clusterização é um tipo de aprendizado não supervisionado de padrões em conjuntos de dados em que cada registro é incluído em um *cluster* de tal forma que todos os registros pertencentes a um mesmo *cluster* possuem mais

similaridades entre si do que com os registros de outros *clusters*. Os algoritmos de clusterização dividem os objetos em grupos nos quais a similaridade intracluster é maximizada e a similaridade intercluster é minimizada. Para este tipo de aprendizado pode ser utilizado o algoritmo *k-means* (MACQUEEN, 1967).

A classificação é um tipo de aprendizado supervisionado que tem por objetivo predizer em que classe previamente conhecida determinado registro ainda não classificado será incluído baseando-se na similaridade das propriedades. Os algoritmos de classificação buscam encontrar alguma correlação entre os atributos do registro ainda não classificado e uma dentre um conjunto finito de classes. Para este tipo de aprendizado pode ser utilizado o algoritmo *J48* (QUINLAN, 1987).

3.3.1. Regras de Associação

Dada uma regra X → Y, a sua medida de suporte (Sup) representa a porcentagem de transações da base de dados que contém os itens de X e de Y, indicando a relevância da mesma.

O suporte de uma regra X → Y, onde X e Y são conjuntos de itens, é dado pela seguinte fórmula:

$$Suporte = \frac{Frequência\ de\ X\ e\ Y}{Total\ de\ T}$$

A fórmula acima demonstra que o suporte é uma relação entre o número de transações em que X e Y ocorrem simultaneamente e o total de transações (T).

A medida de confiança (Conf) representa, dentre as transações que possuem os itens de X, a porcentagem de transações que possuem também os itens de Y, indicando a validade da regra.

A confiança de uma regra X → Y é dado pela seguinte fórmula:

$$Confiança = \frac{Frequência\ de\ X\ e\ Y}{Frequência\ de\ X}$$

A fórmula acima demonstra que a confiança é uma medida que expressa a relação entre o número de transações em que X e Y ocorrem simultaneamente e o total de transações em que X também aparece.

A utilização da técnica de regras de associação em mineração de dados tem por

objetivo encontrar as regras que possuem suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo (SupMin) e uma confiança mínima (ConfMin), especificados pelo usuário interessado no resultado do processo.

3.3.2. O Algoritmo *APRIORI*

O algoritmo *Apriori* é um dos algoritmos mais conhecidos e utilizados quando é utilizada a mineração de regras de associação para solucionar o problema de mineração de *itemsets* frequentes.

Um *itemset* é um conjunto formado por um ou mais dados em uma transação. Um *itemset* que possui *k* elementos é chamado de *k-itemset*.

O algoritmo *Apriori* possui três fases principais: a fase da geração dos *itemsets* candidatos, a fase de poda dos *itemsets* candidatos e fase de cálculo de suporte do *itemset*.

Nas fases de geração de *itemsets* candidatos e da poda dos *itemsets* existe a propriedade de Antimonotonia da relação de inclusão entre os *itemsets* que considera que, dados dois *itemsets* X e Y tal que $X \subseteq Y$, se Y é frequente então X também é frequente, ou seja, um *itemset* é frequente então todos os *itemsets* que são subconjuntos deste *itemset* também são frequentes.

O algoritmo *Apriori* tem um funcionamento de forma iterativa que calcula cada *itemset* frequente de tamanho *k* a partir de *itemsets* frequentes de tamanho $k - 1$ já calculados na iteração anterior, assim, é feita a geração dos *itemsets* candidatos.

Os *itemsets* formados por um único dado são computados considerando-se todos os conjuntos unitários possíveis, de um único item. A seguir, é realizada uma varredura no banco de dados para calcular o suporte de cada um destes conjuntos unitários excluindo-se aqueles que não possuem o suporte igual ou superior ao mínimo definido.

Na fase de poda dos *itemsets* candidatos é feita uma verificação se os *itemsets* possuem algum subconjunto de *itemsets* que não seja frequente.

O cálculo do suporte dos *itemsets* frequentes é realizado varrendo o banco de dados verificando para cada transação *t* presente no banco de dados quais são os candidatos suportados por *t* e para cada candidato é incrementado a unidade do contador do suporte.

3.4. Pós-Processamento

O pós-processamento de dados é a etapa que trabalha diretamente com o conjunto de regras geradas pelos algoritmos utilizados na etapa de mineração de dados antes da avaliação das regras pelo especialista do negócio.

A etapa de pós-processamento se mostra útil porque os algoritmos de mineração de dados podem gerar uma grande quantidade de regras e nem todas possuem potencial para resolver o problema identificado inicialmente na aplicação do processo de KDD e outras não possuem utilidade por serem muito óbvias.

É possível que nesta etapa, como em qualquer outra que compõe o processo de KDD, seja identificada a necessidade de retornar a qualquer das etapas anteriores ou repetir todo o processo de KDD, inclusive com a utilização de outras fontes de informação e mudança do algoritmo utilizado na etapa de mineração.

O pós-processamento é uma etapa de refinamento ao final do processo de KDD para que seja apresentado ao especialista do negócio um número reduzido das regras que estejam dentro do escopo de interesse para facilitar o trabalho de interpretação tornando-o mais produtivo.

3.5. A Ferramenta WEKA

A ferramenta WEKA[5] (*Waikato Environment for Knowledge Analysis*) foi desenvolvida pelos pesquisadores da Universidade de Waikato, na Nova Zelândia. Tem a implementação na linguagem Java que permite a utilização em diversos sistemas operacionais.

WEKA é um software livre de código aberto disponibilizado livremente na sua página para utilização por qualquer usuário de forma gratuita.

A ferramente WEKA possui uma interface gráfica que visa proporcionar facilidade ao usuário nas atividades de aplicação do algoritmo de mineração de dados e de visualização dos padrões encontrados.

A tela inicial da ferramenta WEKA é o *GUI Chooser*, mostrado na figura 7. Nesta tela é possível acessar o WEKA *Explorer*, mostrado na figura 8. É no WEKA

5 Informações sobre a ferramenta WEKA, <http://www.cs.waikato.ac.nz/ml/weka/> - Acesso em: 15 de junho de 2015.

Explorer que o usuário carrega o arquivo de dados e escolhe qual o algoritmo será aplicado.

Figura 7: WEKA GUI Chooser.

O WEKA Explorer é organizado em forma de abas como pré-processamento (abertura do arquivo para processamento), classificação (aplicação do algoritmo de classificação), clusterização (aplicação do algoritmo de agrupamento), associação (aplicação do algoritmo para regras de associação), seleção de atributos (selecionar e definir a relevância de atributos) e visualização (atributos mostrados em coordenadas "x" e "y").

