Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis - IMESA

GLEICE EBILI JUSTE

UMA PROPOSTA DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO REDECA UTILIZANDO A FERRAMENTA *WEKA*

ASSIS

2013

Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis - IMESA

UMA PROPOSTA DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO REDECA UTILIZANDO A FERRAMENTA *WEKA*

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito do Curso de Graduação.

Aluna: Gleice Ebili Juste

Orientador: Dr. Almir Rogério Camolesi

ASSIS

2013

FICHA CATALOGRÁFICA

JUSTE, GleiceEbili

Uma proposta de mineração de dados na base de dados do REDECA utilizando a ferramenta *Weka* – Fundação Educacional do Município de Assis – FEMA – Assis, 2013.
63 p.

Orientador: Dr. Almir Rogério Camolesi
Trabalho de Conclusão de Curso – Instituto Municipal de Ensino Superior de Assis – IMESA.

1. Mineração de Dados 2. REDECA

CDD: 001.6
Biblioteca da FEMA

UMA PROPOSTA DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO REDECA UTILIZANDO A FERRAMENTA *WEKA*

GLEICE EBILI JUSTE

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito do Curso de Graduação, analisado pela seguinte comissão examinadora:

Orientador: Dr. Almir Rogério Camolesi
Analisador (1): Me. Fábio Eder Cardoso
Analisador (2): Esp. Guilherme de Cleva Farto

DEDICATÓRIA

Aos meus pais,
Antonio Aparecido Juste e Maria Cleunice Rosa Juste,
pelo amor incondicional e por sempre acreditarem em mim.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por tudo. Pela vida, pela força e por ter colocado pessoas maravilhosas no meu caminho que me ajudaram nessa jornada. Se eu cheguei até aqui, foi porque Ele me trouxe. Se eu conquistei algo, foi porque Ele me deu.

Agradeço a minha família, meu pai, minha mãe e meu irmão, por todo amor, paciência, apoio e por sempre estarem ao meu lado em todos os momentos. Agradeço também aos demais familiares por todo carinho.

Aos meus amigos e colegas eu agradeço imensamente pela amizade, pelo apoio e por tornar a minha vida mais feliz.

Agradeço a equipe da Associação Filantrópica Nosso Lar e do Projeto Rede Ciranda pela oportunidade concedida.

Ao professor, orientador e amigo, Dr. Almir Rogério Camolesi, pela orientação, pelo apoio, pelas palavras amigas quando precisei e também por sempre acreditar no meu potencial.

E a todos que contribuíram, diretamente ou indiretamente, para a realização desse sonho, não só esse trabalho, mas minha formação acadêmica.

A todos vocês, serei eternamente grata.

(...) Nunca deixe que lhe digam que não vale a pena
Acreditar no sonho que se tem
Ou que seus planos nunca vão dar certo
Ou que você nunca vai ser alguém (...).

(Renato Russo, 1986)

RESUMO

A existência de diagnósticos que mostrem a realidade da criança e do adolescente no município é fundamental para tornar realidade os princípios e objetivos estabelecidos no Estatuto da Criança e do Adolescente. Tais diagnósticos são fundamentais para ter uma visão dos atendimentos e da real situação do município.

Para guardar os dados referentes à criança e ao adolescente da cidade de Assis/SP, os atores do Sistema de Garantias de Direitos da Criança e do Adolescente – SGDCA contam com o sistema REDECA, que permite armazenar todas as informações referentes ao atendimento e à situação que vive as crianças e adolescentes do município.

Nesse contexto, os conceitos de Mineração de Dados aplicados na base de dados do REDECA possibilitarão o acesso a informações fundamentais ao diagnóstico municipal. Tais informações poderão servir de base para a tomada de decisões e o planejamento estratégico das mesmas.

Palavras Chave: Mineração de Dados, REDECA.

ABSTRACT

The diagnosis of the existence that shows the child's and teenager's reality in the city are essential to became reality the main objective established on child's and teenager's constitution .These diagnosis are essentials to get an understand view and the real city's situation.

To save data relative for child and teenager - SGDCA count with the REDECA system which allow save all information relative to the treatment and the situation that the children and teenager live in the city

On these contexts the concept of data mining applied on the REDECA data bases will allow the access to the fundamental information. This information may help to take decisions in its strategic planning.

Key Words: Data Mining, REDECA

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 – INTRODUÇÃO

> É no município que se articula a proteção integral da criança e do adolescente. É para onde deve convergir o diálogo entre todas as instâncias governamentais e não governamentais voltadas para esse propósito. Cada cidade deve buscar o fortalecimento da rede de assistência e garantias de direitos, para que esse esforço se traduza na definição de políticas públicas eficazes e num atendimento de qualidade, objetivando um desenvolvimento maior. (Ribas Junior et al. 2011, p. 5)

Segundo Ribas (2011), o Conselho Municipal dos Direitos da Criança e do Adolescente (CMDCA)[1] deve instaurar um processo tecnicamente qualificado, participativo e transparente de diagnóstico da realidade, definição de prioridades, proposição de ações que respondam às necessidades diagnosticadas e acompanhamento da inclusão de programas de ação no ciclo orçamentário municipal.

Ribas (2011) também traz que para tornar realidade em todo o Brasil os princípios e objetivos estabelecidos no Estatuto da Criança e do Adolescente (ECA) muitos obstáculos ainda precisam ser removidos, sendo que entre os mais importantes está a ausência de diagnósticos mais completos e detalhados sobre as realidades locais que fundamentem a formulação de políticas consistentes. Para mobilizar forças e a articulação entre o Estado e a sociedade civil em busca de melhorias das condições de vida das crianças e adolescentes faz-se necessário bom diagnóstico que revele como os problemas se manifestam em cada contexto.

Para guardar os dados referentes à criança e adolescente da cidade de Assis, os atores do Sistema de Garantias dos Direitos da Criança e do Adolescente (SGDCA) contam com o sistema REDECA[2]. O nome REDECA surgiu da união dos termos Rede e ECA.

O REDECA foi desenvolvido pela Fundação Telefônica em parceria com 8 municípios paulistas e tem por objetivo unificar os registros referentes ao atendimento da criança e do adolescente nas organizações que fazem parte do SGDCA. No município de Assis, o REDECA está em funcionamento desde 2010 em 16 instituições de atendimento à criança e ao adolescente.

---

1. www.cmdca-assis.org.br
2. http://www.redeca.org.br

No entanto, além da preocupação em guardar os dados há também a preocupação do que se fazer com eles e como transformá-los em informações de auxílio às tomadas de decisões.

No final da década de 80 foi proposto por Usama Fayyad a Mineração de Dados, do inglês *Data Mining*. Fayyad (1996, *apud* Navega, 2002, p. 1), definiu como "o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

Para Fayyad (1999, *apud* Camilo e Silva, 2009, p. 3), o modelo tradicional para transformação dos dados em informação, consiste em um processamento manual de todas essas informações por especialistas que, então, produzem relatórios que deverão ser analisados. Na grande maioria das situações, devido ao grande volume de dados, esse processo manual torna-se impraticável. A tentativa de solucionar esse problema é o uso do *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento nas Bases de Dados.

O KDD refere-se ao processo de transformação de dados em informações, já o termo Mineração de dados, a somente uma das etapas desse procedimento.

1.1 – OBJETIVO

Este trabalho tem por objetivo realizar um estudo dos conceitos de Mineração de Dados e propor sua aplicação na base de dados do sistema REDECA para auxiliar no diagnóstico da realidade da criança e do adolescente da cidade de Assis.

1.2 – JUSTIFICATIVAS E MOTIVAÇÕES

A utilização de um sistema de diagnóstico por organizações como CMDCA/Assis, Poder Público e Entidades Sociais é fundamental para a visão dos atendimentos realizados e da real situação dos municípios. Com isso essas organizações podem traçar metas, projetar ações e intervir no andamento das atividades desenvolvidas.

Nesse contexto, os conceitos de Mineração de Dados aplicados na base de dados do REDECA tendem a beneficiar o conselho, entidades envolvidas, associações e empresas que atuam na área e possibilitará acesso a informações específicas para uma determinada situação.

Tais informações podem servir de base para a tomada de decisões e o planejamento estratégico das mesmas, contribuindo assim com o crescimento dos envolvidos.

1.3 – PERSPECTIVAS DE CONTRIBUIÇÃO

Umas das expectativas é que este trabalho contribua com aumento de material bibliográfico para trabalhos futuros na área de Mineração de Dados com a ferramenta *Weka*.

Outra perspectiva de contribuição serão os padrões descobertos com a aplicação da técnica de mineração de dados na base de dados do REDECA que irá colaborar com tomadas de decisões em relação à criança e ao adolescente de Assis.

1.4 – METODOLOGIA DE PESQUISA

A metodologia de pesquisa adotada para este trabalho é a experimental. Será dada continuidade ao levantamento e leitura de material bibliográfico e fundamentado nas leituras realizadas e levantamento de requisitos, será definido um estudo de caso para demonstrar a utilização dos conceitos estudados. Por fim, será aplicada a técnica de mineração de dados na base de dados do sistema REDECA.

1.5 – ESTRUTURA DO TRABALHO

Este trabalho será divido em cinco capítulos como serão descritos a seguir:

No Capítulo 1 foi apresentada a importância do diagnóstico da realidade da criança e do adolescente de Assis/SP, também foram apresentados os objetivos deste trabalho, as justificativas, as motivações, as perspectivas de contribuição e a metodologia adotada.

Os conceitos Descoberta de Conhecimento nas Bases de Dados e a descrição das suas etapas serão apresentados no Capítulo 2. Este capítulo também contará com uma seção que apresentará algumas ferramentas para a aplicação da Mineração de Dados e detalhes da ferramenta escolhida para a realização deste trabalho.

Já no Capítulo 3 será apresentado o Projeto Rede Ciranda da Criança e Adolescente de Assis, mostrando a importância deste Projeto e qual o seu papel na Rede de Atendimento à Criança e ao Adolescente.

No Capítulo 4 será abordado todo andamento da aplicação do processo de descoberta de conhecimentos na base de dados do REDECA, como foi aplicar cada etapa, as dificuldades encontradas e outros detalhes.

As considerações Finais serão apresentadas no Capítulo 5.

Para concluir, serão apresentadas todas as referências bibliográficas dos materiais (livros, artigos, entre outros) utilizados neste trabalho e também será anexado ao final o cronograma de todas as atividades desenvolvidas.

2 – MINERAÇÃO DE DADOS E DESCOBERTA DE CONHECIMENTO NAS BASES DE DADOS

Tão importante quanto guardar dados é saber o que fazer com eles. Há a necessidade de transformar estes dados em informações de apoio a tomadas de decisão, que poderão ser usadas para melhorar procedimentos, detectar tendências e características disfarçadas, e até prevenir ou reagir a um evento que ainda está por vir. (Carvalho, 2003).

A figura 2 demonstra o processo natural de busca de informação.

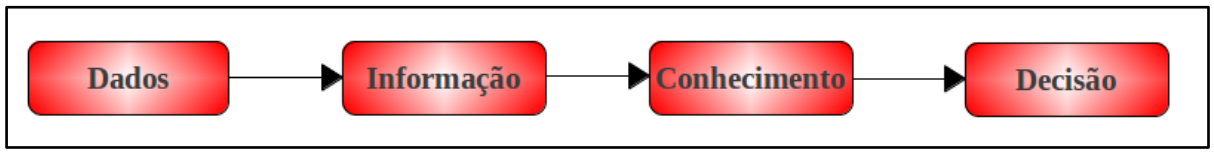


Figura 1 - Processo natural de busca de informação. Inspiração Vasconcelos (2004).

A descoberta de novas informações em termos de padrões ou regras com base em grandes quantidades de dados recebe o nome de Mineração de Dados e para ser útil precisa ser executada de maneira eficiente. (Elmasri e Navathe, 2011).

Mineração de Dados se refere apenas a um dos passos no processo de descobrimento de conhecimento em base de dados. KDD ou, em português, Descoberta de conhecimento em banco de dados, é o termo correto a ser usado para definir todo esse procedimento. A etapa da Mineração de Dados é a aplicação de algoritmos específicos para extrair padrões/modelos de dados e quando aplicada de qualquer maneira e sem o devido preparo poderá conduzir a descoberta de padrões sem nenhum sentido e não levará a descoberta de conhecimento.

Segundo Elmasri e Navathe (2011), a descoberta de conhecimento em Banco de Dados abrange mais do que a Mineração de Dados. O procedimento compreende seis fases: seleção de dados, limpeza de dados, enriquecimento de dados, transformação ou codificação de dados, mineração de dados e relatório e exibição das informações descobertas.

Há certa divergência em relação ao número exato de etapas. Diferentes autores definem um número variável de etapas para descreverem o processo de KDD, porém podemos destacar

três etapas principais e fundamentais, sendo elas: pré-processamento, mineração dos dados e avaliação. (Barioni e Traina Junior, 2002).

A figura abaixo ilustra as etapas principais de KDD.

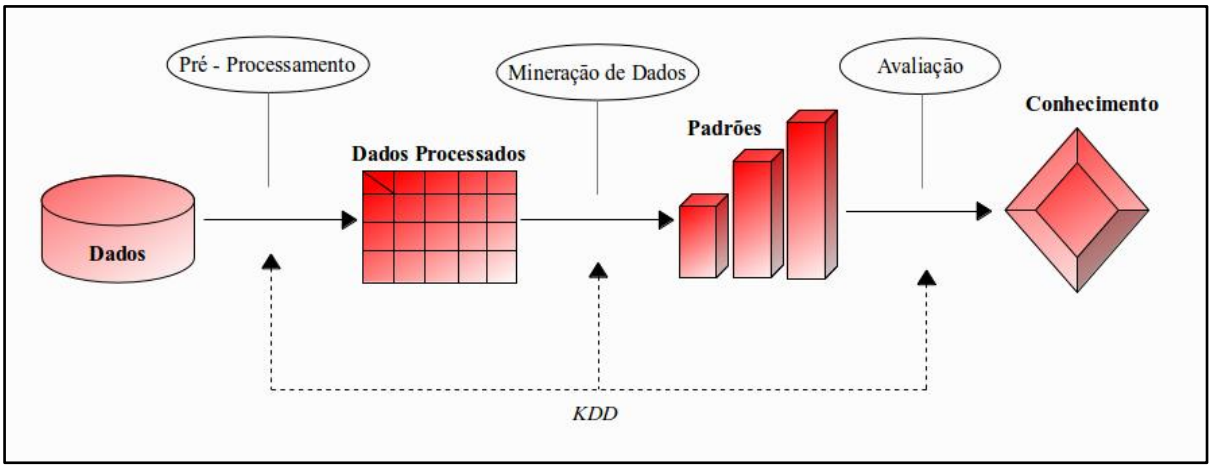


Figura 2 - Principais etapas do KDD.

2.1 – PRÉ-PROCESSAMENTO OU PREPARAÇÃO

É fundamental conhecer o tipo dos dados que serão trabalhados para a escolha adequada da técnica de mineração que será aplicada. Os dados podem ser categorizados em dois tipos: quantitativo, onde são representados por valores numéricos ou qualitativos, onde contêm valores nominais e ordinais (categóricos). No entanto, antes de aplicar os algoritmos de mineração é necessário explorar, conhecer e preparar os dados. (Camilo e Silva, 2009).

Na primeira fase de KDD, os dados serão preparados para serem aplicados às técnicas de mineração de dados. Esta etapa divide-se em três passos, onde os dados serão selecionados, preparados e transformados. A figura 4 representa o procedimento de preparação dos dados:

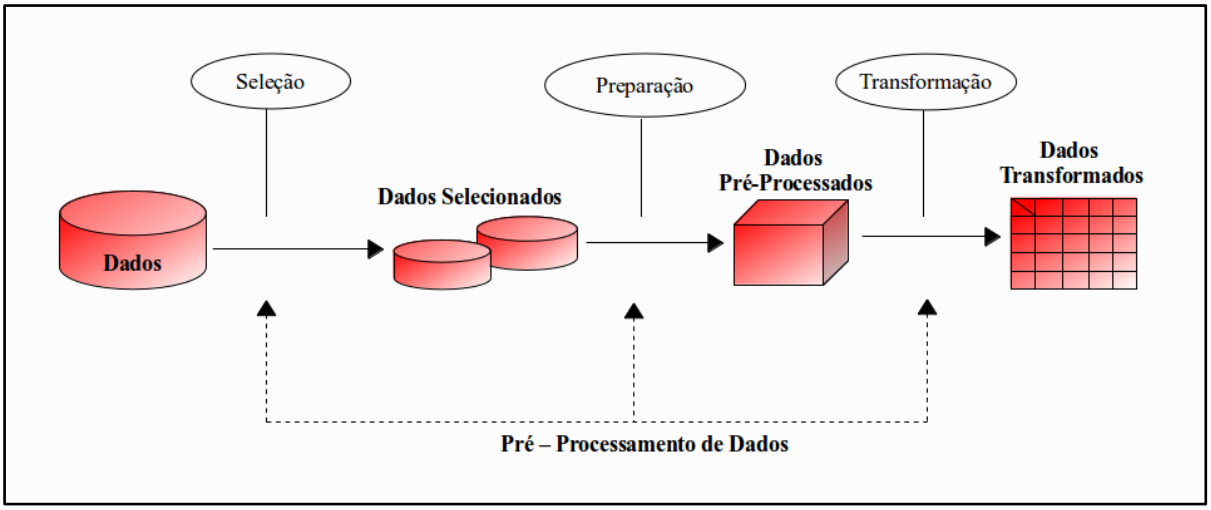


Figura 3 - Etapas do pré-processamento dos dados.

