

UNIVERSIDADE FEDERAL DA PARAÍBA  
CENTRO DE TECNOLOGIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

TARCÍSIO FERREIRA GRILO JÚNIOR

Aplicação de técnicas de *Data Mining* para auxiliar no  
processo de fiscalização no âmbito do Tribunal de Contas  
do Estado da Paraíba

JOÃO PESSOA – PB  
2010

UNIVERSIDADE FEDERAL DA PARAÍBA  
CENTRO DE TECNOLOGIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

TARCÍSIO FERREIRA GRILO JÚNIOR

Aplicação de técnicas de *Data Mining* para auxiliar no  
processo de fiscalização no âmbito do Tribunal de Contas  
do Estado da Paraíba

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do Centro de Tecnologia da Universidade Federal da Paraíba, como requisito para obtenção do grau de Mestre em Engenharia de Produção.

**Área de concentração:** Gestão da Produção

**Subárea:** Tecnologia

**Professor orientador:** Dr. Luiz Bueno da Silva

JOÃO PESSOA – PB  
2010

M357r Grilo Júnior, Tarcísio Ferreira

Aplicação de técnicas de Data Mining para auxiliar no processo de fiscalização no âmbito do Tribunal de Contas do Estado da Paraíba / Tarcísio Ferreira Grilo Júnior- João Pessoa: UFPB, 2010.

102f.

Orientador: Dr. Luiz Bueno da Silva

**Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal da Paraíba, Centro de Tecnologia.**

1. Business Intelligence 2. Data mining 3. Descoberta de conhecimento em bases de dados

I.Título.

CDU 378: 91(043)

# TARCÍSIO FERREIRA GRILO JÚNIOR

## Aplicação de técnicas de *Data Mining* para auxiliar no processo de fiscalização no âmbito do Tribunal de Contas do Estado da Paraíba

Dissertação julgada e **aprovada** em **03** de **setembro** de **2010** como parte dos requisitos necessários para a obtenção do título de Mestre em Engenharia de Produção no Programa de Pós-graduação em Engenharia de Produção da Universidade Federal da Paraíba.

**Área de concentração:** Gestão da Produção

**Subárea:** Tecnologia

### **BANCA EXAMINADORA**

Prof. Luiz Bueno da Silva, Doutor.  
Universidade Federal da Paraíba  
Orientador

Prof. Ricardo Moreira da Silva, Doutor.  
Universidade Federal da Paraíba  
Examinador

Profª. Adriana Zenaide Clericuzi, Doutora.  
Universidade Federal de Pernambuco  
Examinadora Externa

A Deus pela conquista desta etapa em minha vida.

A meus pais, pela dedicação e incentivo.

As minhas filhas, Mariana e Giovanna.

A minha esposa, Sandra, com amor.

## **AGRADECIMENTOS**

Agradeço a Deus pelo dom da vida e pela força atribuída em todos os momentos. Suas bênçãos foram essenciais ao longo desta caminhada.

À minha família, minha mãe, meu pai, minha esposa e minhas filhas.

Ao professor Dr. Fabrício Barth, pela admirável dedicação, pela atenção concedida em todos os momentos e pelas críticas e contribuições a este estudo.

À professora Dra. Natasha Queiroz Lino pelos valiosos ensinamentos, pelas sugestões sempre oportunas e construtivas.

Ao professor Dr. Luiz Bueno da Silva pelos momentos de incentivo, dedicação, entusiasmo, e competência na condução de suas atividades.

Em especial a professora Dra. Adriana Zenaide Clericuzi pela dedicação incondicional ao magistério, a quem expresso minha admiração.

À Ana Araújo e a equipe da secretaria do PPGEP, pela costumeira atenção e contribuição, seja nas funções de rotina, seja com palavras de apoio.

À equipe de dirigentes e servidores do TCE-PB, principalmente ao Diretor Geral Severino Claudino Neto, ao Assessor Técnico Chefe Ed Wilson Fernandes de Santana, as colegas Rodrigo Galvão, Marcos Uchôa e demais membros da Assessoria Técnica do TCE-PB (ASTEC) pelo apoio e contribuição para a realização desta pesquisa.

Agradeço, ainda, às muitas pessoas que não têm seus nomes citados, mas que contribuíram para mais uma realização em minha vida.

“Os computadores são incrivelmente rápidos, precisos e burros;  
os homens são incrivelmente lentos, imprecisos e brilhantes;  
juntos, seu poder pode ultrapassar os limites da imaginação.”

Albert Einstein

“A coisa mais indispensável a um homem é reconhecer o uso  
que deve fazer do seu próprio conhecimento.”

Platão

## RESUMO

Esta pesquisa tem como objetivo validar a hipótese da aplicabilidade das técnicas de mineração de dados na base de dados de Licitação e Contratos gerenciada pelo Tribunal de Contas do Estado da Paraíba (TCE-PB), possibilitando a geração de regras e descoberta de conhecimento oculto ou implícito, contribuindo desta forma com o processo de tomada de decisão, fiscalização e celeridade processual no âmbito desta Corte de Contas.

Para melhor compreensão desse trabalho foi realizada uma revisão de literatura abordando primeiramente um histórico sobre o processo de decisão, bem como a evolução dos estudos deste tema e da relação entre os processos licitatórios enviados ao TCE-PB e o processo de descoberta de indícios de fraudes e irregularidades através do uso de mineração de dados.

São abordados os conceitos sobre a tecnologia de *Business Intelligence* (BI) e dos seus principais componentes, bem como os conceitos de Descoberta de Conhecimentos em Bases de Dados (*Knowledge Discovery in Databases*), e uma comparação das funcionalidades presentes nas ferramentas de mineração de dados. Espera-se com a implantação desta ferramenta de mineração de dados, um ganho de produtividade e um aumento na celeridade do tramite processual decorrentes da análise das contas públicas e na fiscalização do erário.

Palavras-chave: *Business Intelligence*, *Data mining*, Regras de Associação, Controle externo, Descoberta de conhecimento em bases de dados.



## **ABSTRACT**

This search has as goal to validate the hypothesis of the applicability of data mining techniques in Bidding and Contracts' database managed by the Account Court of Paraiba State, enabling the generation of rules and discovery of hidden knowledge or implicit, contributing to the process of decision making, supervision and celerity in this Court of Auditors.

To the best comprehension of this work, It was made a literature revision bringing at first place a historic vision about the decision process, as well as this theme evolution studies and the relation between the tender processes sent to Account Court of Paraiba State and the fraud indication discovery process and irregularities through the data mining process using.

We will bring to light the concept of Business Intelligence (BI) and for it's main components, as well as the concepts of knowledge discovery in database, and a comparing between the using of the instruments of data mining. We expect from this implant of the data mining an increase in the productivity and also an increase in speed of lawsuit process from the public accounts analysis and public money fiscal control.

Keys-words: Business Intelligence, Data Mining, Association Rules, External Control, Database Knowledge Discovery.

## LISTA DE ABREVIACÕES

API	<i>Application Programming Interface</i>
BI	<i>Business Intelligence</i>
CNPJ	Cadastro Nacional da Pessoa Jurídica
CPF	Cadastro de Pessoas Físicas
CRM	<i>Customer Relationship Management</i>
CSV	<i>Comma-Separated Values</i>
DM	<i>Data Mining</i>
EIS	Executive Information System
ETL	<i>Extraction Transforming and Loading</i>
FGV	Fundação Getúlio Vargas
GPL	<i>General Public License</i>
IA	Inteligência Artificial
KDD	<i>Knowledge Discovevy in Databases</i>
LRF	Lei de Responsabilidade Fiscal
OLAP	<i>On-line Analytical Processing</i>
PCP	Planejamento e Controle de Produção
PROMOEX	Programa de Modernização do Sistema de Controle Externo dos Estados, Distrito Federal e Municípios Brasileiros
SAD	Sistema de Apoio à Decisão
SAGRES	Sistema de Acompanhamento da Gestão dos Recursos da Sociedade
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	<i>Structured Query Language</i>
TCE	Tribunal de Contas do Estado
TI	Tecnologia da Informação
XML	<i>Extensible Markup Language</i>

## LISTA DE FIGURAS

Figura 1: Potifólio de metodologias disponíveis em um sistema de <i>Business Intelligence</i> .....	27
Figura 2: Proposta da arquitetura de Inteligência de Negócio .....	28
Figura 3: Ciclo da análise de BI Fonte: Adaptado de Versellis (2009).....	29
Figura 4: O processo de ETL .....	30
Figura 5: Estrutura interna de um DW .....	32
Figura 6: Uma visão geral das etapas que compõem o processo de KDD .....	35
Figura 7: Mineração de dados como uma confluência de muitas disciplinas .....	37
Figura 8: Valor das Informações x Investimentos em BI.....	42
Figura 9: Exemplo de uma Regra de Associação.....	46
Figura 10: Diagrama de Contexto do SAGRES .....	59
Figura 11: Tabela de dados do módulo de Licitações utilizadas pelo SAGRES.....	63
Figura 12: Adaptado do demonstrativo de funcionamento do SAGRES Captura .....	67
Figura 13: Interface principal do aplicativo Sagres Auditor .....	68
Figura 14: Consulta de licitações no módulo Sagres Auditor .....	69
Figura 15: Detalhamento da licitação no módulo Sagres Auditor .....	70
Figura 16: Ambiente de trabalho do RapidMiner .....	72
Figura 17: <i>Data mart</i> de Licitações.....	73
Figura 18: Área de trabalho principal do RapidMiner com o os objetos usados no experimento realizado .....	76
Figura 19: Resultado do processamento dos dados minerados pelo RapidMiner .....	80
Figura 20: Regras de associação do exercício de 2005 geradas pelo RapidMiner.....	80
Figura 21: Grafo gerado pelo RapidMiner com as regras de associação de um proponente .....	82
Figura 22: Conjunto de dados encontrados pelo RapidMiner com os alterações nos parâmetros de filtro .....	85

## LISTA DE QUADROS

Quadro 1: Ferramentas de ETL.....	31
Quadro 2: Comparativo entre ferramentas utilizadas para a mineração de dados.....	43
Quadro 3: Métodos de mineração de dados que podem ser aplicados em cada tarefa de KDD .	44
Quadro 4: Base de dados da tarefa exemplo “Jogar tênis” .....	51
Quadro 5: Um exemplo de transações em cesta de compras. Fonte: Adaptado de Pizzi (2006).	52
Quadro 6: Índícios de irregularidade nos processos licitatórios.....	64
Quadro 7: Processos licitatórios enviados ao TCE-PB por tipo de modalidade de licitação .....	65
Quadro 8: Quantidade de participantes por tipo de modalidade de licitação .....	66
Quadro 9: Relação Licitação versus Proponentes .....	75
Quadro 10: Matriz Esparsa Licitações <i>versus</i> Proponentes .....	75
Quadro 11: Distribuição dos proponentes nos processos licitatórios correspondente ao ano de 2005 .....	79
Quadro 12: Medidas de interesse objetivas geradas no exercício de 2005 .....	82
Quadro 13: Filtro realizado por proponente nas regras de associação geradas .....	82
Quadro 14: Definição dos principais indicadores utilizados para análise dos dados minerados	83
Quadro 15: Medidas de interesse objetivas no exercício de 2006.....	84
Quadro 16: Medidas de interesse objetivas no exercício de 2007.....	84
Quadro 17: Medidas de interesse objetivas no exercício de 2008.....	84
Quadro 18: Medidas de interesse objetivas no exercício de 2009.....	85
Quadro 19: Filtro realizado por proponente nas regras de associação geradas em 2009 .....	86

## SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO.....	14
1.1. Definição do problema de pesquisa .....	14
1.2. Justificativa .....	16
1.3. OBJETIVOS .....	18
1.3.1. Objetivo Geral .....	18
1.3.2. Objetivos Específicos.....	18
1.4. Relevância do estudo .....	18
1.5. Estrutura da dissertação .....	19
CAPÍTULO 2 - REFERENCIAL TEÓRICO.....	21
2.1. A Gestão da Informação e a Inteligência nos Negócios .....	21
2.1.1. A informatização e os grandes volumes de dados .....	22
2.2. <i>Business Intelligence</i> .....	23
2.2.1. Principais componentes da arquitetura de um BI.....	27
2.2.1.1. ETL ( <i>Extraction Transforming and Loading</i> ):.....	30
2.2.1.2. Data Warehouse (DW): .....	31
2.2.1.3. Data Mart.....	32
2.2.1.4. OLAP ( <i>On-line Analytical Processing</i> ): .....	33
2.3. Descoberta de Conhecimentos em Base de Dados .....	34
CAPÍTULO 3 – DATA MINING .....	39
3.1. Principais ferramentas para a mineração de dados.....	43
3.2. Algoritmos para mineração de dados .....	44
3.3. Aplicações de mineração de dados .....	49
3.4. Exemplos de aplicação das técnicas de <i>data mining</i> .....	51
3.5. <i>Cases</i> de sucesso com <i>data mining</i> .....	53
CAPÍTULO 4 - CONTROLE NO TRIBUNAL DE CONTAS .....	55
4.1. O uso das tecnologias de BI nos Tribunais de Contas .....	57
4.2. O Sistema de Acompanhamento da Gestão dos Recursos da Sociedade (SAGRES) .....	58
CAPÍTULO 5 - ASPECTOS METODOLÓGICOS DA PESQUISA.....	61
5.1 Classificação da pesquisa .....	61
5.2 Delimitação do tema .....	62

5.3 Seleção dos dados da pesquisa ou caso.....	62
5.3.1 Principais indícios de fraudes e/ou irregularidades em licitações.....	63
5.3.2 Critérios da amostra ou caso .....	65
5.4 Coleta dos dados .....	65
5.5 Análise dos dados.....	71
5.6 Método abordado .....	72
5.7 Roteiro da pesquisa .....	73
5.7.1 Criação do modelo de dados .....	74
5.7.2 Procedimento realizado .....	75
CAPÍTULO 6 – ANÁLISE DOS RESULTADOS .....	78
6.1 Ganhos esperados com a implantação de técnicas de mineração de dados no Tribunal de Contas do Estado .....	86
CAPÍTULO 7 – CONSIDERAÇÕES FINAIS .....	87
7.1 Recomendações para trabalhos futuros.....	89
REFERÊNCIAS BIBLIOGRÁFICAS .....	90
APÊNDICE I .....	98
APÊNDICE II.....	99

## CAPÍTULO 1 - INTRODUÇÃO

Este capítulo tem a finalidade de apresentar o problema de pesquisa, a justificativa do tema, os objetivos gerais e específicos, a estrutura do trabalho e as limitações da pesquisa.

### 1.1. Definição do problema de pesquisa

Segundo Turban *et al.* (2009), o ambiente de negócios no qual as organizações operam atualmente está se tornando cada vez mais complexo e mutante. As instituições, públicas e privadas, são submetidas a pressões crescentes no sentido de responder rapidamente a condições que estão sempre em transformação.

As organizações governamentais também passaram por profundas transformações em decorrência da globalização da economia e dos avanços tecnológicos, não apenas quanto à produção de bens e serviços, mas especialmente no que dizem respeito às Tecnologias da Informação (TI), no sentido de coordenar e otimizar o fluxo de informações relevantes aos gestores para a tomada de decisão (BEUREN; GIBBON, 2001).

Freitas Jr. (2003) afirma que a Tecnologia da Informação torna-se extremamente útil ao processo decisório, pois possibilita coletar, processar e armazenar dados pela utilização de sistemas de informação.

O papel desempenhado pelos Sistemas de Informação (SI) para se adequar a estas novas exigências empresariais vem se modificando acentuadamente: de um papel operacional, para um papel cada vez mais estratégico (MARCHIORATO, 2006). Segundo O'Brien (2004) os sistemas de informações estratégicos tem a função de apoiar ou moldar a posição e as estratégias competitivas de uma empresa, podendo ser qualquer tipo de SI que contribua para a obtenção de vantagem estratégica ou competitiva, ou pela diminuição de uma desvantagem estratégica que permita à empresa alcançar seus objetivos.

O'Brien (2004) complementa afirmando que, para atender de forma efetiva às crescentes exigências de informação, os sistemas de informação, ao longo do tempo, evoluíram de uma fase inicial na qual eram utilizados prioritariamente para a

informatização de processos (substituição de processos manuais por processos computadorizados), para um papel relevante no suporte à decisão.

Neste cenário, a informação e o conhecimento tem um papel de extrema importância, pois apoia as decisões estratégicas das organizações. Assim, a informação e o conhecimento tem um valor alto, porém exigem das empresas adequação sistematizada e organizada de seu uso, no qual os Sistemas de Informação assumem papéis cada vez mais revelantes e estratégicos para as mesmas.

No contexto do desenvolvimento da tecnologia de informação, soluções de Inteligência de Negócios ou *Business Intelligence* (BI), como são amplamente conhecidos, visam oferecer os meios necessários para a transformação de dados em informação, proporcionando o suporte necessário para o processo decisório e flexibilização das tarefas.

Barbieri (2001) conceitua de forma mais ampla que o *Business Intelligence* pode ser entendido como a utilização de variadas formas de informação para se definir estratégias de competitividade nos negócios da empresa, e aponta o BI como uma evolução e consolidação dos sistemas que envolvem apoio ao processo de tomada de decisões, agregando mais inteligência ao processo.

Uma das etapas do processo de BI é a mineração de dados ou *data mining* (DM), que surgiu no início da década de 1990, como uma alternativa promissora para a análise de grandes volumes de dados (TULER *et al.*, 2006). Segundo Cardoso e Machado (2008), o *data mining* usa ferramentas de análise estatística, assim como técnicas da área de inteligência artificial, ou técnicas baseadas em regras e outras técnicas inteligentes. Ainda segundo os mesmos autores, a mineração dos dados pode dar-se sobre um banco de dados operacional, ou sobre um *data warehouse*, constituindo um sistema de suporte à decisão.

De acordo com Tan, Steinbach e Kumar (2009), as técnicas de mineração de dados podem ser usadas para apoiar uma ampla gama de aplicações de inteligência de negócio como a criação de perfis de clientes, administração do fluxo de trabalho e detecção de fraudes. Conjugando as técnicas provenientes de diversas áreas como estatística, banco de dados e inteligência artificial, a mineração de dados se diferencia das demais técnicas de análise por seu caráter exploratório.



Apesar das diversas técnicas e ferramentas disponibilizadas por diversos fabricantes, os conceitos de *Business Intelligence* não são amplamente conhecidos.

Neste trabalho será estudado o levantamento de possíveis fraudes decorrentes dos processos licitatórios através das técnicas de mineração de dados aplicada nos órgãos de controle externo, com ênfase para a fiscalização e análise no âmbito de competência do Tribunal de Contas do Estado da Paraíba (TCE-PB).

## 1.2. Justificativa

O exame das contas dos entes públicos brasileiros, conforme determinação do artigo 70 da Constituição Federal de 1988, é efetuado pelo controle interno e pelo controle externo. O controle externo é efetivado pelo Poder Legislativo, com auxílio dos Tribunais de Contas, ao passo que o controle interno é levado a efeito por sistema do próprio ente público (CARVALHO, 2002).

O objetivo da maioria das iniciativas de mudança é a melhoria de algum indicador produtivo da organização, elas surgem devido a mudanças nos requisitos de desempenho, novas oportunidades, novas preferências de consumidores, concorrências, novos regulamentos, entre outros. Com isso, durante o ciclo de vida de um negócio, muitas melhorias ou ciclos de transformação podem ser necessários (SOUSA; GROESBECK, 2004).

Diante deste fato, baseado em um estudo feito a partir dos resultados de uma pesquisa realizada pela FIA-USP no ano de 2001, Moraes (2006) afirma que, de forma geral, os Tribunais de Contas apresentam defasagem gerencial e tecnológica, com uso pontual e acessório da tecnologia da informação, traduzindo-se em trabalho excessivamente manual e artesanal (análise e produção de documentos em papel) produzindo estoques de processos, morosidade e ineficácia.

Ainda segundo Moraes (2006), a tecnologia da informação é utilizada fundamentalmente para automatizar etapas dos procedimentos, tais como: receber informações dos jurisdicionados, na automação de escritórios e no acompanhamento de processos; pouco alterando sua eficácia, eficiência e tempo de ciclo. Demonstra desta forma, a predominância do modelo burocrático de trabalho com controles *a posteriori* e tramitação lenta de processos.

Embora esta defasagem venha diminuindo com o passar do tempo, principalmente devido à redução dos custos de aquisição de equipamentos de informática e as diversas ferramentas de software disponíveis no mercado, os Tribunais de Contas passaram para um estágio de informatização avançado, mas ainda mantem muitos dos seus processos com a mesma sistemática de anos anteriores, ou seja, mantendo o fluxo manual de análise, não usufruindo da totalidade dos recursos tecnológicos existentes (GRILO JR.; CLERICUZI, 2008).

Para Vieira (2008), a realidade é que os Tribunais de Contas brasileiros ainda operam com muitas deficiências, evidenciado, por exemplo, pelas pesquisas associadas ao PROMOEX (Programa de Modernização do Sistema de Controle Externo dos Estados, Distrito Federal e Municípios Brasileiros) e realizadas pela FIA-USP e FGV (Fundação Getúlio Vargas).

O PROMOEX tem como objetivo a modernização e o fortalecimento institucional do sistema de controle externo no âmbito dos estados e municípios brasileiros, de forma a garantir a sua integração nacional e aumentar a eficiência e eficácia das ações de fiscalização e controle, contribuindo para a efetiva e regular aplicação dos recursos públicos estaduais e municipais, em benefício da sociedade.

A concepção do programa se deu a partir de um diagnóstico amplo acerca da real situação das Cortes de Contas no Brasil e cada Tribunal de Contas desenvolveu um projeto específico contemplando suas necessidades institucionais e organizacionais observando as linhas gerais estabelecidas no programa (TCE-PI, 2010; TCE-CE, 2010). Apesar dos investimentos que estão sendo feitos pelos TCE's, ainda há muito que se avançar para a implantação de instrumentos de controle, objetivando evitar dano ao erário e ao patrimônio público.

Desta forma esta pesquisa justifica-se por relacionar os possíveis indícios de fraudes em processos licitatórios oriundos da implantação de técnicas de data mining em uma organização do setor governamental, em especial no âmbito do Tribunal de Contas do Estado da Paraíba (TCE-PB), não focando apenas aspectos operacionais do TCE-PB, mas também o valioso processo analítico das operações eletronicamente armazenadas, o acúmulo de informações e o seu uso potencial para análise e tomada de decisão. Desta forma uma pergunta pode ser feita: a implantação de uma ferramenta de

mineração de dados pode contribuir na identificação de fraudes em processos licitatórios analisados pelo TCE-PB?

### 1.3. OBJETIVOS

#### 1.3.1. Objetivo Geral

Este estudo tem como objetivo: Aplicar técnicas de *data mining* no âmbito de competência do Tribunal de Contas do Estado da Paraíba com vistas a relacionar possíveis indícios de fraudes em processos licitatórios.

#### 1.3.2. Objetivos Específicos

Para se atingir os objetivos gerais, propõem-se os seguintes objetivos específicos:

- explorar os conceitos de técnicas de *data mining*;
- realizar uma investigação através de uma pesquisa exploratória com vistas a conhecer os fatos e fenômenos relacionados a implantação das técnicas de *data mining* para a organização;
- identificar e analisar a aplicabilidade das técnicas de *data mining* para o processo de fiscalização nos órgãos de controle externo, com observação especial na aplicação desta sistemática no âmbito de competência do Tribunal de Contas do Estado da Paraíba.

#### 1.4. Relevância do estudo

Segundo Romão (2002), a Engenharia de Produção é uma área de pesquisa multidisciplinar que tradicionalmente envolve disciplinas da produção civil, elétrica e mecânica. Ultimamente tem envolvido a disciplina de produção de software, a qual possui estreita relação com as demais disciplinas das engenharias. Ainda segundo o autor, problemas de produção existem em qualquer empresa pública ou privada.

Essa pesquisa é relevante tanto para o meio acadêmico quanto para o meio governamental. No meio acadêmico, a sua importância reside no fato de que tem validade no entendimento do desempenho da sistemática processual dos auditores da organização. A partir desse entendimento serão possíveis, em trabalhos científicos

posteriores, o desenvolvimento e refinamento através dos resultados obtidos com a implantação desta proposta.

Para o meio governamental, pretende-se proporcionar subsídios aos profissionais da área de auditoria e demais interessados em analisar os dados oriundos desta pesquisa, como por exemplo, os órgãos de fiscalização e controle externo como a Polícia Federal, Controladoria Geral da União, Ministério Público Federal, entre outros órgãos públicos federais, estaduais e municipais que buscam subsídios nos dados do SAGRES, para que a partir de associações e indicadores, proporcionem os meios necessários para uma investigação mais detalhada, de acordo com as interações proporcionadas pela ferramenta de BI adotada.

Neste trabalho, pretende-se também contribuir para aprofundar o conhecimento sobre o uso das técnicas de mineração de dados em organização de controle externo. Neste contexto, a importância deste estudo está em colaborar com a identificação de uma sistemática para a detecção de indícios de irregularidade e fraudes nos processos licitatórios. Esta dissertação pretende agilizar o trabalho realizado pelos auditores e analistas de controle no desempenho do trabalho de fiscalização do erário público.

Além disso, a contribuição deste trabalho se verifica na captação de informações relevantes para a organização em foco neste estudo, que passa a ter meios de melhor compreender o processo de contratação fraudulenta através de licitações que objetivam o desvio de recursos públicos.

### 1.5. Estrutura da dissertação

Esta dissertação foi organizado em 7 (sete) capítulos, distribuídos da seguinte forma:

O Capítulo 1 foi apresentado o problema de pesquisa, os objetivos geral e específicos, sendo feita a justificativa do estudo em termos de relevância e a dificuldade.

O Capítulo 2 apresenta uma revisão bibliográfica sobre a tecnologia de BI, mineração de dados, tarefas de *Knowledge-Discovery in Databases* (KDD). O Capítulo 3 apresenta em detalhes o *Data Mining*, aplicações e as principais ferramentas para mineração de dados disponíveis no mercado, que se fazem necessários à contextualização do objetivo proposto.

Capítulo 4 é apresentado à função constitucional dos Tribunais de Contas, e no Capítulo 5 descreve sobre o procedimento metodológico para a realização da desta pesquisa, o qual evidencia em que consiste e a forma como a pesquisa foi desenvolvida e conduzida, fornecendo detalhes sobre as técnicas de pesquisa utilizadas, de forma a orientar o leitor ou outros pesquisadores acerca dos caminhos percorridos e dos resultados alcançados.

O Capítulo 6 apresenta o resultado da pesquisa, ou seja, visa responder o problema proposto para o presente estudo, descrevendo como ocorre a utilização das ferramentas mineração de dados para a descoberta de conhecimento escondido na base de dados dos processos licitatórios, de forma a apoiar o processo decisório.

Por fim, o Capítulo 7 apresenta as considerações finais sobre este estudo e oferece algumas recomendações para trabalhos futuros, identificadas durante a realização da pesquisa.

## CAPÍTULO 2 - REFERENCIAL TEÓRICO

Este capítulo apresenta o referencial teórico empregado na pesquisa, abordando os seguintes aspectos: a gestão da informação e a inteligência nos negócios, conceito de BI e principais componentes da sua arquitetura, principais ferramentas e algoritmos para a mineração de dados. No Capítulo 3 será apresentada a contextualização do ambiente pesquisado e suas competências constitucionais.

### 2.1. A Gestão da Informação e a Inteligência nos Negócios

Segundo Freitas Jr. (2003), nos últimos anos as organizações tem valorizado a informação como recurso estratégico e fonte de vantagem competitiva, desta forma, a coleta, o processamento, a análise e a disseminação de informação sobre tendências e demandas do mercado, de competidores e das necessidades dos clientes, devem ser apropriadas para os gestores, visando alimentar o planejamento e a execução de suas ações para obter vantagem competitiva.

Ainda segundo Freitas Jr. (2003), a Tecnologia da Informação, pode ser entendida com a utilização de recursos tecnológicos e computacionais para a geração e uso da informação diante do crescente desafio de proporcionar aos tomadores de decisão a apresentação de informações precisas, oportunas e relevantes.

