UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

# MINERAÇÃO DE DADOS EM DATA WAREHOUSE PARA SISTEMA DE ABASTECIMENTO DE ÁGUA

**ROBERTA MACÊDO MARQUES GOUVEIA**

Dissertação de Mestrado

João Pessoa-PB
Maio-2009

**ROBERTA MACÊDO MARQUES GOUVEIA**

# MINERAÇÃO DE DADOS EM DATA WAREHOUSE PARA SISTEMA DE ABASTECIMENTO DE ÁGUA

Dissertação de mestrado apresentada ao Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, como requisito parcial para obtenção do título de Mestre em Informática (Sistemas de Computação).

Orientadora: Professora Dra. Valéria Gonçalves Soares Elias
Co-orientador: Professor Dr. Heber Pimentel Gomes

João Pessoa-PB
Maio-2009

G719m Gouveia, Roberta Macêdo Marques.
Mineração de dados em data warehouse para sistema de abastecimento de água / Roberta Macedo Marques Gouveia. João Pessoa, 2009.
147f. : il.
Orientadora: Valéria Gonçalves Soares Elias.
Co-orientador: Heber Pimentel Gomes.
Dissertação (Mestrado) – UFPB/CCEN
1. Data warehouse – Banco de dados. 2. Mineração de dados. 3. Tecnologias OLAP.

UFPB/BC CDU: 004.65 (043)

Ata da Sessão Pública de Defesa de Dissertação de Mestrado da Roberta Macedo Marques Gouveia, candidata ao Título de Mestre em Informática na Área de Sistemas de Computação, realizada em 29 de maio de 2009.

Aos vinte e nove dias do mês de maio do ano dois mil e nove, às oito horas, na Sala de Reuniões do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar a candidata ao grau de Mestre em Informática, na área de "*Sistemas de Computação*" e na linha de pesquisa "*Computação Distribuída*", a Sra. Roberta Macedo Marques Gouveia. A comissão examinadora foi composta pelos professores doutores: Valéria Gonçalves Soares (DI-UFPB), Orientadora e Presidente da Banca Examinadora, Lucídio dos Anjos Formiga Cabral (DI-UFPB), Ed Porto Bezerra (DI-UFPB) e Heber Pimentel Gomes (UFPB), como examinadores internos e Sônia Virginia Alves França (UFRPE), com examinadora externa. Dando início aos trabalhos, a Profª. Valéria Gonçalves Soares, cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra à candidata para que a mesma fizesse, oralmente, a exposição do trabalho de dissertação intitulado "*MINERAÇÃO DE DADOS EM DATA WAREHOUSE PARA SISTEMA DE ABASTECIMENTO DE ÁGUA*". Concluída a exposição, a candidata foi argüida pela Banca Examinadora que emitiu o seguinte parecer: "*Aprovada*". Assim sendo, deve a Universidade Federal da Paraíba expedir o respectivo diploma de Mestre em Informática na forma da lei e, para constar, eu, professor José Antônio Gomes de Lima, membro do Colegiado deste Programa, representando a coordenação do PPGI, lavrei a presente ata que vai assinada por mim mesmo e pelos membros da Banca Examinadora. João Pessoa, 29 de maio de 2009.

José Antônio Gomes de Lima

Profª. Dra. Valéria Gonçalves Soares
Orientadora (DI-UFPB)

Prof. Dr. Lucídio dos Anjos Formiga Cabral
Examinador Interno (DI-UFPB)

Prof. Dr. Ed Porto Bezerra
Examinador Interno (DI-UFPB)

Prof. Dr. Heber Pimentel Gomes
Examinador Interno (UFPB)

Profª. Dra. Sônia Virgínia Alves França
Examinador Externo (UFRPE)

# AGRADECIMENTOS

**A Deus** pelo dom da vida e pelas oportunidades concedidas em minha vida, permitindo-me enveredar pelo caminho da ciência e do saber, e dando-me o alento necessário para prosseguir. Nossa aliança é eterna!

**À Nossa Senhora,** pelo seu grande exemplo de vida, mostrando-me o caminho da fé, superação, esperança, tolerância, doação e principalmente, seu exemplo de amor.

**Aos meus pais que tanto amo, Severino M. Gouveia e Ilsaira M. M. Gouveia**, pelo exemplo de dedicação, amizade, amor incondicional e investimento dispensado ao longo da minha formação.

**Ao Prof. Dr. Heber Pimentel Gomes** um agradecimento muito especial pelas orientações e pela amizade conquistada ao longo desses dois anos juntos ao Laboratório de Eficiência Energética e Hidráulica em Saneamento - LENHS. Seus ensinamentos e motivações foram significantes para a conclusão deste trabalho.

**À Prof$^{a}$. Dra. Valéria Gonçalves Soares Elias** pelas orientações sugeridas, das quais foram úteis ao desenvolvimento desta pesquisa.

**Aos meus irmãos Bruno M. M. Gouveia e Rafael M. M. Gouveia** pelo apoio e harmônica convivência, me incentivando a seguir em frente e fornecendo todo o sustentáculo.

**Ao meu amado Alexandre Magno Gurgel Fialho** pelo amor, dedicação, apoio, carinho e compreensão em todos os momentos.

**Aos meus amigos e colegas da UFPB,** em especial a toda equipe do LENHS – dentre eles, Moisés M. Salvino, Paulo Sérgio O. Carvalho, Saulo B. de Tarso, Magno J. G. Silva e Wil L. L. Camboim – pelo auxílio, incentivo e companheirismo.

**Ao Governo do Brasil**, pelo apoio financeiro concedido através das Centrais Elétricas Brasileiras S.A. (**ELETROBRÁS**), da Financiadora de Estudos e Projetos (**FINEP**) e do Conselho Nacional de Desenvolvimento Científico (**CNPq**).

**À Companhia de Água e Esgotos da Paraíba (CAGEPA),** em nome dos engenheiros Leonardo L. B. Montenegro e Jaqueline Pequeno, pela disponibilização dos dados necessários ao estudo de caso do trabalho.

**À UFPB**, instituição que, através de seus docentes e funcionários, foi responsável pela minha formação acadêmica. E aqueles que contribuíram de alguma forma para a realização deste trabalho. **Muito Obrigada!**

# RESUMO

Esta dissertação se propõe a utilizar tecnologias de Banco de Dados com a finalidade de oferecer apoio à decisão para os gestores do setor de saneamento, haja vista que os serviços de abastecimento de água para uso da população se constituem em um dos principais indicadores da qualidade de vida da humanidade. A idéia fundamental consiste em coletar os dados operacionais, reduzi-los ao escopo de um problema, organizá-los em um repositório de dados, e finalmente aplicar as tecnologias OLAP e os algoritmos de Mineração de Dados, a fim de obter resultados que proporcionem aos gestores um melhor entendimento do comportamento e perfil da companhia. Para facilitar a aplicação de técnicas de Mineração de Dados é necessário que estes dados estejam armazenados apropriadamente. Neste sentido, uma das alternativas para o aumento da eficiência no armazenamento, gestão e operação dos dados para o suporte a decisão baseia-se no desenvolvimento do *Data Warehouse*. Este ambiente constitui fontes de informações estratégicas do negócio, gerando um diferencial competitivo para a companhia. Diante deste contexto, se fez necessário a implementação do repositório de dados, o *Data Warehouse*, para armazenar, integrar e realizar as consultas multidimensionais sobre os dados extraídos da companhia de abastecimento de água. Portanto, esta dissertação de mestrado tem como objetivos projetar um *Data Warehouse* Departamental referente ao setor comercial, também conhecido como *Data Mart*; aplicar as tecnologias OLAP sobre os cubos de dados multidimensionais; e executar algoritmos de Mineração de Dados visando a geração de um sistema de apoio à decisão para minimização das perdas aparentes no sistema de abastecimento urbano de água.

**Palavras chave:** *Data Warehouse*, OLAP, *Data Mining,* Sistemas de Abastecimento de Água e Perdas Aparentes.

# ABSTRACT

This work propose to use technologies of databases with the aim of providing decision support for managers of sector of sanitation, given that the services of water supply for use of the population are a key indicator of quality of life. The fundamental idea is to collect operational data, reduce them to the scope of the problem, organize them into a repository of data, and finally apply the techniques OLAP and Data Mining algorithms to obtain results that give managers a better understanding of the behavior and profile of the company. To facilitate the application of the techniques of Data Mining is necessary that the data are stored properly. Accordingly, an alternative for increasing the efficiency in storage, management and operation of data to support the decision based on the development of Data Warehouse. This is source of strategic information of the business, creating a competitive differential for the company. In this context, was required to implement the repository of data, Data Warehouse, to store, integrate and carry out consultations on the multidimensional data from the company of water supply. Therefore, this Master's thesis aims to design a Data Warehouse relating to Departmental Business, also known as *Data Mart*; applied the technology on the OLAP multidimensional cubes of data, and run the Data Mining algorithms to the generation of a decision support system to minimize the apparent losses in the urban water supply system.

**Keywords:** Data Warehouse, OLAP, Data Mining, Water Supply Systems and Apparent Losses.

# SUMÁRIO

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS

| | |
|---|---|
| BI | *Business Intelligence* |
| CAGEPA | Companhia de Água e Esgotos da Paraíba |
| DW | *Data Warehouse* |
| EIS | *Executive Information Systems* |
| ETL | *Extraction, Transformation and Load* |
| ID-3 | *Iterative Dichotomiser* |
| JDBC | *Java Database Connectivety* |
| KDD | *Knowledge Discovery in Databases* |
| OLAM | *On-Line Analytical Mining* |
| OLAP | *On-Line Analytical Processing* |
| OLTP | *On-Line Transaction Processing* |
| PNCDA | Programa Nacional de Combate ao Desperdício de Água |
| ROLAP | *Relational On-Line Analytical Processing* |
| SAD | Sistemas de Apoio à Decisão |
| SGBD | Sistema Gerenciador de Banco de Dados |
| SNIS | Sistema Nacional de Informações sobre Saneamento |
| SQL | *Structured Query Language* |
| WEKA | *Waikato Environment for Knowledge Analysis* |
| XML | *Extensible Markup Language* |

# CAPÍTULO 1

***Este capítulo introdutório descreve as principais motivações para realização do trabalho, apresenta os objetivos e a justificativa da pesquisa e, finaliza, expondo a estrutura e organização da dissertação.***

## 1 INTRODUÇÃO

Os sistemas informatizados coletam e armazenam enormes quantidades de dados em seus bancos de dados, aumentando o número de corporações que buscam alternativas para um planejamento, controle e gestão mais eficiente das informações armazenadas, com o melhoramento dos processos de apoio à tomada de decisão e sistemas inteligentes, baseados em descobertas de conhecimento.

Nos dias atuais, com a necessidade de desenvolver sistemas para dar suporte a decisões gerenciais, vem sendo utilizado e aperfeiçoado o *Data Warehouse (DW)*. O DW é um ambiente cuja finalidade é extrair, integrar, limpar e dar consistência aos dados provenientes dos sistemas transacionais da companhia. Além disso, o DW dimensiona e consolida esses dados, organizando-os e melhorando a performance das consultas.

Os primeiros sistemas de suporte à decisão ficaram conhecidos como *Executive Information Systems* (EIS), e tornaram-se muito populares devido à rapidez com que geravam as informações. Contudo, a falta de flexibilidade para realizar consultas *ad hoc* e a necessidade de definição de fórmulas e formatação de novos relatórios por parte do usuário, fizeram com que os EIS ficassem restritos à geração de relatórios corporativos pré-estabelecidos. Visando suprir as necessidades acima citadas surgiram as ferramentas OLAP (*On-Line Analytical Processing*). Elas tornaram viável a construção de um ambiente no qual os analistas de negócio pudessem facilmente navegar pelos dados da companhia, realizando consultas *ad hoc*, fazendo novos cruzamentos entre as dimensões de análise.

Diante deste ambiente empresarial cada vez mais competitivo, a tecnologia da informação, quando bem utilizada, torna-se um importante diferencial entre as empresas que buscam excelência na qualidade do serviço prestado. Neste cenário, surgem as técnicas e aplicações de Mineração de Dados com intuito de descoberta de padrões de comportamento e

de novos conhecimentos sobre os dados armazenados. Portanto, a gestão aleatória baseada na intuição dá lugar a inteligência de negócio.

O presente trabalho expõe uma experiência do Processo de Descoberta do Conhecimento em Banco de Dados, também conhecido com *Knowledge Discovery in Databases* (KDD), a fim de observar a viabilidade e aplicabilidade de um caso real de apoio à decisão. O estudo segue sob a forma da pesquisa bibliográfica, da criação e implementação do *Data Warehouse* Departamental, do uso de tecnologias de análise e recuperação de dados úteis ao processo decisório, conhecidas como OLAP, e da aplicação de técnicas e algoritmos de *Data Mining* para descoberta de novos conhecimento e padrões nos dados.

## 1.1 OBJETIVOS

Os serviços de abastecimento de água para uso da população continuam sendo um dos indicadores da qualidade de vida da população, sendo de fundamental importância à saúde e à alimentação. Estudos recentes comprovam que a água está se tornando mais escassa, e que menos de 1% (um por cento) da água no mundo está diretamente acessível ao homem. Cerca de vinte países, a maioria deles na África e no Oriente Médio, sofrem de escassez crônica de água, causando danos severos à produção de alimentos e atraso no desenvolvimento econômico (JAMES, et al., 2002).

O estudo proposto por esta dissertação pretende provocar o interesse em pesquisadores envolvidos com a produção, implantação, manutenção, gerência e utilização de Sistemas de Informações Gerenciais ou de Apoio à Decisão. Assim, o resultado desse trabalho terá sua validade para todos aqueles profissionais envolvidos, de alguma forma, em projetos de *Data Warehouse e Data Mining*.

Os objetivos gerais do trabalho são:

- Projetar e desenvolver um Sistema de Apoio à Decisão (SAD);
- Aplicar as tecnologias de Banco de Dados voltadas para projetos de suporte a decisão (modelagem multidimensional);
- Organizar os dados do setor do sistema de abastecimento de água em um *Data Warehouse*, para que eles possam ser analisados por tecnologias OLAP;
- Encontrar padrões e conhecimentos nos dados do setor analisado através dos algoritmos de *Data Mining*.

De acordo com as peculiaridades do setor, os objetivos específicos são:

- Determinar o perfil do setor e do consumidor, por meio da verificação dos consumos de água, valores faturados (conta de água) e pontos de utilização de água;
- Verificar e diagnosticar a situação dos medidores (hidrômetros) presentes nos imóveis;
- Encontrar respostas para as anormalidades e irregularidades praticadas pelos consumidores da qual a empresa de abastecimento de água desconhece;
- Avaliar as inadimplências dispostas no setor selecionado para o estudo de caso.

Este trabalho visa contribuir para o uso racional e eficiente dos recursos hídricos, para isso são aplicadas tecnologias de Banco de Dados como *Data Warehouse*, OLAP e *Data Mining*. Tais tecnologias se propõem em fornecer à entidade gestora de um sistema de abastecimento de água um controle maior do comportamento dos consumidores e imóveis, proporcionando tomadas de decisões eficientes que buscam a redução de perdas de água e das perdas econômicas da companhia de saneamento.

Neste trabalho há a necessidade de conhecimentos envolvendo os dados históricos, tais como o tempo em que o cliente se encontra inadimplente junto à operadora de abastecimento de água; dados históricos das contas e consumos de água e esgoto, histórico do hidrômetro (dados relativos à troca do hidrômetro), etc. Os algoritmos de *Data Mining* com dados que variam com o tempo (séries temporais) são utilizados neste trabalho para prever novos conhecimentos a partir dos dados históricos da série. Tais algoritmos analisam a quantidade de dados existentes e fornecem uma previsão do que pode acontecer nos próximos períodos, levando em consideração os dados passados da base temporal.

As tecnologias de *Data Warehouse* serão utilizadas como parte do processo de descoberta de conhecimento na base de dados do setor de saneamento da cidade de João Pessoa-PB. O ambiente de *Data Warehouse* organizará e disponibilizará os dados, visando facilitar os comandos e execuções OLAP e as consultas para o processo de *Data Mining*.

O termo *Data Warehouse Departamental* é sinônimo de *Data Mart*. Já o termo *Data Warehouse Corporativo* é distinto de ambos. Desta forma, ao longo da dissertação serão encontrados os termos *Data Warehouse, Data Warehouse Departamental ou Data Mart*, ambos indicando o mesmo conceito, ou seja, um armazém de dados para o setor de saneamento urbano da cidade de João Pessoa - Paraíba.

O uso das tecnologias OLAP proporcionará as agregações e sumarizações dos dados contidos no *Data Warehouse*, gerando informações úteis ao processo decisório e oferecendo uma análise mais detalhada do setor. A ferramenta OLAP utilizada neste trabalho foi *Pentaho Analysis View*, que por sua vez utiliza a ferramenta *Pentaho Schema Workbench*, ambas serão apresentadas no capítulo 3.

A aplicação do *Data Mining* visa encontrar os consumidores em potencial que apresentam algumas ou todas as características daqueles que já cometeram algum tipo de fraude e/ou inadimplência na rede de distribuição de água, assim como detectar erros e anormalidades na medição do consumo de água por meio dos hidrômetros. Ao constatar tais irregularidades e anormalidades nos consumos e faturas, ações poderão ser tomadas por parte da companhia para eliminá-las, reduzindo o alto índice de perdas de água e consequentemente o alto percentual de perdas de faturamento.

Os resultados obtidos com o *Data Mining* serão utilizados a fim de detectar padrões, descobrir regras significativas e estabelecer relações entre os índices de inadimplências e anormalidades das ligações de água e esgoto dos consumidores, na tentativa de reduzir os índices de perdas aparentes na distribuição de água.

Os dados serão extraídos do *Data Warehouse* Departamental para em seguida alguns algoritmos de *Data Mining* serão aplicados sobre esses dados pelo software *Pentaho* WEKA. Os resultados serão analisados com o propósito de obter medidas corretivas e preventivas para minimizar o problema das perdas aparentes nos sistemas de abastecimento de água. Serão utilizados e comparados entre si três algoritmos de mineração de dados do Aprendizado Indutivo Supervisionado. Quanto ao Aprendizado Indutivo Não-Supervisionado será aplicado um algoritmo que servirá como complemento no processo de descoberta do conhecimento dos dados contidos no *Data Warehouse* (Os tipos de Aprendizado Indutivo serão explanados na seção 2.5.2).

## 1.2 MOTIVAÇÃO DA PESQUISA

As companhias de saneamento no Brasil perdem em média 44,18% da água que corre no seu sistema de abastecimento, de acordo com o Programa Nacional de Combate ao Desperdício de Água (PNCDA), (MARCKA, et al., Revisão 2004). Boa parte desta água se perde antes mesmo de chegar aos imóveis e atender a população, isto é, a água que se perde entre as estações de tratamento (ETA) e a rede de distribuição do consumidor final.

Segundo o Ministério das Cidades, além dos impactos negativos que as perdas hídricas provocam nos custos operacionais, ampliando a necessidade de investimento em novas instalações de produção e tratamento, elas também causam danos à natureza, pelo aumento da demanda, e geram prejuízos à distribuição regional, principalmente para áreas do Nordeste, onde há escassez de recursos hídricos, e também do Sudeste, cuja região concentra a maior parte da população.

O problema das perdas aparentes em sistemas de abastecimento de água é um assunto que está sempre em foco, visto que o uso correto e consciente da água pela população e pela companhia é significante para o desenvolvimento da humanidade. A detecção das perdas aparentes tem sido de grande interesse para diversas companhias de abastecimento de água, uma vez que representam um fator negativo, tanto financeiro quanto ambiental. Foi desta forma que surgiu o interesse de aprofundar nesta área e desenvolver este trabalho de mestrado.

Portanto, a motivação da presente dissertação surge do interesse de investigar mais detalhadamente se as perdas aparentes de água estão distribuídas proporcionalmente pela cidade ou se estão concentradas em áreas específicas, como por exemplo, nos setores onde o poder aquisitivo dos consumidores é baixo. Para o estudo de caso, serão utilizados dados de um setor do saneamento da cidade de João Pessoa - Estado da Paraíba.

A Companhia de Abastecimento de Água da Paraíba (CAGEPA) disponibilizou o setor 64, na cidade de João Pessoa-PB, para o estudo de caso da presente pesquisa. Este setor corresponde ao sistema de abastecimento urbano de água do bairro e comunidade de Miramar e suas proximidades. Ele apresenta realidades sociais distintas, contemplando população de classe alta, média e a população de baixa renda (habitações populares), além de dispor de diversos tipos de estabelecimentos (comercial, público, industrial, residencial, etc.). Este setor possui aproximadamente 17.800 pontos de utilização e 1.300 consumidores.

A solução desenvolvida nesta dissertação poderá ser aplicada para os demais setores da cidade, trazendo como resultado futuro, uma visão geral dos consumidores de todo o setor de saneamento de João Pessoa. A idéia fundamental desta pesquisa de mestrado é traçar e analisar o perfil dos consumidores e dos imóveis quanto à medição e às perdas aparentes em um determinado período de referência contínuo.

## 1.3 JUSTIFICATIVA DO TRABALHO

As perdas de água em sistema de abastecimento de água correspondem ao volume de água retirado dos mananciais, e que se encontra na Estação de Tratamento de Água (ETA), subtraído dos volumes de água medidos nos hidrômetros. As ações que visam o controle e a redução de perdas de água delineiam-se na melhoria da qualidade da operação e gestão dos sistemas de abastecimento de água e, consequentemente, inserem-se no contexto do uso racional da água.

### 1.3.1 Perdas em Sistemas de Abastecimento de Água

Segundo (MARQUES, et al., 2006), o volume de água computado pela companhia de abastecimento de água que não foi faturado corresponde ao índice de perda do sistema. Estas perdas podem ser geradas por vazamentos nas tubulações da rede de distribuição, erros de medição, fraudes nos hidrômetros, erros cadastrais, inadimplências ligações clandestinas de água etc. As perdas são de dois tipos: Reais e Aparentes.

#### 1.3.1.1 Perdas Reais

Segundo (GOMES, et al., 2007), as perdas físicas de água, também chamadas de Perdas Reais, ocorrem em todo o sistema de abastecimento, desde o ponto de captação até os de consumo, passando pela estação de tratamento, de bombeamento, reservatórios, rede de distribuição e ligações prediais. Elas representam a água que efetivamente não chega ao consumidor, em decorrência de vazamentos nas redes de distribuição e seus ramais provocados por deficiência nos equipamentos, envelhecimento das tubulações e conexões, e operação e manutenção inadequada em todo o sistema.

#### 1.3.1.2 Perdas Aparentes

De acordo com a *International Water Association* (IWA), as Perdas Aparentes, também chamadas de Perdas Não Físicas ou Comerciais, referem-se a toda água que não é medida ou que não tenha o seu uso definido. Ocorre com a água que é tratada e fornecida pela companhia, e consumida pelos clientes, porém não é corretamente medida e, portanto não é faturada, nem gera arrecadação correspondente. Estão relacionadas às ligações clandestinas e/ou irregulares, fraudes nos hidrômetros, erros de micro e macromedição, política tarifária, erro cadastral (desatualização do cadastro, inatividade em ligação ativa, ligação não cadastrada por descuido), erro de leitura, etc.

Para (JAMES, et al., 2002), algumas das causas para as Perdas Aparentes são os erros e desatualizações no cadastro de clientes; Fraudes, violação ou danificação de medição nos hidrômetros[1]; e Ligações Clandestinas ou Ligações não Cadastradas.

Segundo estima (QUEYROI, 2007), metade dos problemas no segmento de saneamento estão ligados a vazamento, ou seja, perdas físicas, e a outra metade são decorrentes de falhas na medição, ou seja, perdas aparentes.

De acordo com (SNIS, 2007), as regiões Norte e Nordeste são as áreas onde há maior perda de faturamento e são também onde predominam as menores rendas per capta no país. Isto aponta para dois aspectos possíveis de situações de perdas: um relacionado ao baixo poder de consumo destas populações, altos índices de inadimplência e consequentemente lucros menores e outro relacionado às grandes potencialidades de irregularidades nas redes, com perdas de volumes de água tratada em função das ligações clandestinas.

No que se refere aos dados do (SNIS, 2007), o valor médio das perdas de faturamento para todo o conjunto de prestadores de serviços foi de 39,8%. Ressalta-se, segundo o relatório, que os prestadores com maiores perdas concentraram-se nas regiões Norte (53,4%) seguida do Nordeste (45,1%). A região Sudeste possui índices de perdas em torno de 39,8%, Centro-Oeste de 39,2% e Sul de 26,6%.

A Companhia de Água e Esgotos da Paraíba (CAGEPA), utilizada no estudo de caso, obteve um intervalo de perdas de faturamento entre 40,1 e 50,0 %. Este alto índice reflete-se de forma negativa para o Estado, visto que as perdas de faturamento estão diretamente ligadas às perdas reais e aparentes. Estas, por sua vez, acarretam problemas estruturais, ambientais e sociais para toda a população.

É importante reduzir as perdas aparentes para elevar a eficiência do sistema de abastecimento de água. Na tentativa de minimizar e evitar tais desperdícios, este trabalho empenha-se em investigar e detectar perdas aparentes, e para alcançar este objetivo, utilizou-se o processo de descoberta do conhecimento em base de dados, com ênfase no *Data Mining*.

[1] Por exemplo: rompimento do lacre e inversão do hidrômetro; execução de *by pass (i.e.,* desvio feito no aparelho, evitando que ele meça corretamente o volume consumido); colocação de arame para travar a turbina do hidrômetro etc.

A análise de grande volume de dados permitirá que se observem tendências, que se detectem regiões onde as perdas aparentes e inadimplências dos consumidores são mais freqüentes; quais são categorias de consumo mais suscetíveis às perdas, entre outras ações.

## 1.4 ESTRUTURA DA DISSERTAÇÃO

A presente dissertação está organizada em 7 capítulos, incluindo este introdutório. O Capítulo 2 configura o estado da arte da pesquisa e tem como objetivo apresentar os principais conceitos envolvidos com o tema da dissertação, sob forma de uma revisão bibliográfica.

O capítulo 3 apresenta e caracteriza a companhia de abastecimento de água envolvida no estudo de caso; e relaciona a teoria exposta no capítulo 2 sob a forma de um estudo de caso real. Nele serão discutidas as tecnologias de banco de dados aplicadas ao setor de saneamento, além de descrever os mecanismo de criação e implementação do *Data Warehouse*; a utilização das tecnologias OLAP e de *Data Mining,* apresentando suas principais funções, vantagens e aplicabilidade.

O capítulo 4 apresenta os resultados e discussões do estudo de caso, apresentado as comparações dos algoritmos de *Data Mining* quanto ao seu tipo de aprendizado indutivo.

O capítulo 5 retoma as discussões gerais do trabalho de forma conclusiva, finalizando a dissertação com os resultados e contribuições relevantes, dificuldades encontradas e as indicações para trabalhos futuros. O último capítulo expõe as referências bibliográficas consultadas.

# CAPÍTULO 2

***Este capítulo configura o estado da arte da dissertação e empenha-se em discutir os assuntos e requisitos relacionados aos Sistemas de Apoio à Decisão, Data Warehouse, OLAP e Data Mining. São apresentados os principais conceitos, o histórico e importância de cada um no processo decisório, mostrando sua relevância para o atual mercado competitivo e tecnológico do Business Intelligence.***

## 2 FUNDAMENTAÇÃO TEÓRICA

### 2.1 SISTEMA DE APOIO À DECISÃO

Os Sistemas de Apoio à Decisão (SAD), ou *Decision Support Systems* (DSS), visam proporcionar uma avaliação crítica das informações dos negócios, auxiliando a gerência a definir tendências, apontar problemas e absorver decisões inteligentes.

De acordo com (DATE, 2004), o processo de tomada de decisão com auxílio de computadores iniciou na década de 70, onde os processos começaram a ser informatizados e as informações passaram a ser pré-definidas e selecionadas por meio dos *Executive Information Systems* (EIS). Na fase atual, os processos de tomada de decisão são totalmente informatizados e o gestor define os atributos mais importantes ao processo decisório, recebendo subsídios e informações processadas pelos Sistemas de Apoio à Decisão, através de ferramentas OLAP, que será discutida na seção 2.4.

Nas décadas anteriores, o foco estava voltado ao crescente aumento da quantidade de informação armazenada em formato eletrônico. Segundo (ZARUR, 2005), estima-se que a quantidade de dados duplica a cada um ano e meio e que o tamanho e número de bases de dados crescem a um ritmo ainda mais elevado. Este grande aumento deve-se essencialmente à constante diminuição do custo de armazenamento dos dados e ao efetivo aumento da eficiência dos computadores em manuseá-los.

De acordo com (ELMASRI, et al., 2005), os Bancos de Dados de apoio à decisão costumam ser extensos, fortemente indexados e envolver uma grande quantidade de

redundância, em especial, sob a forma de replicação e de tabelas de totalização. As chaves costumam envolver um componente temporal e as consultas costumam ser complexas.

Certos aspectos dos sistemas de BD para apoio à decisão os distinguem dos sistemas de BD tradicionais, sendo o principal deles o fato dos BD para apoio à decisão serem quase que exclusivamente para leitura/consultas, e dificilmente para atualizações. Como consequência, observa-se as dificuldades em se trabalhar na prática com um grande número de variáveis, que são os atributos do BD, e a grande quantidade de dados históricos. Em virtude desta complexidade, opta-se por extrair apenas as informações mais relevantes da base de dados transacional.

O bom processamento de extração dos dados é a principal razão para o sucesso na tomada de decisão. Esta extração corresponde à cópia dos dados desejáveis do ambiente operacional para o processamento subsequente. Significa que os usuários podem operar sobre os dados extraídos da maneira como desejarem, sem interferência no ambiente operacional.

Após tantos anos de concentração na obtenção de dados, o problema, agora, passa a ser o aproveitamento deste precioso recurso. Reconheceu-se que estes dados propiciam aos indivíduos responsáveis pelas decisões, o planejamento das ações, a definição de estratégias e a eficácia em suas decisões.

O apoio à decisão se utiliza de várias tecnologias, dentre elas, *Data Warehouse*, *Data Mart*, Sistema Gerenciadores de Banco de Dados, Processamento Analítico On-line (OLAP), Banco de Dados Multidimensionais, Mineração de Dados (*Data Mining*) etc.

As Ferramentas de Apoio à Decisão (FAD) fazem parte do conceito de *Business Intelligence* (BI), ou Inteligência de Negócios, e correspondem ao conjunto de tecnologias que permitem o cruzamento de informações e suporte a análise dos indicadores de desempenho de um negócio (COLAÇO, 2004).

Estas ferramentas são softwares desenvolvidos com objetivo de apresentar graficamente (e não apenas numericamente) as informações do negócio, auxiliando a simulação de ocorrências, fornecendo maior capacidade de análise para o descobrimento de novos conhecimentos e padrões.

### 2.1.1 Descoberta de Conhecimento em Banco de Dados

O processo de descoberta de conhecimento em banco de dados se propõe em encontrar e interpretar padrões através das análises nas fontes de dados. O objetivo é extrair de grandes bases de dados, sem nenhuma formulação prévia de hipóteses, as informações desconhecidas, válidas e acionáveis, que poderão ser úteis para a tomada de decisão.

Ficou mais conhecido pelo acrônimo KDD, que em inglês significa *Knowledge Discovery in Database*. O processo de KDD foi proposto para determinar as etapas que produzem conhecimentos a partir dos dados e, principalmente, definir a etapa de *Data Mining* (Mineração de Dados), que é a fase que transforma dados em conhecimento (FAYYAD, et al., 1996).

Como ilustra a Figura 2.1, cada fase da execução do processo KDD possui uma interseção com as demais. Deste modo, os resultados produzidos em uma fase podem ser utilizados para melhorar os resultados das próximas fases. Este cenário revela um processo iterativo, que busca sempre aprimorar os resultados a cada iteração.

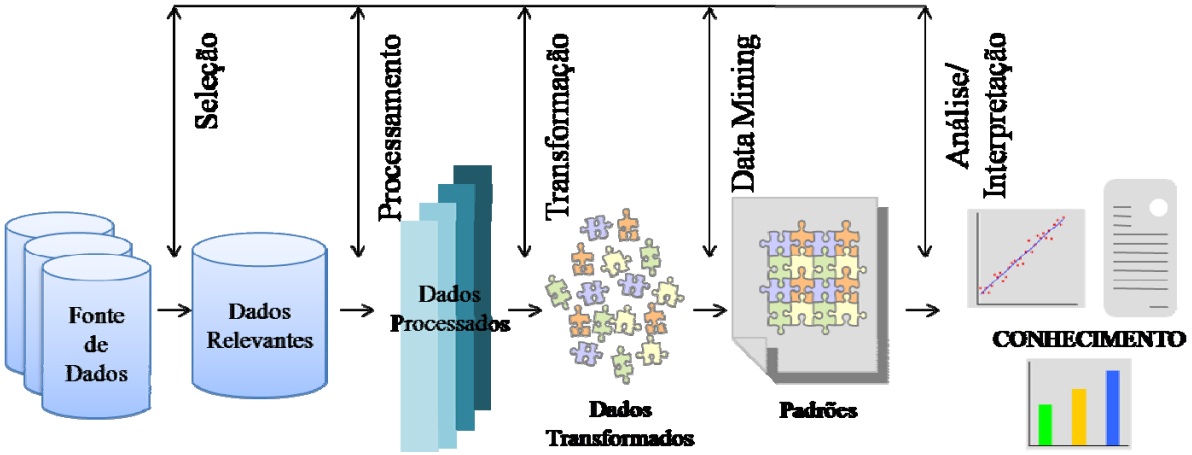



**Figura 2.1 - etapas do processo de KDD**

**Fonte: (Adaptação) (SYMEONIDIS, et al., 2005 p. 14)**

O processo de KDD envolve três etapas iniciais: seleção, (pré) processamento e transformação, as quais compõem a preparação dos dados. Em seguida vem a fase de *Data Mining*, considerada essencial ao processo e foco principal deste trabalho. Por fim, o conhecimento gerado é analisado e assimilado, por meio da etapa de análise e interpretação dos resultados, que se encontra no topo do processo.

## 2.2 DATA WAREHOUSE

Os *Data Warehouses* podem ser traduzidos como Armazéns de Dados e são tipos especiais de banco de dados que se tornaram conhecidos e bastante utilizados a partir da década de 90. Será utilizado o termo em inglês neste trabalho, visto que a maioria dos autores utiliza-o por considerarem mais intuitivo. De acordo com (INMON, 2005), o termo é definido como "um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões da gerência". Onde não volátil significa que, uma vez inseridos, os dados não podem ser alterados, embora possam ser excluídos. O conceito de armazém de dados surgiu por duas razões: primeiro, pela necessidade de fornecer uma origem de dados única, limpa e consistente para fins de apoio à decisão; segundo, pela necessidade de fazê-lo sem causar impacto sobre os sistemas operacionais.

O processo de desenvolver e gerenciar repositórios de dados a partir de várias fontes com o propósito de obter uma visão detalhada e singular de parte ou todo um negócio, é conhecido como *Data Warehousing*. De acordo com (GARDNER, 1998), a concretização do *Data Warehousing* é considerada um dos primeiros passos para tornar factível a análise de grande quantidade de dados no apoio ao processo decisório.

Segundo (PONNIAH, 2001), o *Data Warehousing* não é um software ou produto de hardware que se adquire para fornecer informações estratégicas. É, sim, um ambiente computacional onde os usuários são colocados diretamente em contato com os dados que necessitam para tomar as melhores decisões.

O produto principal obtido de um projeto de *Data Warehousing* é o seu *Data Warehouse* (DW), e cujo objetivo básico é gerar um repositório que contenha dados limpos, agregados e consolidados, podendo este ser analisado por ferramentas do tipo OLAP (*On-Line Analytical Processing*) e *Data Mining* (assuntos abordados nas seções 2.4 e 2.5, respectivamente).

As bases de dados convencionais (relacionais) possuem algumas características, tais como dinamismo, redundâncias, incompletude e ruídos, tornando-as confusas e não viáveis à extração de informações delas próprias. O *Data Warehouse* surgiu com o objetivo de fornecer os subsídios necessários para a transformação de uma base de dados que utiliza *On-Line Transaction Processing* (OLTP) para *On-Line Analytical Processing* (OLAP). A primeira significa os processamentos que executam as operações do dia-a-dia da organização e a

última, os processamentos que suportam a tomada de decisões. Os termos OLTP e OLAP serão detalhados na seção 2.4.

Alguns problemas são apontados por (KIMBALL, et al., 2002; IMHOFF, et al., 2003) quanto ao uso do modelo relacional pra a realização de consultas complexas. A manipulação dos dados, incluindo as consultas, é muito mais rápida e intuitiva no modelo multidimensional em comparação ao modelo relacional.

Enquanto uma busca no modelo relacional exige a navegação entre diversas tabelas, no modelo multidimensional isto não é necessário, o que o torna mais eficiente e com melhor desempenho. Devido ao grande número de tabelas normalizadas do modelo relacional, torna-se inviável a realização das consultas, já que é preciso fazer um grande número de conexões (*inner join*) entre as mesmas.

Os benefícios da modelagem multidimensional é que ela torna os esquemas de dados mais compreensíveis para os usuários finais, e por outro lado, ela permite usar armazenamento específico e técnicas de acesso que melhoram o desempenho de *queries*. A maneira para obter estes benefícios é a simplificação dos esquemas de dados, de forma que eles só contenham as coisas essenciais (i.e. um fato para ser analisado e suas dimensões de análise).

Constantemente há atualização na base de dados e consequentemente as informações históricas são perdidas. Na projeção de bases de dados para *Data Warehouse*s, deve-se quebrar o paradigma dos modelos de dados normalizados utilizados nos BD tradicionais, e buscar armazenamento histórico/temporal. Ao desnormalizar as tabelas, o projetista do DW busca ganhar desempenho nas consultas, contudo, não se deve introduzir redundância em qualquer lugar do modelo.

A idéia dos *Data Warehouses* geralmente se destina a fornecer uma única origem aos dados para todas as atividades de apoio à decisão. O propósito de construir uma espécie de warehouse limitado e de uso especial, adaptado à finalidade imediata, é uma solução aos problemas encontrados com os *Data Warehouses* coorporativos, visto que desta forma é possível o acesso mais rápido aos dados, ao contrário se eles tivessem que ser sincronizados com todos os outros dados a serem carregados no warehouse completo. Essas considerações levaram ao conceito de *Data Marts*, que será apresentado no próximo Item.

Existem três tipos principais de processamentos usados com o *Data Warehouses* (HAN, et al., 2006):

- Processamento de Informação: suporta consultas, análises estatísticas e relatórios;
- Processamento Analítico: ferramentas OLAP e suas operações;
- Processamento de Mineração de Dados: descoberta de conhecimento automatizada, encontrando padrões escondidos nos dados. Pode-se realizar visualizações dos dados, assim como classificações e predições através das técnicas de *Data Mining*.

### 2.2.1 Data Mart

De acordo com (KIMBALL, et al., 2002 p. 36):

> *"Um Data Mart é um Data Warehouse de menor capacidade e complexidade usado para atender a uma unidade específica de negócios. Portanto, são tipicamente mais fáceis de construir e manter."*

Um *Data Mart*, segundo (INMON, 2005) é uma coleção de assuntos organizados para dar suporte à tomada de decisão e estão baseados nas necessidades de um determinado departamento. É geralmente descrito como um subconjunto dos dados extraído para um ambiente separado. Eles são úteis nas seguintes condições:

- Os dados devem estar segregados para melhorar o desempenho do sistema do ponto de vista do usuário.
- Deve existir uma cópia dos dados onde apenas pessoas com autorização podem ter o privilégio de acessá-las.
- Em um ambiente corporativo, é importante fortalecer o conceito de propriedade dentro do banco de dados. Diferentes setores (Financeiro, Marketing, Vendas, etc.) serão responsáveis por diferentes *Data Mart*s.

Um *Data Mart* representa uma área específica a partir de um único processo empresarial, sendo considerado a parte de um todo. É por isso que o *Data Mart*, que é uma abordagem descentralizada do conceito de *Data Warehouse*, não é um "pequeno *Data Warehouse*", mas sim uma unidade lógica de um DW, podendo ser qualificado como um *Data Warehouse* Departamental. A Tabela 2.1 relaciona algumas diferenças entre o ambiente de *Data Mart* e o ambiente de *Data Warehouse*.

**Tabela 2.1 - diferenças entre *data mart* e *data warehouse***

| *Data Mart* | *Data Warehouse* |
|---|---|
| Departamental (única área); | Corporativo (múltiplas áreas); |
| Nível tático; | Nível estratégico; |
| Otimizado para acesso e análise; | Otimizado para armazenamento e gerenciamento de grandes volumes de dados; |
| Poucas fontes de dados; | Muitas fontes de dados; |
| Pequenos estágios de implementação (menor tempo) | Múltiplos estágios de implementação (maior tempo); |

**Fonte: (INMON, 2005)**

Observa-se que as principais diferenças entre *Data Mart* e *Data Warehouse* estão relacionadas ao tamanho e o escopo do problema a ser resolvido. Enquanto um *Data Mart* trata de problema departamental ou local, um *Data Warehouse* envolve o esforço de toda a companhia para que o suporte à decisões atue em todos os níveis da organização. Desta forma, o desenvolvimento de um *Data Warehouse* requer tempo, dados e investimentos gerenciais muito maiores que um *Data Mart*.

De acordo com (INMON, 2005), um dos assuntos em pauta para a área de TI nos últimos anos é decidir qual ambiente de apoio à decisão desenvolver primeiro, o *Data Warehouse* ou os *Data Marts*. A escolha entre um único *Data Warehouse* Corporativo e uma arquitetura consistindo de muitos *Data Mart*s é um ponto de algumas controvérsias entre os pesquisadores. Uma boa parte dos especialistas defende a implementação de *Data Mart*s como passo inicial e existe uma unanimidade de especialistas alertando ao usuário que em momento algum ele pode esquecer o modelo corporativo, sob o risco de obter sérios prejuízos.

Após o levantamento e definição do conjunto de atributos e dados necessários para realização desta pesquisa, optou-se por implementar um *Data Warehouse* Departamental, ou seja, um *Data Mart* do departamento comercial A escolha se deu em virtude dos dados adquiridos corresponderem às informações comerciais dos consumidores e imóveis de um setor da companhia de abastecimento de água. Os resultados obtidos com aplicação das ferramentas OLAP e *Data Mining* sobre o *Data Warehouse* Comercial visam à criação de um novo ambiente computacional com o propósito de fornecer informação estratégica para a companhia de saneamento.

A presença de vários *Data Marts* em uma mesma companhia oferece alto risco de redundância dos dados. Esses ambientes de armazenamento e análises de dados fisicamente distintos trazem benefícios e facilidades, entretanto, existe um preço a se pagar. Desta forma, ao construir *Data Marts* deve-se sempre ter a preocupação de compartilhamento de dados, tabelas e relatórios em comum entre os demais departamentos, consequentemente entre os demais *Data Marts*. Afinal, relatórios em comum não podem possuir valores diferentes entre os departamento.

A separação física dos dados em diferentes grupos, pela presença de vários *Data Marts* em uma única companhia, diminui a habilidade de organização das informações. A dificuldade em evitar a inconsistência dos dados pode ir contra o paradigma de um *Data Warehouse*. Afinal, uma das principais motivações para o surgimento do DW foi eliminar as inconsistências dos dados e agrupá-los em um único ambiente de apoio à decisão.

### 2.2.2 Propriedades do Data Warehouse

De acordo com (INMON, 2005), o DW deve seguir quatro propriedades fundamentais, são elas: Orientado por Temas, Integrado, Variante no Tempo e Não Volátil.

A propriedade "Orientado por Tema", (INMON, 2005) refere-se à importância de organizar as informações pelos temas principais. Para o setor de saneamento, que caracteriza o estudo de caso deste trabalho, os principais temas são: perfil dos consumidores e imóveis, serviço prestado e perdas aparentes.

Cada tema pode envolver várias tabelas e atributos e podem existir dados acumulativos e detalhados. Para o tema perfil dos consumidores, por exemplo, os atributos podem ser os dados cadastrais (nome, endereço, telefone, e-mail), dados das contas e consumos de água, etc. Como exemplo de dados acumulativos tem-se a consulta que retorna o somatório dos consumos descendentes, agrupados por clientes no período de 2007 a 2008.

A propriedade "Integrado" presente em um DW mostra a necessidade de acoplar dados de diferentes formatos. Os dados precisam seguir uma convenção padrão para que desta forma eles possam fornecer significados únicos. Um sistema do setor comercial pode codificar o "indicativo de medidor" como SIM ou NÃO. Onde SIM se refere ao consumidor que possui hidrômetro para medição do consumo de água e NÃO caracteriza o consumidor que não possui hidrômetro para medição. Outro setor da companhia de abastecimento pode

codificar 0 (Tem Hidrômetro) e 1 (Não tem Hidrômetro), assim como S (Tem Hidrômetro) e N (Não tem Hidrômetro). Desta forma, é necessário definir uma única codificação dos dados extraídos para o *Data Warehouse*.

A terceira propriedade "Variante no Tempo" em um ambiente de *Data Warehouse* determina que os dados não sejam atualizáveis e que eles possam ser comparados ao longo do tempo. Os dados são atribuídos como retratos da base de dados operacional atual, onde cada ocorrência e cada mudança são consideradas como um novo registro, pois a informação histórica não é perdida.

Contudo, em um Ambiente Transacional[2] a atualização dos dados ocorre em virtude das mudanças ocorridas. Os dados retornados em consultas correspondem à informação no momento da consulta, e neste caso as consultas históricas não são consideradas[3].

Supondo que desejamos recuperar a quantidade de pontos de consumo do consumidor. Em 2007 o consumidor possuía 20 pontos de consumo em sua residência, já em 2008 passou para 23 pontos de consumo. A consulta retornará apenas a estado atual dos pontos de consumo, ou seja, 23. A informação histórica anterior é perdida. Entretanto, no DW ao consultar os pontos de acesso do cliente em 2007, do exemplo acima, o resultado corresponderá ao valor 20.

A última propriedade proposta por (INMON, 2005), que é a "não volatilidade" dos dados, se verifica em banco de dados que é disposto fisicamente para otimizações de inclusões e consultas. Ou seja, não deve ser um banco preparado para atualizações.

O DW consiste em fornecer apenas acessibilidade aos dados, não permitindo atualizações ou alterações. Ele concede apenas a carga inicial e consulta (acessos) aos dados. Ao contrário, a volatilidade é uma propriedade bastante observada em ambientes operacionais tradicionais, pois os registros dos dados são atualizados constantemente.

---

[2] Conhecido também por "Ambiente Operacional". O termo mais utilizado nesta dissertação é "Ambiente Transacional".
[3] Neste caso não estão sendo mencionados os ambientes que utilizam Banco de Dados Temporais (BDT), apenas os que utilizam Banco de Dados Relacionais.

### 2.2.3 Granularidade

A questão da granularidade é um dos mais importantes aspectos no projeto de *Data Warehouse.* Corresponde ao nível no qual os dados estão sumarizados no *Data Warehouse*, ou seja, refere ao nível de detalhamento das informações armazenadas. Quanto mais detalhados os dados, menor a granularidade do DW (granularidade fina ou baixa). Quanto maior o nível de granularidade, menor será os detalhes dos dados (granularidade grossa ou alta).