Os dados utilizados como fonte de entrada para a ferramenta WEKA possuem um formato de arquivo próprio, o *Atribute Relation File Format* (ARFF), também desenvolvido pela Universidade de *Waikato*, além dele outros formatos podem ser utilizados, como o CSV. Outra forma de trabalhar com dados na ferramenta WEKA é utilizando uma base de dados diretamente através da API (*Application Programming Interface*) JDBC (*Java DataBase Connectivity*) que permite a comunicação do Java com qualquer tipo de banco de dados.

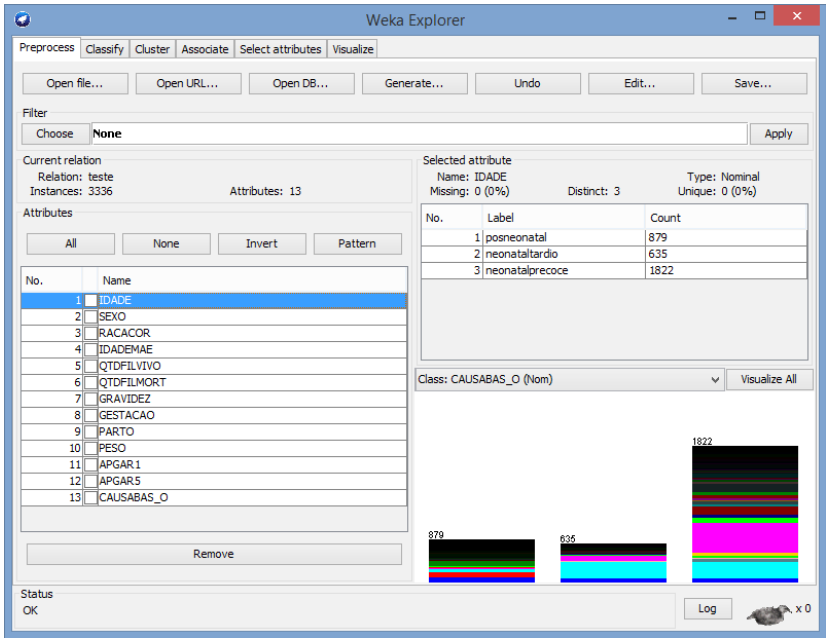


Figura 8: WEKA Explorer

O resultado da aplicação do algoritmo é mostrado no *Explorer* e pode ser salvo pelo usuário para que possa ser visualizado posteriormente, conforme mostrado na figura 9.

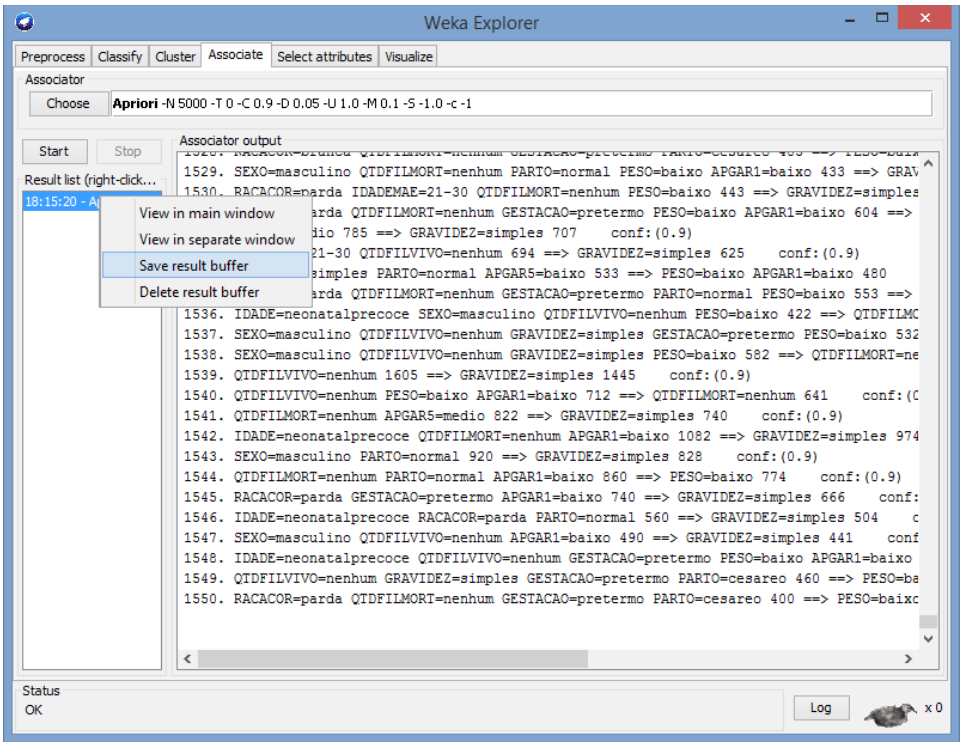


Figura 9: Opção de salvar o resultado do WEKA.

4. A Aplicação de KDD no SIM e no SINASC

Nesta etapa será descrito o conjunto de atividades realizadas para a criação e preparação da base de dados (pré-processamento), a mineração de dados para a extração dos padrões (mineração de dados) e o tratamento, seleção e avaliação das regras geradas (pós-processamento), para atingir o objetivo proposto na seção 1.1. deste trabalho.

4.1. Pré Processamento no SIM e no SINASC

4.1.1. Definição do Problema

O trabalho consiste no estudo do processo de KDD em uma base composta por dados reais, neste caso, a base de dados resultante da integração dos registros de óbito do SIM e dos registros de nascimento do SINASC em busca de extração de padrões de relacionamento de dados que possam descrever a ocorrência de óbitos em crianças de até um ano de idade no Município do Rio de Janeiro através da aplicação de um algoritmo de mineração de dados.

4.1.2. Seleção das Bases de Dados

Os arquivos com os registros de óbito do SIM estão disponibilizados pelo Departamento de Informática do SUS (DataSUS) e contém registros de óbito de todo o território nacional e estão separados por estado e por ano de ocorrência, a partir de 1996 até 2012[6].