Para Gouveia (2002), o problema central da Gestão da Informação parece ser o excesso de dados (*data overload*). Tomando o exemplo da Internet (mais precisamente da *World Wide Web*), é cada vez mais complexo determinar a existência e localização de recursos num repositório que é caótico e desestruturado. Este esforço é ainda maior por ser necessário aprender as diferentes formas de utilização e comportamento de diferentes programas disponíveis, além do excesso de resultados retornados, após cada pesquisa efetuada.

Semelhante modo ao exemplo da internet, da mesma forma as organizações coletam os dados necessários ao seu funcionamento e vão acumulando durante anos, ocasionando um aumento no volume de dados armazenados, forçando-as a adotarem políticas para a guarda e segurança das informações ali contidas. Para Romão (2002), a grande quantidade de informações nos bancos de dados informatizados das instituições pode esconder conhecimentos valiosos e úteis para a tomada de decisão.

No entanto, algumas pessoas questionam se o problema reside de fato no excesso de dados. Pode ser que o verdadeiro problema esteja na multiplicidade de canais de comunicação existentes. De modo diferente de outras eras, as novas tecnologias de informação não estão substituindo, mas complementando a oferta de escolhas (GILSTER, 1997).

O aumento acentuado no volume dos dados, associado à crescente demanda por conhecimento novo para decisões estratégicas, tem provocado o interesse crescente em descobrir conhecimento em banco de dados.

A informação chega agora a todos e de múltiplas formas. Infelizmente, a maioria das práticas de Gestão da Informação são demasiadamente lineares e específicas – foram técnicas desenvolvidas para um fluxo e não para um oceano de dados (ALESANDRINI, 1992), o maior problema reside no excesso de dados disponíveis, que muitas vezes não são utilizados adequadamente ou descartados devido a não apresentarem significado útil para o embasamento da tomada de decisão. Segundo Romão (2002), o aumento no volume dos dados, associado à crescente demanda por conhecimento novo voltado para decisões estratégicas, tem provocado o interesse crescente em descobrir novos conhecimentos em banco de dados.

#### 2.1.1. A informatização e os grandes volumes de dados

Para Araújo Júnior & Alvares (2007), as organizações necessitam administrar o seu fluxo informacional a fim de buscar diferenciação e efetividade em seus processos decisórios. E a sobrecarga de informação, ou seja, o acúmulo aleatório de dados nas organizações, não mais representa vantagem, ao contrário, a coleta sistemática de dados só se justifica se houver aproveitamento em uma situação ou decisão previamente selecionada, senão a acumulação de dados e de papéis sem destinação clara comprometerá a correta gestão do fluxo da informação.

O excesso de informação está associado à perda de controle sobre a informação e à incapacidade em usar efetivamente a informação (TERRA; BAX, 2003), em resumo, pode-se dizer que a informação disponível considerada útil pode se tornar um entrave para a organização e ao processo de tomada de decisão se não for corretamente gerenciada.

Terra e Bax (2003) citam que segundo alguns autores, por volta dos anos 50 do século XX a própria ciência corria risco. Porém, apenas ao final dos anos 50 e início dos

anos 60 é que o estado de sobrecarga informacional é compreendido como um problema real.

Nos anos 90 torna-se um problema grave, particularmente no mundo dos negócios. Tecnologias como as encontradas nas bases de dados, hiperdocumentos, arquivos numéricos, simuladores, captadores numéricos, telepresença, realidades virtuais, inteligência artificial, modelização de fenômenos complexos entre outras, permitem novas formas de acesso à informação e novos estilos de raciocínio e de conhecimento que caracterizam o espaço do saber (FACHINELLI, RECH E MATTIA, 2006). Assim, a Tecnologia da Informação evoluiu de uma orientação tradicional de suporte administrativo para um papel estratégico dentro das organizações.

Para Fachinelli, Rech e Mattia (2006), torna-se evidente que a gestão da informação por si só não é suficiente para os usuários da informação nos mais diferentes níveis decisórios, se o seu produto não fizer sentido, se não for compreendido no contexto da organização. Ainda segundos os autores, esta compreensão está vinculada aos processos de criação de sentido que conduzem para a aprendizagem.

Para O'Brien (2006), é preciso que a organização esteja atenta às novas disponibilidades em Tecnologia da Informação e que haja uma integração entre suas estratégias e sua estrutura, tanto de tecnologia da informação como de negócio. O mercado tecnológico apresenta algumas ferramentas que buscam consolidar os processos e ativos organizacionais pela relação da eficácia e eficiência de gestão, tais como o *Business Intelligence*.

De acordo com Batista (2006) o sistema de *Business Intelligence* consiste de um conjunto de ferramentas e aplicativos que oferece aos tomadores de decisão possibilidade de organizar, analisar, distribuir e agir, ajudando a organização a tomar decisões melhores e mais dinâmicas. Para o autor, as ferramentas de visão completa do negócio ajudam na distribuição uniforme dos dados entre os usuários, não importando onde esses dados estejam. O objetivo das ferramentas é transformar grandes quantidades de dados em informações de qualidade para a tomada de decisões, gerando, então, resultados diretos para a empresa.

## 2.2. *Business Intelligence*

De acordo com Turban *et al.* (2009), o processamento das informações, na estrutura das decisões necessárias, deve ser feito de forma rápida, e comumente com



algum apoio computadorizado. A tecnologia de BI contempla um novo conceito de estruturação das informações, visando gerar valor ao negócio.

Processos baseados em BI utilizam tecnologias para coletar, armazenar, analisar e disponibilizar o acesso aos dados, transformando-os em informação e conhecimento (BARBIERI, 2001). O mesmo autor conceitua, de forma mais ampla, o *Business Intelligence* como a utilização de variadas fontes de informação para se definir estratégias de competitividade nos negócios da empresa, e aponta o BI como uma evolução e consolidação dos sistemas que envolvem apoio ao processo de tomada de decisões, agregando mais inteligência ao processo. Pinto (2010) conceitua BI da seguinte forma:

*Business Intelligence* é um conjunto de processos que você realiza no seu negócio de modo a torná-lo mais eficiente. A eficiência vem do fato de que suas decisões partem de dados que são analisados. Ao invés de decidir por intuição ou pela percepção do que acontece, você passa a decidir baseado em informações, em dados que são colhidos no seu próprio negócio e fora dele. A maioria dos dados é obtida por meio de ferramentas de automação comercial (PINTO, 2010).

Na abordagem administrativa, a Inteligência de Negócio (IN), tradução do termo *Business Intelligence*, é vista como um processo em que os dados internos e externos da empresa são integrados para gerar informação pertinente para o processo de tomada de decisão. O papel da inteligência de negócio é criar um ambiente informacional com processos através dos quais dados operacionais possam ser coletados, tanto dos sistemas transacionais como de fontes externas, e analisados, revelando dimensões “estratégicas” do negócio (PETRINI, 2006). Para Turban *et al.* (2009), BI é uma expressão livre de conteúdo, portanto, significa coisas diferentes para pessoas diferentes.

Segundo Filho (2010), BI é um conceito, um modelo que se presta ao atendimento de pessoas que ocupam posições estratégicas dentro das organizações, que estão diretamente ligadas ao negócio e que possuem poder de decisão e influencia sobre os rumos das empresas.

De acordo com Serra (2002), um sistema de BI deve apresentar as seguintes características:

- extrair e integrar dados de múltiplas fontes;
- fazer uso da experiência;

- analisar dados contextualizados;
- trabalhar com hipóteses;
- procurar relações de causa e efeito;
- transformar os registros obtidos em informação útil para o conhecimento empresarial.

Como parte da solução de TI, o BI vem sendo trabalhado como uma excelente opção a ser utilizada em um Sistema de Apoio a Decisão (SAD), proporcionando ao tomador de decisões os meios necessários para obter os melhores resultados possíveis, além de permitir maior integração entre os diversos atores envolvidos neste processo. Para Vercellis (2009), um sistema de apoio à decisão é um aplicativo interativo baseado em computador que une dados e modelos matemáticos para ajudar os tomadores de decisão de resolver problemas complexos enfrentados na gestão das empresas públicas, privadas e nas organizações.

Segundo Mylius (2004), a sigla BI é uma roupagem nova para os já conhecidos Sistemas de Apoio à Decisão (SAD) baseados em fatos, ou *Executive Information System* (EIS), que existem há muito tempo, mas carregavam no passado um impeditivo para sua difusão em larga escala: o custo das ferramentas. Para Sell (2006), as soluções de BI surgiram como uma evolução dos sistemas de informação para o desenvolvimento das organizações.

Para Favaretto (2007), informações mais abrangentes, elaboradas e integradas (compostas com dados de outros sistemas de informação), não podem ser geradas sem a realização de consolidações manuais ou através do desenvolvimento de novos sistemas de informações por pessoal especializado. Desta forma, no processo de coleta, armazenamento e tratamento dos dados, a figura de um profissional de TI bem qualificado se faz imprescindível para a correta manipulação dos diversos dados que serão coletados, processados e armazenados.

O BI não pode ser representado por um software, ou até mesmo por uma ferramenta de manipulação de dados, mas deve ser considerado um conceito aplicável em uma organização. Observa-se que o interesse pelo BI cresce na medida em que seu emprego possibilita às organizações realizar uma série de análises e projeções, de forma a agilizar os processos relacionados às tomadas de decisão por seus gestores.

Turban *et al.* (2009) afirmam que o BI deve servir como uma forma de mudar a maneira como a empresa conduz suas operações de negócio. Isso é feito através da melhoria dos processos de negócios e da transformação das tomadas de decisão em processos mais orientados aos dados.

De acordo com Dutra (2005), os sistemas de BI tem como principais características:

- a) extrair e integrar dados de múltiplas fontes;
- b) fazer uso da experiência e conhecimento adquirido por seus usuários;
- c) analisar dados dentro de uma cadeia de processos de negócios;
- d) trabalhar com múltiplas hipóteses e simulações;
- e) extrair padrões de comportamento e classificá-los em categorias.

Além das características acima podemos acrescentar que os sistemas de BI também tem como característica a transformação dos registros obtidos em informação útil para o conhecimento organizacional/empresarial.

A implantação de um BI pode trazer inúmeros benefícios, entre eles destacam-se, de acordo com Filho (2010), alguns itens discriminados a seguir:

- a) possibilidade de capturar, criar, organizar e usar todos os ativos de informação da empresa;
- b) antecipação de mudanças de mercado;
- c) antecipação de ações concorrentes;
- d) conhecimento sobre o negócio;
- e) aprendizado pelos sucessos e falhas internos e dos concorrentes;
- f) visão clara sobre novos negócios;
- g) auxílio na implantação de novas ferramentas gerenciais e
- h) conhecimento sobre novas tecnologias, produtos ou serviços, que possam afetar diretamente e indiretamente o negócio.

Além dos benefícios apresentado na visão de Filho (2010), Vercellis (2009), afirma que os sistemas de BI podem ser usados em uma série de aplicações, como marketing relacional, *Balanced Scorecard*, mineração de dados, além de outras nove

constantes na Figura 1, pois nos traz uma visão geral das principais áreas que podem ser trabalhadas com um sistema de *Business Intelligence*, algumas destas áreas tem um caráter metodológico e pode ser usadas em diferentes domínios de aplicação, enquanto outras só podem ser aplicadas a tarefas específicas.

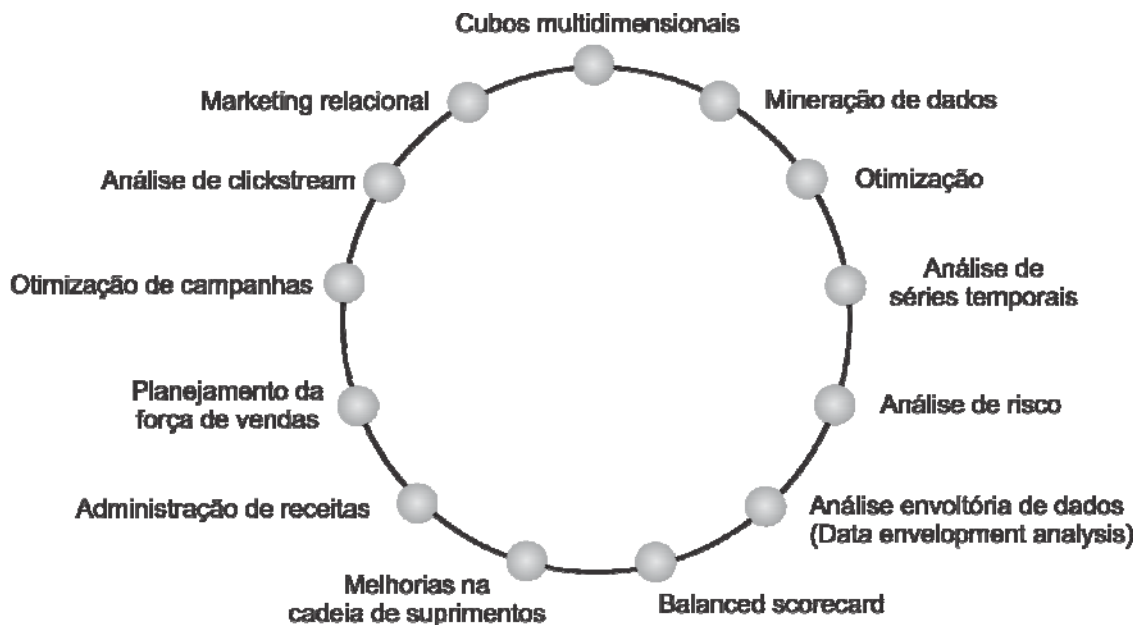


Figura 1: Potifólio de metodologias disponíveis em um sistema de *Business Intelligence*  
 Fonte: Adaptado de Vercellis (2009)

Faz-se necessário esclarecer que a escolha da solução de BI não deve basear-se em tecnologia, e sim num processo que contemple o levantamento detalhado das necessidades dos usuários e provas de conceitos (projetos-piloto) (MYLIUS, 2004).

### 2.2.1. Principais componentes da arquitetura de um BI

Os principais componentes da arquitetura tecnológica de BI são o ETL (*Extraction Transforming and Loading*), *DATA WAREHOUSE (DW)*, *DATA MART*, *OLAP (On-line Analytical Processing)* e *DATA MINING*, os quais serão detalhados a seguir.

Estes cinco componentes são os responsáveis pelo desempenho apropriado da tecnologia de BI em toda a sua abrangência, e sua implantação necessita de profissionais com amplo conhecimento nas regras de negócio, trabalhando ou dando apoio ao projeto. Os componentes de um BI podem ser utilizados de forma

independente um do outro ou na sua totalidade, dependendo dos objetivos estratégicos de cada organização.

Estes componentes da arquitetura do BI são representadas na Figura 2 proposta por Petrini (2006).

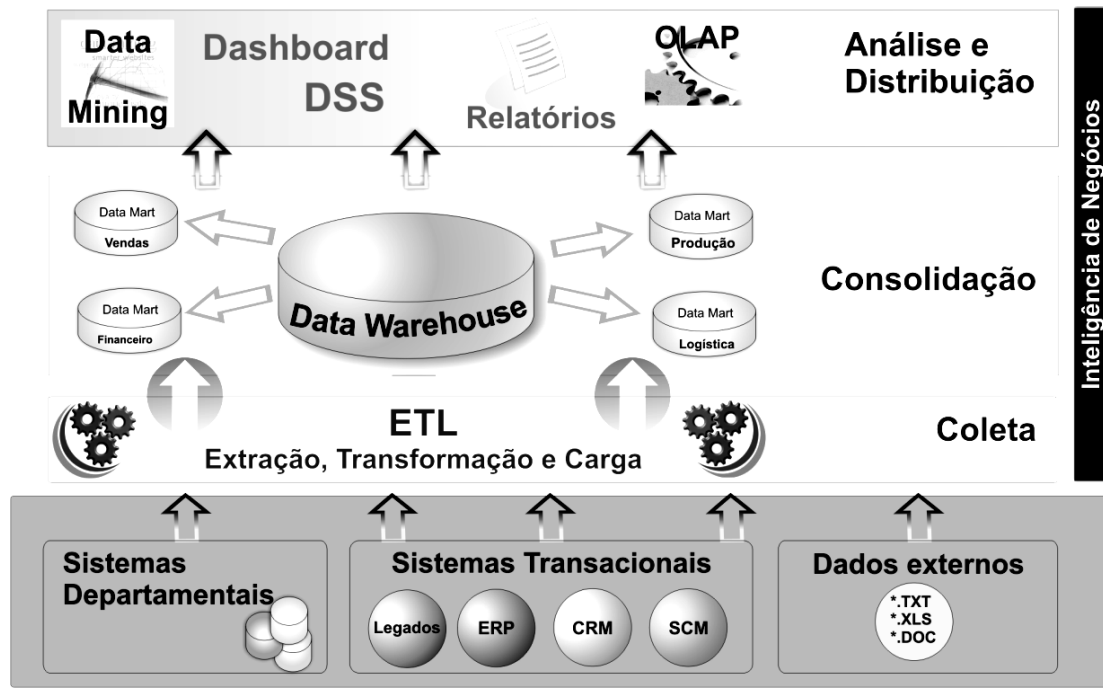


Figura 2: Proposta da arquitetura de Inteligência de Negócio  
Fonte: Adaptado de PETRINI (2006).

Para Petrini (2006), a Figura 2 propõe uma arquitetura de um BI como uma solução completa para utilização desta tecnologia, e representa desde a etapa de coleta dos dados, em que os dados são coletados nas diversas fontes de dados existentes na organização (bancos de dados, planilhas eletrônicas e documentos digitalizados), passa para o processo de ETL para serem armazenados em um *Data Warehouse* ou para os *Data Marts* de forma estruturada e padronizada.

Após a realização destas etapas, o resultado final é a formação de um grande banco de dados consolidando todos os diversos dados oriundos de todas as fontes de dados existentes na organização para, então, disponibilizá-los em forma de novas consultas que poderão alimentar outros sistemas, relatórios gerenciais, ou até mesmo fazer parte de um processo de mineração de dados com o propósito de subsidiar outros processos para a tomada de decisão.

De acordo com Turban *et al.* (2009), o processo do BI baseia-se na transformação de dados em informações, depois em decisões e finalmente em ações.

Segundo Versellis (2009) em um sistema de BI cada análise de inteligência de negócios segue seu próprio rumo de acordo com o domínio da aplicação, a atitude pessoal dos tomadores de decisão e as metodologias analíticas envolvidas no processo. No entanto, é possível identificar um caminho cíclico ideal que caracteriza a evolução de uma análise em uma estrutura de BI, como mostrado na Figura 3.

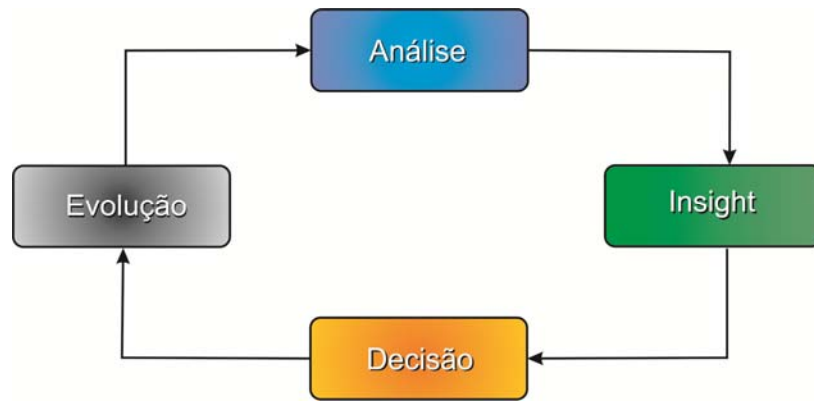


Figura 3: Ciclo da análise de BI  
Fonte: Adaptado de Versellis (2009).

Na Figura 3 os ciclos descritos por Versellis (2009) são:

**Análise:** os tomadores de decisão devem, então, criar uma representação mental do fenômeno a ser analisado, identificando os fatores críticos que são percebidos como mais relevantes. Assim, a primeira fase do ciclo de inteligência de negócios leva os decisores a fazerem várias perguntas e obterem respostas rápidas de uma forma interativa.

**Insight:** esta segunda fase permite aos decisores uma compreensão mais aprofundada do problema na mão. As informações obtidas durante a fase de análise são então transformadas em conhecimento.

**Decisão:** durante a terceira fase, o conhecimento obtido como resultado da fase de introspecção é convertido em decisões e, posteriormente, em ações.

**Avaliação:** finalmente, a quarta fase do ciclo de inteligência de negócios envolve a medição de desempenho e avaliação. Métricas devem ser concebidas para que não sejam exclusivamente limitadas as decisões aos aspectos financeiros, mas também ter em conta os principais indicadores de desempenho definidos para diferentes departamentos da empresa ou organização.

### 2.2.1.1. ETL (*Extraction Transforming and Loading*):

Segundo Turban *et al.* (2009), o processo de ETL (*Extraction Transforming and Loading* – Extração, Transformação e Carga) consiste na extração (leitura de dados dos diversos bancos de dados), transformação (conversão dos dados extraídos de sua forma anterior na forma desejada) e carga (colocação dos dados no DW). A Figura 4 representa um esboço do processo de ETL.

As ferramentas de extração, transformação e carga de dados são responsáveis por capturar os dados oriundos de diversas fontes de dados como planilhas eletrônicas, documentos em texto e dos diversos bancos de dados transacionais existentes na organização e armazená-los em um *Data Warehouse*. Durante este processo, os dados são limpos e organizados de forma a garantir a qualidade e consistência dos mesmos (GRILO JR. *et al.*, 2009).

Segundo Mylius (2003), na camada ETL os dados não integrados dos aplicativos de gestão são convertidos para uma estrutura integrada, onde poderão sofrer severa transformação.

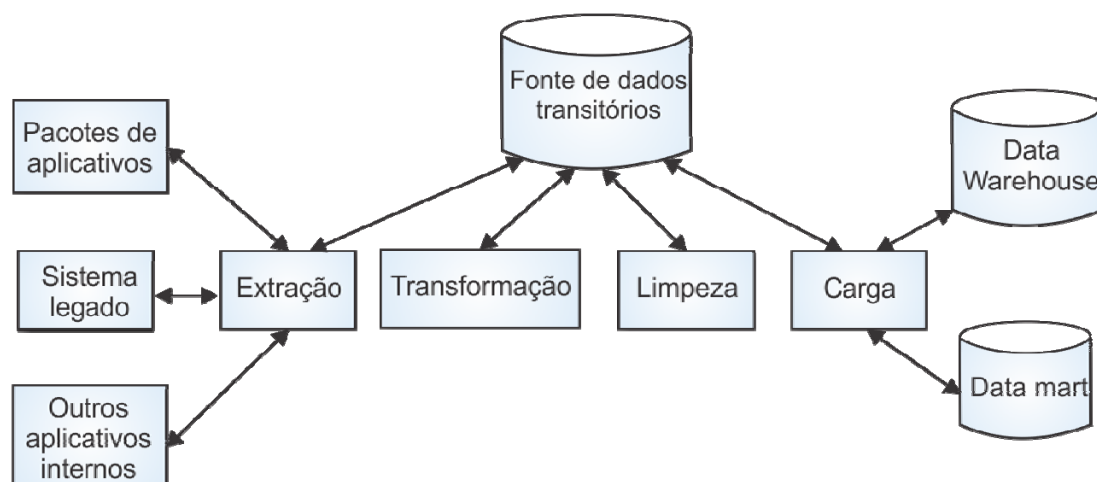


Figura 4: O processo de ETL

Fonte: Adaptado de Turban *et al.* (2009, p. 72)

Algumas das ferramentas de ETL mais utilizadas estão discriminadas no Quadro 1, embora a finalidade destas ferramentas sejam a extração e carga dos dados, existem diferenciais que devem ser ponderados, como por exemplo: o valor do investimento na sua aquisição, treinamentos, facilidade de manuseio e, principalmente, quanto a organização esta disposta a investir em determinado projeto. Apesar de estas ferramentas fazerem grande parte do trabalho, em muitos casos tem-se a necessidade de criar rotinas ou procedimentos computacionais personalizados para a carga dos dados,

objetivando atender a determinadas necessidades surgidas no processo de extração dos dados. Desta forma, recorre-se ao uso de scripts ou comandos na linguagem SQL (*Structured Query Language*, ou Linguagem de Consulta Estruturada).

Ferramenta	Tipo de distribuição	Fabricante
CloverETL	Open Source	OpenSys
Data Stage		IBM
DTS	Open Source	Microsoft
ETI		ETI Corporation
Informatica Power Conect		Informatica
Oracle Warehouse Builder	Open Source	Oracle
Pentaho Data Integration		Pentaho Corporation
Sagent Data Flow		Pitney Bowes Business Insight
Sunopsis	Open Source	Oracle
Talend		

Quadro 1: Ferramentas de ETL

Fonte: GOLDSCHMIDT; PASSOS (2005).

A escolha da ferramenta ideal para a realização da extração e carga dos dados não é uma tarefa trivial, pois existem vários aspectos a considerar, como o objetivo da tarefa a ser realizada, o tamanho da base de dados, a precisão exigida, entre outros. Muitas vezes são necessários testes empíricos com diferentes tipos de algoritmos objetivando descobrir qual deles apresentará o melhor resultado em termos de qualidade e precisão nas atividades realizadas, a execução desta tarefa poderá consumir boa parte do tempo trabalhado no projeto, além de ser considerado de suma importância para o sucesso do mesmo.

#### 2.2.1.2. Data Warehouse (DW):

Segundo Barbieri (2001), um DW pode ser definido como um banco de dados, destinado a sistemas de apoio à decisão, e cujos dados foram armazenados em estruturas lógicas dimensionais. O DW surgiu da necessidade de integrar dados provenientes de diversas origens e também na necessidade de gerenciar um grande volume de dados (FAVARETO, 2007). Para Sampaio, Rocha & Schiel (2000), o DW é um banco de dados multidimensional, em que os fatos mensuráveis de um negócio são registrados por critério de agregação ou dimensão.

Turban *et al.* (2009) conceitua DW como um conjunto de dados produzidos para oferecer suporte à tomada de decisões. O DW é um banco de dados voltado para suporte à decisão de usuários finais, derivado de diversos outros bancos de dados operacionais e tem a tarefa de transformar, consolidar e racionalizar as informações dispersas por



diversos bancos de dados e plataformas, permitindo que sejam feitas análises estratégicas eficazes, em informações antes inacessíveis ou subaproveitadas (SERRA, 2002).

Segundo Lopes e Oliveira (2007) a integração dos diversos sistemas e bases de dados deve ser feita através de uma seleção cuidadosa e otimizada dos dados, já que a prioridade na utilização do ambiente do DW é o processamento de consultas e não o processamento de transações. A Figura 5 representa a estrutura de um DW.

Segundo Serra (2002) um *Data Warehouse* é um conjunto composto por tecnologias e sistemas que não devem apenas permitir o acesso a dados, mas que também permita análises de dados significativas, de tal maneira a transformar dados brutos em informação útil para os processos estratégicos da empresa.



Figura 5: Estrutura interna de um DW  
Fonte: Lopes; Oliveira (2007, p. 2)

O *data warehouse* de um sistema de BI recebe cargas de dados de uma ou várias origens, que geralmente constituem os sistemas de gestão da empresa – contabilidade, faturamento e vendas – e sistemas de produção como PCP (Planejamento e Controle de Produção) e automação industrial (MYLIUS, 2003).

#### 2.2.1.3. Data Mart

Conforme Barbieri (2001), *data mart* significa depósito de dados que atende a áreas específicas da empresa e objetiva auxiliar o processo decisório gerencial. Por ser uma estrutura setorial do *data warehouse*, pode-se ter um *data mart* para cada

departamento de uma empresa, por exemplo: *data mart* comercial, *data mart* financeiro, *data mart* de recursos humanos e etc. (CALDAS, SCANDELARI & PILATTI, 2006).

Serra (2002) conceitua *data Mart* como um pequeno *data warehouse* que fornece suporte à decisão de um pequeno grupo de pessoas. Apesar do conceito de *Data mart* ser um subconjunto de um *data warehouse*, muitas empresas decidem implementar seus sistemas a partir da adoção direta de *data marts*, por serem mais rápidos e fáceis de implementar (PETRINI, 2006).