Segundo (PONNIAH, 2001 p. 23), a granularidade está diretamente ligada ao volume de informações armazenadas e aos tipos de consultas que podem ser realizadas pelo usuário de um DW. Ao definir um nível muito detalhado, o usuário poderá ver a informação em qualquer nível de agregação e maior será o detalhamento das consultas. Contudo, a escolha de um nível baixo demais poderá ocasionar em um aumento do volume de dados armazenado e, consequentemente, afetará a performance do sistema. Por outro lado, ao definir um nível pouco detalhado, o usuário ficará impossibilitado de realizar consultas mais detalhadas, visto que o volume de informações armazenadas é menor, porém, permite maior desempenho e rapidez nas respostas das consultas.

Portanto, quanto mais alto o nível de granularidade, menor o volume de dados e o número de índices e, indiretamente, menor o processamento necessário. O problema existente é que o nível de granularidade é também inversamente proporcional ao número de consultas que podem ser atendidas.

A utilização de apenas um nível de granularidade em projetos de *Data Warehouse* não é recomendada como solução eficiente. Afinal, o nível de granularidade é inversamente proporcional à quantidade de consultas atendidas e/ou desempenho do processamento. O modelo dimensional (ver item 2.3) é o mais utilizado nas aplicações de DW, e este utiliza técnicas de níveis duais de granularidade.

O desenvolvimento de um ambiente com níveis duais de granularidade consiste em ter dados de um mesmo assunto em granularidades diferentes. A opção pelo uso de níveis duais tem como finalidade baixos tempos de resposta nas consultas de granularidade alta e análise dos dados em maior detalhe nas consultas com níveis de granularidade baixa.

A razão pela qual a granularidade é a principal questão de projetos de *Data Warehouses* consiste no fato de que ela afeta profundamente o volume de dados, ao mesmo

tempo afeta no tipo de consulta que pode ser atendida. O volume de dados residentes no DW deve ser balanceado de acordo com o nível de detalhe de uma consulta.

### 2.2.4 Arquitetura do Data Warehouse

Em um ambiente projetado de *Data Warehouse* há duas espécies de dados: Dados Primitivos (operacionais ou atômicos) e Dados Derivados (de apoio à decisão ou sumarizados). Os dados primitivos consistem em valores referentes ao momento presente, e são baseados em aplicações, podem ser atualizados, são detalhados, e processados repetitivamente. Enquanto que os dados derivados são geralmente valores históricos, baseados em assuntos ou negócios, são resumidos, ou refinados, não são atualizados, representam valores de momentos já decorridos ou instantâneos e são processados de forma heurística (INMON, 2005).

A escolha de dados primitivos para o armazenamento em um DW proporciona vários benefícios, porém gera algumas desvantagens. O maior benefício está na possibilidade de se pesquisar em base de dados mais rica, proporcionando uma análise mais aprofundada e cuidadosa nos dados, o que permite a verificação do histórico, de tendências, de previsões e de elaboração de cenários. A principal desvantagem é a necessidade de um espaço muito maior nos dispositivos de armazenamento, assim como uma maior capacidade de processamento para que não haja baixa performance nas consultas e análises dos dados.

A escolha de dados derivados para o armazenamento em DW também traz benefícios e desvantagens. O maior benefício é que os dados já estão sumarizados, ou seja, já estão resumidos e armazenados em um formato no qual são mais consultados. Ocupam menos espaço nos dispositivos de armazenamento e a performance das consultas e das análises dos dados é mais rápida. A desvantagem é que o armazenamento dos dados sumarizados limita bastante a capacidade de pesquisa e de análise. A maioria das empresas opta pelas duas formas de armazenamento simultaneamente. Desta forma, somam-se as vantagens e reduzem-se as desvantagens de ambas.

Segundo (INMON, 2005), com estas diferenças nos dados, tem-se a projeção de quatro níveis do ambiente arquitetural de um DW, são eles: Nível Operacional (ou Transacional), Nível Atômico (ou *Data Warehouse*), Nível Departamental (ou *Data Mart*) e Nível Individual, como mostra a Figura 2.2.

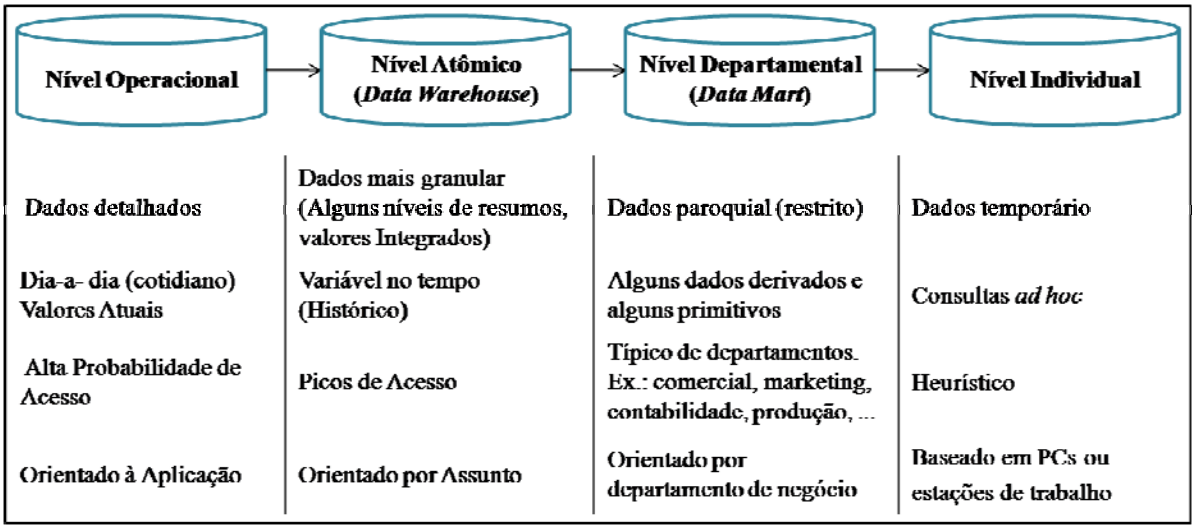



**Figura 2.2 - os quatro níveis de dados do ambiente arquitetural de um *data warehouse***

**Fonte: Adaptação de (INMON, 2005)**

O nível Operacional de dados detém apenas a aplicação orientada a dados primitivos e atende à comunidade de processamento de transações de alta performance. O nível de *Data Warehouse* contém dados primitivos que não são atualizados, além de alguns dados derivados. O nível Departamento contém quase que exclusivamente dados derivados. Este nível é moldado pelas necessidades dos usuários finais adaptadas às necessidades do departamento. E o nível individual de dados é onde muitas das análises heurísticas são realizadas. Segue a Figura 2.3 com exemplos dos quatro níveis de dados.

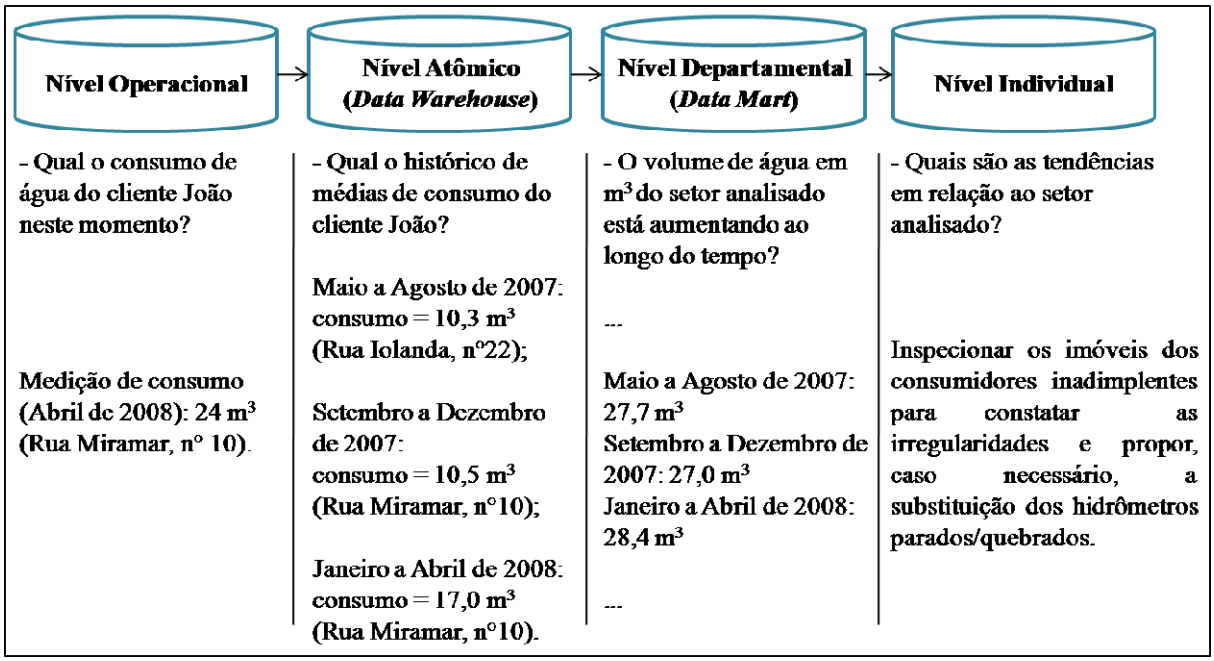



**Figura 2.3 - exemplos de consultas referentes aos quatro níveis de dados**

O Nível Operacional retornará a média na medição de consumo de água do cliente João (nome e endereço fictício) na última medição efetuada, ou seja, em Abril de 2008 e que corresponde a 24 $m^3$de água. O registro neste nível contém os valores recentes do cliente, onde para se conhecer a situação atual dele, é acessado o registro existente neste nível. Para alteração dos dados de João, o registro do nível operacional será alterado, com o objetivo de refletir os novos dados atualizados.

O segundo nível, nível de *Data Warehouse*, resulta no histórico de consumo do consumidor João, isto é: 10,3 $m^3$ entre Maio e Agosto de 2007, 10,5 $m^3$ entre Setembro e Dezembro de 2007 e média de volume de 17,0 $m^3$ entre Janeiro e Abril de 2008. Neste nível existem vários registros do João, apresentando o histórico das informações sobre ele. Não há sobreposição nos registros existentes no ambiente de DW. Quando houve mudança de endereço do consumidor (da Rua Iolanda para Rua Miramar), foi gerado um novo registro no DW, refletindo as datas do período que João residiu naquele local.

O terceiro nível, nível de *Data Mart*, permitirá ao executor extrair informações de maior complexidade e específico do negócio, facilitando as tomadas de decisões. Um exemplo seria uma lista com todos os clientes por categoria, sendo o consumidor João incluído nesse resumo de cada quadrimestre. Como consulta do nível 3 tem-se: "O volume de água em $m^3$ do setor analisado está aumentando ao longo do tempo (relatório quadrimestral)?". O retorno desta consulta são as médias de consumo agrupadas por quadrimestre (Maio a Agosto de 2007; Setembro a Dezembro de 2007 e Janeiro a Abril de 2008).

Por fim tem-se o nível Individual, que possibilita a previsão de informações, fornecendo visões futuras por meio das análises heurísticas. Os dados neste nível são, geralmente, temporários e de pequenas proporções.

No exemplo apresentado na Figura 2.3, ao analisar o setor observou-se que a maioria dos consumidores inadimplentes possui hidrômetros instalados a mais de 10 anos e com capacidade de vazão de até 3 $m^3$. Ainda no nível Individual, verificou-se que aproximadamente metade dos consumidores está com consumo de água igual a zero, o que representa hidrômetro parado. Estes resultados indicam casos onde uma inspeção técnica poderia ser realizada, afinal os equipamentos de medição podem estar defasados e/ou

danificados, gerando perdas aparentes no sistema. Na seção 3.3.3 é proposto um modelo de Mineração de Dados aplicado à inspeção e troca de hidrômetros.

## 2.3 MODELAGEM DIMENSIONAL

A modelagem dimensional[4] é uma metodologia que possibilita que os dados sejam modelados visando aperfeiçoar o desempenho de consultas e oferecer facilidades de utilização a partir de um grupo de eventos simples de medição. A visão dimensional facilita o entendimento e visualização de problemas típicos de sistemas de apoio à decisão, é mais intuitiva e eficaz para o processamento analítico e é utilizada pelas tecnologias OLAP (discutidas na seção 2.4).

Três conceitos estão envolvidos com a modelagem dimensional, são eles: *fatos*, *dimensões* e *métricas* (medidas ou atributos). De acordo com (BALLARD, et al., 1998), um *fato* é uma coleção de itens de dados que consiste de métricas e do contexto do negócio. A *dimensão* é uma coleção de itens do mesmo tipo que representa as visões do negócio. A *métrica* é definida como um atributo numérico de um fato, e representa o comportamento do negócio para as dimensões.

Os fatos são reunidos na tabela de fatos. Segundo (KIMBALL, 1997), as tabelas de fatos normalmente contém dados numéricos e somatórios. Como os Data Warehouses geralmente recuperam muitos registros em uma única consulta, é uma tendência agrupar os dados para análise, pois esta compactação proporciona ganhos de performance. Cada dimensão possui uma tabela de dimensão associada que armazena as descrições textuais das dimensões do negócio. Cada tabela de dimensão tem uma chave primária que corresponde exatamente a um dos componentes da chave composta da tabela de fatos.

A Tabela 2.2 a seguir apresenta o modelo dimensional implementado em SGBD Multidimensional e SGBD Relacional. Os dados da tabela correspondem às médias de consumo em m3 das quadras 010, 015, 020 e 025, agrupadas por categoria de consumo durante o período de 2007 a 2008.

[4] Os termos “modelagem dimensional” e “modelagem multidimensional” são utilizados na literatura para expressar o mesmo conceito. Não há uma definição padrão que indique uma diferença precisa entre os dois termos.

**Tabela 2.2 - exemplo da modelagem dimensional em SGBDS**

| Quadra | Categoria | | |
|---|---|---|---|
| | Comercial | Industrial | Residencial |
| Quadra_010 | 190.0 | - | - |
| Quadra_015 | 34.3 | 23.5 | 114.0 |
| Quadra_020 | 38.2 | - | 88.8 |
| Quadra_025 | - | - | 19.8 |

**Modelagem Dimensional em SGBD Multidimensional**

Painel de saída

Saída de Dados | Explain | Mensagens | Histórico

| | quadra text | categoria text | media_consu numeric |
|---|---|---|---|
| 1 | Quadra_010 | COMERCIAL | 190.0 |
| 2 | Quadra_015 | COMERCIAL | 34.3 |
| 3 | Quadra_015 | INDUSTRIAL | 23.5 |
| 4 | Quadra_015 | RESIDENCIAL | 87.9 |
| 5 | Quadra_020 | COMERCIAL | 25.5 |
| 6 | Quadra_020 | RESIDENCIAL | 88.8 |
| 7 | Quadra_025 | RESIDENCIAL | 19.8 |

**Modelagem Dimensional em SGBD Relacional (PostgreSQL)**

A principal vantagem na utilização de SGBDs Multidimensionais é que eles implementam fisicamente o modelo dimensional. Contudo, uma das desvantagens é a esparsidade, ou seja, células que ocupam espaços em disco, mas não contêm dados cadastrados, como é caso das quadras 010, 020 e 025. Outra desvantagem é considerada quando o modelo dimensional possui um grande número de dimensões, pois traz como consequências, problemas de desempenho e tempo maior de processamento das consultas. Os SGBDs Relacionais possuem uma maior aceitação e utilização, entretanto, exigem adaptações, visto que eles não implementam fisicamente o modelo dimensional.

Existem três esquemas utilizados para modelagem dimensional dos dados, são eles: Esquema Estrela (S*tar Schema*), Esquema Floco de Neve (S*nowflake* S*chema*) e Esquema Constelação de Fatos (*Facts Constallation Schema*).

### 2.3.1 Esquema Estrela

Idealizado e criado por Ralph Kimball, o Esquema Estrela é uma forma de dispor as tabelas do modelo relacional para o modelo dimensional, podendo ser implementado em BD relacionais e principalmente, em BD multidimensional (KIMBALL, et al., 2002).

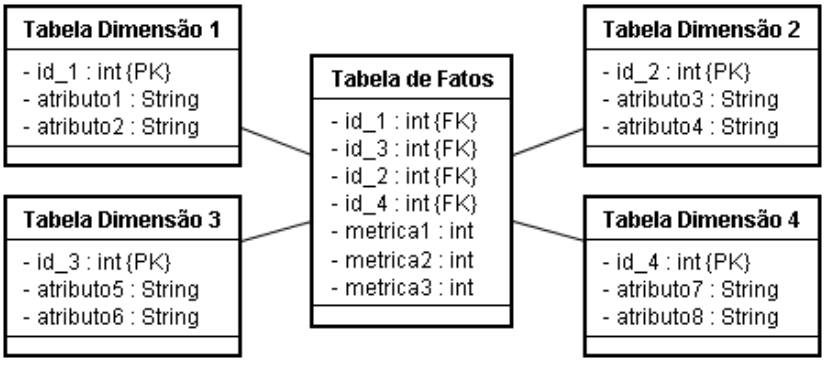



**Figura 2.4 - exemplo geral do esquema estrela**

Conforme ilustra a Figura 2.4, o Esquema Estrela é uma estrutura com tabelas e ligações bem definidas, baseado no formato de uma estrela. É formado por uma tabela central, denominada *tabela de fatos*, a qual possui os dados principais da visão da análise, ou seja, o assunto que está sendo analisado, por exemplo, o consumo, as quantidades de inadimplentes, as quantidades de consumidores, etc. Nela ficam ligadas as tabelas de dimensão, que possuem os aspectos pelos quais se deseja observar as medidas relativas ao processo que se está analisando.

De acordo com (HAN, et al., 2006), as tabelas dimensionais são desnormalizadas para aumentar o desempenho das consultas. A consulta ocorre inicialmente nas tabelas de dimensão e em seguida na tabela de fatos, assegurando a precisão dos dados através de uma estrutura completa de chaves onde não é preciso percorrer todas as tabelas. Isso garante um acesso mais eficiente e um melhor desempenho.

Ao contrário das tabelas de dimensão, a tabela de fatos armazena grandes quantidades de dados históricos, normalmente numéricos, obtidos a partir da interseção de todas as dimensões do Esquema Estrela. Ela também armazena os indicadores de desempenho (medidas) do negócio. Para cada dimensão há uma chave primária que corresponde a um dos campos, chave estrangeira, da chave da tabela de fatos.

A Tabela 2.3 apresenta um comparativo entre os dois tipos de tabelas do Esquema Estrela, mostrando as diferenças entre elas.

**Tabela 2.3 - comparativo entre as tabelas de fatos e dimensão**

| Tabela de Fatos | Tabela de Dimensão |
|---|---|
| Grande volume de dados | Volume comparativamente menor |
| Chave composta | Chave simples |
| Referencia cada tabela de dimensão | Descrevem os fatos |
| Histórica | Atributos usados como filtro nas consultas |
| Agiliza consultas, pois os fatos (variáveis) são usualmente numéricos e tipicamente aditivos | Desnormalizada (redundâncias) |

**Fonte:** (KIMBALL, et al., 2002)

Apesar do Esquema Estrela apresentar desvantagens em termos de espaço de armazenamento devido à redundância dos dados e, principalmente, fazer com que o desempenho diminua nas operações de atualização dos dados, no qual o custo para manter a

integridade é muito alto, esta característica não possui importância em um *Data Mart* por se tratar de uma estrutura de dados que sofre pouca ou nenhuma atualização.

### 2.3.2 Esquema Floco de Neve

O Esquema Floco de Neve é uma extensão do Esquema Estrela e consiste na decomposição de uma ou mais dimensões, formando hierarquias nas dimensões, isto é, normalizando-as. Esse tipo de esquema é utilizado quando se tem dimensões grandes que são estáticas ou semi-estáticas. A Figura 2.5 ilustra um exemplo geral deste tipo de esquema, nele as dimensões 2 e 4 foram normalizadas.

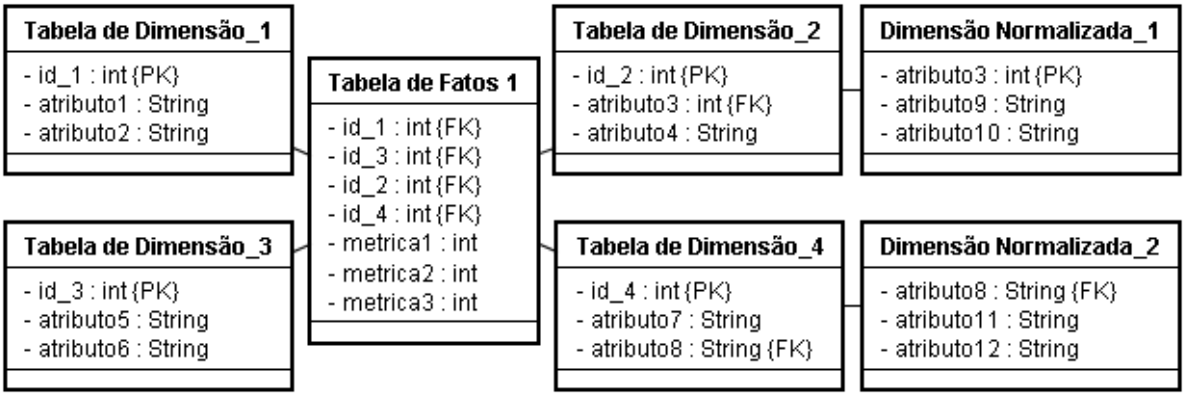



**Figura 2.5 - exemplo geral do esquema floco de neve**

A vantagem do seu uso está na diminuição do volume de dados trazido para a memória, além dos *inner join* com a tabela normalizada ser mais facilmente resolvido. No Esquema Floco de Neve o número de relacionamentos entre as tabelas é maior, fazendo com que o tempo de execução das consultas aumente devido à necessidade de operações de junção. Durante a especificação das tabelas do *Data Mart* é importante levar em consideração estas características de forma a normalizar as tabelas somente nos casos em que não haja uma grande perda de desempenho. Em geral, recomenda-se utilizar o Esquema Estrela ou o Esquema Constelação de Fatos, pois ambos possuem dimensões desnormalizadas.

### 2.3.3 Esquema Constelação de Fatos

O Esquema Constelação de Fatos é constituído de duas ou mais tabelas de fatos que compartilham uma ou mais dimensões. Esse tipo de esquema pode ser visto como uma coleção de esquemas estrelas, conforme ilustra a Figura 2.6, na qual a tabela Dimensão 2 e Dimensão 4 são compartilhadas pela Tabela de Fatos 1 e 2.

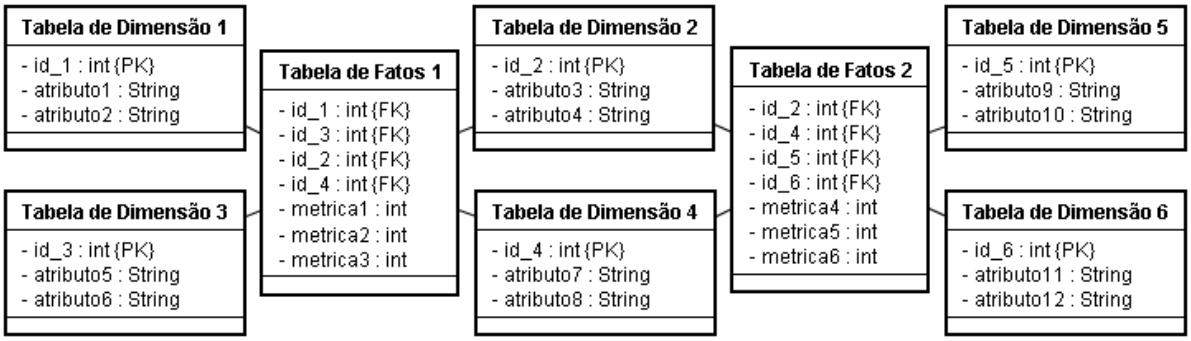



**Figura 2.6 - exemplo geral do esquema constelação de fatos**

Para *Data Warehouses* (ou *Data Marts*), o esquema de Constelação de Fatos é mais comumente utilizado, visto que ele pode modelar assuntos múltiplos e inter-relacionados. Desta forma, o Esquema Constelação de Fatos foi o que apresentou mais adequação para a modelagem dimensional do *Data Mart* implementado para este trabalho, visto que algumas tabelas de dimensão precisaram ser compartilhadas entre as tabelas de fatos. O capítulo 1.3, item 3.2.4, apresenta um exemplo de consulta SQL ao Esquema Constelação de Fatos modelado para o estudo de caso proposto por este trabalho, e apresenta também a tabela resultante com os valores obtidos da consulta.

Na Figura A.1 do APÊNDICE A encontra-se a modelagem completa do Esquema Constelação de Fatos para o Perfil do Setor e das Perdas Aparentes da Companhia de Abastecimento de Água e Esgoto da Paraíba. A Figura A.1 representa a tabela de fatos "Perfil do Setor" e suas 11 dimensões, juntamente com a tabela de fatos "Perdas Aparentes" associada a suas 12 dimensões. Quatro dimensões (Quadra, Matrícula, Inadimplência e Referência de Consumo) são compartilhadas pelas duas tabelas de fatos.

## 2.4 TECNOLOGIAS OLAP

Inicialmente, surgiram as tecnologias conhecidas como *On-Line Transaction Processing* (OLTP) que atendem às necessidades de operações transacionais. Elas denotam as movimentações tradicionais que acessam registros pequenos e individuais. As principais operações neste tipo de processo são alteração, inclusão, exclusão e consultas. Estas operações ocorrem muitas vezes em um mesmo dia e podem ser requisitadas ao sistema simultaneamente por muitos usuários, o que demanda uma resposta quase imediata do sistema. (AURÉLIO, et al., 2000)

As tecnologias *On-Line Analytical Processing* (OLAP), por sua vez, são projetadas para apoiar análises e consultas, além de auxiliar seus usuários a sintetizar informações através de comparações, visões personalizadas e análises históricas. As tecnologias OLAP têm como característica principal permitir uma visão mais fácil e intuitiva dos dados multidimensionais, por meio de análises em diferentes perspectivas (INMON, 2005).

De acordo com (HAN, et al., 2006), OLAP faz parte do processo que habilita usuários a explorar os dados do *Data Warehouse*, fornecendo funcionalidades para análise interativa de dados em diferentes dimensões e granularidades.

Alguns tipos de informações podem ser interessantes ao gerente de uma companhia de abastecimento, como por exemplo: "Qual a quantidade de consumidores, pontos de utilização e quantidade de inadimplências da subcategoria FAVELA, agrupados pelas categorias de consumo (Comercial, Industrial, Público e Residencial), situações da ligação de água (Cortada, Ligada, Suprimida parcial e Suprimida total) e estado de inadimplência (Inadimplência e Adimplência) dos consumidores?", ou ainda, "Qual a média de faturamento das quadras agrupadas pela categoria de consumo comercial e semestres de referência (primeiros seis meses e últimos seis meses de medição)?". Estas e outras consultas utilizando tecnologias OLAP são apresentadas em detalhes a partir da seção 3.2.6, página 93.

O processamento analítico é necessário em diversas situações no qual se deseja obter informações referentes à evolução histórica. Tecnologias OLAP permitem esses tipos de consultas e melhoram o desempenho de tempo em relação àquelas feitas em BD convencionais, ou seja, BD relacionais.

O *On-line Analytical Processing* (OLAP), ou Processamento Analítico On-Line, surgiu pela necessidade de minerar conhecimento e padrões em diferentes níveis de abstração através de análises multidimensionais dos dados, ou seja, uma visão lógica dos dados. É uma análise interativa dos dados, através de agregações em todas as interseções de dimensões necessárias. Permite obter informações sumarizadas e mostrá-las em tabelas 1-D (planilhas), 2-D (dimensões em xy), 3-D (dimensões em xyz), mapas e gráficos, com suporte para modificações dos eixos. Além disso, compõe análises estatísticas (razões, médias, somatórios, mínimos, máximos, contagens, variâncias, etc.) envolvendo quaisquer medidas ou dados numéricos entre muitas dimensões. A Tabela 2.4 mostra as diferenças entre as duas abordagens, OLTP *versus* OLAP.

**Tabela 2.4 - diferenças entre OLAP e OLTP**

| OLAP | OLTP |
|---|---|
| - Relevância para dados históricos; | - Mantém usualmente a situação corrente; |
| - Necessidade de ver o dado sob diferentes perspectivas: aplicações dinâmicas; | - Voltado para velocidade e automação de funções repetitivas; |
| - Atualizações quase inexistentes, apenas novas inserções; | - Atualizações em grande número; |
| - Baseado em dados históricos, consolidados e frequentemente totalizados; | - Baseado em transações; |
| - Operações de agregação e cruzamentos. | - Alto nível de detalhe. |

**Fonte: (COLAÇO, 2004)**

De acordo com (GONZALES, 2003), o termo OLAP também é usado para descrever a estrutura de armazenamento dos dados e os métodos utilizados para acessá-los. OLAP representa diversos tipos de tecnologias que variam no método de acesso. Há três adaptações de métodos de acesso OLAP, que são: OLAP Multidimensional (MOLAP); OLAP Relacional (ROLAP); OLAP Híbrido (HOLAP).

Os métodos de acesso do tipo MOLAP utilizam a estrutura de dados multidimensional e permitem a navegação pelos níveis de detalhamento em tempo real. Utiliza SGBDs Multidimensionais otimizados ao máximo para as consultas OLAP e com tratamento dimensional nativo. Requer migração dos dados do SGBD Relacional para o armazenamento multidimensional e a sua constante atualização. Teoricamente, é a melhor arquitetura de acesso a ambientes multidimensionais, mas na prática deixa a desejar pela falta de SGBDs Multidimensionais mais consolidados, dificultando sua aplicação.

Os métodos de acesso do tipo ROLAP é a solução mais utilizada hoje e surgiram em decorrência do uso consagrado dos SGBDs Relacionais nos BDs operacionais (transacionais), com todas as vantagens da tecnologia aberta e padronizada da linguagem SQL. Os dados obtidos dos bancos fontes são armazenados em SGBDs Relacionais, formando o *Data Warehouse* com tabelas implementadas em estruturas relacionais clássicas. O método de acesso ROLAP foi a solução adotada neste trabalho.

É uma tendência dos SGBDs Relacionais modernos adicionarem uma arquitetura multidimensional para prover facilidades à ambientes de suporte a decisão. Tal conceito fez surgir os métodos de acesso do tipo HOLAP, isto é, mistura do ROLAP com o MOLAP, que proporciona o desempenho e flexibilidade de um BD Multidimensional e mantém a gerenciabilidade, escalabilidade, confiabilidade e acessibilidade conquistadas pelos BDs

Relacionais. A idéia é armazenar dados de maior granularidade do DW em estruturas relacionais normalizadas e os dados agregados de granularidade inferior em estruturas dimensionais nativas.

A visualização multidimensional dos dados através das tecnologias OLAP favorece a análise de várias dimensões em única tela, em virtude da estrutura conceitual conhecida por cubos de dados. A visualização se dá através de configurações tridimensionais de linhas, colunas, operações *Slice and Dice* e gráficos, como mostra a Figura 2.7. Os cubos de dados e operações *Slice and Dice* serão discutidos nas seções 2.4.1 e 2.4.2, respectivamente.

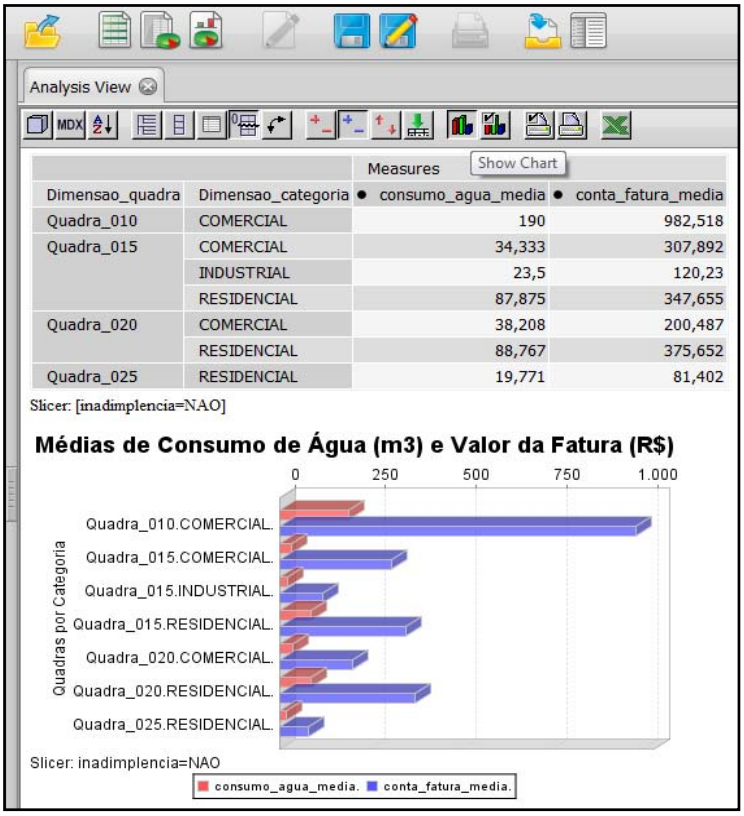



**Figura 2.7 - visualização dos dados através de ferramenta OLAP pentaho analysis view[5]**

**Fonte: Dados do setor de saneamento de João Pessoa.**

Os dados da Figura 2.7 foram obtidos através de uma consulta ao "Esquema Constelação de Fatos" implementado para o estudo de caso deste trabalho. O retorno desta consulta corresponde às médias de consumo de água em $m^3$ e médias da fatura dos

[5] A ferramenta *OLAP Pentaho Analysis View* será discutida com mais detalhes na seção 3.2.6 (página 119).

consumidores adimplentes[6] agrupadas por quadra (010, 015, 020 e 025) e por categoria de consumo durante o período de 2007 a 2008.

A Figura 2.8 ilustra a mesma consulta executada acima, contudo, utilizando o software *pgAdmin III* (desenvolvido para dar suporte ao SGBD *PostgreSQL*).

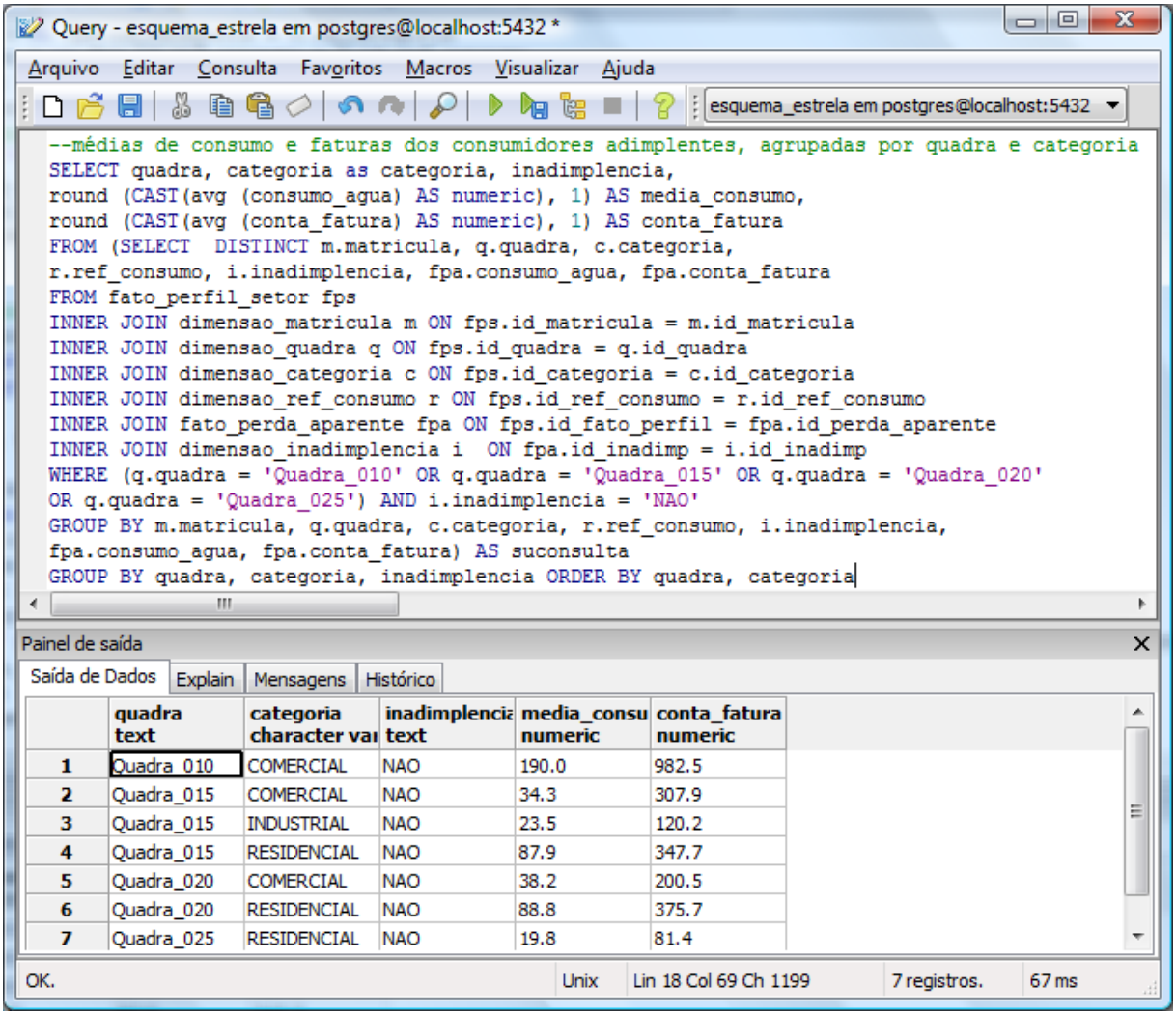



**Figura 2.8 - visualização dos dados através do software PgAdmin**

A principal vantagem em utilizar uma ferramenta OLAP ao invés de uma ferramenta puramente de Banco de Dados, é a facilidade proporcionada pela ferramenta OLAP quanto à visualização e manipulação do modelo dimensional (tabelas de fatos e dimensões). Outra vantagem é que o analista não precisa escrever as *queries* SQL, como ocorre em ambientes puramente de BD, pois a ferramenta OLAP dispõe de *interface* gráfica para dá o suporte a

[6] Inadimplência igual a "NAO" significa que a conta de água foi quitada pelo consumidor junto à companhia de distribuição de água.

realização das consultas. Neste trabalho optou-se por utilizar a ferramenta *OLAP Pentaho Analysis View*, que é apresentada no Capítulo 3, item 3.2.6.

### 2.4.1 Estrutura Multidimensional: Cubo de Dados

A principal característica das tecnologias OLAP é permitir uma visão conceitual multidimensional dos dados de uma empresa. Um cubo de dados é uma estrutura que armazena os dados em formato dimensional. Uma dimensão é uma unidade de análise com dados agrupados.

Por exemplo, a dimensão tempo tem os dados agregados por meses, quadrimestres e semestres. A dimensão categoria tem os dados agregados em comercial, industrial, público e residencial, etc. A Figura 2.9 apresenta os dados modelados numa estrutura conhecida por Cubo, onde cada Dimensão (D1, D2 e D3) representa um tema importante da companhia para realização de análises e comparações. O cubo da Figura 2.9 é "Fato Perfil do Setor" e suas dimensões são Categoria, Status da Água e Status do Esgoto.

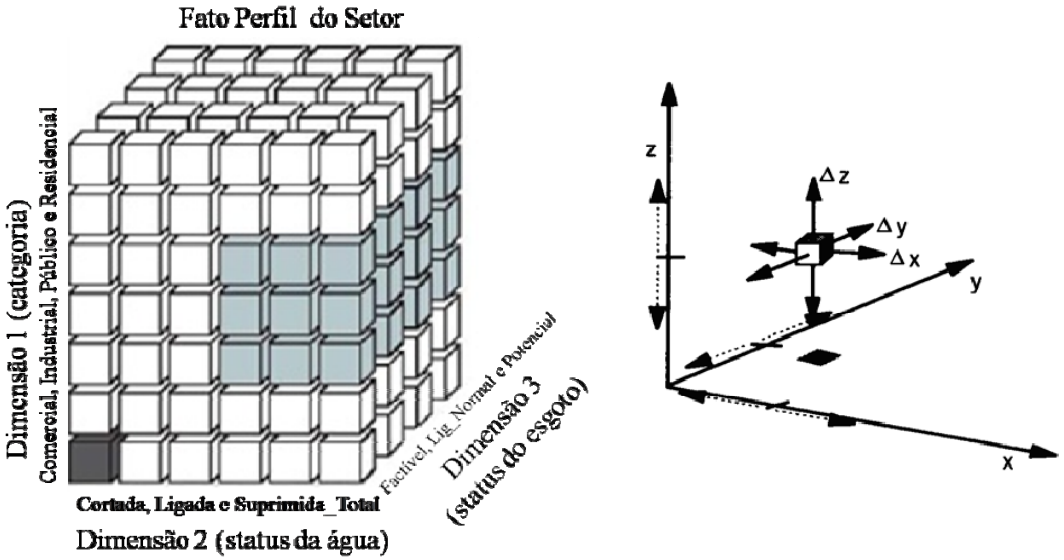



**Figura 2.9 - (a) um cubo de dados com três dimensões. (b) busca tridimensional de células no cubo**

**Fonte: Adaptação de (RAINARDI, 2008).**

A partir da modelagem do Esquema Estrela, Floco de Neve ou Constelação de Fatos pode-se construir os cubos de dados e realizar buscas nesse espaço multidimensional. Os cubos de dados são visões lógicas multidimensionais dos dados com referência hierárquica. As tecnologias OLAP fornecem funcionalidades para análise interativa de dados em diferentes visões e granularidades, permitindo visualizar as hierarquias e navegar pelas dimensões (THOMSEN, 2002).

As operações sobre os cubos de dados foram introduzidas por (GRAY, et al., 1996) visando suportar múltiplas agregações em sistemas de Banco de Dados com suporte a OLAP. O operador Cubo é uma generalização n-dimensional da operação *group-by*, sendo capaz de executar diversos *group-by* correspondentes a diversas combinações.

Na Figura 2.10 é apresentada a idéia envolvendo os operadores de cubo de dados, para isto utilizaram-se as dimensões categoria, situação da água e situação do esgoto, ambas associadas à tabela de fatos "Fato Perfil do Setor" do esquema Constelação de Fatos[7].

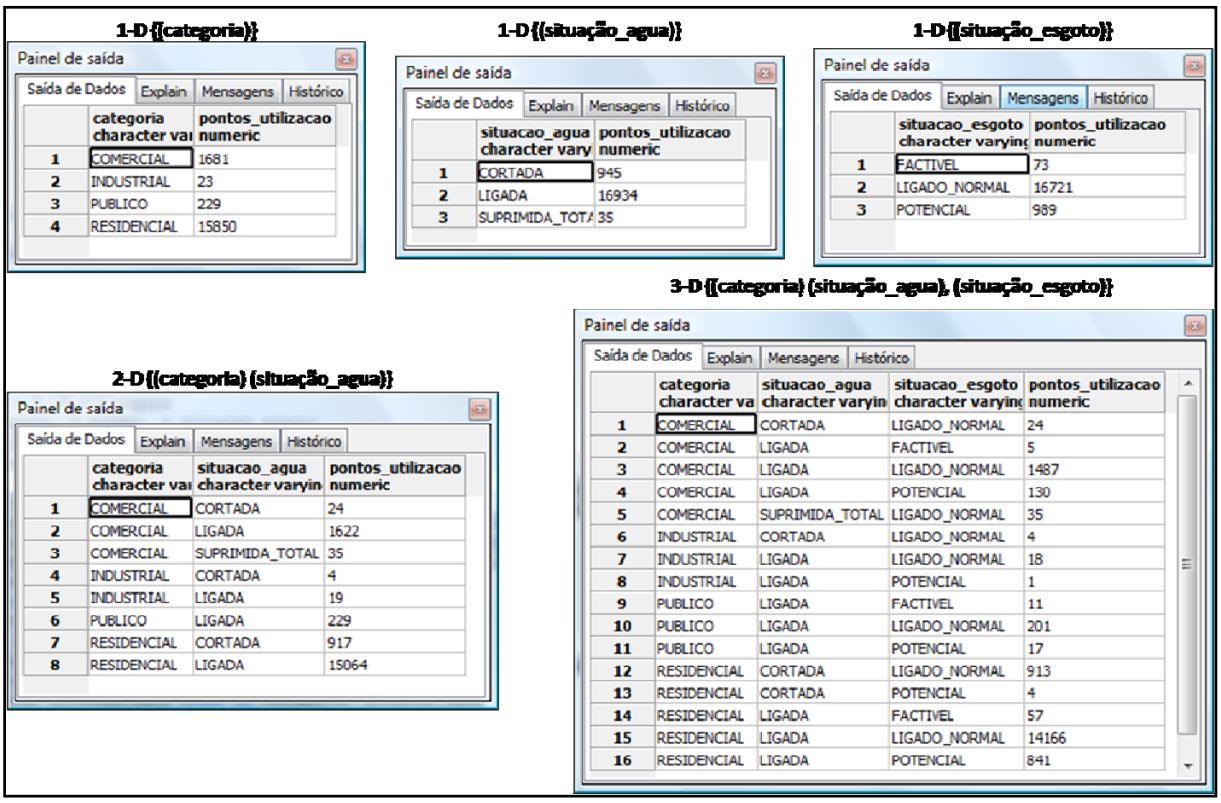



**Figura 2.10 - exemplo de *cuboids* (1-D), (2-D) e (3-D) para o esquema constelação de fatos**

Os agrupamentos das dimensões do esquema constelação de fatos para o perfil do setor geram a computação da ordem $2^3$ agregações, ou seja, 8 *group-by* (*cuboids*) formado pelas combinações 3-D {(categoria) (situação_agua) (situação_esgoto)}; 2-D {(categoria) (situação_agua)}, {(categoria) (situação_esgoto)}, {(situação_agua) (situação_esgoto)}; 1-D {(situação_agua)}, {(situação_esgoto)}, {(categoria)}; e (vazio)[8].

[7] Esquema Constelação de Fatos encontra-se ilustrado na Figura A.1 do APÊNDICE A.
[8] (vazio) representa um *group-by* vazio

No exemplo da Figura 2.10 a dimensão categoria foi associada à dimensão situação da água, o que resultou no *cuboids* de duas dimensões (2-D). A Figura 2.11 ilustra a rede de cubóides completa formada pelas três dimensões agrupadas em cuboids de uma, duas e três dimensões.

