VIVIANE SIMI MAGALHÃES

Avaliação do Processo de Mineração de Dados no Negócio de Cartões de Crédito

Trabalho Final apresentado ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT, para obtenção do título de Mestre em Engenharia de Computação.

São Paulo

2003

VIVIANE SIMI MAGALHÃES

Avaliação do Processo de Mineração de Dados no Negócio de Cartões de Crédito

Trabalho Final apresentado ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT, para obtenção do título de Mestre em Engenharia de Computação. Área de Concentração: Engenharia de Software.

Orientadora: Dra. Edit Grassiani Lino de Campos.

São Paulo

2003

Magalhães, Viviane Simi
Avaliação do processo de mineração de dados no negócio de cartões de crédito/Viviane Simi Magalhães. São Paulo, 2003.
110 p.

Trabalho Final apresentado ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT, para obtenção do título de Mestre em Engenharia de Computação. Área de concentração: Engenharia de Software.

Orientadora: Profa. Dra. Edit Grassiani Lino de Campos

1. Mineração de dados 2. Cartão de crédito 3. Engenharia de Software 4. Tese I. Instituto de Pesquisas Tecnológicas do Estado de São Paulo. Centro de Aperfeiçoamento Tecnológico II. Título

CDU 004.658(043)
M188a

Agradeço muito à Profa. Edit pela grande contribuição e paciência para a realização deste sonho, e, principalmente, por me tornar um ser humano melhor.

Agradeço ainda à administradora de cartões de crédito que cedeu os casos práticos e a experiência profissional cotidiana, ambos vitais para a realização deste trabalho.

Dedico este trabalho a todos os meus familiares e amigos, os maiores responsáveis pela minha formação e pelo meu desenvolvimento pessoal.

Em especial, dedico ao meu marido Eduardo, pela incansável compreensão nos momentos de ausência e de cansaço, sendo minha maior fonte de motivação e de inspiração durante todos estes anos de intensa dedicação.

Sumário

Lista de figuras

Lista de tabelas

Lista de símbolos, siglas e abreviaturas

Abreviatura	Significado
AID	Algoritmo de árvore de decisão, caracterizado pela iteração automática dos grupos (em inglês, ***A**utomatic **I**nteraction **D**etection*).
ARIMA	Modelagem integrada de média móvel
AVR/URA	Ferramenta utilizada em centrais de atendimento para a automatização de processos mais freqüentes, através da unidade remota automatizada.
CRM	Gerenciamento de relacionamento com clientes (em inglês, ***c**ustomer **r**elationship **m**anagement*).
DW	Armazém de dados (em inglês, ***d**ata **w**arehouse*).
GUI	Interface gráfica para usuário (em inglês, ***g**raphical **u**ser **i**nterface*).
KDD	Descoberta de conhecimento de dados (em inglês, ***k**nowledge **d**iscovery in **d**atabases*).
MIS	Sistemas de informações gerenciais (em inglês, ***m**anagement **i**nformation **s**ystems*)
OLAP	Ferramentas que permitem ao usuário a realização de consultas via armazém de dados (em inglês, ***O**n **L**ine **A**nalytical **P**rocess)*
SEMMA	Metodologia implementada pela empresa de software SAS cujas fases são: ***S**ample, **E**xplore, **M**anipulate, **M**odel, **A**ssess.*
SQL	Linguagem estruturada no formato de consultas (em inglês, ***S**tructured **Q**uery **L**anguage)*

Resumo

Ao longo da era da tecnologia, com o crescente desenvolvimento das técnicas e ferramentas computacionais e com a queda no custo de armazenamento, a mineração de dados vem ganhando um número cada vez maior de adeptos corporativos que buscam na aplicação de seu processo garantir maior assertividade em suas decisões organizacionais, visando aumento de lucratividade, prospecção, fidelização e retenção de seus clientes.

Neste contexto, o presente trabalho visa avaliar a aplicação do processo de mineração de dados no ambiente corporativo, especialmente em áreas de administração de cartões de crédito, através da análise dos casos reais, a fim de avaliar a adequação do processo teórico à realidade corporativa (e vice-versa). Esta análise contempla desde as fases do processo até as técnicas e ferramentas utilizadas, buscando a otimização do processo em todo o seu potencial.

Dessa forma, a análise dos casos práticos vivenciados no ambiente corporativo face à metodologia de mineração de dados definida na Literatura possibilita a identificação de melhorias a serem aplicadas no processo prático bem como a sugestão de novas fases ou adaptações no processo teórico, tendo em vista o aprendizado prático.

Por fim, a utilização da Engenharia de Software viabiliza a padronização na especificação de requisitos e na documentação, facilmente utilizada para o processo de mineração de dados, trazendo assim novas e importantes contribuições para a criação de um processo mais sólido e bem sucedido.

Abstract

Throughout the technology era, with the rapid development of computer tools and techniques and storage cost reduction, data mining has been obtaining a crescent number of corporation adepts that want to use this process to guarantee better hit rates in their organizational decisions, in order to increase their profitability, market share, customer acquisition and retention.

In this context, this research analyses the application of data mining process in business environment, specially in credit card management areas, presenting real-world data mining cases, in order to assess the accuracy of the theoretical process in organizational reality (and vice-versa). This analysis contains all process steps, including techniques and tools, and aims to detect possible process optimization opportunities.

Thus, the case studies analyzed and compared to theoretical data mining methodology result to new possibilities to be applied to practical process with the suggestion of new steps or adaptations on the process, in order to the practical learning.

So, the Software Engineering process can be used in the data mining process as a way to facilitate the requirements specification and the documentation, resulting in new and important contribution to develop a more solid and successful process.