A seguir será descrito cada passo que deverá ser feito para obter dados prontos para a aplicação da técnica de mineração.

2.1.1 – Seleção

Segundo Date (2003), a Seleção "é o processo de capturar dados de banco de dados operacionais e outras fontes". (Date, 2003, p.600).

Este procedimento tende a ser muito trabalhoso e com isso interferir **e**m operações de missão crítica, por esse motivo, essa etapa geralmente é realizada em paralelo (como um conjunto de subprocessos paralelos) e em nível físico. As chamadas "extrações físicas" podem causar sérios problemas para o processamento subsequente, pois podem perder muitas informações (geralmente informações sobre relacionamentos) representadas de algum modo físico (por ponteiros ou por proximidade física, por exemplo). Por esse motivo, as ferramentas de extração às vezes oferecem um meio para conservar tais informações, introduzindo número de registros sequenciais e substituindo ponteiros por chaves estrangeiras. (Date, 2003).

Os dados são armazenados em bases de dados operacionais que são utilizadas por sistemas de informação das empresas/instituições e na maioria das vezes não obedecem às exigências definidas pelo domínio apresentado. É necessário juntar essas informações em uma base de dados centralizada, porém, é uma tarefa muito trabalhosa, já que pode envolver dados de

baixo nível em tabelas relacionais ou conjunto de elementos hierárquicos em sistemas relacionais. Quando a base de dados que será trabalhada é muito extensa, recomenda-se trabalhar com amostragem de dados. Essas amostras devem ser selecionadas de maneira randômica, a fim de obter uma ideia do que pode ser esperado. (Cruz, 2000, apud Carvalho *et. al.*, 2003).

Para Batista (2003), a coleta de dados é considerada uma das fases mais trabalhosas de todo processo de KDD, pois essa fase envolve frequentemente extrair dados de sistemas computacionais antigos que geralmente não possuem documentação nem do projeto e nem da arquitetura do sistema. Dessa forma, por mais conhecimento que se tenha na base de dados a ser trabalhada, pode-se não saber onde e de qual forma determinadas informações foram armazenadas.

Uma forma de amenizar os desafios para coletar os dados é utilizar o *Data Warehouse*. "*Data Warehouse* é um repositório de dados geralmente construído para dar suporte às pessoas que tomam decisões, tais como gerentes e diretores" (Batista, 2003, p.38). Essa tecnologia tem sido muito utilizada, pois "os bancos transacionais não são considerados adequados para fornecer respostas para análises estratégicas".(Batista, 2003, p.38). Frequentemente esses modelos de bancos, principalmente de projetos mais antigo, apresentam diversos problemas como, por exemplo: falta de documentação do projeto, problemas com inconsistências e integridade dos dados, entre outros. O *Data Warehouse* é atualizado frequentemente com dados de sistemas transacionais e/ou fontes externas. Antes de serem carregados no *Data Warehouse* os dados extraídos de diferentes bancos de dados integrados e sua consistência é verificada. "Dessa forma, o *Data Warehouse* pode ser uma boa fonte de dados para um projeto de KDD". (Batista, 2003, p. 38).

2.1.2 – Preparação

Esta etapa pode ser chamada também de limpeza dos dados e tem por objetivo eliminar ruídos (dados estranhos e/ou inconsistentes), registros incompletos ou repetidos e ainda resolver problemas de tipagem de dados. O nível de ruídos encontrados no conjunto de dados influencia diretamente em sua qualidade. Esses ruídos podem estar relacionados a erros de

digitação, transmissão de dados que não contém informações suficientes, atributos que irrelevantes a modelagem e/ou dados não atualizados. (Carvalho, 2003).

Quando está trabalhando com pequenas quantidades de dados, onde todos os exemplos são importantes, é necessário substituir os ruídos por valores consistentes ou então gerar dados manualmente. No caso de grandes quantidades de dados, elimina-se (sempre que possível) os dados com ruídos. Em ambas as situações são necessárias utilizar técnicas estatísticas que irá detectar os campos com ruídos e substituí-los ou desconsiderá-los. (Gurek, 2001, *apud* Carvalho, 2003).

Segundo Camilo (2009), esta etapa irá eliminar esses problemas (ruídos) de maneira que não interfiram no resultado dos algoritmos utilizados. As técnicas utilizadas poderão ser: remoção do registro com problemas, atribuição de um valor padrão e até aplicação de técnicas de agrupamentos para auxiliar na descoberta dos melhores valores.

2.1.3 – Transformação

Para facilitar a aplicação da mineração de dados, após passar pela fase de preparação, os dados deverão passar por uma transformação para serem armazenados de maneira adequada. Há diversos tipo de algoritmos e cada um aceita um tipo de entrada diferente, há dados que necessitam ser convertidos, pode haver a necessidade de criar de novas variáveis e categorização de variáveis contínuas. Essa etapa é necessária quando os processos de mineração de dados não estão acoplados ao sistema de bancos de dados. “Em algumas aplicações, ferramentas avançadas de representação de conhecimento podem descrever o conteúdo de um banco de dados por si só, usando esse mapeamento como uma meta-camada para os dados”. (Carvalho, 2003, p.4)

Para Camilo (2009), a etapa de transformação merece destaque, pois alguns algoritmos de mineração de dados trabalham com tipos de dados específicos, como por exemplo, existem algoritmos que trabalham somente com tipos de dados numéricos e outros somente com valores categóricos. Quando isso ocorre, é necessário fazer uma transformação dos dados. Algumas técnicas podem ser aplicadas, como por exemplo:

- Suavização: os valores errados dos dados são removidos;

- Generalização: os valores muito específicos são convertidos para valores mais genéricos;
- Agrupamento: os valores são agrupados em faixas sumarizadas;
- Normalização: as variáveis são colocadas em uma mesma escala;
- Criação de novos atributos: atributos são gerados a partir de outros já existentes.

Nesta etapa de transformação, o uso de *Data Warehouse* aumenta, pois com essa estrutura as informações são armazenadas de uma maneira mais eficiente. O *Data Warehouse* coleta dados de diversas aplicações, integra-os e organiza-os em áreas lógicas, armazena as informações de formas acessíveis e compreensíveis e as disponibilizam para que possam receber as técnicas de análise e extração de dados. (Lemos, 2003).

2.2 – MINERAÇÃO DE DADOS

Somente depois de passar pelo pré-processo é que finalmente os dados estão prontos para receberem a mineração de dados para extrair diferentes regras e padrões.

Segundo Vasconcelos (2004), a mineração de dados é o núcleo do processo de descoberta de conhecimento em bases de dados e consiste no processo de analisar grandes quantidades de dados em diferentes perspectivas, para extrair informações que normalmente não estão visíveis ou que dificilmente são encontradas.

2.2.1 – Técnicas de mineração de dados

A seguir será descrito as principais técnicas utilizadas na mineração de dados, que na maioria das vezes, são aplicadas repetidamente a fim de descobrir padrões e regras escondidos entre os dados.

2.2.1.1 – Classificação

Considerada uma das técnicas mais comuns, a Classificação, tem por objetivo identificar a qual classe pertence determinado registro. As características de um objeto são examinadas e atribui-se a ele uma classe pré-definida. Esta tarefa tem por objetivo construir modelos que permitam o agrupamento de dados em classes. Uma vez que as classes são definidas, pode-se prever a classe de um novo dado. (Garcia, 2008).

Amorim (2006, p. 23), disse que "essa técnica pode ser utilizada tanto para entender dados existentes quanto para prever como novos dados irão se comportar".

Alguns exemplos de aplicação da técnica de Classificação: para diagnosticar a presença de uma determinada doença, para identificar transações financeiras ilegais ou suspeitas de fraudes, identificar quando uma pessoa pode apresentar ameaças em questão de segurança ou detectar possíveis consumidores para determinados produtos.

2.2.1.2 – Estimativa ou Regressão

Muito similar a Classificação, a estimativa é usada quando o registro é identificado por um valor numérico. Sendo assim, é possível estimar o valor de uma determinada variável após análise dos valores das demais variáveis.

A estimativa pode ser aplicada quando se deseja, por exemplo, estimar a quantia gasta por um casal em um restaurante em determinada data.

2.2.1.3 – Previsão

Visa prever o valor futuro de determinado atributo a partir do histórico de dados. Só há uma maneira de avaliar se a previsão foi bem feita ou não: aguardar o acontecimento e comparar com a previsão. A previsão é sem dúvida a tarefa mais difícil, não só na Mineração de Dados, mas também no dia a dia das pessoas. (Amorim, 2006).

Esta técnica pode ser utilizada para prever se o índice Bovespa subirá ou descerá, o número de

clientes que uma empresa ganhará em determinado evento, a população de uma determinada cidade após cinco anos, entre outros exemplos.

2.2.1.4 – Associação

Nesta técnica procura-se estabelecer regras que ligam um conceito ao outro identificando quais atributos estão relacionados.
A associação obedece à regra SE atributo X ENTÃO atributo Y, onde o atributo Y ocorrerá em consequência do atributo X. (Navega, 2002).

Muito utilizada no comércio, pois com a associação é possível identificar quais produtos são comprados em conjunto. Sempre que o cliente leva o produto X ele também leva o produto Y. Com as informações obtidas facilita a elaboração de promoções e a organização dos produtos no estabelecimento.

2.2.1.5 – Série Temporal

A análise de série temporal tem por objetivo analisar grandes séries de dados a fim de encontrar regularidades, sequencias, entre outras características, e levantar padrões sequenciais, tendências e divergências entre os dados.

Na série temporal a preocupação é detectar ligações entre itens ao longo do tempo, por exemplo, um consumidor que compra um computador em uma determinada data comprará também uma impressora em seis meses. Isso permite o planejamento de uma promoção desse produto para atender a todos que estão nesta situação. (Navega, 2002)

2.2.1.6 – Agrupamento

Esta técnica consiste em identificar possíveis agrupamentos de dados semelhantes. Um agrupamento é um conjunto de registros semelhantes entre si, mas diferentes dos outros registros nos demais agrupamentos. A diferença entre a técnica de agrupamento e a técnica de

classificação é que o agrupamento não necessita que os dados sejam categorizados, como ocorre na classificação. O agrupamento somente irá identificar grupos de dados similares, sem a pretensão de classificar ou estimar o valor de uma variável. As aplicações de agrupamento são variadas e inúmeras, um exemplo do uso do agrupamento é na segmentação de mercado para determinado nicho de produtos ou então para detectar comportamentos atípicos como fraudes. O agrupamento de dados pode também, ser usado na fase de preparação dos dados. (Camilo, 2009).

2.3 – PÓS-PROCESSAMENTO OU INTERPRETAÇÃO

Há inúmeras propostas para a mineração de dados, porém tão importante quanto minerar os dados é saber o fazer com os resultados obtidos.

O pós-processamento irá analisar os padrões apontados pela mineração e procurar identificar se os mesmos oferecem a solução do problema que motivou a realização do processo de KDD. Na maioria das vezes esta etapa se aplica, pois o volume de padrões obtido é tão extenso que dificulta a sua análise, inviabilizando assim, seu uso no apoio a tomadas de decisões, que é um dos maiores objetivos de quem recorre ao procedimento de KDD. (Calil *et. al.*, 2008).

Segundo Takamoto (2011), a fase de pós-processamento consiste não só em analisar e interpretar os modelos de conhecimentos gerados na mineração de dados, mas também elaborar gráficos, diagramas e/ou relatórios demonstrativos.

2.4 – FERRAMENTAS PARA MINERAÇÃO DE DADOS

Há inúmeras ferramentas disponíveis no mercado para auxiliar na aplicação da mineração de dados e tornar a tarefa menos técnica. A seguir serão apresentadas algumas opções dessas ferramentas:

- *Entreprise Miner:* desenvolvida pela SAS, para plataformas UNIX, *Windows* e *Mac;*
- *Pentaho:* desenvolvida pela *Pentaho Corporation*, para plataformas UNIX, *Windows* e *Mac;*

- *Intelligent Miner:* desenvolvida pela IBM, para plataforma UNIX;
- *Business Miner:* desenvolvida pela *Business Objects*, para plataforma Windows;
- *Data Survevor:* desenvolvida pela *Data Distilleries*, para plataforma UNIX;
- *MineSet:* desenvolvida pela *Purple Insight*, para plataforma UNIX;
- *Weka:* desenvolvida pela Universidade de *Waikato*, para plataformas *Windows*, UNIX e *Mac*.

Dentre as ferramentas citadas, a escolhida para este trabalho será a ferramenta *Weka* (*Waikato Environment for Knowledge Analysis*), pois é uma ferramenta livre, oferece uma série de algoritmos para as tarefas de mineração que podem ser aplicados diretamente da ferramenta ou por programas em Java (Camilo, 2009). A seguir a ferramenta *Weka* será apresentada com maiores detalhes.

2.5.1 – *Weka*

Desenvolvida pela Universidade de *Waikato* na Nova Zelândia, foi implementada pela primeira vez em 1997. A ferramenta *Weka* é uma ferramenta livre, ou seja, atende as especificações da GPL (*General Public License*), desenvolvida na linguagem Java, possui interface gráfica para interagir com arquivos de dados e produzir resultados visuais, possui uma API para que possa incorporá-la em sua aplicação permitindo assim que as tarefas de mineração de dados sejam automatizadas. (Abernethy, 2010). Além do modo gráfico, a ferramenta também possui o modo console, onde os algoritmos aplicados e os arquivos com os dados são chamados por linha de comandos como mostra a figura abaixo.

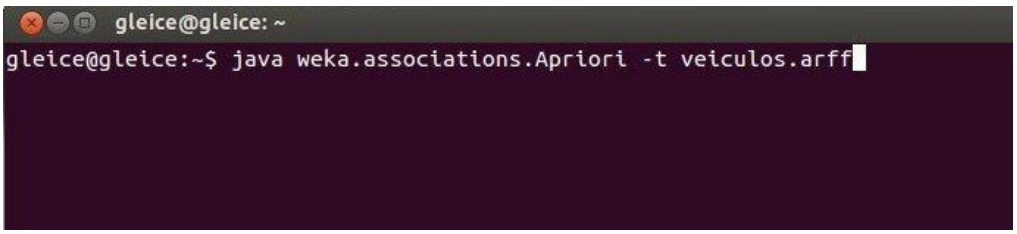


Figura 4 - Uso da ferramenta *Weka* modo Console.

Além de muito utilizada no meio acadêmico, a ferramenta *Weka* é uma das melhores opções para profissionais que desejam aprender os conceitos básicos de mineração de dados. Por meio de sua interface gráfica, também conhecida como *Weka Explorer*, é possível aplicar as técnicas de mineração de forma simples, realizar a avaliação dos resultados obtidos, fazer uma comparação de algoritmos e também executar tarefas relacionadas ao pré-processamento de dados, como por exemplo, seleção e transformação de atributos. (Gonçalves, 2012).

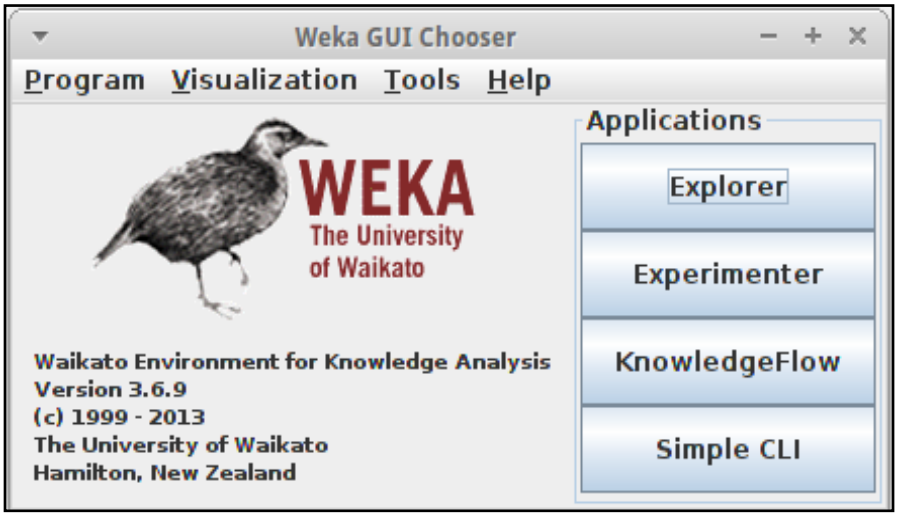


Figura 5 - Tela inicial da ferramenta *Weka*

A figura 5 ilustra a tela de inicial da ferramenta *Weka*, já a figura 6, a tela principal, onde o usuário irá carregar o arquivo ARFF com a base de dados que será utilizado e escolher qual técnica e qual algoritmo será aplicado.