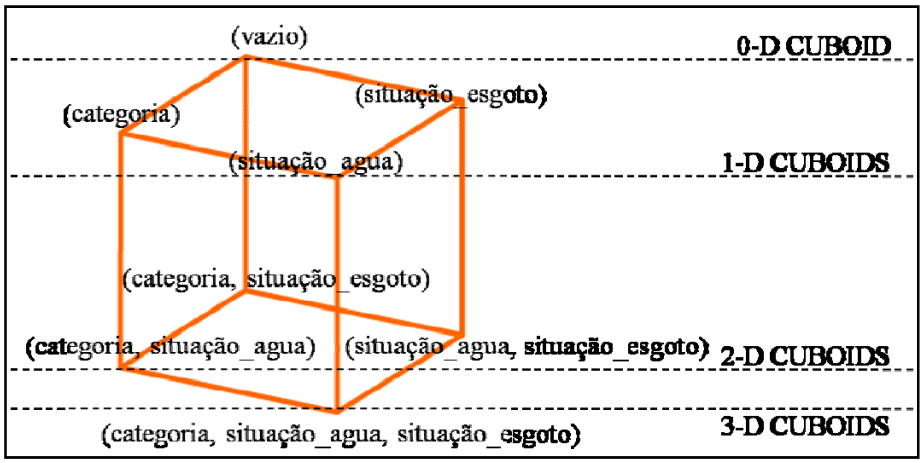



**Figura 2.11 - Rede de *cuboids* para um cubo de três dimensões**

Estudos voltados para manipulação eficiente da estrutura dimensional dos cubos de dados, bem como seus mecanismos de acesso, estão em constante evolução. As pesquisas nesta área buscam otimizar cada vez mais as consultas e operações OLAP, visando o melhor desempenho dos sistemas de apoio à decisão.

### 2.4.2 Conjunto de Operações OLAP

Ao iniciar uma consulta a um *Data Warehouse* é necessário traduzi-la de forma inteligível ao ambiente computacional. Assim, devem ser oferecidos aos analistas meios para realizar eficientemente uma consulta, a fim de obter resultados coerentes. Como solução, os desenvolvedores de ferramentas OLAP fornecem suporte para as operações de derivação de dados complexos, que recebem o nome de *Slice and Dice*.

Segundo (WREMBEL, et al., 2007), o suporte às operações *Slice and Dice* é uma das principais características de uma ferramenta OLAP. A operação *Slice*, suportada pelas ferramentas OLAP, faz restrição de um valor ao longo de uma dimensão. Já a operação *Dice* é mais complexa, pois faz restrições de valores em várias dimensões.

O *Slice and Dice* compreende quatro operações, que são o *Ranging, o Drilling, o Rotation/Pivotting e o Ranking*. A Figura 2.12 ilustra de forma genérica a operação de *Slice, Dice, Rotate, Drill-down e Drill-up/Roll-up.*

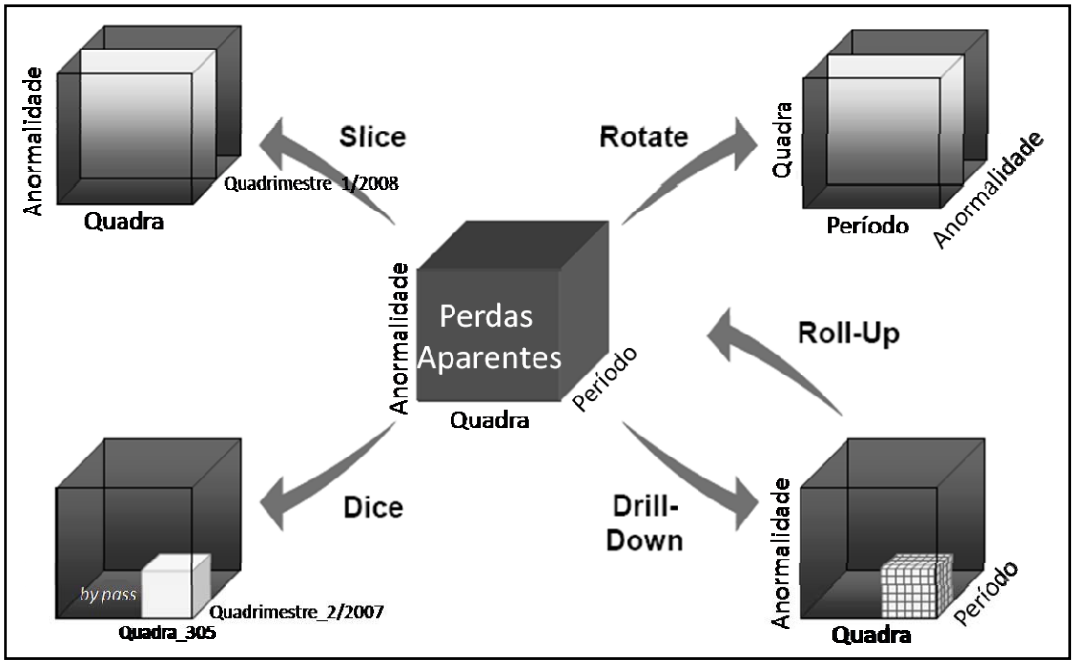



**Figura 2.12 - exemplo da operação slice, dice, drill-down, drill-up e rotate.**

A operação *Ranging* é responsável por, a qualquer momento, alterar o resultado das consultas, inserindo novas posições ou removendo as que estão em foco. Por exemplo, a inserção de uma nova quadra de consumidores em uma consulta representa uma operação de *Ranging*. O resultado dessa operação será considerado para todas as demais, e assim, pode-se encarar o resultado como um novo cubo gerado a partir do cubo original.

A operação *Drilling* consiste em escolher o que deseja analisar, o analista ainda pode mudar o escopo do que está analisando, porém os dados podem encontrar-se agregadas em diversos níveis. A Figura 2.12 ilustra a operação *Drilling.* O *Drilling* permite navegação por entre os níveis. Existem três operações OLAP que permitem mudar o escopo dos dados, são elas: *Drill-Down, Drill-Up* e *Drill-Across*.

A operação *Drill Down* navega verticalmente na hierarquia, no sentido em que os dados são mais atômicos (valores simples, indivisíveis e monovalorados). Consiste em desagregar dimensões. Exemplo: Semestre → Quadrimestre. A operação *Drill-Across* permite navegar transversalmente no eixo da árvore hierárquica. O *Drill-Across* é uma operação de grande utilidade, pois permite inserir e retirar posições do corrente cenário. O *Drill Up* ou *Roll Up* faz parte da operação *Drilling* e realiza a função inversa do *Drill-Down*. Ela permite ao usuário uma visão mais agregada das informações. Exemplo: Quadrimestre → Semestre.

Com esta técnica pode-se navegar nos diversos níveis de maiores detalhes para os níveis mais sumarizados.

A operação *Rotation ou Pivot* além de permitir ao analista mudar as posições das dimensões em foco, e tem também a flexibilidade de alterar o eixo de visualização dos dados, alterando linhas por colunas, com intuito de facilitar a compreensão. Vale salientar que *Rotation* não adiciona nem retira posições do cenário, mas permite ao analista alterar a visão que se tem dos dados. Um exemplo desta operação seria alterar a dimensão quadra da horizontal para a vertical, e desta forma, o cubo rotacionaria no sentido horário.

Com a operação *Ranking* o analista pode filtrar as informações que se deseja obter. É possível fazer uma classificação dos dados adquiridos e operar diretamente sobre os valores das células. Todas as operações anteriores atuavam apenas sobre as posições ou dimensões dos dados, entretanto, através do *Ranking*, o analista pode executar diversos tipos de filtros, eliminando assim os dados desnecessários e inconsistentes.

De acordo com o estudo de caso desenvolvido neste trabalho, as operações OLAP foram aplicadas nas tabelas de fatos e dimensões do *Data Warehouse* Comercial implementado para um setor do saneamento da cidade de João Pessoa. Em todos os casos e exemplos utilizou-se a ferramenta de código aberto *OLAP Pentaho Analysis View*[9].

## 2.5 DATA MINING

As técnicas de *Data Mining* podem ser aplicadas em diversas áreas do conhecimento, dentre elas na Engenharia Hidráulica, que por sua vez, é o objeto do estudo de caso deste trabalho. A sua principal característica é a aplicação dos algoritmos aos dados pré-processados, com o objetivo de auxiliar as companhias, que no caso deste trabalho é a companhia de abastecimento de água e esgoto, a gerar indicadores numéricos, indicadores gráficos e relatórios *ad hoc,* i.e., relatórios onde o analista define o que deseja obter no momento da consulta, através de aplicações que possam servir de apoio à tomada de decisão nos diferentes níveis, sejam eles estratégicos, táticos ou operacionais.

De acordo com (BATISTA, 2003), as etapas de *Data Mining* são:

---

[9] A ferramenta *Pentaho Analysis* faz parte do software livre *Pentaho BI Suite Enterprise Edition* , que se encontra na versão 3.0 disponível em <http://www.pentaho.com/products/analysis/>.

- Escolha da tarefa de *Data Mining*: uma combinação de tarefas deve ser escolhida dentre os vários tipos de tarefas possíveis como: classificação, regressão, associação, *clustering* (Ver Item 2.5.4);
- Escolha do algoritmo de *Data Mining*: de acordo com a tarefa selecionada, um determinado algoritmo, também denominado de técnica, será aplicado nos dados, utilizando-se os modelos e parâmetros mais apropriados (Ver Item 2.5.5);
- Aplicação de *Data Mining*: busca por padrões de interesse particular em uma forma representacional particular ou em um conjunto de aplicações.

### 2.5.1 Metas do Data Mining

Existem duas metas primárias que podem ser alcançadas através de *Data Mining* (FAYYAD, et al., 1996):

- PREVISÃO: antecipar os valores de variáveis desconhecidas ou analisar um possível valor para uma variável com o passar do tempo, utilizando algumas variáveis, como atributos da base de dados. Logo, indica as chances de uma ação ocorrer.
- DESCRIÇÃO: procurar por padrões que descrevem os dados e que sejam de entendimento dos usuários.

A Mineração Preditiva consiste na generalização de exemplos ou experiências passadas com respostas conhecidas ou regras de negócio estabelecidas por especialistas. A Mineração Descritiva consiste na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada.

### 2.5.2 Aprendizado Indutivo

A indução é um meio de inferência lógica que permite que conclusões gerais sejam obtidas de exemplos particulares. É caracterizada como o raciocínio que parte do específico para o geral, do particular para o universal, da parte para o todo.

De acordo com (BATISTA, 2003), um argumento indutivo e correto pode, perfeitamente, admitir uma conclusão falsa, ainda que suas premissas sejam verdadeiras. Se as premissas de um argumento indutivo são verdadeiras, o melhor que pode ser dito é que a sua conclusão é *provavelmente* verdadeira. Desta forma, esse recurso deve ser utilizado com os devidos cuidados, dado que se o número de observações for insuficiente ou se os dados relevantes forem mal escolhidos, as hipóteses induzidas poderão produzir conclusões

errôneas. Apesar disso, a inferência indutiva é um dos principais meios de criar novos conhecimentos e prever eventos futuros.

O *Data Mining* compreende dois tipos de aprendizado indutivo: Supervisionado e Não-Supervisionado. O aprendizado Supervisionado é direcionado a tomada de decisão e é através dele onde se realiza inferências nos dados com o intuito de realizar predições, envolvendo o uso dos atributos para prever o valor futuro. Enquanto que no Aprendizado Não-Supervisionado as atividades são descritivas, o que permite a descoberta de padrões e novos conhecimentos.

#### 2.5.2.1 Aprendizado Supervisionado

O aprendizado supervisionado serve para identificar a classe a que pertence uma nova amostra de dados. Neste tipo de aprendizado é sempre conhecida a classe dos dados que são usados para treino e há um histórico de dados que permite prever sobre dados futuros.

Inicialmente é fornecido ao sistema de aprendizado um conjunto de exemplos E = {$E_1$, $E_2$,..., $E_N$}, onde cada exemplo $E_i \in E$ possui um rótulo associado. Esse rótulo define a classe a qual o exemplo pertence. Formalmente, cada exemplo $E_i \in E$ corresponde a uma tupla $E_i = (\overrightarrow{x_i}, y_i)$. Sendo $\overrightarrow{x_i}$ um vetor de valores que representam as características (atributos) do exemplo $E_i$, e $y_i$ o valor da classe desse exemplo. O objetivo do aprendizado supervisionado é induzir um mapeamento geral dos vetores $\overrightarrow{x_i}$ para valores y. Portanto, o sistema de aprendizado deve construir um modelo, tal que y = f ($\overrightarrow{x_i}$), onde f é uma função desconhecida (função conceito) que permite predizer valores y.

#### 2.5.2.2 Aprendizado Não-Supervisionado

Neste tipo de aprendizado o rótulo da classe de cada amostra de treino não é conhecido e o número de classes a ser treinada pode não ser conhecido a priori. É fornecido ao sistema de aprendizado um conjunto de exemplos E, no qual cada exemplo consiste somente de vetores $\overrightarrow{x_i}$, não incluindo a informação sobre a classe y. O objetivo é construir um modelo que procura por regularidades nos exemplos, formando agrupamentos ou clusters de exemplos com características similares.

O aprendizado não-supervisionado utiliza-se de algoritmos descritivos. As atividades descritivas trabalham com conjuntos de dados que não possuem uma classe determinada e têm

o objetivo de identificar padrões de comportamento semelhantes nestes dados. As tarefas descritivas podem ser divididas em: Associação, Segmentação e Generalização.

A Figura 2.13 apresenta a divisão dos algoritmos de *Data Mining* de acordo com a tarefa (preditiva ou descritiva) da qual fazem parte. Todas as tarefas apresentadas nesta Figura 2.13 serão detalhadas na seção 2.5.4, com ênfase para as tarefas de Classificação e Associação, visto que elas foram utilizadas nos modelos de *Data Mining* aplicados ao estudo de caso.

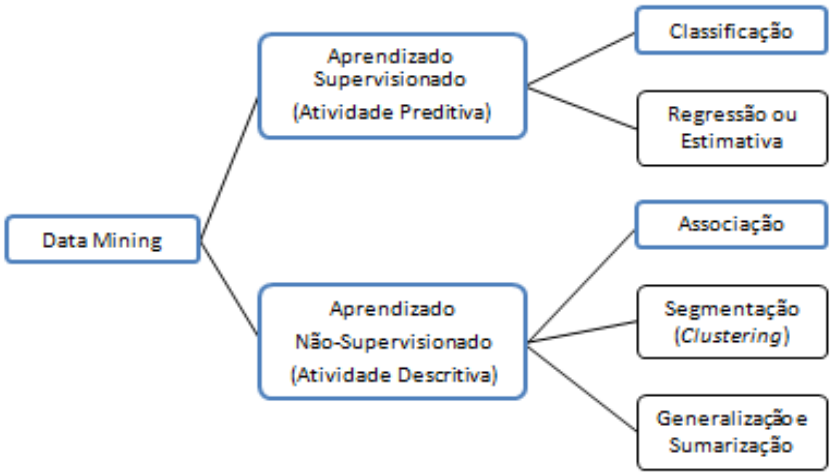



**Figura 2.13 - taxonomia do d*ata mining***

**Adaptação (REZENDE, et al., 2003)**

### 2.5.3 O Processo Iterativo do Data Mining

O primeiro passo no processo de *Data Mining* é a identificação da fonte de dado. A tarefa de identificar os dados começa com a decisão sobre que dados serão necessários para resolver o problema. O próximo passo é o *cleaning*, ou seja, a preparação dos dados. O principal desafio do *cleaning* é formatar os dados de forma compatível com a representação do *software* que será utilizado para mineração.

O terceiro passo é construção do modelo de mineração, e este é constituído das regras que descrevem os dados analisados no banco. Isso é feito automaticamente através de dados analíticos e de algoritmos de *Data Mining*. O quarto passo no processo iterativo de mineração é a avaliação do modelo criado, que consiste em estimar a precisão do modelo e refinar sua compreensão e sua utilidade. Por último surge o desdobramento do modelo, que serve para aplicar o modelo a novos dados, a fim de dar lugar ao surgimento de novas perguntas, trazendo um refinamento adicional às descobertas (SANCHES, 2003).

Ao final do processo de mineração, espera-se, como principal objetivo, o uso das descobertas úteis e suas representações. Desta forma, seguem as ações para a etapa de pós-processamento:

- Interpretação dos Padrões: avaliação e interpretação dos padrões encontrados, a fim de determinar aqueles que terão alguma utilidade e gerarão algum conhecimento. Nesta etapa, pode ocorrer a necessidade de retorno a umas das etapas anteriores;
- Consolidação do Conhecimento: verificação e utilização do novo conhecimento sobre os dados através das ferramentas de visualização. E produção da documentação para auxiliar a compreensão do usuário.

### 2.5.4 Principais Tarefas do Data Mining

A tarefa de *Data Mining* precisa ser definida no início do processo de KDD, no momento em que for decidido o domínio da aplicação (FAYYAD, et al., 1996). Existem diversas tarefas para alcançar as metas de previsão e descrição, discutidas na seção 2.5.1, dentre elas: Classificação; Regressão ou Estimativa; Associação; Segmentação (*clustering*) e Generalização ou Sumarização. Nas seções seguintes serão descritas as duas tarefas de mineração utilizadas no trabalho

#### 2.5.4.1 Classificação

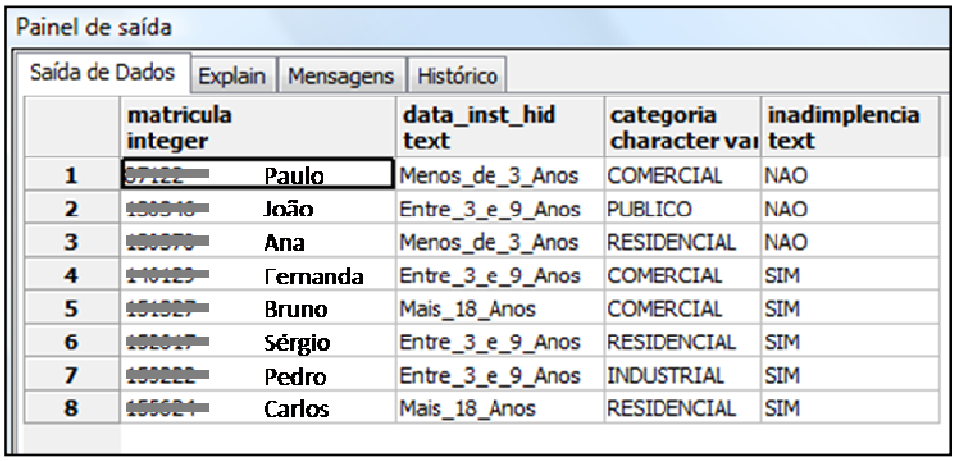
A tarefa de classificação consiste em encontrar propriedades comuns em um determinado conjunto de objetos de um banco de dados e classificá-los em diferentes classes. Os passos para classificação são: definição de um conjunto de exemplos conhecidos (treinamento); treinamento sobre esse conjunto; e geração de regras de classificação ou descrição.

Conforme (BARROSO, et al., 2006), o princípio desta tarefa é descobrir algum tipo de relacionamento entre os atributos preditivos e o atributo objetivo, de modo a descobrir um conhecimento que possa ser utilizado para prever a classe de uma tupla desconhecida, ou seja, que ainda não possui uma classe definida.

O conhecimento descoberto é frequentemente representado na forma de regras SE→ENTÃO. Essas regras são interpretadas da seguinte maneira: “SE os atributos preditivos

de uma tupla satisfazem as condições no antecedente da regra, ENTÃO a tupla tem a classe indicada no consequente da regra".

Como exemplo da tarefa de classificação utilizou-se os dados da Figura 2.14. Ao aplicar a Classificação sobre esses dados, são geradas as regras de classificação, conforme apresenta a Tabela 2.5


Painel de saída

Saída de Dados | Explain | Mensagens | Histórico

| | matricula integer | data_inst_hid text | categoria character var | inadimplencia text |
|---|---|---|---|---|
| 1 | [redacted] Paulo | Menos_de_3_Anos | COMERCIAL | NAO |
| 2 | [redacted] João | Entre_3_e_9_Anos | PUBLICO | NAO |
| 3 | [redacted] Ana | Menos_de_3_Anos | RESIDENCIAL | NAO |
| 4 | [redacted] Fernanda | Entre_3_e_9_Anos | COMERCIAL | SIM |
| 5 | [redacted] Bruno | Mais_18_Anos | COMERCIAL | SIM |
| 6 | [redacted] Sérgio | Entre_3_e_9_Anos | RESIDENCIAL | SIM |
| 7 | [redacted] Pedro | Entre_3_e_9_Anos | INDUSTRIAL | SIM |
| 8 | [redacted] Carlos | Mais_18_Anos | RESIDENCIAL | SIM |

**Figura 2.14 - exemplo de dados utilizados na tarefa de classificação**

Foram geradas quatro regras de classificação. Por exemplo, a primeira regra determina que todos os hidrômetros com mais de 18 anos de funcionamento são de consumidores inadimplentes. Enquanto que a última regra determina que os hidrômetros entre 3 e 9 anos de funcionamento e que pertencem aos consumidores que não sejam da categoria Público, estão inadimplentes.

**Tabela 2.5 - regras de classificação geradas (descobertas) com os dados da Figura 2.14**

| |
|---|
| SE (Instalação_Hid = Mais_18_Anos) ENTÃO Inadimplência = Sim |
| SE (Instalação_Hid = Menos_de_3_Anos) ENTÃO Inadimplência = Não |
| SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria = Público) ENTÃO Inadimplência = Não |
| SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria != Público) ENTÃO Inadimplência = Sim |

No contexto deste trabalho, seguem alguns exemplos onde a tarefa de classificação poderia ser aplicada: classificação do consumidor quanto ao risco (baixo, médio ou alto risco) de inadimplência; classificação do consumidor potencialmente fraudador a julgar pelo seu perfil; classificação das categorias quanto às anormalidades; classificação do tipo de ligações de água quanto legal, clandestina ou suspensa por irregularidade etc.

### 2.5.4.2 Associação

A tarefa de associação foi introduzida por (AGRAWAL, et al., 1993) e tem a finalidade de determinar os grupos de itens que tendem a ocorrer ao mesmo tempo, em uma mesma transação, gerando-se as regras de associação. Elas podem ser vistas como regras do tipo SE-ENTÃO integrada a duas medidas de interesse: *confiança* e *suporte*. A primeira medida corresponde à probabilidade condicional e a segunda corresponde à fração que sustenta a regra. Ambas serão definidas mais adiante.

A regra de associação é um relacionamento X (antecedente) ⇒ Y (consequente), onde X e Y são conjuntos de itens da transação e a interseção X ∩ Y é o conjunto vazio. Cada regra está associada a um Fator de Suporte Superior "Fs" (medida de interesse *suporte*) e a um Fato de Confiança "Fc" (medida de interesse *confiança*). Seguem as fórmulas dos dois fatores:

$$Fs = \frac{|X \cup Y|}{N}, \text{onde N é o número total de tuplas} \qquad Fc = \frac{|X \cup Y|}{X}$$

O fator de suporte[10] pode ser descrito como a probabilidade de uma transação qualquer satisfazer tanto X como Y, ao passo que o fator de confiança[11] é a probabilidade de que uma transação satisfaça Y, dado que ela satisfaça X. A tarefa de descobrir regras de associação consiste em extrair do banco de dados todas as regras com "Fs" e "Fc" maiores ou iguais a um "Fs" e "Fc" especificado pelo analista.

A descoberta de regras de associação segue normalmente em dois passos. Primeiramente, o algoritmo determina todos os conjuntos de itens que têm "Fs" maior ou igual ao "Fs" especificado pelo analista. Estes conjuntos são chamados conjuntos de {Itens Frequentes}. Em seguida, todas as possíveis regras candidatas são geradas e testadas para cada conjunto de {Itens Frequentes} com relação ao "Fc". Apenas as regras candidatas com "Fc" maior ou igual ao "Fc" especificado pelo analista são dadas como saída do algoritmo.

Segue na Tabela 2.6 abaixo, um exemplo do processo de descoberta de regras de associação. A primeira coluna da Tabela mostra o identificador da transação, e as demais

[10] O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador ao total de transações.
[11] O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador se refere à quantidade de transações em que o item X ocorre.

colunas indicam se um determinado item foi ou não localizado na transação correspondente. Suponha que o analista especificou os parâmetros Fs = 0.3 e Fc = 0.8.

**Tabela 2.6 - exemplo de dados para descoberta de regra de associação**

| ID | Consumidor Comercial | Fraude | Corte Ligação | Multa | Parcelamento Fatura | Pagamento Fatura |
|---|---|---|---|---|---|---|
| 1 | N | S | S | S | N | N |
| 2 | S | S | N | S | S | N |
| 3 | N | S | S | S | N | N |
| 4 | S | S | S | S | N | N |
| 5 | N | N | N | N | S | N |
| 6 | N | N | N | S | N | N |
| 7 | N | S | N | N | N | N |
| 8 | N | N | N | N | N | S |
| 9 | N | N | N | N | N | S |
| 10 | N | N | N | N | N | N |

**Tabela 2.7 - descoberta de regras de associação com fs = 0.3 e fc = 0.8**

| |
|---|
| Conjunto de Itens Frequentes: Corte_Ligação, Fraude. Fs = 3/10 = 0.3<br>Regra: SE (Corte_Ligação) ENTÃO (Fraude). Fc = 3/3 = 1. |
| Conjunto de Itens Frequentes: Corte_Ligação, Multa. Fs = 3/10 = 0.3<br>Regra: SE (Corte_Ligação) ENTÃO (Multa). Fc = 3/3 = 1. |
| Conjunto de Itens Frequentes: Fraude, Multa. Fs = 4/10 = 0.4<br>Regra: SE (Fraude) Então (Multa). Fc = 4/5 = 0.8.<br>Regra: SE (Multa) ENTÃO (Fraude). Fc = 4/5 = 0.8 |
| Conjunto de Itens Frequentes: Corte_Ligação, Fraude, Multa. Fs = 3/10 = 0.3<br>Regra: SE (Corte_Ligação E Fraude) ENTÃO (Multa). Fc = 3/3 = 1.<br>Regra: SE (Corte_Ligação E Multa) ENTÃO (Fraude). Fc = 3/3 = 1<br>Regra: SE (Corte_Ligação) ENTÃO (Fraude E Multa). Fc = 3/3 = 1 |

Os atributos “Consumidor Comercial”, “Parcelamento Fatura” e “Pagamento Fatura” possuem Fs = 0.2 e não pertencem ao Conjunto de Itens Frequentes, visto que o Fs deve ser maior ou igual a 0.3. Já os atributos “Fraude”, “Corte Ligação” e “Multa” possuem Fs igual a 0.5, 0.3 e 0.5 respectivamente e desta forma, pertencem ao Conjunto de Itens Frequentes.

A Tabela 2.7 demonstra as regras de associação que são descobertas dos dados oriundos da Tabela 2.6 utilizando-se os valores de Fs e Fc maiores ou iguais aos especificados pelo analista, que foram respectivamente 0.3 e 0.8. Na Tabela 2.7 as regras de associação são agrupadas por três conjuntos de {Itens Frequentes}, sendo dois deles formados pelos itens

frequentes {Corte_Ligação e Fraude} e {Corte_Ligação e Multa}, e o terceiro conjunto formado pelos itens frequentes {Corte_Ligação, Fraude e Multa}.

### 2.5.5 Técnicas de Data Mining

Segundo afirma (BALLARD, et al., 1998), não há uma técnica que resolva todos os problemas de *Data Mining*. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e desvantagens, por isso, é importante conhecer bem o ambiente de aplicação e as técnicas disponíveis para que se possa escolher a mais adequada.

Dentre as técnicas de DM normalmente utilizadas tem-se: Árvores de Decisão; Regras de Associação; Redes Neurais Artificiais; Algoritmos Genéticos e Classificação Bayesiana. O item 3.3 discute a aplicação do algoritmo de mineração que se mostrou mais adequado para escopo do problema abordado neste trabalho.

A Tabela 2.8 apresenta a descrição das principais técnicas e tarefas de DM, e cita alguns algoritmos relacionados com as respectivas técnicas. Nas seções seguintes serão discutidas as técnicas utilizadas neste trabalho, que foram: Árvore de Decisão, Classificação Bayesiana e Regras de Associação.

**Tabela 2.8 - técnicas, tarefas e algoritmos de *data mining***

| Técnica | Descrição | Tarefas | Algoritmos |
|---|---|---|---|
| Árvore de Decisão | Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos. | Classificação Regressão | J4.8, One-R, ID-3, CART, CHAID, C4.5, C5.0, SPRINT, etc. |
| Classificação Bayesiana | Métodos estatísticos que podem predizer a probabilidade de um registro pertencer a uma determinada classe. | Classificação | NaïveBayes |
| Regras de Associação | Estabelece uma correlação estatística entre os atributos de dados e conjunto de dados. | Associação | Apriori, AprioriTid, AprioriHybrid, AIS, SETM e DHP, etc. |
| Redes Neurais Artificiais | Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões. | Classificação Segmentação Regressão | Perceptron, Rede MLP, Redes ART, Rede IAC, Rede BSB, etc. |
| Algoritmos Genéticos | Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes". | Classificação Segmentação | Algoritmo Genético Simples, Genitor, CHC, Algoritmo de Hillis, GANuggets, etc. |

**Fonte: (FAYYAD, et al., 1996)**

### 2.5.5.1 Árvores de Decisão

As árvores de decisão são uma maneira de representar uma série de regras que conduzem a uma classe ou a um valor. De acordo com (SYMEONIDIS, et al., 2005), o objetivo principal de uma árvore de decisão é separar as classes, onde as tuplas de classes diferentes tendem a ser alocadas em subconjuntos diferentes, cada um descrito por regra simples em um ou mais itens de dados. Essas regras podem ser expressas como declarações lógicas, em uma linguagem como SQL, de modo que possam ser aplicadas diretamente a novas tuplas.

Uma das principais vantagens das árvores de decisão é o fato de que o modelo é bem explicável, uma vez que tem a forma de regras explícitas, podendo ser representada como um conjunto de regras (galhos), onde cada nó não terminal representa um teste ou decisão sobre o item considerado.

Na árvore de decisão cada nó não terminal representa um teste ou decisão sobre o item de dado. Assim, os nós representam os atributos, as ligações entre os nós representam os valores dos atributos e as folhas representam as classes. Cada caminho da árvore pode ser convertido numa regra. O nó interno e os valores das setas são convertidos no antecedente da regra (parte SE); o nó folha é convertido no consequente da regra (parte ENTÃO).

Um exemplo poderia ser a classificação de consumidores na categoria "Aceitável" ou "Risco", onde a primeira indica confiança na adimplência do consumidor e a segunda indica risco de inadimplência perante a companhia de abastecimento de água. A Figura 2.15 ilustra uma árvore simplificada de decisão que resolve esta situação.

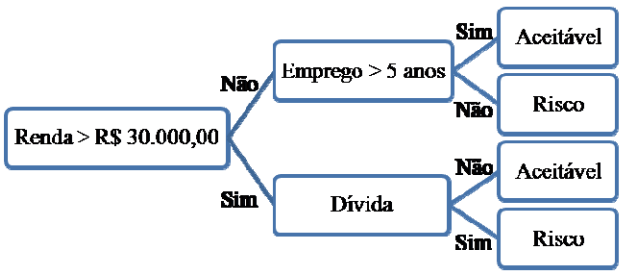



**Figura 2.15 - exemplo de árvore de decisão**

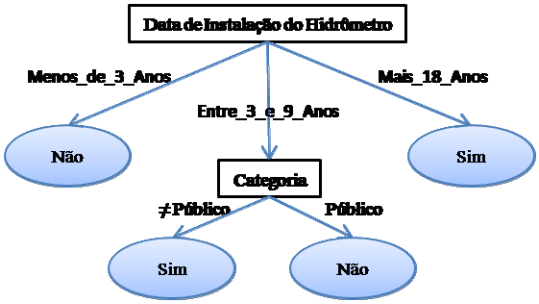



**Figura 2.16 - arvore de decisão gerada com os dados da Figura 2.14**

Na Figura 2.16 tem-se uma árvore de decisão para o exemplo da Figura 2.14. Cada caminho da árvore pode ser convertido numa regra. A tupla <"João", "Entre_3_e_9_Anos",

"Público", ?> identifica um tipo de consumidor. A interrogação (?) representa o valor do atributo objetivo (Estado de Inadimplência), e este é responsável por informar se o consumidor está inadimplente (Sim) ou adimplente (Não) perante a companhia de abastecimento de água.

O nó raiz da árvore de decisão da Figura 2.16 representa o atributo "Data de Instalação do Hidrômetro". Na tupla dada como exemplo, o nó raiz direciona a regra para data de instalação do hidrômetro "Entre_3_e_9_Anos". Seguindo a hierarquia da árvore, a regra passa pelo seu segundo nó, que é o atributo "Categoria", e o seu valor na tupla é "Público". Por fim o algoritmo direciona a regra para o atributo objetivo, que por sua vez, está no nó folha rotulado pelo valor "Não", que indica que João está na classe dos consumidores adimplente.

Geralmente uma árvore de decisão classifica uma nova tupla de maneira *top-down*, utilizando um algoritmo baseado na aproximação "dividir para conquistar" (WITTEN, et al., 2005). Inicialmente todas as tuplas que estão sendo mineradas são associadas ao nó raiz da árvore. Então o algoritmo seleciona uma partição de atributos e divide o conjunto de tuplas no nó raiz de acordo com o valor do atributo selecionado. O objetivo desse processo é separar as classes para que tuplas de classes distintas tendam a ser associadas a diferentes partições. Esse processo é recursivamente aplicado a subconjuntos de tuplas criados pelas partições, produzindo subconjuntos de dados cada vez menores, até que um critério de parada seja satisfeito. A fim de minimizar o tamanho da árvore, sem prejudicar a qualidade da solução, aplica-se algoritmo de poda de árvore de decisão.

As principais vantagens de algoritmos baseados em árvores de decisão são sua eficiência computacional e simplicidade. Devido ao uso da aproximação "dividir para conquistar", entretanto essa aproximação também possui desvantagem. Por exemplo, uma condição envolvendo um atributo que será incluído em todas as regras descobertas. Essa situação possivelmente produz regras com informações irrelevantes, além de desperdício de processamento. Três algoritmos de árvore de decisão serão analisados, são eles: Algoritmo de Indução de Regras, Algoritmo ID-3 e Algoritmo J4.8.

Outro exemplo simples de como funciona um algoritmo de classificação, que apresenta seu resultado sob a forma de árvore de decisão, está ilustrado na Figura 2.17. Neste exemplo, os consumidores podem ser classificados em confiáveis ou não confiáveis junto à

companhia de abastecimento, baseando-se na quantidade de pontos de utilização e o valor da conta (fatura a ser paga).

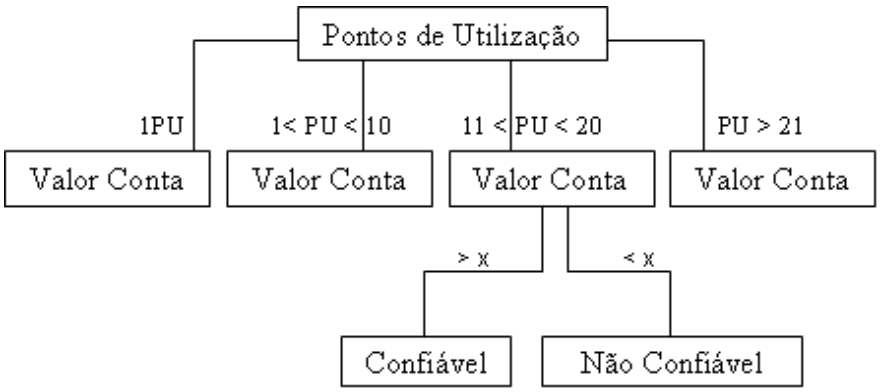



**Figura 2.17 - classificação por árvore de decisão (pontos de utilização *versus* fatura)**

Cada regra tem seu início na raiz da árvore e caminha até suas folhas. A interpretação de um dos galhos é que uma pessoa que possua entre 11 a 20 Pontos de Utilização (PU) de água e o Valor da Conta (VC) maior que "x" (valor em real), onde x representa um intervalo de consumos e contas aceitáveis para uma determinada categoria do sistema. Por exemplo: um consumidor que possui 15 pontos de utilização, pertence à categoria residencial cujo "x" está definido entre [R$ 50,00 e R$ 200,00], e pagou pela conta um valor menor do que R$ 50,00, então ele passará a ser classificado com não confiável. Tal regra extraída da base de dados permite ao gerente tomar a decisão de realizar uma intervenção/vistoria no abastecimento dos consumidores com baixo grau de confiança.

Existem alguns algoritmos de classificação que, ao invés de montarem uma árvore de decisão, expressam o conhecimento extraído através de regras do tipo "SE condição ENTÃO classe" ou X, Y → Z, chamadas simplesmente de Regras de Classificação. Cada galho de uma árvore de classificação representa uma regra. A seguir, são apresentadas as regras da Figura 2.17, sob a forma de regras de classificação:

Se (Pontos de Utilização corresponde ao intervalo [11..20]), (VC < x)) → Não Confiável

Se (Pontos de Utilização corresponde ao intervalo [11..20]), (VC > x)) → Confiável

Os ramos da árvore podem crescer de maneiras diferentes. Por exemplo, caso não exista consumidor com apenas 1 ponto de utilização de água (1° regra mais a esquerda) em todo o setor da rede de distribuição de água, então a regra nunca será aplicada e o ramo ficará estático. Além disso, todas as regras geradas a partir de uma AD terão que conter o atributo

raiz em seu antecedente. No exemplo da Figura 2.17, como "Pontos de Utilização" é o atributo raiz escolhido, não há como se ter uma regra do tipo: Se (" ", (x > 300)) → Confiável.

***a) Algoritmo de Indução de Regras***

Este tipo de algoritmo é baseado em duas idéias chaves: Estado e Operador. Um Estado é a descrição da situação de um problema num dado instante e um Operador é um procedimento que transforma um estado em outro. Resolver um problema utilizando esse algoritmo consiste em encontrar uma sequência de operadores dos quais transformam um estado inicial num estado objetivo, ou estado meta. Um estado corresponde a uma regra candidata e os operadores correspondem a operações de generalização e/ou especialização que transformam uma regra candidata em outra (LAROSE, 2005).

A principal vantagem desse algoritmo é que geralmente ele produz conhecimento compreensível e o conhecimento descoberto está na forma de regras "SE→ENTÃO", desse modo as regras podem ser facilmente entendidas e validadas pelo usuário.

Um exemplo da utilização do algoritmo de indução de regras se encontra na Tabela 2.9. Este exemplo remete-se aos dados da Figura 2.14, e que também foram utilizados na tarefa de classificação.

**Tabela 2.9 - operações de especialização e generalização por indução de regras**

| Especializando uma regra pela adição da conjunção em seu antecedente |
|---|
| Regra Original:<br>SE (Instalação_Hid = Menos_de_3_Anos) ENTÃO Inadimplência = Não<br>Regra Especializada:<br>SE (Instalação_Hid = Menos_de_3_Anos e Categoria = Comercial) ENTÃO Inadimplência = Não |
| Generalizando uma regra relaxando uma condição no antecedente |
| Regra original:<br>SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria = Comercial) ENTÃO Inadimplência = Sim<br>Regra Generalizada:<br>SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria != Público) ENTÃO Inadimplência = Sim |

A operação de especialização mostra que a regra pode ser especializada pela adição de novas condições ao antecedente. Note que a nova regra é uma especialização da original porque o antecedente da regra é satisfeito por um número menor de tuplas no banco de dados. A regra original atende 2 registros (Bruno e Carlos), enquanto a regra especializada, com a conjunção "Categoria = Público", atende apenas 1 registro (Bruno).

Na generalização a idéia é estender o intervalo de valores cobertos pelo atributo "Categoria", relaxando-o de modo que o antecedente da regra satisfaça um número maior de tuplas na base de dados. No exemplo dado, os registros atendidos passou de 1 (Fernanda) para 3 (Fernanda, Sérgio e Pedro).

***b) Algoritmo ID-3***

O *Iterative Dichotomiser* (ID-3) proposto por J. Ross Quinlan é um dos mais conhecidos algoritmos destinados a construção de árvore de decisão para a tarefa de classificação. O algoritmo cria a árvore de decisão a partir dos exemplos de treinamento utilizando o método de indução *top-down induction of decision trees* (TDIDT). A evolução do ID-3 são os algoritmos ID-4, ID-6, C4.5 e C5.0 (QUINLAN, 1993). No tópico seguinte se dará uma maior atenção ao algoritmo C4.5[12], visto que é um dos algoritmos utilizados como técnica de *Data Mining* proposta por este trabalho.

Ele recebe como entrada um conjunto de tuplas para treinamento, chamado exemplos; um atributo objetivo, chamado meta; e um conjunto de atributos preditivos, chamado atributos. Não é considerado um algoritmo incremental, pois todos os exemplos de treinamento devem estar disponíveis no início do processo.

Para a geração da árvore, são utilizados exemplos de treinamento rotulados que possuem os atributos e a classe a que pertence. Aplica-se uma função de avaliação para cada atributo verificando aquele que discrimina melhor os conceitos positivos dos negativos, deixando na raiz da árvore de decisão o atributo mais informativo.

O processo é recursivo, gerando sub-árvores até que se atenda um critério de parada, que no caso ideal seria obter nós contendo apenas exemplos de uma mesma classe (KANASHIRO, 2007). De uma maneira geral, os passos do algoritmo ID-3 são apresentados na Tabela 2.10.

**Tabela 2.10 - passos para construção da árvore de decisão através do ID-3**

| |
|---|
| 1. Dado um NÓ na árvore e todas as tuplas do conjunto de treinamento S;<br>2. Selecione o melhor atributo A para esse nó;<br>3. Para cada valor $v_i$ de A, cresça uma subárvore, ou uma folha, sob o nó. |

[12] É similar ao algoritmo J4.8. A diferença é que o C4.5 foi desenvolvido na linguagem C e o J4.8 em Java.

***c) Algoritmo J4.8***

É um algoritmo baseado na implementação do algoritmo *C4.5 release 8*, e este por sua vez é uma evolução do algoritmo *ID-3*, ambos foram desenvolvidos por Ross Quinlan (QUINLAN, 1993). A versão mais recente desta classe de algoritmos é C5.0, contudo, este algoritmo não será discutido neste trabalho por se tratar de uma implementação proprietária e que é disponibilizada apenas comercialmente.

O algoritmos *J4.8* surgiu da necessidade de recodificar o algoritmo C4.5, que originalmente é escrito na linguagem C, para a linguagem Java. (WITTEN, et al., 2005). Ele tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste.

Um dos aspectos para a grande utilização do algoritmo *J4.8* pelos especialista em *Data Mining* é que o mesmo mostra-se adequado para os procedimentos envolvendo as variáveis (dados) qualitativas e variáveis quantitativas contínuas e discretas presentes nas Bases de Dados.

### 2.5.5.2 Classificação Bayesiana

Os classificadores Bayesianos são classificadores estatísticos que podem predizer a probabilidade de um registro pertencer a uma determinada classe. Eles oferecem uma simples, porém poderosa técnica de classificação supervisionada assumindo que todos os atributos de entrada possuem a mesma importância e são independentes entre si.

***a) Algoritmo NaïveBayes***

A classificação Bayesiana é baseada no teorema de *Bayes* e no algoritmo de classificação, conhecido como *NaïveBayes* (WEISS, et al., 1991).

O princípio básico desse método está fundamentado na teoria da Probabilidade Bayesiana (SHEN, 1993), como mostra a Equação 2.1.

$$P(AB|C) = P(A|C)P(B|AC) = P(B|C)P(A|BC)$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

**Equação 2.1 - Formulação da Probabilidade Bayesiana**

Onde P refere-se a probabilidade de um evento; A, B, e C são subconjuntos do espaço de amostras do problema, isto é, a base de dados; e a notação $P(AB|C)$ significa a probabilidade dos eventos A e B acontecerem dado que o evento C acontece. De modo análogo $P(A|C)$ significa a probabilidade do evento A acontecer dado que C acontece.

A Tabela 2.11 mostra um exemplo de uma base de dados que servirá como entrada para extração de conhecimento utilizando classificadores bayesianos.

**Tabela 2.11 - exemplo de dados para classificação bayesiana**

| ID | Categoria | Medição | Pagamento | Consumo | Inspecionar |
|---|---|---|---|---|---|
| 1 | Residencial | Regular | Inadimplente | Normal | Não |
| 2 | Residencial | Regular | Inadimplente | Abaixo | Não |
| 3 | Comercial | Regular | Inadimplente | Normal | Sim |
| 4 | Pública | Ausente | Inadimplente | Normal | Sim |
| 5 | Pública | Suspensa | Adimplente | Normal | Sim |
| 6 | Pública | Suspensa | Adimplente | Abaixo | Não |
| 7 | Comercial | Suspensa | Adimplente | Abaixo | Sim |
| 8 | Residencial | Ausente | Inadimplente | Normal | Não |
| 9 | Residencial | Suspensa | Adimplente | Normal | Sim |
| 10 | Pública | Ausente | Adimplente | Normal | Sim |
| 12 | Residencial | Ausente | Adimplente | Abaixo | Sim |
| 12 | Comercial | Ausente | Inadimplente | Abaixo | Sim |
| 13 | Comercial | Regular | Adimplente | Normal | Sim |
| 14 | Pública | Ausente | Inadimplente | Abaixo | Não |

Visando exemplificar o uso dos classificadores Bayesianos pode-se constatar a probabilidade de um acontecimento (verificação/inspeção do imóvel por um técnico) vir a acontecer com base nos atributos da base de dados apresentados na Tabela 2.11. Segue duas perguntas que poderiam ser feitas:

$P_1$: Qual a probabilidade de *inspecionar* o imóvel dado que a categoria é residencial, a medição é regular, o pagamento está atrasado (inadimplente) e o consumo está normal?

Em termos probabilísticos essa pergunta equivalente a $P_1$(INSPECIONAR = Sim | [Residencial, Regular, Inadimplente, Normal]).

$P_2$: Qual a probabilidade de *não inspecionar* o imóvel dado que a categoria é residencial, a medição é regular, o pagamento está atrasado (inadimplente) e o consumo está normal?

Em termos probabilísticos essa pergunta equivale a $P_2$(INSPECIONAR = Não | [Residencial, Regular, Inadimplente, Normal]).

**Tabela 2.12 - cálculo das probabilidades dos dados da Tabela 2.11 utilizando classificadores bayesianos**

| Formulação da Pergunta | Probabilidade |
|---|---|
| P(**INSPECIONAR = Sim**) | 9/14 |
| P(**INSPECIONAR = Não**) | 5/14 |
| P(**CATEGORIA = Residencial**\|**INSPECIONAR = Sim**) | 2/9 |
| P(**CATEGORIA = Residencial**\|**INSPECIONAR = Não**) | 3/5 |
| P(**MEDIÇÃO = Regular**\|**INSPECIONAR = Sim**) | 2/9 |
| P(**MEDIÇÃO = Regular**\|**INSPECIONAR = Não**) | 2/5 |
| P(**PAGAMENTO = Inadimplente**\|**INSPECIONAR = Sim**) | 3/9 |
| P(**PAGAMENTO = Inadimplente**\|**INSPECIONAR = Não**) | 4/5 |
| P(**CONSUMO = Normal**\|**INSPECIONAR = Sim**) | 6/9 |
| P(**CONSUMO = Normal**\|**INSPECIONAR = Não**) | 2/5 |

Uma vez calculadas as probabilidades de cada termo, pode-se estimar o P1 e P2 por meios da Equação 2.1 (definida na página 62). Desta forma, encontra-se:

$P_1$(INSPECIONAR = Sim | [Residencial, Regular, Inadimplente, Normal]) =

$$\frac{[(2/9)\ (2/9) * (3/9) * (6/9)] * (9/14)}{[(5/14) * (4/14) * (7/14) * (8/14)]} = \frac{(108/6561) * (9/14)}{(1120/38416)} = 0{,}3630 = 36{,}30\%$$

$P_2$(INSPECIONAR = Não | [Residencial, Regular, Inadimplente, Normal]) =

$$\frac{[(3/5) * (2/5) * (4/5) * (2/5)] * (5/14)}{[(5/14) * (4/14) * (7/14) * (8/14)]} = \frac{(48/625) * (5/14)}{(1120/38416)} = 0{,}9498 = 94{,}98\%$$

Logo, a classificação Bayesiana estimou 36,30% de probabilidade de acontecimento aplicada ao $P_1$. Enquanto para $P_2$ resultou em 94,98% de probabilidade, o que significa que os consumidores residenciais com a medição calculada corretamente e apresentando uma média de consumo em relação aos últimos meses, não necessitam de inspeção técnica, mesmo possuindo histórico de inadimplência perante a companhia.