#### 2.2.1.4. OLAP (*On-line Analytical Processing*):

O Processamento Analítico *On-Line* (OLAP) é uma forma de acessar as informações estruturadas multidimensionalmente e carregá-las pelo ETL em um *Data Warehouse* (LEITE, 2007). As ferramentas de OLAP possibilitam tarefas como análises e cruzamentos de informações, emissão de relatórios das informações analisadas, gravação dos relatórios no servidor de aplicações OLAP e exportação das informações para planilhas eletrônicas e arquivos (AFFELDT, VANTI E RAUTER, 2005).

Segundo Giudici (2003), OLAP é geralmente um instrumento gráfico utilizado para destacar as relações entre as variáveis disponíveis, seguindo a lógica de um relatório bidimensional, sendo uma ferramenta importante para a inteligência de negócios. As ferramentas de consulta e relatórios descrevem o que um banco de dados contém (no sentido mais amplo), sendo usado para explicar por que existem certas relações, ou seja, o usuário faz suas próprias hipóteses sobre as possíveis relações entre as variáveis e ele observa para a confirmação da sua opinião ao analisar os dados. OLAP pode ser utilizada nas fases de pré-processamento da mineração de dados. Isso nos faz entender mais facilmente os dados, porque se torna possível focar os dados mais importantes, identificando casos especiais ou à procura de relações principais.

Segundo Sampaio, Rocha e Schiel (2000), a expressão “aplicações OLAP” vem sendo utilizada para caracterizar uma nova gama de aplicações voltadas para a atividade de análise de negócios. Ainda segundo os autores, um usuário executivo de média ou alta gerência que utiliza uma aplicação OLAP, deve receber informações sumarizadas e históricas do seu negócio, que o capacitem a tomar decisões no seu nível de gerência.

Os dados usados por essas aplicações são chamados de dados analíticos; são dados operacionais otimizados para a análise, e não para transações (CALDAS, SCANDELARI E PILATTI, 2006).

A extração de relatórios e dados sumarizados também poderá ser realizada sem o uso de ferramentas de OLAP em banco de dados de uso rotineiro dentro da organização, que são chamados de base de dados de produção, porém será necessário um esforço muito maior da equipe de desenvolvimento para preparar estas consultas e integrá-las aos aplicativos usados pelos tomadores de decisão. Por este motivo recomenda-se o uso de um *Data Warehouse* para que os dados possam ser acessados de forma consolidada e com o tempo de resposta otimizado.

Para Giudici (2003), ao contrário do OLAP, a mineração de dados reúne todas as variáveis disponíveis e combina-as de diferentes maneiras. Também significa que podemos ir além da representação visual dos resumos em aplicações OLAP, criando modelos úteis para o mundo dos negócios.

### 2.3. Descoberta de Conhecimentos em Base de Dados

Segundo constantes avanços na área de TI tem viabilizando o armazenamento de grandes e múltiplas bases de dados. A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais adequadas, tornando-se necessário o desenvolvimento de ferramentas computacionais que auxiliem o homem de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados a fim de desenvolver e selecionar estratégias de ação em cada contexto de aplicação.

Diante desta necessidade surge uma nova área denominada Descoberta de Conhecimentos em Bases de Dados (*Knowledge Discovery in Databases – KDD*).

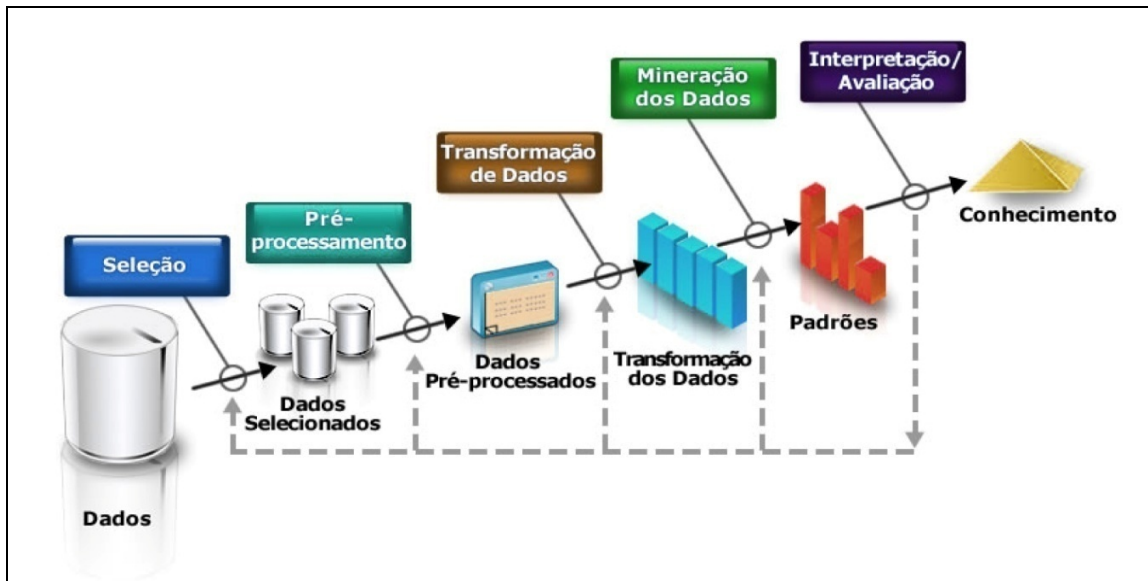


Figura 6: Uma visão geral das etapas que compõem o processo de KDD

Fonte: Adaptado de Fayyad *et al.* (1996)

Segundo Fayyad *et al.* (1996), o processo KDD refere-se ao processo global de descoberta de conhecimentos úteis a partir de dados, o mesmo é composto por 5 etapas, como pode ser visualizado na Figura 6, a mineração de dados refere-se a quarta etapa deste processo.

Estas etapas para Fayaad (1996) são distribuídas em 3 grupos: pré-processamento (seleção de dados, pré-processamento e tratamento de dados), mineração de dados e pós-processamento (interpretação e avaliação). Para melhor entendimento sobre estas etapas, as mesmas serão descritas abaixo:

**Seleção de dados:** esta etapa do processo de seleção é bastante complexa, uma vez que os dados podem vir de uma série de fontes diferentes (*data warehouses*, planilhas, sistemas legados) e podem possuir os mais diversos formatos.

**Pré-processamento:** esta etapa visa eliminar os dados que não se adequam às informações, com base em algoritmos específicos verifica-se a existência de dados não padronizados, um problema bastante comum nesta fase é a ausência de valores para determinadas variáveis, ou seja, registro com dados incompletos seja por falhas no processo de seleção ou de revisão, então se procura atribuir um valor padrão ou descartá-lo.

**Transformação dos dados:** nesta etapa os dados já existentes poderão ser transformados em novos dados mediante a consolidação ou desdobramento dos mesmos de acordo com o alvo que se deseja alcançar.

**Mineração de dados:** nesta etapa poderá ser utilizado um conjunto de ferramentas que, através do uso de algoritmos de aprendizado ou baseada em redes neurais e estatística, permite buscar em um determinado conjunto de dados por padrões/relações e regularidades. Para o autor, *data mining* é a etapa-chave no processo de descoberta de conhecimento em bases de dados.

**Interpretação e/ou Avaliação:** nesta etapa verifica-se a qualidade do conhecimento descoberto, procurando identificar se o mesmo auxilia a resolução do problema originalmente proposto para a realização do processo de KDD.

Os pesquisadores desta área do conhecimento buscam produzir técnicas, metodologias e instrumentos que apoiem a descoberta de conhecimento inédito e acrescentem informações a um domínio de aplicação específica. A extração de conhecimento de bases de dados é uma área multidisciplinar que visa incorporar e relacionar técnicas adotadas em diversas áreas do conhecimento.

De acordo com Fayyad *et al.* (1996), o processo de descoberta de conhecimento em bases de dados evoluiu, e continua evoluindo, através da interseção de vários campos de pesquisa: aprendizado de máquina, reconhecimento de padrões, bancos de dados, estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados, e computação de alto-desempenho. Desse modo, as técnicas utilizadas em KDD não devem ser vistas como substitutas de outros métodos de análise, mas sim como aplicações que buscam melhorar os resultados das explorações feitas com ferramentas convencionais.

Para Tan, Steinbach e Kumar (2009), uma quantidade de outras áreas também desempenham papéis chave, como a computação de alto desempenho (paralela) e a computação distribuída, conforme visualizado na Figura 7. Segundo os mesmos autores, as técnicas de computação de alto desempenho são muitas vezes importantes para abordar o tamanho volumoso de alguns conjuntos de dados, e as técnicas distribuídas também podem auxiliar a abordar a questão do tamanho e são essenciais quando os dados não podem ser juntados em um único local.

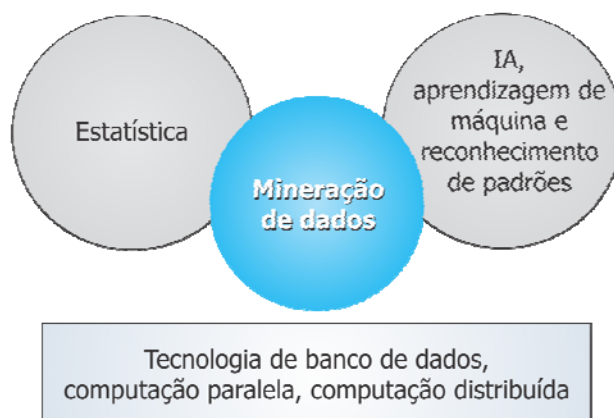


Figura 7: Mineração de dados como uma confluência de muitas disciplinas

Fonte: Adaptado de Tan, Steinbach e Kumar (2009, p. 8).

De acordo com as definições de Shapiro e Matheus (1992), a Extração de Conhecimento de Bases de Dados é um processo não convencional de identificar padrões desconhecidos, potencialmente úteis e compreensíveis, embutidos em bases de dados. Extrair conhecimento dessas bases de dados envolve, entre outras técnicas, a utilização de algoritmos de aprendizado de máquina com a capacidade de generalizar os fatos encontrados em um grande repositório de dados na forma de regras de alto nível, passíveis de compreensão pelo ser humano.

Essas técnicas incentivam pesquisadores a participarem do processo de KDD oferecendo contribuições científicas substancialmente maiores. Ao se tentar realizar descobrimento de conhecimento em aplicações do mundo real, observa-se que as técnicas podem ser extremamente complexas e que as tarefas de descoberta de novas informações representam apenas uma porção de um extenso processo.

O processo de extração de conhecimento pode ser aplicado de forma iterativa e iterativa, envolvendo diversas fases de preparação que englobam a definição e compreensão do domínio de aplicação, seleção e amostragem dos dados, limpeza, enriquecimento e preparação dos mesmos, extração de padrões, até a etapa em que os resultados são interpretados, avaliados e consolidados pelos especialistas do domínio. O processo de descoberta de conhecimento em bases de dados é considerado iterativo, pois pode conter repetições entre qualquer dos passos, buscando sempre aprimorar os resultados a cada iteração.

As etapas do processo KDD são geralmente realizadas de forma sequencial, ou seja, é necessário obter compreensão a respeito do domínio de aplicação, selecionar e

transformar os dados para, somente depois, encontrar padrões. Mesmo observando estas recomendações, podem ocorrer iterações entre as tarefas. Por se tratar de um processo interativo, o grupo de pessoas envolvidas na sua realização deve possuir um canal de comunicação que maximize a troca de informações. Estes usuários devem possuir diferentes especialidades, necessidades e expectativas, desempenhando o seu papel dentro do processo.

Os dados que fornecerão a base para a realização do processo de KDD podem ser provenientes de diversas fontes de dados, internas ou externas. Os dados de origem interna normalmente são fornecidos por repositórios de dados que já estão incorporados ao sistema de aplicação do domínio em questão e, normalmente são constituídas por *Data Warehouse*, bases de dados operacionais e dados históricos. Os dados de origem externa são compostos por todos os outros tipos de localidades que habitualmente não estão incorporados ao sistema de aplicação, como por exemplo, documentos e informações do especialista do domínio.

Um problema que deve ser resolvido envolve o tamanho do conjunto de dados, pois a grande quantidade de dados reunida pode inviabilizar a realização do processo de KDD, neste caso recomenda-se uma filtragem na amostra com os dados que são relevantes para o estudo.

Assim, devem ser utilizadas técnicas de amostragem de dados que reduzam o tamanho do conjunto de dados, obtendo um subconjunto relevante e representativo para toda a base de dados. Esse processo de amostragem é crítico dentro do processo de KDD, pois, a não utilização de técnicas estatísticas adequadas pode gerar um subconjunto de dados de características não representativas, comprometendo as análises, que não representariam a verdadeira situação dos fatos registrados.

## CAPÍTULO 3 – DATA MINING

Segundo Turban *et al.* (2009) *data mining* é uma classe de análise de informações, baseada em bancos de dados, a qual procura padrões ocultos em uma coleção de dados que podem ser usados para prever comportamentos futuros. As ferramentas de *data mining* são usadas para substituir ou aprimorar a inteligência humana devido a sua capacidade de verificar enormes armazéns de dados.

Carvalho (2005) define o processo de *data mining* como sendo a forma de descobrir conhecimento novo escondido em grandes massas de dados, e Witten e Frank (2000), de forma sucinta, afirmam que a mineração de dados pode ser definida como a extração de informações implícitas, previamente desconhecidas e potencialmente úteis de grandes bases de dados. Para Giudici (2003), *data mining* é o processo de seleção, exploração e modelagem de grandes quantidades de dados para descobrir regularidades ou relações que são desconhecidas com o objetivo de obter resultados claros e úteis para o proprietário do banco de dados.

Han e Kamber (2006) conceituam a mineração de dados como a tarefa de descoberta de padrões interessantes extraídas de um grande volume de dados armazenados em base de dados, *data warehouse* ou outro repositório.

O processo de mineração de dados faz parte da categoria das ferramentas de análise, em vez de se fazer uma pergunta, entrega-se grandes quantidades de dados e pergunta-se se existe algo de interessante (que pode ser uma tendência ou um agrupamento). O *data mining* pode extrair conhecimento que está escondido ou informações de prognóstico do *data warehouse* ou de uma outra base de dados sem a necessidade de consultas específicas ou requisições. Esse processo de mineração usa técnicas avançadas como redes neurais, heurísticas, descoberta por regra e detecção de desvio.

Segundo Giudici (2003), ao contrário de relatórios e consultas cujos relacionamentos já se conhecem, o trabalho de *data mining* é descobrir o que não se sabe que existe no banco de dados. São exemplos de aplicações de *data mining*: identificar padrões de compra dos clientes, identificar correlações escondidas entre diferentes indicadores financeiros, e identificar superfaturamento em grandes obras públicas. De acordo com Dutra (2005), o propósito de *data mining* é extrair



conhecimento de base de dados cuja dimensão, complexidade e volume de dados seriam proibitivos a um observador humano.

Para Gonçalves (2005), a mineração de dados é realizada por meio de estratégias automatizadas para a análise de grandes bases de dados, procurando extrair das mesmas informações que estejam implícitas, que sejam previamente desconhecidas e potencialmente úteis. Em geral, o conhecimento descoberto através de processos de mineração de dados é expresso na forma de regras e padrões. A utilização de técnicas de data mining se propõe a encontrar padrões dentro desse gigantesco volume de dados que compõe o banco de dados das organizações na busca de revelar surpresas e detalhes sobre o negócio que não eram conhecidos ou evidenciados.

Segundo Barcelos Tronto *et al.* (2003), um dos grandes problemas dos especialistas em análise de informação é a transformação de dados em informação. Como fazer isso de uma forma automatizada, e no menor tempo possível? Uma das respostas para isso é a combinação de estatística convencional e técnicas de inteligência artificial, que resulta no *data mining*. Segundo os autores, em qualquer projeto de mineração de dados é obrigatório a existência de um profissional com o conhecimento do negócio, pois ele possui domínio total do assunto, evitando que o resultado da modelagem não tenha utilidade para apoiar uma decisão.

Mineração de dados é um termo genérico que identifica a etapa do processo de KDD de transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação como regras e relações entre dados (REZENDE, 2005). Para Tan, Steinbach e Kumar (2009), a mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. O produto gerado pelas ferramentas de *data mining* está relacionado com o tratamento especial da informação, ao contrário das estruturas de dados (BARBIERI, 2001).

O'Brien (2004) enfatiza que o software de DM faz uso de algoritmos avançados de reconhecimento de padrões, assim como uma variedade de técnicas matemáticas e estatísticas para selecionar “montanhas” de dados e extrair as informações estratégicas anteriormente desconhecidas. Para Romão (2002), o *data mining* é especialmente útil em casos onde não se conhece a pergunta, mas, mesmo assim, existe a necessidade de respostas.

Os sistemas de mineração são baseados principalmente em sistemas de arquivos ‘*stand-alone*’, estruturas de dados especializadas, e estratégias locais de gerência de ‘*buffers*’. No máximo, os dados para mineração são importados ou extraídos de um Sistema Gerenciador de Banco de Dados (SGBD) e armazenados localmente (‘*cache-mining*’). Dessa forma, elimina-se a necessidade de recuperar dados várias vezes do SGBD, melhorando o desempenho da aplicação (VASCONCELOS, 2002).

De acordo com SPSS (2010), a maioria dos analistas separa os *softwares* de *data mining* em dois grupos:

- ferramentas de mineração de dados e;
- aplicativos de mineração de dados.

As ferramentas de mineração de dados fornecem um número de técnicas que podem ser aplicadas em qualquer problema de negócios. Já os aplicativos de mineração de dados, por outro lado, encaixam técnicas dentro de uma aplicação customizada endereçada a um problema de negócio específico. Tanto as ferramentas como as aplicações de mineração de dados são valiosas e cada vez mais, as empresas estão usando-as em conjunto dentro de um ambiente integrado para análise preditiva.

Técnicas de *data mining* podem ser aplicadas para gerar diversos ganhos como traçar o perfil de um cliente que tem um bom relacionamento no mercado, verificar possíveis fraudes, análise de dados de vendas e ainda ganhos sociais como identificar uma fatia da população que precise de maior assistência em um determinado ramo social (MACHADO; JAMIL, 2008). Ainda segundo os mesmos autores, com isso, os dados se tornam um patrimônio importantíssimo para a empresa e explorá-los de forma sábia pode ser um fator decisivo, para o crescimento da empresa, sustentação de uma posição no mercado ou até mesmo decisões importantes de investimentos.

Segundo Murayama (2002), as informações descobertas precisam ser verdadeiras e relevantes para o contexto da exploração daquilo que se está procurando. A regra final é que a descoberta possa ser trabalhada, passível de uma ação estratégica e que traga benefícios para a organização, como por exemplo:

- otimização das campanhas de marketing;
- derivação de uma lista negra com fatores para combater fraudes e evitar riscos;
- promoção de produtos/serviços;

- aumento de postos de trabalhos;
- abertura de novas filiais e etc.

Sistemas de BI se destinam a atender usuários da área de negócios que esperam extrair informações a partir dos dados coletados. Muitas vezes milhares de transações são compiladas para se calcular um único indicador (ou métrica) (CRIVELINI, 2006).

Segundo Mylius (2003), o volume de investimentos alocado nos projetos de *Business Intelligence* está diretamente ligado ao retorno obtido através do “valor das informações”. No nível mais baixo está o acesso a relatórios antes não disponíveis devido aos dados não estarem integrado ou na mesma base de dados e o ponto máximo do valor das informações seria o *data mining*, onde os prognósticos são viabilizados e revelados através da exploração da base de dados e utilização de algoritmos matemáticos, conforme ilustrado na Figura 8.

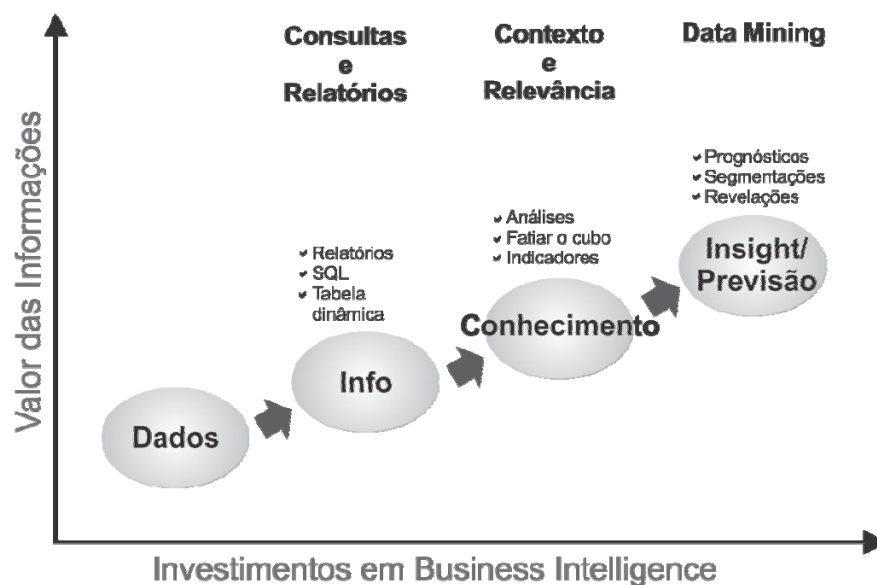


Figura 8: Valor das Informações x Investimentos em BI  
 Fonte: Adaptado de MYLIUS (2003, p. 51)

Na utilização das técnicas de *Data Mining* busca-se mais que a interpretação dos dados existentes, pretende-se realizar inferências, buscando correlações nas informações não explicitadas no volume de dados de um *Data Warehouse* ou *Data Mart* (PINTO, SANTOS & OLIVEIRA, 2008). As técnicas de mineração de dados são organizadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis e recentes que poderiam, de outra forma, permanecer ignorados (TAN, STEINBACH E KUMAR, 2009).

### 3.1. Principais ferramentas para a mineração de dados

Nesta sessão serão abordadas algumas das principais ferramentas utilizadas para a mineração de dados.

No Quadro 2 são relacionadas às principais ferramentas (*softwares*) para mineração de dados, com suas características e principais tarefas realizadas no processo de descoberta de conhecimento, além de informar alguns domínios onde as mesmas estão sendo utilizadas.

Ferramenta	Características	Tarefa de KDD	Domínios utilizados	Fabricante
<b>SPSS/Clementine</b>	Permite o desenvolvimento rápido de modelos preditivos para as operações da corporação, melhorando a tomada de decisão.	Classificação, Regras de Associação, Clusterização, Sequência e Detecção de Desvios.	Associação Comercial de São Paulo, Credicard, CTBC Telecom, DirecTV, Globo.com, entre outras.	SPSS Inc. www.spss.com
<b>PolyAnalyst</b>		Classificação, Regressão, Regras de Associação, Clusterização, Sumarização e Detecção de Desvios.		Megaputer Intelligence www.megaputer.com
<b>Intelligent Miner</b>		Classificação, Regras de Associação, Clusterização e Sumarização.		IBM Corp. www.ibm.com
<b>WizRule</b>		Sumarização, Classificação e Detecção de Erros.		WizSoft Inc. www.wizsoft.com
<b>SAS Enterprise Miner</b>		Classificação, Regras de Associação, Clusterização, Agrupamento, .	Bank of America, Telefonica O2, Korea Customs Service, Australian Bureau of Statistics, entre outros.	SAS Corp. www.sas.com
<b>Tamanduá</b>		Associação, Agrupamento e Classificação	Auditoria-Geral do Estado de Minas Gerais; Sec.de Log.e TI do Min. do Planej., Orçamento e Gestão; Min.da Justiça; CGU.	Depart. Ciência da Comput. da UFMG <a href="http://tamandua.speed.dcc.ufmg.br">http://tamandua.speed.dcc.ufmg.br</a>
<b>Oracle Data Mining</b>		Classificação, Regressão, Associação, Clusterização e Mineração de Textos.		Oracle www.oracle.com
<b>WEKA</b>	API e ambiente de testes com algoritmos de mineração de dados e aprendizado por computador.	Classificação, Regressão e Regras de Associação, Clusterização.		University of Waikato www.cs.waikato.ac.nz
<b>RapidMiner (antigo YALE)</b>	Derivado do WEKA, é um pacote mais completo de mineração de dados.	Classificação, Regressão e Regras de Associação, Clusterização.	Ford, Honda, Nokia, Miele, Philips, IBM, HP, Cisco, Bank of America, entre outras.	Rapid-I Rapid-i.com

Quadro 2: Comparativo entre ferramentas utilizadas para a mineração de dados  
Fonte: Baseado em GOLDSCHMIDT; PASSOS (2005); TAMANDUÁ (2010).

### 3.2. Algoritmos para mineração de dados

No Quadro 3 está discriminando alguns métodos de mineração de dados que podem ser aplicados para cada as tarefas de KDD, não foi abordado todos os algoritmos pertencentes a cada uma das tarefas, mas os principais métodos utilizados por elas.

Tarefa de KDD	Métodos de mineração de dados
Descoberta de associações	Basic, Apriori, DHP, Partition, DIC, ASCX-2P
Descoberta de associações generalizadas	Basic, Apriori, DHP, Partition, DIC, ASCX-2P
Descoberta de sequências	GSP, MSDD, SPADE
Descoberta de sequências generalizadas	GSP, MSDD, SPADE
Classificação	Redes Neurais (Ex.: Back-Propagation, RBF), C4.5, Rough, Sets, Algoritmos genéricos (Ex.: Rule Evolver), CART, K-NN, Classificadores Bayesianos.
Regressão	Redes Neurais (Ex.: Back-Propagation), Lógica Nebulosa
Sumarização	C4.5, Algoritmos genéricos (Ex.: Rule Evolver)
Clusterização	K-Means, K-Modes, K-Prototypes, Fuzzy K-Means, Algoritmos genéricos, Redes Neurais (Ex.: Kohonen)
Previsão de Séries Temporais	Redes Neurais (Ex.: Back-Propagation), Lógica Nebulosa (Ex.: Wang-Mendel)

Quadro 3: Métodos de mineração de dados que podem ser aplicados em cada tarefa de KDD  
Fonte: GOLDSCHMIDT; PASSOS (2005).

Com a diversificação das atividades relacionadas à mineração de dados, podem ser extraídos diferentes tipos de conhecimento. Faz-se necessário decidir no início do processo de mineração de dados qual o tipo de tarefa que se deseja executar e qual o tipo de conhecimento que o algoritmo de mineração de dados deve extrair ou quais padrões escondidos nos dados poderão ser descobertos. Segundo Fayyad *et al.* (1996), não ha um método de mineração de dados universal e a escolha de um algoritmo particular é de certa forma uma arte.

De acordo com Tan, Steinbach e Kumar (2009), as tarefas de mineração de dados são geralmente divididas em duas categorias principais: tarefas de previsão e tarefas descritivas. As tarefas de previsão objetivam prever o valor de um determinado atributo baseados em valores de outros atributos, já as tarefas descritivas objetivam derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumam os relacionamentos adjacentes dos dados. Para os autores existem quatro tarefas centrais da mineração de dados: modelagem de previsão, análise de associação, análise de agrupamentos e detecção de anomalias. A seguir são descritas cada uma destas tarefas:

A **modelagem de previsão** se refere à tarefa de construir um modelo para a variável alvo como uma função das variáveis explicativas. Neste processo há dois tipos de tarefas: classificação, usada para variáveis alvo discretas, e regressão, que é usada

para variáveis alvo contínuas. O objetivo e ambas as tarefas é aprender um modelo que minimize o erro entre os valores previsto e real da variável alvo. Um exemplo de uso é a identificação de clientes que responderão a uma campanha de vendas ou avaliar se um paciente possui uma determinada doença baseado nos resultados de exames médicos.

A **análise de associação** é usada para descobrir padrões que descrevam características altamente associativas dentro dos dados, os padrões descobertos são normalmente apresentados na forma de regras de implicação ou subconjuntos de dados. Um exemplo de uso inclui a descoberta de genes que possuem funcionalidade associada, ou a análise de uma cesta de compras.