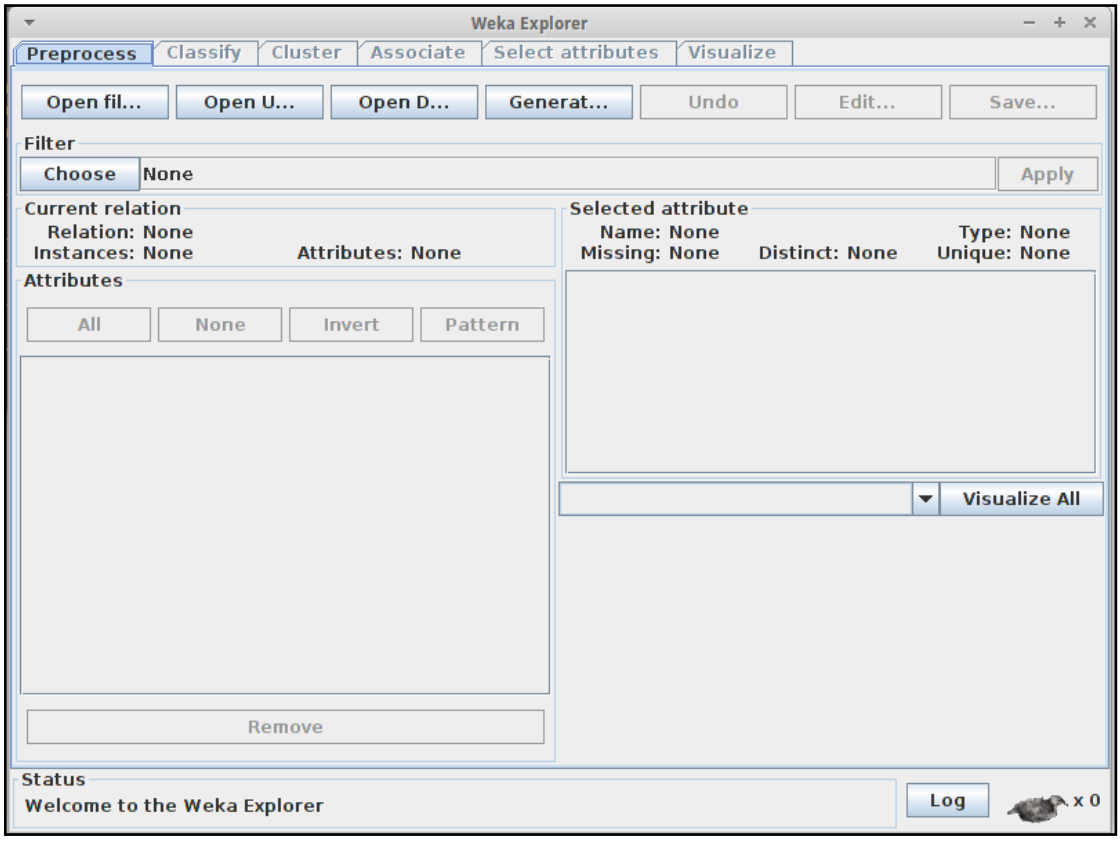


Figura 6 - Tela principal da ferramenta *Weka*.

A ferramenta trabalha preferencialmente com arquivos textos no formato ARFF (*Atribute Relation File Format*). Este formato é utilizado como padrão para estruturar as bases de dados que serão manipuladas pela *Weka*.

O arquivo ARFF é composto por um cabeçalho e um conjunto de instâncias. O cabeçalho contém a declaração da relação que o arquivo representa (@relation), uma lista de atributos (@attribute) e a relação dos valores que cada atributo pode assumir. Os dados são precedidos da *tag* @data, cada linha representa uma instância e os valores dentro de cada instância deve ser separados por uma vírgula. (Perboni, 2013).

A figura 7 mostra a estrutura de um arquivo ARFF.

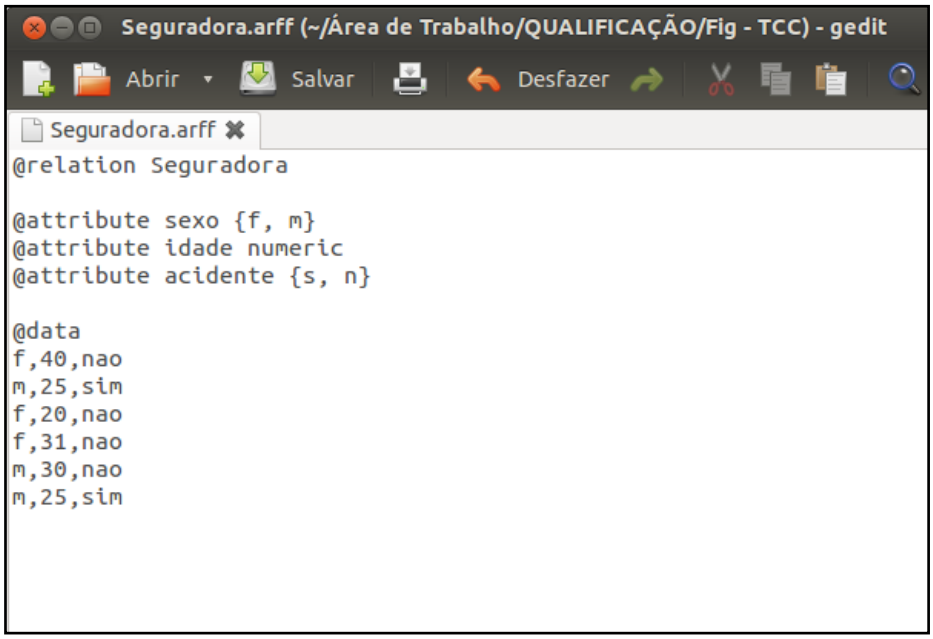


Figura 7 - Estrutura de um arquivo ARFF.

Neste capítulo foram abordados os conceitos de mineração de dados, detalhando suas principais etapas e as técnicas de mineração. Foram abordadas, também, algumas ferramentas utilizadas para tais tarefas, em específico a ferramenta *Weka*, bem como o seu funcionamento.

3. PROJETO REDE CIRANDA DA CRIANÇA E ADOLESCENTE DE ASSIS

O Projeto Rede Ciranda da Criança e Adolescente de Assis teve início no ano de 2010 por meio de uma parceria entre a Associação Filantrópica Nosso Lar e a Fundação Telefônica com o objetivo inicial de promover um diagnóstico da realidade da criança e adolescente de Assis. Porém, foram detectadas algumas lacunas em relação aos serviços existentes. Dentre essas lacunas, a necessidade de um alinhamento entre as organizações e ações que eram desenvolvidas. Com isso, surgiu então a proposta do trabalho em rede, onde o objetivo central é fortalecer a base de atendimento ao público infanto-juvenil.

O Projeto faz a integração das áreas de assistência social, educação, saúde, segurança pública e organizações do terceiro setor. Com essa integração, o Rede Ciranda oferece alguns benefícios às entidades associadas, dentre eles, cursos de capacitação e encontros para que seja potencializado a atuação das entidades e projetos sociais, melhoria da atuação do Poder Público, otimização e soluções e realizar o diagnóstico sobre a realidade da criança e do adolescente de Assis que sustenta e apóia o CMDCA na formulação de políticas públicas. (Xavier, *et. al.*, 2012).

Várias instituições fazem parte do Rede Ciranda, entre elas: Associação de Pais e Amigos dos Excepcionais – APAE, Associação Beneficente de Assis – SIM, Associação Filantrópica Nosso Lar, Unidade de Prestação de Serviço Especial de Reabilitação – SER, Casa da Menina São Francisco de Assis, Casa da Criança Dom Antônio José dos Santos, Associação Amigos de Santa Cecília, Círculo dos Amigos dos Pobres do Pão de Santo Antônio – CAPSA, entre outras.

Segundo Xavier *et. al.* (2012), o objetivo maior do Projeto é garantir os direitos e o desenvolvimento integral das crianças e dos adolescentes. Com isso, o Rede Ciranda fica responsável por algumas ações, sendo elas: realizar manutenção no sistema REDECA, oferecer cursos de aprimoramento aos atores sociais da rede; dar visibilidade ao CMDCA com apoio para divulgação de suas ações, auxílio para organização de eventos, entre outros; promover condições para a realização de estudos sobre os problemas que envolvem a criança e o adolescente; realizar encontros para os gestores e funcionários

das organizações; promover campanhas sociais de conscientização sobre temas de interesse da infância e juventude; realizar um diagnóstico da realidade de atendimento à criança e ao adolescente de Assis.

3.1 – REDECA

O REDECA é um sistema livre que tem por objetivo unificar e permitir o acesso dos atores do SGDCA aos registros de atendimentos de crianças e adolescentes que foram realizados no município e, por ser um sistema livre, pode ser adaptado de acordo com a necessidade local.

O sistema é fruto de uma parceria da Fundação Telefônica com oito municípios paulistas (Araçatuba, Bebedouro, Diadema, Guarujá, Itapecerica da Serra, Mogi das Cruzes, São Carlos e Várzea Paulista) e mais de 400 organizações governamentais e não governamentais que atuam no setor.

Em funcionamento desde 2010 em, ao menos, 16 instituições, o REDECA possui mais de 24 mil pessoas cadastradas, sendo crianças, adolescentes e familiares de crianças e adolescentes atendidos no município.

A implantação do sistema na cidade de Assis ocorreu no ano de 2010 com a parceria do Projeto Rede Ciranda e a Fundação Educacional do Município de Assis – FEMA. Estagiários do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas e Ciência da Computação visitaram as instituições para realizarem a implantação do sistema e o treinamento do técnico e coordenador responsável pelo uso do sistema.

Entre os benefícios que o REDECA oferece, podemos destacar: agilidade no atendimento a crianças e adolescentes; estabelecer relação com outros bancos de dados de âmbito municipal, estadual ou da união, acompanhamento das vagas referentes às atividades e benefícios oferecidos na rede; permitir que cada instituição tenha mais controle de seus atendimentos, entre outros. (Xavier *et. al*. 2012).

A figura 9 mostra a interface inicial do REDECA logo após autenticação do usuário. Por meio desta, o usuário autenticado poderá executar algumas ações, entre elas: pesquisar entidades, atividades; manter (cadastrar, editar, pesquisar) dados referentes a crianças e

adolescentes; caso seja usuário com permissões de coordenador, editar as informações referentes à entidade que ele pertence, editar os dados dos técnicos da sua entidade, manter atividades da sua entidade, entre outras.

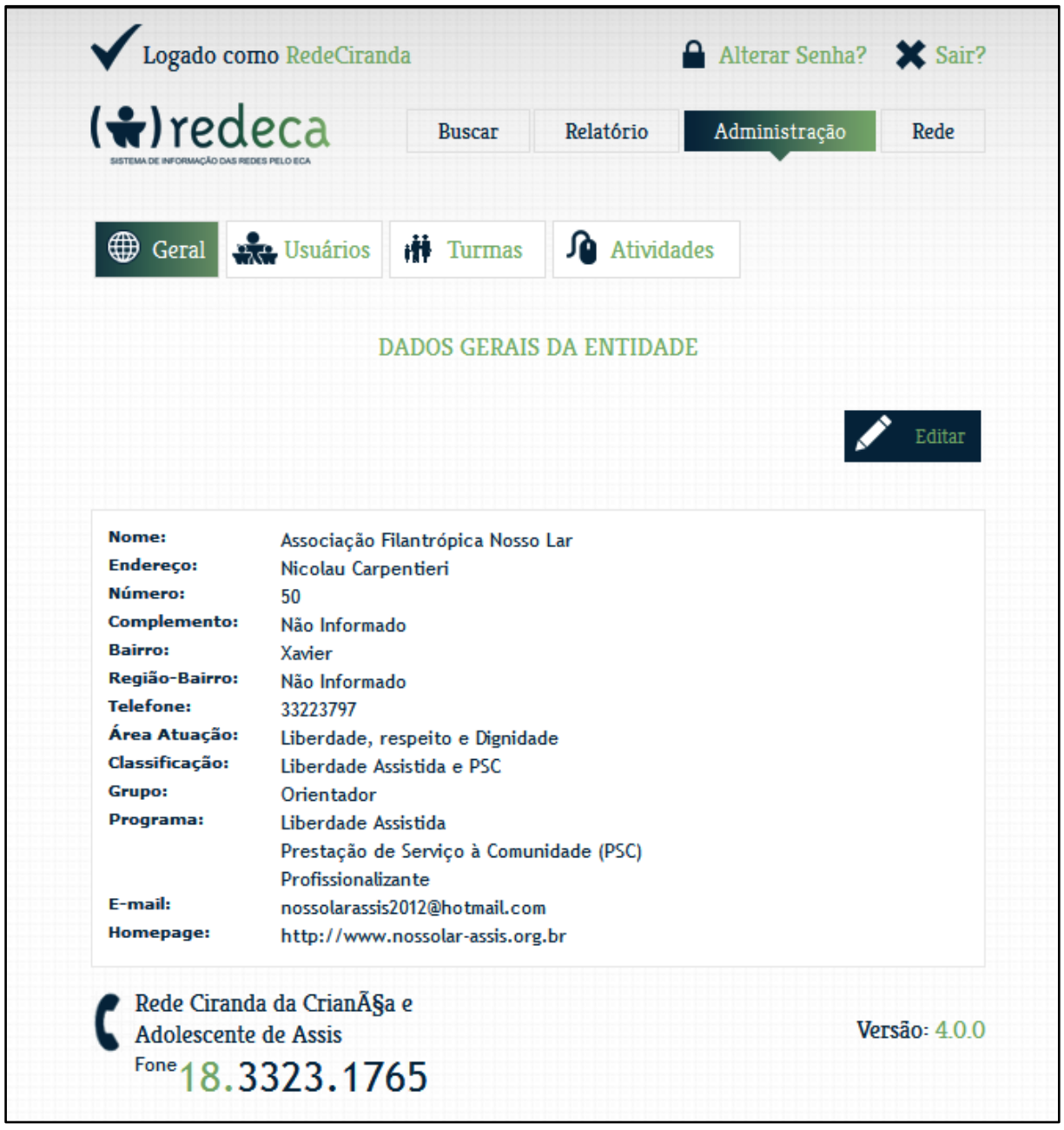


Figura 8 – Interface Gráfica Inicial do REDECA

Neste capítulo foi apresentado o Projeto Rede Ciranda, o seu papel e a sua importância no atendimento à criança e ao adolescente. Foi apresentado, também, o sistema REDECA, sua importância e alguns dos benefícios que tal sistema oferece.

4. APLICANDO O PROCESSO DE DESCOBERTA DE CONHECIMENTOS NA BASE DE DADOS DO REDECA

Neste capítulo será abordado todo o procedimento da aplicação das técnicas de descoberta de conhecimentos na base de dados do sistema REDECA.

4.1 – TECNOLOGIAREDECA

O REDECA é um sistema *Web*, desenvolvido na linguagem PHP e utiliza o banco de dados MySQL. A base de dados do REDECA é muito complexa, atualmente está em uso a versão 4.0e conta com 108 tabelas para guardar informações sobre diversas áreas referente à criança e adolescente. Essas áreas são: identificação, escolaridade, saúde, renda, família, moradia, atendimentos e entidades. A figura abaixo ilustra essas áreas por meio das guias: Geral, Educação, Saúde, Benefício, Renda, Moradia e Atendimento, em destaque na figura 10.

Figura 9–Guias que identificam as áreas de abrangência do REDECA

No Diagrama Entidade-Relacionamento (DER) podem-se identificar essas áreas por cores. A seguir será identificada cada área no diagrama e destacado os principais campos de cada uma delas.