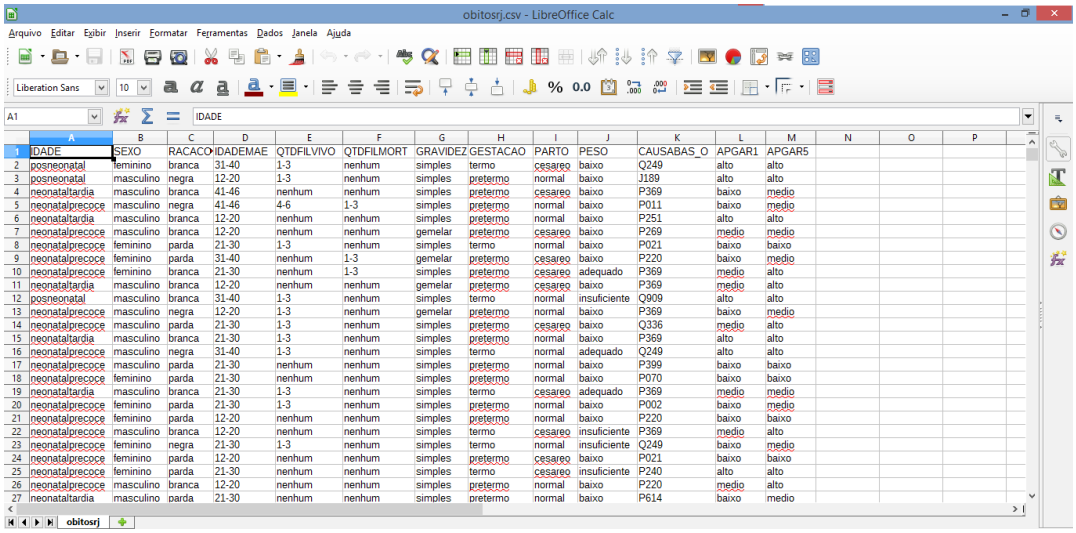
Os arquivos com os registros de nascimento do SINASC estão disponibilizados pelo DataSUS e contém os registros de nascimentos ocorridos em todo o território

6 Até a realização do trabalho não haviam dados do ano de 2013 em diante.

nacional, separados por estado e por ano de ocorrência, a partir de 1996 até 2012[7].

Foram selecionados os arquivos correspondentes aos últimos cinco anos disponíveis de registros de óbito, a partir de 2008 até 2012, e os arquivos de registros de nascimento necessários para fazer a associação com os registros de óbito, a partir de 2008 até 2012, inclusive registros de nascimento do ano de 2007 associados a óbitos que ocorreram no ano de 2008, ambos correspondentes ao estado do Rio de Janeiro.

Tanto os registros de óbito quanto os registros de nascimento são armazenados em formato de banco de dados relacional. Os arquivos encontram-se em formato comprimido ".dbc". Foi utilizado o programa de tabulação para Windows (TabWin) para expandir os arquivos para um formato ".dbf" e, posteriormente, exportados no formato ".csv" para utilização na ferramenta LibreOffice Calc, conforme figura 10.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	IDADE	SEXO	RACACOR	IDADEMAE	QTDFILVIVO	QTDFILMORT	GRAVIDEZ	GESTACAO	PARTO	PESO	CAUSABAS_O	APGAR1	APGAR5
2	posneonatal	feminino	branca	31-40	1-3	nenhum	simples	termo	cesareo	baixo	Q249	alto	alto
3	posneonatal	masculino	negra	12-20	1-3	nenhum	simples	pretermo	normal	baixo	J189	alto	alto
4	neonataltardia	masculino	branca	41-46	nenhum	nenhum	simples	pretermo	cesareo	baixo	P369	baixo	medio
5	neonatalprecoce	masculino	negra	41-46	4-6	1-3	simples	pretermo	normal	baixo	P011	baixo	medio
6	neonataltardia	masculino	branca	12-20	nenhum	nenhum	simples	pretermo	normal	baixo	P251	alto	alto
7	neonatalprecoce	masculino	branca	12-20	nenhum	nenhum	gemelar	pretermo	cesareo	baixo	P269	medio	medio
8	neonatalprecoce	feminino	parda	21-30	1-3	nenhum	simples	termo	normal	baixo	P021	baixo	baixo
9	neonatalprecoce	feminino	parda	31-40	nenhum	1-3	gemelar	pretermo	cesareo	baixo	P220	baixo	medio
10	neonatalprecoce	feminino	branca	21-30	nenhum	1-3	simples	pretermo	cesareo	adequado	P369	medio	alto
11	neonataltardia	masculino	branca	12-20	nenhum	nenhum	gemelar	pretermo	cesareo	baixo	P369	medio	alto
12	posneonatal	masculino	branca	31-40	1-3	nenhum	simples	termo	normal	insuficiente	Q909	alto	alto
13	neonatalprecoce	masculino	negra	12-20	1-3	nenhum	gemelar	pretermo	normal	baixo	P369	baixo	medio
14	neonatalprecoce	masculino	parda	21-30	1-3	nenhum	simples	pretermo	cesareo	baixo	Q336	medio	alto
15	neonataltardia	masculino	branca	21-30	1-3	nenhum	simples	pretermo	normal	baixo	P369	alto	alto
16	neonatalprecoce	masculino	negra	31-40	1-3	nenhum	simples	termo	normal	adequado	Q249	alto	alto
17	neonatalprecoce	masculino	parda	21-30	nenhum	nenhum	simples	pretermo	normal	baixo	P399	baixo	baixo
18	neonatalprecoce	feminino	parda	21-30	nenhum	nenhum	simples	pretermo	normal	baixo	P070	baixo	baixo
19	neonataltardia	masculino	branca	21-30	1-3	nenhum	simples	termo	cesareo	adequado	P369	medio	medio
20	neonatalprecoce	feminino	parda	21-30	1-3	nenhum	simples	pretermo	normal	baixo	P002	baixo	medio
21	neonatalprecoce	feminino	parda	12-20	nenhum	nenhum	simples	pretermo	normal	baixo	P220	baixo	baixo
22	neonatalprecoce	masculino	branca	12-20	nenhum	nenhum	simples	termo	cesareo	insuficiente	P369	medio	alto
23	neonatalprecoce	feminino	negra	21-30	1-3	nenhum	simples	termo	normal	insuficiente	Q249	baixo	medio
24	neonatalprecoce	feminino	parda	12-20	nenhum	nenhum	simples	pretermo	cesareo	baixo	P021	baixo	baixo
25	neonatalprecoce	feminino	parda	21-30	nenhum	nenhum	simples	termo	cesareo	insuficiente	P240	alto	alto
26	neonatalprecoce	masculino	branca	12-20	nenhum	nenhum	simples	pretermo	normal	baixo	P220	medio	alto
27	neonataltardia	masculino	parda	21-30	nenhum	nenhum	simples	pretermo	normal	baixo	P614	baixo	medio

Figura 10: Tela do LibreOffice

O TabWin é um software desenvolvido pelo DataSUS e permite a realização de operações aritméticas, estatísticas, elaboração de gráficos, dentre outras tarefas.