### 2.5.5.3 Regras de Associação

De acordo com (AGRAWAL, et al., 1996), a descrição formal do problema de mineração envolvendo regras de associação é dado a seguir:

Sejam $I = \{i_1, i_2, i_3, ..., i_n,\}$ um conjunto de n itens distintos e D uma base de dados formada por um conjunto de transações, onde cada transação T é composta por um conjunto de itens, chamado *itemset*, tal que $T \subseteq I$. Uma regra de associação é uma expressão na forma $X \Rightarrow Y$, onde $X \subseteq I$, $Y \subseteq I$, $X \neq \emptyset$, $Y \neq \emptyset$, $X \cap Y \neq \emptyset$. X é denominado antecedente e Y denominado consequente da regra. Tanto o antecedente, quanto o consequente de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens. A quantidade de itens pertencentes a um conjunto de itens é chamada de comprimento do conjunto. Um conjunto de itens de comprimento k costuma ser referenciado como um *k-itemset*.

A regra $X \Rightarrow Y$ é válida no conjunto de transações D com grau de confiança "c", se c% das transações em D que contêm X também contêm Y. E a regra $X \Rightarrow Y$ tem suporte "s" em D, se s% das transações em D contêm $X \cup Y$.

***a) Algoritmo Apriori***

O algoritmo *Apriori* é considerado um clássico na extração de Regras de Associação e foi concebido pelo centro de pesquisa da IBM − "*The Quest Data Mining System, IBM Almaden Research Center*". Esse algoritmo emprega busca em profundidade e utiliza os conjuntos de itens de tamanho k para gerar os conjuntos de itens de tamanho (k + 1). O primeiro passo do algoritmo é encontrar os conjuntos de itens frequentes com 1 item. Este conjunto é denominado de $L_1$. O conjunto de $L_1$ é usado para gerar $L_2$, que representa os conjuntos de itens frequentes com 2 itens, e assim por diante, até que nenhum conjunto de itens frequentes possa ser gerado (NONG, 2003 p. 28).

O algoritmo *Apriori* usa o principio de que cada subconjunto de um conjunto de itens frequentes também deve ser frequente. Esta regra é utilizada para reduzir o número de candidatos a serem comparados com cada transação no banco de dados. Todos os candidatos gerados que contém algum subconjunto que não seja frequente são eliminados, ou utilizando a terminologia do algoritmo, podados.

Cada passo inicia com um conjunto semente de itens, e esse conjunto semente gerará novos conjuntos potenciais, chamados conjunto de itens candidatos. Enquanto o conjunto de itens candidatos não ficar vazio, o algoritmo armazena esses conjuntos e para cada tupla do banco de dados testa se um conjunto candidato está ou não contido na tupla. Caso um conjunto candidato esteja contido na tupla, então incrementa um contador. Se ao final do teste

para cada tupla da base de dados uma regra candidata tiver um suporte mínimo especificado, então ela é inserida no novo conjunto semente, que são os itens candidatos.

Suponha um banco de dados cujo conjunto de itens I = {a, b, c, d, e} e um conjunto de transações T = {1, 2, 3, 4, 5, 6}, conforme mostra Tabela 2.13. O objetivo do algoritmo Apriori é determinar os *itemsets* com MinSup igual a 50%, ou seja, que ocorram em pelo menos três transações, haja vista que 50% de 6 transações corresponde a 3 transações.

**Tabela 2.13 - exemplo de uso do algoritmo *apriori***

| Itens da Base de Dados | | | | | | |
|---|---|---|---|---|---|---|
| T | 1 | 2 | 3 | 4 | 5 | 6 |
| Itens | abde | bce | abde | abce | abcde | bcd |

O algoritmo prossegue sua execução, conforme mostra a Figura 2.14.

**Tabela 2.14 - passos da execução do algoritmo *apriori***

| $C_1 = L_1$ | | |
|---|---|---|
| Itemset | Suporte | |
| a | 67% | 4/6 |
| b | 100% | 6/6 |
| c | 67% | 4/6 |
| d | 67% | 4/6 |
| e | 83% | 5/6 |

| $C_2$ | | |
|---|---|---|
| ab | 67% | 4/6 |
| ac | 33% | 2/6 |
| ad | 50% | 3/6 |
| ae | 67% | 4/6 |
| bc | 67% | 4/6 |
| bd | 67% | 4/6 |
| be | 83% | 5/6 |
| cd | 33% | 2/6 |
| ce | 50% | 3/6 |
| de | 50% | 3/6 |

| $L_2$ | | |
|---|---|---|
| ab | 67% | 4/6 |
| ad | 50% | 3/6 |
| ae | 67% | 4/6 |
| bc | 67% | 4/6 |
| bd | 67% | 4/6 |
| be | 83% | 5/6 |
| ce | 50% | 3/6 |
| de | 50% | 3/6 |

| $C_3$ | | |
|---|---|---|
| abc | 17% | 1/6 |
| abd | 50% | 3/6 |
| abe | 67% | 4/6 |
| acb | 33% | 2/6 |
| acd | 17% | 1/6 |
| ace | 33% | 2/6 |
| ade | 67% | 4/6 |
| bce | 50% | 3/6 |
| bde | 50% | 3/6 |
| cde | 17% | 1/6 |

| $L_3$ | | |
|---|---|---|
| abd | 50% | 3/6 |
| abe | 67% | 4/6 |
| ade | 50% | 3/6 |
| bce | 50% | 3/6 |
| bde | 50% | 3/6 |

| $C_4$ | | |
|---|---|---|
| abcd | 17% | 1/6 |
| abce | 33% | 2/6 |
| acde | 17% | 1/6 |
| abde | 50% | 3/6 |
| bcde | 17% | 1/6 |

| $L_4$ | | |
|---|---|---|
| abde | 50% | 3/6 |

| RESULTADO – Algoritmo Apriori | | |
|---|---|---|
| Suporte | | itemset |
| 100% | 6/6 | b |
| 83% | 5/6 | e, be |
| 67% | 4/6 | a, c, d, ab, ae, bc, bd, abe. |
| 50% | 3/6 | ad, ce, de, abd, ade, bce, bde, abde. |

O algoritmo gera os conjuntos candidatos ($C_1$, $C_2$, $C_3$ e $C_4$) e a partir destes descobrindo os *itemsets* frequentes ($L_1$, $L_2$, $L_3$ e $L_4$) com suporte mínimo de 50%, como mostra a tabela RESULTADO. O conjunto candidato (Cn) é formado por todas as combinações do *itemset*. O

conjunto do "*itemset freqüente*" ($L_n$) é formado pelos valores do Cn que possuem suporte mínimo.

### 2.5.6 Visão Hierárquica do KDD

Segundo (CARVALHO, 2001), muitas vezes os termos *Data Mining* e "Descoberta de Conhecimento em Banco de Dados - KDD" são confundidos como sinônimos. Porém, é importante frisar que o termo KDD é empregado para descrever todo o processo de extração de conhecimento de um conjunto de dados, enquanto que o termo *Data Mining* refere-se a uma das etapas deste processo. Na etapa de *Data Mining* é possível descobrir informações escondidas do usuário, diferentemente das tradicionais consultas que utilizam a *Structured Query Language* (SQL), nas quais resultam apenas de dados e relatórios pré-determinados.

A Figura 2.18 retoma de forma hierárquica e sistemática uma visão das principais fases, tarefas e algoritmos do KDD discutidos nas seções 2.5.4 e 2.5.5.

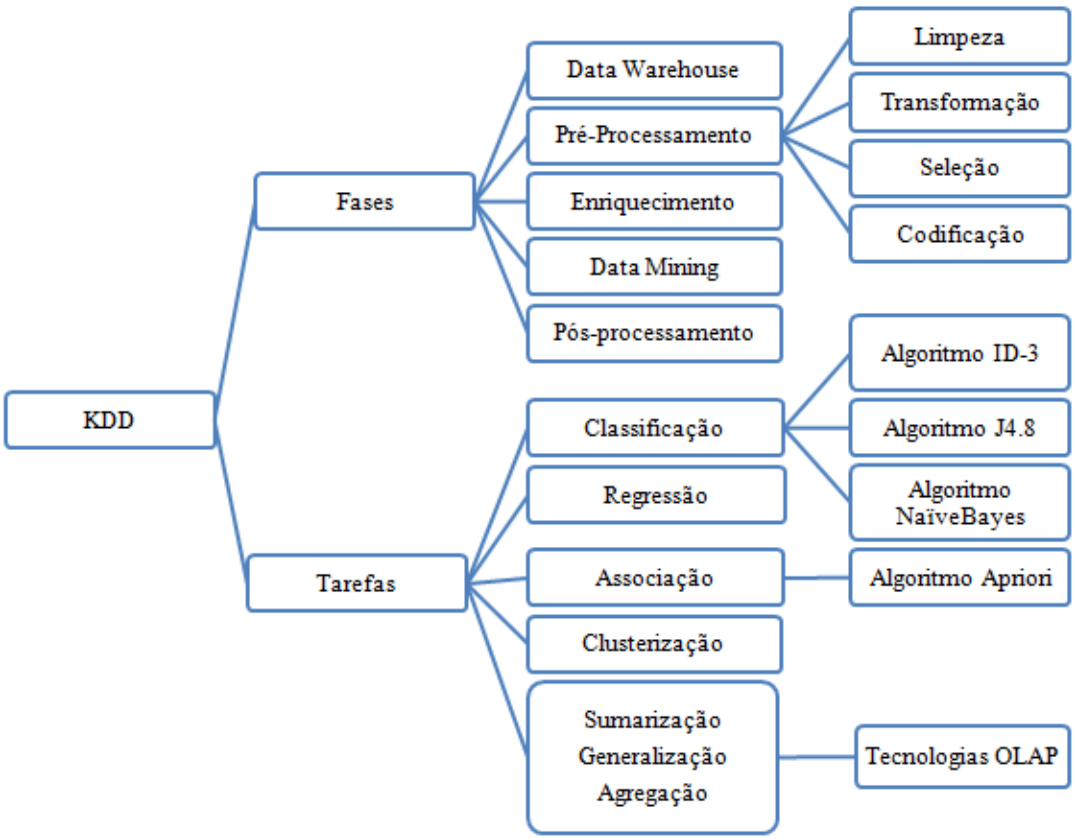



**Figura 2.18 - taxonomia do processo de descoberta do conhecimento em banco de dados**

O KDD está relacionado com diversos domínios de aplicações: marketing, análises corporativas, astronomia, medicina, biologia, engenharia, entre outros. Deste modo, identificam-se diversas tarefas de KDD que são, principalmente, dependentes do domínio da

aplicação e do interesse do analista. Cada tarefa de KDD extrai um tipo diferente de conhecimento do banco de dados, logo cada tarefa requer um algoritmo diferente para a extração de conhecimento (SANCHES, 2003).

### 2.5.7 Ferramentas de Data Mining

Existem várias ferramentas disponíveis no mercado que implementam características de ambientes integrados de KDD, afim de facilitar a execução das etapas e a minimização das dificuldades operacionais de suporte a decisão. A Tabela 2.15 apresenta algumas das ferramentas mais populares, as tarefas de *Data Mining* que elas implementam e a empresa que desenvolveu (QUITÉRIO, et al.; PEREIRA, 2002).

**Tabela 2.15 - ferramentas de *data mining* - apoio à KDD**

| Ferramenta | Tarefas Implementadas | Desenvolvedor |
|---|---|---|
| WEKA | Classificação, Regressão e Regras de Associação | University of Waikato www.cs.waikato.ac.nz |
| *Intelligent Miner* | Classificação, Regras de Associação, Clusterização e Sumarização | IBM Corp. www.ibm.com |
| *Oracle Data Mining* | Classificação, Regressão, Associação, Clusterização e Mineração de Textos | Oracle www.oracle.com |
| *Enterprise Miner* | Classificação, Regras de Associação, Regressão e Sumarização | SAS Inc. www.sas.com |
| *Clementine* | Classificação, Regras de Associação, Clusterização, Sequência e Detecção de Desvios | SPSS Inc. www.spss.com |
| *Darwin* | Classificação | Thinking Machines |
| *Business Objects* | Classificação, Regras de Associação, Clusterização e Sumarização | Business Objects$^{R}$ www.businessobjects.com |
| *Microsoft Data Analyser* | Classificação e Clustering | Microsoft Corp. www.Microsoft.com |

**Fonte: Adaptação (REZENDE, 2003).**

Algumas das ferramentas apresentadas na Tabela 2.15 foram comparadas entre si quanto ao âmbito da aplicação, formato dos dados input/output, algoritmos suportados, visualização dos dados e características positivas e negativas, foram selecionadas cinco ferramentas: WEKA, *Business Objects*, *Enterprise Miner, Darwin e Intelligent Miner*.

Quanto ao âmbito da aplicação, as ferramentas *Business Objects*, *Enterprise Miner* e *Intelligent Miner* dão suporte a OLAP e *Data Mining* para todos os tipos de aplicações, apesar

do suporte ao DM pela ferramenta *Business Objects* ser bastante limitado. As ferramentas WEKA e *Darwin* suportam apenas *Data Mining*.

Quanto ao tipo e/ou formatos de dados de entrada e modelos de saída, as cinco ferramentas possuem como opções de input de dados o padrão ODBC/JDBC e *drivers* da base de dados nativa. No que diz respeito ao output de modelos, as cinco ferramentas mencionadas fornecem relatórios e apenas a ferramenta *Enterprise Miner* não fornece código fonte.

Com relação aos algoritmos de aprendizagem suportados, as cinco ferramentas implementam Árvore de Decisão. Nenhuma delas implementam Algoritmo Genético e apenas *Business Objects* não implementa Redes Neurais. As ferramentas WEKA, *Enterprise Miner e Darwin* também suportam algoritmos de Indução de Regras.

De acordo com a capacidade de visualização, i.e., os modelos visuais fornecidos, as cinco ferramentas dispõem de histogramas e gráficos de linhas. E apenas *Business Objects* fornece "regiões de decisão e classificação".

No que diz respeito à usabilidade, a ferramenta *Enterprise Miner*, *Darwin* e *Intelligent Miner* são consideradas satisfatórias quanto ao "carregamento e manipulação de dados", "construção de modelos", "compreensão de modelos" e "suporte técnico". Já as ferramentas WEKA e *Business Objects* são classificadas como boas quanto aos itens mencionados, menos com relação ao item "suporte técnico".

O resumo da avaliação comparativa, pontos fortes e fracos, das ferramentas WEKA, *Business Objects, Enterprise Miner, Darwin e Intelligent Miner* é apresentado na Tabela 2.16.

**Tabela 2.16 - avaliação comparativa entre as ferramentas de *data mining***

| Ferramenta | Pontos Fortes | Pontos Fracos |
|---|---|---|
| WEKA | Interface gráfica amigável e principais algoritmos de mineração implementados. | Alguns parâmetros só podem ser executados via linha de comando (script) |
| *Business Objects* | Usabilidade e escalabilidade; | Algoritmo fraco para *Data Mining*; |
| *Enterprise Miner* | Quantidade de algoritmos e interface visual; | Difícil de usar; |
| *Darwin* | Eficiência e interface com usuário; | Visualização limitada; |
| *Intelligent Miner* | Quantidade de algoritmos, output gráfico e volume de dados tratáveis; | Falta de flexibilidade dos algoritmos e pouca automação; |

A ferramenta de *Data Mining* selecionada para o estudo de caso e experimentação do trabalho foi WEKA. Ela faz parte do ambiente *Pentaho Open BI Suite* e será apresentada com maiores detalhes no capítulo 4, item 4.1.1. Um dos motivos para escolha desta ferramenta se deu pelo fato dela ser livre, i.e., de código aberto, alem de ser consideradamente indicada na literatura acadêmica.

Outro benefício da ferramenta WEKA é que ela possibilita a descoberta de padrões de comportamento e conhecimento dos dados através de vários algoritmos de exploração de dados. O WEKA oferece uma interface intuitiva e uniforme para os algoritmos implementados (associação, classificação, regressão e agrupamento), além de fornecer visualização gráfica dos dados minerados.

### 2.5.8 Relação entre Data Warehouse, OLAP e Data Mining

Os sistemas de apoio à decisão ou *Decision Support System* (DSS) agregam importante diferencial competitivo nas organizações, ajudando-as na tomada de decisão. A implantação dos DSS ocorre principalmente pelo uso de ferramentas *On-line Analytical Processing* (OLAP) e *Data Mining*, que por sua vez fazem acesso aos dados do *Data Warehouse.*

De acordo com (SANCHES, 2003), existe uma relação simbólica entre a atividade de *Data Mining* e *Data Warehouse*. Os DW organizam os dados para um efetivo processo de mineração, porém, a exploração de dados através da mineração pode ser aplicada onde não exista nenhum DW. O uso do DW aumenta significativamente as chances de sucesso do *Data Mining,* visto que o DW dispõe de dados integrados; dados detalhados e resumidos; dados históricos e metadados. A utilização desses tipos de dados melhora o desempenho e o resultado do processo de mineração.

Dados integrados permitem ao analista, que é o agente minerador, visualizar de forma rápida e fácil os dados. Desta forma, o agente minerador pode concentrar-se integralmente nos algoritmos de mineração. Na ausência de integração entre os dados, o agente minerador necessitaria de uma quantidade de tempo maior para condicionar e refinar os dados antes do processo de mineração.

Dados detalhados são necessários quando o agente minerador deseja examinar os dados de forma mais granular. Algumas vezes o nível de exploração dos dados requer a

análise cuidadosa destes dados, mas geralmente os dados resumidos asseguram que uma prévia já foi feita e evitam muito processamento desnecessário e repetitivo. Dados históricos são importantes porque grande quantidade de informações fica implicitamente armazenada. Trabalhar somente com informações atuais pode impedir que se detectem tendências e padrões de comportamento ao longo do tempo. Informações históricas são necessárias para o entendimento das circunstâncias dos negócios.

Quanto aos metadados, eles ajudam a descrever não só o conteúdo dos dados, mas o contexto das informações. À medida que a informação passa a ser examinada, o contexto passa a ser mais importante do que o conteúdo, revelando explicações a respeito do significado dos dados. Desta forma, esta importante relação entre *Data Mining* e *Data Warehouse*, quando utilizados em conjunto, maximizam os resultados do processo de *Data Mining*.

Segundo (KIMBALL, 1997), enquanto OLAP é dedutivo e guiado por especialistas, *Data Mining* é indutivo e guiado pelos próprios dados. Ambas necessitam de dados limpos e consistentes. E neste caso, o *Data Warehouse* é capaz de fornecer dados para as duas tecnologias, o que o torna a principal fonte de dados para OLAM, cujo termo refere-se à junção de OLAP e *Data Mining*.

Segundo (HAN, et al., 2006), OLAM significa minerar interativamente em diferentes porções dos dados e em diferentes níveis de agregação, utilizando operações OLAP, podendo-se escolher as funções de *Data Mining* e algoritmos dinamicamente, além de poder navegar pelos resultados da mineração.

## 2.6 TRABALHOS RELACIONADOS

Empresas de vários segmentos, tais como do saneamento urbano, energia elétrica, telecomunicações, transporte, educação e saúde, estudam a viabilidade de implantar novos processos computacionais de BI em seus negócios por meio das tecnologias de *Data Warehouse,* OLAP e *Data Mining,* buscando o armazenamento, extração e análise automática de conhecimento.

O Trabalho de mestrado "Aplicação de Modelo de Mineração de Dados em um Sistema de Apoio à Decisão para Empresas de Saneamento" foi um dos trabalhos significativos da revisão bibliográfica. O autor (QUEYROI, 2007) apresentou de forma clara

e concisa a metodologia utilizada, fato que contribuiu bastante para o entendimento e clareza de detalhes para realização deste trabalho.

De acordo com (QUEYROI, 2007), o trabalho propôs dois objetivos, sendo o primeiro, estudar possibilidades de modernização para o segmento de saneamento no Brasil. E o segundo, abordar técnicas, metodologias e sistemas de apoio à decisão suportados principalmente por ferramentas de BI e mineração de dados.

Assim como nesta dissertação, o processo de mineração que foi conduzido no trabalho de (QUEYROI, 2007) utilizou Árvores de Decisão pelo aspecto da facilidade de visualização por parte dos especialistas, proporcionando o aprofundamento das análises dos resultados.

Como resultado, (QUEYROI, 2007) conclui que o modelo de descoberta de conhecimento aplicado ao problema acrescentou ao conhecimento do especialista, padrões de atuação que podem ser levados em conta no futuro. Por exemplo, em trocas futuras de hidrômetros, serão privilegiados os agrupamentos onde os modelos apresentaram maiores índices de precisão e não trocar, necessariamente um determinado conjunto de equipamentos (hidrômetros), onde se tem maior imprecisão de resultados.

A tese de doutorado "Ambiente para Extração de Informações através da Mineração das Bases de Dados do Sistema Único de Saúde (SANTOS, 2007)" apresentou a definição, implantação e avaliação de um ambiente de extração de informação a partir da mineração das bases de dados do SUS. A dificuldade na extração de informações gerenciais a partir da exploração das bases de dados do Sistema Único de Saúde (SUS) foi uma das questões motivadoras do trabalho, levando o autor a criar um ambiente computacional para extração de informações utilizando técnicas de mineração de dados.

Os resultados da tese de (SANTOS, 2007) confirmam a coerência da informação produzida nas questões elaboradas, comprovando a capacidade do ambiente em extrair informações úteis à gestão da Saúde Pública. Os resultados permitiram concluir e comprovar que o ambiente atendeu à premissa de prover uma forma simples de minerar as bases de dados do SUS pelo usuário final e extrair informações gerenciais. O trabalho deixou duas contribuições: a metodologia para a construção deste ambiente, e o ambiente implementado e disponível para uso.

O trabalho de mestrado de (PRADO, 2006), intitulado "*Data Warehouse* para Apoio à Gestão da Operação em Empresas do Transporte Rodoviário Interestadual de Passageiros", desenvolveu um *Data Warehouse* para apoio à gestão da operação em empresas operadoras do transporte rodoviário interestadual de passageiros. O modelo proposto foi desenvolvido a partir da proposta de um Sistema de Informações Gerenciais, utilizando tecnologia de *Data Warehouse* e ferramenta OLAP para análise de dados.

Como resultado da dissertação de (PRADO, 2006), foi implantado um *Data Mart* para a gerência de operações, sendo que o mesmo foi concebido com o objetivo principal de proporcionar a geração de consultas *ad hoc* utilizando ferramenta OLAP. Esse sistema foi aplicado em um estudo de caso envolvendo uma empresa de transporte rodoviário, e conclusivamente apresentou boa capacidade na constituição do diagnóstico dos procedimentos da operação e apoio adequado ao processo decisório.

A dissertação de mestrado "Metodologia de Mineração de Dados para Detecção de Desvio de Comportamento do Uso de Energia em Concessionária de Energia Elétrica" de (MINUSSI, 2008), empenhou-se em minimizar as perdas comerciais e maximizar os lucros em concessionárias de energia, haja vista que quanto menos se perde, menos precisa ser gerado, e menos se desperdiça recursos naturais. Visando solucionar este problema foi desenvolvido um método de mineração de dados para detecção de desvio de comportamento no uso de energia em concessionária de energia elétrica.

Na elaboração do método, (MINUSSI, 2008) compreendeu etapas de análise e avaliação dos dados, assim como construção de um *Data Warehouse*. Foram analisadas curvas de cargas dos clientes e através dessa análise observou-se o perfil de consumo dos mesmos, embasados na análise foram aplicados os algoritmos de mineração de dados, como o algoritmo de associação *Apriori* para fornecer padrões de indicadores de perfil dos consumidores bem como os algoritmos de Árvore de Decisão e Classificadores Bayesianos.

Os resultados obtidos por (MINUSSI, 2008) validam o método desenvolvido e implementado, permitindo sua utilização em uma concessionária de energia elétrica. Os algoritmos foram aplicados e foi traçado um comparativo dos mesmos, sendo que todos resultaram em boas respostas de mineração e poucas diferenças entre os resultados.

O trabalho "Modelo para Elaboração de Cenários do Setor Energético, Utilizando Técnicas de *Data Mining*" de (ZARUR, 2005) determinou um modelo computacional,

fazendo uso de técnicas de computação de alto desempenho, em especial a mineração de dados, para definição, prospecção e acompanhamento da conjuntura do setor energético, através da elaboração de cenários futuros, baseado no cenário atual, em passados e através de amostras estatísticas representativas, caracterizando uma base fundamental para qualquer exercício de planejamento.

O trabalho desenvolvido por (KANASHIRO, 2007), "Um *Data Warehouse* de Publicações Científicas: Indexação Automática da Dimensão Tópicos de Pesquisa dos *Data Marts*", insere-se no contexto do projeto de uma Ferramenta Inteligente de Apoio à Pesquisa, chamado de FIP.

A ferramenta proposta por (KANASHIRO, 2007) teve como propósito a recuperação, organização e mineração de grandes conjuntos de documentos científicos da área de computação. Nesse contexto, foi projetado um *Data Warehouse* de artigos para a ferramenta FIP e, adicionalmente, realizado experimentos com técnicas de mineração e Aprendizado de Máquina para automatizar o processo de indexação das informações, descoberta de tópicos e documentos armazenados no DW.

Como resultados do trabalho de (KANASHIRO, 2007), as consultas multidimensionais realizadas com as ferramentas FIP mostraram-se adequadas para análises de tendências e evolução nas pesquisas devido ao desempenho, facilidade de uso, flexibilidade na manipulação dos dados e variedade de operações sobre os mesmos. O levantamento sobre repositórios de materiais científicos permitiu identificar as informações importantes que são armazenadas e utilizadas, como os relacionados aos tópicos de pesquisa, que possibilitam a recuperação e exploração em grandes quantidades de publicações.

O trabalho de mestrado de (QUISPE, 2003), "Técnicas e Ferramentas para a Extração Inteligente e Automática de Conhecimento em Banco de Dados", apresentou uma visão geral e exploratória de técnicas e ferramentas para a extração inteligente e automática de conhecimento em grandes bancos de dados (EIACB), combinadas aos sistemas de base de dados, almoxarifado de dados, estatística, aprendizagem de maquina, visualização de dados e recuperação de informação. O trabalho expõe a construção do *Data Warehouse* para a EIACBD e o uso de OLAP.

Comparando com o que foi desenvolvido nos trabalhos de (SANTOS, 2007), (PRADO, 2006), (MINUSSI, 2008), (ZARUR, 2005), (KANASHIRO, 2007) e (QUISPE,

2003), esta dissertação possui algumas similaridades nos conteúdos abordados, uma delas é a implementação do ambiente de *Data Warehouse,* associado às tecnologias OLAP; assim como a aplicação de algoritmos de mineração de dados no estudo de caso.

Contudo, em (SANTOS, 2007) a pesquisa foi direcionada para o setor de saúde; em (PRADO, 2006) para o setor de transportes; (MINUSSI, 2008) e (ZARUR, 2005) para o setor de energia elétrica; em (KANASHIRO, 2007) para o setor de educação; em (QUISPE, 2003) não há uma área determinada a priori; e neste trabalho a pesquisa é fundamentada para o segmento do saneamento urbano.

O artigo "*Data Mart* para Apresentação dos Resultados Econômico-financeiros da Setorização – Estudo de Caso SANASA" de (PASSINI, et al., 2004), apresentou o projeto desenvolvido na SANASA (Sociedade de Abastecimento de Água e Saneamento – Campinas SP) para tratamento de informações econômico-financeiras dos diversos setores de abastecimento e respectivos índices de perdas, utilizando a tecnologia de *Data Marts*. O projeto permitiu que as informações fossem visualizadas através de gráficos de tendência, que evidenciaram as prioridades para tomada de decisão e implantação de projetos no combate as perdas físicas e de faturamento no saneamento. O monitoramento informatizado viabilizou o confronto das informações antes e depois das ações preventivas e corretivas.

O artigo "Mineração de Dados para Detecção de Fraudes em Ligações de Água", também de (PASSINI, et al., 2002), expôs um projeto piloto desenvolvido na SANASA, onde utilizou-se o software *DB2 Intelligent Miner*, da IBM, a fim de se detectar fraudes em ligações de água através de técnicas de mineração de dados, tendo como principal motivação o combate às perdas físicas e o enfoque ao crescente número de ligações irregulares.

Durante o levantamento da pesquisa bibliográfica encontraram-se na literatura outros trabalhos significativos na área de extração do conhecimento em bases de dados para o setor de saneamento, além dos já discutidos, dentre eles, (TEIXEIRA, 2006; ZIULKOSKI, 2003; PASSINI, 2002).

## 2.7 CONSIDERAÇÕES FINAIS

As tecnologias OLAP e *Data Mining* transformaram-se em abordagens estereotipadas no cenário de Bancos de Dados, e os especialistas desta área buscam cada vez mais utilizá-las de forma eficiente em sistemas de suporte à decisão. Vimos que OLAP não é uma tecnologia

nova e que alguns produtos existem no mercado, com destaque para o software *OLAP Pentaho Analysis View* que foi utilizado neste trabalho. OLAP é qualquer sistema que capture dados sumarizados, e permita que as agregações sejam apresentadas com suporte as operações de dados complexos por meio do *Slice and Dice*.

Em geral, as empresas necessitam de um repositório exclusivo que armazene adequadamente os dados extraídos das diversas fontes de dados, a fim de disponibilizá-lo para análises e melhor entendimento do negócio. O BD deste ambiente, que surgiu como solução para prover integração de informações gerenciais para a tomada de decisões pelas ferramentas OLAP e *Data Mining*, que são tecnologias complementares, é denominado *Data Warehouse*.

As ferramentas OLAP são utilizadas para a criação de relatórios gerencias, facilitando a formatação multidimensional e análise dos dados. Enquanto que as ferramentas de *Data Mining* utilizam técnicas estatísticas com a finalidade de encontrar correlações e padrões entre os dados que ajudem nas decisões estratégicas da empresa.

Os trabalhos relacionados apresentados neste capítulo reforçam a importância de estudos acadêmicos utilizando ambientes de *Data Warehou*se, tecnologias OLAP e *Data Mining,* voltados para resolver problemas de áreas que provêem serviços essenciais e de interesse de toda a população, como é o caso das áreas do saneamento urbano, energia elétrica, telecomunicações, transporte, educação e saúde.

# CAPÍTULO 3

***Descreve a metodologia proposta da pesquisa, proveniente da implementação do Sistema de Apoio à Decisão (SAD) para o setor de saneamento, este que é composto pelo Data Warehouse, tecnologias OLAP e Data Mining. O capítulo prossegue com a apresentação da companhia de abastecimento de água utilizada no estudo de caso e com as discussões de alguns dos resultados extraídos do ambiente de apoio à decisão, Pentaho, quanto à aplicação da tecnologia OLAP.***

## 3 PROJETO E IMPLEMENTAÇÃO DO SAD

O controle e o uso eficaz dos dados armazenados é um dos grandes desafios enfrentados atualmente pelas empresas de TI. Vário softwares ainda são produzidos sem preocupação com a geração futura de informações integradas e estratégicas. Diante deste contexto surgem sistemas isolados, com cada setor dentro da organização (comercial, financeiro, operacional etc.) possuindo sua base de dados desintegrada das demais.

Um Sistema de Apoio à Decisão (SAD) consiste em um ambiente projetado para apoiar, contribuir e influenciar no processo de tomada de decisão. Conforme ilustra a Figura 3.1, o SAD utilizado e implementado nesta pesquisa é formado por três componentes: os Dados (Dispostos no *Data Warehouse Departamental*), o SGBD, e as Ferramentas de Apoio à Decisão.

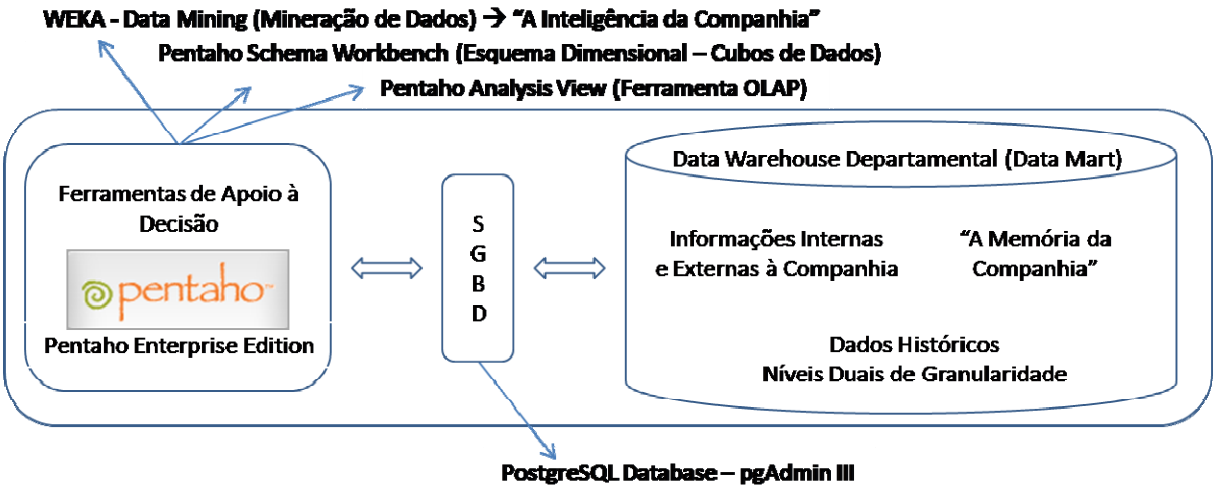



**Figura 3.1 - componentes do ambiente de apoio à decisão**

A solução para sistemas desintegrados se constitui em reorganizar a maneira como a informação é armazenada, disponibilizada e acessada. Este cenário favorece o desenvolvimento de *Data Warehouse* que é um ambiente de suporte à decisão propício a unir dados armazenados em diferentes fontes, organizá-los e entregá-los aos "tomadores" de decisões.

A coleção de dados inter-relacionados é formada por informações internas e externas à organização, por dados históricos e níveis duais de granularidade. O papel do SGBD em ambientes de apoio à decisão é permitir que os usuários definam, construam e manipulem o Banco de Dados com dados integrados e compartilhados. Um SGBD pode representar a unificação de diversos arquivos, que, de outra forma, seriam distintos, eliminando-se total ou parcialmente a redundância entre os mesmos. Já o compartilhamento não significa apenas que as aplicações existentes podem compartilhar dados do Banco de Dados, mas também que novas aplicações podem ser desenvolvidas para operar sobre os mesmos dados armazenados.

Quando se deseja extrair informações apenas de uma determinada área da empresa opta-se por implementar *Data Warehouse Departamental* (*Data Mart)* ao invés de *Data Warehouse Corporativo*. Desta forma, o *Data Warehouse Departamental* deste trabalho foi implementado no SGBD *PostgreSQL,* utilizando a modelagem dimensional do esquema Constelação de Fatos. Este *Data Warehouse* corresponde aos dados do setor do saneamento 64 (bairro de Miramar e proximidades) da cidade de João Pessoa.

As Ferramentas de Apoio à Decisão são softwares utilizados para manipular os dados extraídos do *Data Warehouse* através da estrutura de cubos de dados, de funções de agregações (sumarizações, médias, mínimos, máximos, *count*, etc.), de funções estatísticas ou de funções gráficas. Elas auxiliam na simulação e análise dos dados, proporcionando a descoberta de novos conhecimentos. As ferramentas de apoio à decisão utilizadas neste trabalho foram:

- *Pentaho Schema Workbench:* ferramenta responsável pela criação dos cubos de dados (tabelas de fatos), dimensões (tabelas de dimensões) e métricas do esquema dimensional Constelação de Fatos. Neste trabalho foram criados dois cubos de dados, o *cubo_perfil_setor* e o *cubo_perda_aparente*. A Figura 3.2 ilustra o esquema dimensional e o código XML correspondente ao *cubo_perda_aparente.* A seção 3.2.5 aborda a criação e publicação dos cubos de dados.

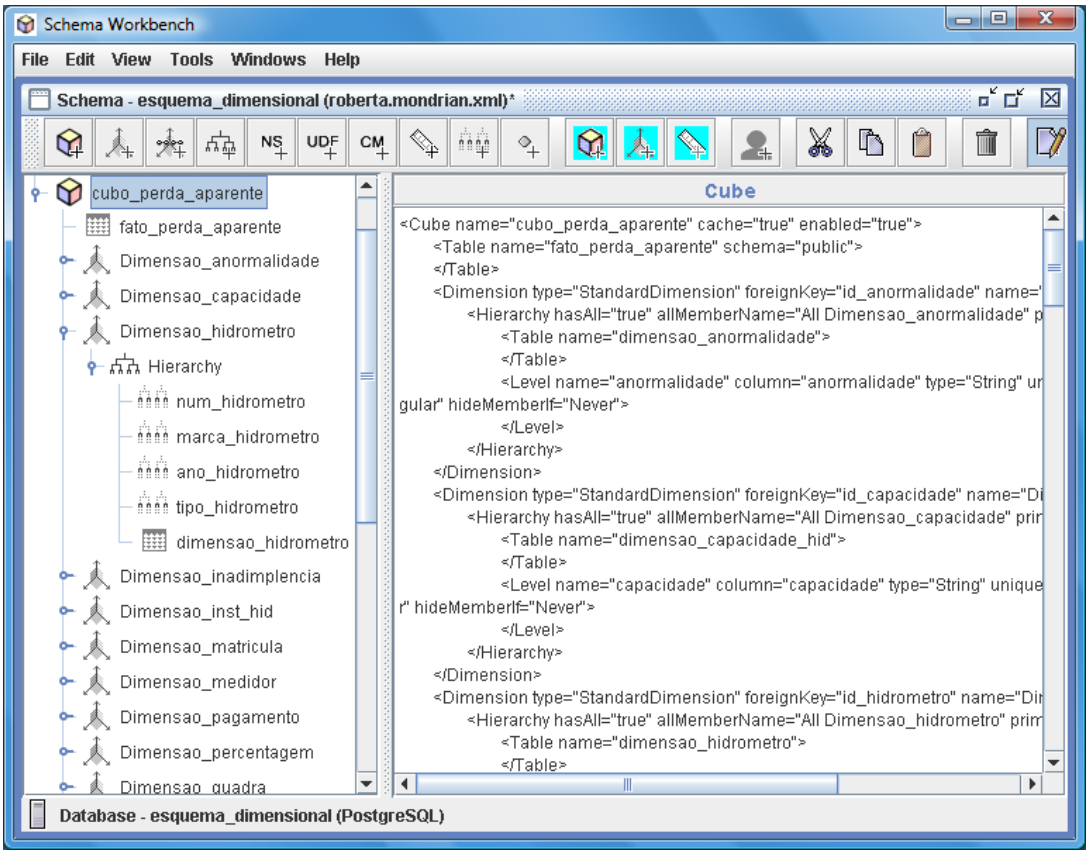



**Figura 3.2 - criação dos cubos de dados pela ferramenta schema workbench**

- *Pentaho Analysis View: f*erramenta OLAP que executa as operações *Slice and Dice* sobre o arquivo XML do esquema dimensional, conforme mostra a Figura 3.3. A seção 3.2.6 aborda as consultas e análises das dimensões dos cubos de dados.

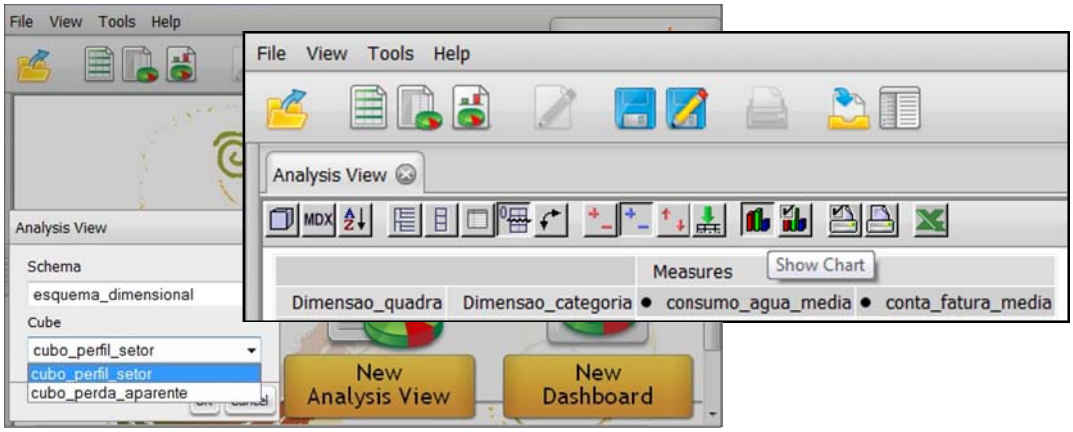



**Figura 3.3 - tela inicial da ferramenta OLAP *pentaho analysis view***

- WEKA (*Waikato Environment for Knowledge Analysis)*: ferramenta que implementa os principais algoritmos *de Data Mining*. A Figura 3.4 mostra a tela inicial de pré-mineração dos dados.

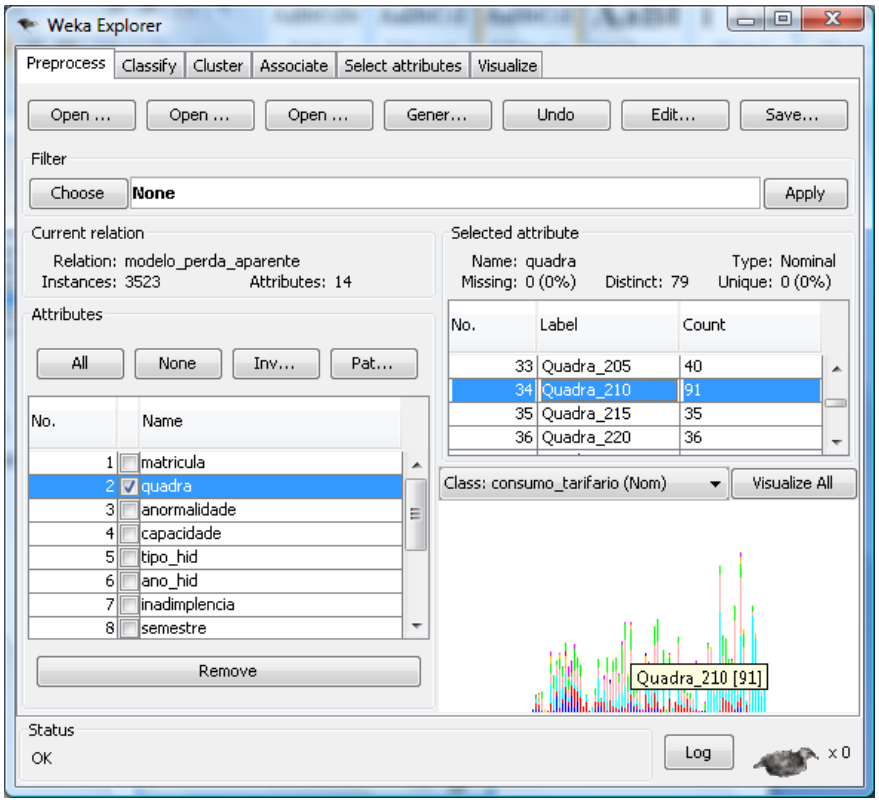



**Figura 3.4 - mineração de dados pela ferramenta WEKA**

A seção 4.1.1 aborda os algoritmos de *Data Mining* aplicados sobre os modelos de mineração de dados desenvolvidos para o estudo de caso deste trabalho.

## 3.1 O ESTUDO DE CASO

A Companhia de Água e Esgota da Paraíba (CAGEPA/PB) foi utilizada como estudo de caso desta pesquisa. Ela é responsável pelo abastecimento de água e coleta de esgotos em 185 dos 223 municípios paraibanos. As duas principais atividades desenvolvidas pela empresa são: distribuição de água tratada e a coleta e tratamento de esgotos.

O atendimento nos municípios do estado é feito através das 6 Unidades de Negócio (UN), são elas: Litoral, Brejo, Borborema, Espinhara, Rio do Peixe e Alto Piranhas. A UN Litoral faz parte da pesquisa de campo realizada por este trabalho. Segundo dados da CAGEPA, a UN Litoral envolve 23 municípios; 262.374 ligações de água; 72.825 ligações de esgoto; e abrange uma população urbana de 1.091.361 pessoas, destas, 998.738 são abastecidas por água e 33,16% servida pelo sistema de esgoto.

A UN Litoral possui vários setores, dentre eles, o setor 64 da localidade 001, que corresponde ao bairro de Miramar e suas proximidades, na cidade de João Pessoa. Ele foi escolhido para aplicação da metodologia de detecção de perdas aparentes proposto por este

trabalho. Um dos motivos que levou ao estudo desta área foi que ela contempla as distintas realidades sociais: habitações populares; classe alta, classe média, além de possuir diversos tipos de estabelecimentos (comercial, industrial, público, praças, terrenos, etc.). A maioria dos consumidores desta região é de classe média alta e a minoria é de classe baixa. Outra razão para a escolha foi à condescendência por parte da gerência da companhia em disponibilizá-lo.