- Amarelo – pessoa: nome, apelido, sexo, tatuagem, país de origem, data de nascimento, nacionalidade, raça, estado civil, documentação, deficiência;

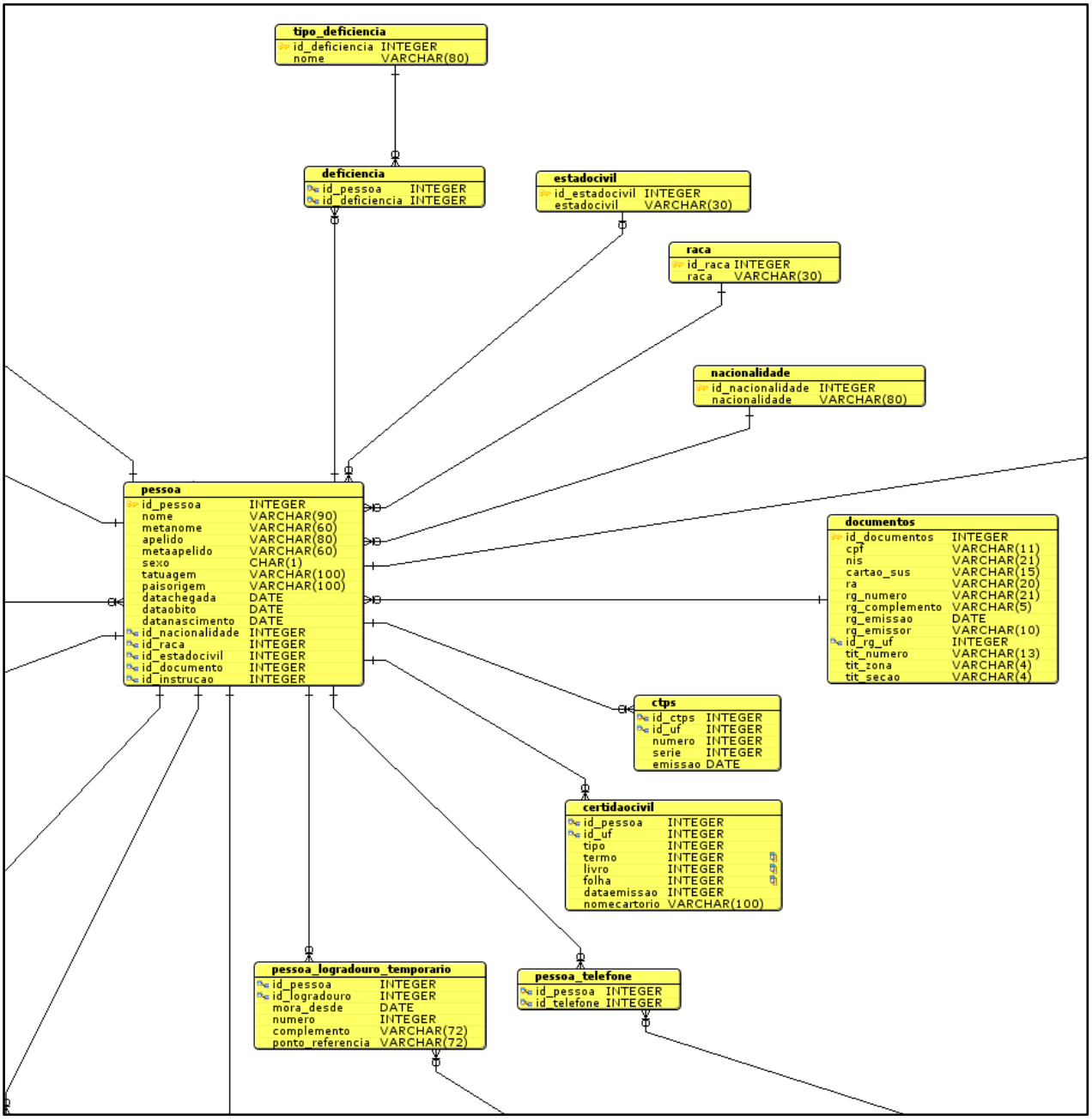


Figura 10 - DER: Pessoa

- Verde – educação: escola, tipo de escola, série, período, grau de instrução, frequência;

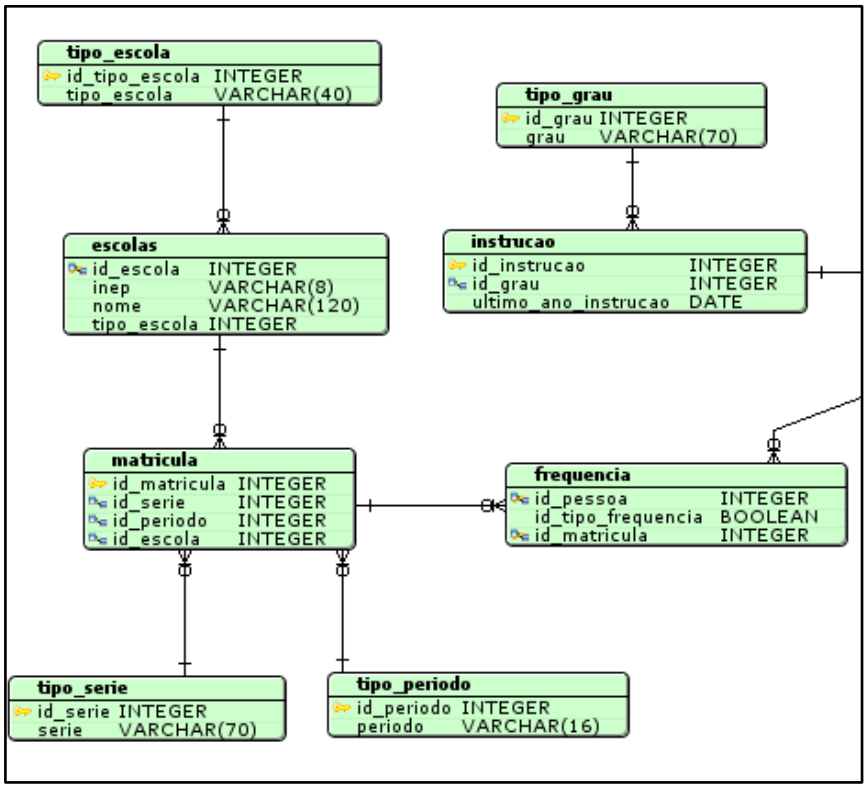


Figura 11 - DER: Educação

- Laranja – família: representante familiar, parentesco;

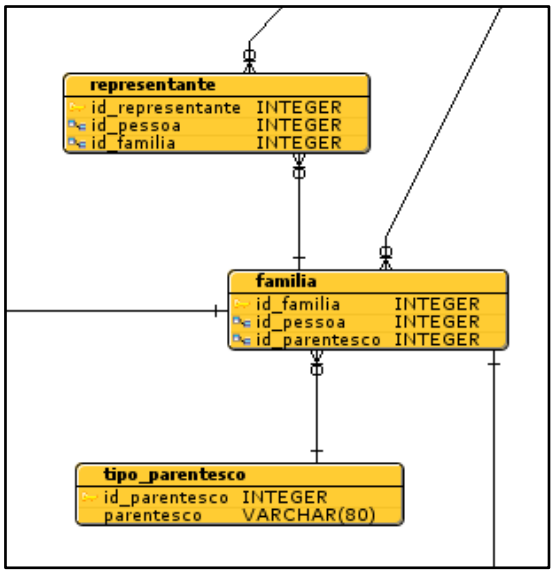


Figura 12 - DER: Família

- Branco – saúde: quadro de saúde, gestantes, vacinação, convênio médico, uso de drogas;

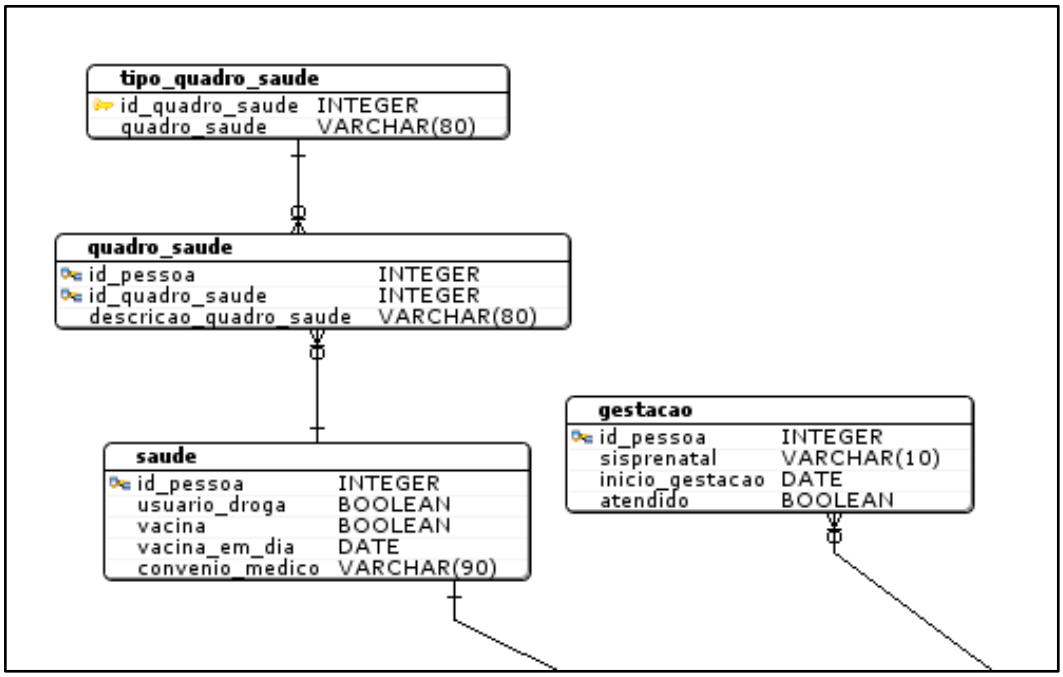


Figura 13 - DER: Saúde

- Verde cítrico – programas sociais: programa social, benefício;

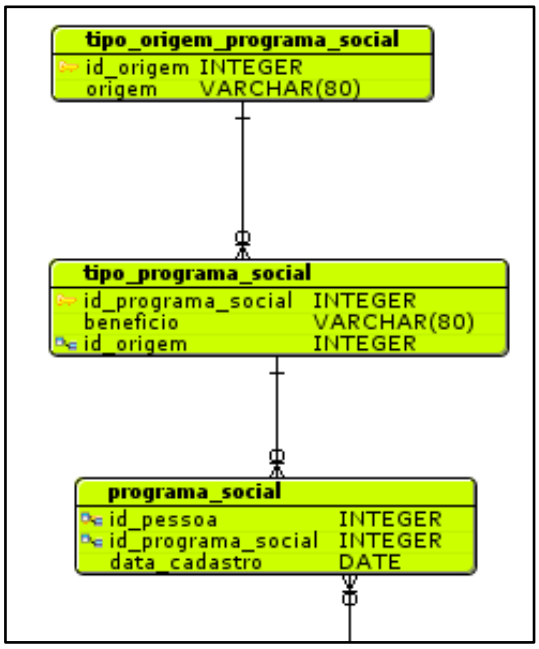


Figura 14 - DER: Programas Sociais

- Verde escuro – moradia: tipo de moradia, ponto referencia, situação da moradia, localidade, tipo de construção, tipo de abastecimento de água, tipo de sanitário, tipo de iluminação, tipo de coleta de lixo;

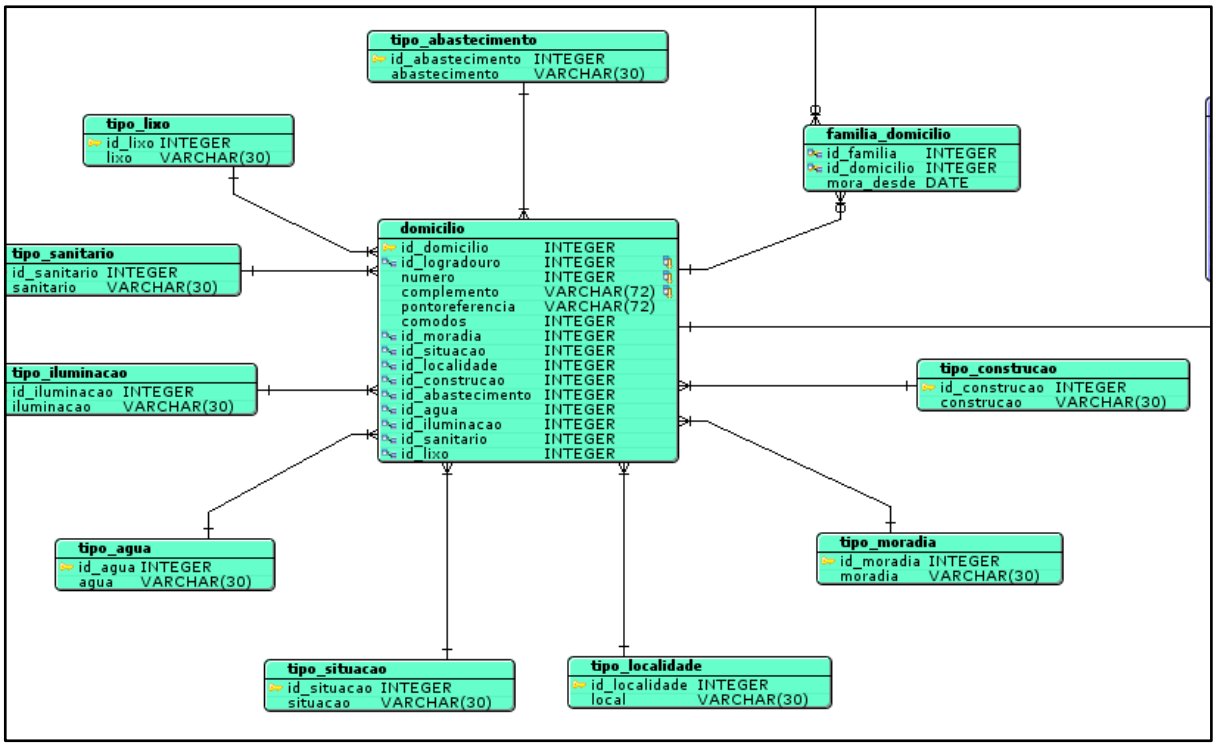


Figura 15 - DER: Moradia

- Lilás – atendimento: pessoa, entidade, usuário, programa, início, término previsto, encerramento real;

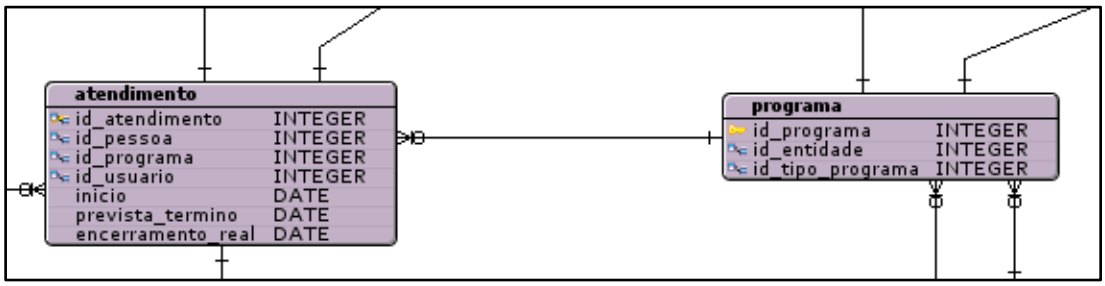


Figura 16 - DER: Atendimento

- Salmão – despesas familiares: tipo de despesa, valor;

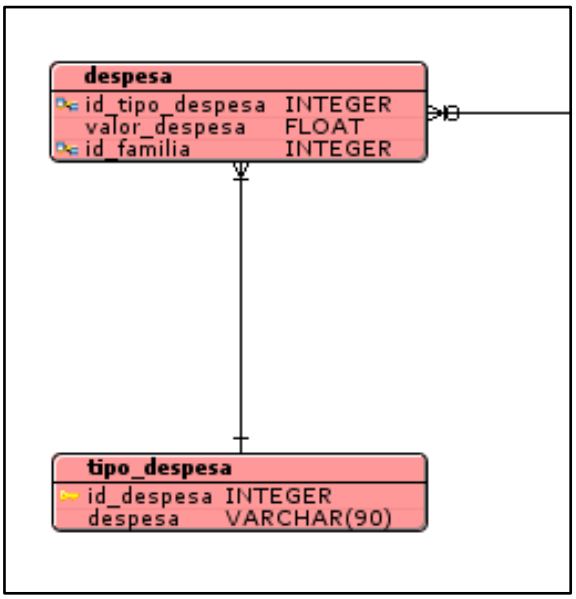


Figura 17 - DER: Despesas Familiares

- Roxo – renda e emprego: tipo de renda, situação de emprego, empresa, endereço, telefone, valor da renda;