A análise de **agrupamentos** ou *clustering* procura encontrar grupos de observações intimamente relacionadas, de modo que observações que pertençam ao mesmo grupo sejam mais semelhantes entre si do que com as que pertençam a outros grupos. Segundo Fayyad *et al.* (1996), o agrupamento é uma tarefa onde se procura identificar um conjunto finito de categorias ou agrupamentos para descrever os dados. Um exemplo de uso é o agrupamento de documentos, outro exemplo é juntar conjuntos de clientes relacionados ou com as mesmas afinidades.

A **detecção de anomalias** é a tarefa de descobrir grupos utilizando a similaridade dos valores de seus atributos cujas características sejam significativamente diferentes do resto dos dados. O objetivo de um algoritmo de detecção de anomalias é descobrir as anomalias verdadeiras e evitar rotular erroneamente objetos normais como anômalos (TAN, STEINBACH e KUMAR, 2009). Um exemplo de uso é a detecção de fraudes em cartões de crédito, padrões incomuns de doenças e perturbações no meio ambiente.

Com a finalidade de detalhar melhor alguns algoritmos de mineração de dados, a tarefa de análise de associação segundo Goldschmidt e Passos (2005) é uma operação de associação que consiste em encontrar um conjunto de itens que ocorram simultaneamente e de forma frequente em um banco de dados. A tarefa de descoberta de associações, segundo os mesmos autores, pode ser definida como a busca por regras de associação frequentes e válidas em um banco de dados. Desse modo, o objetivo das regras de associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento nos dados analisados.

Segundo Gonçalves (2005), os algoritmos de Regras de Associação descrevem padrões de relacionamento entre itens de uma base de dados. Uma de suas típicas aplicações é a análise de transações de compras. Este processo examina padrões de compras de consumidores para determinar produtos que costumam ser adquiridos em conjunto. Da mesma forma que podemos identificar padrões em cestas de compras, também podemos identificar padrões constantes em processos licitatórios, objeto deste estudo, já que estes dados são enviados em formato padronizado ao TCE-PB, o que facilita o seu manuseio, já que estes dados são capturados por aplicativos disponibilizados com esta finalidade.

Segundo Silveira (2003), a técnica de descoberta de regras de associação estabelece uma relação entre certos itens em um conjunto de dados. Para a autora, a descoberta de associações em itens de cestas de compras deve não apenas evidenciar as associações triviais conhecidas, como por exemplo, quem costuma comprar leite também costuma comprar pão, mas sim aquelas que não são óbvias e que podem se tornar importante fonte de informação na tomada de decisão. Uma regra de associação possui duas partes: a condição (X) e o resultado (Y) ou:  $(X_1, X_2, \dots, X_n) \Rightarrow Y$ ; onde os itens  $X_1, X_2, \dots, X_n$  preveem a ocorrência de Y, onde a probabilidade de encontrar Y por esta regra, é chamada de grau de certeza ou fator de confiança. A Figura 9 exemplifica bem esta condição.

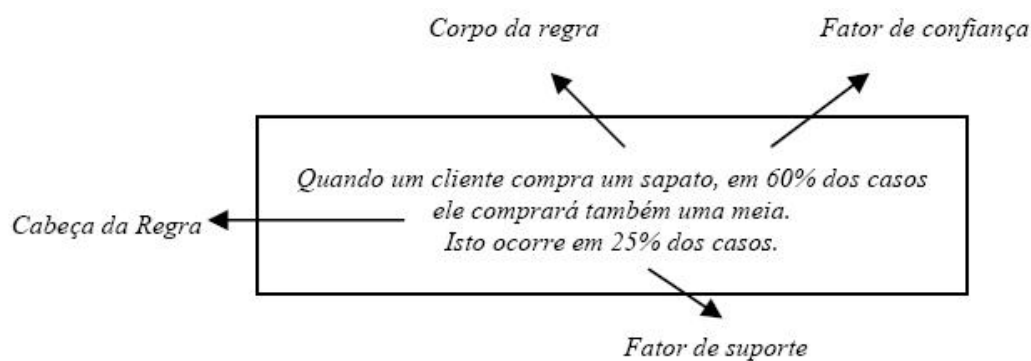


Figura 9: Exemplo de uma Regra de Associação  
Fonte: Silveira (2003).

De acordo com Tan, Steinbach e Kumar (2009), o **fator de suporte** determina a frequência na qual uma regra é aplicável a um determinado conjunto de dados, enquanto o **fator de confiança** determina a frequência na qual os itens em Y aparecem em transações que contenham X.

Segundo Agrawal, Imielinski e Srikant (1993), as regras de associação podem ser entendidas da seguinte forma: sejam  $I = \{i_1; i_2, \dots, i_m\}$  um conjunto de  $m$  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 (*itemset*), tal que  $T \subseteq I$ . Uma regra de associação é uma expressão na forma  $A \Rightarrow B$ , onde  $A \subset I$ ,  $B \subset I$ ,  $A \neq \emptyset$ ,  $B \neq \emptyset$  e  $A \cap B = \emptyset$ .  $A$  é denominado antecedente e  $B$  denominado conseqüente da regra. Tanto o antecedente, quanto o conseqüente 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*.

Ainda segundo os autores, o suporte de um conjunto de itens  $Z$ ,  $Sup(Z)$ , representa a porcentagem de transações da base de dados que contêm os itens de  $Z$ . O suporte de uma regra de associação  $A \Rightarrow B$ ,  $Sup(A \Rightarrow B)$ , é dado por  $Sup(A \cup B)$ . Já a confiança desta regra,  $Conf(A \Rightarrow B)$ , representa, dentre as transações que contêm  $A$ , a porcentagem de transações que também contêm  $B$ , ou seja,  $Conf(A \Rightarrow B) = Sup(A \cup B) \div Sup(A)$ .

De acordo com Pizzi (2006), uma regra de associação pode ser caracterizada como unidimensional, quando os itens envolvidos na mesma derivam de um único atributo, ou multidimensional, quando existem dois ou mais atributos envolvidos na regra. A autora complementa informando que, além disso, as regras de associação podem ser caracterizadas quanto aos valores de seus atributos, podendo ser booleana, quando os atributos são categóricos; quantitativa, quando os atributos são numéricos, ou nebulosa, quando os atributos envolvem conceitos nebulosos.

Segundo Gonçalves (2005), o modelo típico para mineração de regras de associação em bases de dados consiste em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo (*SupMin*) e uma confiança mínima (*ConfMin*), especificados pelo usuário. Por este motivo, o modelo costuma ser referenciado na literatura como Modelo Suporte/Confiança.

No Modelo Suporte/Confiança, para que uma regra seja considerada forte, contendo informação interessante, é necessário que ela apresente bons valores de suporte e confiança. A decisão sobre quais regras devem ser mantidas e quais deverão ser descartadas durante o processo de mineração é baseada nos valores destes dois



índices. Isto significa que o suporte e a confiança atuam como medidas de interesse no processo de mineração de regras de associação.

Para Tan, Steinbach e Kumar (2009), o suporte é uma medida importante porque uma regra que tenha baixo suporte pode acontecer simplesmente por coincidência, por esta razão, o suporte é muitas vezes usado para eliminar regras sem interesse. A confiança, por outro lado, mede a confiabilidade de inferência feita por uma regra, por exemplo: para uma determinada regra  $X \rightarrow Y$ , quanto maior a confiança, maior a probabilidade de que Y esteja presente em transações que contenha X. Ainda segundo os autores, a confiança também fornece uma estimativa da probabilidade condicional de Y dado X.

As aplicações contendo regras de associação podem ser usadas em várias áreas de aplicação, segundo Versellis (2009) ela é particularmente recomendada para:

**Análise de cesta de compras (*Maket basket analysis*):** as compras realizadas nos diversos pontos de vendas são registradas, uma lista de itens comprados é armazenada junto com o preço, horário e local da transação. Estas operações são reunidas em um conjunto maciço de dados, que pode ser explorado para executar uma análise de mineração de dados visando identificar regras recorrentes relacionadas à compra de um produto, ou grupo de produtos, para a compra de outro produto, ou grupo de produtos. As regras de associação para a análise de cesta de compras podem ser bastante úteis para os gerentes de marketing no planejamento de iniciativas de promoção ou a definição do sortimento e localização dos produtos nas prateleiras.

**Web Mining:** as análises de mineração web é particularmente útil para compreender o padrão estabelecido nos caminhos percorridos durante a navegação e da frequência com que as combinações de páginas são visitadas por um indivíduo durante uma única sessão ou sessões consecutivas. As regras de associação deste tipo podem influenciar a estrutura das ligações entre as páginas, para facilitar a navegação e recomendar caminhos de navegação específicos, ou colocar *banners* publicitários e outras mensagens promocionais.

**Compras com cartão de crédito (*Purchases with a credit card*):** as regras de associação são também utilizadas para analisar as compras feitas por detentores de cartão de crédito a fim de direcionar promoções futuras. Neste caso, cada operação consiste na compra e os pagamentos feitos por um usuário de cartão de crédito.

**Detecção da fraude (*Fraud detection*):** na identificação de fraudes em seguros, as operações consistem nos relatórios de incidentes e nos pedidos de indenização pelos danos sofridos. A existência de combinações específicas podem revelar comportamentos potencialmente fraudulentos e, portanto, justificam uma investigação aprofundada por parte da companhia de seguros.

As regras representam um paradigma clássico de representação de conhecimento popular, devido à sua estrutura simples e intuitiva, o que os torna facilmente compreensível e lógica semelhante aos regimes típicos do raciocínio humano.

Ainda segundo Versallis (2009), as normas destinadas a extração de conhecimento para uma análise da inteligência de negócios deve ser não-trivial e interpretável, para que possam ser potencialmente úteis para os trabalhadores do conhecimento e fáceis de serem traduzidas em planos de ação concretos.

### 3.3. Aplicações de mineração de dados

De acordo com Vercellis (2009) as metodologias de mineração de dados podem ser aplicadas a uma variedade de domínios, no marketing, no controle do processo de fabricação, e para o estudo de fatores de risco em diagnósticos médicos a partir da avaliação da eficácia de novas drogas e também para a detecção de fraude. Abaixo são detalhadas algumas destas áreas segundo o autor:

**Marketing relacional (*relational marketing*):** aplicações de mineração de dados na área de marketing relacional contribuíram significativamente para o aumento na popularidade dessas metodologias. Algumas aplicações relevantes dentro do marketing relacional são:

- a identificação de segmentos de clientes que são mais propensos a responder a campanhas de marketing, tais como *cross-selling* e *up-selling*;
- a identificação de segmentos de clientes alvo para campanhas de retenção;
- previsão da taxa de respostas positivas às campanhas de marketing;
- interpretação e compreensão do comportamento de compra dos clientes;
- análise dos produtos adquiridos pelos clientes em conjunto, conhecido como análise de cesta de compras.

O *up-selling* é uma estratégia de venda na qual um vendedor ou um site de comércio eletrônico sugere um adicional para o produto ou serviço que está sendo comprado, levando o cliente a comprar bens/serviços mais caros de forma a produzir uma venda mais rentável. Já no *cross-selling* é uma técnica de vendas onde são sugeridos produtos complementares.

Segundo Giudici (2003) a análise de cestas de compras diz respeito a estudar a compreensão das associações entre comportamentos de compra. Um tipo semelhante de análise pode ser aplicada a problemas em que o objetivo principal é *cross-selling* para aumentar o número de produtos que são comprados em uma determinada unidade comercial (supermercado, um banco, uma agência de viagens, ou de forma generalizada, uma empresa oferecendo mais de um produto ou serviços). Uma classe de problemas relacionados surge em campanhas promocionais: é desejável colocar em promoção o menor número possível de produtos, mas que possibilite obter o maior número possível de benefícios em outros produtos disponibilizados para venda.

**Detecção da fraude (*fraud detection*):** a detecção de fraudes é um outro campo relevante nas aplicações de mineração dados. A fraude pode afetar diferentes setores, desde o setor de telefonia até o de seguros, também pode ocorrer na utilização ilegal de cartões de crédito e cheques bancários, além de operações monetária fraudulentas.

**Avaliação de riscos (*risk evaluation*):** o objetivo da análise de risco é estimar o risco ligado com as futuras decisões, que muitas vezes assumem uma forma dicotômica. Por exemplo, um banco pode desenvolver um modelo preditivo para determinar se é conveniente conceder um empréstimo monetário ou um empréstimo à habitação, com base nas características do pretendente.

**Mineração de texto (*text mining*):** a mineração de dados pode ser aplicada a diferentes tipos de textos, que representam dados não estruturados, a fim de classificar os artigos, livros, documentos, e-mails e páginas web. Exemplos são os motores (robôs) de pesquisa na web ou a classificação automática de comunicados de imprensa para fins de armazenamento. Outras aplicações de mineração de texto incluem a geração de filtros de mensagens de *e-mail* e *newsgroups*.

**Reconhecimento de imagem (*image recognition*):** o tratamento e classificação de imagens digitais, tanto estáticas quanto dinâmicas, é um assunto muito interessante, tanto para estudo como para um grande número de aplicações que as utilizam. É útil

para reconhecer caracteres escritos, comparar e identificar rostos humanos, aplicar filtros de correção para equipamento fotográfico e detectar comportamentos suspeitos através de vigilância por câmeras de vídeo.

**Web mining:** aplicações de mineração na Web são destinadas à análise dos chamados *clickstreams* ou sequência de cliques – são as sequências de páginas visitadas e as escolhas feitas por um internauta. Eles podem ser úteis para a análise de sites de comércio eletrônico, ao oferecer a flexibilidade e páginas personalizadas para os surfistas.

**Diagnóstico médico (*medical diagnosis*):** modelos de aprendizagem é uma ferramenta valiosa na área médica para a detecção precoce de doenças usando os resultados dos testes clínicos. A análise de imagens para fins de diagnóstico é outro campo de investigação que está em expansão.

### 3.4. Exemplos de aplicação das técnicas de *data mining*

Nesta sessão serão apresentados alguns exemplos de aplicações do uso de técnicas de mineração de dados para os casos de modelagem de previsão, análise de associação e agrupamento. A finalidade deste tópico é dar um melhor entendimento sobre a utilização e aplicabilidade destas técnicas em benefícios de seus usuários.

**Modelagem de previsão** → Para ilustrar a aplicação deste método, considere os dados constantes no Quadro 4:

Aparência	Temperatura	Umidade	Vento	Jogar tênis?
Ensolarado	Quente	Alta	Fraco	Não
Ensolarado	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Moderado	Alta	Forte	Sim
Chuvoso	Fresco	Normal	Fraco	Sim
Chuvoso	Fresco	Normal	Forte	Não
Nublado	Fresco	Normal	Forte	Sim
Ensolarado	Moderado	Alta	Fraco	Não
Ensolarado	Fresco	Normal	Fraco	Sim
Chuvoso	Moderado	Normal	Fraco	Sim
Ensolarado	Moderado	Normal	Forte	Sim
Nublado	Moderado	Alta	Forte	Sim
Nublado	Quente	Normal	Fraco	Sim
<b>Chuvoso</b>	Moderado	Alta	Forte	Não

Quadro 4: Base de dados da tarefa exemplo “Jogar tênis”  
Fonte: Adaptado de Goldschmidt e Passos (2005, p. 101)

Nesta ilustração de Goldschmidt e Passos (2005), iremos considerar o atributo “Jogar tênis” como o objetivo da classificação, este problema tem duas classes: “jogar = sim” e “jogar = não”. Se desejarmos saber se devemos ou não jogar em determinadas circunstâncias basta fazer uma pergunta e inserir os dados no algoritmo para obter o resultado, por exemplo: devo jogar tênis em dia ensolarado, quente, de alta umidade e com vento fraco?

No exemplo podemos utilizar o Teorema de Bayes, relacionado ao cálculo de probabilidades condicionais. A teoria desenvolvida por Bayes pode ser aplicada às mais diversas áreas do conhecimento, inclusive nas atividades cotidianas. Pelo Teorema de Bayes podemos afirmar que a  $P(A|B) = (P(B|A) \cdot P(A)) / P(B)$ , assim, substituindo os valores da nossa problemática na fórmula temos:

$$P(\text{jogar} = \text{sim} \mid \text{ensolarado, quente, alta umidade, vento fraco}) = P(\text{ensolarado} \mid \text{jogar} = \text{sim}) * P(\text{quente} \mid \text{jogar} = \text{sim}) * (\text{alta umidade} \mid \text{jogar} = \text{sim}) * P(\text{vento fraco} \mid \text{jogar} = \text{sim}) = 0,0071$$

$$P(\text{jogar} = \text{não} \mid \text{ensolarado, quente, alta umidade, vento fraco}) = P(\text{ensolarado} \mid \text{jogar} = \text{não}) * P(\text{quente} \mid \text{jogar} = \text{não}) * (\text{alta umidade} \mid \text{jogar} = \text{não}) * P(\text{vento fraco} \mid \text{jogar} = \text{não}) = 0,0274$$

Portanto, a resposta do algoritmo seria **Jogar = Não**.

**Análise de associação** → Uma tarefa de associação busca por padrões que demonstrem o relacionamento entre conjuntos de itens, para ilustrar a aplicação deste método, considere os dados apresentados no Quadro 5.

Identificador	Item
100	Pão, leite, manteiga.
200	Pão, requeijão, leite.
300	Manteiga, farinha, leite.
400	Manteiga, pão, refrigerante.
500	Bolacha, leite, manteiga.

Quadro 5: Um exemplo de transações em cesta de compras.  
Fonte: Adaptado de Pizzi (2006).

De acordo com Pizzi (2006), analisando a regra *manteiga* → *pão*, nota-se que dentre as 5 transações existentes, pão e manteiga ocorrem em 2 transações. Além disso, dentre as 4 transações em que manteiga ocorre, pão ocorre em 2 dessas transações. Com isso, pode-se dizer que a regra *manteiga* → *pão* possui suporte de 40% e confiança de

50%, o que demonstra que essa regra pode revelar um padrão de comportamento dos clientes: “clientes que compram manteiga tendem a comprar pão”.

Algoritmos de análise de associação possui um potencial de gerar uma variedade enorme de padrões com a combinação dos itens analisados, conforme são ajustados os limites de suporte e confiança.

**Agrupamentos** → O método de armazenamento em *cluster* permite que um usuário faça grupos de dados para determinar padrões a partir dos dados coletados ou classificados, criando um número específico de grupos, dependendo de suas necessidades de negócio. Neste tipo de aplicação os dados são divididos em um banco de dados por segmentos, onde seus membros compartilham características semelhantes comportamentos similares.

Um exemplo do uso de agrupamento são os empregados na construção de um CRM (*Customer Relationship Management*), que são aplicações que gerenciam todos os modos como às empresas lidam com seus clientes atuais e potenciais, objetivando desenvolver estratégias específicas para grupos de clientes de acordo com o padrão identificado nestes grupos. Pode ser utilizado também no sistema financeiro para discriminar e classificar bons e maus pagadores.

Os exemplos aqui ilustrados sevem para dar uma dimensão do uso das técnicas de mineração de dados nos mais diversos segmentos, para uma gama de aplicações variadas, servindo como um poderoso auxílio tecnológico em processos não triviais para identificar padrões válidos e potencialmente úteis para as organizações.

### 3.5. Cases de sucesso com *data mining*

Alguns *cases* de sucesso com o uso de técnicas de mineração de dados são facilmente encontrados na literatura e principalmente quando fazemos uma consulta rápida na internet. A seguir serão listadas algumas aplicações de *data mining* e quais os benefícios oriundos com a análise dos seus resultados:

#### - Wal-Mart

Este exemplo clássico cita o caso da rede de supermercados *Wal-Mart*, que identificou um hábito curioso dos seus consumidores, segundo Murayama (2002), foi

procurado relações entre o volume de vendas e os dias da semana, o *software* utilizado para exploração de dados apontou que, às sextas-feiras, as vendas de cerveja cresciam na mesma proporção que as de fraldas. Uma investigação mais detalhada revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cervejas para o final de semana. A nova informação foi confirmada e uma ação válida foi o lançamento imediato da promoção e localização próxima de fraldas e cervejas – nunca venderam tanto. Este exemplo pode ser utilizado para exemplificar os recursos que uma ferramenta de *data mining* e *data warehouse* podem oferecer.

#### **- Bank of America**

Segundo Braz *et al.* (2009), um outro caso de sucesso é o do *Bank of America*, que utilizou técnicas de mineração de dados para selecionar dentre seus clientes aqueles com menor possibilidade de não honrar com empréstimos. Com base nessa informação, foi feito um trabalho de marketing na oferta de crédito aos correntistas selecionados. Em apenas três anos, o *Bank of America* aumentou sua lucratividade em cerca de trinta milhões de dólares.

A detecção de fraudes é uma das aplicações mais visadas pelos gerentes que procuram por soluções em *data mining*. Diversos bancos recorrem a esse recurso para avaliar a credibilidade de seus clientes.

#### **- Outras aplicações**

No setor público, a mineração de dados é utilizada, principalmente, como forma de descobrir fraudes. Ainda segundo Braz *et al.* (2009), os governos brasileiro e norte americano, por exemplo, monitoram transações financeiras procurando detectar padrões similares à lavagem de dinheiro do narcotráfico. Na área de segurança pública a mineração de dados pode ser utilizada para: determinar os locais com maior criminalidade, definir perfis de vítimas e criminosos, identificar a existência de quadrilhas e *serial killers*, detectar quais dias da semana em que ocorrem mais delitos e até mesmo as suas causas, entre tantas outras possibilidades. Os benefícios em utilizar mineração de dados são muitos, por exemplo, descobrir os bairros que são mais perigosos pode induzir a uma melhor alocação dos policiais, bem como definir o perfil das vítimas pode ajudar a traçar campanhas preventivas focadas no público que se encaixa nesse perfil.

## **CAPÍTULO 4 - CONTROLE NO TRIBUNAL DE CONTAS**

Conforme determinação constitucional, art. 70 da Constituição Federal de 1988 (BRASIL, 2009), o exame das contas dos entes públicos brasileiros é efetuado pelo controle interno e pelo controle externo. O controle externo se efetivará pelo Poder Legislativo, com auxílio dos Tribunais de Contas, ao passo que o controle interno será levado a efeito por sistema do próprio ente público (CARVALHO, 2002).

O controle externo é exercido por órgão diverso, não pertencente à estrutura do responsável pelo ato controlado, como o controle que o Poder Judiciário pode vir a realizar sobre atos do Poder público, o controle direto ou indireto a cargo do Congresso Nacional e o realizado pelo Tribunal de Contas (NETO, 2004).

O controle constitucional terá como um de seus propósitos, executar a fiscalização financeira, orçamentária, operacional e patrimonial da administração, inclusive no tocante à legalidade do gasto público (DECOMAIN, 2005). Segundo o mesmo autor, em suma, o controle a cargo dos Tribunais de Contas, exercido nos termos do Art. 71, da mesma Constituição Federal de 1988, alcança autarquias, empresas públicas, sociedades de economia mista e fundações, instituídas ou simplesmente mantidas pelo Poder Público.

Segundo Santana (2008), as competências desse controle externo que são entre outras as de apreciar e julgar as contas prestadas pelos administradores públicos, aplicar sanções, assinar prazo para cumprimento de suas decisões e sustar determinados atos dos gestores públicos.

Os Tribunais de Contas, constitucionalmente, são os responsáveis pela fiscalização contábil, financeira, operacional e patrimonial do uso e aplicação dos recursos públicos brasileiros, exercendo o papel da fiscalização da eficiência no uso do dinheiro público. Desta forma, a função dos Tribunais de Contas ganha dimensões significativas, necessitando de instrumentos eficazes para a detecção e correção da utilização imprópria dos recursos públicos (CARVALHO, 2002).

Cabe aos Tribunais de Contas a função de acompanhar a gestão, fiscalizar e analisar os processos de prestações de contas, atos de admissão de pessoal, aposentadorias e pensões, licitações, contratos e convênios (TCE-PB, 2009). No Tribunal de Contas, os dados técnicos apresentados, ou seja, os demonstrativos



contábeis deverão necessariamente ser tabulados, analisados e confrontados com a legislação regente e emitido parecer sobre sua aprovação ou desaprovação (MELO, 2005).

De acordo com Santana (2008), os Tribunais de Contas, atualmente, possuem as seguintes atribuições que estão estabelecidas na nossa Constituição Federal nos artigos 70 a 75:

- apreciar as contas do chefe do Executivo;
- julgar as contas dos Administradores Públicos e demais responsáveis por dinheiro, bens e valores públicos da administração direta e indireta;
- apreciar, para fins de registro, a legalidade da contratação de pessoal, concessões de aposentadoria, reforma e revisão;
- realizar auditoria e inspeções de natureza contábil, financeira, orçamentária, operacional e patrimonial nas entidades da administração direta e indireta, quanto à legalidade, legitimidade, economicidade, aplicação das subvenções e renúncia de receitas;
- fiscalizar a aplicação de recursos públicos repassados;
- aplicar sanções em caso de irregularidades;
- determinar a adoção de providências para o exato cumprimento da Lei;
- sustar ato impugnado se não atendido;
- representar ao Poder competente sobre irregularidades ou abusos apurados.

Segundo o mesmo autor, além destas atribuições, a Lei de Responsabilidade Fiscal (LRF), desde que foi sancionada, estabelece, no seu Capítulo VIII, seção VI, que o Tribunal de Contas deverá fiscalizar a gestão fiscal dos Órgãos e Poderes Públicos de maneira concomitante.

De acordo com Figueirêdo (2002), aos Tribunais de Contas compete verificar o cumprimento da LRF, que está erigida sobre alguns pilares, dentre os quais o da transparência. O objetivo mais nobre do princípio da transparência é permitir e estimular o exercício do controle social, a mais eficaz das formas de controle da conduta do gestor público.

Segundo Xavier (2003), o controle externo abordado até o momento é também conhecido como institucional, ou seja, é aquele que obrigatoriamente é realizado pelas instituições públicas ou estatais de controle dos atos da administração pública, em virtude de normas constitucionais ou infraconstitucionais.

Além do Poder Legislativo e dos Tribunais de Contas, também integram este controle as Ouvidorias e o Ministério Público. O grande diferencial é que esta fiscalização não depende da vontade política das autoridades a ser fiscalizadas, ou das autoridades fiscalizadoras quando atuam, mediante denúncia, representação ou o conhecimento de eventuais irregularidades, mas sim, cumprem determinações legais e não podem deixar de fazê-las, sob pena de prevaricação.

#### 4.1. O uso das tecnologias de BI nos Tribunais de Contas

Segundo Slack *et al.* (2007) O processo de tomada da decisão em organizações sem fins lucrativos, como no caso dos Tribunais de Contas, pode ser mais complexo e envolvem uma mistura de objetivos políticos, econômicos, sociais ou ambientais, por causa disso, existe maior chance de que as decisões de operações ocorram em condições de objetivos conflitantes.

A adoção da tecnologia de BI pode dar uma nova dinâmica ao processo de análise das contas públicas, e consequentemente proporcionar um aumento na produtividade através de respostas mais rápidas, reduzindo o tempo total de análise dos processos, envolvendo desde a fase de análise inicial do processo até o parecer final destas contas pelos Tribunais de Contas.

As contas públicas são consideradas regulares, quando expressarem, de forma clara e objetiva, a exatidão dos demonstrativos contábeis, e atenderem aos princípios da legalidade, legitimidade, economicidade e razoabilidade dos atos de gestão do responsável (TCE-SE, 2009).

Atualmente os Tribunais de Contas buscam ferramentas que auxiliem o processo de fiscalização, transparência e para a tomada da decisão. Observem os seguintes casos:

Segundo Sousa *et al.* (2008) um estudo acerca da utilização de *data mining* foi realizado no Tribunal de Contas do Estado de Pernambuco (TCE-PE), que apresentou uma solução de mineração de dados, com a extração do conhecimento contido na própria base de dados que compõe o estoque de processos do TCE-PE, para a

proposição de um Sistema de Apoio à Decisão (SAD) como instrumento de atuação efetiva na celeridade de execução dos processos daquela Corte de Contas.

Santana (2008) apresenta o Sistema de Acompanhamento da Gestão dos Recursos da Sociedade (SAGRES), criado e implantado pelo Tribunal de Contas do Estado da Paraíba (TCE-PB) em 2002, que possui uma ferramenta de consulta utilizando conceitos básicos de mineração de dados e *data warehouse*. Para o autor, este sistema objetivou proporcionar meios de realizar a contento as novas atribuições que a Lei de Responsabilidade Fiscal (LRF) impôs.