Logo após a escolha das bases de dados que serão utilizadas no trabalho, é necessário escolher os registros que estão dentro do escopo do trabalho.

4.1.3. Seleção dos Registros

O processo de seleção dos registros levou em consideração as duas principais características necessárias à realização do trabalho aplicadas inicialmente sobre os

7 Até a realização do trabalho não haviam dados do ano de 2013 em diante.

registros de óbito do SIM para eliminar os registros que não atendiam ao objetivo do trabalho.

Devido à quantidade de registros de óbito foi aplicado um filtro sobre o campo IDADE na ferramenta LibreOffice Calc para selecionar os registros que caracterizavam óbitos em crianças menores de um ano de idade.

Sobre este conjunto de registros selecionados foi aplicado um segundo filtro no campo CODMUNOCOR para selecionar os registros de óbitos ocorridos no município do Rio de Janeiro. Foram identificados 7.077 (sete mil e setenta e sete) registros de óbito.

Esta etapa é necessária para identificar os respectivos registros de nascimento correlacionando com as informações de óbito possibilitando a criação de uma base de dados única para a realização da etapa de mineração de dados.

4.1.4. Integração entre os Registros

Para a realização desta atividade, o SINASC possui como atributo identificador para seus registros o número da declaração de nascido vivo (NUMERODV) e o SIM, a partir do ano de 2008, possui o mesmo campo para preenchimento em caso de ocorrência de óbito em menor de um ano de idade, o que permite a associação entre os respectivos registros e a criação da base de dados única.

No contexto dos sistemas de informação utilizados pelo SUS, esses dois sistemas representam uma exceção. Isso porque a integração de dados dos SIS, em geral, é algo muito difícil de ser feito diretamente especialmente pela falta de um identificador único para o paciente. O cadastro dos pacientes é feito por diversos sistemas que não se integram gerando um conjunto de bancos de dados desconexos (LEÃO, 2004).

Além disso, o prontuário médico, a principal fonte de informações do paciente no período de internação hospitalar, ainda é utilizado em forma impressa, o que torna o processo de estudo, integração e extração de informações ainda mais difícil.

O fato de ambos possuírem um campo comum permite que sejam utilizados ferramentas ou métodos automáticos para facilitar a tarefa de integração das bases de dados de forma direta entre os respectivos registros.

Os arquivos “.csv” foram importados para o sistema gerenciador de banco de dados MySQL. Foram definidas consultas SQL para relacionar os registros de óbito do

SIM – pelo campo NUMERODV – e os registros de nascimento do SINASC.

Nos casos em que o valor do campo da declaração de nascido vivo não estava preenchido no SIM foi necessário utilizar um conjunto de valores de atributos identificadores secundários, como por exemplo, data de nascimento, sexo, raça/cor, para fazer o relacionamento com os respectivos registros do SINASC.

Considerando que não foi possível relacionar todos os registros de óbito com os respectivos registros de nascimento devido ao campo NUMERODV ou campos identificadores secundários não estarem preenchidos, identificou-se para os registros de óbito suas respectivas informações de nascimento. Assim, foram relacionados 3410 (três mil, quatrocentos e dez) registros.

É notável que houve uma perda considerável de registros, o que demanda um estudo sobre as dificuldades encontradas nesta atividade e o planejamento de ações para um processo mais eficiente de coleta de dados e consequente melhora na qualidade dos registros desses dois sistemas.

Assim foi criada uma base de dados única contendo os registros de nascimento de crianças menores de um ano de idade que foram a óbito no município do Rio de Janeiro.

Continuando a tarefa de preparação da base de dados para a aplicação do algoritmo de mineração de dados, após a seleção dos registros que serão utilizados, podem ser definidos quais os atributos serão importantes para a sequência do trabalho.

4.1.5. Seleção dos Atributos

Após a escolha dos registros que possibilitou a criação da base de dados única utilizada neste trabalho procedeu-se à escolha dos atributos necessários para a realização do trabalho. Foram excluídos os atributos que não atendiam aos objetivos do trabalho como, por exemplo, número da declaração de nascido vivo, data de nascimento, etc.

Foram selecionados os atributos que continham a descrição das características do paciente (menor de um ano de idade), da mãe, da gestação e do parto, demonstrados a seguir.

O campo **SEXO** é um atributo categórico que define o sexo do recém-nascido, conforme abaixo:

- **0:** Ignorado, não informado;
- **1:** Masculino;
- **2:** Feminino;

A escala de *apgar*[*] é o método que é empregado para avaliar o ajuste imediato do recém-nascido à vida extrauterina. Consiste na avaliação de 5 itens do exame físico do recém-nascido. Os aspectos avaliados são: frequência cardíaca, esforço respiratório, tônus muscular, irritabilidade reflexa e cor da pele. Para cada um dos 5 itens é atribuída uma nota de 0 a 2. Somam-se as notas de cada item e o total pode dar uma nota mínima de 0 e máxima de 10. A Tabela 3 demonstra os parâmetros utilizados para a medida do *apgar*.

Pontos	0	1	2
Frequência Cardíaca	Ausente	<100/minuto	>100/minuto
Respiração	Ausente	Fraca, Irregular	Forte/Choro
Tônus Muscular	Flácido	Flexão de pernas e braços	Ativo/Boa Flexão
Irritabilidade Reflexa	Ausente	Algum movimento	Espirros/Choro
Cor	Cianótico/pálido/arroxeado	Cianose das Extremidades (mãos e pés roxas)	Rosado

Tabela 3: parâmetros para pontuação do *apgar*.

O campo **APGAR1** é um atributo numérico que descreve a avaliação do *apgar* no 1º minuto e o campo **APGAR5** descreve a avaliação do *apgar* no 5º minuto.

O campo **PESO** é um atributo numérico que define o peso do recém-nascido no momento do nascimento.

O campo **RACACOR** é um atributo categórico que descreve a raça/cor do recém-nascido, conforme abaixo:

- **1:** Branca;
- **2:** Preta;
- **3:** Amarela;
- **4:** Parda;
- **5:** Indígena;

O campo **IDADE** é um atributo que descreve a idade da criança no momento do óbito, utilizando um sistema de representação composto por dois subcampos. O primeiro, de um dígito, indica a unidade da idade, o segundo, de dois dígitos, indica a quantidade de unidades, conforme os valores a seguir:

* O método de apgar foi desenvolvido pela médica norte-americana Virginia Apgar.