Capítulo 1: Introdução

1.1 Motivação

Com o crescimento da competitividade entre as empresas, um processo decisório eficaz torna-se cada vez mais vital para o sucesso organizacional. Visando facilitar a tomada de decisões e, principalmente, a obtenção de diferencial estratégico, a mineração de dados revela-se uma ferramenta muito poderosa, uma vez que possibilita a extração de conhecimento através das informações existentes nos bancos de dados da empresa [1,8,12,22,23,24].

Dessa forma, muitas organizações, entre as quais Wal Mart, Banco Itaú e Embratel, têm buscado otimizar seus processos e estratégias, através do instrumento de mineração de dados que viabiliza a identificação de tendências de clientes, facilitando a tomada de decisões e realização de ações, de forma mais focada, conseqüentemente com redução de custos e maior taxa de assertividade.

Sendo assim, a análise do processo de mineração de dados no ambiente corporativo torna-se muito necessária, pois possibilita a avaliação crítica e detalhada das metodologias comumente utilizadas na prática cotidiana pelas empresas, a fim de melhorá-las e otimizá-las continuamente.

Como fatores motivacionais para a realização deste trabalho citam-se:

- Aprofundamento dos conhecimentos teóricos de mineração de dados dada à experiência vivida profissionalmente nos casos práticos a serem abordados.
- Avaliação dos pontos fortes e fracos do processo de mineração de dados num cenário real e dinâmico.
- Sugestões de melhoria ao processo, dada a análise crítica dos casos práticos e de futuras pesquisas neste segmento de extração de conhecimento através de dados.
- Melhoria no desempenho dos profissionais das áreas de negócio e tecnologia na realização de projetos relacionados à informação.
- Aproveitamento de conceitos da Engenharia de Software para melhoria do processo de mineração de dados.

Neste contexto, o negócio de cartões de crédito representa uma fonte rica de informações práticas relacionadas ao processo de mineração de dados, uma vez que concentra dentre suas atividades principais a definição de estratégias e políticas para

maximização de rentabilidade com administração de risco de inadimplência e fraude. Para uma administradora de cartões de crédito, cujo objeto de existência corresponde à concessão de crédito à população, a análise de informações históricas para previsão de dados futuros baseados no comportamento de clientes revela-se importante ferramenta para a manutenção do sucesso do negócio.

Observa-se ainda que o processo de mineração de dados pode ser comparado ao processo de Engenharia de Software, sendo que em ambos os casos, as etapas do processo (especialmente, análise de requisitos e documentação) devem ser cumpridas para obtenção de um bom produto final. Dessa forma, a análise destes dois processos estruturados (Engenharia de Software e mineração de dados) deve encontrar muitas semelhanças, salvo particularidades do processo de mineração de dados, dada sua maior especificidade.

1.2 Objetivo

O presente trabalho tem como objetivo principal realizar uma avaliação do processo de mineração de dados vivido na prática cotidiana das organizações, especialmente em áreas de análise de risco, de forma a validar os conceitos existentes, confirmando-os ou sugerindo melhorias, de acordo com a experiência obtida nos casos práticos. Esta avaliação visa, principalmente, o amadurecimento das empresas quanto aos processos de mineração de dados.

Para a realização desta análise serão utilizados como referências o processo de mineração de dados definido teoricamente [2,3] e o processo de mineração de dados identificado na utilização prática em área de administração de cartões de crédito. Como referência, será utilizado o processo de Engenharia de Software. A partir das comparações entre estas estruturas e seus respectivos resultados, uma avaliação de processo de mineração de dados será criada, a fim de potencializar o processo já existente ou simplesmente validar a metodologia definida teoricamente.

Deste modo, o foco principal deste trabalho é o processo de mineração de dados, porém não há como discorrer sobre o mesmo, sem avaliar todos os aspectos que lhe dizem respeito, tais como técnicas, plataformas e algoritmos, pois para um processo funcionar de forma efetiva e eficiente, todos os sub-processos devem ser estabelecidos de forma harmônica e coerente.

Sendo assim, os casos práticos avaliados sob os pontos de vista técnico e de negócio, servem de base a fim de garantir a melhor utilização desta poderosa

ferramenta de transformação de dados em informação e, conseqüentemente de estratégia empresarial, de modo a permitir a avaliação da utilização dos conceitos, técnicas e ferramentas de extração de conhecimento.

A diversidade representada pelos cinco estudos de caso possibilita a avaliação do processo de mineração de dados sob diversas óticas (focos de negócio distintos ou diversas técnicas escolhidas), de forma a confrontar a metodologia de forma unificada, através de realidades diferentes, para se ter uma visão global mais acurada e aplicável para outros tipos de organizações.

Este ambiente organizacional refere-se a uma administradora de cartões de crédito, tendo como principal responsabilidade os controles de todos os processos, políticas e processos de concessão de crédito, manutenção de contas, cobrança e prevenção a fraudes.

Dentre os estudos de caso destacados para análise citam-se:

1. **Sistema de informações gerenciais de autorização:** projeto de construção de um sistema de informações gerenciais, a fim de suprir as necessidades da gerência e da diretoria para o embasamento das decisões relacionadas às políticas de autorização de venda.
2. **Venda de carteira de crédito em liquidação**: projeto para análise de carteira de clientes inadimplentes, a fim de identificar o perfil de recuperação dos mesmos, garantindo uma previsão de preço de venda adequado à cessão de direitos.
3. **Modelo de renda presumida**: projeto de construção de um modelo estatístico, baseado nos dados demográficos e profissionais informados na proposta para a obtenção de cartão de crédito, visando otimizar o processo de aprovação de contas.
4. **Análise de perfil de cadastro**: projeto de desenvolvimento de uma estrutura de análise automática de cadastros, a fim de identificar clientes potenciais para prospecção e abordagem para venda e oferta de produtos e / ou serviços, considerando as variáveis fornecidas em cada arquivo x o perfil histórico dos clientes.
5. **Sistema de prevenção a fraudes em autorizações**: projeto para implantação de regras de detecção de fraudes em compras nos cartões de crédito, baseado no comportamento histórico das ocorrências de fraude.