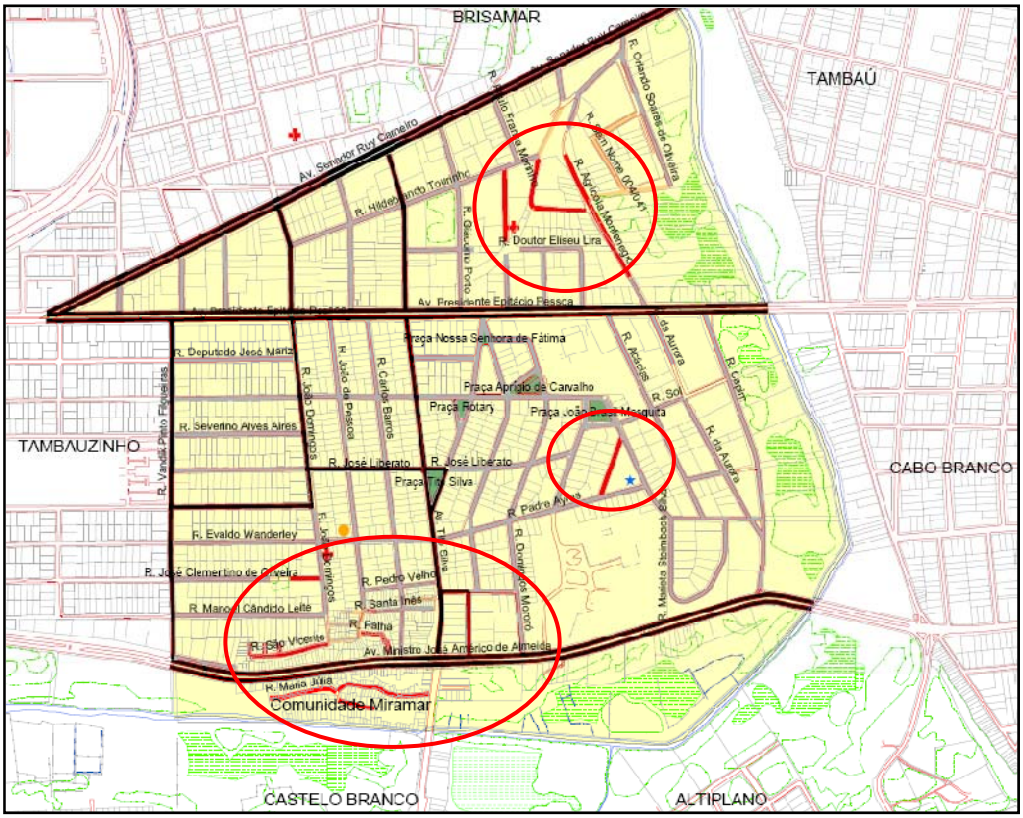



**Figura 3.5 - sistemas de logradouros de João Pessoa - setor Miramar**

**Fonte: (Planejamento/JP, 2006)**

A Figura 3.5 ilustra o mapa referente ao bairro Miramar. Este setor possui aproximadamente 17.800 pontos de ligações de água. A Comunidade Miramar, que se encontra destacada em vermelho, próximo ao Castelo Branco, representa a porção da classe de baixo poder aquisitivo. Existem outros dois pontos no mapa, em vermelho, que também representam a classe baixa.

O BD da CAGEPA está lotado na cidade de Recife-PE, onde se encontra a empresa que administra o seu sistema de informação. Os dados adquiridos da companhia para o estudo de caso são referentes aos cadastros dos consumidores; dos registros de consumo; dos imóveis, de contas/faturas; dos hidrômetros e movimentações dos hidrômetros; entre outros. Segue na Tabela 3.1 o dicionário de dados dos principais atributos adquiridos.

**Tabela 3.1 - dicionário de dados. Fonte: CAGEPA**

| Informações (Atributos) | Layout dos dados (Descrição) |
|---|---|
| Matricula do consumidor | Cada consumidor possui uma matrícula que é única. |
| Inscrição: 111.22.333.4444 | Quatro atributos numéricos que identificam a localização física do imóvel.<br>Exemplo:111 = localidade; 22 = setor de faturamento; 333 = número da quadra; e 444 = número do lote. |
| Nome do consumidor | Atributo alfanumérico informando o nome do proprietário do imóvel. |
| Endereço do imóvel | Nome da Rua/Avenida, Bairro, número e complemento. |
| Situação de Água | Ligada/Cortada/Suprimida Total ou Parcial. |
| Situação de Esgoto | Potencial/Ligado Normal/Factível. |
| Pontos de Utilização | Quantidade de pontos de água disponíveis m cada imóvel. |
| Categoria | Atributo alfanumérico que identifica o tipo de economia (número de domicílios atendidos pelo hidrômetro) ao qual o consumidor pertence. Comercial/Residencial/Industrial/Público |
| Subcategoria | Atributo alfanumérico que identifica o subconjunto do tipo de economia ao qual o consumidor está inserido.<br>Casa/Loja/Edifício/Escola/Terreno...<br>Existem 41 subcategorias distintas no setor. |
| Indicativo de medidor | SIM – possui hidrômetro<br>NÃO – não possui |
| Referência do consumo | Identifica os ciclos de medição das leituras nos hidrômetros. Cada matrícula totaliza 12 tuplas, visto que são dados obtidos no período de 1 ano. |
| Consumo do mês de referência | Atributo numérico que identifica o valor total de consumo de água em $m^3$/hora do mês de referência. |
| Valor total da conta | Atributo numérico que identifica o valor total de consumo em reais (R$) do mês de referência.<br>Soma do valor de água + valor de esgoto + valor de serviço - valor de crédito |
| Data de Pagamento | "9999-31-12" significa que a conta não foi paga. |
| Data de Vencimento | Data de vencimento da fatura. |
| Número do Hidrômetro | "**********" significa que não há hidrômetro. |
| Anormalidade de leitura | Observações sobre o hidrômetro ou imóvel.<br>Existem 25 anormalidades distintas no setor. |
| Data de Instalação do Hidrômetro | "9999-31-12" significa que há hidrômetro. |
| Capacidade do Hidrômetro | Vazão do hidrômetro em $m^3$/hora. |
| Ano de Fabricação do Hidrômetro | "999999" significa que não há hidrômetro. |
| Tipo de Hidrômetro | Especifica o modelo do hidrômetro. Existem 6 tipos diferentes no setor. |

Esta pesquisa nasceu da necessidade de investigar se as perdas aparentes estavam distribuídas proporcionalmente pelo setor selecionado para o estudo de caso ou se elas estavam concentradas em áreas específicas, como exemplo, em áreas com níveis sociais similares. Além disso, tem como objetivo realizar uma avaliação através da análise do perfil do consumidor e do imóvel de acordo com a categoria de consumo, situação da água e esgoto, avaliando os consumos e valores medidos e faturados por mês, durante um período de um ano, entre Maio de 2007 a Abril de 2008.

Após o levantamento das condições necessárias para elaboração dos modelos de *Data Mining*, especificou-se o conjunto de atributos requeridos à companhia de abastecimento de água relativos ao período de 12 meses. Em seguida, foram definidas as ferramentas para implementação do sistema de apoio à decisão, com capacidade de fornecer resultados confiáveis para o estudo das perdas aparentes. O entendimento dos dados e termos técnicos da Engenharia Hidráulica, especificamente da área de Saneamento, também se fez necessário. Os dados precisaram ser tratados para que os resultados esperados fossem alcançados, visto que inconsistências e algumas carências nas informações foram encontradas e precisaram ser eliminadas.

O modelo de *Data Mining* desenvolvimento foi dirigido no sentido de que os resultados finais espelhassem tanto a posição de um único setor (64 – Bairro Miramar), como também, permitissem que se estendesse para toda a cidade, proporcionando as visões gerenciais de toda setorização de consumo do Estado atualmente atendida pela companhia de abastecimento.

O *Data Warehouse Departamental* foi criado a partir dos dados cadastrais dos hidrômetros (capacidade, tipo, número, marca, ano e data de instalação), dos indicativos de medição, dos volumes micromedidos ou estimados, dos valores faturados (vencimento/pagamento), das quantidades de ligações e de economias, das categorias (Comercial, Industrial, Pública e Residencial) e subcategorias de consumo, das percentagens de perdas, dos períodos de referência (Semestres, Quadrimestres e Meses), dos consumos de água e esgoto, dos tipos e quantidades de anormalidades e das inadimplências. Os dados foram orientados de modo a permitir os agrupamentos principalmente por matrículas e quadras, visando às informações referentes às perdas aparentes (atributos do tipo: anormalidade, inadimplência, percentagem de aumento ou diminuição de consumo de água durante os 12 meses de medição etc.).

O motivo principal que levou ao desenvolvimento de um ambiente de *Data Warehouse* ao invés de um ambiente de Banco de Dados tradicional reside no fato dos ambientes de suporte a decisão e extração do conhecimento em bases de dados serem caracterizados pela não-volatilidade dos dados e pela complexidade das consultas *ad hoc*.

Os projetos envolvendo ambientes de *Data Warehouse* são fortemente influenciados visando o *modelo de dados* eficiente, ou seja, facilidade em manipular funções de agregações, associações, classificações etc., ao contrário do que ocorre com aplicações tradicionais (Banco do Dados Relacionais), nas quais o foco reside na definição de um *modelo de transações* eficiente. As técnicas tradicionais de modelagem de dados (Modelo Entidade-Relacionamento, E/R) são inadequadas para sistemas de apoio à decisão (projeto de *Data Warehouse)* pelas seguintes razões:

- A estrutura do banco de dados resultante é muito complexa para que usuários finais possam compreender. No melhor dos casos, o modelo gerado é composto por várias tabelas interligadas por uma complexa rede de relacionamentos. Até mesmo consultas simples requerem o estabelecimento de *joins* (junções) entre múltiplas tabelas, o que se constitui num mecanismo propenso a erro e inapropriado ao acesso eficiente dos grandes volumes de dados;
- Os modelos tradicionais enfatizam a normalização como forma de evitar dados redundantes e consequentes anomalias de atualizações. Em ambientes de *Data Warehouse*, porém, o dado não é atualizado e sua redundância demonstra ser, ao contrário, um artifício para maximizar a eficiência das consultas ao repositório;
- O modelos E/R não são extensíveis o suficiente para acomodar mudanças nos requisitos de negócio do sistema. Em consequência, toda a estrutura de entidades e relacionamentos, bem como as funcionalidades da aplicação nela apoiadas, precisam ser revistas e adaptadas quando há inclusão de novos elementos ou mudança em requisitos de projeto. Em contrapartida a essa característica, em ambientes de *Data Warehouse* os requisitos do usuário estão sujeitos a constantes mudanças, tornando a evolução do esquema conceitual um aspecto primordial para o sucesso do projeto.

De acordo com (KIMBALL, 1997; KIMBALL, et al., 2002; PAIM, 2003), existe um consenso na comunidade científica que indica os modelos relacionais tradicionais, gerados a partir da técnica de E/R, como inadequados ao projeto de sistemas de apoio à decisão. Os esquemas relacionais resultantes de uma modelagem tradicional, onde as tabelas do modelo

são normalizadas, contrariam uma premissa básica dos *Data Warehouses* (esquemas dimensionais), que é a recuperação intuitiva e em alta performance dos dados. Desta forma, optou-se por um *Data Warehouse.*

## 3.2 PROCESSO DE EXTRAÇÃO DO CONHECIMENTO: FASE 1

### 3.2.1 Implementação do Data Warehouse

A modelagem dimensional do *Data Warehouse (Data Mart Comercial)* desenvolvida neste trabalho foi implementada fisicamente no SGBD Relacional *PostgreSQL*, conforme ilustra a Figura 3.6.

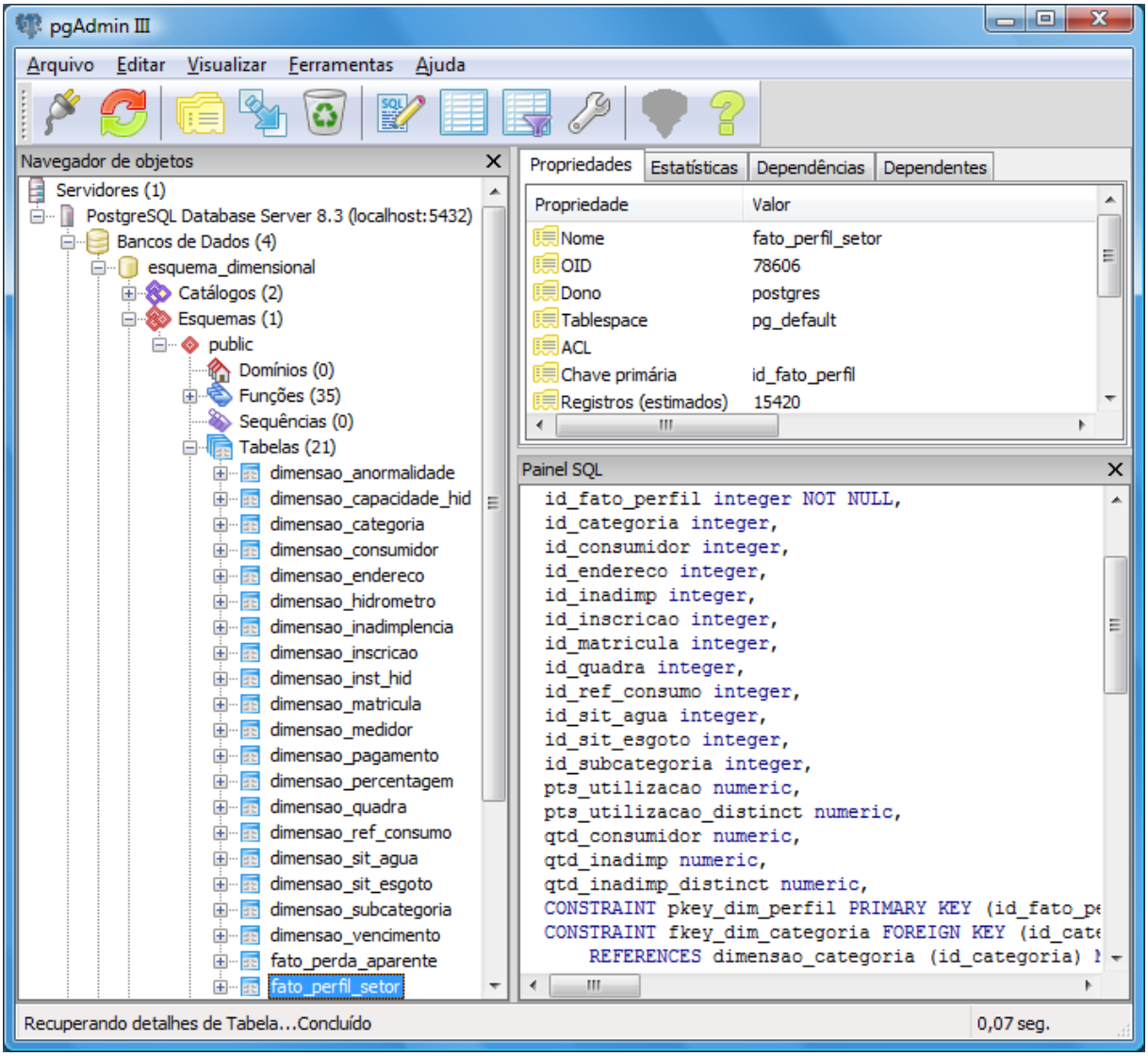



**Figura 3.6 - desenvolvimento da modelagem dimensional no SGBD *postgresql***

Para atender as necessidades de análise das informações, o SAD utiliza o *Data Warehouse* Departamental para dar suporte às operações OLAP do tipo *Slice and Dice* e

também para dar suporte às técnicas de *Data Mining*. Na concepção do DW se fez necessário definir a área de atuação (setor comercial), as tabelas de fatos e suas dimensões.

A modelagem lógica do *Data Warehouse* possui duas tabelas de fatos ("Perfil do Setor" e "Perdas Aparentes") e 19 tabelas de dimensões (dentre elas: Dimensão Anormalidade, Dimensão Categoria, Dimensão Inadimplência, Dimensão Referência de Consumo, etc.), conforme Figura A.1 do APÊNCIDE A.

O *Data Warehouse* se apresentou de forma satisfatória para realização das consultas OLAP e para aplicação das técnicas de *Data Mining*, conforme será discutido mais adiante.

### 3.2.2 Pré-Processamento: Limpeza e Enriquecimento

A etapa de Pré-processamento e Transformação serve para detectar os erros de cadastros e inconsistência dos dados extraídos do ambiente operacional. É realizada a limpeza dos dados a fim de adequar e carregar apenas os dados necessários no *Data Warehouse*. Esta adequação dos dados aos algoritmos de mineração se dá através da integração de dados heterogêneos, remoção de dados incompletos, eliminação de repetição dos dados e dos problemas de tipagem.

Alguns registros da base de dados foram excluídos por não apresentarem informações concisas. Nesta situação se encontravam as instâncias que possuíam o valor de consumo de água igual a zero, o valor da conta igual a zero e a data de pagamento igual '9999-12-31' (inadimplência). Estes casos correspondem aos consumidores que se encontram inadimplente por uma conta no valor de R$ 0,0, ou seja, um débito inexistente. Desta forma, não foram consideradas e possivelmente representam erros no cadastro da base de dados da companhia de abastecimento de água.

Houve limpeza e transformação dos dados com as datas de pagamento que se encontravam nulas (em branco) e que foram atualizadas para '9999-12-31', pois de acordo com a gerência comercial da companhia CAGEPA (usuária do estudo de caso), tais circunstâncias correspondem aos consumidores que estão com a conta em aberto, ou seja, estão inadimplentes. Nos atributos "Ano de Instalação do Hidrômetro" e "Ano do Hidrômetro" os valores que ultrapassavam quatro dígitos foram truncados para '9999' e os anos cadastrados com valor zero também foram atualizados para '9999'. Os anos com apenas dois dígitos foram completados para quatro dígitos (ex. 95 → 1995).

Ainda na fase de pré-processamento, foram encontradas 23 matrículas com mais de 12 meses de referência de consumo, entretanto, verificou-se que os registros estavam duplicados e alguns até triplicados. Neste caso, foi necessário eliminar das 23 matrículas, os registros excedentes e carregar no *Data Warehouse Departamental* apenas os 12 meses.

Outra situação de limpeza dos dados foi aplicada ao atributo "Pontos de Utilização" e "Valor da Conta". As instâncias referentes aos pontos de utilização de água cadastradas com zero não foram consideradas no estudo, pois de acordo com a gerência da CAGEPA, cada matrícula (consumidor) deve possuir pelo menos 1 ponto de utilização de água por imóvel cadastrado.

Para os casos onde o atributo "Valor da Conta" estava cadastrado com o valor zero, foi preciso verificar se na tupla em questão o atributo "Situação da Água" encontrava-se cadastrado com o valor "Cortada". Caso contrário, as matrículas analisadas não poderiam ser consideradas e por isso, deveriam ser eliminadas da base de dados.

Todas as ocorrências apresentadas acima, e que foram ajustadas na fase de pré-processamento, configuravam erros de cadastro no banco de dados da companhia, e por isso tiveram que ser desconsideradas no *Data Warehouse*, pois caso não fossem realizadas as limpezas necessárias, iriam provocar erros de semântica nas informações a serem mineradas.

### 3.2.3 Transformação, Seleção e Integração dos Dados

Os dados selecionados para dar suporte na fase de *Data Mining* correspondem aos agrupamentos dos setores e logradouros característicos da cidade de João Pessoa, Estado da Paraíba. Foram 82 quadras disponibilizadas para o estudo de caso, sendo que três delas não puderam ser utilizadas, visto que elas não possuíam os 12 meses completos de medições. Desta forma, 79 quadras foram detalhadas e tiveram seus dados utilizados e aplicados aos modelos de *Data Mining* propostos.

Ao todo foram 15.420 registros correspondentes às 1.285 matrículas dos consumidores com 12 meses (de 05/2007 a 04/2008) contínuos de registro de consumo e faturamento junto à companhia de abastecimento. Nos modelos propostos para aplicação da mineração de dados, utilizaram-se todo o conjunto das 79 quadras. Contudo, nem todos os atributos foram considerados, visto que alguns deles não agregariam valor para o processamento da mineração pela ferramenta WEKA. Como exemplo, tem-se os atributos nome e endereço do consumidor

que não foram carregados para mineração, afinal, não era importante saber estas informações e a identificação dos consumidores precisava ser preservada.

Visando apresentar e gerar resultados mais claros e concisos, além de minimizar o tempo de processamento da mineração, alguns valores dos atributos foram agrupados. Neste caso estão os atributos "Data de Instalação do Hidrômetro", "Capacidade do Hidrômetro" e "Referência de Consumo".

No caso do atributo "Data de Instalação do Hidrômetro" foram geradas 4 classificações: "Mais de 18 anos", que corresponde ao intervalo de datas entre 1976 e 1989; "Entre 10 e 18 anos", que corresponde ao intervalo de datas entre 1990 e 1998; "Entre 3 e 9 anos", que corresponde ao intervalo de datas entre 1999 e 2005; e "Menos de 3 anos", que corresponde ao intervalo de datas entre 2006 e 2008. O mesmo ocorreu com o atributo "Ano de fabricação do hidrômetro", ou seja, foram agrupados em "1984_a_1988", "1989_a_1993", "1994_a_1998", "1999_a_2003", "2004_a_2008" e "nao_informado"

Para o atributo "Capacidade do Hidrômetro" foram geradas 4 classificações: "Até 3 $m^3$/hora", que corresponde aos hidrômetros com vazões de 1,5 e 3 $m^3$/hora; "De 5 a 10 $m^3$/hora", que corresponde aos hidrômetros com vazões de 5, 7 e 10 $m^3$/hora; "Acima de 10 $m^3$/hora", que corresponde aos hidrômetros com vazões de 20 e 30 $m^3$/hora; e "Não Informado", que corresponde as vazões dos hidrômetros não cadastradas, ou seja, nulas.

O atributo "Referência de Consumo" foi agregado em "Primeiro_Semestre" e "Segundo_Semestre". Onde a primeira classificação corresponde aos meses de Maio de 2007 a Outubro de 2007 e a segunda classificação corresponde aos meses de Novembro de 2007 a Abril de 2008. Foi também agregado por quadrimestres de consumo, isto é, "Quadrimestre_1" (Maio a Agosto de 2007), "Quadrimestre_2" (Setembro a Dezembro de 2007) e "Quadrimestre_3" (Janeiro a Abril de 2008).

O cálculo para diagnosticar o aumento ou não, em percentagem, da fatura do consumidor é apresentado na Equação 3.1.

$$\frac{[(M_U - M_P) \times 100]}{M_P}$$

**Equação 3.1 - percentagem de faturamento do consumidor**

Onde, $M_P$ significa a média de faturamento do primeiro semestre, ou seja, os primeiros 6 meses de consumo e $M_U$ significa a média de faturamento do segundo semestre, ou seja, os últimos 6 meses de consumo, ambos por consumidor. Se o resultado da equação for maior que zero, então houve aumento na conta (aumento de consumo de água) nos últimos 6 meses. Se resultado for menor que zero, então houve diminuição na conta do consumidor nos últimos 6 meses. Sendo este último caso de maior interesse para a detecção das perdas aparentes, visto que uma diminuição drástica no consumo pode representa problemas, irregularidades ou fraudes nos medidores, e nestes casos uma inspeção (visita técnica) no imóvel é indicada para comprovar ou não o pré-diagnóstico.

Ao gerar as médias de faturamento dos consumidores referentes ao primeiro e segundo semestre não foram consideradas as contas com valor zero, visto que a média tenderia para menos e ao realizar os comparativos e percentagens de consumo, o valor não representaria a situação real do consumidor. O mesmo aconteceu com os valores de consumo de água igual a zero, que não foram considerados na geração das respectivas médias. O consumo igual a 0 indica que foi atribuída a média de consumo do consumidor. Em geral, quando consumo é igual a 0, o hidrômetro se encontra quebrado/parado.

### 3.2.4 Utilização do Esquema Constelação de Fatos

A principal razão em utilizar o Esquema Constelação de Fatos para modelagem do Banco de Dados é que ele favorece a otimização das consultas fazendo com que os comandos *SQL* resultantes tenham o mínimo de junções possível. Este esquema auxilia o desempenho na fase de atualização do DW devido à desnormalização das tabelas de dimensão.

Visando construir o Esquema Constelação de Fatos otimizado para consultas e que representasse de forma clara e concisa o setor estudado, foram definidas duas tabelas de fatos e suas respectivas dimensões. Cada tabela de dimensão é definida com uma única chave primária (PK), baseada na integridade do relacionamento com a tabela de fatos, isto é, a chave primária da dimensão é chave estrangeira na tabela de fatos (FK). A modelagem das tabelas de fatos e suas dimensões encontram-se ilustradas na Figura A.1 do APÊNDICE A. Seguem as definições de ambas as tabelas de fatos:

- Tabela de Fatos Perfil do Setor

É a tabela (entidade) responsável pelo reconhecimento e análise dos consumidores e imóveis através dos dados cadastrais referentes ao setor 64, Bairro Miramar. Possui 11

dimensões associadas, são elas: Categoria, Consumidor, Endereço, Inadimplência, Inscrição, Matrícula, Quadra, Referência de Consumo, Situação da Água, Situação do Esgoto e Subcategoria.

- Tabela de Fatos Perdas Aparentes

É a tabela responsável pelas verificações e análises das irregularidades na rede de distribuição de água, visando detectar perdas aparentes no setor 64. Possui 12 dimensões associadas, são elas: Matrícula, Quadra, Anormalidade, Capacidade do Hidrômetro, Hidrômetro, Inadimplência, Data Instalação do Hidrômetro, Medidor, Pagamento, Vencimento, Percentagem e Referência de Consumo.

A Figura 3.7 ilustra parte do Esquema Constelação de Fatos correspondente ao setor estudado da Companhia de Abastecimento de Água. Nesta figura encontram-se as tabelas de fatos Perfil do Setor e Perdas Aparentes associadas às quatro Tabelas de Dimensão (Matrícula, Quadra, Categoria e Referência de Consumo). Ambas as tabelas possuem a indicação de suas chaves primárias e estrangeiras.

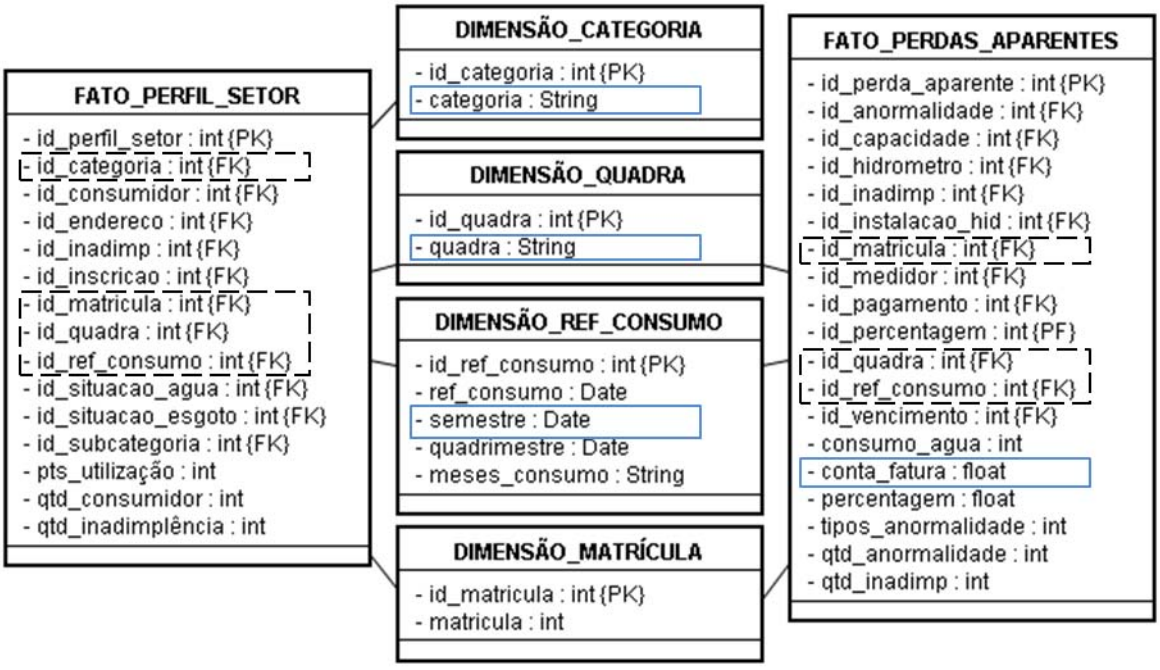



**Figura 3.7 - parte do esquema constelação de fatos para o setor de saneamento**

Na Figura 3.8 há uma consulta SQL ao Esquema Constelação de Fatos da Figura 3.7, utilizando uma ferramenta de BD tradicional (p*gAdmin/postgreSQL)* que é projetada para trabalhar com dados relacionais. Esta consulta retorna as médias de faturamento das quadras

(010, 015, 020, 030 e 050) agrupadas pela categoria de consumo comercial e semestres de referência (primeiros seis meses e últimos seis meses de medição).

Por exemplo, a média de faturamento da Quadra_015 foi de RS 35,1 no primeiro semestre de referência de consumo e de R$ 580,7 para o segundo semestre. Isto denota que houve uma maior arrecadação no segundo semestre em relação ao primeiro semestre por parte da companhia de abastecimento, e que não há necessidade de inspeção no local, nem de troca do hidrômetro. A situação se encontra estável, indicando apenas que houve aumento de consumo de água na quadra.

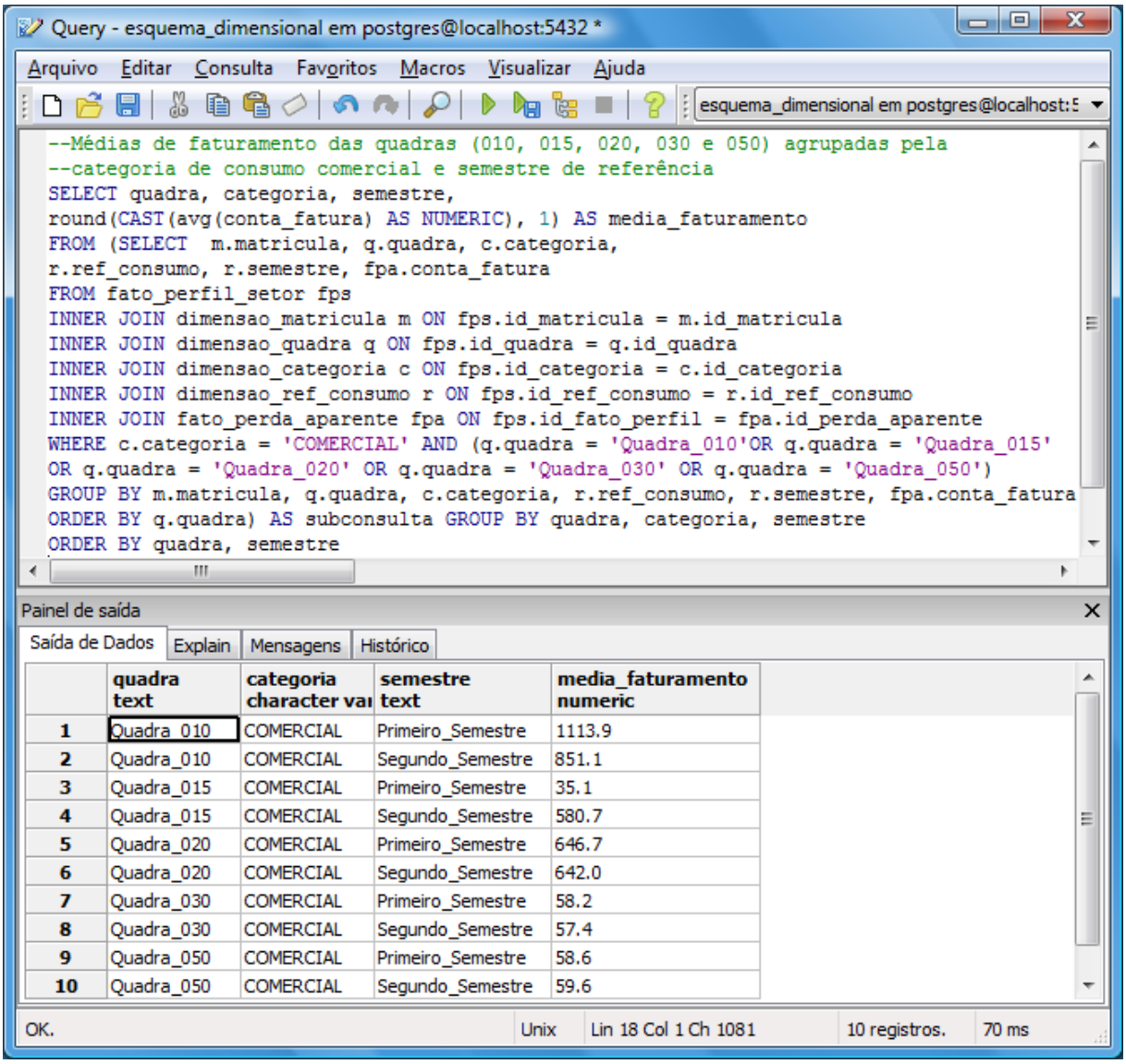



**Figura 3.8 - consulta ao esquema constelação de fatos da Figura 3.7**

Mais adiante é apresentada na Figura 3.11 (página 96) esta mesma consulta, porém, utilizando uma ferramenta OLAP que é desenvolvida para dar apoio às modelagens

dimensionais de cubos de dados. A idéia consiste em evidenciar que o uso da ferramenta OLAP torna bem mais inteligível a análise do negócio, facilitando o suporte à decisão, que ademais, não ocorre com ferramentas de BD tradicionais, visto que elas não oferecem suporte gráfico. A consulta precisa ser formulada manualmente, ficando sujeita a erros e a um tempo de formulação bem maior.

### 3.2.5 Pentaho Schema Workbench – Modelagem Dimensional

O Esquema Constelação de Fatos foi modelado utilizando o módulo *Schema Workbench* da plataforma de código aberto do *Business Intelligence*, *Pentaho,* conforme ilustra a Figura 3.9.

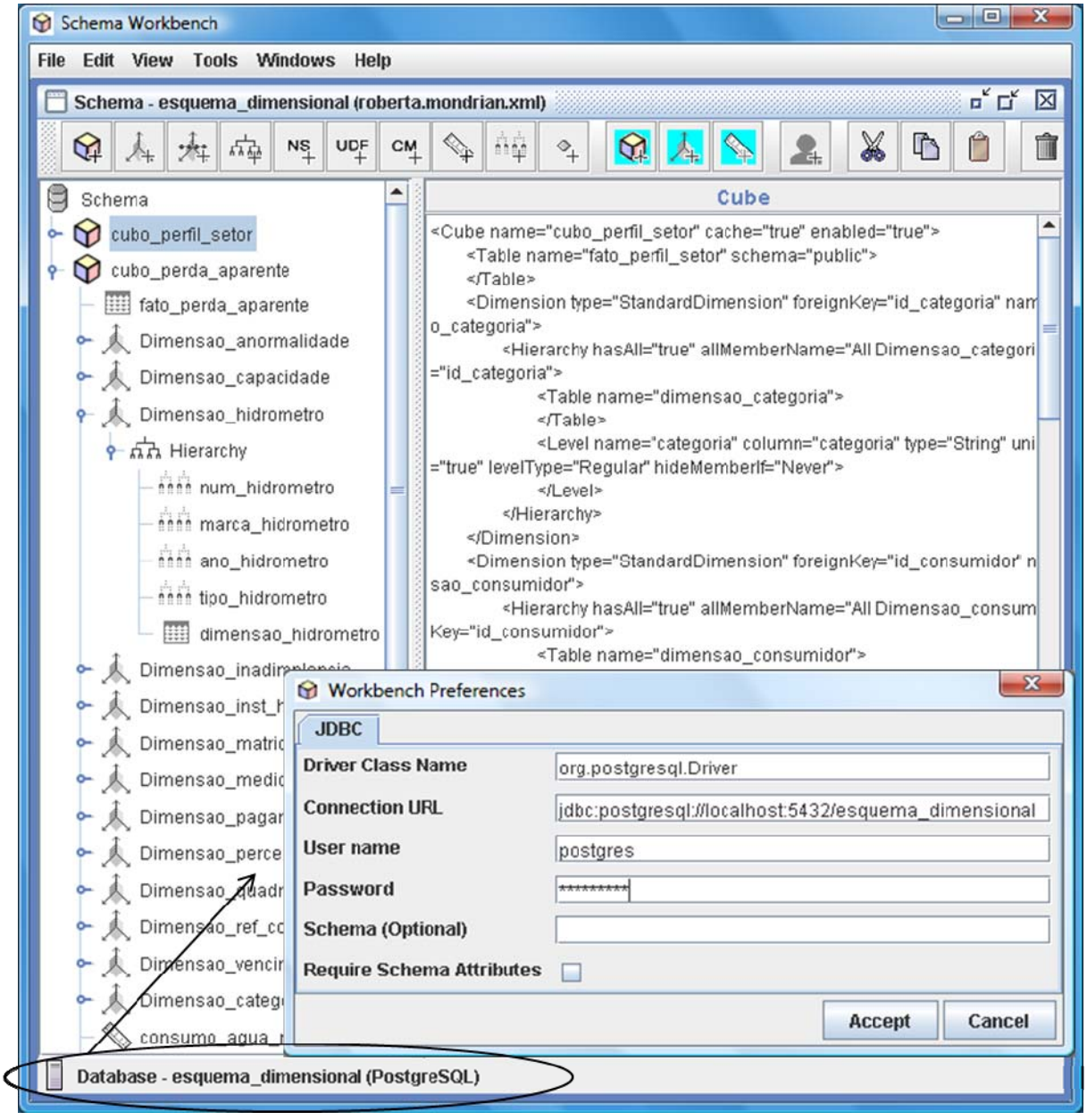



**Figura 3.9 - criação do esquema constelação de fatos através da ferramenta *schema workbench***

A ferramenta *Schema Workbench* está incorporada na plataforma do *Pentaho,* e proporciona a geração dos cubos de dados OLAP. Ela tem uma interface visual para navegar entre as definições do cubo, permitindo criar métricas, dimensões e hierarquias, que proporcionam a correta utilização e exploração do cubo de dados OLAP.

A Figura A.1 do APÊNDICE A ilustra as tabelas de fatos, suas métricas e as suas respectivas dimensões, sendo elas essenciais para formação dos cubos de dados OLAP.

Segue na Figura 3.9 a estrutura dos cubos de dados contendo as tabelas de fatos, suas dimensões, hierarquias e métricas, gerada pela ferramenta *Schema Workbench,* e que corresponde ao Esquema Constelação de Fatos modelado para este trabalho. A conexão com o Banco de Dados *postgreSQL* ocorreu via *driver* JDBC, como mostra a Figura 3.9.

Os dois cubos de dados (cubo_perfil_setor e cubo_perda_aparente) implementados pela ferramenta *Schema Workbench* são salvos no formato *XML* e precisam ser publicados para que as tecnologias OLAP (operações *Slice and Dice*) sejam aplicadas através da ferramenta OLAP Pentaho Analysis View.

### 3.2.6 Pentaho Analysis View - OLAP

Após a publicação dos cubos de dados pela ferramenta *Schema Workbench*, as operações OLAP do tipo *Slice and Dice* estarão dispostas para serem utilizadas pela ferramenta *Pentaho Analysis View,* cuja finalidade é facilitar a execução e visualização multidimensional dos dados contidos no *Data Warehouse*, fornecendo ao analista informações detalhadas, sumarizadas ou agregadas do setor.

A ferramenta OLAP além de fornecer os dados da consulta, possibilita a criação de gráficos que facilitam a compreensão das informações. A Figura 3.10 ilustra uma consulta executada sobre o cubo de dados “Perfil do Setor” do esquema Constelação de Fatos através da ferramenta *Pentaho Analysis View*.

As dimensões necessárias para execução da consulta da Figura 3.10 foram: Categoria, Subcategoria, Água e Inadimplência. As métricas utilizadas foram: quantidade de consumidores, quantidade de pontos de utilização e quantidade de inadimplências.

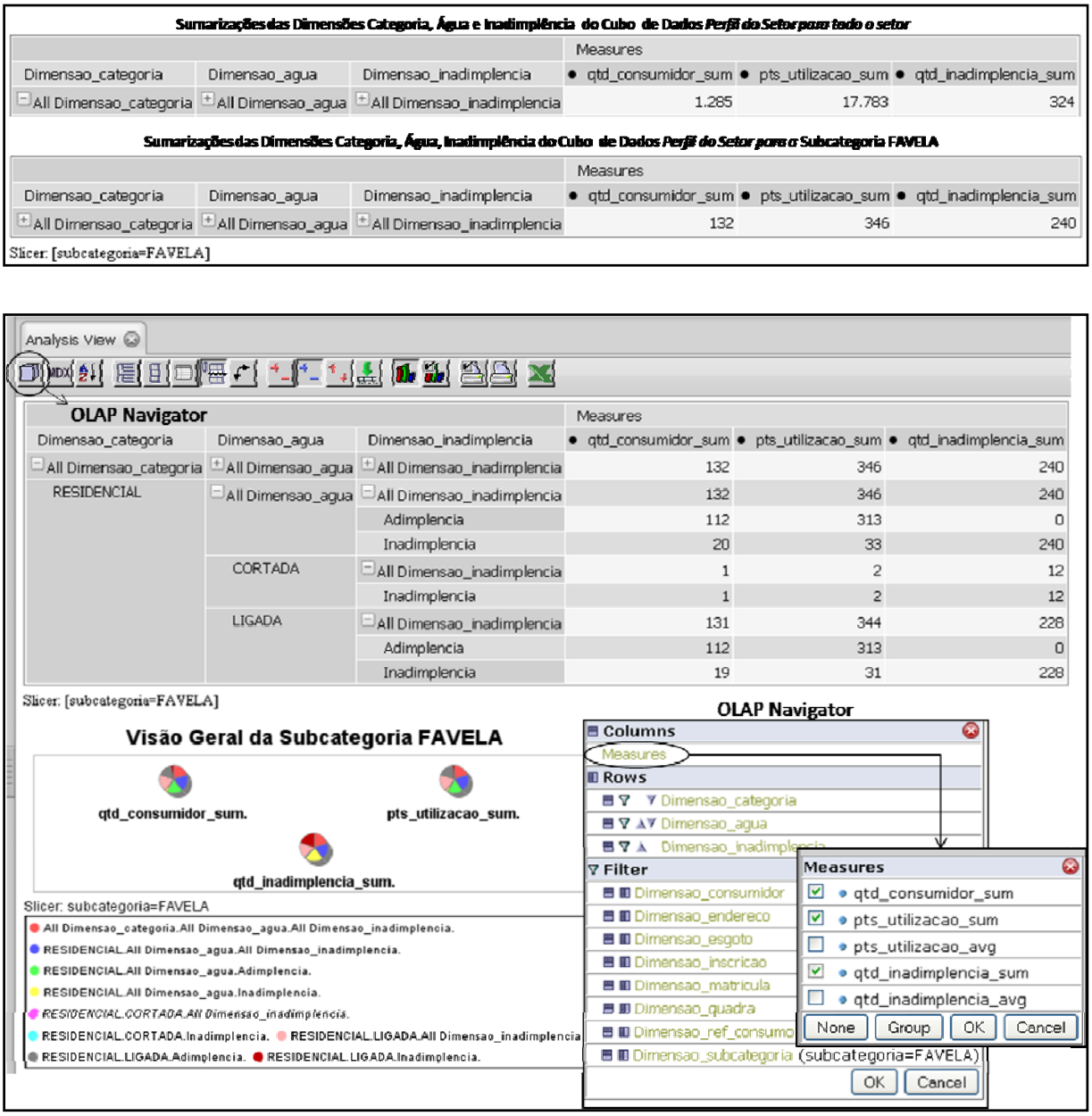



**Figura 3.10 - consulta sobre o perfil do consumidor de baixa renda quanto a inadimplência**

Esta consulta determina a quantidade de consumidores, pontos de utilização e quantidade de inadimplências da subcategoria FAVELA, associando-os com os agrupamentos das categorias de consumo (Comercial, Industrial, Público e Residencial), situações da ligação de água (Cortada, Ligada, Suprimida parcial e Suprimida total) e estado de inadimplência (Adimplência e Inadimplência) dos consumidores.

Conforme a Figura 3.10, o resultado da consulta informa que todos os consumidores da categoria FAVELA estão agrupados apenas pela categoria RESIDENCIAL e situação da água CORTADA e LIGADA.

Dos 1.285 consumidores de todo o setor 64, 132 estão na subcategoria FAVELA. Das 324 inadimplências de todo o setor, 240 encontram-se na subcategoria FAVELA, i.e., 74,1%. Na situação da água CORTADA há um consumidor com 12 inadimplências. Já na ligação da água LIGADA, há consumidores inadimplentes (19) e adimplentes (112). Dos 17.783 pontos de utilização de água de todo o setor, 346 pertencem a subcategoria FAVELA, sendo 33 de consumidores inadimplentes.

Em um ambiente empresarial, as consultas são realizadas todas as vezes que se pretende obter informações específicas de uma ou várias dimensões. Afinal, é uma ferramenta gerencial que proporciona a gestão e criação de relatórios *ad hoc.*

Foram realizadas diversas consultas ao *Data Warehouse* e várias manipulações e análises aos dois cubos de dados através da ferramenta *Pentaho Analysis View.* Nesta seção são discutidas algumas consultas e mostrada algumas das potencialidades que a ferramenta proporciona aos analistas.

Uma consulta SQL ao Esquema Constelação de Fatos[13] foi exemplificada na página 90. Esta consulta retornou as médias de FATURAMENTO das quadras (010, 015, 020, 030 e 050) agrupadas pela categoria de consumo COMERCIAL e SEMESTRES de referência.

O resultado da consulta foi apresentado na Figura 3.8 pela ferramenta de BD (p*gAdmin/postgreSQL),* projetada para trabalhar com dados relacionais. O mesmo exemplo da página 90 foi feito, porém, desta vez utilizando a ferramenta OLAP, que é apropriada e específica para manipular a estrutura dimensional de cubos de dados.

No exemplo da Figura 3.11 a operação OLAP *rotate*, também conhecida por *pivot,* foi executada sobre as dimensões. A Figura 3.11 também apresenta o *menu* de navegação da ferramenta *Pentaho Analysis View* com as operações OLAP *Slice and Dice* e gráficas (*Show Chart*), estas que favorecem a exploração dos cubos de dados de forma mais intuitiva.

---

[13] Parte do Esquema Constelação de Fatos contendo duas tabelas de fatos (Perfil do Setor e Perdas Aparentes) associadas às dimensões matrícula, quadra, categoria e referência de Consumo.