- **0:** Idade ignorada, o segundo subcampo é 00;
- **1:** Horas, o segundo subcampo varia de 00 a 23;
- **2:** Dias, o segundo subcampo varia de 01 a 29;
- **3:** Meses, o segundo subcampo varia de 01 a 11;
- **4:** Anos, o segundo subcampo varia de 00 a 99;
- **5:** anos (mais de 100 anos), o segundo subcampo varia de 0 a 99.

O campo **CAUSABAS** é um atributo alfanumérico que define a causa básica da morte, informada pelo médico responsável pelo atendimento, conforme os códigos constantes da Classificação Internacional de Doenças – 10ª revisão (CID-10).

O campo **IDADEMAE** é um atributo numérico que define a idade da mãe em anos. O campo **QTDFILVIVO** é um atributo numérico que define a quantidade de filhos vivos que a mãe possui considerando as gestações anteriores e excluindo-se a gestação atual. O campo **QTDFILMORT** é um atributo numérico que define a quantidade de filhos mortos que a mãe possui considerando as gestações anteriores e excluindo-se a gestação atual.

O campo **GESTACAO** é um atributo categórico que define em número de semanas a idade gestacional, conforme abaixo:

- **1:** menos de 22 semanas;
- **2:** de 22 a 27 semanas;
- **3:** de 28 a 31 semanas;
- **4:** de 32 a 36 semanas;
- **5:** de 37 a 41 semanas;
- **6:** 42 ou mais semanas;
- **9:** Ignorado;

O campo **GRAVIDEZ** é um atributo categórico que define o tipo de gravidez, conforme abaixo:

- **1:** Única;
- **2:** Dupla;
- **3:** Tripla ou mais;
- **9:** Ignorada;

O campo **PARTO** é um atributo categórico que define o tipo de parto realizado, conforme abaixo:

- **1:** Vaginal (ou Normal);
- **2:** Cesáreo;
- **9:** Ignorado;

A verificação da consistência dos dados informados foi realizada comparando os valores dos campos comuns dos dois sistemas. A mesma estratégia foi utilizada para o preenchimento dos campos dos registros com dados faltantes.

Foram descartados os registros cuja correspondência entre nascimento e óbito não foi possível de ser realizada devido a problemas nos dados armazenados, principalmente, com relação ao campo NUMERODV e quando muitos dos campos restantes que serviriam como parâmetro para relacionar os registros estavam em branco.

Também foram descartados os registros cujos valores coletados pelos dois sistemas eram conflitantes, como por exemplo, informação divergente quanto ao sexo correspondente ao paciente. Ao final, restaram 3.336 (três mil, trezentos e trinta e três) registros para a aplicação da etapa de mineração de dados.

4.1.6. Transformação dos Dados

Após a definição dos registros que serão possíveis de serem utilizados na etapa de mineração foram realizados os ajustes nos atributos. Os atributos categóricos SEXO, RACACOR, GRAVIDEZ e PARTO foram representados pelos nomes de suas categorias e os atributos PESO, IDADE e GESTAÇAO foram discretizados, conforme explicado abaixo:

O atributo **SEXO** foi representado em suas duas categorias nominais:

- masculino;
- feminino;

O atributo **PESO** foi discretizado em quatro categorias nominais:

- baixo: peso até 2.499g;
- insuficiente: peso entre 2.500g e 2999g;

- adequado: peso entre 3000g e 3999g;
- excesso: peso acima de 4000g;

O atributo **RACACOR** foi representado em suas cinco categorias nominais:

- branca;
- preta;
- amarela;
- parda;
- indígena;

O atributo **IDADE** foi discretizado em três categorias nominais:

- neonatal precoce: entre 0 e 6 dias de vida;
- neonatal tardio: entre 7 e 27 dias de vida, inclusive;
- pós neonatal: entre 28 e 364 dias de vida, inclusive.

O atributo **GESTACAO** foi discretizado em três categorias:

- pré-termo: até 36 semanas de gestação;
- a termo: entre 37 e 41 semanas de gestação, inclusive;
- pós-termo: 42 ou mais semanas de gestação;

O atributo **GRAVIDEZ** foi discretizado em três categorias:

- simples: um único filho;
- gemelar: dois filhos;
- trigemelar: três ou mais filhos;

O atributo **PARTO** foi representado em suas duas categorias nominais:

- normal: partos vaginais;
- cesáreo: partos cesarianos;

Os atributos APGAR1, APGAR5, IDADEMAE, QTDFILVIVO e QTDFILMORT foram discretizados conforme abaixo:

Atributos	Cenário 1	Cenário 2
APGAR1 e APGAR5	• Baixo: até 5 pontos; • Médio: 6 e 7 pontos; • Alto: de 8 a 10 pontos;	• zero • 1 e 2 pontos • 3 e 4 pontos • 5 e 6 pontos • 7 e 8 pontos • 9 e 10 pontos
IDADEMAE	• 12 a 20 anos • 21 a 30 anos • 31 a 40 anos • 41 a 46 anos	• 12 a 15 anos • 16 a 19 anos • 20 a 23 anos • 24 a 27 anos • 28 a 31 anos • 32 a 35 anos • 36 a 39 anos • 40 a 43 anos • 44 a 46 anos
QTFILVIVO	• Nenhum • 1 a 3 filhos vivos • 4 a 6 filhos vivos • 7 a 9 filhos vivos • 10 a 12 filhos vivos	• Nenhum • 1 e 2 filhos vivos • 3 e 4 filhos vivos • 5 e 6 filhos vivos • 7 e 8 filhos vivos • 9 e 10 filhos vivos • 11 e 12 filhos vivos
QTDFILMORT	• Nenhum • 1 a 3 filhos mortos • 4 a 6 filhos mortos • 7 e 8 filhos mortos	• Nenhum • 1 e 2 filhos mortos • 3 e 4 filhos mortos • 5 a 8 filhos mortos

Tabela 4: Cenários de discretização dos atributos.

Os valores do atributo CAUSBAS foram apresentados conforme a codificação da tabela de Classificação Internacional de Doenças (CID-10). Em um cenário alternativo (cenário 3), o mesmo atributo foi agrupado conforme a tabela CID-10, por exemplo, o grupo J95 – J99 representa todos os CIDs que iniciam com J95 até os CIDs que iniciam com J99.

Após a exposição da técnica utilizada para a transformação dos dados, a base de dados integrada está pronta para a próxima etapa que é a aplicação do algoritmo de mineração de dados para extração dos padrões de relacionamento que, neste caso, será utilizado o algoritmo *Apriori* para encontrar regras de associação entre o conjunto de dados utilizados.