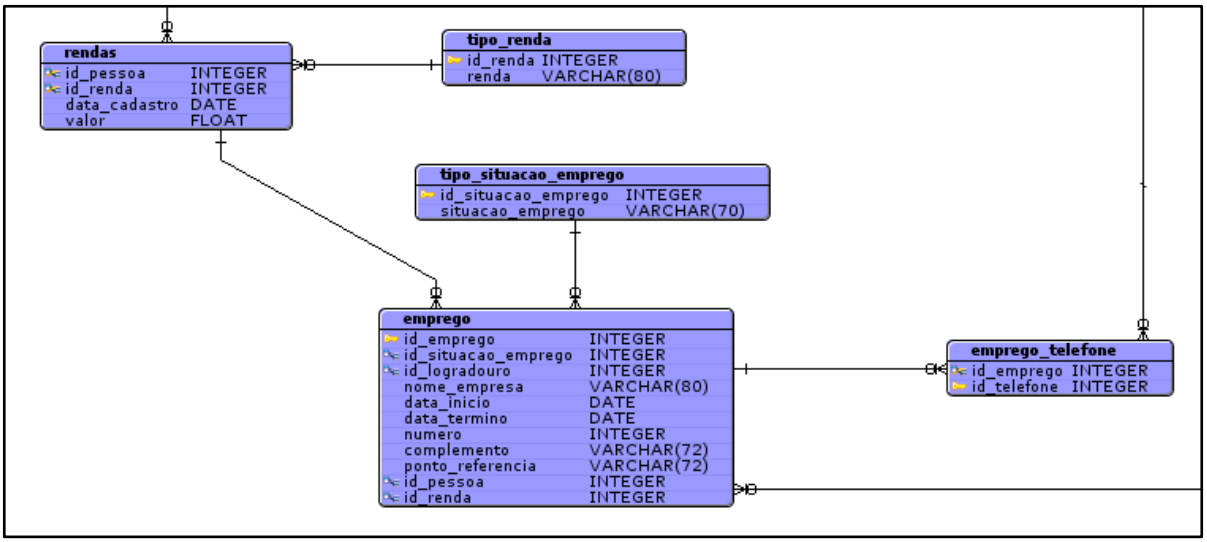


Figura 18 - DER: Renda e Emprego

- Verde musgo: entidades; nome, *e-mail*, *homepage*, telefone, área de atuação;

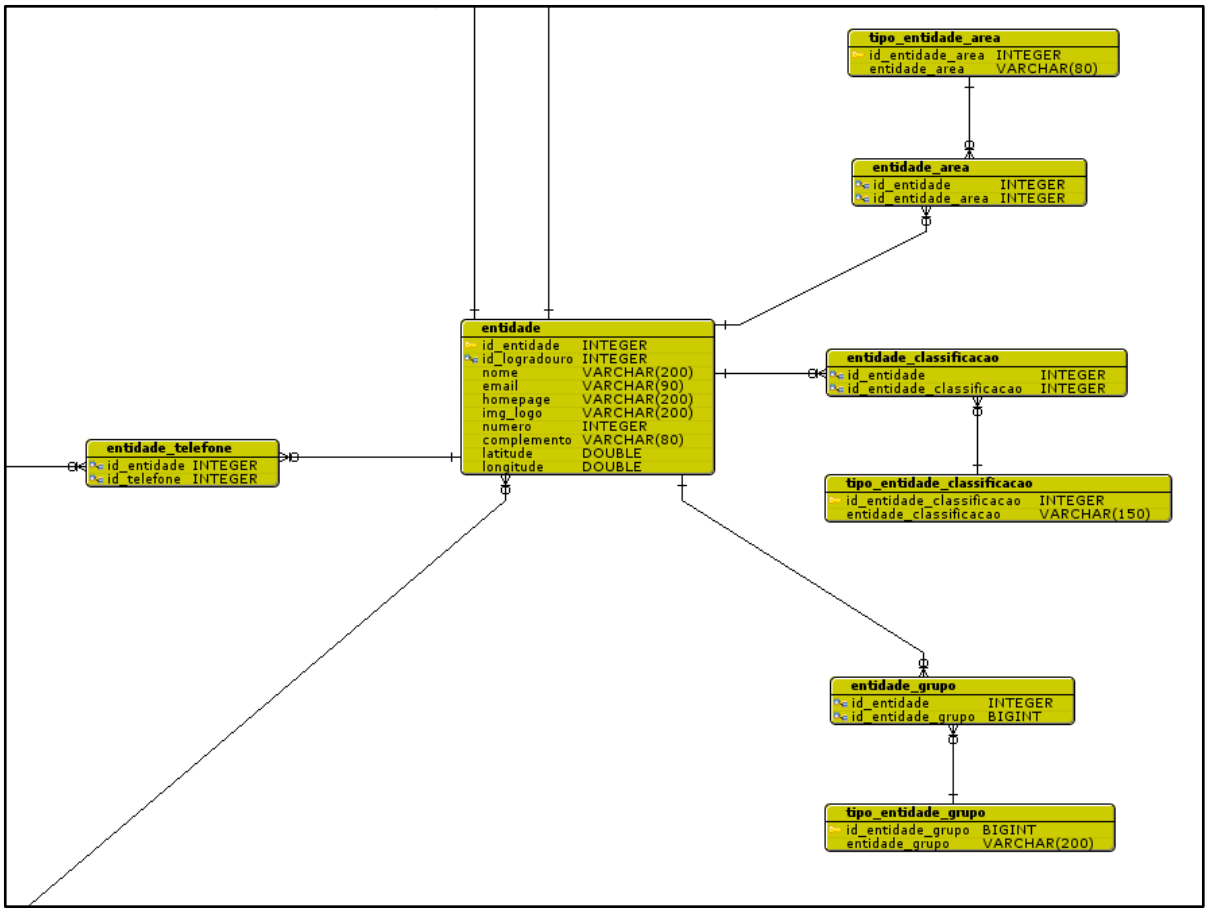


Figura 19 - DER: Entidade

- Cinza claro – telefone e logradouro: logradouro, tipo logradouro, bairro, região, telefone;

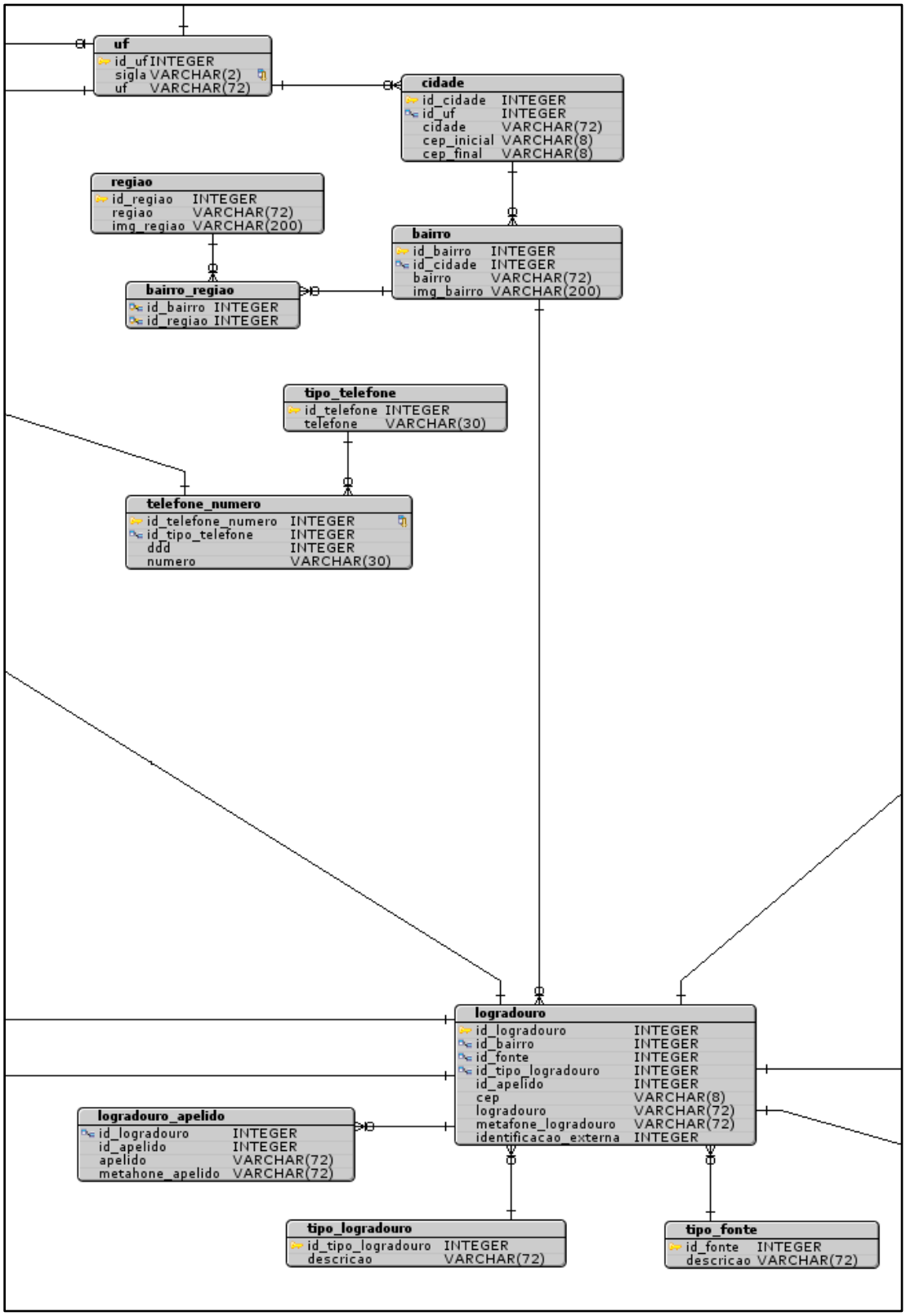


Figura 20 - DER: Logradouro e Telefone

- Rosa – atendimento especial: ofício, processo, decisão judicial, ano processo, número processo, data;

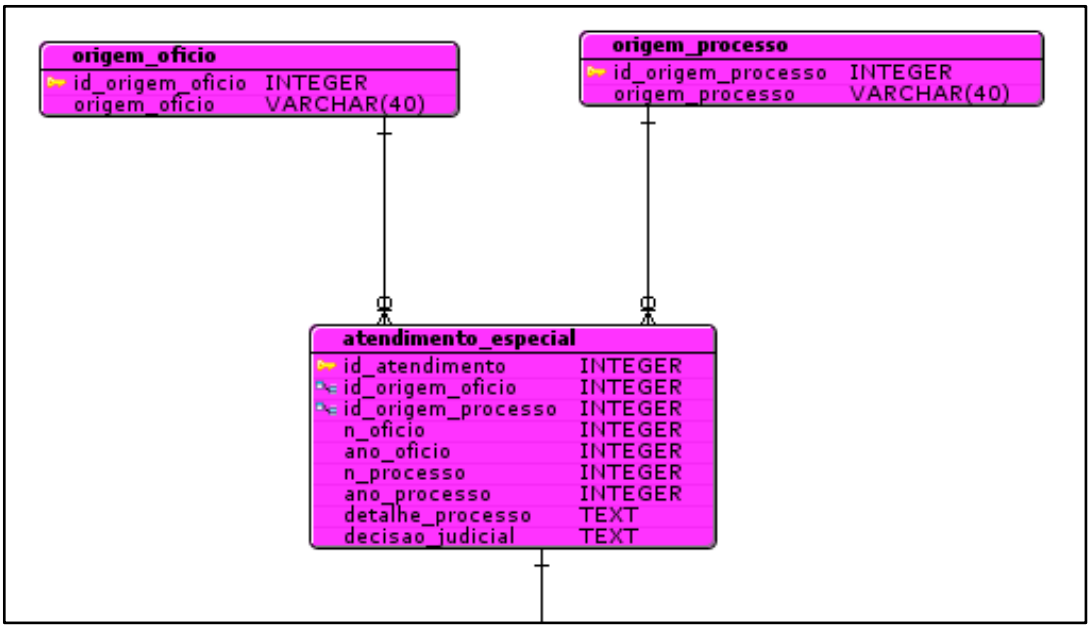


Figura 21 - DER: Atendimento Especial

- Vermelho – usuário: nome, CPF, senha, *login*, entidade;

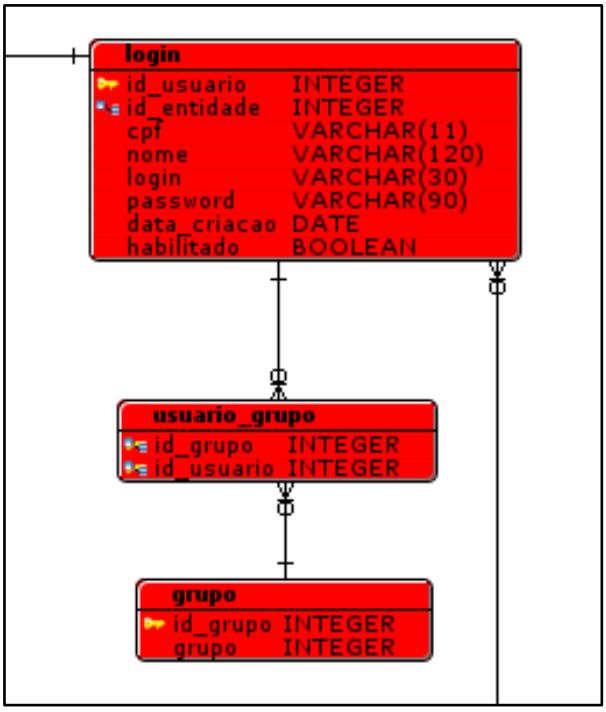


Figura 22 – DER: Usuário

- Verde limão – atividades: turmas, vagas, horário, atendimento, data início, data término, detalhe atividades, carga horária, categoria;