Logo, o principal objetivo deste trabalho consiste em avaliar o processo de mineração de dados através da análise de cinco casos práticos vivenciados em uma

administradora de cartões crédito, de forma a comparar a metodologia teórica com a experiência prática, garantindo a aderência dos processos práticos à definição da Literatura. Esta avaliação visa sugerir melhorias ao processo utilizado pelas empresas, a fim de amadurece-las em mineração de dados, assim como já ocorre em Engenharia de Software.

1.3 Metodologia

A metodologia utilizada para o desenvolvimento do trabalho baseia-se em pesquisas relacionadas ao tema, através da leitura de livros e de artigos divulgados na Internet, e em estudos de casos práticos desenvolvidos no ambiente organizacional.

Com estes instrumentos teóricos e práticos, é possível realizar-se a análise crítica do processo de mineração de dados realizado nas empresas bem como das técnicas e ferramentas utilizadas, avaliando-se a adequação dos processos para a tomada de decisões organizacionais.

O trabalho realiza a conexão entre os conceitos teóricos extraídos dos livros e artigos com a vivência prática trazida pelos estudos e visitas, de forma a garantir foco real, atendendo as necessidades das empresas, especialmente em áreas de administração de cartões de crédito.

1.4 Organização do trabalho

O presente trabalho é organizado da seguinte forma:

- O capítulo 1 contém a introdução ao assunto a ser tratado bem como o objetivo e a metodologia utilizada para a confecção do texto, trazendo também a motivação para a realização da pesquisa;
- O capítulo 2 descreve os aspectos teóricos que abordam a mineração de dados em seu estado da arte. São apresentadas as definições, conceitos, técnicas, algoritmos, ferramentas e plataformas, bem como o processo de mineração de dados em si com suas fases de execução. Nesta seção será realizada a comparação entre o processo de mineração de dados definido teoricamente e o processo de Engenharia de Software.
- O capítulo 3 apresenta o ambiente corporativo de análise de risco e suas necessidades e os casos práticos levantados no cotidiano da empresa, descrevendo todo o processo de mineração de dados realizado para a

obtenção dos resultados definidos nos objetivos. Neste capítulo, há também o detalhamento dos dados operacionais e das técnicas utilizadas nos casos práticos, classificando-os de acordo com os conceitos e definições de mineração de dados existentes.

- O capítulo 4 apresenta a análise dos casos práticos quanto à aderência aos conceitos teóricos e práticos do processo de mineração de dados, à observação sobre os resultados e ao alcance dos objetivos de negócio, através do levantamento dos pontos fracos e fortes. Com base nesta análise, será detalhada uma avaliação do processo de mineração de dados, a fim de validar a metodologia existente e melhorar a utilização da mesma de acordo com os casos práticos.
- O capítulo 5 traz a conclusão quanto ao estudo realizado, contendo um sumário geral do assunto, as possíveis contribuições e as sugestões para futuros trabalhos bem como os cuidados a serem tomados para a obtenção de melhores resultados na utilização da extração de conhecimento para a tomada de decisões e transformação dos padrões e tendências descobertos em ações rentáveis para a organização.

Capítulo 2: Visão Geral sobre Mineração de dados

2.1 Introdução

Neste capítulo, serão abordados os conceitos teóricos de mineração de dados, de acordo com pesquisa em artigos e livros sobre o assunto. Dentre os conceitos importantes, cita-se o detalhamento do processo, das plataformas, das técnicas e das ferramentas.

2.2 O conceito da mineração de dados

Nos negócios, dados trazem informações sobre os mercados críticos, concorrentes e clientes. Já em indústrias, dados capturam oportunidades em eficiência e otimização, assim como chaves para a melhoria de processos e resolução de problemas.

Dados primitivos têm raramente benefício direto, seu real valor está na habilidade de extrair informação útil para o suporte a decisões ou exploração e entendimento dos fenômenos que governam os dados originalmente.

Conforme definição de Robert Groth [2], mineração de dados é o processo de descoberta de tendências e padrões existentes, através da manipulação dos dados e da extração de conhecimento relevante ao processo decisório organizacional.

Segundo Jiawei Han [3], mineração de dados refere-se à extração ou mineração de conhecimento de grandes quantidades de dados. O termo de "mineração" corresponde à extração de ouro das rochas. Assim, a mineração de dados pode ser mais apropriadamente denominada de "mineração" de conhecimento através de dados. Existem outros termos que carregam um significado semelhante ou levemente diferente para a mineração de dados como descoberta de conhecimento de banco de dados (*knowledge discovery in databases*, em inglês, usualmente conhecido como KDD), extração de conhecimento, análise de dados e padrões, arqueologia de dados, dragagem de dados, dentre outros termos [28].

Tradicionalmente, análise era um processo manual restrito. Um ou mais analistas poderiam tornar-se intimamente familiares com os dados e, com o auxílio de técnicas estatísticas, proverem resumos e gerarem relatórios.

Na realidade, os analistas atuavam como sofisticados processos de consultas. Porém, como o crescimento da quantidade de dados e das dimensões, este processo

perdeu sua eficiência. Quem poderia entender milhões de casos, cada qual com centenas de campos? Para complicar ainda mais a situação, a quantidade de dados vem crescendo continuamente e com tal rapidez, de forma que análises manuais tornam-se inviáveis.

Questões específicas de aplicações KDD em análise de dados científicos são elaboradas com exemplos ilustrativos por Fayyad, Haussler e Stolorz [27, 28, 30], que defendem o uso de KDD para analisar conjuntos massivos de dados para que os cientistas fiquem livres para focar suas atividades na formação de hipóteses e teorias e na derivação de insights para a elucidação de fenômenos.

De acordo com [28], Etzioni explora os desafios e oportunidades apresentadas na descoberta de informações úteis nos vastos recursos da Internet, concluindo que a efetividade da mineração de dados na *Web* é viável na prática.