4.2. Mineração de Dados sobre a Base de Dados Única

Para realizar a extração das regras, conforme definido no objetivo do trabalho pela aplicação do processo de KDD, foi escolhida a técnica descritiva de busca por regras de associação (Seção 3.3.1), foi escolhido o algoritmo Apriori (Seção 3.3.2), utilizando-se a ferramenta WEKA (Seção 3.5.), versão 3.6.12.

4.3. Pós-Processamento de Dados (Regras)

Nesta etapa foram definidas algumas métricas para a redução do número de regras para avaliação no processo final do KDD.

Foram definidas as seguintes regras para exclusão das regras que podem ser consideradas sem muita relevância para a análise dos resultados:

- regras do tipo A → B, em que A e B são conjuntos unitários de dados, por exemplo, {APGAR1 = baixo → PESO = baixo};
- regras do tipo A → B, em que esta regra seja subconjunto de uma outra regra maior X→ Y, em que A ⊂ X e B ⊂ Y, por exemplo, regra 1 = {APGAR1 = alto, APGAR5 = alto → PESO = adequado} e regra 2 = {APGAR1 = alto, APGAR5 = alto, PARTO = normal → PESO = adequado};
- regras que possuíam valores muito óbvios para evidenciar um conhecimento novo, por exemplo: APGAR1, APGAR5 e PESO com valores baixos e IDADE neonatal precoce, por exemplo, {APGAR1 = baixo, APGAR5 = baixo, PESO = baixo → IDADE = neonatal precoce};

A seguir são apresentadas algumas regras geradas após o filtro baseado nas métricas definidas anteriormente e, posteriormente, são tecidos comentários com relação a essas regras contextualizando com algumas observações citadas ao longo do trabalho.

Foram feitos experimentos em três cenários diferentes em que foram utilizadas as formas de discretização dos atributos APGAR1, APGAR5, IDADEMAE, QTDFILVIVO, QTDFILMORT e CAUSABAS, bem como, para cada um dos cenários foram modificados os parâmetros do algoritmo para verificação das regras geradas.

Neste ponto, algumas regras geradas são apresentadas bem como o suporte e a confiança de cada uma delas nos diferentes cenários de discretização dos atributos utilizados neste trabalho:

As regras geradas através da aplicação do algoritmo *Apriori* no cenário 1, conforme explicado anteriormente, pode ser conferido através da Tabela 5, abaixo demonstrada:

	Regras	Suporte	Confiança
Cenário 1	1) Nenhum filho morto de gestações anteriores, Gravidez simples, Parto cesareano, Apgar no 1º minuto alto (de 7 a 10 pontos), associados a Apgar no 5º minuto alto (de 7 a 10 pontos).	10%	100%
	2) Gestação a termo (37 a 41 semanas), Peso adequado (entre 3000g e 3999g), Apgar alto no 5º minuto (de 7 a 10 pontos), associados a Gravidez simples.	10%	99%
	3) Nenhum filho vivo de gestações anteriores, nenhum filho morto de gestações anteriores, Apgar alto no 1º minuto (de 7 a 10 pontos) associados a Apgar alto no 5º minuto (de 7 a 10 pontos).	10%	99%
	4) Mãe com idade entre 12 e 20 anos, nenhum filho vivo, nenhum filho morto, criança com peso baixo (menor que 2500g), associados a Gravidez simples.	12%	91%
	5) Idade neonatal precoce (até 6 dias de vida), raça/cor parda, nenhum filho vivo de gestações anteriores, nenhum filho morto de gestações anteriores, associados a Gravidez simples.	12%	91%
	6) Criança do sexo masculino, nenhum filho vivo de gestações anteriores, Gravidez simples, criança com Peso baixo (menor que 2500g), Apgar no 1º minuto baixo (até 5 pontos) associados a Gestação Pré termo (menos de 36 semanas)	10%	93%
	7) Idade neonatal precoce, mãe com idade entre 12 e 20 anos, nenhum filho morto de gestações anteriores, criança com Peso baixo (menor que 2500g), associados a gravidez simples.	11%	91%

Tabela 5: Regras geradas com aplicação do algoritmo no cenário 1.

As regras geradas através da aplicação do algoritmo *Apriori* no cenário 2, conforme explicado anteriormente, pode ser conferido através da Tabela 6, abaixo demonstrada:

	Regras	Suporte	Confiança
Cenário 2	1) Idade Pós neonatal, Nenhum filho morto em gestações anteriores, Apgar no 5º minuto entre 9 e 10 pontos associados a Gravidez simples.	12%	96%
	2) Mãe com idade entre 16 e 19 anos, nenhum filho vivo de gestações anteriores, Gravidez simples, associados a nenhum filho morto de gestações anteriores.	11%	96%
	3) Mãe com idade entre 20 e 23 anos, nenhum filho morto de gestações anteriores, Gestação pré-termo (menos de 36 semanas), associados a gravides simples.	11%	90%
	4) Raça/cor parda, nenhum filho morto de gestações anteriores, Apgar no 5º minuto entre 9 e 10 pontos, associados a gravidez simples.	12%	94%
	5) Criança do sexo feminino, Gestação Pré-termo (menos de 36 semanas), Parto normal, associados a Peso baixo (menor que 2500g).	15%	96%

Tabela 6: Regras geradas com aplicação do algoritmo no cenário 2.

As regras geradas através da aplicação do algoritmo *Apriori* no cenário 3, conforme explicado anteriormente, pode ser conferido através da Tabela 6, abaixo demonstrada:

	Regra	Suporte	Confiança
Cenário 3	1) Gestação pré-termo (menos de 36 semanas), Parto normal, Apgar no 1º minutos entre 1 e 2 pontos, associados a Peso baixo (menor que 2500g).	15%	98%
	2) Nenhum filho morto, Gestação a Termo (37 a 41 semanas) , Apgar no 5º minuto entre 9 e 10 pontos associados a Gravidez simples.	13%	98%
	3) Raça/cor branca, nenhum filho morto de gestação anteriores, Gestação pré-termo (menos de 36 semanas), Parto normal, associados a Peso baixo (menor que 2500g);	10%	95%
	4) Criança do sexo masculino, nenhum filho morto de gestações anteriores, Apgar no 1º minuto entre 7 e 8 pontos, associados a Gravidez simples.	10%	94%

Tabela 7: Regras geradas com aplicação do algoritmo no cenário 3.

As regras 1 e 2 (cenário 1), regras 1, 2 e 4 (cenário 2), regras 2 e 4 (cenário 3)

demonstram boas condições de nascimento da criança. Este fato fica evidenciado pelas boas condições de saúde da criança ao nascer (alto valor de apgar nos 1º e 5º minutos e peso adequado), bem como, boas condições relacionadas a gravidez/gestação (gravidez simples, gestação a termo) e o fato de não haver ocorrência de óbito em gestações anteriores.