A LRF normatiza os princípios do equilíbrio - não gastar mais do que se arrecada; do planejamento - planejar para executar; da transparência - prestar contas e informar à sociedade; e da participação - governar com o cidadão.

#### 4.2. O Sistema de Acompanhamento da Gestão dos Recursos da Sociedade (SAGRES)

O SAGRES é um sistema computadorizado que objetiva dar apoio ao exercício do controle externo que tem como base a coleta mensal e a análise e disponibilização para a sociedade de dados informatizados sobre a execução orçamentária e financeira, sobre licitações, contratos administrativos e sobre as despesas com pessoal das unidades gestoras municipais e estaduais.

O SAGRES dispõe de um módulo de captura, armazenamento e consultas dos dados dos seus jurisdicionados, e disponibiliza no web site denominado SAGRES ONLINE alguns módulos deste sistema para que a sociedade possa exercer a fiscalização dos gastos públicos realizados pelas Prefeituras, Câmaras Municipais e Órgãos da Administração Indireta municipal, além das despesas realizadas pelo Governo do Estado (GRILO JR.; CLERICUZI, 2008). A Figura 10 mostra o contexto onde o *software* SAGRES está inserido para a coleta, análise e disponibilização de dados para a sociedade.



Figura 10: Diagrama de Contexto do SAGRES  
 Fonte: Adaptado de Santana (2008, p. 67)

Analisando a Figura 10, identificamos o fluxo dos dados da execução orçamentária e financeira, que são coletados pelo módulo de Captura do Sagres e enviados mensalmente ao TCE-PB por seus jurisdicionados. Então estes dados são tratados, armazenados e disponibilizados para consulta através do módulo Auditor, para os usuários internos, e pelo *web site* do Tribunal de Contas (Sagres On-Line), para a consulta pública para o cidadão e por órgãos fiscalizadores do controle externo.

O SAGRES possui dois mecanismos principais de acompanhamento da Gestão Pública: o cadastramento das informações relativas a Gestão, que constituem por si só a formação de um banco de dados do TCE-PB, fonte de consulta ilimitada para toda a sociedade, e o acompanhamento da aplicação dos recursos públicos dentro dos ditames da legislação vigente (XAVIER, 2003).

O SAGRES foi adotado pelo Tribunal de Contas do Estado do Piauí (TCE-PI) em 2006 através da celebração de um termo de convênio de cooperação com o TCE-PB, e em maio de 2010 foi celebrado outro convênio, desta vez com o Tribunal de Contas de do Estado de Pernambuco (TCE-PE) para a implantação e adoção do *software* desenvolvido pelo TCE-PB, para a realização da captura (coleta) dos dados dos jurisdicionados, uma vez que o SAGRES já é um sistema estável, de fácil operacionalização e manuseio.

Para o TCE-PI (2010), a implantação do SAGRES tem por objetivo dar celeridade à apreciação das contas anuais, bem como auxiliar o acompanhamento e a fiscalização das gestões orçamentária e financeira da administração municipal, com

vistas ao recebimento e sistematização eletrônica da coleta e da remessa de dados necessários à composição da prestação de contas dos agentes públicos.

A prestação de contas é exigida no parágrafo único do Artigo 70 da Constituição Federal que estabelece que deverão prestar contas qualquer pessoa física ou jurídica, pública ou privada, que utilize, arrecade, guarde, gerencie ou administre dinheiros, bens e valores públicos ou pelos quais a União responda, ou que, em nome desta, assuma obrigações de natureza pecuniária (BRASIL, 2009).

## CAPÍTULO 5 - ASPECTOS METODOLÓGICOS DA PESQUISA

Gil (1991) considera a pesquisa como um procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos.

Todos os conceitos de pesquisa, de uma forma ou de outra, apontam seu caráter racional, como conceituado por Andrade (2009) que define pesquisa como um conjunto de procedimentos sistemáticos, baseados no raciocínio lógico, que tem por objetivo encontrar soluções para problemas propostos, mediante a utilização de métodos científicos. Para Lakatos e Marconi (2007) os procedimentos metodológicos estão associados aos caminhos, formas, maneiras, métodos para atingir determinado fim.

Há vários métodos para a realização de uma pesquisa, e este capítulo tem o objetivo de apresentar a classificação e os procedimentos escolhidos para a aplicação desta, ou seja, explicar como foram obtidos, tratados e analisados os dados deste estudo.

### 5.1 Classificação da pesquisa

Quanto à natureza, esta pesquisa pode ser classificada como aplicada, segundo Lakatos e Marconi (2007), a pesquisa aplicada é caracterizada por seu interesse prático, ou seja, uma aplicação ou a utilização imediata dos resultados, na solução de problemas que ocorrem na realidade.

Para Menezes e Silva (2001), uma pesquisa de natureza aplicada objetiva gerar conhecimento para aplicar práticas dirigidas à solução de problemas específicos.

A pesquisa de natureza aplicada faz referência ao processo investigativo e desenvolvimento de novos conhecimentos ou a compreensão dos já existentes, necessários para determinar os meios pelos quais se pode desenvolver e aprimorar produtos, processos ou sistemas, com vistas à satisfação de uma necessidade específica e reconhecida. Visto que este estudo objetivou gerar conhecimentos para a aplicação prática relacionada à utilização de ferramentas de BI, especificamente usando técnicas de *data mining* para a detecção de indícios de fraudes e irregularidades em processos licitatórios no âmbito de competência do Tribunal de Contas do Estado da Paraíba, sendo classificada desta forma.

Do ponto de vista da abordagem do problema, foi classificada com sendo uma pesquisa quantitativa, pois considera que tudo pode ser quantificável, o que significa

traduzir em números opiniões e informações para classificá-las e analisá-las. Para Teixeira (2002), a pesquisa qualitativa utiliza a descrição matemática como linguagem para descrever as características de um fenômeno. O método quantitativo descreve os fatos por meio de métodos estatísticos.

Quanto aos objetivos classificou-se como uma pesquisa exploratória e descritiva. Segundo Gil (1991), a pesquisa exploratória visa proporcionar maior familiaridade com o problema com vistas a torná-lo explícito ou a construir hipóteses. Envolve levantamento bibliográfico; entrevistas com pessoas que tiveram experiências práticas com o problema pesquisado; e análise de exemplos que estimulem a compreensão. Ainda segundo o autor, esta pesquisa assume, em geral, as formas de pesquisas bibliográficas e estudos de caso.

A pesquisa foi descritiva, pois procurou descrever e caracterizar os processos adotados pela organização para o uso das técnicas de *data mining*, com o intuito do acesso às informações necessárias para a tomada de decisão, bem como relatar os benefícios que poderão ser trazidos pela implantação de tal ferramenta na operacionalização das atividades dos técnicos do TCE-PB, especificamente na análise das informações coletadas para a elaboração do relatório de auditoria e pareceres.

## 5.2 Delimitação do tema

Este estudo de caso foi realizado a partir dos dados coletados pelo sistema SAGRES Captura na base de dados do TCE-PB, abrangendo principalmente os dados constantes na tabela de Licitações. Os dados analisados foram enviados ao Tribunal de Contas pelos jurisdicionados municipais através do referido *software*. Não fazem parte do escopo desta pesquisa às demais informações constantes neste banco de dados, como os dados das receitas, despesas orçamentárias e extra orçamentárias, dados da folha de pessoal e da execução de obras.

## 5.3 Seleção dos dados da pesquisa ou caso

A legislação que rege o processo de compras e contratações governamentais é a Lei nº 8666/93 – “Lei de Licitações e Contratos Administrativos” (BRASIL, 2010), que regulamenta o Artigo 37, inciso XXI, da Constituição Federal, e institui as normas para licitações e contratos da Administração Pública, que vincula no seu Artigo 113 a

competência dos Tribunais de Contas para a fiscalização da execução dos contratos firmados, a saber:

O controle das despesas decorrentes dos contratos e demais instrumentos regidos por esta Lei será feito pelo Tribunal de Contas competente, na forma da legislação pertinente, ficando os órgãos interessados da Administração responsáveis pela demonstração da legalidade e regularidade da despesa e execução, nos termos da Constituição e sem prejuízo do sistema de controle interno nela previsto (BRASIL, 2010). (grifo nosso)

Como o objeto deste estudo é a identificação dos indícios de fraudes por irregularidades e desvio de recursos públicos oriundos de processos licitatórios com a administração pública, a seleção dos dados compreende as tabelas de Licitação, Participantes, Propostas, Fornecedores, Contratos e Aditivos, que foram extraídas do Sistema de Gerenciamento de Banco de Dados (SGBD) MS-SQL Server 2000, que compõe o a base de dados do SAGRES, conforme visualizadas na Figura 11.

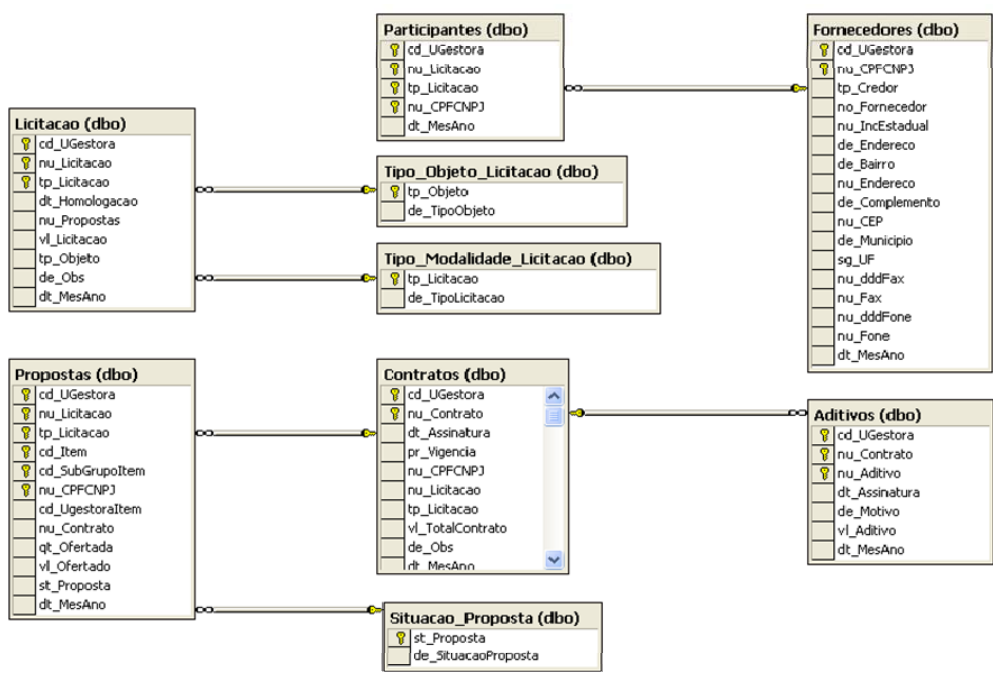


Figura 11: Tabela de dados do módulo de Licitações utilizadas pelo SAGRES  
Fonte: TCE-PB (2010).

### 5.3.1 Principais indícios de fraudes e/ou irregularidades em licitações

A Lei nº 8666/93 (BRASIL, 2010), ressalta no Artigo 1º sobre a sua competência e no Artigo 4º sobre a forma de proceder e fiscalizar as licitações realizadas:



Art. 1º: Esta Lei estabelece normas gerais sobre licitações e contratos administrativos pertinentes a obras, serviços, inclusive de publicidade, compras, alienações e locações no âmbito dos Poderes da União, dos Estados, do Distrito Federal e dos Municípios.

(...)

Art.4 º: Todos quantos participem de licitação promovida pelos órgãos ou entidades a que se refere o artigo 1º têm direito público subjetivo à fiel observância do pertinente procedimento estabelecido nesta lei, podendo qualquer cidadão acompanhar o seu desenvolvimento, desde que não interfira de modo a perturbar ou impedir a realização dos trabalhos.

Como o objetivo de identificar irregularidades nos processos licitatórios, a utilização dos algoritmos de mineração de dados se propõe a pesquisar em uma base de dados a procura de padrões evidentes e ocultos que possam ajudar a identificar os indícios de irregularidades, como exemplo, no Quadro 6 estão listados algumas situações que demonstram irregularidades em processos licitatórios de acordo com os procedimentos normatizados pela Lei nº 8666/93.

Item	Padrões Identificados	Interpretação
1	Participação do mesmo grupo de fornecedores (proponentes) em outros processos licitatórios, alguma vez só alterando o nome do licitante vencedor.	<ul style="list-style-type: none"> <li>· Inexistência de competitividade</li> <li>· Favorecimento de fornecedores</li> <li>· Direcionamento de resultados</li> </ul>
2	Realização de aquisições por dispensa de licitação de um mesmo material várias vezes em um mesmo ano, e para um determinado proponente.	<ul style="list-style-type: none"> <li>· Inexistência de competitividade</li> <li>· Favorecimento de fornecedores</li> <li>· Direcionamento de resultados</li> <li>· Fracionamento de Licitação</li> <li>· Não aplicação dos princípios relativos às compras governamentais</li> </ul>
3	Realização de aquisições por dispensa de licitação para determinado proponente, ultrapassando o limite permitido estabelecido em lei.	<ul style="list-style-type: none"> <li>· Inexistência de competitividade</li> <li>· Favorecimento de fornecedores</li> <li>· Direcionamento de resultados</li> <li>· Não aplicação dos princípios relativos às compras governamentais</li> </ul>

Quadro 6: Indícios de irregularidade nos processos licitatórios

Fonte: Adaptado de Rezende (2005); O Autor.

Os indícios acima expostos representam apenas algumas das irregularidades que

podem ocorrer em um processo licitatório, não esgotando o levantamento e identificação de outras ocorrências.

### 5.3.2 Critérios da amostra ou caso

Este estudo considera como ambiente pesquisado os dados provenientes das licitações realizados pelos 223 municípios paraibanos, compreendendo as Prefeituras Municipais, as Câmaras Municipais e os Órgãos da Administração Indireta Municipal (Autarquias, Fundações, Empresas e Sociedades de Economia Mista), correspondendo ao período de 2005 a 2009. Os dados das licitações homologadas são enviados mensalmente ao TCE-PB através do *software* Sagres Captura. Segundo Lakatos e Marconi (2007) o universo em uma pesquisa é o conjunto de seres animados ou inanimados que apresentam pelo menos uma característica em comum.

### 5.4 Coleta dos dados

A coleta de dados é o meio utilizado para o qual o pesquisador obtém as informações sobre o problema da pesquisa. Baseado neste fato, este estudo considera os dados extraídos da base de dados do SAGRES, especificamente relacionados ao módulo de licitações do TCE-PB. O Quadro 7 traz um extrato da quantidade de processos licitatórios constantes no banco de dados, catalogadas por tipo de modalidade de licitação, compreendendo o período de 2005 a 2009.

Modalidade de Licitação	2005	2006	2007	2008	2009	Total	%
<b>Pregão (Eletrônico e Presencial)</b>	211	593	978	1408	2068	5258	10,68%
<b>Concorrência</b>	23	55	50	50	52	230	0,47%
<b>Tomada de Preço</b>	541	859	887	873	968	4128	8,38%
<b>Convite</b>	6336	6458	6130	5539	6333	30796	62,53%
<b>Concurso</b>	4	5	2	0	2	13	0,03%
<b>Leilão</b>	25	31	19	10	26	111	0,23%
<b>Dispensa por outros motivos</b>	232	278	311	313	575	1709	3,47%
<b>Inexigível</b>	642	726	1630	1432	2457	6887	13,98%
<b>Adesão a Registro de Preço</b>	0	0	0	8	110	118	0,24%
<b>Subtotal</b>	10.019	11.011	1.2014	11.641	14.600	49.250	100,00%

Quadro 7: Processos licitatórios enviados ao TCE-PB por tipo de modalidade de licitação

Fonte: TCE-PB (2010).

Baseado no quadro acima, se observa uma predominância pela modalidade de licitação Carta Convite, representando 62,53% do total dos processos licitatórios realizados. A modalidade de licitação Carta Convite é utilizada para contratações de menor vulto, ou seja, para a aquisição de materiais e serviços até o limite de R\$ 80.000,00 (oitenta mil reais), e para a execução de obras e serviços de engenharia até o valor de R\$ 150.000,00 (cento e cinquenta mil reais). No quadro 8 está a discriminação das modalidades de licitação por participantes, ou seja, a quantidade de proponentes para cada modalidade coletada pelo SAGRES.

Modalidade de Licitação	2005	2006	2007	2008	2009	Total	%
<b>Pregão (Eletrônico e Presencial)</b>	784	1832	2982	3778	6428	15804	10,52%
<b>Concorrência</b>	95	270	277	235	259	1136	0,76%
<b>Tomada de Preço</b>	2651	3440	3669	3536	3984	17280	11,51%
<b>Convite</b>	19749	19572	18772	16908	19379	94380	62,85%
<b>Concurso</b>	10	18	6	0	9	43	0,03%
<b>Leilão</b>	80	63	45	19	64	271	0,18%
<b>Dispensa por outros motivos</b>	500	572	532	421	1436	3461	2,30%
<b>Inexigível</b>	826	878	1736	1580	2569	7589	5,05%
<b>Adesão a Registro de Preço</b>	0	0	0	13	164	177	0,12%
<b>Subtotal</b>	26.700	28.651	30.026	28.498	36.301	150.176	100,00%

Quadro 8: Quantidade de participantes por tipo de modalidade de licitação  
Fonte: TCE-PB (2010).

Analisando os dados constantes dos Quadros 7 e 8, obtivemos o total de 49.200 processos licitatórios no período compreendido por esta pesquisa, totalizando 150.176 proponentes distribuídos pelas nove modalidades discriminadas nos quadros acima mencionados. Observa-se que o tipo de modalidade Adesão a Registro de Preço só apresenta dados a partir de 2008, pois antes desta data não eram exigidos e capturados pelo TCE-PB, por este motivo é que estão com os valores zerados no período de 2005 a 2007.

No Sagres a análise dos dados enviados é realizada através de um módulo do sistema que recebe e valida o arquivo contendo os dados provenientes dos sistemas contábeis e de gestão das unidades gestoras, conforme demonstrado na Figura 12. Estas informações são enviadas pelos jurisdicionados através do Sagres Captura e após a verificação de possíveis inconsistências, os mesmos são validados e armazenados no

banco de dados do TCE-PB. Na etapa seguinte, os dados válidos são liberados para a consulta dos técnicos do Tribunal de Contas e posteriormente disponibilização na internet para a consulta pública.



Figura 12: Adaptado do demonstrativo de funcionamento do SAGRES Captura  
Fonte: SANTANA (2008)

Até o ano de 2008, o envio dos dados era realizado através de disquetes, a partir desta data passaram a ser enviados ao TCE-PB através do portal web institucional denominado Portal do Gestor. O acesso aos dados capturados pelo Sagres é realizado pelos técnicos do TCE-PB através do aplicativo Sagres Auditor, que disponibiliza diversas consultas e verificações para que seus usuários do sistema possam realizar as devidas averiguações. Na Figura 13 visualizamos a tela do formulário principal do Sistema Sagres – Módulo Auditor, com destaque para o módulo de Licitações.

Neste aplicativo, estão disponíveis os dados do Governo Estadual e Municipal, além de rotinas específicas para a auditoria e emissão de relatórios gerenciais e de acompanhamento da gestão.

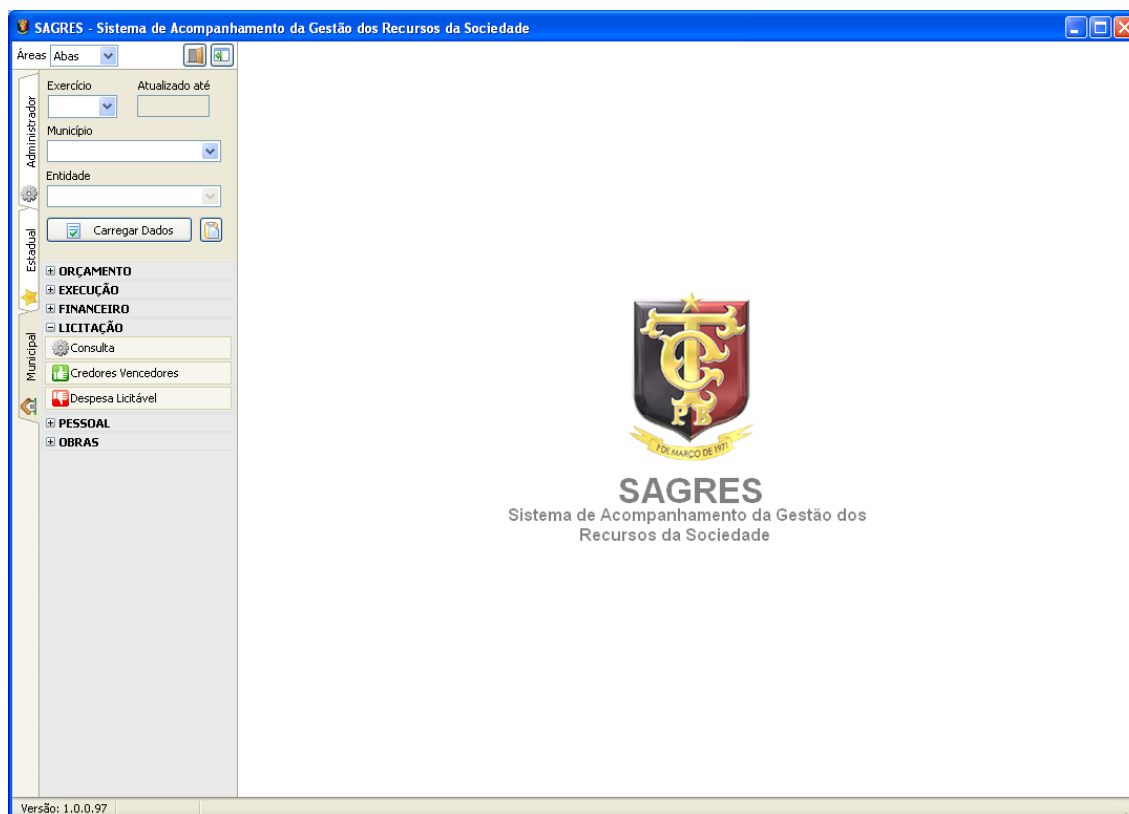


Figura 13: Interface principal do aplicativo Sagres Auditor  
 Fonte: SAGRES

Através da funcionalidade da consulta no módulo de Licitações, os usuários do sistema tem acesso ao formulário com a listagem de todas as licitações realizadas pelos jurisdicionados, conforme mostrado na Figura 14. Para ter acesso a esta funcionalidade do sistema, antes deve ser indicada qual a unidade gestora que se deseja analisar.

No quadro seguinte os dados são exibidos de forma consolidada, agrupados por exercício. Observe que na expansão do detalhamento de cada exercício, são exibidas as modalidades de licitação que foram realizadas trazendo o total licitado e a quantidade de processos.

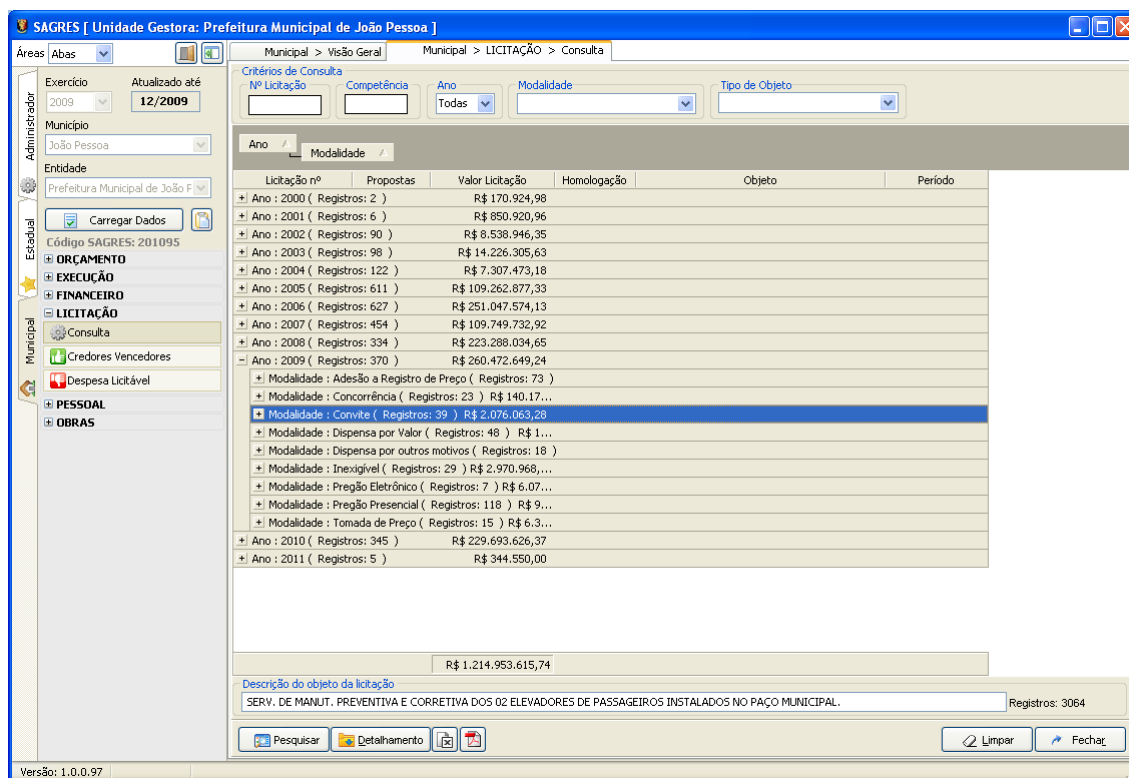


Figura 14: Consulta de licitações no módulo Sagres Auditor  
Fonte: SAGRES

O detalhamento da licitação é visualizado através dos botões de acesso disponíveis no formulário de consulta visualizado na Figura 14. Ao ser pressionado o botão “Detalhamento” o sistema Sagres Auditor exibe a tela constante na Figura 15, que exibe os participantes da licitação selecionada, o valor de cada proposta e a situação da mesma (vencedora ou perdedora), além da opção de consultar onde estes participantes estão concorrendo em outras licitações, dando uma visão ampliada da área de atuação de cada proponente.

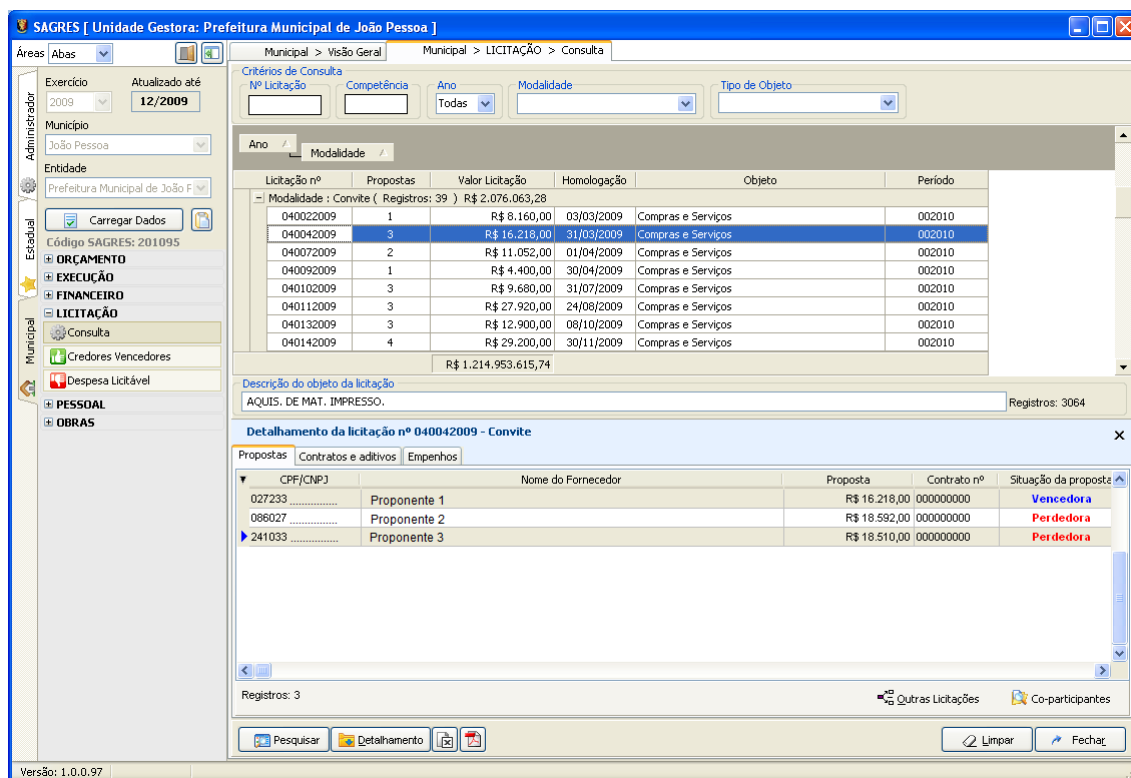


Figura 15: Detalhamento da licitação no módulo Sagres Auditor  
Fonte: SAGRES

Na Figura 15, ao ser pressionado o botão “Outras Licitações” é exibido onde o participante selecionado estão participando em outros processos licitatórios na base de dados do SAGRES, esta consulta busca as informações em toda a base de dados, independente do exercício (ano) e de unidade gestora (entidade). Observem que no quadro acima, o nome do fornecedor proponente foi omitido para não expor nesta pesquisa dados de uma ou outra empresa, assim preservando a sua identidade. Esta consulta poderá ser realizada sem restrições no Sagres *On-Line*, disponível em <http://portal.tce.pb.gov.br>.