Ainda segundo [28], trabalhos endereçados ao problema vital de *KDD* estão em desenvolvimento: representação, complexidade de busca e uso de conhecimento prioritário para auxiliar na busca e na inferência estatística remanescente. Independentemente, aplicações bem sucedidas aparecem continuamente, dirigidas principalmente para o contexto aglutinado de bancos de dados que tem claramente matéria prima humana com habilidades de processamento. Conduzindo o crescimento de cada campo está as forças econômica e social, resultando no fenômeno de carregamento excessivo de dados a que todos estão familiarizados.

Estatística está no centro do problema de inferência de dados. Através da validação da hipótese e análise exploratória dos dados, técnicas estatísticas têm sua importância fundamental. Em reconhecimento de padrões e inteligência artificial, os algoritmos são baseados na idéia de que os dados podem ser carregados na memória principal de um computador, conforme artigos de Imielinsk e Mannila [32], que atualmente estudam o desafio de tecnologias de grandes bancos de dados em memória principal.

Outro aspecto vital refere-se à qualidade dos dados, crítica para a análise dos dados. Immon [14] revela a importância deste assunto e da necessidade da construção de *data warehouses* no processo de *data mining*. Outros estudiosos no assunto na esfera industrial são Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro e Simoudis [28, 31,33], que destacam como a mineração de dados influencia a forma como as empresas guiam seus negócios e os desafios práticos para a utilização de aplicações KDD.

A mineração de dados surgiu da fusão dos conceitos da Estatística e da Inteligência Artificial juntamente com os recursos computacionais e as aplicações de negócio [4], como mostrado na Figura 2.1.

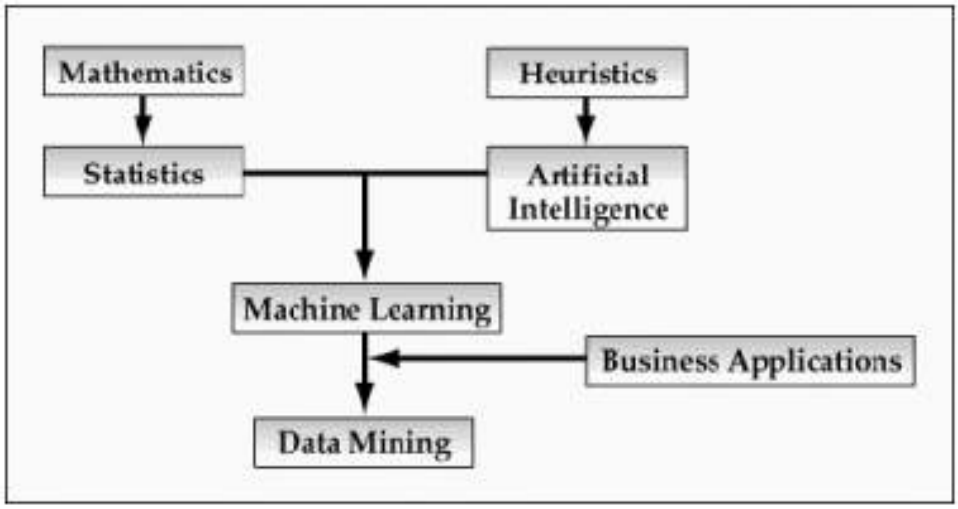

Figura 2.1: Origens do processo de mineração de dados [26]

De acordo com [47], o processo de mineração de dados é composto por três pilares, responsáveis pela realização da identificação de padrões e tendências em grandes quantidades de dados, conforme mostrado na figura 2.2.

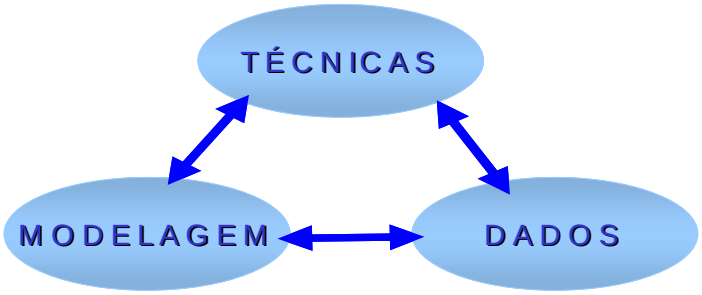

Figura 2.2: Pilares do processo de mineração de dados [47]

Os **dados** servem de base para a escolha e aplicação das **técnicas**, através de **modelos** preditivos ou descritivos, promovendo a obtenção de conhecimento. Esta aquisição de conhecimento consiste na transformação dos dados em informações úteis para análise e síntese, conforme demonstrado na figura 2.3.

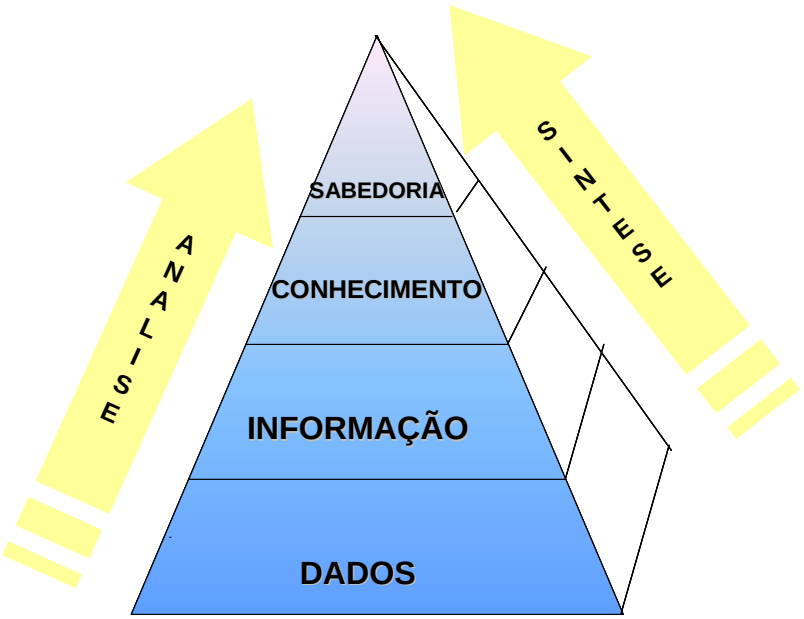

Figura 2.3: Hierarquia de transformação de dados, através da mineração

De acordo com a figura 2.3, observa-se que a transformação de dados em conhecimento ocorre em diversos níveis, passando pelos processos de análise e síntese comumente realizados pela mente humana dentro do processo de mineração de dados, mas também automatizados ou semi-automatizados através de sistemas específicos para administração de conhecimento.