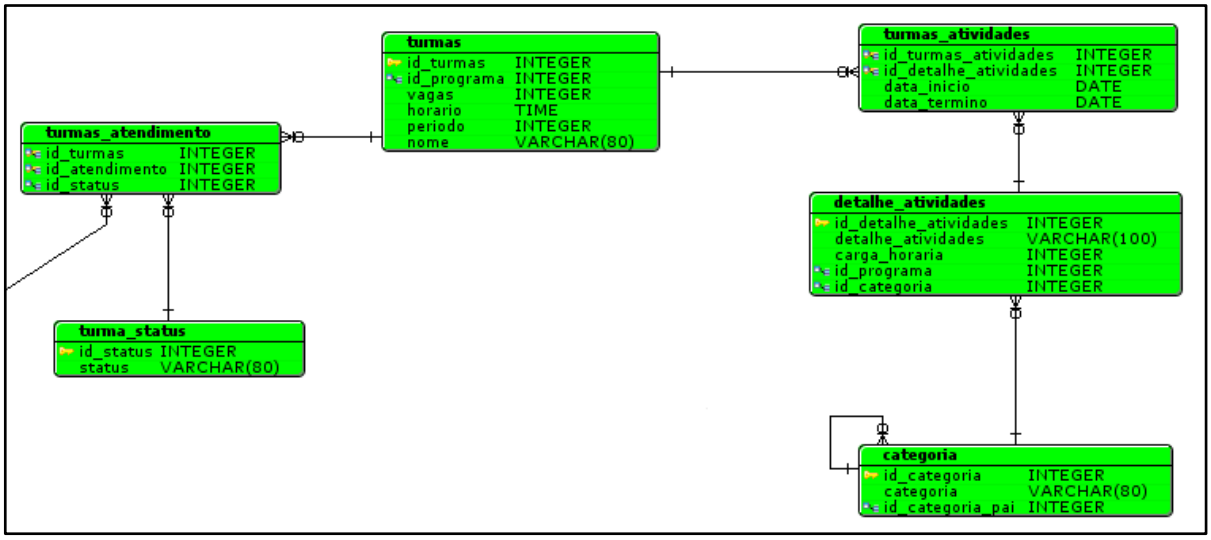


Figura 23 - DER: Atividades

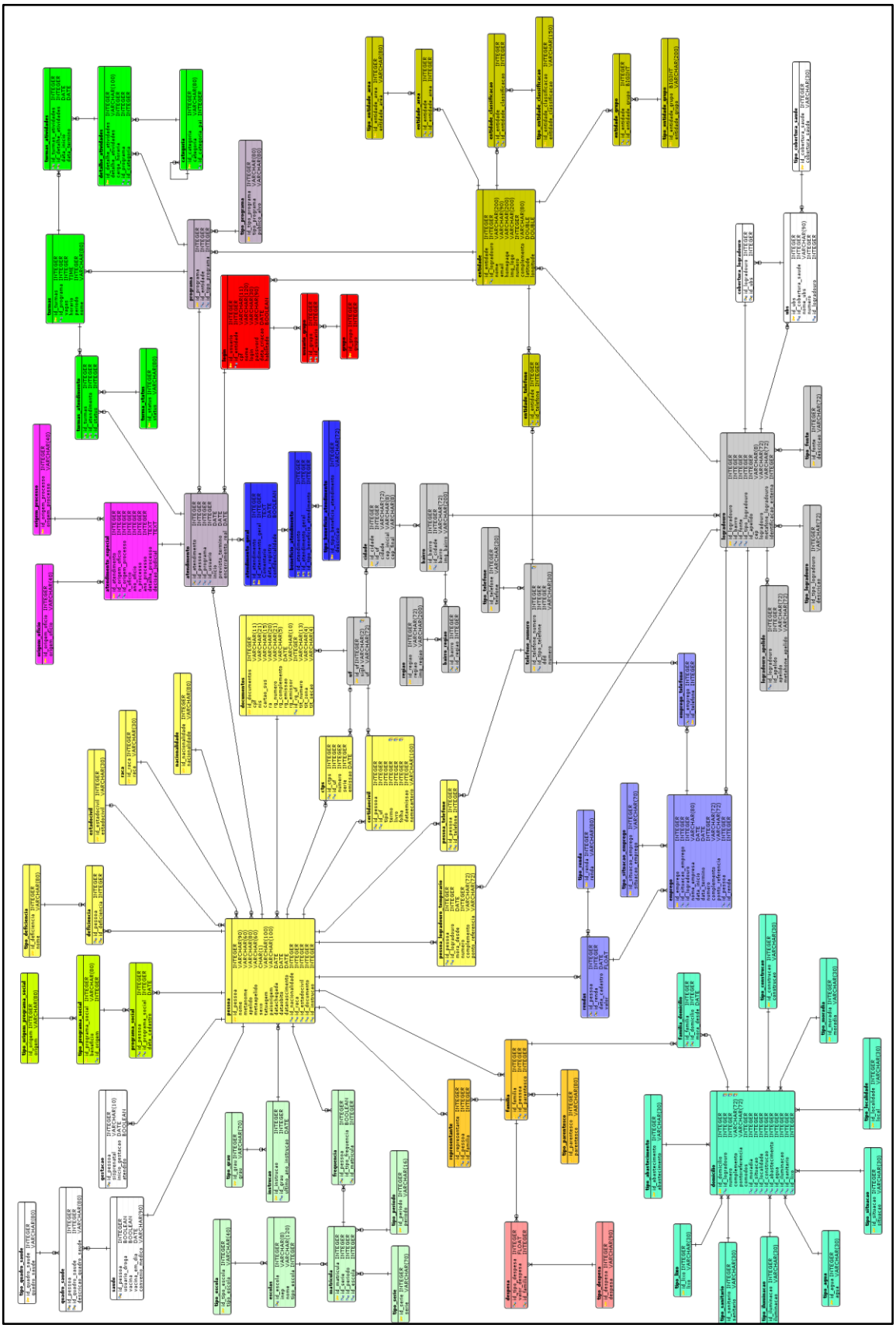
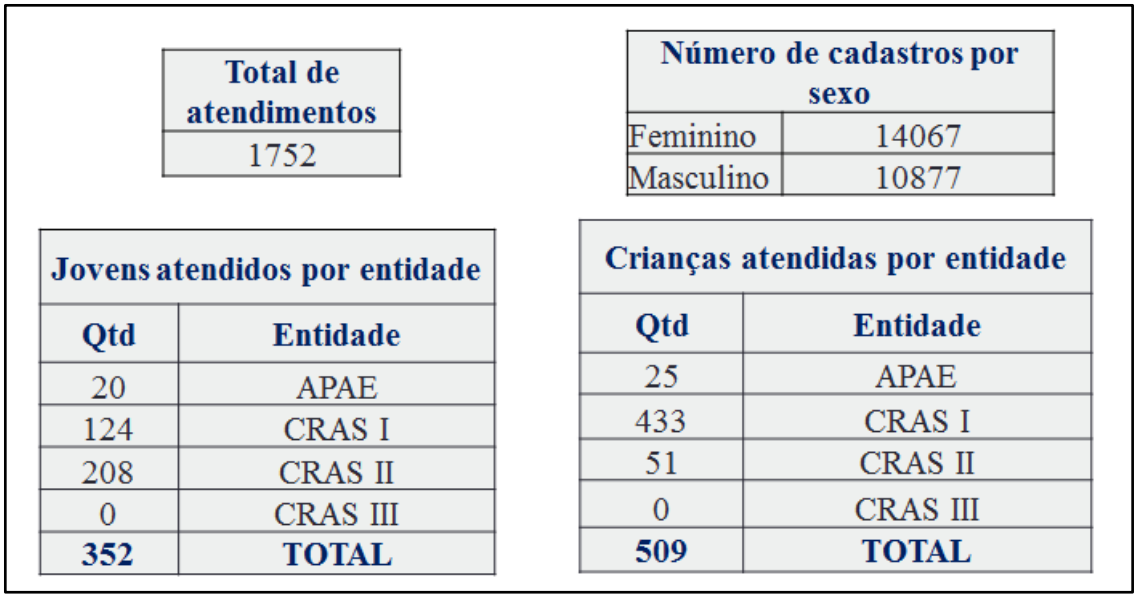


Figura 24 - Diagrama entidade-relacionamento do REDECA

Além dos 1.019 cadastros realizados pelos técnicos no sistema, foram importados 19.650 registros do Cadastro Único[3] e 890 registros do projeto Legião Mirim[4], sendo assim obtém-se um total de 27.559 registros. Porém, com a importação dos registros a base de dados do REDECA passou a contar com muitos registros duplicados. Com isso, foi necessário realizar uma manutenção nessa base de dados que passou a contar com 24.994 registros, um número bem pequeno comparado com a quantidade de habitantes da cidade.

Em maio deste ano foram realizados alguns relatórios com dados do REDECA, as figuras a seguir mostram alguns números obtidos.



Total de atendimentos
1752

Número de cadastros por sexo	
Feminino	14067
Masculino	10877

Jovens atendidos por entidade	
Qtd	**Entidade**
20	APAE
124	CRAS I
208	CRAS II
0	CRAS III
352	**TOTAL**

Crianças atendidas por entidade	
Qtd	**Entidade**
25	APAE
433	CRAS I
51	CRAS II
0	CRAS III
509	**TOTAL**

Figura 25: Relatórios de Número de atendimentos e cadastros por sexo

A Figura 25 nos traz o número de cadastros por sexo, o número total de atendimentos e os atendimentos de jovens e crianças das Instituições APAE[5], CRAS I, CRAS II e CRAS III[6]. A divisão de atendimentos funciona da seguinte forma: cada instituição atua em uma (ou mais) área.Cada área de atendimento possui um ou vários programas.Cada programa possui uma ou várias atividades, e, por fim, cada atividade possui atendimentos.

3. http://www.cadastrounico.caixa.gov.br
4. http://www.legiaomirimassis.org.br
5.http://www.apaeassis.org.br
6.http://www.assistenciasocial.net/

Exemplo: A instituição do Nosso Lar[7] atua na área de Atendimento a jovens e adolescentes, dentro dessa área atua no Programa Jovem em Ação, que conta com atividades como natação, cursos profissionalizantes, entre outras atividades. Cada dia que um adolescente participa de uma aula de natação ou de uma aula de algum curso profissionalizante oferecidos por essa instituição é considerado um atendimento.

Na figura 26 pode-se observar o número de atendimentos por atividades.

Atendimentos por atividades	
Qtd	**Atividade**
31	Academia com Saúde
22	Acompanhamento Familiar
31	Atendente de Farmácia
81	Cabeleireiro
81	Culinária
188	Defesa Direitos Criança/Adolescente
161	Departamento Pessoal
100	Educação Ambiental/Viveirismo
81	Educação Especial
31	Hardware
161	Informática
161	Inglês
31	Lan House Social
130	Marketing Pessoal
130	Orientação para o Trabalho
62	Projeto 100% Criança
...	
1482	**TOTAL**

Figura 26: Número de atendimentos por atividades

Com esses números em mãos já foi possível prever que a base de dados a ser trabalhada estaria muito instável e que problemas como dados em branco e com ruídos seriam alguns dos desafios a serem enfrentados.

7. http://www.nossolar-assis.org.br/

4.2 – APLICANDO A ETAPA DE PRÉ-PROCESSAMENTO

A primeira ação executada nesta etapa foi a preparação dos dados como será descrito a seguir:

Foi realizado o preenchimento dos campos vazio com informações óbvias, por exemplo, registros de crianças com menos de 12 anos e com o campo correspondente a estado civil em branco, o campo foi preenchido com "solteiro"; ou então crianças que frequentam o berçário o campo referente a grau de escolaridade recebeu "naoAlfabetizado" e o preenchimento dos demais campos vazios com o valor "naoConsta";

Foi realizado também um tratamento no campo referente ao bairro. No sistema REDECA além da busca de endereço pelo CEP há a possibilidade do usuário digitar o bairro. Com isso, o sistema grava na base de dados o bairro conforme foi digitado, inclusive com erros de digitação ou grafia. A figura 27 exibe alguns exemplos.

Atendimentos por bairro

Qtd	Bairro	Qtd	Bairro
5	Assis 3	15	Eldorado
34	Assis III	70	Jardim Eldorado
4	Colinas	9	JD ELDORADO
21	Park Colinas	10	conjunto habitacional nova assis
7	Park da Colinas	33	Nova Assis
12	PARK DAS COLINAS	35	Parque Universitario
6	Parque Colinas	4	PQ UNIVERSITARIO
16	Parque das Colinas	7	Prudenciana
1	Parque da Colinas	41	Vila Prudenciana
14	Residencial Colinas	2	Conjunto Habitação Assis IV
1	San Fernando Valey	18	Conjunto Habitacional Assis IV
15	San Fernando Valley	1	Vila Santa Rita
34	Jardim 3 Americas II	11	Vila Santa Rita
1	Jardim III Amercias 2	2	Parque Acacias
5	Maria Izabel	1	parque da Acacias
58	Vila Maria Izabel	24	Parque das Acacias
1	vlia Maria Izabel	**113 Bairros → 82 Bairro**	

Figura 27: Quantidade de atendimentos por bairro

Pode-se observar pequenas variações como "Assis 3" e "Assis III", ou "San Fernando Valey" e "San Fernando Valley" e também variações maiores como por exemplo para o

bairro “Park Colinas” que foi encontrado até 8 variações que correspondiam ao mesmo bairro.Com isso, foi necessário substituir os bairros que estavam com a grafia errada.

Foram eliminadas as chaves–estrangeiras de todas as tabelas que seriam utilizadas, por exemplo, na tabela de endereço o campo correspondente ao bairro armazenava o código do bairro. Foi necessário fazer uma substituição do código do bairro pelo nome do bairro.

Esse tratamento foi realizado em todos os campos que foram encontrados uma situação semelhante à situação citada.

Também foi necessário eliminar/substituir todos os caracteres especiais, acentos e espaços dos dados utilizados, como por exemplo, o campo que armazena os nomes das escolas com o valor “Dona Carolina Francine Burali” recebeu “carolinaBurali”, ou então, o campo série escolar com o valor “1º ano do ensino fundamental” recebeu “1EF” e assim sucessivamente em todos os outros campos que repetiam tal situação.

Para facilitar o processo de mineração e interpretação, campos com informações referentes a valor da renda e ano de nascimento receberam valores padronizados, por exemplo, todas as pessoas nascidas entre o ano de 1991 a 2000, receberam o valor 2000 e assim sucessivamente. Além disso, para os campos que armazenam datas (data de nascimento, data de início da gestação) foi considerado somente o ano, para facilitar o processo de mineração.

A imagem a seguir mostra como foram atribuídos esses valores padrões.

Intervalo – Renda Familiar (R$)	Valor Atribuído
até 100,00	100,00
101,00 – 300,00	300,00
301,00 – 500,00	500,00
501,00 – 1.000,00	1.000,00
1.001,00 – 1.500,00	1.500,00
1.501,00 – 2.000,00	2.000,00
acima de 2.001,00	2.500,00

Intervalo - Ano de nascimento	Valor Atribuído
até 1920	1920
1921 – 1930	1930
1931 – 1940	1940
1941 – 1950	1950
1951 – 1960	1960
1961 – 1970	1970
1971 – 1980	1980
1981 – 1990	1990
1991 – 2000	2000
2001 – 2010	2010
2011 – 2013	2013

Figura 28: Padronização de Valores

Foram descartados os campos que armazenavam nome, número de documento, número do telefone, nome da rua, entre outros que foram considerados irrelevantes para a mineração de dados.

Uma observação muito importante a se fazer é que para se realizar tais procedimentos foi mantida a base de dados original e trabalhado somente com uma cópia.

Vale ressaltar também que com exceção, de raras situações, que foi necessário realizar o tratamento dos dados "manualmente", na maioria dos casos foi utilizada uma SQL para tal tarefa, como mostra a figura 29.

```
218
219  UPDATE PESSOA
220  SET tipoCasa = "naoConsta"
221  WHERE tipoCasa IS NULL;
222
223  UPDATE PESSOA
224  SET situacaoCasa = "naoConsta"
225  WHERE situacaoCasa IS NULL;
226
```

Figura 29: Exemplo de SQL utilizada

A figura 29 exibe duas sqls utilizadas, a primeira, completa com o valor "naoConsta" se o campo "tipoCasa" estiver vazio, e a segunda, completa o campo "situacaoCasa" com o valor "naoConsta" caso também esteja vazio.

Todas SQLS utilizadas foram guardadas em um documento de texto comum para facilitar o processo de tratamento e modelagem em novas aplicações.

Após realizar todas as substituições necessárias, finalizando assim, o tratamento dos dados, passou-se então para a fase de modelagem.

Como foi dito anteriormente, a ferramenta *Weka* trabalha, preferencialmente, com arquivos do tipo ARFF. Para a elaboração de tal arquivo foi criada uma tabela com o nome "PESSOA" que seria a base de tal arquivo.

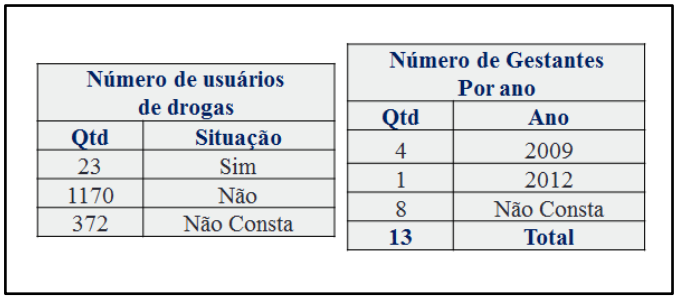
Na tabela criada, foram adicionados todos os campos que seriam utilizados no processo de mineração. A figura 30 nos ilustra a SQL utilizada para criar tal tabela.

```
delimiter $$
CREATE TABLE `PESSOA` (
  `id_person`        int(10) unsigned NOT NULL AUTO_INCREMENT,
  `sexo`             char(1),
  `anoNascimento`    varchar(4),
  `nacionalidade`    varchar(20),
  `raca`             varchar(20),
  `estadoCivil`      varchar(15),
  `grauInstrucao`    varchar(25),
  `anosEscolar`      varchar(15),
  `periodo`          varchar(10),
  `escola`           varchar(150),
  `tipoEscola`       varchar(50),
  `usuarioDrogas`    varchar(10),
  `vacina`           varchar(10),
  `planoSaude`       varchar(10),
  `gestante`         varchar(10),
  `prenatal`         varchar(10),
  `anoGravidez`      varchar(4),
  `bairro`           varchar(50),
  `tipoCasa`         varchar(20),
  `situacaoCasa`     varchar(20),
  `localidade`       varchar(15),
  `tipoConstrucao`   varchar(25),
  `abastAgua`        varchar(25),
  `tratAgua`         varchar(25),
  `iluminacao`       varchar(25),
  `tratEsgoto`       varchar(25),
  `coletaLixo`       varchar(25),
  `ocupacao`         varchar(25),
  `situacaoEmprego`  varchar(25),
  `empresa`          varchar(100),
  `valorRenda`       varchar(10),
  `tipoRenda`        varchar(30),
  `despesa`          varchar(200),
  `valorDespesa`     varchar(10),
  `deficiencia`      varchar(45),
  `repFamilia`       VARCHAR(20),
  `papelFamilia`     VARCHAR(20),
  `entidade`         VARCHAR(150),
  `programa`         VARCHAR(150);
PRIMARY KEY (`id_person`)) ENGINE=InnoDB AUTO_INCREMENT=37916 DEFAULT CHARSET=latin1
COMMENT='Registro dos dados cadastrais das pessoas'$$
```

Figura 30: SQL utilizada para criar a tabela PESSOA

Após a criação da tabela foram adicionados todos os valores em cada campo e os campos vazios receberam o valor “naoConsta”.

Com a finalização desta tarefa já teríamos a base de dados pronta para trabalhar. Seria necessário apenas gerar o arquivo ARFF com tal base de dados. Porém, durante a etapa de tratamento e modelagem foram encontrados muitos campos não preenchidos. As figuras a seguir ilustram algumas das situações encontradas.



Número de usuários de drogas	
Qtd	**Situação**
23	Sim
1170	Não
372	Não Consta

Número de Gestantes Por ano	
Qtd	**Ano**
4	2009
1	2012
8	Não Consta
13	**Total**

Figura 31: Número de usuários de drogas e gestantes

Sabe-se que não são necessários relatórios muito elaborados para observarmos que as informações referentes ao número de usuário de drogas e gestantes estão muito fora da realidade. Na figura 32 é possível verificar uma pequena parte da base de dados com muitos dados não preenchidos que receberam o valor “naoConsta”.

sexo	anoNascimento	nacionalidade	raca	estadoCivil	grauInstrucao	anoEscolar	periodo	escola	tipoEscola	usuarioDrogas
m	1940	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1992	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
f	1969	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
f	2002	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1980	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	2005	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1984	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1985	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1959	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1999	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1956	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1992	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1971	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1967	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
f	1997	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1995	brasileiro	parda	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
f	1967	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1990	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	2000	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
f	1997	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
f	1979	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
f	1976	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1999	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta
m	1997	brasileiro	branca	solteiro	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta	naoConsta

Figura 32:Dados não preenchidos na base de dados do REDECA

Nota-se, na figura 32, que somente estão preenchidas, em todos os registros, os primeiros campos, que são referentes às informações básicas da pessoa. Tais informações, além de estarem na primeira parte do cadastro, são de preenchimento obrigatório e campos como nacionalidade, raça e estado civil já vem com valores pré-definidos.