O fato de o nascimento ter ocorrido através de parto cesariano (regra 1, cenário 1) pode indicar alguma complicação no estado de saúde da mãe que de alguma forma pode ter gerado consequência para a saúde da criança.

No caso da regra 3 (cenário 2), a mãe não apresentava nenhum filho morto em gestações anteriores e a gestação foi simples. Nestas regras, a gestação pré-termo, ou seja, menos de 36 semanas de gestação é um ponto que precisa ser melhor investigado para verificar possíveis ocorrências de complicações durante o período gestacional.

No caso da regra 2 (cenário 2), a baixa idade da mãe (entre 16 e 19 anos) é um fator que merece um pouco mais de atenção para a verificação da ocorrência de óbito em crianças com boas condições de nascimento.

As regras 3, 4 e 5 (cenário 1), regra 2 (cenário 2) apresentam como característica similar o fato de o óbito está acontecendo na primeira gestação. Nestas regras, está explícito que não há nenhum filho vivo e nenhum filho morto de gestações anteriores.

A consulta às informações coletadas no prontuário da mãe no momento da internação para o parto pode apontar algumas possíveis causas para a ocorrência do óbito, o que reforça a importância de os sistemas de informação em saúde possuírem um campo identificador para realizar a integração e a necessidade da utilização da forma eletrônica do prontuário para facilitar a consulta.

Tanto o fato da ocorrência do óbito na primeira gestação e de pouca idade da mãe com boas condições de nascimento da criança já sinalizam alguma atenção por parte dos profissionais de saúde com relação a este tipo de gravidez.

As regras 6 e 7 (cenário 1), regra 5 (cenário 2), regras 1 e 3 (cenário 3) apresentam algumas características que podem demonstrar alguma evidência de má condição no nascimento da criança, como baixo peso, baixo apgar, e alguma complicação durante a gestação, no caso, devido a gestação pré-termo.

No caso da regra 6 (cenário 1) e da regra 4 (cenário 3), há padrão de ocorrência de óbito em crianças do sexo masculino. No caso da regra 5 (cenário 2), há padrão de ocorrência de óbito em crianças do sexo feminino.

No caso da regra 7 (cenário 1), um indicativo de óbito nos primeiros 6 dias de

vida evidenciam a importância de um acompanhamento da gestação no período pré-natal. Nestes casos, a análise de informações coletadas no período de realização do pré-natal podem ser utilizadas para enriquecer a caracterização deste tipo de ocorrência de óbito, principalmente para interpretação da ocorrência do óbito na idade neonatal precoce e uma possível identificação de uma gestação de alto risco, conforme dito na seção 2.1.

Após a demonstração e os comentários a respeito das regras, fica evidente a importância da utilização de diversas fontes de sistemas de informação em saúde para uma melhor interpretação dos acontecimentos em um contexto mais amplo.

Considerando a própria natureza do sistema público de saúde brasileiro que permite que um cidadão receba atendimento em mais de um estabelecimento de saúde, é necessário que as informações sejam compartilhadas e acessíveis para melhor fornecer subsídios para o estudo e planejamento de ações que permitam atender de forma efetiva as necessidades da população.

5. Trabalhos Relacionados

Este capítulo tem por objetivo a aplicação do processo de KDD em algumas áreas de saúde de acordo com os objetivos e técnicas escolhidos.

5.1. O Processo de KDD Aplicado à Saúde

O processo de KDD, e dentro dele especialmente a etapa de mineração de dados, é uma excelente alternativa para exploração e análise de dados em busca de respostas fundamentais para o processo de tomada de decisão. Assim, tem aplicação em diversas áreas da atividade humana, inclusive, em diversas situações na saúde para o entendimento e a proposta de solução para os mais variados tipos de problemas.

Realizando um estudo no campo da saúde bucal, (MATTOS, 2004) propôs uma aplicação de mineração de dados para verificar a incidência de cárie dental e traçar um perfil da população composta por crianças de 6 a 12 anos de idade da região sul do país utilizando os dados das escolas públicas das capitais dos estados do sul: Curitiba, Florianópolis e Porto Alegre. Foi utilizada a técnica de clusterização através do algoritmo *k-means* (aprendizado não supervisionado) para realizar o agrupamento de dados em grupos similares.

A base de dados do Sistema de Informações de Agravos de Notificações (SINAN) foi a fonte de dados utilizada por (TRINDADE *et al*, 2004) para traçar o perfil epidemiológico das hepatites virais considerando a população da cidade de Curitiba no ano de 2003, tendo como atributos a idade, vacinação contra a hepatite B, forma clínica, evolução e diagnóstico da doença. Foi utilizado o algoritmo de classificação *C.45* (QUINLAN, 1987) para gerar as regras em forma de árvore de decisão (aprendizado supervisionado).

A descoberta de conhecimento através de regras de associação foi utilizada por (SILVA, 2004) em busca de padrões de relacionamento referentes à situação

socioeconômica dos pacientes vinculados aos procedimentos a que foram submetidos durante as internações hospitalares. Neste trabalho, foram utilizadas a base de dados do CadSUS (Cartão Nacional do Saúde) que registra os dados socioeconômicos da população e a base de dados do CLEITOS (Sistema de Central de Leitos) que registra alguns dados de internações de pacientes. Foi desenvolvida uma ferramenta de mineração de dados baseada no algoritmo *apriori*.

Nestes três exemplos, observa-se uma aplicação da mineração de dados em busca de conhecimento que descreva algum padrão sobre a situação de saúde dos pacientes em busca de respostas para o planejamento de ações específicas para esta faixa da população.

Além destes casos, a mineração de dados foi aplicada para outros problemas relacionados à área da saúde mas que não tinham por objetivo o conhecimento a respeito da saúde dos pacientes para propor tratamentos.

Visando otimizar o processo de auditoria sobre os valores cobrados nas guias de procedimentos hospitalares (BAURAKIADES, 2012) utilizou a mineração de dados para posicionar as guias de cobrança em 3 intervalos os quais determinavam quais faturas seriam encaminhadas para apreciação da auditoria e quais não seriam enviados para auditoria, procedimento necessário para a verificação da conformidade do processo de cobrança de procedimentos.

Considerando que o modelo assistencial do SUS pode ser baseado em gestão de saúde a partir de uma abordagem territorial, (MALUCELLI, 2010) utilizou a mineração de dados para a descoberta de regras referentes às condições do ambiente físico para a classificação de microáreas de risco.