É importante salientar que os dados somente devem ser transformados em conhecimento quando as informações obtidas através deste processo sejam adequadamente avaliadas e definidas como de interesse para o gerenciamento do negócio, pois atualmente existe imensa quantidade de dados disponíveis nos sistemas organizacionais, tornando-se tarefa essencial definição de escopo para o uso dos mesmos, para que não se perca tempo e dinheiro em dados sem valor agregado ao negócio.

Os sistemas especialistas e baseados em conhecimento, desenvolvidos e pesquisados em Inteligência Artificial [45], possuem um processo de desenvolvimento semelhante ao processo de mineração de dados, conforme figura 2.4.

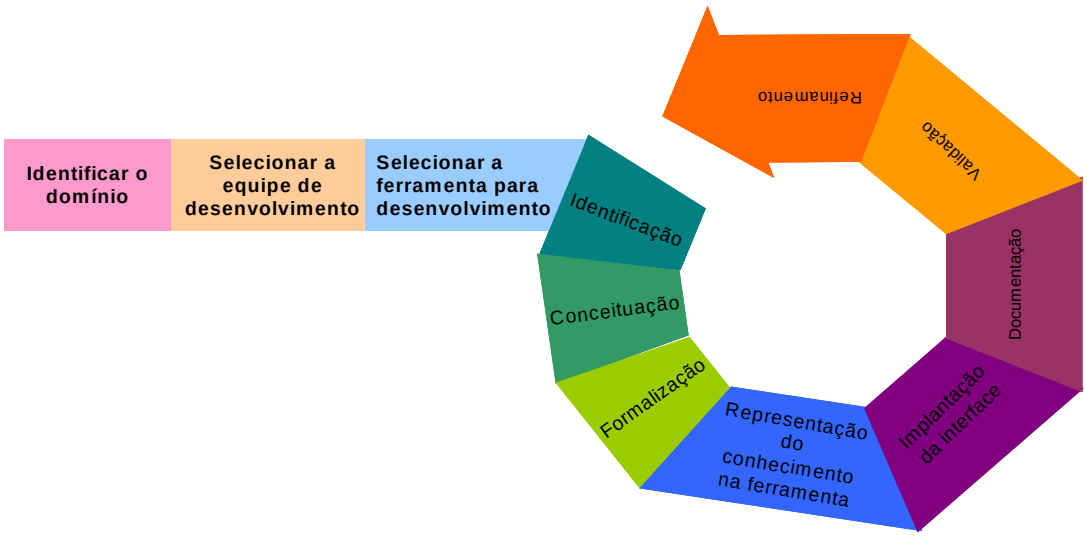

Figura 2.4: Processo de desenvolvimento de sistema baseado em conhecimento [45]

Dada a proximidade conceitual com a mineração de dados, para o desenvolvimento de sistemas baseados em conhecimento são realizadas atividades comuns nos projetos de desenvolvimento de sistemas, destacando-se apenas algumas particularidades da Inteligência Artificial, como, por exemplo, a etapa de representação do conhecimento na ferramenta, garantindo a compreensão e o interesse dos usuários destes tipos de sistemas (normalmente, médicos, cientistas e engenheiros).

Sendo assim, mineração de dados é um processo de descoberta de conhecimento a partir de grandes quantidades de dados armazenados em bancos de dados, *data warehouses* ou outras formas de repositórios [2, 3, 4]. Envolve a integração de múltiplas disciplinas, tecnologias de bancos de dados, conceitos de Estatística, aprendizado de máquina, computação de alta velocidade, reconhecimento de padrões, redes neurais, visualização de dados, filtragem de informações, processamento de imagem e sinal e análise de dados espaciais. A descoberta de conhecimento pode ser aplicada para a tomada de decisões, controle de processos, gerenciamento de informações e processamento de consultas. Devido à sua grande versatilidade e abrangência, a mineração de dados é considerada hoje uma das áreas mais promissoras da indústria da informação e apresenta sucessos práticos, como nos casos dos supermercados Wal Mart [22, 26], onde o consumo cresceu em 30% às sextas feiras com a redefinição da distribuição física dos produtos, baseada na

conexão das hipóteses desenvolvidas pela mineração de dados; e o Banco Itaú [23,26] que aumentou a taxa de retorno de malas diretas em 30%, reduzindo o custo de expediência de correspondências em 80%, devido à implantação de um processo de mineração de dados eficiente.

2.3 O processo de mineração de dados

A mineração de dados refere-se a um processo [2, 3, 17] que consiste dos seguintes passos, como mostrado na Figura 2.5.