Diante de tal situação, foi considerado trabalhar com dados aleatórios. A ideia é preservar essa base de dados original e gerar dados aleatórios para serem minerados. Com isso será possível mostrar aos usuários do sistema REDECA a importância de cadastrar, e manter atualizadas as informações referentes à criança e ao adolescente de Assis. Isso será possível pelo fato do usuário ver as informações que podem ser extraídas dessa base de dados, como que a base de dados precisa estar para obter tais informações e a base de dados como realmente está.

Para gerar os dados aleatórios foi utilizada uma função chamada *rand*, disponível no banco de dados MYSQL, que permite atribuir valores aleatórios. A imagem a seguir mostra a função utilizada para gerar valores aleatórios no campo "anoEscolar".

```
UPDATE PESSOA SET anoEscolar = (SELECT FLOOR(0+(RAND()*(2-0)))) WHERE anoEscolar = "naoConsta" ;
UPDATE PESSOA SET anoEscolar = "naoEstuda" WHERE anoEscolar = "0";
UPDATE PESSOA SET anoEscolar = "naoConsta" WHERE anoEscolar = "1";
UPDATE PESSOA SET anoEscolar = "1EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2007";
UPDATE PESSOA SET anoEscolar = "2EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2006";
UPDATE PESSOA SET anoEscolar = "3EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2005";
UPDATE PESSOA SET anoEscolar = "4EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2004";
UPDATE PESSOA SET anoEscolar = "5EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2003";
UPDATE PESSOA SET anoEscolar = "6EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2002";
UPDATE PESSOA SET anoEscolar = "7EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2001";
UPDATE PESSOA SET anoEscolar = "8EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "2000";
UPDATE PESSOA SET anoEscolar = "9EF" WHERE anoEscolar = "naoConsta" AND anoNascimento = "1999";
UPDATE PESSOA SET anoEscolar = "1EM" WHERE anoEscolar = "naoConsta" AND anoNascimento = "1998";
UPDATE PESSOA SET anoEscolar = "2EM" WHERE anoEscolar = "naoConsta" AND anoNascimento = "1997";
UPDATE PESSOA SET anoEscolar = "3EM" WHERE anoEscolar = "naoConsta" AND anoNascimento = "1996";
```

Figura 33: Exemplo de função utilizada para gerar valores aleatórios

Na situação apresentada na figura acima, a função atribuiu valores aleatórios maior que 0 e menor que 2, no campo "anoEscolar" quando o mesmo possuía o valor "naoConsta"; os valores 0 e 1 foram substituídos pelos valores "naoEstuda" e "nãoConsta"; o campo "anoEscolar" recebeu os valores correspondente ao ano escolar com base no ano de nascimento, para não ocorrer situações impossíveis, como por exemplo, ano de nascimento = 1996 e ano escolar = berçário.

Foi realizado esse procedimento em todos os campos da tabela "PESSOA" para gerar os dados aleatórios. Após a finalização destes procedimentos obtemos uma base de dados completa e foi possível então gerar o arquivo ARFF.

Para a elaboração de tal arquivo foi necessário realizar uma consulta à base de dados que retornava todos os registros, exportar o resultado para um arquivo no formato CSV e converter o arquivo CSV para ARFF.

Com o arquivo no formato adequado, ARFF, foi necessário adicionar o cabeçalho do arquivo. Como explicado anteriormente, tal arquivo é composto por cabeçalho e um conjunto de instâncias. No cabeçalho declara-se a relação que o arquivo representa, lista de atributos e a relação dos valores que poderá conter em cada atributo. Abaixo do cabeçalho, estão as instancias com os dados, separados por vírgulas. Cada linha representa uma instancia/registro. A figura 34ilustra parte do arquivo “PESSOAS.ARFF”.

```
@relation PESSOAS

@attribute sexo {m, f}
@attribute anoNascimento string
@attribute nacionalidade {brasileiro, estrangeiro}
@attribute raca {branca, parda, negra, amarela, indigena}
@attribute estadoCivil {solteiro, casado, viuvo, divorciado, separado}
@attribute escolaridade {4SerieCompleta, 4SerieIncompleta, 8SerieCompleta, 8SerieIncompleta, analfabet
@attribute serieEscolar {naoEstuda, bercario, MI, MII, JI, JII, 1EF, 2EF, 3EF, 4EF, 5EF, 6EF, 7EF, 8EF
@attribute periodo {naoEstuda, integral, manha, tarde, noite}
@attribute escola string
@attribute tipoEscola {publicaEstadual, publicaMunicipal, publicaFederal, naoEstuda, particular}
@attribute usuarioDrogas {sim, nao}
@attribute vacina {sim, nao}
@attribute planoSaude {sus, unimed}
@attribute getante {sim, nao}
@attribute prenatal {naoConsta, sim, nao}
@attribute anoGravidez string
@attribute bairro string
@attribute tipoCasa {casa, comodos, apartamento}
@attribute situacaoCasa {proprio, alugado, financiado, cedido, arrendado}
@attribute localidade {urbana, rural}
@attribute tipoConstrucao {materialAproveitado, adobe, taipaRevestida, taipaNaoRevestida, tijoloAlvena
@attribute abastAgua {redePublica, pocoNascente}
@attribute tratAgua {cloracao, semTratamento}
@attribute iluminacao {relogioProprio, relogioCompartilhado, vela, lampiao}
@attribute tratEsgoto {redePublica, ceuAberto, fossaSeptica, fossaRudimentar, vala}
@attribute coletaLixo {coletado, enterrado, ceuAberto, queimado}
@attribute ocupacao string
@attribute situacaoEmprego {naoTrabalha, aposentadoPensionista, assalariadoComCarteira, assalariadoSem
@attribute valorRenda {0, 100, 1000, 1500, 2000, 2500, 300, 500}
@attribute deficiencia {nao, intelectual, fisica, surdez, mental}
@attribute repFamilia {sim, nao}
@attribute papelFamilia{avo, sobrinho, mae, adotivo, pai, filho, neto, primo, genro, irmao, enteado, e
@attribute entidade {amigosEscola, apae, bemMeQuer, bracosAbertos, camisa10, capsa, casaAbrigo, casaCr

@data
m,1940,brasileiro,branca,solteiro,4SerieIncompleta,naoEstuda,naoEstuda,naoEstuda,naoEstuda,sim,sim,sus
m,1992,brasileiro,parda,solteiro,8SerieIncompleta,naoEstuda,naoEstuda,naoEstuda,naoEstuda,sim,nao,sus,
f,1969,brasileiro,negra,solteiro,4SerieIncompleta,naoEstuda,naoEstuda,naoEstuda,naoEstuda,sim,nao,sus,
```

Figura 34: Parte do arquivo "PESSOA.ARFF"

4.3 – APLICANDO A ETAPA DE MINERAÇÃO DE DADOS

Após passarem pela fase de pré-processamento, os dados estão prontos para serem minerados. Com o arquivo ARFF pronto foi necessário abrir o mesmo na ferramenta *Weka*. Quando o arquivo apresenta alguma falha ou não corresponde aos padrões exigidos, é exibida uma mensagem de erro indicando qual erro e em qual linha do arquivo ele se encontra. A figura 35 apresenta a tela principal do *Weka* com o arquivo "PESSOAS.ARFF" pronto para receber a mineração.

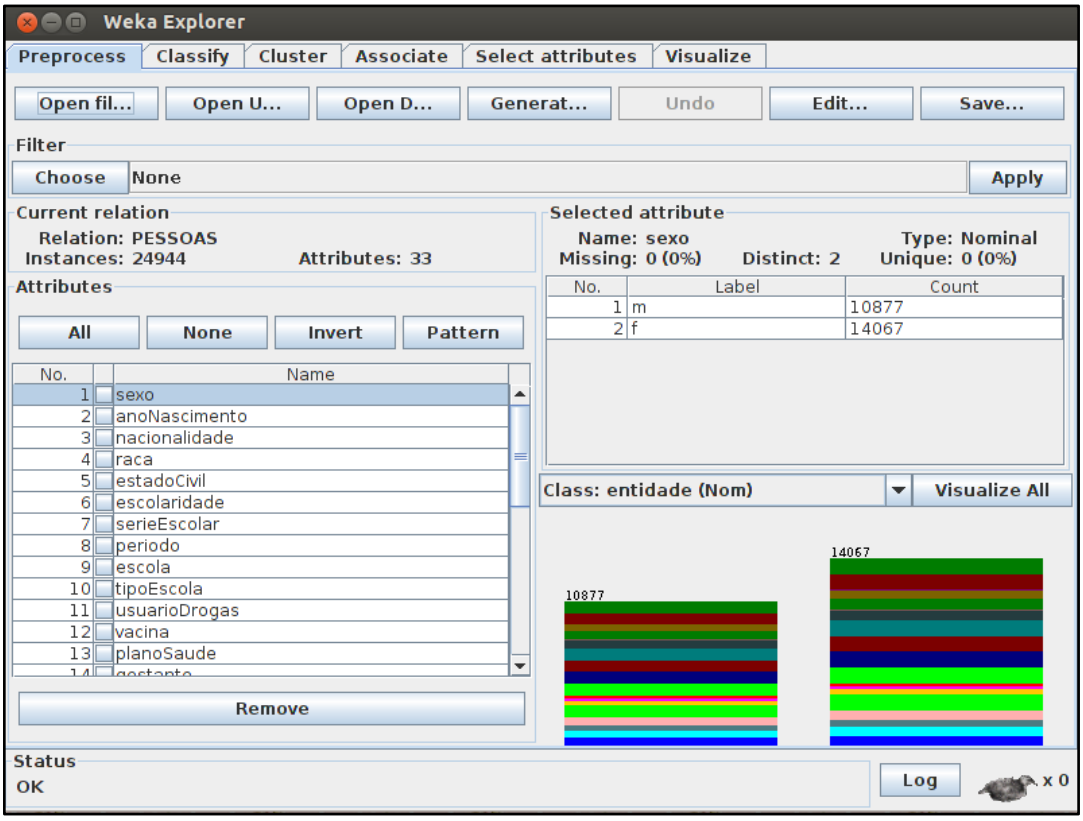


Figura 35: Ferramenta *Weka* com arquivo "PESSOAS.ARFF"

Com o arquivo previamente carregado é necessário escolher qual técnica será aplicada por meio das guias apresentadas ao usuário na parte superior da interface gráfica e posteriormente o algoritmo que será aplicado. A figura 36 ilustra a aba de classificação e o algoritmo Id3 selecionado para exemplificar como deve ser feita a escolha da técnica e do algoritmo que serão aplicados.

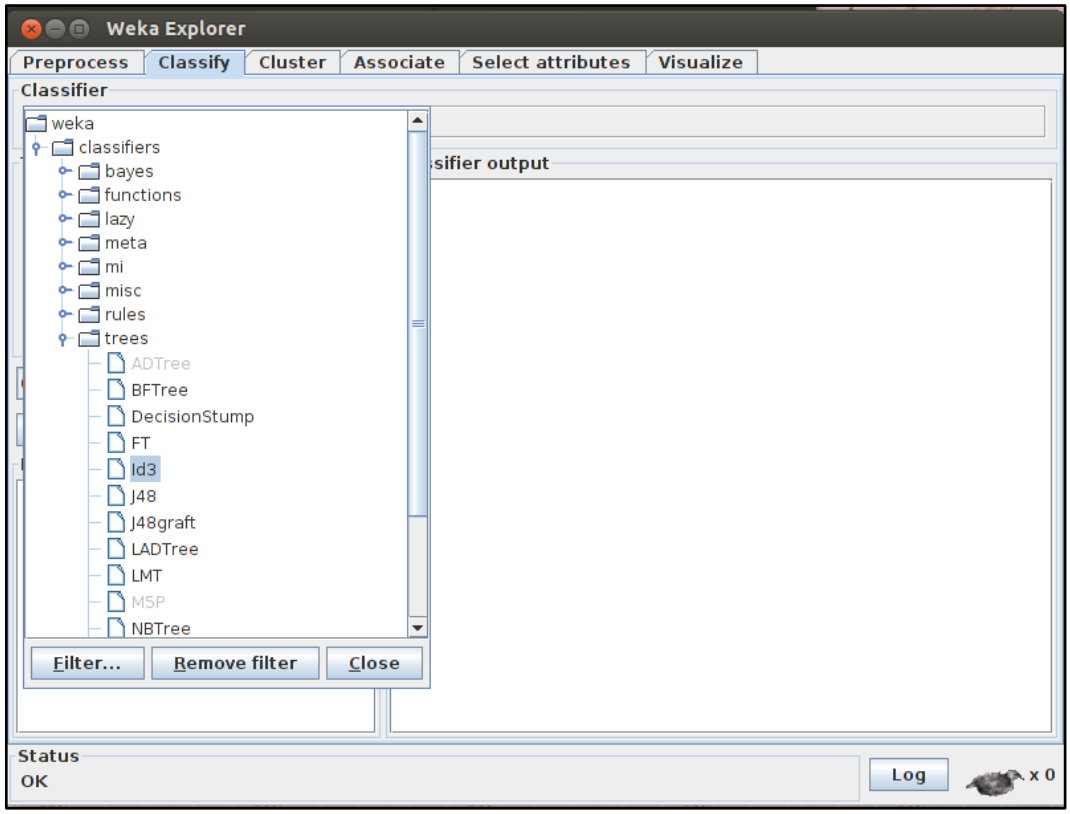


Figura 36: Escolha da técnica de mineração

Para a realização deste trabalho foi utilizada a técnica de Classificação, pois ela nos permite identificar a qual classe pertence determinados registros e dependendo do algoritmo, no caso os algoritmos de árvore de decisão, a ferramenta *Weka*, possibilita verificar a árvore de decisão criada, com base em na entrada, cada nó na árvore representa um ponto no qual se deve tomar uma decisão. As folhas dessa árvore representam a saída prevista de acordo com a entrada definida. Após escolher a técnica, foi necessário escolher o algoritmo que seria aplicado. Para a escolha do(s) algoritmo(s), foram realizados alguns testes para verificar qual/quais apontariam o maior número de instancias classificadas corretamente ou incorretamente, de acordo com a base selecionada.

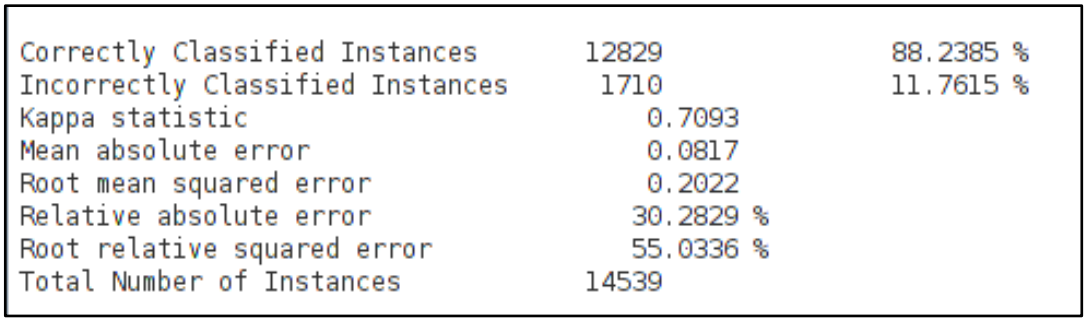
Após constatar-se muita dificuldade na interpretação dos resultados devido ao grande número de registros e atributos definidos os atributos foram separados por áreas e aplicados alguns algoritmos de mineração e observado o porcentual de classificação

correta ou incorreta dos dados e selecionados os algoritmos que apresentaram maior número de aproveitamento dos dados, nesse caso, o algoritmo J48.

4.4 – APLICANDO A ETAPA DE INTERPRETAÇÃO

Foi elaborado um arquivo ARFF somente com os atributos referentes a uso de drogas, gestação, pré-natal e escola. Nesta situação também foram descartados os registros referente ao sexo masculino e que a idade não correspondiam idade de uma adolescente.

Com a aplicação da mineração, foi possível obter os seguintes resultados: a 37 exibe, nas duas primeiras linhas, o número e porcentagem de dados classificados corretamente e o número e porcentagem dos dados desconsiderados, pois apresentaram algum tipo de erro identificado pelo algoritmo aplicado. É importante ressaltar que quanto maior for o porcentual dos dados classificados corretamente, melhor é a qualidade da base de dados minerada. Neste exemplo, pode-se perceber que temos uma base de dados relativamente boa, pois foram descartados apenas 11,76% dos dados.