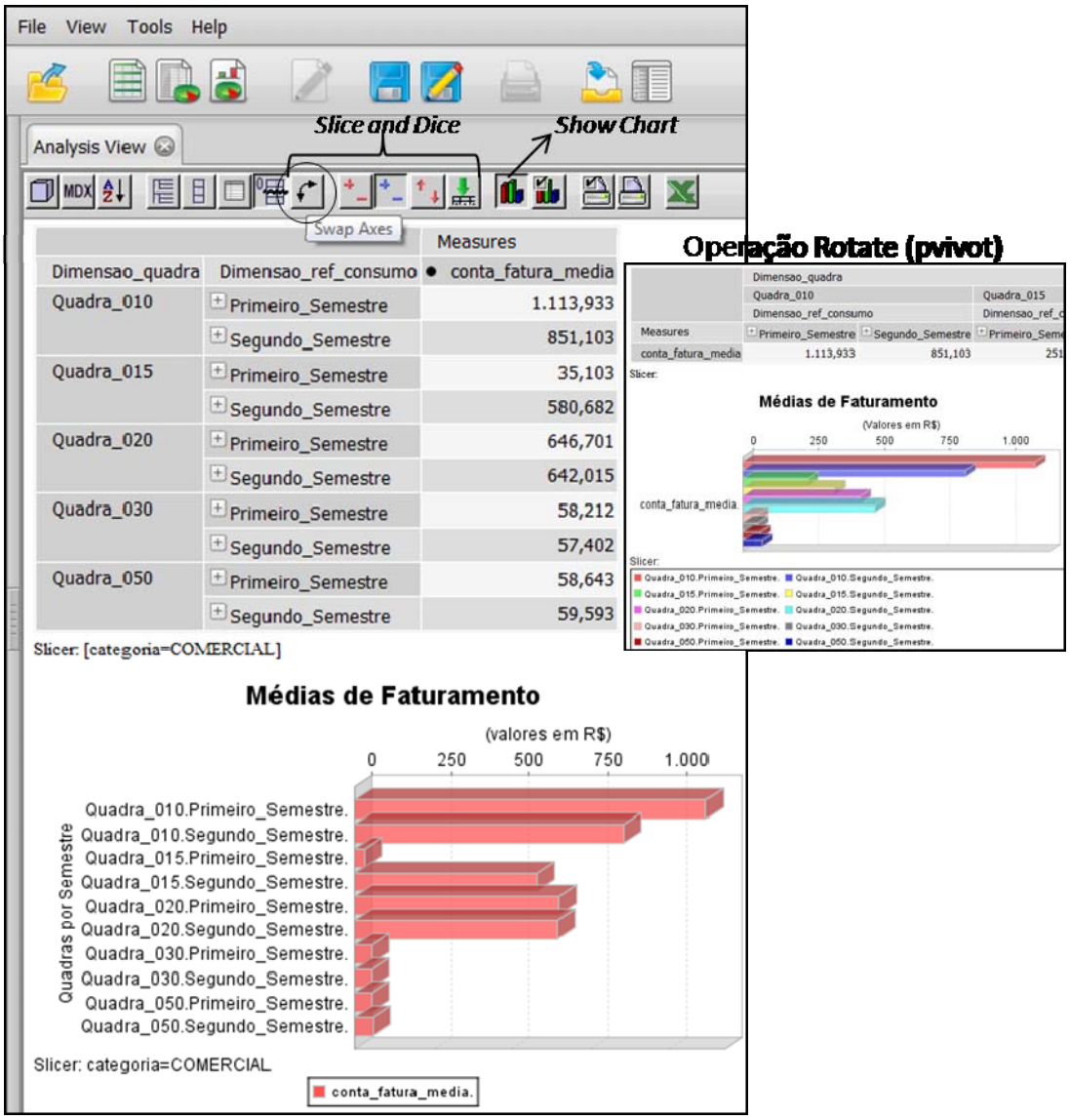



**Figura 3.11 - exemplo de consulta ao esquema constelação de fatos da Figura 3.7**

No capítulo2, seção 2.4.1 (página 44), foi apresentada a estrutura multidimensional do cubo de dados. O exemplo apresentado na Figura 3.12 retoma o que foi discutido na seção 2.4.1, mostrando a funcionalidade do *operador Cubo* sobre o cubo_perfil_setor através da ferramenta OLAP.

O *operador Cubo* representa a generalização n-dimensional da operação *group-by.* Os agrupamentos das dimensões categoria, água e esgoto podem gerar a computação de até 8 *group-by* (*cuboids*), sendo eles de três dimensões (3-D), duas dimensões (2-D), uma dimensão (1-D) e zero dimensão (0-D).

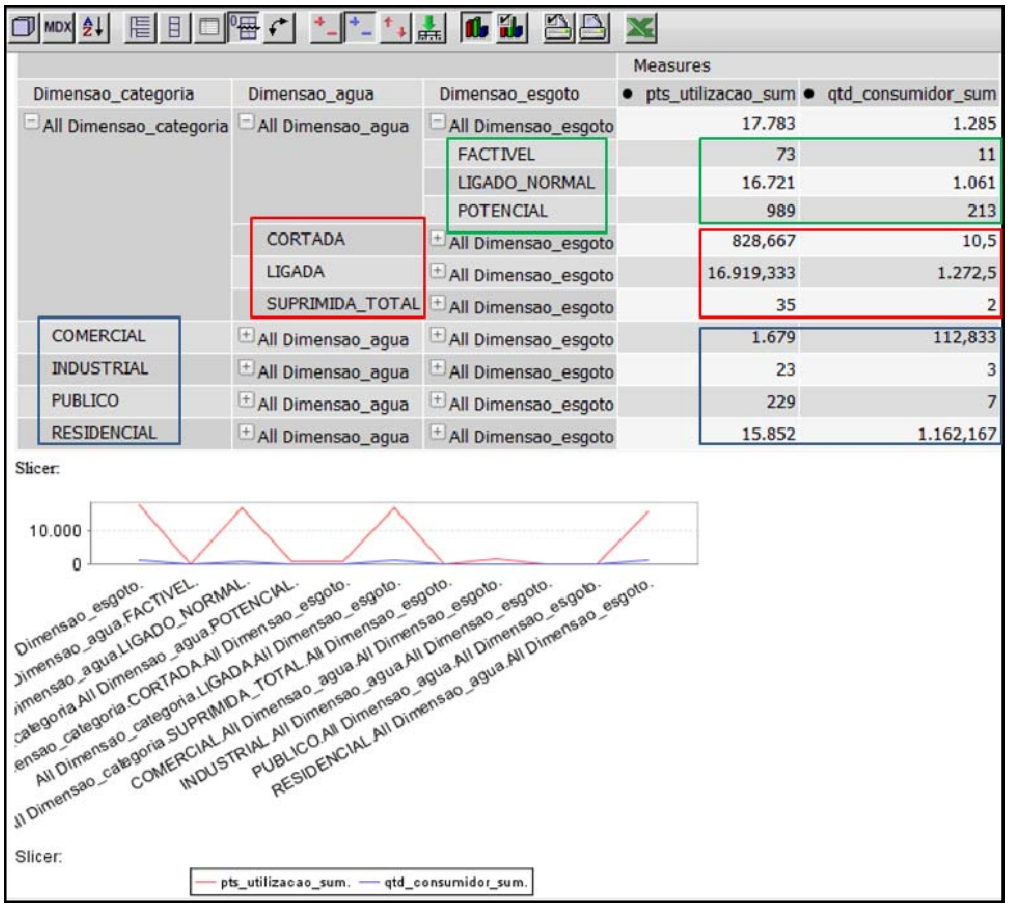



**Figura 3.12 - consulta ao cubo de dados "fato perfil do setor" (*cuboids* 1-D)**

No exemplo da Figura 3.12 as dimensões foram agrupadas individualmente, ou seja, representam *cuboids* de uma dimensão (1-D). No caso da Figura 3.13 a dimensão categoria foi associada à dimensão estado da ligação da água, o que representa o *cuboid* de duas dimensões (2-D). Desta forma, é possível verificar os pontos de utilização e quantidade de consumidores por categoria e estado da ligação da água em uma única consulta.

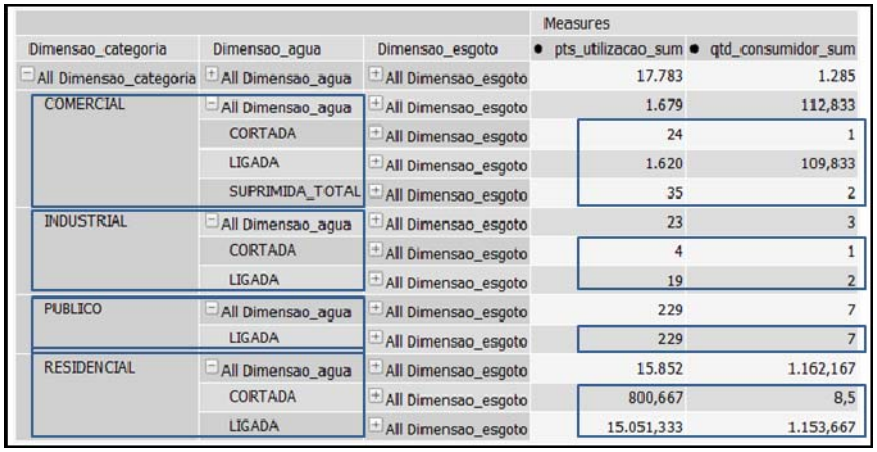



**Figura 3.13 - consulta ao cubo de dados "fato perfil do setor" (*cuboids* 2-D)**

## 3.3 PROCESSO DE EXTRAÇÃO DO CONHECIMENTO: FASE 2

### 3.3.1 Utilização do Data Mining

Esta seção discute uma das principais fases de extração de conhecimento em banco de dados que é o *Data Mining* (Mineração de Dados). Os algoritmos de *Data Mining* interpretam os dados a fim de produzir uma quantidade de padrões úteis, válidos e de fácil entendimento. Os resultados gerados podem ser usados para predições e têm por finalidade conduzir a tomadas de decisões inteligentes. O fator humano faz parte de todo o processo, por isso não pode ser uma ação totalmente automatizada.

Os algoritmos de mineração de dados favorecem a extração de informações de grandes volumes de dados e a análise estatística desses dados permite que se observem tendências e respostas para situações do tipo: encontrar e detectar as regiões onde a perdas aparentes são mais frequentes; determinar os tipos de perdas e anormalidades que ocorrem mais frequentemente nas diversas categorias de consumo; associar o perfil de consumidores inadimplentes etc.

O emprego de *Data Mining* para identificar as perdas aparentes proposto por este trabalho surgiu pelos seguintes motivos: a disponibilidade de grandes quantidades de dados; a existência de dados históricos armazenados ao longo de 1 ano; e a possibilidade de encontrar um perfil de comportamento típico.

A detecção de perdas aparentes bem como sua prevenção configura-se em um problema complexo. Mesmo que os históricos e o perfil de comportamento de um consumidor apresentem claros indícios de uso indevido da água − tais como: cortes nos ramais de água, muitas retificações nas contas, parcelamentos, grande número de inspeções no imóvel, ocorrências de mesma leitura, mais de 50% de variações no consumo e na fatura, entre outros – é importante que uma segunda investigação seja realizada. Afinal, nenhum teste é completo e 100% suficiente para se detectar a causa da perda aparente. Portanto, as informações geradas pelos sistemas de apoio à decisão também precisam ser compatibilizadas com outras variáveis do sistema, para que cobranças errôneas não sejam aplicadas aos consumidores.

Alguns pré-requisitos são essenciais para o sucesso da mineração de dados. Por isso foram construídos modelos baseados em metas preditivas e descritivas. As metas preditivas visam traçar o perfil do consumidor que dispõe de algum tipo de irregularidade no seu

consumo de água, assim como visam traçar o perfil do consumidor inadimplente. Diante das metas preditivas, tem-se, por exemplo, a utilização da tarefa de Classificação por Árvore de Decisão. Quanto às metas descritivas tem-se a utilização das Regras de Associação.

O procedimento de *Data Mining* se dará em duas etapas: a modelagem realizada e a construção das tarefas de mineração.

### 3.3.2 Modelagem Realizada

Inicialmente foi criado um BD com os dados extraídos da companhia de abastecimento de água, onde todos os atributos e instâncias foram inseridos em uma única tabela. Em seguida, foi projetado e desenvolvido o ambiente de *Data Warehouse* utilizando a modelagem dimensional do esquema Constelação de Fatos, útil e essencial para realização otimizada das consultas OLAP e para a seleção dos dados a serem minerados.

As instâncias do *Data Warehouse* foram modeladas de acordo com o esquema Constelação de Fatos e se resumem em dados do tipo Cadastrais, Relacionamento e Padrão de Comportamento. Os dados do tipo Cadastrais especificam cada cliente (nome, endereço, pontos de utilização, matrícula, etc.) e são praticamente estáticos, ou seja, se modificam pouco. Os dados do tipo Relacionamento correspondem ao relacionamento da empresa com o cliente (exemplo: vazão do hidrômetro instalado, qualidade e tempo do ramal, qualidade e tempo do hidrômetro, padrão da ligação, multas aplicadas, etc.). E os dados do tipo Padrão de Comportamento são do tipo: consumo de água mensal, percentual de variação de consumo (menor, maior e média), irregularidades, anormalidades e inadimplências, ou seja, o padrão de comportamento do consumidor perante a companhia.

As instâncias de comportamento são as mais adequadas para encontrar um padrão de comportamento que identifique perdas aparentes no sistema de abastecimento. Contudo, as outras não foram descartadas, visto que ajudam na interpretação dos resultados.

Após análises e experimentações de ferramentas de *Data Mining* (WITTEN, et al., 2005), observou-se que o software WEKA ajusta-se adequadamente na modelagem proposta por este trabalho, haja vista a variedade de funcionalidades para realizar classificação, associação, descoberta de sucessões, série temporais, agrupamento e regressão, além de suportar várias fontes de dados e os principais algoritmos de mineração indicados na literatura pelos especialistas.

Para dar início ao processo de *Data Mining*, a extração e o carregamento dos dados são realizados. Nesta etapa, antes de utilizar os algoritmos de mineração, o software WEKA permite configurar os dados através de várias funções de filtragens, dentre eles, normalizações, adições, compartilhamentos, junções, eliminações, conversões de tipos e formatos, funções estatísticas, etc. Os filtros normalmente são utilizados quando os dados são extraídos de ambientes transacionais. Neste trabalho, contudo, os dados foram extraídos do *Data Warehouse*, e por isso os filtros não precisaram ser utilizados, afinal os dados já se encontravam preparados no *Data Warehouse* e prontos para serem minerados pelo software WEKA.

### 3.3.3 Abordagem do Data Mining Aplicada aos Hidrômetros

Diante das perdas aparentes surgem questões de decisão a respeito do corte de fornecimento dos consumidores inadimplentes e fraudadores, bem como a instalação e substituição periódica dos hidrômetros.

A aplicação de *Data Mining* para apoio à tomada de decisão em relação aos medidores (hidrômetros) consiste em formar classes de decisões do tipo "Substituir o hidrômetro", caso se verifiquem anormalidades e irregularidades de consumo, faturamento e consequentemente arrecadação. Com isto, a companhia de abastecimento aumentaria o volume de água medido ($m^3$), diminuiria as perdas aparentes e consequentemente obteria a correta arrecadação pelo serviço prestado.

A troca de hidrômetros é proposta baseada na percentagem de arrecadação e volume medido. Os atributos da base de dados dispostos para o estudo de caso correspondem ao período de 12 meses. Desta forma, a abordagem consiste em comparar os seis primeiros meses de consumo com os seis últimos meses de consumo de cada consumidor. Ao se detectar que houve uma arrecadação do volume de água menor que 50% nos últimos seis meses, que corresponde a um decréscimo de 50% ou mais no volume de água, isto é, o hidrômetro está medindo a metade ou mais da metade do volume de água que o consumidor consumia nos seis primeiros meses, então o hidrômetro deverá ser verificado, visto que ele apresenta fortes indícios de comprometimento da medição.

Caso se confirme o inadequado uso do hidrômetro ele estará apto para ser substituído, e esta troca beneficiar a companhia de abastecimento de água. A utilização deste critério permitirá mais uma ação preditiva na substituição dos hidrômetros, sem o empirismo de

propor trocá-lo apenas por tempo de uso. Entretanto, nos casos em que não detectar diminuição na arrecadação nos últimos seis meses (segundo semestre) em relação ao primeiro semestre, então o hidrômetro não estará apto para ser vistoriado nem substituído, visto que se encontra funcionando corretamente e a troca acarretaria gastos financeiros desnecessários à companhia. E nos casos onde se detectar que a diminuição na arrecadação nos últimos seis meses em relação ao primeiro semestre está no intervalo de [-50 e -25%], então o hidrômetro precisa ser analisado mais detalhadamente. A Figura 3.14 apresenta os intervalos de aumento ou diminuição do faturamento no segundo semestre de consumo e a melhor ação a ser tomada.

O atributo "decisão" pertence à Dimensão Percentagem, conforme ilustra a Figura A.1 do APÊNDICE A. De acordo com a Figura 3.14, o atributo "decisão foi definido para identificar o comportamento e procedimento para tomada de decisão quanto à situação do hidrômetro de cada consumidor, a fim de propor medidas corretivas e preventivas no que diz respeito à correta medição e consumo de água.

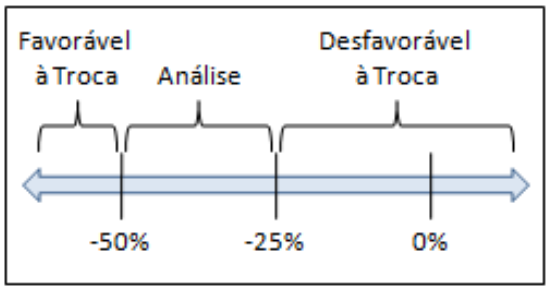



**Figura 3.14 - intervalos de valores percentuais do faturamento no último semestre**

Os procedimentos a serem realizados pelo atributo "decisão" são:

- Favorável à Troca (constatação): Realizar a verificação e posterior troca do hidrômetro se o faturamento do segundo semestre em relação ao primeiro semestre for menor que -50%. A não realização deste procedimento possivelmente indica perdas aparentes para a companhia de saneamento.
- Desfavorável à Troca (nenhuma ação): Não realizar a troca do hidrômetro se o faturamento do segundo semestre em relação ao primeiro semestre for maior que -25%. A situação se encontra estável, indicando apenas que houve aumento de consumo de água.
- Indiferente à Troca (em análise): Nenhuma ação deve ser realizada a priori caso o faturamento do segundo semestre em relação ao primeiro semestre estiver no intervalo de [-50% a -25%]. Contudo, nesta situação deve-se verificar o comportamento do

consumidor em um período maior, além de analisar outras variáveis como situação do imóvel (por exemplo, se ele encontra-se fechado), diminuição de moradores no imóvel, diminuição dos pontos de utilização de água, etc.

### 3.3.4 Construção das Tarefas de Mineração

Das várias tarefas de *Data Mining* definidas na literatura, duas delas se mostraram adequadas aos resultados a que se pretende obter com este pesquisa de detecção de perdas aparentes. As tarefas aplicadas foram: Classificação por Árvore de Decisão e Classificação Bayesiana (Aprendizado Supervisionado); e Geração de Regras de Associação (Aprendizado Não Supervisionado). Ambas foram detalhadas anteriormente nas seções 2.5.4.1 e 2.5.4.2.

#### 3.3.4.1 Classificação

Os algoritmos aplicados na Classificação dos dados fornecidos pela companhia de abastecimento de água CAGEPA foram o algoritmo ID-3, J4.8 e *NaiveBayes*. Todos eles fazem parte da biblioteca de algoritmos implementados pelo software de *Data Mining* WEKA. A utilização desses algoritmos ao estudo de caso será apresentada no capítulo 4.

A matriz de confusão faz parte dos algoritmos classificadores, i.e., da tarefa de classificação do *Data Mining*. A matriz oferece meios efetivos para a avaliação do modelo com base em cada classe. Cada elemento da matriz mostra o número de exemplos para os quais a classe verdadeira é a linha e a classe predita é a coluna. A diagonal principal da matriz (elementos (i,j), onde i = j) representa os acertos do modelo, enquanto os demais elementos representam os erros, discriminados para cada classe.

A Tabela 3.2 ilustra um exemplo da Matriz de Confusão para um problema com duas classes, denominadas Adimplente (classe positiva) e Inadimplente (classe negativa). Existem quatro possibilidades de acertos e de erros do classificador, são eles: VP, FN, VN e FP.

**Tabela 3.2 - matriz de confusão para a classificação com duas classes**

| | | Preditiva | | |
|---|---|---|---|---|
| | | Adimplente | Inadimplente | |
| Verdadeira | Adimplente | 278 (VP) | 3 (FN) | Erros |
| | Inadimplente | 15 (FP) | 4 (VN) | Acertos |

Verdadeiros positivos (VP) ocorrem quando os exemplos pertencem à classe Adimplente e foram preditos como pertencentes a essa mesma classe. Falsos negativos (FN)

ocorrem quando os exemplos pertencem à classe Adimplente e foram preditos como pertencentes à classe Inadimplente. Verdadeiros negativos (VN) ocorrem quando os exemplos pertencem à classe Inadimplente e foram preditos como pertencentes a essa mesma classe. Falsos positivos (FP) ocorrem quando os exemplos pertencem à classe Inadimplente e foram preditos como pertencentes à classe Adimplente.

A Taxa de Erro (Equação 3.2) e a Taxa de Acurácia (Equação 3.3) são as medidas de avaliação mais utilizadas para os modelos de classificação. A taxa de erro é a proporção de erros de predição sobre um conjunto de exemplos em que se conhece o valor do atributo meta. A acurácia e a taxa de erro são estimativas do percentual de acertos e de erros do classificador, respectivamente, na predição da classe de novos exemplos.

$$\text{Taxa de Erro} = \left(\frac{\text{FP} + \text{FN}}{\text{VP} + \text{FN} + \text{FP} + \text{VN}}\right) \times 100$$

**Equação 3.2 - taxa de erro para matriz de confusão com duas classes**

$$\text{Taxa de Acurácia (Precisão)} = 1 - \text{Taxa de Erro}$$

**Equação 3.3 - taxa de acurácia para matriz de confusão com duas classes**

**Fonte: (WITTEN, et al., 2005; HAN, et al., 2006).**

No exemplo da Tabela 3.2 existem 300 instâncias a serem mineradas. As instâncias classificadas corretamente foram VP (278 instâncias) e VN (4 instâncias), totalizando 282 instâncias (278 + 4). As instâncias classificadas incorretamente foram FP (15 instâncias) e FN (3 instâncias), totalizando 18 instâncias (15 + 3). Portanto, a Taxa de Erro corresponde a $(18 \div 300) \times 100 = 6\%$ e a Taxa de Acurácia corresponde a $(282 \div 300) \times 100 = 94\%$.

Através da Classificação por Árvores de Decisão observa-se que uma regra de classificação terá sempre no seu consequente uma resposta ao fato das condições satisfazerem ou não a uma determinada classe previamente definida. O uso desta tarefa para o setor de saneamento objetiva predizer informações, gerando dados futuros dos consumidores com risco de infringir a companhia de saneamento através de inadimplências e/ou fraudes, bem como identificar possíveis padrões nos processos da companhia. A meta de predição é a habilidade de elaborar cenários diferentes para antecipar certos resultados que possivelmente irão ocorrer, caso a taxa acurácia da regra seja elevada.

#### 3.3.4.2 Associação

Há um número muito grande de regras de associação encontradas ao aplicar a tarefa de associação em um *Data Warehouse.* Contudo, muitas dessas regras não são exploradas e não interessam ao processo analítico de descoberta de conhecimento. A fim de minimizar a geração de regras de associação desnecessárias, são introduzidas as duas medidas de interesse, o *suporte* e a *confiança*. O *suporte* indica a frequência com que uma regra aparece no *Data Warehouse* e a *confiança* indica o grau de acerto da regra. Estas medidas foram apresentadas em detalhes na seção 2.5.4.2.

Nem todas as regras geradas pelo *Data Mining* são consideradas relevantes para o processo de extração do conhecimento em banco de dados, visto que o especialista precisa interpretá-las no contexto em que o seu negócio está inserido e só depois aplicá-las, afinal o fator humano também faz parte do processo. Desta forma, o especialista do negócio precisa avaliar as regras para que o resultado seja aplicável na prática.

O algoritmo utilizado para gerar as regras de associação (Aprendizado Não-Supervisionado) com os dados fornecidos pela CAGEPA foi o *Apriori,* que faz parte da biblioteca de algoritmos implementados pelo software de *Data Mining* WEKA. A utilização do algoritmo *Apriori* ao estudo de caso será apresentada no capítulo 4.

## 3.4 CONSIDERAÇÕES FINAIS

Ao longo dos anos o processo de extração de conhecimento em banco de dados conseguiu atingir um ótimo grau de aperfeiçoamento e as experiências e pesquisas nesta área têm proporcionado resultados satisfatórios para as empresas que o adotam. Por sua vez, os processos computacionais que visam apoio à decisão em bases de dados se mostram adequados também para evidenciar usos indevidos de água em redes de abastecimento de água, tendo como principal motivação o combate às perdas aparentes, devido às crescentes irregularidades e anormalidades das ligações e medições de água,.

A ferramenta de criação dos cubos de dados (*Schema Workbench*), a ferramenta OLAP de análise e consulta ao *Data Warehouse* (*Analysis View*), e a ferramenta de mineração de dados (WEKA), proporcionaram a adequada e eficiente utilização do *Data Warehouse* e dos algoritmos de *Data Mining,* e juntas formaram o Sistema de Apoio à Decisão do setor estudado.

# CAPÍTULO 4

***Este capítulo apresenta a aplicação dos algoritmos de Data Mining ao estudo de caso, fazendo um comparativo entre os algoritmos do aprendizado indutivo supervisionado. Discute os resultados extraídos do ambiente de mineração WEKA, apresentado as demonstrações e análises referentes aos conhecimentos e padrões adquiridos da base de dados, o Data Warehouse, do setor de saneamento.***

## 4 DATA MINING APLICADO AO ESTUDO DE CASO

Cada Sistema de Apoio à Decisão possui suas peculiaridades quanto ao objeto de estudo. Neste trabalho o objeto de estudo é o setor de saneamento e a metodologia proposta foi desenvolvida para ser extensível e aplicável também a outros segmentos (e.g., saúde, educação, transporte, etc.) que buscam informatizar e otimizar seus processos de descoberta de padrões e conhecimento. A metodologia para desenvolver o SAD, utilizando tecnologias de Banco de Dados e da inteligência de negócio, foi discutida nas seções anteriores. A fim de realizar a análise detalhada do setor 64 do sistema de abastecimento urbano da Paraíba, este capítulo discute como foi aplicada a Mineração de Dados sobre os dados contidos no *Data Warehouse*.

### 4.1 ETAPA DE DATA MINING

Na etapa de *Data Mining* utilizou-se o programa computacional WEKA, versão 3. Os dados operacionais do setor analisado foram extraídos do SGBD IBM-DB2 e fornecidos pela CAGEPA em um arquivo de texto padrão. Para o desenvolvimento do *Data Warehouse Departamental* utilizou-se o SGBD *PostgreSQL 8.3.1* e a ferramenta *PgAdmin 1.8* para reorganizar os dados em tabelas de dimensões e em tabelas de fatos, utilizando o esquema dimensional Constelação de Fatos.

Os dados necessários para aplicação dos modelos de *Data Mining* "Perfil do Setor" e "Perdas Aparentes" foram extraídos do *Data Warehouse* e exportados para um arquivo no formato ".csv". Este por sua vez foi reestruturado e transformado em um arquivo do tipo ".arff", que é a extensão padrão utilizada pelo software WEKA para realização das tarefas de *Data Mining*.

Durante a execução da mineração para detecção de perdas aparentes, foram aplicados os dois tipos de aprendizado indutivo: o aprendizado supervisionado (método preditivo) e o aprendizado não-supervisionado (método descritivo). Para o aprendizado supervisionado utilizou-se a tarefa de Classificação e as técnicas de Árvores de Decisão e Classificação Bayesiana através dos algoritmos *ID-3, J4.8* e *NaiveBayes*. As técnicas de Árvore de Decisão e Classificação Bayesiana foram escolhidas neste trabalho por serem de fácil percepção a análise e visualização dos resultados. Para o aprendizado não-supervisionado utilizou-se a técnica de Regras de Associação através do algoritmo *Apriori.*

Os imóveis atendidos pela companhia de abastecimento de água estão distribuídos por localidade, setor, quadra e lote. Desta forma, alguns agrupamentos foram formados por quadras de acordo com o número da quadra que imóvel pertence. O número da quadra está presente no número de inscrição do consumidor. Por exemplo: o consumidor João possui número de inscrição "001.64.350.0174". Desta forma, ele pertence à quadra 350 do setor 64 (bairro Miramar) da localidade 001 (cidade de João Pessoa).

Na fase de extração e carga dos dados, os atributos e instâncias do modelo são carregados pela ferramenta *Explorer,* do software WEKA, para que os algoritmos de classificação e de associação sejam aplicados. Por fim, após a mineração dos dados, são realizadas as comparações quanto à taxa de acurácia e a taxa de erro entre os resultados dos algoritmos e modelos minerados.

### 4.1.1 Software de Data Mining: WEKA

O *Waikato Environment for Knowledge Analysis* (WEKA) é um software de código aberto, distribuído sob *GNU General Public License* , que implementa vários algoritmos de *Data Mining.* Foi desenvolvido na linguagem de programação *Java* pelos pesquisadores da Universidade de Waikato na Nova Zelândia. Através do software WEKA é possível descobrir vários tipos de padões de comportamento e conhecimento dos dados, visto que ele dispõe de ferramentas de exploração de dados, tais como regras de associação, classificação, regressão, agrupamento e visualização de dados.

O WEKA oferece uma interface intuitiva e uniforme para diferentes algoritmos de aprendizado de máquina. A interface é composta por quatro módulos: *SimplesCLI*, *Explorer*, *Experimenter* e *KnowledgeFlow*. O *Explorer* é o módulo mais comumente utilizado, pois enquadra separadamente as etapas de pré-processamento (filtros), *Data Mining* (associação,

agrupamento, regressão e classificação) e pós-processamento (apresentação e avaliação de resultados). O módulo *Explorer*, versão 3.6 de 2008, foi utilizado para aplicação do *Data Mining* deste estudo de caso.

Os principais arquivos de dados utilizados como entrada pelo software WEKA são: arquivo no formato "arff" (*Attribute Relation Format File*), arquivo ".csv", arquivo URL no formato "arff"; e tabelas de banco de dados via JDBC. A sintaxe "arff" possui suporte somente aos tipos de dado numérico e nominal. O acesso aos dados pela ferramenta de *Data Mining* se deu pela execução de consultas ao *Data Warehouse.* Os dados retornados das consultas foram exportados para um arquivo do tipo ".csv", que por sua vez foi transformado em um arquivo estruturado para o WEKA, do tipo ".arff".

A estrutura do arquivo "arff" é composta de três partes: Relação, Atributos e Dados. A relação (*@relation*) é a primeira linha do arquivo, e deve conter a palavra-reservada @relation seguida de uma palavra-chave que identifique a tabela/relação ou a tarefa que está sendo analisada. Os atributos (*@attribute*) formam um conjunto de linhas onde cada uma contem a palavra-reservada @attribute seguida do nome do atributo e do seu tipo, que pode ser nominal ou numérico. A última parte do arquivo "arff" corresponde ao conjunto de instâncias de dados (*@data*), inseridos logo após a definição dos atributos.

## 4.2 RESULTADOS E DISCUSSÕES

Visando identificar com mais precisão o perfil do consumidor, o perfil dos imóveis e as abordagens referentes às perdas aparentes pertencente ao setor 64, bairro de Miramar, no que diz respeito às características envolvidas no serviço prestado pela companhia de abastecimento de água e esgoto da Paraíba, foram desenvolvidos dois modelos de *Data Mining*. São eles: Modelo do Perfil do Setor (corresponde ao perfil do consumidor e perfil dos imóveis) e o Modelo das Perdas Aparentes.

O software WEKA fornece funcionalidades para realização de análises prévias dos dados (pré-mineração) e obtenção de informações relevantes para o apoio à decisão, por meio de representações gráficas, antes mesmo de aplicar os algoritmos de *Data Mining*.

### 4.2.1 Pré-Mineração do Modelo Perfil do Setor

Os atributos dimensionais do *Data Warehouse* requeridos como entrada nos algoritmos de *Data Mining* para geração e diagnóstico deste primeiro modelo foram:

matrícula do consumidor, quadra, situação da ligação de água, situação da ligação do esgoto, categoria de consumo, subcategoria de consumo, inadimplência do consumidor e referência de consumo por semestre.

Os oito (8) atributos mencionados e suas instâncias para identificação do Perfil do Setor foram extraídos do ambiente de *Data Warehouse* através da tabela de fatos *fato_perfil_setor* e suas dimensões.

A Figura 4.1 e Figura 4.2 apresentam a visão geral dos atributos do modelo Perfil do Setor. Ao todo foram analisadas 1.285 matrículas de consumidores, 79 quadras, 2.583 instâncias e 8 atributos. O gráfico (A) da Figura 4.1 apresenta as quadras e a quantidade de registro (instâncias) por quadra. Por exemplo, a Quadra_390 é a que possui o maior número de instâncias, 126 instâncias, representando a quadra onde há o maior número de consumidores (63 consumidores)[14].

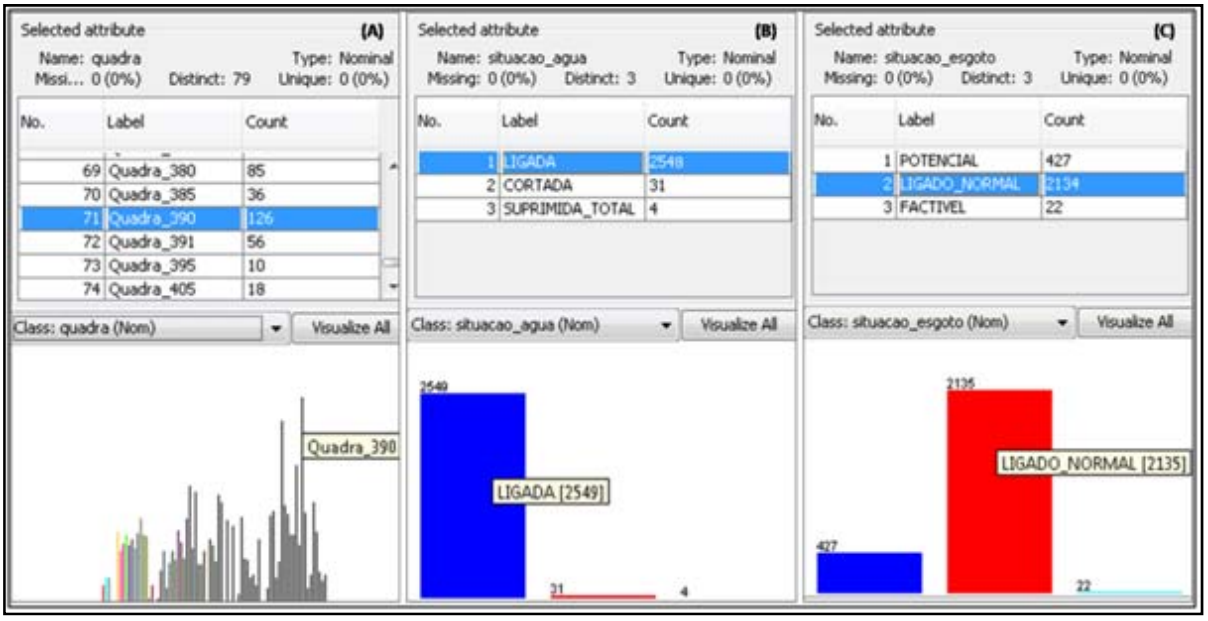



**Figura 4.1 - visão geral dos atributos do modelo perfil do setor. (A-C)**

Os gráficos (B) e (C) da Figura 4.1 ilustram os três tipos de classificação para situações da água e esgoto, sendo os tipos "ligada" e "ligado normal" os mais encontrados na base de dados, com 2.549 e 2.135 instâncias respectivamente. Existem 1.273 consumidores

[14] Os valores representados nos gráficos da Figura 4.1 e Figura 4.2 correspondem ao número de instâncias (2.583 instâncias). Já os valores referentes à quantidade de consumidores por atributo foram obtidos através da análise dos dados pela ferramenta OLAP.

com situação da ligação de água "ligada", 10 com ligação "cortada" e 2 com ligação "suprimida_total". Quanto à situação do esgoto, 1.061 consumidores estão com ligação do tipo "ligado_normal", 213 do tipo "potencial" e 11 com ligação "factível".

O gráfico (D) da Figura 4.2 ilustra as quatro categorias de consumo, com predominância no setor 64 da categoria "residencial" (2.337 instâncias), dispondo de 1.162 consumidores, seguida pela categoria "comercial" com 113 consumidores. O gráfico (E) da Figura 4.2 ilustra o atributo subcategoria, sendo "casa" a subcategoria mais predominante no setor (1965 instâncias), com 976 consumidores, seguida pela subcategoria "favela" (264 instâncias), com 132 consumidores.

De acordo com o gráfico (F) da Figura 4.2, 54 instâncias registram inadimplências e 2.530 registram adimplências, sendo 1.258 consumidores dispostos como adimplentes e 27 inadimplentes.

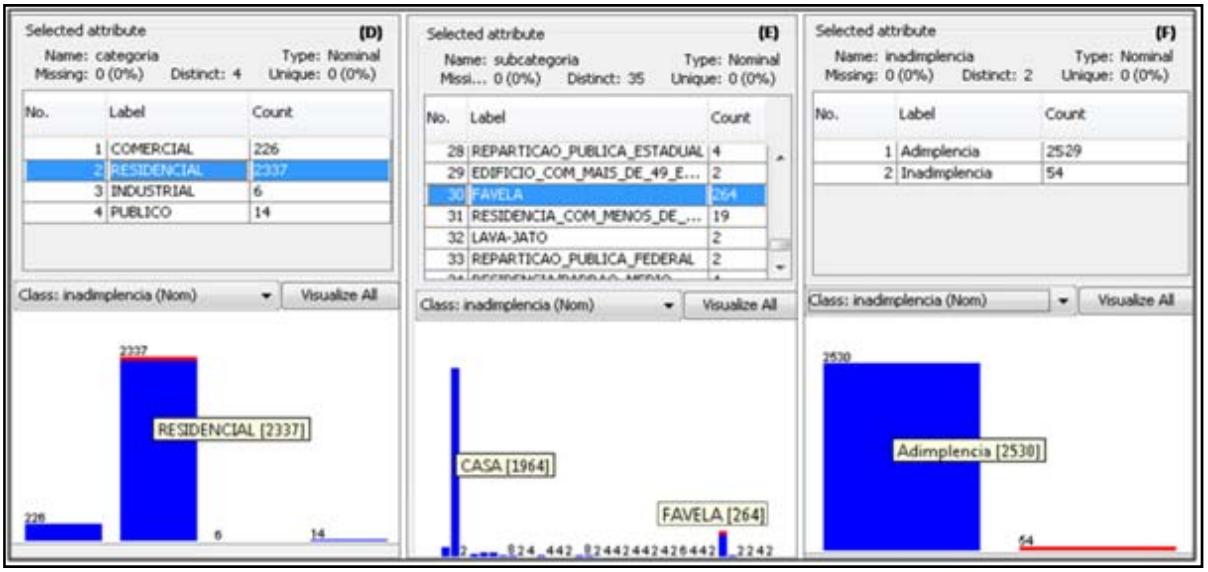



**Figura 4.2 - visão geral dos atributos do modelo perfil do setor. (D-F)**

O atributo inadimplência foi utilizado como *atributo classe*[15], e desta forma todos os atributos do modelo estabelecem relação direta com este atributo. A representação gráfica de cada atributo associado ao *atributo classe* (inadimplência) encontra-se destacado em vermelho na Figura 4.3 e Figura 4.4.

[15] O atributo classe corresponde ao atributo que irá se relacionar com todos os demais atributos do modelo a ser minerado. De acordo com o arquivo "arff", o último atributo antes do @data é sempre considerado o *atributo classe* padrão.

O gráfico (A) da Figura 4.3 ilustra as 79 quadras relacionando-as com o atributo inadimplência dos consumidores. A maior quantidade de inadimplências (96 das 324 totais) foi encontrada na quadra 415, seguida pelas quadras 410 e 120 que possuem 48 inadimplências cada uma. As informações referentes à quantidade de inadimplências por quadra foram obtidas com ferramenta OLAP.

O gráfico (B) da Figura 4.3 ilustra que os consumidores com situação da água "ligada" são os que mais possuem inadimplências (240) perante a companhia, seguido por situação de água "cortada" (60 inadimplências). O gráfico (C) da Figura 4.3 ilustra que de todos os consumidores inadimplentes, nenhum deles se encontra inseridos na situação de esgoto "factível" ou "potencial".

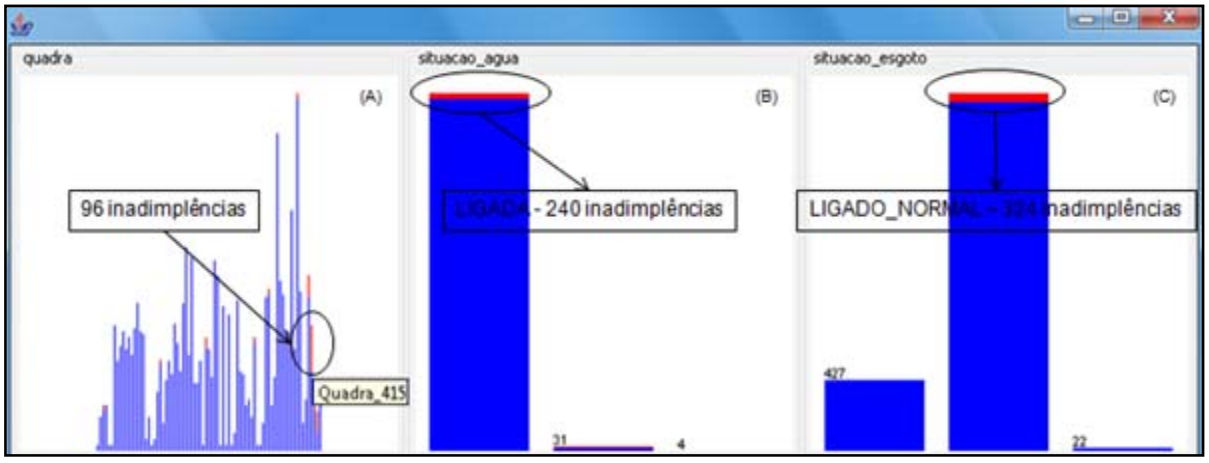



**Figura 4.3 - visão geral do perfil do setor 64 quanto à inadimplência. (A-C)**

De acordo gráfico (D) da Figura 4.4, a categoria "residencial" possui o maior índice de inadimplências (276), seguido pela categoria "comercial" com 36 inadimplências. A categoria "público" possui 100% de adimplência.

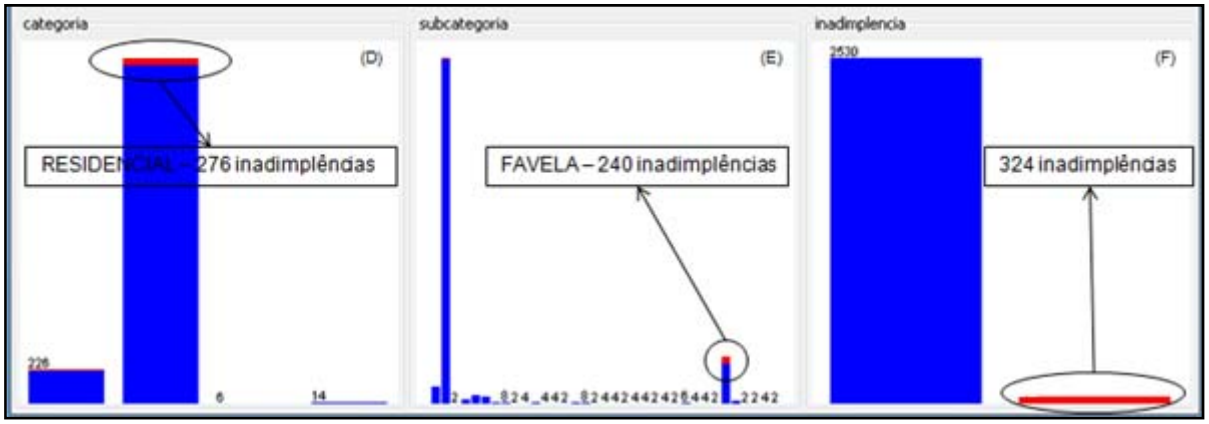



**Figura 4.4 - visão geral do perfil do setor 64 quanto à inadimplência. (D-F)**

O gráfico (E) da Figura 4.4 relaciona a quantidade de inadimplência com as subcategorias do setor, sendo a subcategoria "favela" a que possui maior quantidade de inadimplências (240), seguido pela subcategoria casa, com 36 inadimplências.

A Figura 4.4 (F) apresenta o *atributo classe*, e este não é associado a nenhum outro atributo. A informação obtida por meio deste gráfico é a quantidade total de inadimplência do setor, isto é, 324 inadimplências detectadas no período de um ano.

### 4.2.2 Pré-Mineração do Modelo Perdas Aparentes

As informações referentes às perdas aparentes foram trabalhadas segundo as abordagens envolvidas com a micromedição (medição dos hidrômetros) do setor 64. Os atributos e instâncias para identificação das Perdas Aparentes foram extraídos do ambiente de *Data Warehouse* através da tabela de fatos *fato_perda_aparente* e suas dimensões.

Os atributos dimensionais do *Data Warehouse* requeridos como entrada nos algoritmos de *Data Mining* para geração e diagnóstico deste segundo modelo foram: matrícula do consumidor, quadra, anormalidade, capacidade do hidrômetro em $m^3$, tipo do hidrômetro, ano de fabricação do hidrômetro, inadimplência do consumidor, referência de consumo por semestre, médias de consumo de água, médias dos valores da conta de água, consumo baseado na estrutura tarifária, indicador de medidor (hidrômetro instalado), período de instalação do hidrômetro e decisão de substituição do hidrômetro.

A Figura 4.5 e Figura 4.6 mostram a representação gráfica gerada pela ferramenta WEKA de alguns atributos do modelo Perda Aparente. Ao todo foram analisadas 1.285 matrículas, 79 quadras, 3.523 instâncias e 14 atributos.

Todos os gráficos da Figura 4.5 estão associados ao *atributo classe decisão*, sendo a cor azul correspondente ao tipo "desfavorável_à_troca", a cor vermelha correspondente ao tipo "favorável_à_troca" e a cor verde correspondente ao tipo "análise_mais_detalhada", conforme ilustra o gráfico (A) da Figura 4.5. Das 3.523 instâncias, 3.141 são classificadas como "desfavorável_à_troca", 306 classificadas como "análise_mais_detalhada" e 76 "favorável_à_troca". A idéia acerca do atributo decisão foi discutida em detalhes na seção 3.3.3, página 100.

O gráfico (B) da Figura 4.5 ilustra as 24 anormalidades presentes no setor 64 e gera-nos como informação que a maioria das instâncias do modelo (2.448) não possui

anormalidades. O tipo de anormalidade mais encontrada no setor é "imóvel_ou_portão_fechado", com 420 instâncias. Este tipo de anormalidade impossibilita o leiturista de obter a correta medição de consumo de água.