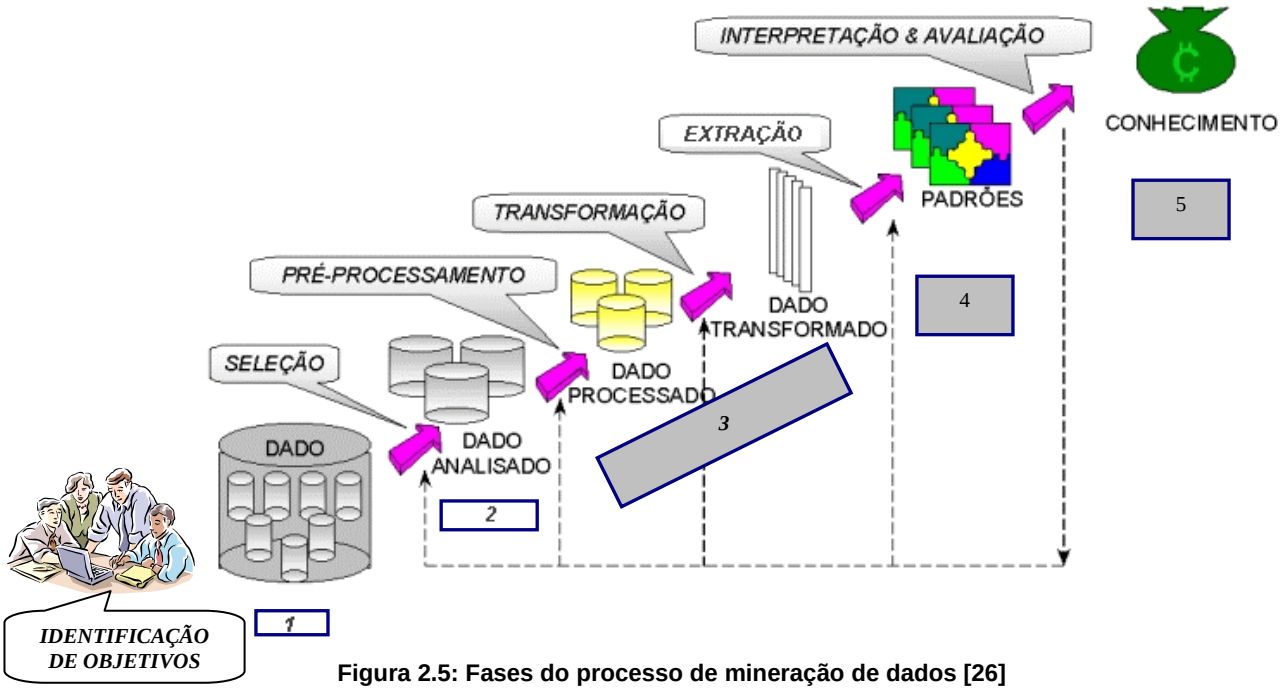

Figura 2.5: Fases do processo de mineração de dados [26]

1. **Identificação dos objetivos:** consiste em definir com precisão o problema a ser estudado ou o objetivo a ser alcançado, de forma a garantir que o mesmo seja solucionado ou atingido como resultado final do processo de mineração de dados, evitando re-execução, sobrecarga de processamento ou falta de objetividade e foco na realização do processo.

Dentre as atividades [3] que fazem parte desta etapa citam-se:

- **definição de escopo do estudo:** tendo como base os objetivos a serem alcançados, há necessidade de uma definição detalhada do escopo do estudo.

- **entendimento dos limites do estudo:** uma vez definido o escopo do estudo, deve-se ponderar os limites do estudo, visando focar esforços de forma mais efetiva.
- **escolha de bons estudos a serem realizados:** levantamento de propostas de estudos, baseadas no escopo previamente estabelecido. Como exemplos de tipos de estudos, citam-se hábitos de consumo de clientes, perfil demográfico de clientes, estudos dependentes de tempo, gerenciamento de cancelamento de clientes, previsão de risco, análise de rentabilidade, análise de tendência, estudos de empregabilidade, estudos regionais, dentre outros.
- **determinação dos elementos corretos para o estudo:** refere-se à escolha das variáveis a serem analisadas no estudo, de forma a selecioná-las previamente para amostra, evitando reprocessamento.
- **entendimento da amostra:** corresponde à compreensão das variáveis da amostra, a fim de evitar utilização de variáveis com conteúdos não confiáveis ou incorretos, e, principalmente, para utilizar os campos corretos de acordo com a concepção do estudo.

Vale ressaltar que a etapa de identificação dos objetivos deve ser realizada em conjunto com os profissionais envolvidos na tomada de decisão, visando esclarecimento correto e adequado do problema a ser solucionado.

2. **Seleção & Coleta de dados:** refere-se à extração de dados necessários para análise, através da definição de uma amostra que atenda aos objetivos identificados. Nesta fase, realiza-se a seleção das variáveis (campos contendo domínios interessantes para a analise) e sistemas adequados à mineração de dados, como por exemplo, para uma análise de rentabilidade de clientes, seleciona-se desde variáveis obtidas na proposta até variáveis de risco e comportamento dos clientes. Estas variáveis podem estar alocadas em diferentes sistemas da empresa. Sendo assim, ocorre nesta etapa, a seleção de variáveis e sistemas a serem analisados.

Dentre as principais atividades [17] da etapa de seleção e coleta de dados, tem-se:

- Definição de variáveis a serem utilizadas e seus sistemas envolvidos, ou seja, escolha de dados demográficos ou comportamentais, de acordo com o tipo de análise a ser realizada.
- Estabelecimento de sistemas a serem acessados: *data warehouses*, bancos de dados relacionais, arquivos textos. Esta atividade consiste na

definição da plataforma a ser acessada para o processo de mineração de dados.

- Verificação da adequação dos dados na descrição do problema.
- Estabelecimento da recência e do período dos dados a serem coletados.
- Validação da consistência e redundância dos dados disponíveis.
- Verificação dos cruzamentos necessários entre diferentes arquivos/tabelas.

Nesta etapa, o foco principal consiste na escolha dos dados a serem selecionados ou coletados para a formação da base de dados ou amostra para a realização da mineração de dados. Sendo assim, todos os aspectos referentes à elaboração da massa de dados devem ser abordados e apontados nesta fase.