Após as considerações relacionadas ao processo de descoberta de conhecimento, considerando suas atividades e tendo sido apresentada uma ferramenta que possibilita a aplicação do processo de KDD bem como a exposição de exemplos de aplicação de KDD na área da saúde, segue no próximo capítulo a aplicação do processo de KDD nas bases de dados do SIM e do SINASC.

6. Conclusões

6.1. Objetivos do Trabalho

O objetivo principal do trabalho foi definido na seção 1.1 como sendo a aplicação do processo de descoberta de conhecimento para a extração de padrões descritivos de óbitos em crianças de até um ano de idade ocorridos no município do Rio de Janeiro.

Assim, após a definição dos conceitos importantes para o entendimento do trabalho demonstrados nos capítulos 2 e 3, procedeu-se a aplicação do processo nas fontes de dados, descrevendo-se as atividades realizadas.

6.2. Dificuldades Encontradas

A atividade de integração das bases de dados descrita na seção 4.1.4, demonstrou um pouco do problema relacionado à qualidade das informações coletadas por ambos os sistemas determinando a exclusão de um número significativo de registros.

Além disso, a natureza dos sistemas de informação em saúde não contempla, em geral, um campo identificador para facilitar a integração entre eles. Assim, a contextualização de algumas causas de óbito não pode ser totalmente definida com relação à informações do período pré-natal e/ou do período pós-parto.

6.3. Contribuições do Trabalho

Além de demonstrar a possibilidade de geração de regras através de KDD conforme estabelecido no objetivo do trabalho, podem ser citadas as seguintes observações citadas ao longo da realização do trabalho:

- A importância de um campo identificador único que possibilite a integração das

bases de dados dos sistemas de informação em saúde para a realização de estudos em um contexto mais amplo;

- A realização de um estudo das causas que impõem dificuldades na alimentação dos sistemas de informação em saúde para possibilitar a melhora da qualidade das informações coletadas;
- A utilização da forma impressa do prontuário que dificulta a utilização de informações importantes coletadas durante o período de internação do paciente.

6.4. Trabalhos Futuros

A ampliação do contexto de estudo do trabalho com a utilização de outras fontes de dados coletados pelos sistemas de informação em saúde, utilizando técnicas que possibilitem a integração entre os dados, mesmo que não haja um campo identificador comum.

A ampliação do número de registros utilizados no trabalho considerando outros municípios e estados analisando as regras sob o ponto de vista nacional e/ou regional que permita realizar uma comparação entre eles.

Referências

AGRAWAL, R.; SRIKANT, R. - Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

BAURAKIADES, Emanuele; DALAGASSA, Marcelo Rosano; BOHN, Giselle Carla. Modelo de Reconhecimento de Padrões de Valores para a Elaboração de Pacotes Hospitalares. n: XIII Congresso Brasileiro de Informática em Saúde, 2012, Curitiba / PR. XIII Congresso Brasileiro de Informática em Saúde, 2012.

BERRY, Michael J. A.; LINOFF, Gordon. Data Mining Techniques: For Marketing, Sales, and Customer Support. New York: Wiley Computer Publishing, 1997.

BRASIL. Conselho Federal de Medicina. Resolução nº. 1779, de Novembro de 2005. Regulamenta a responsabilidade médica no fornecimento da Declaração de Óbito. Publicado no Diário Oficial da União em 05 de Dezembro de 2005. Disponível em <http://www.portalmedico.org.br/resolucoes/cfm/2005/1779_2005.htm>. Acesso em: 20/03/2015.

BRASIL. Constituição (1988). Constituição da República Federativa do Brasil. Brasília, DF: Senado, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm>. Acesso em 20/03/2015.

BRASIL. Ministério da Saúde. Agenda de Compromissos para a Saúde Integral da Criança e Redução da Mortalidade Infantil. 1ª Edição, 2004.

BRASIL. Ministério da Saúde. Portaria nº. 116, de 11 de Fevereiro de 2009. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/svs/2009/prt0116_11_02_2009.html – acesso em: 20/03/2015.

BRASIL. Lei n.º 8.080, de 19 de setembro de 1990. Lei Orgânica da Saúde. Brasília: Ministério da Saúde, 1990. Dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8080.htm>. Acesso em: 20/03/2015.

BRASIL. Ministério da Saúde. Portaria nº. 72, de 11 de Janeiro de 2010. Estabelece que a vigilância do óbito infantil e fetal é obrigatória nos serviços de saúde (públicos e privados) que integram o Sistema Único de Saúde. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2010/prt0072_11_01_2010.html>. Acesso em: 20/03/2015.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.

MACQUEEN, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations" in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. {{{booktitle}}} 1: 281–297, University of California Press. Página visitada em 2012-11-14.

MALUCELLI, Andreia *et al*. Classificação de Microáreas de Risco com uso de Mineração de Dados. Rev. Saúde Pública vol.44 nº.2 São Paulo. Apr. 2010.

MATTOS, Merisandra Côrtes de; SIMÕES, Priscyla Waleska Targino de Azevedo; SELINGER, Tarcísio Cardoso.Data Mining em saúde bucal por meio da técnica de Clusterização e do Algoritmo *K-Means*. In: IX Congresso Brasileiro de Informática em Saúde, 2004, Ribeirão Preto / SP. IX Congresso Brasileiro de Informática em Saúde, 2004.

MINISTÉRIO DA SAÚDE. Manual de Vigilância do Óbito Infantil e Fetal e do Comitê de Prevenção do Óbito Infantil e Fetal. 2ª Edição, 2009. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/vigilancia_obito_infantil_fetal.pdf>. Acesso em 20/03/2015.

Portal da Saúde – Sistema Único de Saúde. Vigilância do Óbito. <http://svs.aids.gov.br/cgiae/vigilancia/>. Acesso em 20/03/2015.

QUINLAN, J. R.; C4.5: *Programs for machine learning. Morgan Kaufmann Publishers* Inc., San Francisco, CA, USA, 1993.

SILVA, Glauco Carlos. Mineração de Regras de Associação Aplicado a Dados da Secretaria Municipal de Saúde de Londrina – PR. 2005. 76fls. Programa de Pós-Graduação em Ciência da Computação. Instituto de Informática. Universidade Federal do Rio Grande do Sul. Porto Alegre.

MITCHELL, T.M. (1997) – *Machine Learning*. McGraw-Hill Science/Engineering/Math, 432.

TRINDADE, C.M. *et al*. Descoberta de Padrões de Comportamento das Hepatites Virais Aplicando Data Mining. In: CBIS 2004 – IX Congresso Brasileiro de Informática em Saúde, 2004, Ribeirão Preto, 2004.