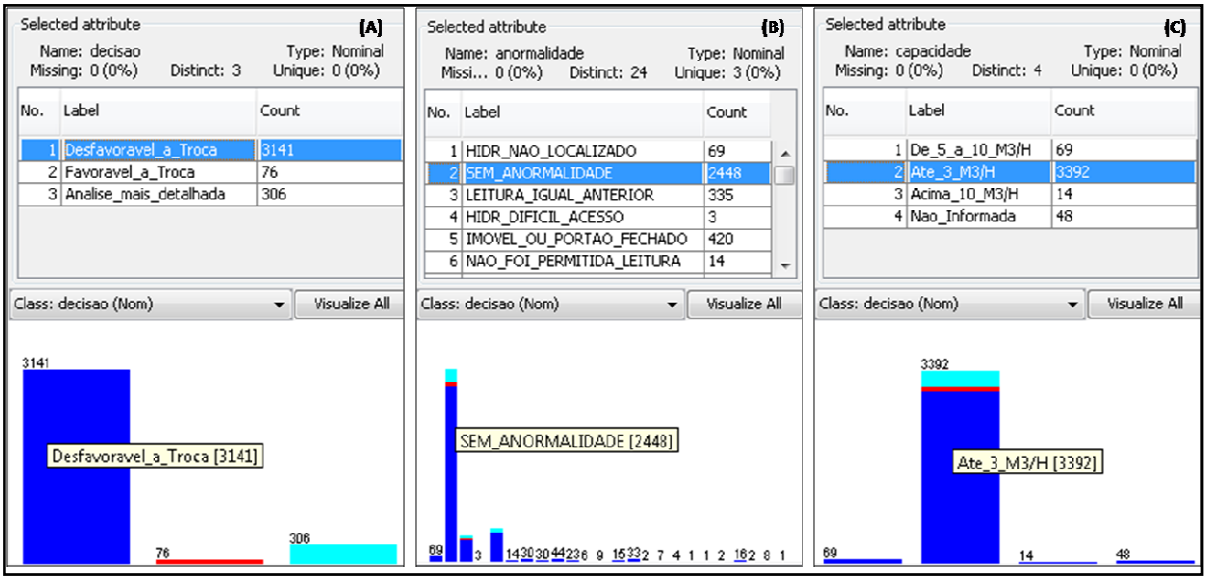



**Figura 4.5 - atributos do modelo perdas aparentes associados ao *atributo classe* decisão. (A-C)**

O gráfico (C) da Figura 4.5 e os Gráficos (D), (E) e (F) da Figura 4.6, correspondem, respectivamente, aos atributos capacidade do hidrômetro, tipo do hidrômetro, ano de fabricação do hidrômetro e data de instalação do hidrômetro. Nos quatro atributos verifica-se a presença do tipo "não_informada(o)", este por sua vez, possui valor 48 nos gráficos (C), (D) e (F), e indica que não há hidrômetro instalado, ou seja, o atributo indicador de medidor nessas instâncias é sempre do tipo "sem_medidor".

O atributo ano de fabricação do hidrômetro representado no gráfico (E) da Figura 4.6 indica que os hidrômetros do setor 64 não são considerados velhos, afinal a maioria deles (1.252 instâncias) possui ano de fabricação entre 2004 e 2008 e apenas 17 instâncias estão com ano de fabricação entre 1984 e 1988. Esta informação é reforçada com os dados do gráfico (F), onde há a indicação que das 3.523 instâncias analisadas, 2.329 possuem período de instalação do hidrômetro entre 3 e 9 anos, o que é considerado um bom período de uso, visto que a média de vida útil dos hidrômetros é de aproximadamente 5 anos. Como a capacidade de vazão da maioria dos hidrômetros do setor 64 é de até $3m^3$ (3.392 instâncias, conforme indica o gráfico (C) da Figura 4.6), então os desgastes no equipamento de medição são bem menores se comparados a um hidrômetro de grande capacidade, desta forma, o

tempo de utilização e vida útil do equipamento aumenta. Apenas 14 instâncias estão associadas à capacidade de vazão acima de 10 $m^3$.

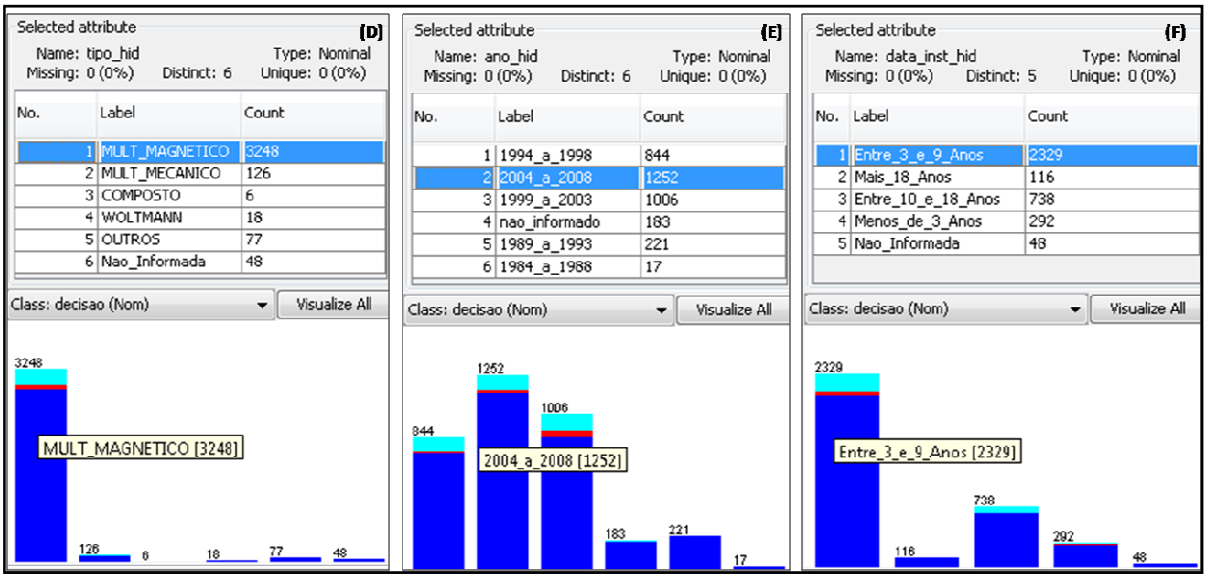



**Figura 4.6 - atributos do modelo perdas aparentes associados ao *atributo classe* decisão. (D-F)**

As 3.523 instâncias estão distribuídas proporcionalmente entre os dois períodos de referência, conforme ilustra o gráfico (G) da Figura 4.7 através do atributo referência de consumo por semestre.

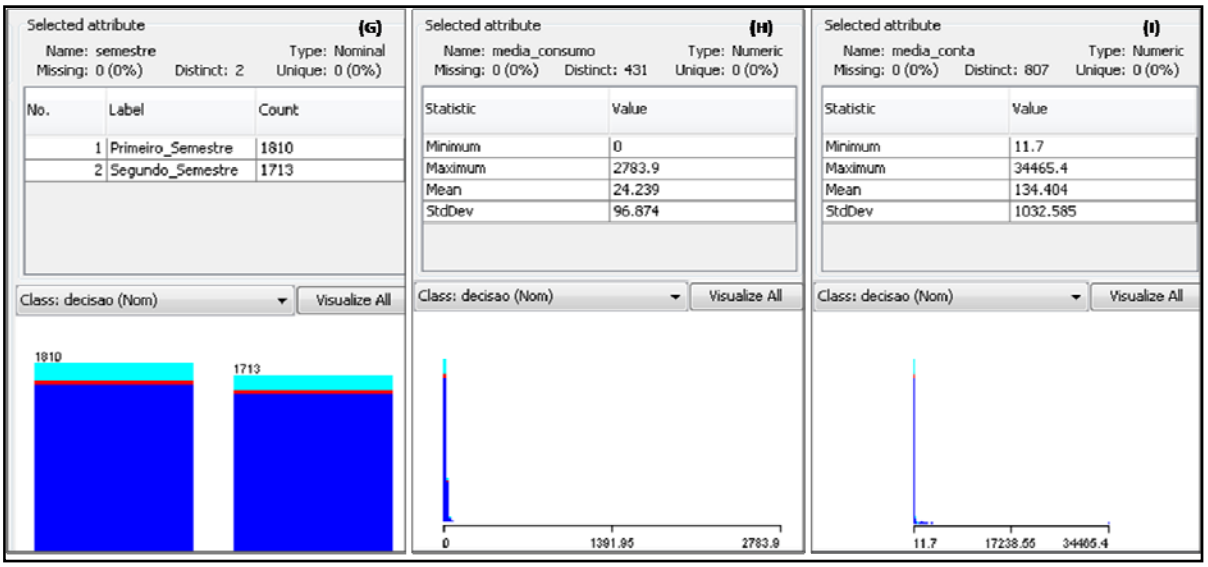



**Figura 4.7 - atributos do modelo perda aparente associados ao *atributo classe* decisão. (G-I)**

Os gráficos (H) e (I) da Figura 4.7 representam os atributos média de consumo de água por consumidor em $m^3$ e média da conta de água em reais. As curvas de tendência dos dois gráficos são similares, afinal o valor da conta e do consumo de água é diretamente

proporcional. A média de consumo de água mínimo[16] por consumidor em todo o setor foi 0 $m^3$ e o valor máximo foi de 2.783,9 $m^3$. Já o valor mínimo da média de faturamento por consumidor foi de R$ 0,00 e máximo de R$ 34.465,4.

De acordo com o Gráfico (J) da Figura 4.7, o consumo médio baseado na estrutura tarifária da CAGEPA[17] permite-nos conhecer o perfil tarifário mais comumente encontrado no setor 64. Tal perfil foi encontrado e corresponde aos consumidores da categoria residencial que consomem até 10 $m^3$/mês de água (1.141 instâncias), seguido novamente pela categoria residencial que consome entre 10 e 20 $m^3$/mês de água (1.017 instâncias).

O gráfico (L) da Figura 4.8 corresponde ao atributo indicador de medidor, onde a grande maioria do setor possui hidrômetro em seu imóvel (3.475 instâncias).

O último gráfico analisado refere-se ao atributo inadimplência, gráfico (M) da Figura 4.8, nele podemos verificar que das 3.523 instâncias, apenas 94 tem inadimplência.

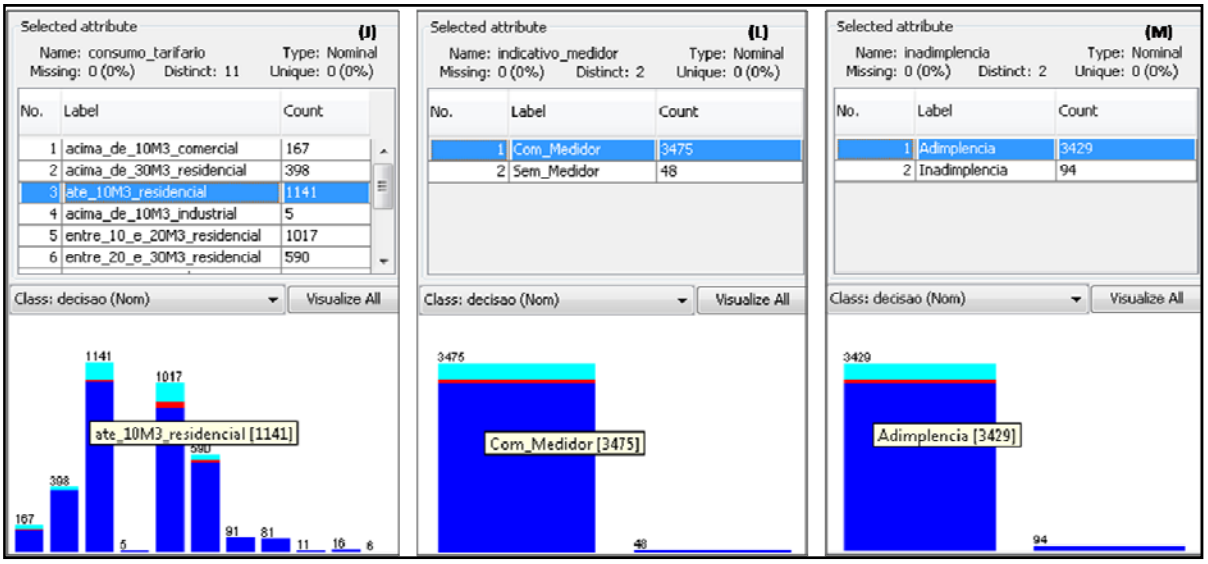



**Figura 4.8 - atributos do modelo perdas aparentes associados ao *atributo classe* decisão. (J-M)**

## 4.3 INTERPRETAÇÃO E AVALIAÇÃO DOS RESULTADOS

Através dos modelos de *Data Mining* propostos (Perdas Aparentes e Perfil do Setor) foram realizadas comparações entre quatro algoritmos implementados pelo software WEKA. Para a tarefa de Classificação (Aprendizado Supervisionado) foram utilizados os algoritmos

[16] Consumo de água igual a zero indica que o hidrômetro se encontra quebrado/parado.
[17] Informações sobre a estrutura tarifária: <http://www.cagepa.pb.gov.br/v4/informacoes_tarifas.php>

*ID-3, J4.8 e NaiveBayes*. E para a tarefa de Associação (Aprendizado Não Supervisionado) foi utilizado o algoritmo *Apriori*[18], conforme ilustra a Figura 4.9.

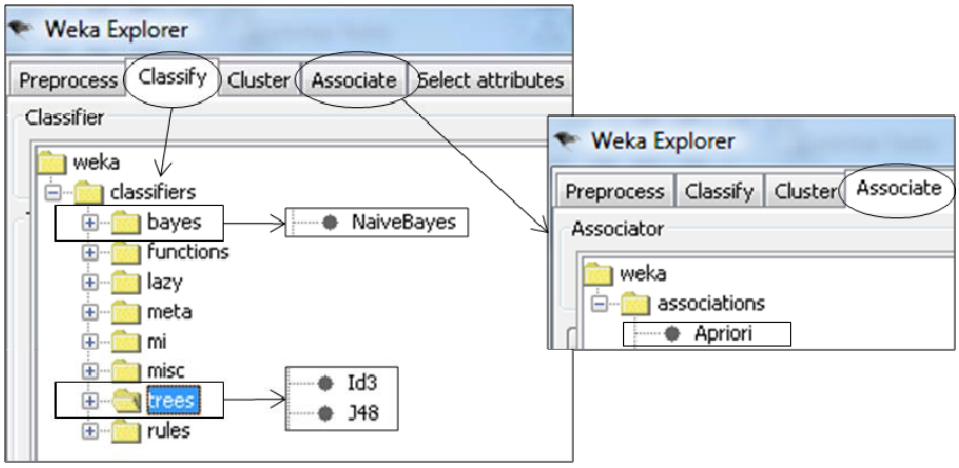



**Figura 4.9 - seleção dos algoritmos de *data mining* pela ferramenta WEKA**

Os algoritmos ID-3, J4.8 e NaiveBayes foram comparados entre si quanto à facilidade de utilização, taxa de acurácia e erro, visualização gráfica, tempo de processamento (desempenho). As comparações foram realizadas tanto para o modelo Perfil do Setor quanto para o modelo Perdas Aparentes. Já o algoritmo *Apriori* foi utilizado com o objetivo de gerar as 10 melhores regras para o modelo Perfil do Setor e Perdas Aparentes, isto é, os padrões que descrevem o setor 64.

Além das comparações computacionais entre os algoritmos, também serão apresentados alguns dos conhecimentos descobertos pelos mesmos. Tais conhecimentos gerados pela ferramenta objetivam fornecer tomadas de decisões para minimização das perdas aparentes do setor de abastecimento de água.

A Tabela B.1 do APÊNDICE B mostra a estrutura do arquivo "arff" para o modelo Perfil do Setor, que foi utilizada como *input* para execução do *Data Mining,* através das tarefas de Classificação e Associação. Na Tabela B.2 do APÊNDICE B, consta a estrutura do arquivo "arff" para o Modelo Perdas Aparentes utilizada como *input* para execução do *Data Mining* através das tarefas de Classificação e Associação.

[18] Apenas os algoritmos ID-3 e Apriori não dão suporte ao tipo de dado numérico.

O funcionamento dos algoritmos foi explanado e exemplificado no capítulo 2, nas seções 2.5.5.1b); 2.5.5.1c); 2.5.5.2a) e 2.5.5.3a). Nesta seção serão apresentados os resultados obtidos pela execução dos quatro algoritmos selecionados.

No Aprendizado Supervisionado, além dos dados de entrada, é necessário também definir o *atributo classe* a ser utilizado pelos três algoritmos (*ID-3, J4.8 e NaiveBayes*). Para o modelo Perfil do Setor o atributo classe selecionado foi INADIMPLÊNCIA e para o modelo Perdas Aparentes o *atributo classe* selecionado foi DECISÃO. A escolha desses atributos se deu por eles serem os mais representativos de cada modelo.

### 4.3.1 Execução do Data Mining: Modelo Perfil do Setor

O algoritmo ID-3 suporta apenas o tipo de dado nominal, desta forma, dos 8 (oito) atributos de entrada discutidos na seção 4.2.1, o atributo numérico matrícula do consumidor precisou ser eliminado. Os algoritmos J4.8 e NaiveBayes dão suporte ao tipo de dado numérico, contudo o atributo matrícula também foi eliminado como entrada de dados para esses algoritmos, visto que para realizar as comparações entre os três algoritmos, as instâncias de entrada precisam ser iguais. O atributo quadra também não foi considerado como entrada para mineração dos dados, visto que o objetivo não foi a análise das quadras especificamente e sim de todo o setor. De modo que para o modelo Perfil do Setor foram considerados 6 atributos de entrada, são eles: situação da ligação de água, situação da ligação do esgoto, categoria de consumo, subcategoria de consumo, inadimplência do consumidor e referência de consumo por semestre).

#### 4.3.1.1 Algoritmo ID-3

Ao executar o algoritmo ID-3 com as instâncias do modelo Perfil do Setor, o *software* WEKA gerou as informações da mineração, conforme mostra na Tabela 4.1 (apenas as informações mais relevantes são apresentadas). Ao todo foram utilizadas 2.583 instâncias de treinamento para a classificação, sendo 2.533 instâncias classificadas corretamente (taxa de acurácia[19] 98,06%), 46 instâncias classificadas incorretamente (taxa de erro 1,78%) e 4 instâncias não classificadas (0,16%). A taxa de acurácia de 98,06% indica alta precisão na

[19] As equações para obtenção dos valores da taxa de erro e acurácia foram definidas na seção 3.3.4 (página 101).

classificação, refletindo resultados confiáveis e satisfatórios sobre os dados de treinamento do modelo minerado. O tempo de processamento foi de 0.04 segundos.

**Tabela 4.1 - algoritmo ID-3 aplicado ao modelo perfil do setor**

| |
|---|
| === Run information ===<br>Scheme: weka.classifiers.trees.Id3<br>Relation: modelo_perfil_setor-weka<br>Instances: 2583<br>Attributes: 6<br>Time taken to build model**: 0.04 seconds**<br>=== Summary ===<br>Correctly Classified Instances 2533 98.06 %<br>Incorrectly Classified Instances 46 1.78 %<br>UnClassified Instances 4 0.1549%<br>=== Confusion Matrix ===<br>a b ← classified as<br>2523 2 \| a = Adimplencia<br>44 10 \| b = Inadimplencia |
| Descoberta de Conhecimento no *Data Warehouse* Através do algoritmo ID-3 |
| …<br>**subcategoria = FAVELA**<br>\| situacao_esgoto = POTENCIAL: Adimplencia<br>\| **situacao_esgoto = LIGADO_NORMAL**<br>\| \| situacao_agua = LIGADA: Adimplencia<br>\| \| **situacao_agua = CORTADA: Inadimplencia**<br>\| \| situacao_agua = SUPRIMIDA_TOTAL: null<br>\| situacao_esgoto = FACTIVEL: Adimplencia<br>…<br>**subcategoria = ESCRITORIO/ASSOCIACAO_COM_ATIVIDADE_COMERCIAL**<br>\| situacao_agua = LIGADA: Adimplencia<br>**\| situacao_agua = CORTADA: Inadimplencia**<br>**\| situacao_agua = SUPRIMIDA_TOTAL: Inadimplencia** ... |

As classificações em forma de árvore de decisão são geradas como resultado da execução do algoritmo de *Data Mining*, e estas correspondem ao conhecimento descoberto da base de dados. Abaixo, serão discutidas algumas das regras de associações geradas pelo software.

A regra formada pelo nó raiz subcategoria, passando pelo nó situação do esgoto, nó situação da água e nó folha inadimplência se encontra em negrito na Tabela 4.1. Esta regra indica que todos os consumidores da subcategoria FAVELA, com situação do esgoto LIGADO_NORMAL e situação da água CORTADA se encontram inadimplentes junto à companhia de abastecimento de água. Outra descoberta no setor foi que todos os

consumidores da subcategoria ESCRITORIO/ASSOCIACAO_COM_ATIVIDADE_COMERCIAL e situação da água CORTADA OU SUPRIMIDA_TOTAL se encontram inadimplentes.

A *Confusion Matrix* (Matriz de Confusão) produzida pelo algoritmo classificador ID-3 mostra que das 2.523 instâncias classificadas como Adimplência, 44 foram classificadas incorretamente, visto que a classificação deveria ter sido de Inadimplência. Das 10 instâncias classificadas como Inadimplência 2 delas foram classificadas incorretamente.

#### 4.3.1.2 Algoritmo J4.8

Ao executar o algoritmo J4.8 com os dados do modelo Perfil do Setor, o *software* WEKA gerou as informações da mineração conforme mostra a Tabela 4.2. Ao todo foram utilizadas 2.583 instâncias de treinamento para a classificação, sendo 2.533 instâncias classificadas corretamente (taxa de acurácia 98,06 %) e 50 classificadas incorretamente (taxa de erro 1,94%). O tempo de processamento foi de 0.07 segundos.

**Tabela 4.2 - algoritmo J4.8 aplicado ao modelo perfil do setor**

| |
|---|
| === Run information ===<br>Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2<br>Relation: modelo_perfil_setor-weka<br>Instances: 2583<br>Attributes: 6<br>Time taken to build model: 0.07 seconds<br>=== Summary ===<br>Correctly Classified Instances 2533 98.06 %<br>Incorrectly Classified Instances 50 1.94 %<br>=== Confusion Matrix ===<br>a b <-- classified as<br>2527 2 \| a = Adimplencia<br>48 6 \| b = Inadimplencia |
| Descoberta de Conhecimento no *Data Warehouse* Através do Algoritmo J4.8 |
| situacao_agua = LIGADA: Adimplencia (2548.0/40.0)<br>situacao_agua = CORTADA<br>\| subcategoria = ESCRITORIO/ASSOCIACAO_COM_ATIVIDADE_COMERCIAL: Inadimplencia (2.0)<br>…<br>\| subcategoria = INDUSTRIA_QUE_USA_AGUA_NO_PROCESSO_INDUSTRIAL: Inadimplencia (2.0)<br>situacao_agua = SUPRIMIDA_TOTAL: Inadimplencia (4.0) |

Partindo-se do nó raiz obtemos como padrão de comportamento que todos os consumidores com situação da água LIGADA e SUPRIMIDA_TOTAL são classificados

como Adimplentes e Inadimplentes respectivamente. Quando a situação da água é CORTADA outros nós (atributos) precisam ser verificados. Por exemplo, todos os consumidores da subcategoria "escritorio/associacao_com_atividade_comercial" e "industria_que_usa_agua_no_processo_industrial" que possuem situação da água CORTADA são classificados como Inadimplentes.

A Matriz de Confusão produzida pelo algoritmo classificador J4.8 mostra que das 2.527 instâncias classificadas como Adimplencia, 48 foram classificadas incorretamente. E das 6 instâncias classificadas como Inadimplencia duas delas foram classificadas incorretamente.

Uma das principais vantagens do algoritmo J4.8 é a árvore de decisão fornecida graficamente pela software WEKA, facilitando o entendimento e melhor análise dos resultados do *Data Mining*. A Figura 4.10 ilustra árvore de decisão do modelo Perfil do Setor.

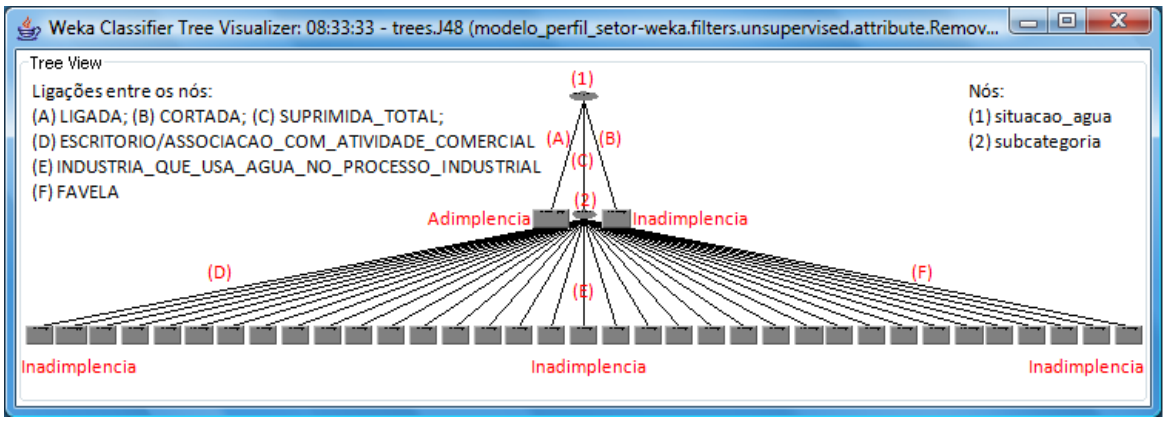



**Figura 4.10 - árvore de decisão para o modelo perfil do setor**

### 4.3.1.3 Algoritmo NaiveBayes

Ao executar o algoritmo *NaiveBayes* com os dados do modelo Perfil do Setor, o *software* WEKA gerou informações da mineração, conforme mostra a Tabela 4.3. Foram utilizadas 2.583 instâncias de treinamento para a classificação, sendo 2.533 instâncias classificadas corretamente (taxa de acurácia de 98,06%) e 50 classificadas incorretamente (taxa de erro de 1,94%). O tempo de processamento foi de 0.02 segundos.

Os algoritmos anteriores utilizam árvore de decisão, enquanto que o algoritmo *NaiveBayes* utiliza classificados estatísticos para predizer a probabilidade de um registro pertencer ao atributo classe. Na Tabela 4.3 encontram-se destacado a quantidade de inadimplências por atributo. Pode-se concluir que há predominância de inadimplências em

situação da água LIGADA, situação do esgoto LIGADA_NORMAL, categoria RESIDENCIAL e subcategoria FAVELA. Quanto ao atributo semestre verificou-se que em ambos os semestre a quantidade de inadimplência foi a mesma.

**Tabela 4.3 - algoritmo *naivebayes* aplicado ao modelo perfil do setor**

Scheme: weka.classifiers.bayes.NaiveBayes Relation: perfil_setor
Instances: 2583 Attributes: 6
Time taken to build model: 0.02 seconds
=== Summary ===

| | | |
|---|---|---|
| Correctly Classified Instances | 2533 | 98.0643 % |
| Incorrectly Classified Instances | 50 | 1.9357 % |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 2523 | 6 | a = Adimplencia |
| 44 | 10 | b = Inadimplencia |

Descoberta de Conhecimento no Data Warehouse Através do Algoritmo NaiveBayes

| Attribute | Class Adimplencia (0.98) | Inadimplencia (0.02) |
|---|---|---|
| situacao_agua | | |
| LIGADA | 2509.0 | 41.0 |
| CORTADA | 22.0 | 11.0 |
| SUPRIMIDA_TOTAL | 1.0 | 5.0 |
| [total] | 2532.0 | 57.0 |
| situacao_esgoto | | |
| POTENCIAL | 428.0 | 1.0 |
| LIGADO_NORMAL | 2081.0 | 55.0 |
| FACTIVEL | 23.0 | 1.0 |
| [total] | 2532.0 | 57.0 |
| categoria | | |
| COMERCIAL | 221.0 | 7.0 |
| RESIDENCIAL | 2292.0 | 47.0 |
| INDUSTRIAL | 5.0 | 3.0 |
| PUBLICO | 15.0 | 1.0 |
| [total] | 2533.0 | 58.0 |
| subcategoria | | |
| LOJA/SUPERMERCADO_E_SIMILARES | 95.0 | 1.0 |
| CASA | 1959.0 | 7.0 |
| ... | | |
| FAVELA | 225.0 | 41.0 |
| RESIDENCIA_COM_MENOS_DE_80_M2 | 20.0 | 1.0 |
| ... | | |
| [total] | 2564.0 | 89.0 |
| semestre | | |
| Primeiro_Semestre | 1259.0 | 28.0 |
| Segundo_Semestre | 1272.0 | 28.0 |
| [total] | 2531.0 | 56.0 |

A Matriz de Confusão produzida pelo algoritmo classificador NaiveBayes nos informa que das 2.523 instâncias classificadas como Adimplência, 44 delas foram classificadas incorretamente. E das 10 instâncias classificadas como Inadimplência, 6 foram classificadas incorretamente.

#### 4.3.1.4 Algoritmo Apriori

Ao executar o algoritmo *Apriori* sobre os dados do modelo Perfil do Setor, o *software* WEKA gerou as 10 melhores regras de associação do modelo, conforme mostra a Tabela 4.4. O valor do suporte mínimo e da confiança (ambos explicado na seção 2.5.5.3a) foi 75% e 0,9, respectivamente. Ao diminuir o valor do suporte mínimo o algoritmo gera mais regras de associações, contudo, a confiança das regras tendem a diminuir. O tempo de processamento do algoritmo *Apriori* não é informado pelo software.

**Tabela 4.4 - algoritmo *apriori* aplicado ao modelo perfil do setor**

=== Run information ===
Scheme: weka.associations.Apriori
Relation: modelo_perfil_setor_apriori-weka.
Instances: 2583
Attributes: 6
=======
Minimum support: 0.75 (1937 instances) //75% das 2.583 instâncias de treinamento
Minimum metric <confidence>: 0.9
=======
**BEST RULES FOUND (As 10 melhores regras de associação descobertas no setor 64):**
1. subcategoria=CASA inadimplencia=Adimplencia 1958 ==> categoria=RESIDENCIAL 1958 conf:(1)
2. situacao_esgoto=LIGADO_NORMAL semestre=Primeiro_Semestre inadimplencia=Adimplencia 1034 ==> situacao_agua=LIGADA 1030 conf:(1)
3. situacao_agua=LIGADA subcategoria=CASA semestre=Segundo_Semestre 972 ==> categoria=RESIDENCIAL 972 conf:(1)
4. situacao_agua=LIGADA subcategoria=CASA inadimplencia=Adimplencia 1942 ==> categoria=RESIDENCIAL 1942 conf:(1)
5. situacao_esgoto=LIGADO_NORMAL inadimplencia=Adimplencia 2080 ==> situacao_agua=LIGADA 2060 conf:(0.99)
6. subcategoria=CASA semestre=Primeiro_Semestre 976 ==> situacao_agua=LIGADA inadimplencia=Adimplencia 971 conf:(0.99)
7. situacao_esgoto=LIGADO_NORMAL subcategoria=CASA 1671 ==> situacao_agua=LIGADA categoria=RESIDENCIAL inadimplencia=Adimplencia 1650 conf:(0.99)
8. situacao_esgoto=LIGADO_NORMAL 2134 ==> situacao_agua=LIGADA 2100 conf:(0.98)
9. situacao_esgoto=LIGADO_NORMAL categoria=RESIDENCIAL 1936 ==> situacao_agua=LIGADA inadimplencia=Adimplencia 1870 conf:(0.97)
10. situacao_agua=LIGADA situacao_esgoto=LIGADO_NORMAL inadimplencia=Adimplencia 2060 ==> categoria=RESIDENCIAL 1870 conf:(0.91)

O algoritmo gera as dez melhores regras do setor através das associações dos dados de entrada. As regras com conf:(1) significam que a confiança é de 100%, como é o caso da regra número 2 da Tabela 4.4. Através desta regra pode-se dizer com 100% de acerto que os consumidores com situação do esgoto LIGADO_NORMAL, situação da água LIGADA e o

semestre de referência PRIMEIRO_SEMESTRE, estão adimplentes. Já a regra número 10 possui uma confiança de 91%, e neste caso pode-se encontrar consumidores (apenas 9%) que não satisfaçam a regra. Isto é, pode haver consumidor com situação da água LIGADA, situação do esgoto LIGADO_NORMAL, categoria RESIDENCIAL, mas esteja inadimplente.

### 4.3.2 Execução do Data Mining: Modelo Perdas Aparentes

Assim como no Modelo Perfil do Setor os atributos matrícula do consumidor e quadra não foram considerados, no Modelo Perdas Aparentes eles também não serão necessários. Os atributos "médias de consumo" e "médias da conta de água" foram eliminados por serem do tipo numérico. De modo que dos 14 atributos discutidos na seção 4.2.2, foram utilizados 10 atributos como entrada, são eles: anormalidade, capacidade do hidrômetro em $m^3$, tipo do hidrômetro, ano de fabricação do hidrômetro, inadimplência do consumidor, referência de consumo por semestre, consumo baseado na estrutura tarifária, indicador de medidor (hidrômetro instalado), período de instalação do hidrômetro e decisão de substituição do hidrômetro.

#### 4.3.2.1 Algoritmo ID-3

Ao executar o algoritmo ID-3 com os dados do modelo Perdas Aparentes, o *software* WEKA gerou as informações da mineração conforme mostra a Tabela 4.5 de forma sucinta. Ao todo foram utilizadas 3.523 instâncias e 10 atributos de treinamento para a classificação, sendo 3.095 instâncias classificadas corretamente (taxa de acurácia 87,85%), 385 classificadas incorretamente (taxa de erro 10,93%) e 43 instâncias não classificadas (1,22%).

**Tabela 4.5 - algoritmo ID-3 aplicado ao modelo perda aparente**

```
=== Run information ===
Scheme:       weka.classifiers.trees.Id3
Relation:     modelo_perda_aparente-weka
Instances:    3523
Attributes:   10
Time taken to build model: 0.09 seconds
=== Summary ===
Correctly Classified Instances        3095               87.8513 %
Incorrectly Classified Instances       385               10.9282 %
UnClassified Instances                  43                1.2206 %
=== Confusion Matrix ===
   a     b    c   <-- classified as
 3075    7   23 |    a = Desfavoravel_a_Troca
```

```
  72    2    1 |   b = Favoravel_a_Troca
 282    0   18 |   c = Analise_mais_detalhada
```

Descoberta de Conhecimento no *Data Warehouse* Através do Algoritmo ID-3

```
...
consumo_tarifario = ate_10M3_residencial
| ano_hid = 1994_a_1998
| | tipo_hid = MULT_MAGNETICO
| | | anormalidade = IMOVEL_OU_PORTAO_FECHADO
| | | | inadimplencia = Inadimplencia: Desfavoravel_a_Troca
…
consumo_tarifario = ate_10M3_residencial
| ano_hid = 2004_a_2008
| | anormalidade = SEM_ANORMALIDADE
| | | data_inst_hid = Entre_3_e_9_Anos
| | | | semestre = Primeiro_Semestre: Desfavoravel_a_Troca
| | | | semestre = Segundo_Semestre: Desfavoravel_a_Troca
…
consumo_tarifario = entre_10_e_20M3_residencial
| anormalidade = HIDROMETRO_SOTERRADO
| | tipo_hid = MULT_MAGNETICO
| | | ano_hid = 1994_a_1998: Favoravel_a_Troca
...
consumo_tarifario = ate_10M3_residencial
| ano_hid = 2004_a_2008
| | anormalidade = BY_PASS: Analise_mais_detalhada
…
consumo_tarifario = entre_10_e_20M3_residencial
| anormalidade = HIDROMETRO_QUEBRADO: Analise_mais_detalhada
…
consumo_tarifario = entre_10_e_20M3_residencial
| anormalidade = HIDROMETRO_VIOLADO
| | ano_hid = 1999_a_2003: Analise_mais_detalhada
…
consumo_tarifario = ate_10M3_industrial
| ano_hid = 1994_a_1998: Analise_mais_detalhada
| ano_hid = 2004_a_2008: Desfavoravel_a_Troca …
```

As regras de classificação correspondem ao conhecimento descoberto e são geradas em forma de árvore de decisão, conforme mostra a Tabela 4.5. A regra formada pelo nó raiz *consumo tarifário* residencial entre 10 e 20 $m^3$, nó *anormalidade* HIDROMETRO SOTERRADO, nó *tipo do hidrômetro* MULT_MAGNETICO e nó *ano de fabricação do hidrômetro* entre 1994 a 1998 está associadas ao nó folha *decisão* Favorável a Troca. Desta forma, Todos os consumidores que estão

associados a esta regra podem estar causando perdas aparentes no sistema de abastecimento de água e a solução proposta para o problema é a troca do hidrômetro.

Outro padrão encontrado no setor 64 foi que os consumidores pertencentes ao consumo tarifário residencial de até 10 $m^3$, ano de fabricação do hidrômetro entre 2004 e 2008 e anormalidade do tipo BY_PASS estão associados ao tributo decisão *analise mais detalhada*. Esta regra indica que os dados referentes a um ano não foram suficientes para propor a troca ou não do hidrômetro, por isso esta regra está associada ao atributo classe decisão do tipo *analise mais detalhada*.

A Matriz de Confusão produzida pelo algoritmo classificador ID-3 mostra que 3.075 instâncias foram classificadas corretamente como Desfavoravel_a_Troca e 354 foram classificadas incorretamente como Desfavoravel_a_Troca, visto que deveria ter sido classificada como Favoravel_a_Troca ou Analise_mais_detalhada. Quanto ao tipo Favoravel_a_Troca, duas instâncias foram classificadas corretamente e 7 incorretamente. E para o tipo Analise_mais_detalhada, 18 instâncias foram classificadas corretamente e 24 classificadas incorretamente.

#### 4.3.2.2 Algoritmo J4.8

Ao executar o algoritmo J4.8 com os dados do modelo Perdas Aparentes, o *software* WEKA gerou as informações da mineração conforme mostra a Tabela 4.6. Ao todo foram utilizadas 3.523 instâncias e 10 atributos de treinamento para a classificação, sendo 3.136 instâncias classificadas corretamente (taxa de acurácia 89,02%), 387 classificadas incorretamente (taxa de erro 10,99%). O tempo de processamento foi de 0.06 segundos.

**Tabela 4.6 - algoritmo J4.8 aplicado ao modelo perda aparente**

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     modelo_perda_aparente-weka
Instances:    3523
Attributes:   10
Time taken to build model: 0.06 seconds
=== Summary ===
Correctly Classified Instances        3136               89.015  %
Incorrectly Classified Instances       387               10.985  %
=== Confusion Matrix ===
   a      b    c   <-- classified as
 3121    3   17 |    a = Desfavoravel_a_Troca
```

```
  75     0    1 |   b = Favoravel_a_Troca
 291     0   15 |   c = Analise_mais_detalhada
```

Descoberta de Conhecimento no *Data Warehouse* Através do Algoritmo J4.8

```
...
tipo_hid = MULT_MAGNETICO
|   consumo_tarifario = acima_de_10M3_comercial
|   |   anormalidade = SEM_ANORMALIDADE
|   |   |   data_inst_hid = Menos_de_3_Anos
|   |   |   |   ano_hid = 1994_a_1998: Analise_mais_detalhada (1.0)
|   |   |   |   ano_hid = 2004_a_2008: Desfavoravel_a_Troca (19.0/6.0)
|   |   |   |   ano_hid = 1999_a_2003: Desfavoravel_a_Troca (3.0)
...
tipo_hid = MULT_MAGNETICO
|   consumo_tarifario = entre_10_e_20M3_residencial
|   |   anormalidade = HIDROMETRO_SOTERRADO: Favoravel_a_Troca (3.0/1.0)
...
tipo_hid = MULT_MAGNETICO
|   consumo_tarifario = acima_de_30M3_residencial
|   |   ano_hid = 1994_a_1998
|   |   |   anormalidade = SEM_ANORMALIDADE
|   |   |   |   data_inst_hid = Entre_10_e_18_Anos
|   |   |   |   |   capacidade = De_5_a_10_M3/H: Analise_mais_detalhada (2.0)
|   |   |   |   |   capacidade = Ate_3_M3/H: Desfavoravel_a_Troca (30.0/2.0) …
```

A Matriz de Confusão produzida pelo algoritmo classificador J4.8 mostra que 3.121 instâncias foram classificadas corretamente como Desfavoravel_a_Troca e 366 foram classificadas incorretamente como Desfavoravel_a_Troca, pois deveriam ter sido classificadas como Favoravel_a_Troca ou Analise_mais_detalhada. Quanto ao tipo Favoravel_a_Troca, nenhuma instância foi classificadas corretamente e 3 foram classificadas incorretamente. E para o tipo Analise_mais_detalhada, 15 instâncias foram classificadas corretamente e 18 classificadas incorretamente.

A navegação pela árvore de decisão gerada pela execução do algoritmo J4.8 determina as regras de classificação descobertas. Algumas delas estão apresentadas na Tabela 4.6, como é o caso da regra iniciada no nó raiz tipo do hidrômetro (tipo_hid) MULT_MAGNETICO, consumo tarifário acima_de_10M3_comercial, SEM_ANORMALIDADE, data de instalação do hidrômetro (data_inst_hid) *Menos_de_3_Anos,* com ano de fabricação do hidrômetro (ano_hid) entre 1999_a_2003 e 2004_a_2008 e nó folha Desfavoravel_a_Troca. Esta regra informa que os hidrômetros estão medindo o consumo de água corretamente, não há anormalidades e indícios de irregularidades e fraudes. Já os hidrômetros com ano de

fabricação entre 1994_a_1998 precisam de uma análise mais detalhada, visto que as perdas aparentes podem ser encontradas nos consumidores com este perfil.

Cada caminho do nó raiz (tipo_hid) ao nó folha (atributo decisão: Favoravel_a_Troca, Desfavoravel_a_Troca e Analise_mais_detalhada) é uma regra de associção.

A interpretação da toda a árvore de decisão determina todos os padrões descobertos da base de dados. A Figura 4.11 ilustra toda a árvore de decisão do modelo Perdas Aparentes, contudo, a regra de associação comentada anteriormente está representada em vermelho.

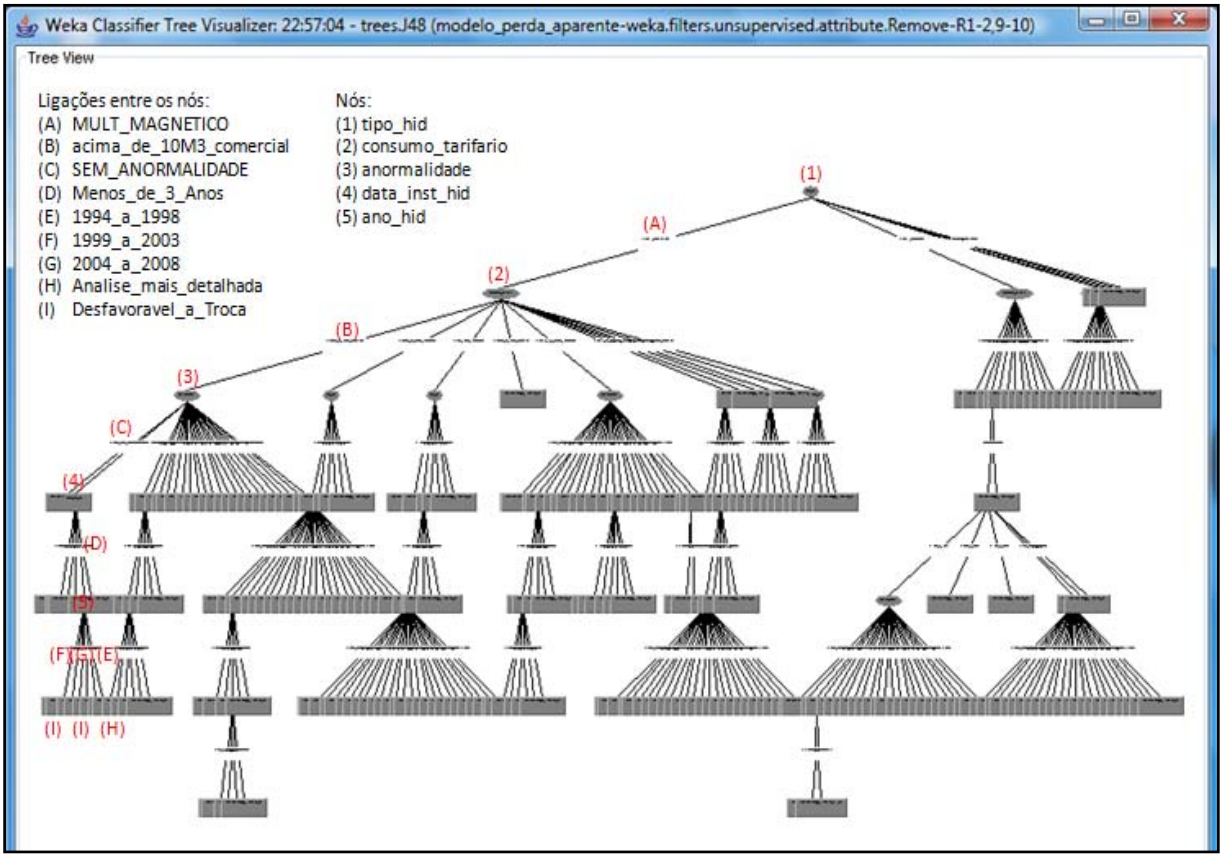



**Figura 4.11 - árvore de decisão para o modelo perda aparente**

### 4.3.2.3 Algoritmo NaiveBayes

Ao executar o algoritmo *NaiveBayes* sobre os dados do modelo Perda Aparente, o *software* WEKA gerou as informações da mineração, conforme mostra a Tabela 4.7. Das 3.523 instâncias de treinamento para a classificação, 3.131 instâncias foram classificadas corretamente (taxa de acurácia de 88,87%) e 392 classificadas incorretamente (taxa de erro de 11,13%). O tempo de processamento do algoritmo foi de 0.01 segundos.

**Tabela 4.7 - algoritmo *naivebayes* aplicado ao modelo perda aparente**