3. **Preparação de dados (Preprocessamento, Transformação e Extração):** para garantir a utilização de dados válidos e consistentes, a etapa de preparação de dados consiste nas fases limpeza, integração, transformação e redução de dados, visando à eliminação de "ruídos" e dados inconsistentes, padronização dos dados de diferentes fontes, normalização dos dados para melhorar a atuação dos algoritmos e diminuição da quantidade de dados, através de classificações e segmentações, respectivamente.

Dentre as principais atividades [2, 3, 17], citam-se:

- Limpeza dos dados: envolve a remoção de "ruídos" e a correção de inconsistência dos dados, tais como valores não preenchidos.
- Integração dos dados: refere-se ao cruzamento dos dados de origens distintas para um armazenamento coerente, tais como um *data warehouse,* uma tabela relacional ou um cubo agregado de dados.
- Transformação dos dados: corresponde a utilização de técnicas de normalização, tratamento de campos, dentre outras transformações nos dados que se façam necessárias.
- Redução dos dados: envolve a redução do tamanho dos dados, através de agregações, eliminação de informações redundantes, segmentações e compressões.
- Extração dos dados: corresponde à extração propriamente dita dos dados dos arquivos produtivos para as bases de dados previamente definidas na etapa de seleção e coleta, com os dados já preparados e tratados, conforme atividades mencionadas acima. Normalmente, esta etapa de

extração ocorre da alta plataforma (mainframe) para a baixa plataforma (micro) ou *data warehouses*.

Vale ressaltar que esta etapa pode ser suprimida ou automatizada na existência de *data warehouses*, pois todo o trabalho de seleção e coleta dos dados bem como de tratamento dos campos já foi previamente definido e implantado, minimizando de forma significativa os esforços de mineração de dados e, principalmente, a ocorrência de erros oriundos da falta de conhecimento dos sistemas produtivos da organização.

4. **Extração de padrões:** consiste na mineração de dados propriamente dita, contemplando a aplicação de técnicas e a seleção de ferramentas para a descoberta de padrões e regras, a fim de atingir os objetivos predefinidos e solucionar os problemas levantados.

Esta etapa ocorre de forma cíclica, de acordo com os objetivos técnicos da mineração de dados, conforme a técnica e o algoritmo selecionado para a obtenção dos resultados, através da análise dos dados. Dessa forma, dado o caminho a ser utilizado para a mineração de dados, tem-se o tratamento dos dados necessário para aplicação da técnica, utilizando ferramentas que possuam os algoritmos capazes de executá-la.

Dentre as principais atividades da etapa de análise de dados, tem-se:

- **Seleção de técnicas**: corresponde à escolha das técnicas a serem utilizadas para a análise dos dados, visando alcance dos objetivos identificados na fase inicial do processo. Nesta atividade, é importante o envolvimento de profissionais técnicos (estatísticos, matemáticos, engenheiros), a fim de garantir a utilização de técnicas adequadas ao objetivo proposto.
- **Escolha das ferramentas**: de acordo com as técnicas selecionadas, esta atividade consiste na escolha das ferramentas disponíveis para aplicação da técnica em questão. A seleção das ferramentas é altamente dependente da atividade a ser realizada pelo processo de mineração de dados.

  Através da seleção da tecnologia correta (contendo o algoritmo apropriado), as características e a estrutura dos dados também precisam ser considerados, tais como número de campos com valores contínuos,

número de variáveis dependentes, número de campos categóricos, tamanhos e tipos de registros.

Outros aspectos a serem considerados na seleção das ferramentas são: escalabilidade, precisão, formatos, recursos de pré-processamento (limpeza, integração, transformação e transformação), conectividade, recursos de importação e exportação de dados, gerenciamento de memória, eficiência, tolerância a ruídos e eficiência.

- **Tratamento dos dados:** semelhante à etapa de pré-processamento de dados, citada na fase anterior, porém de forma reduzida, esta etapa tem como objetivo prepará-los para a aplicação das técnicas selecionadas. Como exemplo, para a realização de modelagem estatística baseada em regressão linear ou logística, existem passos anteriores à regressão que devem ser seguidos para que os dados a serem inseridos na ferramenta, estejam legíveis para o bom funcionamento da técnica, ou seja, na regressão linear, as variáveis participantes devem estar formatadas de forma binária, enquanto que na regressão logística, as variáveis podem possuir domínios contínuos.
- **Análise humana**: refere-se à intervenção humana para a efetiva análise dos dados resultados, a fim de retroalimentar o processo até o alcance de resultados satisfatórios do ponto de vista decisório. De acordo com esta análise, que deve envolver desde os analistas técnicos (estatísticos, matemáticos ou de sistemas) até os analistas de negócio para garantir uma boa interpretação dos resultados da aplicação das técnicas em congruência com a realidade do negócio, o processo evolui para a fase seguinte ou retorna a etapas anteriores para a realização de ajustes ou melhorias.

5. **Interpretação e avaliação do conhecimento:**

Nesta etapa final, são avaliados os resultados finais do processo de mineração de dados, de forma a avaliar se os objetivos iniciais foram atingidos plenamente. Com isso, o conhecimento obtido deve ser transformado em ações ou decisões, sob a forma de implantações de políticas e estratégias.

Dentre as principais atividades da etapa de análise de dados, tem-se:

- **Avaliação dos resultados:** nesta fase, os profissionais envolvidos na fase inicial de identificação de objetivos devem tomar conhecimento dos resultados obtidos pelo processo de mineração de dados, buscando

definir os próximos passos a serem realizados do ponto de vista de negócio, tendo em vista que a mineração de dados somente tem sentido quando convertida em ações e decisões estratégicas e de melhoria para a empresa.