Na opção “Co-participantes” a consulta exhibe os processos licitatórios onde todos os participantes da licitação selecionada estão participando juntos em outros processos licitatórios. Esta consulta exhibe uma visão bastante interessante, pois quando há um retorno de registros indica um provável indício de irregularidade no processo licitatório, mas esta consulta não é realizada de forma automática pelo sistema, tem que ser acionada manualmente e se limita ao número máximo de até quatro participantes por processo pesquisado, devido ao tempo de processamento. Aliado a tudo isso, ocorre que esta consulta não poder ser ranqueada para trazer um escore ao técnico que está fazendo a análise.

A proposta, neste caso, é disponibilizar estes indícios de irregularidades e/ou fraudes nos processos licitatórios a todos os usuários do SAGRES através do processamento realizado pela ferramenta de mineração de dados escolhida.

Com a utilização de técnicas de mineração de dados, os técnicos do TCE-PB poderão receber alertas sobre a possibilidade de ocorrência de fraudes, seja caracterizada por uma situação de fraude cadastral ou fraude transacional (comportamental). Os resultados esperados são de coibir desvios de recursos públicos causados por fraudes, e ainda dispor de um maior controle sobre estas operações, possibilitando assim que a unidade gestora (prefeitura, câmara municipal, órgão da administração indireta municipal) possa ser examinada “*in loco*” para averiguação da veracidade sobre a irregularidade apontada pelo *software* de mineração de dados.

Este procedimento poderá trazer um ganho de tempo na análise dos dados provenientes dos processos licitatórios, ocasionando consequentemente um ganho de produtividade e direcionando esforços para uma fiscalização mais eficiente.

## 5.5 Análise dos dados

Para a realização desta pesquisa foi utilizado o aplicativo RapidMiner. A versão utilizada nos experimentos foi a *Free 5.3*, licenciada sob GPL (*General Public License*) versão 3. O ambiente de trabalho da ferramenta pode ser visualizado na Figura 16.

O RapidMiner é uma ferramenta com interface gráfica que conta com muitos recursos para diversas tarefas para a descoberta de conhecimento e mineração de dados, de acordo com Rapidminer (2010) este *software* é disponibilizado como uma aplicação *stand-alone* para a análise e mineração de dados e possui um engenho de busca que pode ser integrado em produtos próprios, tais como sistemas em uso na organizações, através de uma API (*Application Programming Interface*) que permite a troca comunicação entre aplicativos diferentes.

Uma API é um conjunto de rotinas e padrões estabelecidos por um *software* para a utilização das suas funcionalidades por outros aplicativos fornecendo um nível de abstração para garantir a portabilidade do código (FOLDOC, 2010).



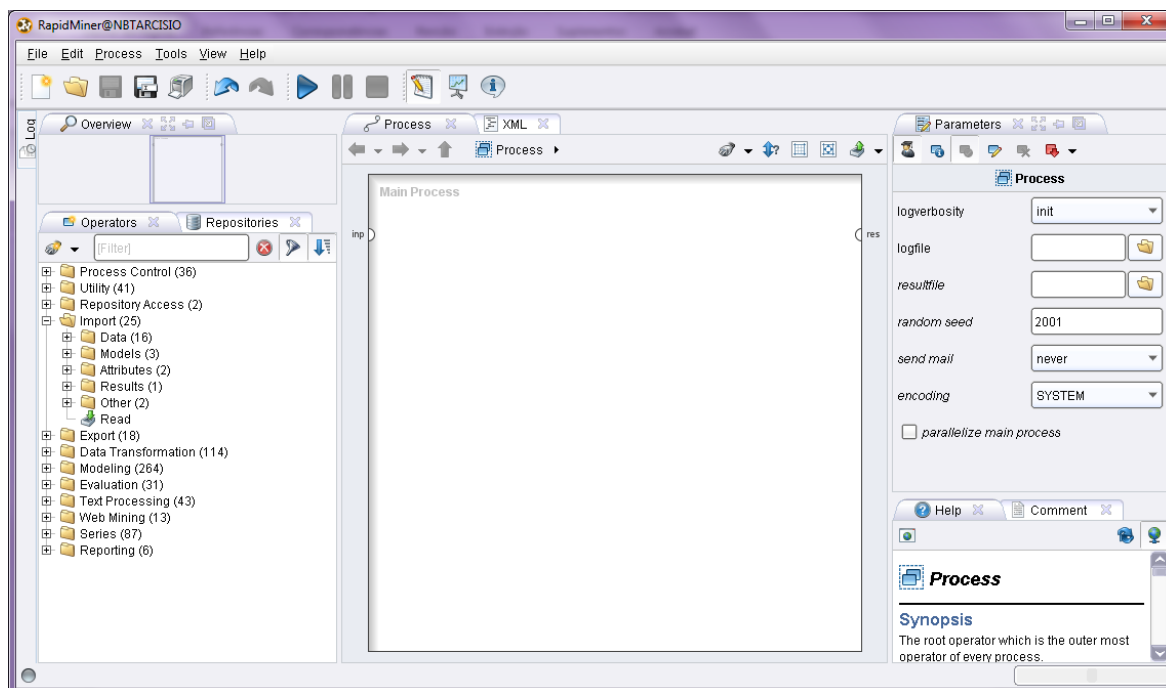


Figura 16: Ambiente de trabalho do RapidMiner

Segundo Coelho (2008), as operações no RapidMiner se dão por processos, isto significa que, para realizar uma classificação de dados nesta ferramenta, faz-se necessário incluir operadores para cada processo que antecede a tarefa final desejada, sendo que alguns operadores possuem parâmetros configuráveis, necessitando que o usuário informe os valores iniciais para alguns parâmetros. A lista de operadores, sua sequência de execução e parâmetros são armazenados em um arquivo do tipo XML (*Extensible Markup Language*), o qual permite ser editado para os ajustes quando necessários, ou mesmo para que sejam gravadas as configurações do experimento realizado para reutilização futura.

O RapidMiner conta com dezenas de operadores para as mais diversas finalidades nas organização, tornando-se uma das opções mais viáveis para usuários de todos os níveis de conhecimentos devido a sua *interface* gráfica amigável e bastante funcional.

## 5.6 Método abordado

Conforme mostrado no Quadro 5, podem ser utilizados algoritmos de mineração de dados objetivando encontrar padrões existentes no banco de dados do SAGRES do Tribunal de Contas do Estado da Paraíba na busca da identificação de padrões e

extração de regras que possam ser representativas para esta pesquisa e que evidencie aspectos que possam representar indícios de irregularidade em processos licitatórios.

Desta forma, foram escolhidos os algoritmos para Regras de Associação, a tarefa de descoberta de associações, como mencionada anteriormente por Goldschmidt e Passos (2005), pode ser definida como a busca por regras de associação frequentes e válidas em um banco de dados. Desse modo, o objetivo das regras de associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento nos dados analisados.

### 5.7 Roteiro da pesquisa

Os dados necessários para o referido estudo foram coletados da base de dados do TCE-PB, de acordo com as tabelas de dados visualizadas na Figura 17. Após a coleta dos dados, a fase seguinte da pesquisa é a análise e interpretação dos mesmos. Segundo Marconi e Lakatos (2007), antes da análise e interpretação dos dados, é preciso organizá-los de forma a fazer a seleção (exame minucioso dos dados); codificação (técnica operacional utilizada para categorizar os dados) e a tabulação (disposição dos dados em tabelas e gráficos).

Os resultados esperados nesta pesquisa são a redução do desvio de recursos públicos causados por procedimentos fraudulentos, além de um maior controle sobre estas operações. De posse de informações adicionais sobre indícios de irregularidade, poderá ser realizada uma diligência para a verificação da irregularidade apontada pelo algoritmo de mineração de dados.

Para a fase de Extração, Transformação e Carga (ETL) foi criado um *data mart* setorial com os dados dos processos licitatórios, envolvendo os proponentes e suas respectivas propostas, conforme visualizado na Figura 17.



Figura 17: Data mart de Licitações

Nesta fase foi realizada uma limpeza para descartar os dados que apresentaram algum tipo de inconsistência e para a remoção dos dados que não fizessem parte do escopo desta pesquisa, com os dados do exercício de 2003 e de 2004, que fazem parte da base de dados completa, mas foram descartados na fase de transferência e carga dos dados, consideramos para este estudo os anos de 2005 a 2009. O intervalo selecionado apresentou dados mais concisos, tendo em vista a maturidade do programa de captura de dados e do aprendizado do uso do mesmo por parte dos jurisdicionados municipais e de uma maior fiscalização realizada pelo TCE-PB.

Como as tarefas de mineração de dados exigem muito recurso de processamento, então a maior parte dos experimentos iniciais foram realizados em um computador DELL Intel Xeon, com dois processadores de 2,27 Ghz, sistema operacional Windows-64 bits e 32 GB de memória RAM, dos quais foram alocados 30 GB para o uso do RapidMiner. Alguns testes também foram realizados em um computador com processador Intel Core i3, 4 GB de RAM, o que ocasionou um tempo maior na execução das tarefas e em muitos casos até travamento no computador por falta de mais recursos de memória livre, mas no decorrer dos testes realizados foram feitos ajustes na combinação dos algoritmos de tal forma que os procedimentos foram otimizados e resolvida a questão dos travamentos que ocorreram nos primeiros experimentos.

A escolha por um computador mais robusto deu-se aos muitos testes realizados para a escolha do melhor método para a mineração dos dados estudados e que dependiam de uma resposta rápida para a análise e validação do experimento.

#### 5.7.1 Criação do modelo de dados

A partir da base de dados de licitações, visualizada na Figura 17, foi gerada uma relação de Licitação *versus* Proponente, formando uma tabela como ilustrada no Quadro 9. Esta consulta foi realizada usando comandos de consulta diretamente no *data mart* de Licitações, selecionando os atributos *Licitação* e *Proponente*. Em seguida estes dados foram exportados para um arquivo no formato CSV (*Comma-Separated Values*), este tipo é um formato de arquivo de valores separados por vírgula é usado para armazenar dados tabulares.

A opção para a utilização do arquivo no formato CSV foi simplesmente pelo fato do banco de dados estar localizado em um equipamento diferente do que foi realizado o procedimento de mineração de dados e assim facilitar o seu manuseio.

Licitação	Proponente
lic-001	prop-001
lic-001	prop-002
lic-001	prop-003
lic-002	prop-004
lic-002	prop-005
...	...
lic-00n	prop-00m

Quadro 9: Relação Licitação versus Proponentes

A tabela acima é então transformada em uma matriz Licitação *versus* Proponente, como ilustrada no Quadro 10. Dada à característica esparsa desta tabela, a mesma é armazenada em um arquivo texto que grava apenas os atributos com valor diferente de zero. Uma matriz esparsa é implementada através de um conjunto de listas ligadas que apontam para elementos que possuem valores, os registros de valor zero não são armazenados, desta forma otimizamos o tamanho do arquivo.

No Quadro 10 podemos visualizar a matriz Licitação *versus* Proponente. Para um melhor entendimento, o valor “1” no campo  $\{i; j\}$  significa que o prop-00i participou da licitação lic-00j, e o valor “0” no campo  $\{i; j\}$  significa que o prop-00i não participou da licitação lic-00j.

Licitações	prop-001	prop-002	prop-003	prop-004	prop-00m
lic-001	1	1	1	0	...
lic-002	0	0	0	1	...
lic-00n	...	...	...	...	...

Quadro 10: Matriz Esparsa Licitações *versus* Proponentes

### 5.7.2 Procedimento realizado

Nesta etapa apresentamos um procedimento para minerar dados a partir do arquivo com os dados extraído no formato CSV, conforme explicado no item anterior. Dentre outros algoritmos que são recomendados para tarefas de associação, optamos pelo algoritmo *FP-Growth*, conforme etapas constantes na Figura 18, a qual ilustra os procedimentos que incluem os objetos para a leitura do arquivo contendo os dados a serem pesquisados, a preparação dos mesmos e as tarefas de associação com a geração e apresentação das regras obtidas com os dados processados.

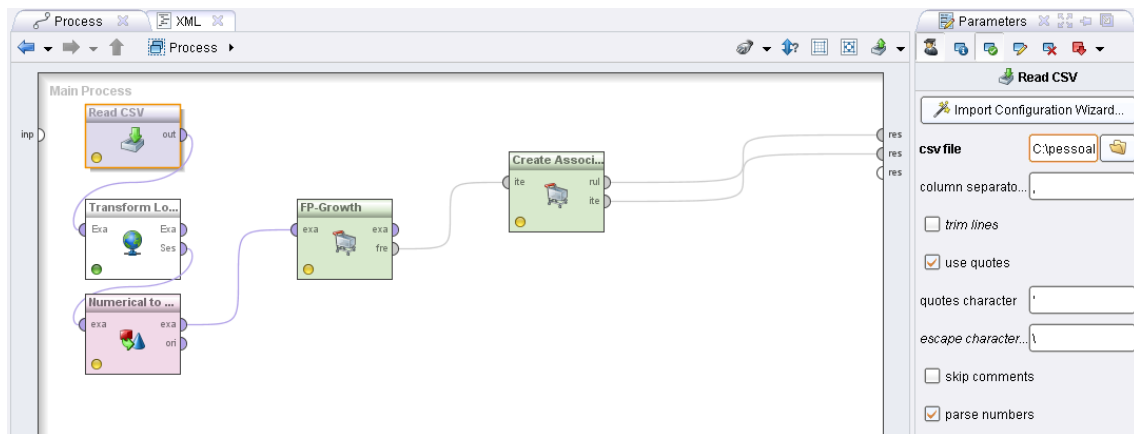


Figura 18: Área de trabalho principal do RapidMiner com o os objetos usados no experimento realizado

O processo do RapidMiner para este experimento possui as seguintes etapas:

- **Read CSV:** lê o conteúdo do arquivo no formato CSV que contém os dados das licitações com seus respectivos proponentes.
- **Transform Log to Session:** transforma os dados das propostas na matriz esparsa. O atributo de seção (linha da matriz) está configurado para ser o atributo *Licitação*. O atributo recurso (coluna da matriz) está configurado para ser o *idProponente* (identificador do proponente na tabela de dados).
- **Numerical to Binomial:** converte todos os atributos numéricos para binários.
- **FP-Growth:** calcula todos os conjuntos de itens frequentes dos dados fornecidos. O algoritmo *FP-Growth* (*Frequent Pattern Growth*) é um método eficiente e escalável para a mineração de padrões frequentes sejam eles curtos ou longos. Utiliza uma estrutura chamada *FP-Tree*, baseada no crescimento de fragmentos de padrões, que armazena informação quantitativa sobre padrões frequentes de forma comprimida.

O único parâmetro ajustável é o *min support*, que é o valor do suporte mínimo a ser considerado de acordo com as características do problema a ser resolvido.

- **Create Association Rules:** este operador gera regras de associação para um dado conjunto de itens frequentes. No RapidMiner, o processo de mineração para o conjunto de itens frequentes é dividido em duas partes:

primeiro, a geração de conjuntos de itens frequentes e, segundo, a geração de regras de associação a partir desses conjuntos.

Para reproduzir o experimento descrito acima basta utilizar o código constante do Apêndice I e executar o procedimento no RapidMiner associando a uma base de dados para testes. Quanto ao arquivo CVS, o mesmo possui a seguinte estrutura:

Cabeçalho: 'licitacao', 'idProponente'

Registros: 'identificador da licitação', 'identificador do proponente'

O identificador da licitação é uma concatenação do número da licitação, tipo de modalidade de licitação e unidade gestora que foi realizado o processo licitatório; já a identificação do proponente é um código inteiro. Para um melhor entendimento sobre o identificador de cada licitação, veja como exemplo este registro: '00006200903201095', 2, onde:

000062009 - Número da licitação, indica que é a licitação nº 6 do ano de 2009

03 - Identifica o tipo de modalidade da licitação (Carta Convite)

201095 - Código da Unidade Gestora (no SAGRES este código corresponde a Prefeitura Municipal de João Pessoa-PB)

2 - Código do proponente na base de dados, foi omitido nº do CPF/CNPJ para preservar a identidade do participante.

## CAPÍTULO 6 – ANÁLISE DOS RESULTADOS

Após a realização do processamento dos dados descritos no capítulo anterior, nos deparamos com um conjunto de informações decorrentes do processamento pelo RapidMiner que necessitam de atenção especial para a sua interpretação e validação, esta etapa corresponde a fase de interpretação e avaliação dos padrões encontrados no processo KDD, conforme pode ser visualizada na Figura 6, constante do capítulo 2.

Os Tribunais de Contas fazem a fiscalização dos gastos públicos por exercício fiscal, compreendendo o período de 01 de janeiro até o dia 31 de dezembro de cada ano. Foram realizados experimentos por ano de competência e também abrangendo todo o período da pesquisa, ou seja, os anos de 2005 a 2009.

A pesquisa foi realizada com a utilização de alguns filtros para otimizar os recursos computacionais e para a obtenção de resultados satisfatórios. Nos processos licitatórios, dependendo do tipo de modalidade de licitação, poderá conter de 1 (um) a infinitos ( $\infty$ ) proponentes, neste caso, um filtro foi aplicado e selecionado os registros que continham de 2 (dois) a 10 (dez) proponentes por processo licitatório.

Os processos que continham mais de 10 proponentes foram desprezados pelo motivo de não representarem relevância para este estudo, além da possibilidade muito remota de ocorrer algum tipo de fraude com esta quantidade de proponentes. Já nos registros que continham apenas 1 (um) proponente, foi desprezada pelo filtro por ser considerada falta de competitividade por não ter outros interessados ou dependendo do tipo de modalidade de licitação, ter sido algum erro na informação enviada, o que facilmente poderá ser analisada pelos técnicos do Tribunal de Contas através de consultas simples ao banco de dados.

Selecionamos os tipos de modalidade de licitação Carta Convite e Tomada de Preços, identificadas pelos técnicos do TCE-PB como as que apresentam maior índice de fraudes e que correspondem juntas a 70,91% de todos os processos de licitatórios cadastrados na base de dados do SAGRES para o período compreendido por esta pesquisa.

No ano de 2005, por exemplo, a distribuição de proponentes nos processos licitatórios para as duas modalidades selecionadas (Carta Convite e Tomada de Preços) está representada no Quando 10. Observando a distribuição nos processos que contém

de 2 a 10 proponentes, em destaque no Quadro 11, juntos correspondem a 90,69% de todos os processos analisados neste primeiro experimento.

Quantidade de Participantes	Processos Licitatórios	%	% Total	Quantidade de Participantes	Processos Licitatórios	%	% Total
1	513	7,62%	<b>7,62%</b>	21	3	0,04%	<b>99,58%</b>
3	5566	82,70%	<b>90,33%</b>	22	1	0,01%	<b>99,60%</b>
4	301	4,47%	<b>94,80%</b>	23	6	0,09%	<b>99,69%</b>
5	92	1,37%	<b>96,17%</b>	24	2	0,03%	<b>99,72%</b>
6	53	0,79%	<b>96,95%</b>	25	2	0,03%	<b>99,75%</b>
7	26	0,39%	<b>97,34%</b>	26	2	0,03%	<b>99,78%</b>
8	21	0,31%	<b>97,65%</b>	27	1	0,01%	<b>99,79%</b>
9	24	0,36%	<b>98,01%</b>	28	1	0,01%	<b>99,81%</b>
10	20	0,30%	<b>98,31%</b>	30	1	0,01%	<b>99,82%</b>
11	16	0,24%	<b>98,54%</b>	31	1	0,01%	<b>99,84%</b>
12	10	0,15%	<b>98,69%</b>	32	1	0,01%	<b>99,85%</b>
13	8	0,12%	<b>98,81%</b>	33	1	0,01%	<b>99,87%</b>
14	15	0,22%	<b>99,03%</b>	35	2	0,03%	<b>99,90%</b>
15	10	0,15%	<b>99,18%</b>	36	1	0,01%	<b>99,91%</b>
16	11	0,16%	<b>99,35%</b>	37	1	0,01%	<b>99,93%</b>
17	7	0,10%	<b>99,45%</b>	47	1	0,01%	<b>99,94%</b>
19	2	0,03%	<b>99,48%</b>	49	2	0,03%	<b>99,97%</b>
20	4	0,06%	<b>99,54%</b>	55	1	0,01%	<b>99,99%</b>
21	3	0,04%	<b>99,58%</b>	71	1	0,01%	<b>100,00%</b>
22	1	0,01%	<b>99,60%</b>	<b>Total</b>	<b>6730</b>		

Quadro 11: Distribuição dos proponentes nos processos licitatórios correspondente ao ano de 2005

Após a realização do processamento pelo RapidMiner, o resultado com a geração de dados consolidados está mostrado na Figura 19, o campo CPF/CNPJ do proponente foi substituída por um código sequencial numérico para facilitar no processamento e preservar a sua identidade.

No experimento foi usado o suporte mínimo de 0.005 (0,05%) e confiança mínima de 0.60 (60%). Com estes parâmetros foram confirmadas 48 regras de associação com um total de 113 conjuntos de itens frequentes encontrados. O suporte e a confiança atuam como medidas de interesse no processo de mineração de regras de associação. Para esta aplicação um suporte mínimo alto não é relevante porque senão estaríamos dizendo que a maioria das licitações possui algum tipo de fraude, o que não é verdade, por este motivo, foi escolhido um suporte mínimo muito baixo devido às características dos dados estudados.



O suporte corresponde à frequência com que  $A$  e  $B$  ocorrem em uma base de dados, e a confiança representa a frequência com que  $B$  ocorre, dentre os eventos que contém  $A$ .

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	27577.0	20445.0, 16560.0	0.005	0.604	0.997	-0.012	0.005	118.698	2.511
2	23132.0	25179.0	0.009	0.607	0.995	-0.020	0.008	37.792	2.502
3	13689.0	14372.0	0.007	0.609	0.996	-0.015	0.007	62.775	2.531
4	7198.0, 27960.0	16979.0	0.005	0.611	0.997	-0.012	0.005	103.905	2.556
5	27577.0	16560.0	0.005	0.623	0.997	-0.012	0.005	97.926	2.633
6	16795.0	21917.0	0.007	0.641	0.996	-0.014	0.006	33.867	2.730
7	6933.0	21154.0	0.010	0.667	0.995	-0.019	0.009	58.250	2.966
8	23735.0	24162.0	0.005	0.681	0.998	-0.010	0.005	57.110	3.096
9	13154.0, 21225.0	21781.0	0.005	0.681	0.998	-0.010	0.005	77.877	3.106
10	14372.0	13689.0	0.007	0.689	0.997	-0.013	0.007	62.775	3.175
11	24357.0	10574.0	0.008	0.700	0.997	-0.014	0.008	47.352	3.284
12	21638.0	10574.0	0.008	0.707	0.997	-0.015	0.008	47.803	3.359
13	21781.0	21225.0	0.006	0.709	0.997	-0.011	0.006	54.401	3.393
14	14648.0	8356.0	0.006	0.709	0.997	-0.011	0.006	91.039	3.411
15	20445.0	16560.0	0.005	0.711	0.998	-0.009	0.005	111.840	3.440
16	20445.0	27577.0, 16560.0	0.005	0.711	0.998	-0.009	0.005	135.564	3.443
17	16991.0	13732.0	0.006	0.740	0.998	-0.010	0.006	103.452	3.819
18	13154.0, 21781.0	21225.0	0.005	0.744	0.998	-0.009	0.005	57.094	3.858
19	14454.0	18741.0	0.006	0.745	0.998	-0.009	0.005	53.236	3.862

Figura 19: Resultado do processamento dos dados minerados pelo RapidMiner

Diante das regras geradas pelo algoritmo de associação, na Figura 20 são apresentadas as regras que possuem um indicador de confiança elevado. O Apêndice II traz a relação com todas as regras geradas neste experimento.

Association Rules
...
[16560.0] --> [27577.0] (confidence: 0.825)
[21154.0] --> [6933.0] (confidence: 0.833)
[23157.0, 13689.0] --> [14372.0] (confidence: 0.846)
[6030.0] --> [15729.0] (confidence: 0.860)
[27545.0] --> [13979.0] (confidence: 0.864)
[20445.0] --> [27577.0] (confidence: 0.889)
[16979.0] --> [7198.0, 27960.0] (confidence: 0.892)
[16979.0] --> [7198.0] (confidence: 0.919)
[16979.0] --> [27960.0] (confidence: 0.919)
[19639.0] --> [21917.0] (confidence: 0.925)
[27577.0, 16560.0] --> [20445.0] (confidence: 0.970)
[7198.0, 16979.0] --> [27960.0] (confidence: 0.971)
[27960.0, 16979.0] --> [7198.0] (confidence: 0.971)
[23157.0, 14372.0] --> [13689.0] (confidence: 1.000)
[20445.0, 16560.0] --> [27577.0] (confidence: 1.000)

Figura 20: Regras de associação do exercício de 2005 geradas pelo RapidMiner

A Figura 20 mostra as regras confirmadas, ou seja, o RapidMiner confirma a associação entre concorrentes, ressaltando a co-ocorrência de proponentes nos dados

minerados. Por exemplo: o concorrente *16560* sempre está associado ao proponente *27577* em 82,5% dos casos. O concorrente *21154* aparece em 83,3% das vezes que o proponente *6933* aparece e assim sucessivamente.

Analisando estas regras, identificamos a ocorrência de vários *insight* que permitem aos decisores uma compreensão mais aprofundada do problema na mão, estas informações são então transformadas em conhecimento.

Segundo Gonçalves (2005), existem dois tipos de medidas de interesse: objetivas e subjetivas. As medidas de interesse objetivas empregam índices estatísticos para avaliar a força de uma regra. Já as medidas de interesse subjetivas consideram principalmente a opinião de um analista para determinar a força da regra. Segundo Bürkle (2006), o suporte é interpretado como uma medida da significância estatística da regra, enquanto que a confiança pode ser interpretada como uma medida da força da regra.

Os valores do suporte e a confiança são exemplos de medidas de interesse objetivas que devem ser analisadas para apoiar a tomada de decisão. Vejamos o Quadro 12, onde estão listadas as regras com maior índice de confiança com os dados processados pelo RapidMiner.