```
=== Run information ===
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     modelo_perda_aparente-weka
Instances:    3523
Attributes:   10
Time taken to build model: 0.01 seconds
=== Summary ===
Correctly Classified Instances        3131               88.8731 %
Incorrectly Classified Instances       392               11.1269 %
=== Confusion Matrix ===
   a    b   c   <-- classified as
3129    8   4 |   a = Desfavoravel_a_Troca
  76    0   0 |   b = Favoravel_a_Troca
 304    0   2 |   c = Analise_mais_detalhada
```

Descoberta de Conhecimento no *Data Warehouse* Através do Algoritmo NaiveBayes

(A) Desfavoravel_a_Troca; (B) Favoravel_a_Troca; (C) Analise_mais_detalhada

| Attribute | (A) (0.89) | (B) (0.02) | (C) (0.09) |
|---|---|---|---|
| Anormalidade: SEM_ANORMALIDADE | 2217.0 | 52.0 | 182.0 |
| LEITURA_IGUAL_ANTERIOR | 282.0 | 13.0 | 43.0 |
| ... | ... | ... | ... |
| [total] | 3165.0 | 100.0 | 330.0 |

| | | | |
|---|---|---|---|
| Capacidade: De_5_a_10_M3/H | 66.0 | 1.0 | 5.0 |
| Ate_3_M3/H | 3015.0 | 77.0 | 303.0 |
| ... | ... | ... | ... |
| [total] | 3145.0 | 80.0 | 310.0 |

| | | | |
|---|---|---|---|
| tipo_hid : MULT_MAGNETICO | 2894.0 | 77.0 | 280.0 |
| MULT_MECANICO | 103.0 | 1.0 | 25.0 |
| ... | ... | ... | ... |
| [total] | 3147.0 | 82.0 | 312.0 |

| | | | |
|---|---|---|---|
| ano_hid : 1994_a_1998 | 748.0 | 11.0 | 88.0 |
| 2004_a_2008 | 1135.0 | 20.0 | 100.0 |
| 1999_a_2003 | 852.0 | 48.0 | 109.0 |
| ... | ... | ... | ... |
| [total] | 3147.0 | 82.0 | 312.0 |

| | | | |
|---|---|---|---|
| Inadimplência: Adimplencia | 3048.0 | 77.0 | 307.0 |
| Inadimplencia | 95.0 | 1.0 | 1.0 |
| [total] | 3143.0 | 78.0 | 308.0 |

| | | | | |
|---|---|---|---|---|
| Semestre: Primeiro_Semestre | 1612.0 | 37.0 | 164.0 | |
| Segundo_Semestre | 1531.0 | 41.0 | 144.0 | |
| [total] | 3143.0 | 78.0 | 308.0 | |
| consumo_tarifario: ate_10M3_residencial | 1027.0 | 8.0 | 109.0 | |
| entre_10_e_20M3_residencial | 862.0 | 44.0 | 114.0 | |
| ... | ... | ... | ... | |
| [total] | 3152.0 | 87.0 | 317.0 | |
| indicativo_medidor: Com_Medidor | 3094.0 | 77.0 | 307.0 | |
| Sem_Medidor | 49.0 | 1.0 | 1.0 | |
| [total] | 3143.0 | 78.0 | 308.0 | |
| data_inst_hid: Entre_3_e_9_Anos | 2065.0 | 56.0 | 211.0 | |
| Entre_10_e_18_Anos | 655.0 | 9.0 | 77.0 | |
| ... | | | | |
| [total] | 3146.0 | 81.0 | 311.0 | |

A Matriz de Confusão produzida pelo algoritmo classificador NaiveBayes mostra que 3.129 instâncias foram classificadas corretamente como Desfavoravel_a_Troca e 380 foram classificadas incorretamente, isto é, deveriam ter sido classificadas como Favoravel_a_Troca ou Analise_mais_detalhada. Quanto ao tipo Favoravel_a_Troca, nenhuma instância foi classificadas corretamente e 8 foram classificadas incorretamente. E para o tipo Analise_mais_detalhada, 2 instâncias foram classificadas corretamente e 4 classificadas incorretamente.

Na Tabela 4.7 encontram-se as quantidades de instâncias Desfavoravel_a_Troca, Favoravel_a_Troca e Analise_mais_detalhada associadas a cada atributo do modelo Perdas Aparentes. Pode-se concluir que há predominância do tipo Desfavoravel_a_Troca (89% das instâncias) em todos os nove atributos, seguida por Analise_mais_detalhada (9% das instâncias). O conhecimento descoberto com esta informação permite-nos avaliar como boa a qualidade dos micromedidores (hidrômetros), assim como diagnosticar que as perdas aparentes do setor 64 relativas à medição do consumo de água são pouco significativas, haja vista que apenas 2% das instâncias analisadas sugerem a substituição do hidrômetro.

#### 4.3.2.4 Algoritmo Apriori

Ao executar o algoritmo *Apriori* sobre os dados do modelo Perdas Aparentes, o *software* WEKA gerou as melhores regras de associação do modelo, conforme mostra a

Tabela 4.8. Foram 3.523 instâncias de treinamento para a aplicação da tarefa de associação, com valor do suporte mínimo e da confiança de 60% e 0,9, respectivamente.

**Tabela 4.8 - algoritmo *apriori* aplicado ao modelo perda aparente**

```
=== Run information ===
Scheme:       weka.associations.Apriori
Relation:     modelo_perda_aparente-weka
Instances:    3523
Attributes:   10
=======
Minimum support: 0.6 (2114 instances) //60% das 3.523 instâncias de treinamento
Minimum metric <confidence>: 0.9
=======
BEST RULES FOUND (As 10 melhores regras de associação descobertas no setor 64):
1. capacidade=Ate_3_M3/H inadimplencia=Adimplencia 3320 ==>
indicativo_medidor=Com_Medidor 3320    conf:(1)
2. tipo_hid=MULT_MAGNETICO inadimplencia=Adimplencia 3217 ==>
indicativo_medidor=Com_Medidor 3217    conf:(1)
3. inadimplencia=Adimplencia decisao=Desfavoravel_a_Troca 3047 ==>
indicativo_medidor=Com_Medidor 3017    conf:(0.99)
4. capacidade=Ate_3_M3/H indicativo_medidor=Com_Medidor 3392 ==>
inadimplencia=Adimplencia 3320    conf:(0.98)
5. capacidade=Ate_3_M3/H tipo_hid=MULT_MAGNETICO 3169 ==>
inadimplencia=Adimplencia 3138    conf:(0.99)
6. tipo_hid=MULT_MAGNETICO indicativo_medidor=Com_Medidor 3248 ==>
capacidade=Ate_3_M3/H inadimplencia=Adimplencia 3138    conf:(0.97)
7. decisao=Desfavoravel_a_Troca 3141 ==> inadimplencia=Adimplencia
indicativo_medidor=Com_Medidor 3017    conf:(0.96)
8. decisao=Desfavoravel_a_Troca 3141 ==> capacidade=Ate_3_M3/H
indicativo_medidor=Com_Medidor 3014    conf:(0.96)
9. capacidade=Ate_3_M3/H inadimplencia=Adimplencia decisao=Desfavoravel_a_Troca 2942
==> tipo_hid=MULT_MAGNETICO indicativo_medidor=Com_Medidor 2787    conf:(0.95)
10. 200. anormalidade=SEM_ANORMALIDADE 2448 ==> tipo_hid=MULT_MAGNETICO
inadimplencia=Adimplencia 2307    conf:(0.94)
```

Os novos conhecimentos adquiridos com o algoritmo *Apriori* estão representados nas dez regras de associação fornecidas na Tabela 4.8. A regra número 1 gera a informação que todos os consumidores com capacidade do hidrômetro de até 3 $m^3$ estão adimplentes. Esta regra tem confiança de 100%. A regra número 7 possui confiança de 96% e determina que os consumidores que possuem medidores de consumo e que se encontram adimplentes estão no grupo de hidrômetros Desfavoravel_a_Troca. Já a regra número 10 indica que todos os consumidores sem anormalidade e tipo do hidrômetro MULT_MAGNETICO estão adimplentes, com confiança de 94%.

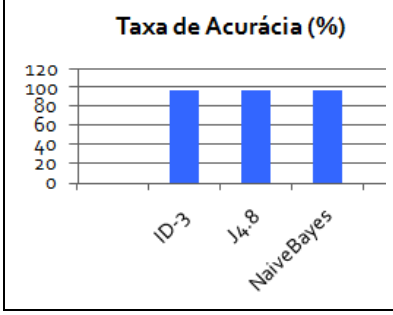
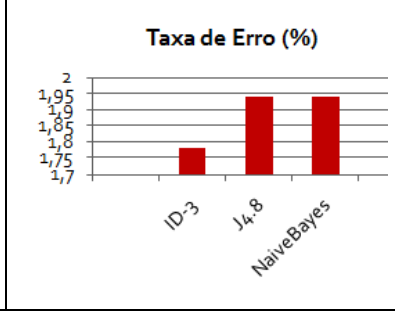
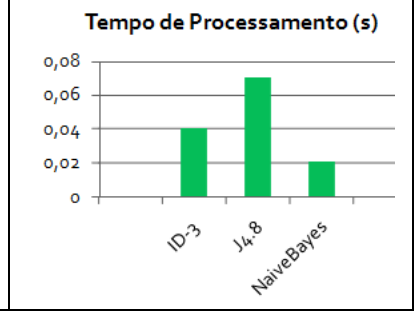
## 4.4 CONSIDERAÇÕES FINAIS

Os algoritmos de *Data Mining* foram utilizados para fornecer as informações desconhecidas do setor 64, úteis ao processo de apoio à decisão, visando a gestão automática dos processos relacionados à detecção e minimização das perdas aparentes no sistema de abastecimento de água. Estas informações foram geradas pelos algoritmos em forma de árvores de decisões, análise estatística e regras de associação ao longo deste capítulo.

A fim de determinar um algoritmo de *Data Mining* adequado para aplicações na área do saneamento é que foram aplicadas e comparadas técnicas e algoritmos distintos de DM sobre o ambiente de DW implementado. Com isto, este trabalho conseguiu extrair conhecimento e padrões desconhecidos entre os dados, como também comparou algoritmos de *Data Mining* e baseado nos resultados obtidos pelo mesmos, elegeu um deles como o mais indicado em projetos na área de KDD para o segmento do saneamento.

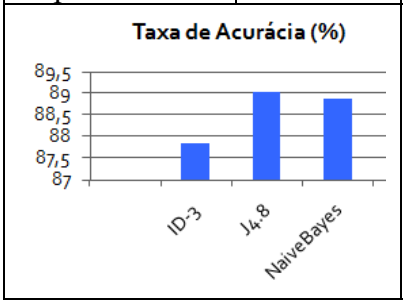
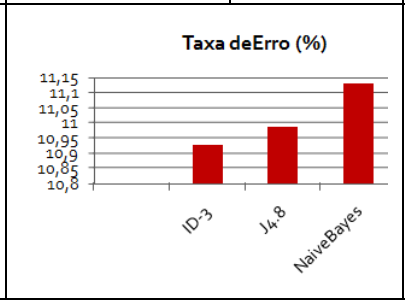
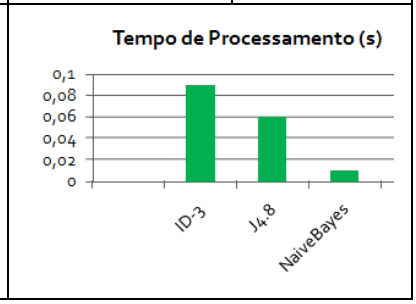
Foram utilizados quatro algoritmos tanto para o modelo Perfil do Setor quanto para o modelo Perdas Aparentes. A Tabela 4.9 e Tabela 4.10 apresentam os comparativos entre os algoritmos *ID-3, J4.8, NaiveBayes e Apriori* quanto à taxa de acurácia (acerto), taxa de erro, visualização gráfica e tempo de processamento (desempenho).

**Tabela 4.9 - comparativo entre os algoritmos de *data mining* aplicados ao modelos perfil do setor**

| MODELO PERFIL DO SETOR | | | | | |
|---|---|---|---|---|---|
| **Aprendizado Indutivo** | **Algoritmo** | **Taxa de Acurácia (%)** | **Taxa de Erro (%)** | **Tempo de Processamento (s)** | **Visualização Gráfica** |
| Supervisionado | ID-3* | 98,06 | 1,78 | 0,04 | - |
| | J4.8 | 98,06 | 1,94 | 0,07 | Árvore de Decisão |
| | *Naive Bayes* | 98,06 | 1,94 | 0,02 | - |
| Não Supervisionado | *Apriori* | - | - | - | - |
|  | |  | |  | |

*0,16% não classificadas

**Tabela 4.10 - comparativo entre os algoritmos de *data mining* aplicados ao modelo perdas aparentes**

| MODELO PERDAS APARENTES | | | | | |
|---|---|---|---|---|---|
| **Aprendizado Indutivo** | **Algoritmo** | **Taxa de Acurácia (%)** | **Taxa de Erro (%)** | **Tempo de Processamento (s)** | **Visualização Gráfica** |
| Supervisionado | ID-3* | 87,85 | 10,93 | 0,09 | - |
| | J4.8 | 89,02 | 10,99 | 0,06 | Árvore de Decisão |
| | *Naive Bayes* | 88,87 | 11,13 | 0,01 | - |
| Não Supervisionado | *Apriori* | - | - | - | - |
|  | |  | |  | |

*1,22% não classificadas

Os algoritmos do modelo Perfil do Setor obtiveram taxas de acurácia iguais; o ID-3 foi o que obteve a menor taxa de erro (1,78%) e o algoritmo *NaiveBayes* foi o que obteve o menor tempo de processamento (0,02s). Enquanto no modelo Perdas Aparentes, o algoritmo J4.8 obteve a melhor taxa de acurácia (89,02%); o ID-3 obteve também a menor taxa de erro (10,93%); e o algoritmo *NaiveBayes* obteve o menor tempo de processamento (0,01s).

Os dois aprendizados indutivos têm objetivos similares, i.e., empenham-se em descobrir regras e informações desconhecidas a priori do conjunto de instâncias submetidas aos algoritmos de *Data Mining*. O que difere é a necessidade de determinar o atributo classe (nó folha de uma árvore de decisão) no aprendizado supervisionado. Já no não supervisionado não precisa definir o atributo classe. Desta forma, os dois tipos puderam ser utilizados neste trabalho sem nenhuma objeção. A finalidade não foi compará-los e sim acrescentar mais conhecimento e informações de suporte a decisão para o setor 64.

As comparações sugeridas no início da seção 4.3, correspondem aos três algoritmos do aprendizado supervisionado. No que diz respeito a geração de novos conhecimentos e padrões, os dois aprendizados foram satisfatórios neste estudo de caso. Quanto aos três algoritmos do aprendizado supervisionado, o J4.8 foi o que mais se sobressaiu. Por sua vez, o que tornou o J4.8 superior aos demais foi a facilidade que ele proporciona em analisar as

regras através da visualização gráfica da árvore de decisão. Além de possuir taxas de acuraria e erro próximas aos demais algoritmos verificados e um tempo de processamento baixo.

A taxa de acurácia e erro dos algoritmos ficaram na média de 98% e 1,8% para o modelo Perfil do Setor e 88% e 11% para o modelo Perdas Aparentes. O modelo Perdas Aparentes obteve uma percentagem menor da taxa de acurácia e uma taxa maior de erro em relação ao modelo Perfil do Setor. Esta característica foi identificada e é válida, visto que o modelo Perdas Aparentes possui uma quantidade maior de atributos e instâncias (10 e 3523, respectivamente) em relação ao modelo Perfil do setor (6 atributos e 2583 instâncias). Ao aumentar o número de instâncias do modelo, a tendência é aumentar a taxa de erro e diminuir o taxa de acurácia, e foi justamente isto que ocorreu nos modelos minerados. Deste modo, os padrões e descobertas dos algoritmos foram satisfatórios e expressaram a realidade do setor.

Portanto, a aplicação dos algoritmos de *Data Mining,* visando os objetivos propostos no início desta dissertação foi alcançada, proporcionando dentre outras funcionalidades e benefícios, a análise e descoberta do perfil do setor e dos consumidores; apresentação da situação dos micromedidores presentes nos imóveis dos consumidores para medição do consumo de água; detecção das anormalidades e irregularidades relacionando-as ao perfil do consumidor e imóveis.

O estudo do setor e os resultados alcançados precisam ser validados pelos especialistas da companhia de abastecimento de água, i.e., àqueles que possuem domínio de informações e resultados históricos, porém com comportamento apenas introspectivos e conhecimento tácitos. Após a validação do especialista, o estudo poderá ser colocado em prática pela companhia. Contudo, este tema evidencia-se a necessidade de se trabalhar com uma equipe interdisciplinar, agregando conhecimentos dos especialistas com os conhecimentos dos analistas, i.e., àqueles que dominam as tecnologias, em especial as técnicas de *Data Mining*, provocando assim a explicitação do conhecimento. Com esta dissertação conclui-se a etapa do analista.

Com a ferramenta OLAP só é possível realizar sumarizações e agregações nos dados contidos no *DW*. Descobrir novos conhecimentos e padrões nos dados não é possível utilizando apenas OLAP ou a pré-mineração. Por isso que o *Data Mining* é a etapa mais importante do processo de descoberta do conhecimento (KDD), pois transforma dados em conhecimento.

# CAPÍTULO 5

## 5 CONCLUSÃO

Uma das principais motivações para realização deste trabalho foi propor uma abordagem para facilitar a extração e geração de informações gerenciais a partir da exploração automatizada dos dados do sistema de abastecimento de água. A fim de se detectar perdas aparentes e verificar o perfil do consumidor e do setor selecionado para o estudo de caso, este trabalho direcionou-se para a criação de um ambiente computacional de armazenamento de dados, o *Data Warehouse*, para em seguida serem aplicadas as tecnologias OLAP e técnicas de *Data Mining* sobre estes dados que foram obtidos do ambiente OLTP da companhia de abastecimento da Paraíba.

O *Data Warehouse* implementado neste trabalho se mostrou uma solução eficiente quanto ao objetivo que lhe é destinado, isto é, fornecer um repositório contendo dados limpos e agregados para serem geradas e analisadas as informações e conhecimento úteis ao processo de gestão da companhia de abastecimento. Portanto, verifica-se que os objetivos contemplados neste estudo quanto ao uso do DW foram alcançados, podendo-se elencar, entre outros benefícios:

- Recuperação intuitiva e fácil dos dados consultados pelos cubos de dados através das ferramentas OLAP, em virtude de ser uma ferramenta com uma boa e eficiente interface gráfica, como foi apresentado na seção 3.2.6 do capítulo 3;
- Visão e análise dos dados em várias dimensões, através do modelagem dimensional utilizada na construção do *Data Warehouse*;
- Facilidade na obtenção das instâncias dos dados necessárias como entrada para os algoritmos de *Data Mining* dos modelos Perfil do Setor e Perdas Aparentes, conforme apresentado a partir do item 4.2.
- Melhor produtividade dos gestores pela utilização da tecnologia OLAP e de *Data Mining*, evitando trabalhos desnecessários de buscas e formulação de consultas a vários sistemas de Banco de Dados transacionais, eliminado as redundâncias e inconsistências nos dados.

Os resultados do trabalho, fornecidos no capítulo 4, contemplam o que foi proposto como objetivos iniciais da dissertação, entre eles:

- Detecção e geração de informações gerenciais úteis à entidade gestora do sistema de abastecimento de água, através das descobertas de conhecimento no *Data Warehouse*. Este objetivo foi verificado ao longo da seção 4.2 (resultados e discussões).
- Proporcionar tomadas de decisões e um maior controle do comportamento dos consumidores e dos imóveis, visando à redução de perdas de água e das perdas econômicas pela companhia de saneamento e assim contribuir para o uso racional da água. Nas análises dos resultados (seção 4.3) obtidos pelos algoritmos de *Data Mining*, este objetivo foi contemplado.
- Ao gerar as regras de classificação e associação dos dados, geram-se também novos padrões dos dados, desta forma, o conhecimento descoberto com as novas informações, se aplicadas ao setor analisado, serão úteis para propor medidas corretivas e preventivas para minimizar o problema das perdas aparentes nos sistemas de abastecimento de água.
- As comparações entre os algoritmos de *Data Mining* do Aprendizado Supervisionado foram realizadas na seção 4.4 (Considerações Finais do capítulo 4). O Aprendizado Não Supervisionado através do algoritmo Apriori foi utilizado como neste trabalho para mostrar mais uma forma de gerar conhecimento através de algoritmos de *Data Mining*.

Portanto, as interpretações dos resultados permitem concluir que os objetivos do trabalho foram alcançados. O ambiente de *Data Warehouse* para extração dos dados e a aplicação da tecnologia OLAP e algoritmos de *Data Mining* sobre estes dados foram definidos, implementados e avaliados. Optou-se pela elaboração de *Data Warehouse* Departamental por ser mais adequado às aplicações departamentais, no caso deste trabalho, o departamento comercial da companhia de abastecimento de água.

As ferramentas OLAP e de *Data Mining* podem ser aplicadas isoladamente e independentemente da implantação do *Data Warehouse*. Porém, sem o *Data Warehouse*, a obtenção dos dados seria feita diretamente sobre o ambiente OLTP da companhia, que se encontra em constantes atualizações dos dados, e com isto não representaria a realidade do setor. Foi justamente para evitar problemas neste ambiente inconstante e suas consequências inoportunas que se optou por implementar o *Data Warehouse*.

As principais contribuições deste trabalho foram:

- Definição, projeto e desenvolvimento de um sistema computacional de apoio à decisão, iniciando-o com a construção da modelagem lógica dimensional e em seguida com a implementação física de um *Data Warehouse,* para o armazenamento e integração dos dados comerciais correspondentes ao período de um ano de um sistema urbano de abastecimento de água.
- Aplicação de tecnologias e ferramentas OLAP para formulação de consultas e análises no *Data Warehouse*, com possibilidade de visualização dos dados em várias dimensões, através dos cubos de dados. E estes que por sua vez possibilitam que as agregações e sumarizações sejam realizadas por meio de operações *Slice and Dice*. Desta forma, o analista pode navegar por todas as informações e gerar seus próprios relatórios, conforme seu interesse e necessidade da empresa.
- Utilização e comparação de algumas técnicas e algoritmos de Mineração de Dados sobre o ambiente de *Data Warehouse,* a fim de extrair o conhecimento e padrões desconhecidos entre os dados, os quais podem ajudar a empresa a gerar novas estratégias para o setor.
- Elencar o algoritmo do aprendizado indutivo supervisionado que melhor se adequou a pesquisa, baseando-se nas funcionalidades e desempenhos apresentados nos resultados de três algoritmos de *Data Mining* (ID-3, J4.8 e *NaiveBayes*) aplicados ao estudo de caso.
- Dentre as contribuições oferecidas para a comunidade científica de Banco de Dados, pode-se destacar como a principal delas, a comprovação e validação que um SAD pode ser concebido e se tornar um poderoso ambiente de diagnóstico de problemas e análises de dados dentro de qualquer empresa que dispõe de dados históricos em suas bases de dados.

Alguns desafios e dificuldades foram encontrados durante o desenvolvimento deste sistema de apoio à decisão, entre eles:

- A dificuldade inicial neste trabalho foi definir os atributos e o intervalo dos dados necessários, haja vista que a empresa que gerencia os dados da CAGEPA se encontra na cidade de Recife-PE, e a falta de algum atributo poderia comprometer os resultados a serem alcançados;

- Por ser um trabalho que envolve outra área do conhecimento, no caso a Engenharia Hidráulica, se fez necessário o estudo específico de como se comporta um sistemas de abastecimento de água e suas peculiaridades;
- Poucos softwares livre de *Business Intelligence* disponíveis para fins técnicos e acadêmicos;

Após o estudo abordado nesta dissertação, estabelecem-se algumas recomendações para pesquisas similares. Alguns assuntos merecem aprofundamento em pesquisas ou trabalhos futuros. Os principais são:

- Utilização de outras técnicas de *Data Mining* não contempladas neste estudo, como por exemplo, Redes Neurais, Clusterizações e Algoritmos Genéticos;
- O estudo de mecanismos inteligentes para detecção do tamanho da amostra de dados e dos parâmetros ideais para serem aplicados aos algoritmos de *Data Mining*;
- Processamento geográfico multidimensional do setor e consultas do tipo SOLAP (*Spatial OLAP*) através de um *Data Warehouse Geográfico*.

Finalmente, espera-se com este trabalho contribuir significativamente para o aumento da qualidade e eficiência na gestão de informações de apoio à decisão para o segmento do saneamento urbano, visando à economia, bom uso e administração adequada e racional de um dos maiores bens da humanidade, a água.

# CAPÍTULO 6

## 6 BIBLIOGRAFIA


AGRAWAL, R., T. IMIELINSKI, e A. SWAMI. *Mining Association Rules Between Sets of Itens in Large Databases.* Editora: Int. Conf. Management of Data, 1993.

AGRAWAL, Rakesh, e Usama M. FAYYAD. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, 1996.

ALMEIDA, Eduardo Cunha de. *Estudo de Viabilidade de uma Plataforma de Baixo Custo para Data Warehouse.* Dissertação de Mestrado, Curitiba, 2004.

ALVES, R., e P. FERREIRA. *Discovering Telecom Fraud Situations through Mining Anomalous Behavior Patterns.* Artigo, In Proceedings of 1st Workshop on Data Mining for Business Applications Philadelphia, USA: ACM SIGKDD, 2006.

ARBEX, Eduardo Compasso, Alexandre de Paiva SABOREDO, e Dhalila MIRANDA. *Implementação e Estudo de caso do algoritmo Apriori para Mineração de Dados.* Artigo, Curso de Sistemas de Informação, Associação Educacional Dom Bosco, Simpósio de Excelência em Gestão e Tecnologia (SEGeT), Resende - Rio de Janeiro, 2004.

AURÉLIO, Marco, Marley VELLASCO, e Carlos Henrique LOPES. *Descoberta de Conhecimento e Mineração de Dados.* Artigo, ICA – Laboratório de Inteligência Computacional Aplicada, Departamento de Engenharia Elétrica/PUC–Rio, 2000.

BALLARD, Chuck, e Dirk HERREMAN. *Data Modeling Techniques for Data Warehousing.* IBM, International Technical Support Organization, 1998.

BARROS, Monica Coutinho de. *Warehouse Management System (WMS): Conceitos Teóricos e Implementação em um Centro de Distribuição.* Dissertação de Mestrado, Departamento de Engenharia Industrial, PUC, Rio de Janeiro, 2005.

BARROSO, Bruno da Costa, e Pedro Nolasco Neto FERREIRA. *Descoberta de Conhecimento na Base de Dados de uma Locadora.* Monografia, Ciência da Computação, Universidade Federal do Pará, Belém, 2006.

BATISTA, G. E. A. P. A. & MONARD, M. C. *Descrição da Arquitetura e do Projeto do Ambiente Computacional Discover Learning Environment — DLE.* Artigo, ICMC-USP, 2003.

BATISTA, Gustavo Enrique de Almeida Prado Alves. *Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado.* Tese de Doutorado, Ciências de Computação e Matemática Computacional, Instituto de Ciência Matemática e de Computação - ICMC-USP, São Carlos - SP, 2003, 232.

BISPO, Carlos Alberto Ferreira. *Uma análise da nova geração de sistemas de apoio à decisão.* Dissertação de Mestrado, Escola de Engenharia de São Carlos, 1998, 174.

CARVALHO, Luís Alfredo Vidal de. *DataMining: a Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.* São Paulo: Editora: Ciência Moderna, 2001.

COLAÇO, Methanias Júnior. *Projetando Sistemas de Apoio a Decisão Baseados em Data Warehouse.* Rio de Janeiro: Editora: Axcel Books, 2004.

COME, Gilberto de. *Contribuiç ão ao Estudo da Implementação de Data Warehousing: Um Caso no Setor de Telecomunicações.* Dissertação de Mestrado, Departamento de Administração (Métodos Quantitativos e Informática), USP (Universidade de São Paulo), São Paulo, 2001, 144.

DATE, C. J. *Introdução a Sistemas de Bancos de Dados.* 8ª Edição. Editora: Campus, 2004.

ELMASRI, Ramez E., e Shamkant B. NAVATHE. *Sistemas de Banco de Dados [Trad].* 4ª Edição. Editora: Pearson, 2005.

FAYYAD, U., G. PIATETSKY-SHAPIRO, e P. SMYTH. *Advances in Knowledge Discovery and Data Mining.* Califórnia América Association for Artificial Inteligence, 1996.

FONSECA, Marcello Porto Alegre da. *Classificação Bayesiana de grandes massas de dados em ambientes ROLAP.* Tese de Doutorado, Ciências em Engenharia Civil, Universidade Federal do Rio de Janeiro, 2007, 117.

GARDNER, S. R. *Building the data warehouse.* Communications of the ACM, 1998.

GILBERTO, de Come. *Contribuição ao Estudo da Implementação de Data Warehousing: Um Caso no Setor de Telecomunicações.* Dissertação de Mestrado, São Paulo, 2001.

GOLDSCHMIDT, Ronaldo, e Emmanuel PASSOS. *Data Mining: Um Guia Prático.* 1ª Edição. Editora: Campus, 2005.

GOUDA, K., e M. J. ZAKI. *Efficiently Mining Maximal Frequent Itemsets.* Artigo, IEEE International Conference on Data Mining, Washington, DC, USA, 2001, p. 163-170.

GOMES, H. P. *Sistemas de Abastecimento de Água: Dimensionamento Econômico e Operação de Redes e Elevatórios.* 2a. Edição. Revisada e ampliada. João Pessoa: Editora Universitária da UFPB, 2004.

GOMES, H. P., Rafael Pérez GARCÍA, e Pedro L. Iglesias REY. *Abastecimento de Água - O estado da arte e técnicas avançadas.* 1ª edição. 2007.

GONZALES, Michael L. *IBM® Data Warehousing with IBM Business Intelligence Tools.* Canada: Wiley Publishing, Inc., 2003.

GRAY, Jim, Surajit CHAUDHURI, Adam BOSWORTH, Andrew LAYMAN, Don REICHART, e Murali VENKATRAO. *Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals.* Edição: Usama Fayyad. 1996.

HAN, Jiawei, e Micheline KAMBER. *Data Mining: Concepts and Techniques.* Second Edition. Elsevier Science & Technology Books, 2006.

IMHOFF, Claudia, Nicholas GALEMMO, e J. G. GEIGER. *Mastering Data Warehouse Design: Relational and Dimensional Techniques.* Indianapolis: Wiley Publishing, 2003.

INMON, William H. *Building the Data Warehouse: Gettiing Started.* 4ª Edição. Editora: Wiley Publishing, Inc, 2005.

JAMES, Kevin, Sephanie L. CAMPBELL, e Christophe GODLOVE. *Água e Energia.* Editora: Alliance – Aliança pra Conservação de Energia, 2002.

KANASHIRO, Augusto. *Um data warehouse de publicações científicas: indexação automática da dimensão tópicos de pesquisa dos data marts.* Dissertação de Mestrado, USP - São Carlos, São Paulo, SP, 2007.

KIMBALL, Ralph. *Digging into data mining - your data warehouse is your data mining platform.* DBMS and Internet Systems, 1997.

KIMBALL, Ralph, e Margy ROSS. *The data warehouse toolkit : the complete guide to dimensional modeling.* 2nd ed. John Wiley and Sons, 2002.

KORTH, H. F., S. SUDARSHAN, e A. SILBERSCHATZ. *Sistema de Banco de Dados.* 5ª Edição. Editora: Campus, 2006.

KUTOVA, Marcos André S., e Clodoveu A. Jr. DAVIS. *Mineração de regras de associação para seleção de oferta de cursos de especialização. III Workshop em Algoritmo e Aplicações de Mineração de Dados - WAAMD*, 2007.

LAROSE, Daniel T. *Discovering knowledge in data: an introduction to data mining.* New Jersey, Central Connecticut State University: A John Wiley & Sons, Inc, 2005.

LOPES, Maurício Capobianco, e Percio Alexandre de OLIVEIRA. *Ferramenta de Construção de Data Warehouse.* Artigo, Departamento de Sistemas e Computação, Universidade Regional de Blumenau - FURB, Blumenau - SC, 2006.

MARCKA, Estanislau, Ricardo Toledo SILVA, e João Gilberto Lotufo CONEJO. *Indicadores de Perdas nos Sistemas de Abastecimento de Água.* DTA A2 - Documento Técnico de Apoio, Secretaria Nacional de Saneamento Ambiental, PNCDA - Programa Nacional de Combate ao Desperdício de Água, Revisão 2004.

MARQUES, Alfeu, e Joaquim José de Oliveira SOUSA. *Hidráulica Urbana: Sistemas de Abastecimento de Água.* 2006.

MARQUES, Rodrigo Noli da Silva. *Uma Contribuição para o Estabelecimento de uma Modelagem de um armazém de Dados como Base para um Sistema de Informação Gerencial Logístico Aplicada ao Transporte Aéreo.* Dissertação de Mestrado, Ciências em Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2007, 96.

MENDES, Ilza Maria B., Alexandre PLASTINO, e Luiz Satoru OCHI. *Regras de Associação: suas Diferentes Formas e seus Algoritmos de Mineração.* Artigo, Instituto de Computação, Universidade Federal Fluminense - UFF, Niterói, RJ, 2002.

MINUSSI, Marlon Mendes. *Metodologia de Mineração de Dados para Detecção de Desvio de Comportamento do Uso de Energia em Concessionária de Energia Elétrica.* Dissertação de Mestrado, Programa de Pós-graduação em Engenharia Elétrica, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2008, 124.

MONTEIRO, Rodrigo Salvador. *Dwfist: Uma Abordagem Baseada em Data Warehouse para Exploração e Análise de Conjuntos de Itens Frequentes.* Tese de Doutorado, Engenharia de Sistemas e Computação, COPPE/UFRJ, Rio de Janeiro, 2005, 142.

NONG, YE. *The Handbook of Data Mining.* New Jersey, Arizona State University: Lawrence Erlbaum Associates, Inc., 2003.

OMIECINSKI, E. R. *Alternative interest measures for mining associations in databases.* Artigo, IEEE Transactions on Knowledge and Data Engineering, 2003, v. 15, p. 57-69.

PAIM, Fábio Rilston Silva. *Uma Metodologia para Definição de Requisitos em Sistemas Data Warehouse.* Centro de Informática - PE, Recife, 2003, 198.

PASSINI, Sílvia Regina Reginato. *Mineração de Dados para Detecção de Fraudes em Ligações de Água.* Dissertação de Mestrado, PUC - Campinas, São Paulo, 2002.

PASSINI, Sílvia Regina Reginato, e Ana Kelly NAIME. *Data Mart para apresentação dos resultados econômico-financeiros da setorização.* Artigo, Sociedade de Abastecimento de Água e Saneamento - SANASA, Campinas – São Paulo, 2004.

PASSINI, Sílvia Regina Reginato, e Carlos Miguel Tobar TOLEDO. *Mineração de Dados para Detecção de Fraudes em Ligações de Água.* Artigo, XI SEMINCO - Seminário de computação - 2002, Campinas, 2002.

PEREIRA, Celina Maria Rodrigues. *Comparação de Ferramentas de Data Mining.* Monografia, Departamento de Engenharia Informática, Instituto Politécnico do Porto, 2002.

PINHO, Selma Foligne Crespio de. *Uma Metodologia de Apoio à Decisão para Priorização de Projetos de Tecnologia da Informação.* Tese de Doutorado, Ciências em Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006, 164.

PONNIAH, Paulraj. *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals.* John Wiley & Sons, Inc., 2001.

PRADO, Marcelo Vinaud. *Data warehouse para apoio a gestão da operação em empresas do transporte rodoviário interestadual de passageiros.* Dissertação de Mestrado, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, 2006.

QUEYROI, Roberto. *Aplicação de Modelo de Mineração de Dados em um Sistema de Apoio a Decisão para Empresas de Saneamento.* Dissertação de Mestrado, Sistemas Computacionais em Engenharia Civil, Universidade Federal do Rio de Janeiro - UFRJ, Rio de Janeiro, 2007, 112.

QUINLAN, J. R. *C4.5: Programs for machine Learning.* San Mateo, CA. Editora: Morgan Kaufman, 1993.

QUISPE, Newton Roy Pampa. *Técnicas e Ferramentas para a Extração Inteligente e Automática de Conhecimento em Banco de Dados.* Dissertação de Mestrado, Engenharia Elétrica e de Computação, Universidade Estadual de Campinas - UNICAMP, Campinas, São Pasulo, 2003, 104.

QUITÉRIO, João, Nuno MARTINS, Paulo FERREIRA, e Pedro VIEIRA. *Análise comparativa de ferramentas de Data Mining* e OLAP. Artigo, Instituto Superior Técnico - IST.

RAINARDI, Vincent. *Building a Data Warehouse: With Examples in SQL Server.* Edição: Editora: Apress. 2008.

REIS, J., E. M. CONTIJO, E. MANIZA, J. E. CABRAL, e J. O. P. PINTO. *Fraud Indentification in Eletricity Company Customers using Decision Tree.* Artigo, IEEE International Conference on Systems, Man and Cybernetics, 2004.

REZENDE, S. O. *Sistema Inteligentes: Fundamentos e Aplicações.* 1ª ed. São Carlos: Manole, 2003.

REZENDE, S. O., J. B. PUGLIESI, E.A. M., e M.F P. *Mineração de Dados.* Editora Manole, 2003.

ROSS, D., e A. SCHOMAN. *Structured Analysis for Requirements Definition.* Artigo, 3(1):6-15, IEEE Transactions on Software Engineering (special issue on requirements analysis), 1977.

SANCHES, André Rodrigo. *Uma visão Geral sobre Mineração de Dados.* Relatório de Estudo - Tópicos em Ciência da Computação, Dept. Ciência da Computação, Universidade de São Paulo - USP, São Paulo, 2003.

SANTOS, Miriam Oliveira dos, e Maria Cláudia CAVALCANTI. *Uma estratégia baseada em técnicas de KDD para apoiar o Projeto Físico em SGBD's XML Nativos.* Artigo do *XXII SBBD - Simpósio Brasileiro de Banco de Dados*, 2007: 15.

SANTOS, Ricardo da Silva. *Ambiente para Extração de Informações através da Mineração das Bases de Dados do Sistema Único de Saúde.* Tese de Doutorado, Programa de pós-graduação em Informática em Saúde, Universidade Federal de São Paulo - USP, São Paulo, 2007, 254.

SANTOS, R. S., A. L. ALMEIDA, U. TACHINARDI, e M. A. GUTIERREZ. *Data Warehouse para a Saúde Pública: Estudo de Caso SES-SP*. Artigo, Anais do X Congresso Brasileiro de Informática em Saúde, Florianópolis, 2006, p. 53-58.

SHEN, W. M. *Bayesian probability theory – A general method for machine learning.* Microelectronics and Computer Technology Corporation, Austin, 1993.

SNIS, Sistema Nacional de Informações sobre Saneamento. *Diagnóstico dos Serviços de Água e Esgoto.* Secretaria Nacional de Saneamento Ambiental, Ministério das Cidades, Brasília, 2007.

SOUSA, Sidney Roberto. *Modelagem e Construção de um ambiente de Data Warehouse para o banco de dados do Censo Indígena.* Ciência da Computação, Universidade Estadual de Mato Grosso do Sul, Dourados - MS, 2007.

STOLTE, C., D. TANG, e P. Polaris HANRAHAN. *A System for Query, Analysis, and Visualization of Multidimensional Relational Databases.* Artigo, IEEE Transactions on Visualization and Computer Graphics, 2002, V. 8, n. 1, p. 52-65.

SUMATHI, S., e S.N. SIVANANDAM. *Introduction to Data Mining and its Applications.* Editora Springer-Verlag Berlin Heidelberg, 2006.

SYMEONIDIS, Andreas L., e Pericles A. MITKAS. *Agent Inteligence Through Data Mining.* Vol. 14. Aristotle University ofThessaloniki, 2005.

TANIAR, David. *Data Mining and Knowledge Discovery Technologies.* IGI Publishing, 2008.

TEIXEIRA, Cristina Josefa Santos. *Descoberta de conhecimento em bases de dados como suporte a actividades de business intelligence: aplicação na área da distribuição da água.* Dissertação de Mestrado, Tecnologias e Sistemas de Informação, Universidade do Minho, Portugal, 2006.

THOMSEN, E. *OLAP : Construindo Sistemas de Informações Multidimensionais.* 2º edição. Rio de Janeiro: Editora Campus, 2002.

VERGILIO, Silvia Regina. *Utilizando Técnicas de Programação Lógica Indutiva para Mineração de Banco de Dados Relacional.* Dissertação de Mestrado, Pós-Graduação em Informática - Ciências Exatas, Universidade Federal do Paraná, Curitiba – PR, 2001, 86.

WANG, John. *Encyclopedia of Data Warehousing and Data Mining.* Montclair State University, USA: Idea Group Inc, 2006.

WEISS, S. M., e C. A. KULILOWSKI. *Computer Systems That Learn.* Editora: Morgan Kaufman, 1991.

WEISS, S. M., e N. INDURKHYA. *Predictive Data Mining: A Pratical Guide.* 1ª. Edição. Editora: Morgan Kaufmann Publishers, Inc., 1998.

WITTEN, Ian H., e Eibe FRANK. *Data mining : practical machine learning tools and techniques.* 2nd ed. Elsevier Inc, 2005.

WREMBEL, Robert, e Christian KONCILIA. *Data Warehouses and OLAP: Concepts, Architectures and Solutions.* Edição: IRM Press. Hershey PA: Idea Group Inc., 2007.

ZARUR, Marco Aurélio Fernandes. *Modelo para Elaboração de Cenários do Setor Energético, Utilizando Técnicas de Data Mining.* Dissertação de Mestrado, Ciências em Engenharia Civil, COPPE/UFRJ, Rio de Janeiro, 2005, 107.

ZIULKOSKI, L. C. C. *Coleta de Requisitos e Modelagem de Dados para Data Warehouse: um estudo de caso utilizando Técnicas de Aquisição de Conhecimento.* - Relatório. 2003. http://www.inf.ufrgs.br (acesso em 04 de Abril de 2008).

# APÊNDICE

**APÊNDICE A:** Modelagem Dimensional Lógica do Esquema Constelação de Fatos do *Data Warehouse* para o Setor de Saneamento

**APÊNDICE B:** Arquivo de entrada do tipo ARFF para Utilização dos Algoritmos de *Data Mining* pelo Software WEKA

# APÊNDICE A

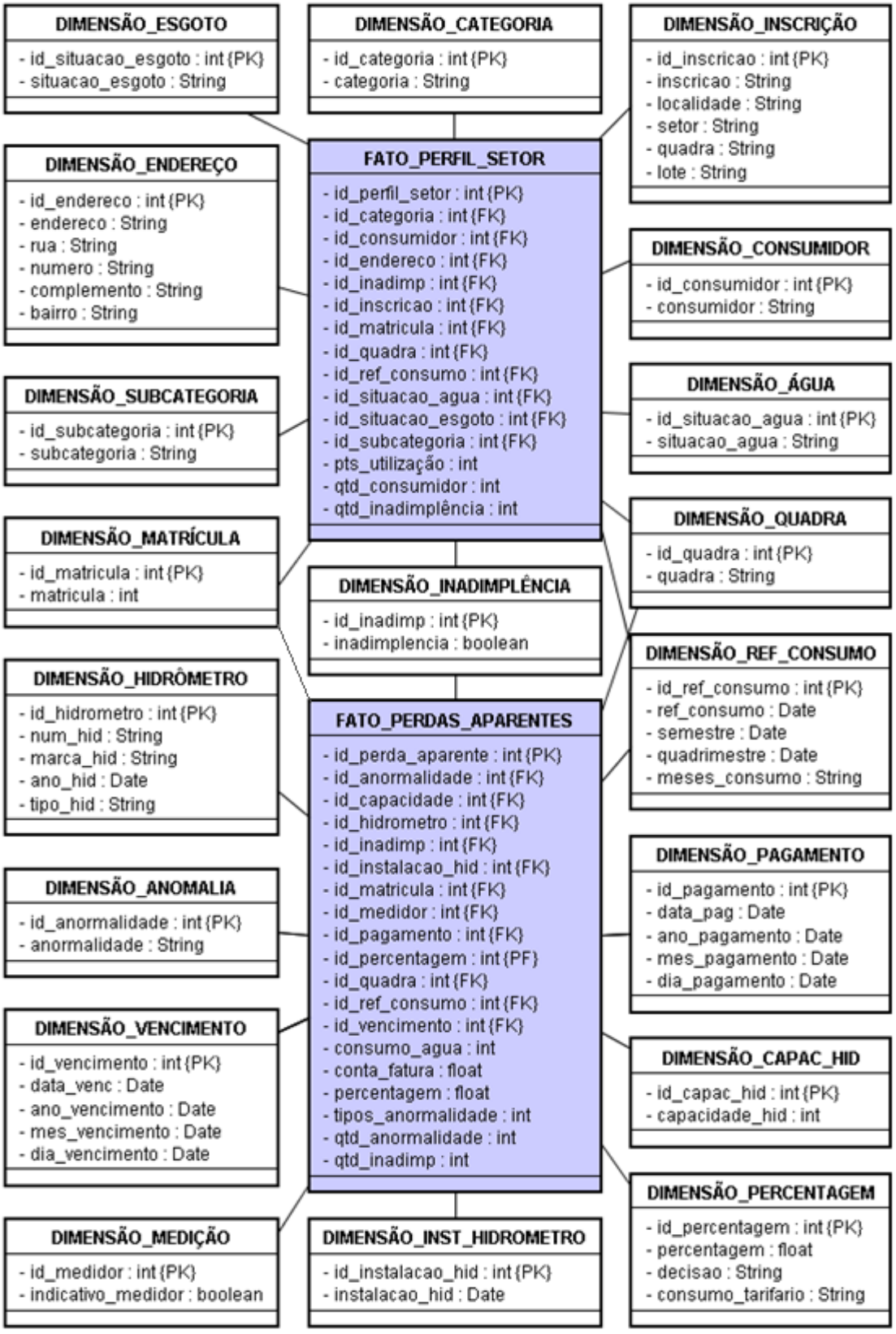



**Figura A.1 - modelagem dimensional do esquema constelação de fatos do *data warehouse***

# APÊNDICE B

**Tabela B.1 - arquivo arff do modelo de *data mining* perfil do setor**

@relation modelo_perfil_setor

@attribute matricula numeric // são 1285 matrículas

@attribute quadra {...} // são79 quadras

@attribute situacao_agua {ligada, cortada, suprimida_total}

@attribute situacao_esgoto {potencial, ligado_normal, factivel}

@attribute categoria {comercial, residencial, industrial, publico}

@attribute subcategoria {...} // são 35 subcategorias, dentre elas: casa, edifício, favela, hospital particular, hotel e repartição pública

@attribute inadimplencia {adimplencia, inadimplencia}

@attribute semestre {primeiro_semestre, segundo_semestre}

@data //conjunto de instâncias a serem mineradas (2.583 instâncias)

161950,quadra_015,cortada,ligado_normal,residencial,casa,adimplencia,primeiro_semestre...

152617,quadra_145,ligada,ligado_normal,residencial,casa,inadimplencia,segundo_semestre...

**Tabela B.2 - arquivo arff do modelo de *data mining* perdas aparentes**

@relation modelo_perda_aparente

@attribute matricula numeric // são 1285 matrículas

@attribute quadra {...} // são79 quadras

@attribute anormalidade {...} // são 24 tipos de anormalidades, dentre eles: sem_anormalidade, hidr_nao_localizado, hidr_impedido_provisoriamente, hidrometro_quebrado, hidrometro_violado, hidrometro_parado e by_pass

@attribute capacidade {ate_3_m3/h, de_5_a_10_m3/h, acima_10_m3/h, nao_informada}

@attribute tipo_hid {mult_magnetico, mult_mecanico, composto, woltmann, outros, nao_informada}

@attribute ano_hid {1984_a_1988, 1989_a_1993, 1994_a_1998, 1999_a_2003, 2004_a_2008, nao_informado}

@attribute inadimplencia {adimplencia, inadimplencia}

@attribute semestre {primeiro_semestre, segundo_semestre}

@attribute media_consumo numeric

@attribute media_conta numeric

@attribute consumo_tarifario {ate_5m3_comercial, ate_10m3_comercial, acima_de_10m3_comercial, ate_10m3_residencial, entre_10_e_20m3_residencial, entre_20_e_30m3_residencial, acima_de_30m3_residencial, , ate_10m3_industrial, acima_de_10m3_industrial, ate_10m3_publico , acima_de_10m3_publico}

@attribute indicativo_medidor {com_medidor, sem_medidor}

@attribute data_inst_hid {menos_de_3_anos, entre_3_e_9_anos, entre_10_e_18_anos,

@attribute decisao {desfavoravel_a_troca, favoravel_a_troca, analise_mais_detalhada}

mais_18_anos, nao_informada}

@data // 3.523 instâncias

1088866, quadra_010, sem_anormalidade, de_5_a_10_m3/h, mult_magnetico, 1994_a_1998, adimplencia, primeiro_semestre, 190, 982.5, com_medidor, entre_3_e_9_anos, desfavoravel_a_troca ...

154385, quadra_305, by_pass, ate_3_m3/h, mult_magnetico, 2004_a_2008, adimplencia, primeiro_semestre, 6.1, 28.3, ate_10m3_residencial, com_medidor, entre_3_e_9_anos, analise_mais_detalhada ...