Correctly Classified Instances	12829	88.2385 %
Incorrectly Classified Instances	1710	11.7615 %
Kappa statistic	0.7093	
Mean absolute error	0.0817	
Root mean squared error	0.2022	
Relative absolute error	30.2829 %	
Root relative squared error	55.0336 %	
Total Number of Instances	14539	

Figura 37: Número dos dados classificados corretamente e número dos dados desconsiderados

Como resultado da mineração, foi possível identicar quais as escolas apresentam adolescentes grávidas, usuárias ou não usuárias de drogas e que fazem ou não acompanhamento de pré-natal. A figura 38 ilustra as escolas que mais apresentaram tais problemas. Foi possível identificar que em escolas do centro da cidade (Professor Carlos Alberto de Oliveira, Rua Dr.

Luiz Pizza, 220, Centro e Dr. Clybas Pinto Ferraz, Rua Santa Cecília, 709, Vila Boa Vista) apresentaram um número alto de adolescentes grávidas que não fazem acompanhamento médico de pré-natal e muitas delas são usuárias de drogas. O primeiro número dentro do parêntese refere-se ao valor apontado pelo atributo, por exemplo: gestante = sim : não (6/2), o número 6 faz referência a adolescentes gestantes que não fazem acompanhamento de pré-natal e o número 2 refere-se ao número de adolescentes gestantes que fazem acompanhamento de pré-natal.

```
escola = CarlosAlberto
|   usuarioDrogas = sim
|   |   gestante = sim : nao (6/2)
|   |   gestante = nao : naoConsta (40/0)
|   usuarioDrogas = nao
|   |   gestante = sim : sim (13/6)
|   |   gestante = nao : naoConsta (38/0)
escola = CarolinaBurali
|   gestante = sim
|   |   usuarioDrogas = sim : nao (11/3)
|   |   usuarioDrogas = nao : sim (12/2)
|   gestante = nao : naoConsta (83/0)
escola = Cleophania
|   gestante = sim
|   |   usuarioDrogas = sim : sim (4/2)
|   |   usuarioDrogas = nao : nao (4/1)
|   gestante = nao : naoConsta (87/0)
escola = Clybas
|   usuarioDrogas = sim
|   |   gestante = sim : nao (19/8)
|   |   gestante = nao : naoConsta (40/0)
|   usuarioDrogas = nao
|   |   gestante = sim : sim (6/3)
|   |   gestante = nao : naoConsta (61/0)
```

Figura 38: Gestantes e usuárias de drogas por escola.

Já na figura 39 é possível vizualizar uma parte das informações obtidas por meio de uma árvore de decisão que nos mostra o número de adolescentes grávidas por escola que fazem, ou não, acompanhamento de pré-natal e se são, ou não, usuárias de drogas.

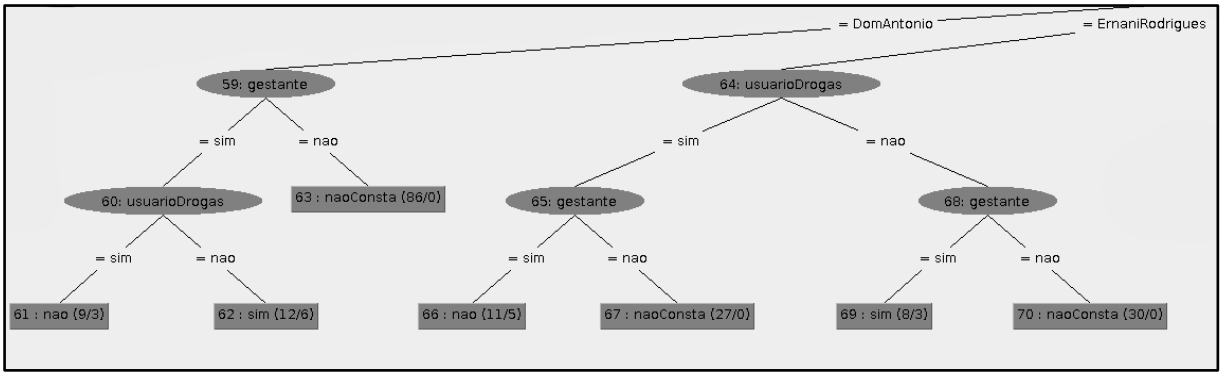


Figura 39: Gestantes que não fazem acompanhamento pré-natal por escola

Em outro exemplo de aplicação da mineração de dados nos campos referentes a uso de drogas, bairro, e condições de moradia, obtiveram-se as seguintes informações: foi possível observar uso de drogas por adolescentes que residem na Vila Prudenciana, em casas feitas com os seguintes materiais: adobe, taipa (revestida ou não revestida), madeira; com iluminação a vela ou relógios compartilhados, ou então adolescentes que moram em apartamentos, como mostra a figura a seguir:

```
|   bairro = vilaPrudenciana
|   |   abastAgua = redePublica
|   |   |   tipoConstrucao = materialAproveitado : nao (1/0)
|   |   |   tipoConstrucao = adobe : sim (2/1)
|   |   |   tipoConstrucao = taipaRevestida : sim (2/0)
|   |   |   tipoConstrucao = taipaNaoRevestida : sim (2/1)
|   |   |   tipoConstrucao = tijoloAlvenaria
|   |   |   |   tipoCasa = casa
|   |   |   |   |   iluminacao = relogioProprio : sim (1/0)
|   |   |   |   |   iluminacao = relogioCompartilhado : sim (1/0)
|   |   |   |   |   iluminacao = vela : sim (2/0)
|   |   |   |   |   iluminacao = lampiao : nao (1/0)
|   |   |   |   tipoCasa = comodos
|   |   |   |   |   iluminacao = relogioProprio : nao (1/0)
|   |   |   |   |   iluminacao = relogioCompartilhado : sim (0/0)
|   |   |   |   |   iluminacao = vela : sim (2/1)
|   |   |   |   |   iluminacao = lampiao : sim (0/0)
|   |   |   |   tipoCasa = apartamento : sim (6/3)
|   |   |   tipoConstrucao = madeira
|   |   |   |   tipoCasa = casa : nao (1/0)
|   |   |   |   tipoCasa = comodos : sim (2/1)
```

Figura 40: Usuários de drogas por bairro e condições de moradia

Utilizando os atributos referentes ao valor da renda, uso de drogas, entidade e escola, foi possível identificar que em famílias que possuem renda entre R$500,00 à R$ 1.000,00 possuem um grande número de usuários de drogas e um grande número de crianças e adolescentes que não frequentam a escola. Também foi possível identificar quais entidades atendem crianças e adolescentes com esse perfil, como mostra a figura 41:

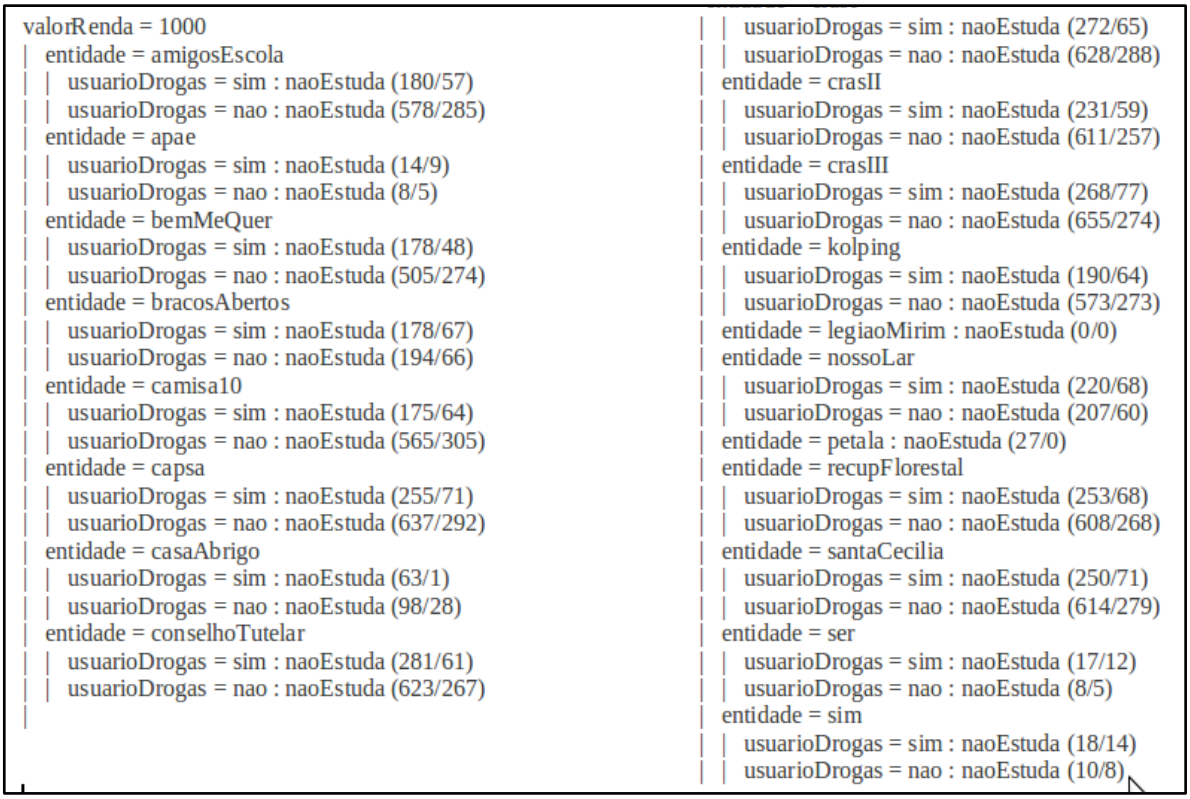


Figura 41:Usuários de drogas e frequência à escola por Entidade e Valor da Renda

Neste capítulo foi apresentado o procedimento de descoberta de conhecimentos na base de dados do sistema REDECA, desde a aplicação da etapa de preparação, as dificuldades encontradas, as soluções propostas à etapa de interpretação e os resultados obtidos.

5. CONCLUSÕES

Para a realização deste trabalho, inicialmente foi necessário realizar um estudo sobre a o Sistema REDECA, da sua estrutura de funcionamento e sua base de dados. Também foi necessário estudar como é feito o diagnóstico sobre a realidade da criança e do adolescente de Assis, quais são os dados necessários e como esse diagnóstico influencia o CMDCA e demais atores que trabalham em benefício à infância e juventude no município.

Com essas informações em mãos, foi realizado um estudo do procedimento de descoberta de conhecimentos em base de dados, das técnicas de mineração de dados e das ferramentas disponíveis no mercado para aplicar tais técnicas.

Após concluir esta etapa de estudos iniciou-se então a aplicação da etapa de pré-processamento, onde os dados foram selecionados e preparados. No andamento desta fase foram constatados muitos dados com ruídos e muitos registros incompletos, o que tornou tal fase muito trabalhosa e extensa. Devido a grande quantidade de dados incompletos foi constatado que não seria possível extrair as riquezas de tais dados. Sendo assim, foi cogitado preservar a base de dados original e trabalhar com dados aleatórios gerados por uma função SQL. Com isso foi possível obter uma base de dados completa, onde foi possível extrair informações interessantes. Ao trabalhar com uma base composta por dados aleatórios, será possível mostrar para os usuários do Sistema REDECA quais informações seriam possíveis obter caso houvesse uma base de dados íntegra e completa. Sendo assim será possível fazer um trabalho de conscientização e mostrar a importância de cadastrar, e manter atualizadas, as informações referentes às crianças e aos adolescentes do município.

Se houvesse uma base de dados íntegra, seria possível identificar, por exemplo, quais escolas e entidades apresentam maior número de adolescentes grávidas e/ou usuárias de drogas, ou então, quais bairros da cidade, condições de moradia e valores das rendas possuem mais adolescentes envolvidos com drogas e que não frequentam à escola.

5.1 – TRABALHOS FUTUROS

Para dar início ao trabalho de conscientização às entidades que atendem crianças e adolescentes na cidade de Assis, serão apresentados na reunião mensal do CMDCA, no mês de dezembro, os resultados obtidos neste trabalho, como forma de mostrar as informações relevantes para auxiliar no diagnóstico municipal. Para isso serão elaborados alguns gráficos que facilitará a visualização das informações obtidas. A equipe do Projeto Rede Ciranda estará disponível para realizar reuniões e visitas mediante determinação do CMDCA ou solicitação de cada Instituição de atendimento à criança e adolescente para apresentar os resultados do trabalho desenvolvido, bem como os resultados que serão possíveis obter por meio da colaboração e participação de todos envolvidos.

Para o diagnóstico municipal dos próximos anos serão incorporadas ao Sistema de Diagnóstico, que já está em fase de desenvolvimento pela equipe do Rede Ciranda, as informações extraídas da base de dados original do REDECA e oferecer ao usuários relatórios completos e de fácil interpretação com das informações obtidas por meio da mineração de dados. Portanto, é muito importante trabalhar com uma base de dados íntegra e completa.

REFERÊNCIAS BIBLIOGRÁFICAS

ABERNETHY, Michael. **Mineração de dados com WEKA, Parte 1: Introdução e regressão.** Disponível em <http://www.ibm.com/developerworks/br/opensource/library/os-weka1/>. Acesso em: 12 mai 2013.

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados.** 2006. 50 p. Trabalho de Conclusão de Curso. Universidade Federal de Pernambuco, PE, Recife, 2006.

BARIONI, Maria Camila Nardini; TRAINA JÚNIOR, Caetano. **Visualização de Operações de Junção em Sistemas de Bases de Dados para Mineração de Dados.** 2002. 9 p. Dissertação de Mestrado. Centro de Informática de São Carlos – USP, SP, São Carlos, 2002.

BATISTA, Gustavo Enrique de Almeida Prado Alves. Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. 2003. 232 p. Tese de doutorado. Serviço de Pós Graduação do ICMC – USP, SP, São Carlos, 2003.

CALIL, Leonardo Aparecido de Almeida, et. al.. **Mineração de dados e pós-processamento em padrões descobertos.** Ponta Grossa: UEPG, 2008.

CAMILO, Cássio Oliveira; SILVA, João Carlos da.**Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** Goiânia: Ufg, 2009.

CARVALHO, Deborah Ribeiro et. al.. **Ferramenta de Pré e Pós-processamento para Data Mining.** In: XII SEMINÁRIO DE COMPUTAÇÃO, 2003, Blumenau, SC. Anais do XII SEMINCO, 2003, p. 131-139

DATE, C. J. **Introdução a Sistemas de Banco de Dados**. 8ª Edição. Rio de Janeiro: Elsevier, 2003 – 7ª reimpressão. 870 p.

ELMASRI, Ramez; NAVATHE, Shamkant, B. **Sistemas de Banco de Dados**. 6ª Edição. São Paulo: Pearson EducationBr, 2011. 788 p.

GARCIA, Edi Wilson. **Pesquisar e avaliar técnicas de Mineração de Dados com o uso da ferramenta Oracle *Data Mining*.**2008. 66 p. Trabalho de Conclusão de Curso. Instituto Municipal de Ensino Superior de Assis, SP, Assis, 2008.

GONÇALVES, Eduardo Corrêa. **Mineração de dados no MySQL com ferramenta Weka.** Disponível em <http://www.devmedia.com.br/mineracao-de-dados-no-mysql-com-a-ferramenta-weka/26360>. Acesso em 12 mai 2013.

LEMOS, Eliane Prezepiorski. **Análise de crédito bancário com o uso de *Data Mining:* Redes Neurais e Árvores de Decisão.** 2003. 147 p. Dissertação de Mestrado. Departamento de Matemática da Universidade Federal do Paraná, PR, Curitiba, 2003.

NAVEGA, Sergio. **Princípios Essenciais do Data Mining.** São Paulo: Cenadem, 2002. Disponível em <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 08 mar. 2013.

PERBONI, Marcos. **Mineração de dados na prática com Weka API**. Disponível em <http://marcosvperboni.wordpress.com/2013/02/15/mineracao-de-dados-na-pratica-com-weka-api/>. Acesso em: 20 mai 2013.

RIBAS JUNIOR, Fábio et al.. **Conhecerpara transformar: Guia para diagnóstico e formulação da política municipal de proteção integral das crianças e adolescentes**. São Paulo: Fundação Telefônica, 2011. 332 p.

TAKAMOTO, Miriam. **Aplicação de técnicas de Mineração de Dados para planejamento agrícola no estado de São Paulo.** Indaiatuba, SP, 2011.

VASCONCELOS, L. M. R. de; CARVALHO, C. L. de. **Aplicação de Regras de Associação para Mineração de Dados na Web.** Goiânia: Ufg, 2004.

XAVIER, Ana Lúcia Pintar et al.. **Rede Ciranda: Desenhando novos caminhos para o trabalho social com crianças e adolescentes.** São Carlos: Pedro & João Editores, 2012. 204 p.

ZANUSSO, Maria Bernadete. **Trabalho Sobre Data Mining**. Disponível em <http://www.dct.ufms.br/~mzanusso/Data_Mining.htm/>. Acesso em: 20 mai. 2013.