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
48	20445 e 16560	27577	0,51%	100,00%	118.698	Infinito
47	23157 e 14372	13689	0,52%	100,00%	91.174	Infinito
46	27960 e 16979	7198	0,52%	97,06%	84.805	33.611
45	7198 e 16979	27960	0,52%	97,06%	91.134	33.638
44	27577 e 16560	20445	0,51%	96,97%	135.564	32.764
43	19639	21917	0,59%	92,50%	48.901	13.081
42	16979	27960	0,54%	91,89%	86.282	12.202
41	16979	7198.0	0,54%	91,89%	80.291	12.192
40	16979	7198 e 27960	0,52%	89,19%	103.905	9.171
39	20445	27577	0,64%	88,89%	105.509	8.924
38	27545	13979	0,60%	86,36%	110.880	7.276
37	6030	15729	0,59%	86,05%	112.775	7.112
36	23157 e 13689	14372	0,52%	84,62%	87.265	6.437
35	21154	6933	0,95%	83,33%	58.250	5.914

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
34	16560	27577	0,52%	82,50%	97.926	5.666

Quadro 12: Medidas de interesse objetivas geradas no exercício de 2005

Interpretando as regras extraídas no Quadro 12, podemos afirmar que, quando o proponente 20445 e o proponente 16560 aparecem juntos em uma licitação, podemos afirmar que em 100% dos casos o proponente 27577 aparece também. E quando o proponente 19630 aparece em uma licitação, em 92,50% dos casos o proponente 21917 também aparece e assim sucessivamente.

Podemos ainda analisar apenas as regras em que um proponente faz parte da conclusão da mesma, independente se aparece nas premissas da regra. Foi selecionado como exemplo o proponente 27577, conforme visualizado no Quadro 13.

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
16	20445	27577 e 16560	0,51%	71,11%	135.564	3.443
30	16560	27577 e 20445	0,51%	80,00%	125.820	4.968
34	16560	27577	0,52%	82,50%	97.926	5.666
39	20445	27577	0,64%	88,89%	105.509	8.924
48	20445 e 16560	27577	0,51%	100,00%	118.698	Infinito

Quadro 13: Filtro realizado por proponente nas regras de associação geradas

A Figura 21 mostra estas regras geradas no Quadro 13 visualizadas no formato de grafo, desta forma ficam bem evidentes os relacionamentos, e consequentemente facilita a análise das correlações.

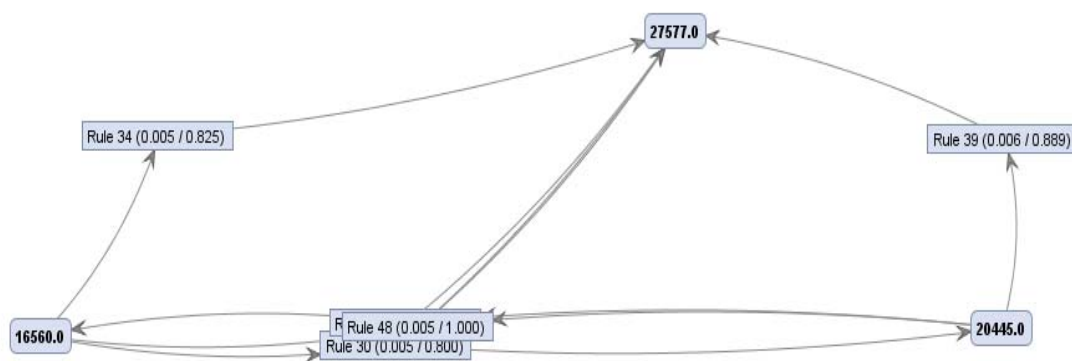


Figura 21: Grafo gerado pelo RapidMiner com as regras de associação de um proponente

Outra medida de interesse importante a ser considerada é o *Lift*. Dada uma regra de associação  $A \Rightarrow B$ , esta medida indica o quanto mais frequente torna-se  $B$  quando  $A$  ocorrer. Esta medida é útil para regras com suporte baixo (como é este caso) e

representa a dependência positiva e negativa entre o predecessor e sucessor da regra. Se o valor foi igual a zero isto significa que não existe dependência. Se o valor for negativo significa que existe dependência negativa. Se o valor for positivo significa que existe dependência positiva.

Observe que para a maioria das regras geradas a partir do experimento, o *Lift* tem um valor muito alto. Isto significa que todos os proponentes relacionados na regra tem uma dependência grande.

A variável *Convicção* (*conviction*) tem como objetivo avaliar uma regra de associação como uma verdadeira implicação. Quando um valor muito alto aparece significa que  $A \Rightarrow B$  é muito forte, mas  $B \Rightarrow A$  não necessariamente será.

A medida de *convicção* varia de 0 a infinito ( $\infty$ ), apresentando valor 1 quando os conjuntos  $A$  e  $B$  são independentes, e infinito quando o valor da confiança for igual a 100%.

Tanto o valor do *Lift* quanto da *Convicção* indicam a forte relação entre as regras geradas, isso indica que as regras são consistentes e que podem representar segurança para quem está analisando.

Ou seja, podemos resumir estes conceitos conforme apresentado no Quadro 14:

	<b>Suporte</b>	<b>Confiança</b>	<b><i>Lift</i></b>	<b>Convicção</b>
<b>Definição</b>	O suporte de um conjunto de itens $Z$ , $Sup(Z)$ , representa a porcentagem de transações da base de dados que contém os itens de $Z$ .	A confiança da regra, $Conf(A \Rightarrow B)$ , representa, dentre as transações que contém $A$ , a porcentagem de transações que também contém $B$ .	Utilizado para verificação do índice de dependência do consequente em relação ao seu antecedente.	Utilizada para indicar o grau de implicação do antecedente no consequente.

Quadro 14: Definição dos principais indicadores utilizados para análise dos dados minerados

Os experimentos realizados nos demais anos compreendidos por esta pesquisa são descritos a seguir, nos Quadros 15, 16, 17, e 18. No Quadro 15, são exibidas somente 12 regras das 27 geradas.

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
16	16979	6724	0,53%	87,18%	156.245	7.756
17	16979	27960 e 6724	0,53%	87,18%	156.245	7.756
18	27960 e 16979	6724	0,53%	87,18%	156.245	7.756
19	21714 e 21917	16795	0,71%	92,00%	65.229	12.323
20	23814	18211	0,54%	92,11%	156.385	12.592
21	18211	23814	0,54%	92,11%	156.385	12.592
22	6724	16979	0,53%	94,44%	156.245	17.891
23	6724	27960 e 16979	0,53%	94,44%	156.245	17.891
24	27960 e 6724	16979	0,53%	94,44%	156.245	17.891
25	16979	27960	0,60%	100,00%	99.261	Infinito
26	6724	27960	0,56%	100,00%	99.261	Infinito
27	16979 e 6724	27960	0,53%	100,00%	99.261	Infinito

Quadro 15: Medidas de interesse objetivas no exercício de 2006

Com os valores informados nos experimentos anteriores, para o ano de 2007 só foram geradas 5 regras de associação para os 82 itens frequentes encontrados, conforme Quadro 16. Para a geração e exibição de mais regras com os itens encontrados faz-se necessário que o valor de suporte e da confiança sejam alterados no aplicativo de mineração de dados utilizado.

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
1	1197	25541	1,01%	62,14%	35.529	2.594
2	6100	27467	0,68%	64,18%	56.575	2.759
3	27545	13979	0,50%	69,57%	137.978	3.269
4	6030	15729	0,52%	80,49%	81.088	5.074
5	13979	27545	0,50%	100,00%	137.978	Infinito

Quadro 16: Medidas de interesse objetivas no exercício de 2007

No Quadro 17 são exibidas todas as 11 regras geradas para os 83 itens frequentes encontrados para o exercício de 2008.

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
1	19786	21316	0,53%	62,50%	57.036	2.637
2	15729	17316	0,69%	65,00%	85.527	2.835
3	27577	24076	0,58%	73,33%	118.548	3.726
4	13979	2673	0,51%	78,38%	147.821	4.600
5	13979	27545	0,55%	83,78%	135.442	6.128
6	27545	13979	0,55%	88,57%	135.442	8.692
7	14307	12302	0,51%	90,63%	170.918	10.610
8	17316	15729	0,69%	90,70%	85.527	10.636
9	24076	27577	0,58%	94,29%	118.548	17.360
10	2673	13979	0,51%	96,67%	147.821	29.803
11	12302	14307	0,51%	96,67%	170.918	29.830

Quadro 17: Medidas de interesse objetivas no exercício de 2008

Observe que foram geradas no Quadro 18 somente 2 regras para o exercício de 2009 com o mesmo valor do suporte e confiança informados nos experimentos anteriores.

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
1	4373	14538	0,58%	64,41%	31.367	2.752
2	3418	14538	0,69%	75,00%	36.526	3.918

Quadro 18: Medidas de interesse objetivas no exercício de 2009

Observem que a relação entre os proponentes diminuiu ao longo dos anos quando foram mantidos os mesmos parâmetros informados nesta pesquisa (Confiança 60% e Suporte de 0,005%), este fato pode ser:

- 1) consequência de um cuidado maior pelas empresas para a manipulação e direcionamento de licitações, ou;
- 2) maior transparência na publicação de editais para os processos licitatórios, e consequentemente, um aumento do número de empresas interessadas em prestar serviços a órgãos públicos.

Na busca de novos resultados foram realizados outros experimentos com os dados do exercício de 2009. Quando alteramos o valor do suporte mínimo para 0,001% e o valor da confiança permanecendo em 60%, a quantidade de itens frequentes passou de 52 para 1061, confirmando um conjunto de 741 regras de associação, conforme visualizado na Figura 22.

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
469	24093.0, 27369.0	12208.0	0.001	1	1	-0.001	0.001	466.143	∞
470	12208.0, 27369.0	24093.0	0.001	1	1	-0.001	0.001	435.067	∞
471	9265.0, 22302.0	11575.0	0.002	1	1	-0.002	0.002	543.833	∞
472	9265.0, 11575.0	22302.0	0.002	1	1	-0.002	0.002	543.833	∞
473	13037.0, 20314.0	19921.0	0.001	1	1	-0.001	0.001	502	∞
474	4072.0, 5019.0	15159.0	0.001	1	1	-0.001	0.001	725.111	∞
475	4072.0, 15159.0	5019.0	0.001	1	1	-0.001	0.001	725.111	∞
476	5019.0, 15159.0	4072.0	0.001	1	1	-0.001	0.001	593.273	∞
477	15688.0	10368.0, 12628.0	0.002	1	1	-0.002	0.002	652.600	∞
478	10368.0, 15688.0	12628.0	0.002	1	1	-0.002	0.002	652.600	∞
479	12628.0	10368.0, 15688.0	0.002	1	1	-0.002	0.002	652.600	∞
480	10368.0, 12628.0	15688.0	0.002	1	1	-0.002	0.002	652.600	∞
481	15688.0, 12628.0	10368.0	0.002	1	1	-0.002	0.002	593.273	∞
482	23367.0, 24276.0	16009.0	0.001	1	1	-0.001	0.001	725.111	∞
483	24276.0, 16009.0	23367.0	0.001	1	1	-0.001	0.001	652.600	∞
484	17400.0	733.0, 16767.0	0.001	1	1	-0.001	0.001	725.111	∞
485	733.0, 17400.0	16767.0	0.001	1	1	-0.001	0.001	725.111	∞
486	16767.0, 17400.0	733.0	0.001	1	1	-0.001	0.001	725.111	∞
487	8382.0	6246.0, 2258.0	0.001	1	1	-0.001	0.001	932.286	∞
488	6246.0, 8382.0	2258.0	0.001	1	1	-0.001	0.001	932.286	∞
489	2258.0	6246.0, 8382.0	0.001	1	1	-0.001	0.001	932.286	∞
490	6246.0, 2258.0	8382.0	0.001	1	1	-0.001	0.001	932.286	∞
491	8382.0, 2258.0	6246.0	0.001	1	1	-0.001	0.001	815.750	∞

Figura 22: Conjunto de dados encontrados pelo RapidMiner com os alterações nos parâmetros de filtro

Quando listamos e analisamos este novo conjunto com 491 regras, identificamos a presença do proponente 16979 participando de várias licitações, conforme visualizado no Quadro 19. Este mesmo proponente ao ser consultado na base de dados foi identificada a presença de correlações desde o ano de 2005 nas regras de associação já confirmadas pelo *software* de mineração de dados.

Regra	Premissas	Conclusões	Suporte	Confiança	Lift	Convicção
137	6724	16979	0,15%	71,43%	186.457	3.486
216	27960 e 6563	16979	0,21%	77,78%	203.031	4.482
248	6563	16979	0,32%	89,78%	210.840	5.180

Quadro 19: Filtro realizado por proponente nas regras de associação geradas em 2009

O estudo realizado neste trabalho de dissertação corresponde a uma pequena parte das possibilidades oriundas do uso adequado das técnicas de *data mining*, existindo uma gama de possibilidades para a sua utilização e aplicabilidade.

As regras que descritas foram validadas junto ao aplicativo SAGRES e mostraram-se coerentes nos testes realizados. Isso não significa que existam irregularidades em todos os processos licitatórios em que ocorreu a co-ocorrência ou co-participação dos proponentes listados nas regras geradas e exibidas nos quadros acima, mas proporciona uma nova visão aos técnicos que fazem a análise dos processos licitatórios, de forma que procedam uma investigação minuciosa afim de constatarem a veracidade destes fatos relatados.

## 6.1 Ganhos esperados com a implantação de técnicas de mineração de dados no Tribunal de Contas do Estado

Com a implantação desta técnica de trabalho e dos algoritmos de mineração de dados aplicado ao contexto de competência de controle realizado pelo TCE-PB, espera-se um ganho de produtividade e celeridade no tramite processual decorrentes da análise das contas pelos setores responsáveis pela fiscalização das unidades gestores sob sua jurisdição.

Outro ganho esperado é a possibilidade da realização de novos cruzamentos com os dados armazenados no banco de dados com o uso de outros algoritmos na busca de conhecimentos escondidos nestes dados, que poderá trazer novas visões e validar outras já encontradas.

## CAPÍTULO 7 – CONSIDERAÇÕES FINAIS

O processo de descoberta de conhecimento em bases de dados (KDD) envolve uma série de etapas, envolvendo desde a coleta, preparação e limpeza dos dados, consolidação e extração dos padrões, seguindo até a avaliação do quanto estes padrões e/ou regras descobertas agregam valor ao que o gestor já conhecia sobre o problema em questão.

O presente estudo permitiu compreender que o recurso da tecnologia da informação é essencial no apoio à fiscalização do erário, principalmente para as organizações que atuam em um ambiente dinâmico, exigindo constante monitoramento de suas estruturas internas, e para satisfazer aos anseios da sociedade, quanto à correta aplicação destes recursos, garantindo um transparente, eficiente e eficaz sistema de fiscalização da gestão pública.

No entanto, a informação gerada pelos sistemas de controle e de banco de dados por si só pode não ser adequada para determinada necessidade, sendo por isso oportuno atentar para quais informações complementares seriam necessárias para apoiar uma tomada de decisão.

O monitoramento, fiscalização e análise dos dados pelos Tribunais de Contas, em todas as esferas, funcionam de forma a centralizar as informações pertinentes à sua atividade de auditar as contas públicas e emitir parecer sobre os fatos analisados em consonância com a legislação vigente, além do provimento de informações consistentes e tempestivas ao processo decisório, apoiada em sistemas de informações e no controle organizacional.

Nesse âmbito, o presente estudo teve como propósito geral investigar a contribuição da aplicação das técnicas de *data mining* para a área de auditoria, objetivando exercer sua função de fiscalização com vistas a identificar e relacionar possíveis indícios de fraudes em processos licitatórios. Para atendê-la foram traçados os seguintes objetivos específicos: explorar os conceitos de técnicas de *data mining*; realizar uma investigação através de uma pesquisa exploratória com vistas a conhecer os fatos e fenômenos relacionados à implantação das técnicas de *data mining* para a organização; e identificar e analisar a aplicabilidade das técnicas de *data mining* para o processo de fiscalização nos órgãos de controle externo, com observação especial na



aplicação desta sistemática no âmbito de competência do Tribunal de Contas do Estado da Paraíba.

Procurou-se atingir a esses objetivos, geral e específicos, por meio de pesquisa bibliográfica e trabalho de campo, voltados a responder ao seguinte questionamento: a implantação de uma ferramenta de mineração de dados pode contribuir na identificação de fraudes em processos licitatórios analisados pelo TCE-PB?

As conclusões decorrentes do trabalho desenvolvido, e a resposta para a questão formulada acima, são apresentadas a seguir, devidamente norteadas pelos objetivos geral e específicos propostos.

Considerando o objetivo de explorar os conceitos de técnicas de *data mining*, este estudo buscou na literatura conceitos sobre o tema em autores clássicos e pesquisadores conceituados, com vistas a dar uma visão geral sobre *Business Intelligence* e seus principais componentes, com ênfase para os algoritmos e procedimentos usados para a mineração de dados.

Em relação ao item realizar uma investigação através de uma pesquisa exploratória com vistas a conhecer os fatos e fenômenos relacionados à implantação das técnicas de *data mining* para a organização, este estudo mostrou que o uso de tais técnicas torna viável a aquisição de programas para esta finalidade, além de investimentos em capacitação de servidores para uso adequado desses programas computadorizados.

Em relação ao item identificar e analisar a aplicabilidade das técnicas de *data mining* para o processo de fiscalização nos órgãos de controle externo, com observação especial na aplicação desta sistemática no âmbito de competência do Tribunal de Contas do Estado da Paraíba, pôde-se observar que a implantação de tal metodologia trouxe a tona uma gama de informações e correlações que antes só eram encontradas após um grande esforço em para tentar obtê-las, o que tende a proporcionar um ganho considerável de tempo na análise de processos licitatórios e na elaboração de pareceres técnicos, além da possibilidade de ser estendido este estudo para as demais áreas de trabalho do Tribunal de Contas, servindo inclusive de referencial para outros órgãos que desempenham as funções de controle e fiscalização semelhantes.

As regras de associação constituem uma técnica muito popular da mineração de dados. Embora muitas vezes apresentem uma grande quantidade de associações inúteis,

o que somente prejudica o trabalho dos analistas do negócio que tentam encontrar oportunidades e/ou correlações consistentes nos dados armazenados, mesmo assim, seus benefícios são superiores e as possibilidades de uso infinitas.

Finalmente, dadas às evidências obtidas nessa pesquisa, conclui-se que a utilização adequada de técnicas de mineração de dados, pode auxiliar na análise das contas públicas, especialmente na sua função de prover informações confiáveis, úteis e tempestivas requeridas pelo processo decisório, tornando-se mais um instrumento confiável e com benefícios mensuráveis para a realização do exercício constitucional delegado aos Tribunais de Contas.

### 7.1 Recomendações para trabalhos futuros

Esta pesquisa é um marco teórico para o Tribunal de Contas do Estado da Paraíba, e a partir dela haverá uma validação mais ampla destas técnicas com a ampliação das possibilidades de uso e perspectivas para novas pesquisas para as quais se apresentam indicações para trabalhos futuros.

A presente pesquisa certamente não esgotou o assunto sobre a contribuição das ferramentas para mineração de dados (*data mining*), objetivando o controle eficiente dos recursos públicos e para coibir práticas de eventos fraudulentos nos processos licitatórios.

Estudos futuros poderão apresentar continuidade, efetuar outros tipos de testes complementares ou reforçar os resultados encontrados aqui.

Para trabalhos futuros sugere-se:

- a análise da aplicabilidade das técnicas de *data mining* em outras áreas de competência do Tribunal de Contas, como a análise de obras e serviços de engenharia;
- estudos de caso múltiplos, ou seja, testar em mais de um Tribunal de Contas se o uso das ferramentas de mineração de dados contribui positivamente e igualmente para cada um deles;
- estudos sobre as limitações que o modelo de gestão dos Tribunais de Contas pode impor ao uso das ferramentas de *Business Intelligence*, especificamente das ferramentas de *data mining*.

## REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R.; IMIELINSKI, T.; SRIKANT, R. *Mining Association Rules between Sets of Items in Large Databases*. Proc. of the ACM SIGMOD Int'l Conference on Management of Data, Washington D.C., 207-216. May, 1993

ANDRADE, Maria Margarida de. *Introdução ao Método científico: elaboração de trabalhos na graduação*. 9. ed. – São Paulo: Atlas, 2009.

AFFELDT, Fabrício Sobrosa; VANTI, Adolfo Alberto; RAUTER, André. *Proposta de configuração informacional, com a utilização de business intelligence (BI), com base no alinhamento estratégico entre TI e gestão de negócios*. XII SIMPEP - Bauru, SP, 2005.

ALESANDRINI, K. *Survive Information Overload*. Homewood, IL: Business One Irwin, 1992.

ARAÚJO JÚNIOR, Rogério Henrique de; ALVARES, Lillian. *Gerenciamento estratégico da informação: a convergência a partir da Sociedade da Informação*. Parcerias Estratégicas, Vol. 12, nº 25, 2007.

BARBIERI, Carlos. *BI – Business Intelligence: modelagem e tecnologia*. 2ª ed. Rio de Janeiro: Axcel Books do Brasil Editora, 2001.

BARBOSA, G. R. *Sistemas de Apoio a Decisão sob o enfoque de Profissionais de Tecnologia da Informação e Decisores*. Dissertação de Mestrado, PPGE, UFPE, 2003.

BARCELOS TRONTO, I. F. ; ARAUJO, A. C. ; SIMOES, J. D. S.; SANT'ANNA, N. *Business Intelligence: Inteligência nos Negócios*. In: III Workshop dos Cursos de Computação Aplicada do INPE, 2003, São José dos Campos. Anais do III Worcap-Workshop dos Cursos de Computação Aplicada do INPE, 2003. v. 3. p. 187-192.

BATISTA, Emerson de Oliveira. *Sistema de Informação: o uso consciente da tecnologia para o gerenciamento*. São Paulo: Saraiva, 2006.

BRAZ, Lucas M.; FERREIRA, Rafael; Dermeval, Diego; VÉRAS, Douglas; LIMA, Marcilio; TIENGO, Willy. *Aplicando Mineração de Dados para Apoiar a Tomada de Decisão na Segurança Pública do Estado de Alagoas*. XXIX Congresso da Sociedade Brasileira de Computação. Bento Gonçalves - RS, 2009.

BELLOQUIM, A. *Porque o software não faz o que eu pedi?* Developers, n. 35, ano 3, p. 54, jul, 1999.

BEUREN, Ilse Maria; GIBBON, Artur Roberto de Oliveira. *Considerações acerca do sistema de informação gerencial como suporte ao controle de gestão: uma abordagem da gestão do conhecimento*. XXI Encontro Nacional de Engenharia de Produção. Salvador, Bahia, 2001.

BRASIL. Constituição de 1988. Constituição da República Federativa do Brasil. Disponível em: <<http://www.planalto.gov.br>>. Acesso em: 15 nov. 2009.

\_\_\_\_\_. Lei Federal nº 8.666, 1993, Lei de LICITAÇÕES. Brasília: Congresso Nacional. Disponível em: <[http://www.planalto.gov.br/ccivil\\_03/Leis/L8666cons.htm](http://www.planalto.gov.br/ccivil_03/Leis/L8666cons.htm)>. Acesso em: 10 mar. 2010.

BÜRKLE, Paula Yamada. Um método de pós-processamento de Regras de Associação com base nas relações de dependência de atributos. Dissertação (Mestrado) – Universidade Federal Fluminense. Niterói/RJ. 2006.

CALDAS, Marcos Paulo Kohler; SCANDELARI, Luciano; KOVALESKI, João Luiz. Aplicações sobre uma *Data Warehouse* no ambiente das organizações e suas vantagens. XIII SIMPEP - Bauru, SP, 2006.

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. Gestão do conhecimento usando *data mining*: estudo de caso na Universidade Federal de Lavras. Rev. Adm. Pública v.42 n.3 Rio de Janeiro maio/jun. 2008

CARVALHO, Luís Alfredo Vidal de. *Datamining* – A mineração de dados no marketing, medicina, economia engenharia e administração. Rio de Janeiro: Editora Ciência Moderna, 2005.

CARVALHO, José Tanajura. Acessibilidade às informações do controle externo: um instrumento para o controle social do estado. Revista do Tribunal de Contas do Estado de Minas Gerais. Ed nº 01 de 2002 - Ano XX.

CLERICUZI, Adriana Zenaide. Desenvolvimento adaptativo para Sistema de Apoio a Decisão específico. 175 f. Tese (Doutorado) – Universidade Federal de Pernambuco, Recife, 2006.

COELHO, Fabrício Luís. Classificação semi-automática de monografias. Novo Hamburgo, RS: 2008. 71p. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Feevale, 2008.

COSTA, Rodrigo Bernardes Ribeiro da. Aplicação do Processo de Mineração de Dados para Auxílio à Gestão do Pronto-Socorro de Clínica Médica do Hospital Universitário de Brasília. 82 f. Monografia. Brasília: UnB, 2007.

CRIVELINI, Wagner. Mitos e verdades sobre modelagem multidimensional. SQL Magazine. Rio de Janeiro, ano 3, 39ª ed., p.6-14, 2006.

DECOMAIN, Pedro Roberto; TRIBUNAIS DE CONTAS NO BRASIL: natureza jurídica e alcance das suas funções, à luz dos princípios constitucionais da Administração Pública. 397 f. Dissertação (Mestrado). Universidade do Vale do Itajaí – UNIVALI. Itajaí. 2005.

DIAS, Fernando Sckackauskas. Avaliação de sistemas de informação: revisão de publicações científicas no período de 1985-2005. Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação. Belo Horizonte. 2006.

DUTRA, Rogério Garcia. Aplicação de Métodos de Inteligência Artificial em Inteligência de Negócios. XXV ENEGEP. Porto Alegre - RS, 2005.

FACHINELLI, Ana Cristina; RECH, Jane; MATTIA, Olivar Maximino. Organizações Inteligentes: da gestão da informação para as organizações que aprendem. III CONVIBRA (Congresso Virtual Brasileiro de Administração), 2006.

FAVARETTO, Fábio. Melhoria da qualidade da informação no controle da produção: estudo exploratório utilizando Data Warehouse. Revista Produção. Vol. 17, n. 2, p. 343-353, Maio/Ago, 2007.

FAYYAD, U.; SHAPIRO, G. P.; SMYTH P. *From data mining to knowledge discovery in databases*. AI Magazine, 17(3): 37-54, Fall 1996.

FIGUEIRÊDO, Carlos Maurício C. Ética na gestão pública e exercício da cidadania: o papel dos tribunais de contas brasileiros como agências de *accountability* - O caso do Tribunal de Contas de Pernambuco. VII *Congreso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública*, Lisboa, Portugal, 8-11 Oct. 2002

FILHO, Trajano Leme. B.I. – *Businnes Intelligence* no Excel. Rio de Janeiro. Editora Novaterra, 2010.

FOLDOC. *Free On-Line Dictionary Of Computing*. Disponível em: <<http://foldoc.org>>. Acesso em: 02/02/2010.

FREITAS JR., O. G. Um modelo de Sistema de Gestão do Conhecimento para Grupos de Pesquisa e Desenvolvimento. 296 f. Tese (Doutorado em Engenharia) – Universidade Federal de Santa Catarina, Florianópolis, 2003.

GIL, A. C. Como elaborar projetos de pesquisa. 3. ed. São Paulo: Atlas, 1991.

GILLMEISTER, Paulo Ricardo Guglieri; CAZELLA, Sílvio César. Uma análise comparativa de algoritmos de regras de associação: minerando dados da indústria automotiva. Escola Regional de Banco de Dados (ERBD). Universidade de Caxias do Sul (UCS), Caxias do Sul, RS. 2007.

GILSTER, Paul. *Digital Literacy*. New York: Wiley, 1997.

GIUDICI, Paolo. *Applied data mining: statistical methods for business and industry*. John Wiley & Sons Ltd. 2003.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. *Data mining: um guia prático*. Rio de Janeiro: Elsevier, 2005, 3ª reimpressão.

GONÇALVES, E. C. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. INFOCOMP (UFLA), v. 4, p. 26-35, 2005.

GOUVEIA, Luis Borges. Gestão da Informação: competências críticas para a Sociedade da Informação e do conhecimento. 2002. Disponível em:

<[http://www2.ufp.pt/~lmbg/formacao/msc\\_competencias\\_book.pdf](http://www2.ufp.pt/~lmbg/formacao/msc_competencias_book.pdf)>. Acesso em 01/02/2010.

GRILO JÚNIOR, Tarcísio Ferreira, XAVIER, G. B., CLAUDINO NETO, Severino, CLERICUZI, A. Z. Proposta de um framework aplicando técnicas de data mining para o controle externo no Tribunal de Contas do Estado da Paraíba In: 6º CONTECSI - Congresso Internacional de Gestão da Tecnologia e Sistemas de Informação, 2009, São Paulo - SP.

GRILO JR., Tarcísio Ferreira; CLERICUZI, Adriana Zenaide. Otimização do processo produtivo e de tomada de decisão através de *Business Intelligence* aplicado ao controle externo. III SEPRONe – Juazeiro, BA. 2008.

GRILO JR., Tarcísio Ferreira; SILVA, José Gentil Bezerra da; CLERICUZI, A. Z. . Business Intelligence: uma abordagem com ênfase no controle externo. XV Simpósio de Engenharia de Produção - SIMPEP. Baurú/SP, 2008.

HAMMOND, J. S.; KEENEY, R. L.; RAIFFA, H. Decisões Inteligentes. Ed. Campus, São Paulo, 2004.

HAN, Jiawei; KAMBER, Micheline. *Data mining: concepts and techniques*. 2ª ed. San Francisco: Morgan Kaufmann Publishers, 2006.

INMON, William H. Como Construir o *Data Warehouse*. 2ª. ed., Editora Campus, Brasil, 1997.

LAKATOS, E. M.; MARCONI, M. A. Fundamentos de metodologia científica. São Paulo: Atlas, 2007.

LEITE, Fabiano Luiz Caldas. Utilização de Business Intelligence para Gestão da Área Operacional de Agências Bancárias: um estudo de caso. 167 f. Dissertação (MPA) - Escola de Administração de Empresas de São Paulo. São Paulo. 2007.

LAUDON, Kenneth. LAUDON Jane – Sistemas de informação gerenciais. 7ª. edição São Paulo: Pearson Prentice Hall, 2007.

LOPES, Mauricio. C.; OLIVEIRA, Pércio A. . Ferramenta de Construção de Data Warehouse. XVI Seminc, Blumenau (SC), 2007.

NOGUEIRA, Marcelo; VENDRAMETTO, Oduvaldo; ABE, Jair Minoru. Qual a importância da adoção do *business intelligence* para obtenção de vantagem competitiva? XII SIMPEP - Bauru, SP. 2005.

MACHADO, Bruno de Abreu; JAMIL, George Leal. Mineração de dados como ferramenta de tomada de decisão: estudo de caso de uma universidade. 5º Congresso Internacional de Gestão da Tecnologia e Sistema de Informação - CONTECSI. São Paulo. 2008.

MARCHIORATO, Haroldo José. Percepção de gestores de TI/SI e executivos sobre o papel da TI/SI e sua utilização estratégica: um estudo exploratório em indústrias de autopeças. 114 f. Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia de Produção. Recife. 2006.

MELO, Renato Sérgio Santiago. Tribunais de Contas: uma teoria geral. João Pessoa: A União, 2005.

MYLIUS, Marcos. *Business Intelligence*: mais fácil do que você imagina. São Paulo: Edições Inteligentes Editora, 2003.

MORAES, Tiago Cacique. O processo de modernização dos Tribunais de Contas no contexto da reforma do Estado no Brasil. 109 f. Dissertação (mestrado) – Escola de Administração de Empresas de São Paulo. 2006.

MURAYAMA, Antonio Carlos. Técnicas Gerenciais Aplicadas em Medição de Desempenho e Gestão Estratégica nas Organizações. Dissertação (mestrado) – Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT. 2002.

NETO, Jayme Baleeiro. As organizações e o controle dos Tribunais de Contas. Rio de Janeiro: Renovar, 2004.

O'BRIEN, James A. Sistemas de informação e as decisões gerenciais na era da internet. 2ª ed., São Paulo: Saraiva, 2004.

PADILHA, THAIS CÁSSIA CABRAL; MARINS, FERNANDO AUGUSTO SILVA. Sistemas ERP: características, custos e tendência. Revista Produção. Vol. 15, n. 1, p. 102-113, Jan./Abr. 2005.

PETRINI, Maira. Incorporando a Gestão da Sustentabilidade aos Sistemas de Inteligência de Negócio. 157 f. Tese (doutorado) - Escola de Administração de Empresas de São Paulo. 2006.

PINTO, Cesar. BI – *Business Intelligence* ou Inteligência nos Negócios. Disponível em < <http://www.mainretail.com.br/artigos/business-intelligence.htm> >. Acesso em: 13 set. 2010.

PINTO, Claudia Cristini Crisp Porto; SANTOS, Vilma da Silva & OLIVEIRA, Edson Aparecida de Araújo Querido. A ferramenta *Business Intelligence* como estratégia de Tecnologia da Informação: um estudo de caso. 5º Congresso Internacional de Gestão da Tecnologia e Sistema de Informação - CONTECSI. São Paulo. 2008.

PIZZI, Luciene Cristina. Mineração multi-relacional: o algoritmo GFP-*growth*. 107 p. Dissertação (Mestrado) - Universidade Federal de São Carlos. São Carlos, 2006.

PROJETO TAMANDUÁ. Disponível em: <<http://tamandua.speed.dcc.ufmg.br>>. Acesso em: 09 mar. 2010.

RAPIDMINER. Disponível em: <<http://rapid-i.com>>. Acesso em: 15 mar. 2010.

ROMÃO, Wesley. Descoberta de conhecimento relevante em banco de dados sobre Ciência e Tecnologia. 253 f. Tese (Doutorado em Engenharia) – Universidade Federal de Santa Catarina, Florianópolis, 2002.

SAGRES. Sistema de Acompanhamento da Gestão dos Recursos da Sociedade. Disponível em: <<http://www.tce.pb.gov.br>>. Acesso em: 18 jan. 2009.

SANTOS, Celso Bilynkievycz dos; SCANDELARI, Luciano; CARVALHO, Deborah Ribeiro de; VAZ, Maria Salete Marcon Gomes; GOMES, João Carlos. Data Mining para Classificação das Funções de Uma Instituição Pública a Partir das Semelhanças Entre Suas Competências. XIV SIMPEP. Bauru. 2007.

SAMPAIO, M. C.; ROCHA, A. B.; SCHIEL, U. . Guardando Histórico de Dimensões em Data Warehouses. Semana de Informática da Bahia, Salvador - BA, p. 212-230. 2000.

SANTANA, ED W. F. Uma análise sobre a utilização da informação contábil como um instrumento de controle social dos municípios paraibanos com mais de 50.000 habitantes. 116 f. Dissertação (Mestrado). Programa Multiinstitucional e Inter-Regional de Pós-Graduação em Ciências Contábeis da Universidade de Brasília, da Universidade Federal da Paraíba e da Universidade Federal do Rio Grande do Norte. 2008.

SCHEFFER, Tobias. Finding association rules that trade support optimally against confidence: principles of data mining and knowledge discovery. 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), v. 2168, pages 424-435. 2001.

SELL, Denilson. Uma arquitetura para *Business Intelligence* baseada em tecnologias semânticas para suporte a aplicações analíticas. 210 f. Tese (doutorado) - Universidade Federal de Santa Catarina. Programa de Pós-graduação em Engenharia de Produção. Florianópolis, 2006.

SERRA, Laércio. A essência do *Business Intelligence*. São Paulo: Berkeley, 2002.

SILVA, Edna Lúcia da. MENEZES, Estera Muszkat. Metodologia da Pesquisa e Elaboração de Dissertação. UFSC. Programa de Pós-Graduação em Engenharia de Produção. Florianópolis, 2001.

SILVEIRA, Rosimeri de Freitas. Mineração de Dados aplicada à definição de índices em Sistemas de Raciocínio Baseado em Casos. UFRGS, 2003.



SLACK, Nigel; CHAMBERS, Stuart; JOHNSTON, Robert. Administração da Produção. Tradução de Maria Teresa Corrêa de Oliveira e Fábio Alher. 2ª ed. – 7ª reimpressão. São Paulo: Atlas, 2007.

SOUSA, G. W. L. & GROESBECK, R. L. Enterprise Engineering: Managing Dynamic Complexity and Change at the Organizational Level. Proceedings of the 2004 American Society for Engineering Management Conference, out 2004.

SOUSA, Maria Uilma R. S. DE; SILVA, Kelly Patrícia D; ADEODATO, Paulo Jorge L; ARNAUD, Adrian L.; CARVALHO, Francisco de Assis Tenório de. Mineração de dados aplicada à celeridade processual do Tribunal de Contas do Estado de Pernambuco. 5º Congresso Internacional de Gestão da Tecnologia e Sistema de Informação - CONTECSI. São Paulo. 2008.

SPSS. Disponível em: <<http://www.spss.com.br>>. Acesso em: 08 fev. 2010.

REZENDE, S. O. Mineração de Dados. In: Anais do 15º Congresso da Sociedade Brasileira de Computação, pp. 397-433, São Leopoldo-RS, 2005.

TEIXEIRA, Elizabeth. As três metodologias: acadêmica da ciência e da pesquisa. 4. ed. Belém: UNAMA, 2002.

TCE-CE. Tribunal de Contas do Estado do Ceará. Disponível em: <<http://www.tce.ce.gov.br/sitetce/arq/PROMOEX/Index.htm>>. Acesso em: 01 ago. 2010.

TCE-PB. Tribunal de Contas do Estado da Paraíba. Disponível em: <<http://www.tce.pb.gov.br>>. Acesso em: 10 jan. 2009.

TCE-PI. Tribunal de Contas do Estado do Piauí. Disponível em: <<http://www.tce.pi.gov.br>>. Acesso em: 14 jul. 2010.

TAN, Pang-Ning; STEINBACH, Michael & KUMAR, Vipin. Introdução ao *Data Mining*. Rio de Janeiro: Editora Ciência Moderna, 2009.

TCE-SE. Tribunal de Contas do Estado de Sergipe: Regimento Interno. Disponível em: < [http://www.tce.se.gov.br/site/lei\\_organica\\_regimento.php](http://www.tce.se.gov.br/site/lei_organica_regimento.php) >. Acesso em: 10 jan. 2009.

TEIXEIRA, M. G. S. A recuperação da iconografia do Tribunal de Contas do Estado da Bahia. Capítulo: Caminho de Contas. Uma história do Tribunal de Contas da Bahia. Tribunal de Contas do Estado da Bahia. Salvador. 2002.

TERRA, J. C.; BAX, M. P. Portais corporativos: instrumento de gestão de informação e de conhecimento. In: Isis Paim. (Org.). A Gestão da Informação e do Conhecimento. 1º ed. Belo Horizonte, p. 33-53, 2003.

TULER, Elisa; PRATES, Raquel O.; ALMIR, Fernando; ROCHA, Leonardo; MEIRA JR., Wagner. Caracterizando desafios de interação com Sistemas de Mineração de

Regras de Associação. IHC 2006 – VII Simpósio Sobre Fatores Humanos em Sistemas Computacionais. Natal, Rio Grande do Norte. 2006.

TURBAN, Efrain; SHARDA, Ramesh; ARONSON, Jai E. & KING, David. *Business Intelligence: um enfoque gerencial para a inteligência do negócio*. Tradução Fabiano Bruno Gonçalves. Porto Alegre: Bookman, 2009.

VASCONCELOS, Benitz de Souza. Mineração de Regras de Classificação com Sistemas de Banco de Dados Objeto-Relacional. Estudo de Caso: Regras de Classificação de Litofácies de Poços de Petróleo. 127 p. Dissertação (Mestrado), Universidade Federal de Campina Grande, Campina Grande. 2002.

VERCELLIS, Carlo. *Business Intelligence: data mining and optimization for decision making*. John Wiley & Sons Ltd, 2009.

Vieira, Luiz Sergio Gadelha. O futuro dos Tribunais de Contas no Brasil. Disponível em: <<http://www.controlepublico.org.br/index.php/publicacoes/artigos-tecnicos>>. Acesso em: 19 jul. 2010.

WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers Inc., 2000.

XAVIER, Glaúcio Barreto. Sistema de Acompanhamento da Gestão dos Recursos da Sociedade – SAGRES: interface entre os controle externos e social na Paraíba. 57 f. Monografia (Especialização em Gestão Pública). Escola de Contas Cons. Otacílio da Silveira - ECOSIL/Centro Universitário de João Pessoa – UNIPÊ, João Pessoa. Paraíba. 2003.

YIN, R. K. Estudo de Caso: planejamento e método. Porto Alegre: Bookman. 2005.

## APÊNDICE I

Código XML gerado pelo RapidMiner para o procedimento de mineração de dados com Regras de Associação:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.1.006">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.1.006" expanded="true" name="Process">
    <parameter key="logfile" value="C:\Program Files\Rapid-I\arquivos\log.log"/>
    <process expanded="true" height="341" width="547">
      <operator activated="true" class="read_csv" compatibility="5.1.006" expanded="true" height="60"
name="Read CSV" width="90" x="45" y="30">
        <parameter key="csv_file" value="C:\Program Files\Rapid-I\arquivos\nomedoArquivo.csv"/>
        <parameter key="column_separators" value=","/>
        <parameter key="quotes_character" value=""/>
        <list key="annotations"/>
        <list key="data_set_meta_data_information"/>
      </operator>
      <operator activated="true" class="web:transform_log_to_session" compatibility="5.1.000"
expanded="true" height="76" name="Transform Log to Session" width="90" x="45" y="120">
        <parameter key="session_attribute" value="licitacao"/>
        <parameter key="resource_attribute" value="idProponente"/>
      </operator>
      <operator activated="true" class="numerical_to_binominal" compatibility="5.1.006"
expanded="true" height="76" name="Numerical to Binominal" width="90" x="45" y="210">
      <operator activated="true" class="fp_growth" compatibility="5.1.006" expanded="true" height="76"
name="FP-Growth" width="90" x="246" y="30">
        <parameter key="find_min_number_of_itemsets" value="false"/>
        <parameter key="min_support" value="0.0010"/>
      </operator>
      <operator activated="true" class="create_association_rules" compatibility="5.1.006" expanded="true"
height="76" name="Create Association Rules" width="90" x="447" y="30">
        <parameter key="min_confidence" value="0.4"/>
      </operator>
      <connect from_op="Read CSV" from_port="output" to_op="Transform Log to Session"
to_port="Example Set"/>
      <connect from_op="Transform Log to Session" from_port="Session Example Set" to_op="Numerical
to Binominal" to_port="example set input"/>
      <connect from_op="Numerical to Binominal" from_port="example set output" to_op="FP-Growth"
to_port="example set"/>
      <connect from_op="FP-Growth" from_port="frequent sets" to_op="Create Association Rules"
to_port="item sets"/>
      <connect from_op="Create Association Rules" from_port="rules" to_port="result 1"/>
      <connect from_op="Create Association Rules" from_port="item sets" to_port="result 2"/>
      <portSpacing port="source_input 1" spacing="0"/>
      <portSpacing port="sink_result 1" spacing="0"/>
      <portSpacing port="sink_result 2" spacing="0"/>
      <portSpacing port="sink_result 3" spacing="0"/>
    </process>
  </operator>
</process>
```

## APÊNDICE II

Resultado da saída do processamento da mineração dos dados do período estudado por esta pesquisa usando Regras de Associação com o RapidMiner. Para obter estes resultados foi utilizado o suporte de 0.05 (5%) e a medida de confiança 0.6 (60%).

### Association Rules – 2005

```
[27577.0] --> [20445.0, 16560.0] (confidence: 0.604)
[23132.0] --> [25179.0] (confidence: 0.607)
[13689.0] --> [14372.0] (confidence: 0.609)
[7198.0, 27960.0] --> [16979.0] (confidence: 0.611)
[27577.0] --> [16560.0] (confidence: 0.623)
[16795.0] --> [21917.0] (confidence: 0.641)
[6933.0] --> [21154.0] (confidence: 0.667)
[23735.0] --> [24162.0] (confidence: 0.681)
[13154.0, 21225.0] --> [21781.0] (confidence: 0.681)
[14372.0] --> [13689.0] (confidence: 0.689)
[24357.0] --> [10574.0] (confidence: 0.700)
[21638.0] --> [10574.0] (confidence: 0.707)
[21781.0] --> [21225.0] (confidence: 0.709)
[14648.0] --> [8356.0] (confidence: 0.709)
[20445.0] --> [16560.0] (confidence: 0.711)
[20445.0] --> [27577.0, 16560.0] (confidence: 0.711)
[16991.0] --> [13732.0] (confidence: 0.740)
[13154.0, 21781.0] --> [21225.0] (confidence: 0.744)
[14454.0] --> [18741.0] (confidence: 0.745)
[7198.0] --> [27960.0] (confidence: 0.750)
[27577.0] --> [20445.0] (confidence: 0.755)
[15729.0] --> [6030.0] (confidence: 0.771)
[13979.0] --> [27545.0] (confidence: 0.776)
[1930.0] --> [19355.0] (confidence: 0.780)
[21781.0] --> [13154.0] (confidence: 0.782)
[13689.0, 14372.0] --> [23157.0] (confidence: 0.786)
[8356.0] --> [14648.0] (confidence: 0.796)
[16560.0] --> [20445.0] (confidence: 0.800)
[27577.0, 20445.0] --> [16560.0] (confidence: 0.800)
[16560.0] --> [27577.0, 20445.0] (confidence: 0.800)
[27960.0] --> [7198.0] (confidence: 0.806)
[21225.0, 21781.0] --> [13154.0] (confidence: 0.821)
[13732.0] --> [16991.0] (confidence: 0.822)
[16560.0] --> [27577.0] (confidence: 0.825)
[21154.0] --> [6933.0] (confidence: 0.833)
[23157.0, 13689.0] --> [14372.0] (confidence: 0.846)
[6030.0] --> [15729.0] (confidence: 0.860)
[27545.0] --> [13979.0] (confidence: 0.864)
[20445.0] --> [27577.0] (confidence: 0.889)
[16979.0] --> [7198.0, 27960.0] (confidence: 0.892)
[16979.0] --> [7198.0] (confidence: 0.919)
[16979.0] --> [27960.0] (confidence: 0.919)
[19639.0] --> [21917.0] (confidence: 0.925)
[27577.0, 16560.0] --> [20445.0] (confidence: 0.970)
[7198.0, 16979.0] --> [27960.0] (confidence: 0.971)
[27960.0, 16979.0] --> [7198.0] (confidence: 0.971)
```

[23157.0, 14372.0] --> [13689.0] (confidence: 1.000)  
[20445.0, 16560.0] --> [27577.0] (confidence: 1.000)

## Association Rules - 2006

[27960.0] --> [16979.0] (confidence: 0.600)  
[16795.0] --> [21917.0] (confidence: 0.604)  
[6933.0] --> [21154.0] (confidence: 0.609)  
[21638.0] --> [10574.0] (confidence: 0.617)  
[15729.0] --> [15519.0] (confidence: 0.623)  
[10574.0] --> [21638.0] (confidence: 0.627)  
[21154.0] --> [6933.0] (confidence: 0.629)  
[21917.0] --> [21714.0, 16795.0] (confidence: 0.630)  
[15729.0] --> [6030.0] (confidence: 0.642)  
[21917.0] --> [21714.0] (confidence: 0.685)  
[21917.0] --> [16795.0] (confidence: 0.753)  
[6030.0] --> [15729.0] (confidence: 0.810)  
[16795.0, 21917.0] --> [21714.0] (confidence: 0.836)  
[21714.0, 16795.0] --> [21917.0] (confidence: 0.868)  
[15519.0] --> [15729.0] (confidence: 0.868)  
[16979.0] --> [6724.0] (confidence: 0.872)  
[16979.0] --> [27960.0, 6724.0] (confidence: 0.872)  
[27960.0, 16979.0] --> [6724.0] (confidence: 0.872)  
[21714.0, 21917.0] --> [16795.0] (confidence: 0.920)  
[23814.0] --> [18211.0] (confidence: 0.921)  
[18211.0] --> [23814.0] (confidence: 0.921)  
[6724.0] --> [16979.0] (confidence: 0.944)  
[6724.0] --> [27960.0, 16979.0] (confidence: 0.944)  
[27960.0, 6724.0] --> [16979.0] (confidence: 0.944)  
[16979.0] --> [27960.0] (confidence: 1.000)  
[6724.0] --> [27960.0] (confidence: 1.000)  
[16979.0, 6724.0] --> [27960.0] (confidence: 1.000)

## Association Rules - 2007

[1197.0] --> [25541.0] (confidence: 0.621)  
[6100.0] --> [27467.0] (confidence: 0.642)  
[27545.0] --> [13979.0] (confidence: 0.696)  
[6030.0] --> [15729.0] (confidence: 0.805)  
[13979.0] --> [27545.0] (confidence: 1.000)

## Association Rules – 2008

[19786.0] --> [21316.0] (confidence: 0.625)  
[15729.0] --> [17316.0] (confidence: 0.650)  
[27577.0] --> [24076.0] (confidence: 0.733)  
[13979.0] --> [2673.0] (confidence: 0.784)  
[13979.0] --> [27545.0] (confidence: 0.838)  
[27545.0] --> [13979.0] (confidence: 0.886)  
[14307.0] --> [12302.0] (confidence: 0.906)  
[17316.0] --> [15729.0] (confidence: 0.907)  
[24076.0] --> [27577.0] (confidence: 0.943)  
[2673.0] --> [13979.0] (confidence: 0.967)  
[12302.0] --> [14307.0] (confidence: 0.967)

## Association Rules – 2009

[4373.0] --> [14538.0] (confidence: 0.644)  
[3418.0] --> [14538.0] (confidence: 0.750)

## Association Rules - 2009

Valor do suporte de 1% e confiança de 60%, foram gerados 1061 itens frequentes com 491 regras geradas

[24357.0] --> [9764.0] (confidence: 0.600)  
[24444.0] --> [20509.0] (confidence: 0.600)  
[5227.0] --> [23977.0] (confidence: 0.600)  
[19178.0] --> [27329.0] (confidence: 0.600)  
[21528.0] --> [7198.0] (confidence: 0.600)  
[27960.0] --> [6563.0] (confidence: 0.600)  
[27960.0] --> [16979.0] (confidence: 0.600)  
[9297.0] --> [27279.0] (confidence: 0.600)  
[6874.0] --> [7090.0] (confidence: 0.600)  
[9631.0] --> [9699.0] (confidence: 0.600)  
[14760.0] --> [20842.0] (confidence: 0.600)  
[12787.0, 14146.0] --> [27798.0] (confidence: 0.600)  
[15906.0] --> [22202.0, 23884.0] (confidence: 0.600)  
[1165.0] --> [20572.0, 19052.0] (confidence: 0.600)  
[14039.0] --> [24615.0] (confidence: 0.609)  
[14039.0] --> [12762.0] (confidence: 0.609)  
[28227.0, 25102.0] --> [25845.0] (confidence: 0.609)  
[17759.0] --> [23977.0] (confidence: 0.611)  
[17759.0] --> [27329.0] (confidence: 0.611)  
[4293.0] --> [12226.0] (confidence: 0.611)  
[4293.0] --> [11905.0] (confidence: 0.611)  
[19195.0] --> [18857.0] (confidence: 0.611)  
[15224.0] --> [4861.0] (confidence: 0.611)  
[3334.0] --> [20187.0] (confidence: 0.615)  
[5818.0] --> [3334.0] (confidence: 0.615)  
[19895.0] --> [25094.0] (confidence: 0.615)  
[23414.0] --> [3273.0] (confidence: 0.615)  
[23228.0] --> [7325.0] (confidence: 0.615)  
[28648.0, 1646.0] --> [24037.0] (confidence: 0.615)  
[19815.0, 2581.0] --> [14571.0] (confidence: 0.615)  
[19815.0, 14571.0] --> [2581.0] (confidence: 0.615)  
[877.0] --> [19763.0] (confidence: 0.618)  
[24421.0] --> [16062.0] (confidence: 0.619)  
[22279.0] --> [8318.0] (confidence: 0.619)  
[20966.0] --> [25918.0] (confidence: 0.621)  
[27798.0] --> [12787.0] (confidence: 0.625)  
[16155.0] --> [14648.0] (confidence: 0.628)  
[16292.0] --> [17044.0] (confidence: 0.630)  
[3382.0] --> [24545.0] (confidence: 0.632)  
[13979.0] --> [27545.0, 2673.0] (confidence: 0.632)  
[3295.0] --> [14648.0] (confidence: 0.636)  
[14506.0] --> [582.0] (confidence: 0.636)  
[11394.0] --> [22423.0] (confidence: 0.636)  
[15048.0] --> [8447.0] (confidence: 0.636)  
[22400.0] --> [17796.0] (confidence: 0.636)  
[15048.0] --> [10400.0] (confidence: 0.636)  
[19763.0, 21638.0] --> [877.0] (confidence: 0.636)  
[14506.0] --> [582.0, 8172.0] (confidence: 0.636)  
[8172.0, 14506.0] --> [582.0] (confidence: 0.636)  
[27279.0] --> [10447.0, 9297.0] (confidence: 0.636)  
[6563.0, 6724.0] --> [16979.0] (confidence: 0.636)

```

[26463.0] --> [14538.0] (confidence: 0.639)
[2479.0] --> [23354.0] (confidence: 0.640)
[5290.0] --> [26874.0] (confidence: 0.643)
[8356.0] --> [14648.0] (confidence: 0.643)
[28867.0] --> [2140.0] (confidence: 0.643)
[23955.0] --> [15949.0] (confidence: 0.643)
[6859.0] --> [22395.0] (confidence: 0.643)
[9699.0] --> [9631.0] (confidence: 0.643)
[20842.0] --> [14760.0] (confidence: 0.643)
[398.0] --> [3523.0] (confidence: 0.643)
[28930.0] --> [3669.0] (confidence: 0.643)
[19614.0] --> [20307.0] (confidence: 0.643)
[19468.0] --> [9556.0] (confidence: 0.643)
[19468.0] --> [15979.0] (confidence: 0.643)
[15729.0, 8100.0] --> [17316.0] (confidence: 0.643)
[4373.0] --> [14538.0] (confidence: 0.644)
[9539.0] --> [3223.0] (confidence: 0.647)
[9539.0] --> [26214.0, 3223.0] (confidence: 0.647)
[203.0] --> [8283.0, 16924.0] (confidence: 0.654)
[2135.0] --> [20187.0] (confidence: 0.667)
[582.0] --> [12711.0] (confidence: 0.667)
[27467.0] --> [27796.0] (confidence: 0.667)
[24971.0] --> [18577.0] (confidence: 0.667)
[5207.0] --> [24470.0] (confidence: 0.667)
[24481.0] --> [22423.0] (confidence: 0.667)
[22279.0] --> [13560.0] (confidence: 0.667)
[13560.0] --> [22279.0] (confidence: 0.667)
[1165.0] --> [20572.0] (confidence: 0.667)
[1165.0] --> [19052.0] (confidence: 0.667)
[20958.0] --> [4674.0] (confidence: 0.667)
[18417.0] --> [15683.0] (confidence: 0.667)
[1588.0] --> [9020.0] (confidence: 0.667)
[11801.0] --> [21478.0] (confidence: 0.667)
[3418.0, 4373.0] --> [14538.0] (confidence: 0.667)
[24971.0] --> [17044.0, 18577.0] (confidence: 0.667)
[14648.0, 15288.0] --> [16155.0] (confidence: 0.667)
[7198.0, 21528.0] --> [2248.0] (confidence: 0.667)
[6563.0, 16979.0] --> [27960.0] (confidence: 0.667)
[27467.0] --> [27796.0, 4783.0] (confidence: 0.667)
[9297.0] --> [10447.0] (confidence: 0.680)
[16924.0] --> [8283.0, 203.0] (confidence: 0.680)
[27279.0] --> [9297.0] (confidence: 0.682)
[10574.0, 24461.0] --> [9764.0] (confidence: 0.682)
[1646.0] --> [28648.0] (confidence: 0.684)
[6933.0] --> [6698.0] (confidence: 0.684)
[6933.0] --> [11042.0] (confidence: 0.684)
[11042.0] --> [6933.0] (confidence: 0.684)
[26214.0] --> [3223.0] (confidence: 0.688)
[26214.0] --> [9539.0, 3223.0] (confidence: 0.688)
[27796.0] --> [4783.0] (confidence: 0.690)
... 391 other rules ...

```