- **Implantação:** consiste na etapa de transformação do estudo realizado em ações e regras de negócio. Isto pode ser alcançado seja através da utilização da técnica de visualização com da implantação do novo processo, da nova política ou estratégia em ambiente produtivo. Como exemplo, no caso da análise de retenção de clientes, utilizam-se os meios pré-estabelecidos (malas diretas, brindes, etc), tendo como base à análise dos clientes mais propensos ao cancelamento. Nesta etapa, as atividades principais variam de acordos com as ações pré-estabelecidas para cada problema, porém se referem à implementação dos resultados em ambiente produtivo.
- **Documentação:** consiste na fase de finalização da documentação do projeto, a fim de assegurar o registro do aprendizado obtido, independente da transformação do mesmo em regras para o negócio. A documentação deve ser confeccionada durante todo o processo, sendo apenas concluída e arquivada nesta etapa, conforme os padrões e normas específicos de cada empresa.
- **Acompanhamento:** esta fase se faz necessária para avaliar periodicamente a eficácia da implantação das regras de negócio, através de novos processos, políticas e estratégias, tendo em vista as alterações e mudanças nas tendências de mercado e de economia, fatores diretamente relacionadas ao comportamento de clientes, logo do negócio de cartões de crédito. Nesta etapa, a freqüência e a forma de acompanhamento variam de acordo com as necessidades de negócio e com o estilo de gerência. Como exemplo, o acompanhamento pode ser realizado através do desenvolvimento e análise de relatórios de eficácia das políticas implantadas, fornecidos pelos sistemas de informações gerenciais.

Outro aspecto crucial para o sucesso de qualquer projeto de mineração de dados citado por [47] refere-se à necessidade do contato com pessoas que entendam sobre o negócio, durante todo o projeto. Estes profissionais são denominados especialistas de domínio e possuem tanto o conhecimento necessário para a

realização da mineração de dados quanto à intuição pertinentes às tendências e padrões do negócio ao qual fazem parte.

Por tratar-se de um processo iterativo, ou seja, com muita retroalimentação e reprocessamentos, [47] também definiu um fluxo de construção de um processo de mineração de dados, conforme figura 2.6. Este fluxo definido de forma mais detalhada e específica, apresenta uma rotina de desenvolvimento de modelos estatísticos, desde a identificação dos requisitos e dados necessários para a modelagem até a avaliação dos resultados obtidos, dada a técnica escolhida.

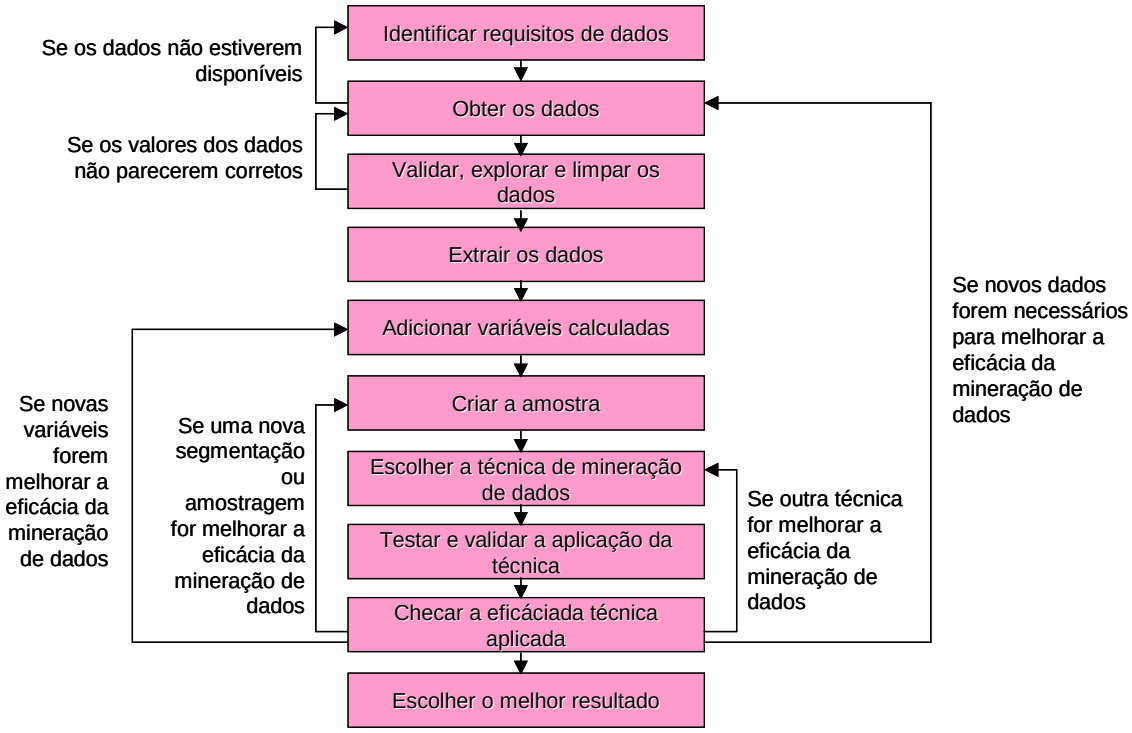

Figura 2.6: Construção do processo de mineração de dados como um processo iterativo [47]

Conforme [3], a princípio, a mineração de dados pode ser aplicada a todos os tipos de repositórios, incluindo bancos de dados relacionais, *data warehouses*, bancos de dados transacionais, aplicações avançadas de bancos de dados, arquivos planos e *Web*. Aplicações avançadas de bancos de dados contemplam bancos de dados orientados a objetos, objeto-relacionais e bancos de dados orientados a aplicações especiais, tais como espaciais, seqüenciais, texto e multimídia. Os desafios de mineração de dados diferem de acordo com o tipo de repositório.

2.4 Exemplo de metodologia de mineração de dados

Há alguns exemplos práticos de utilização do processo de mineração de dados, como no caso da SAS Institute, empresa de software, que utiliza a metodologia SEMMA (Sample Explore Manipulate Model Assess), como mostrada a figura 2.7.

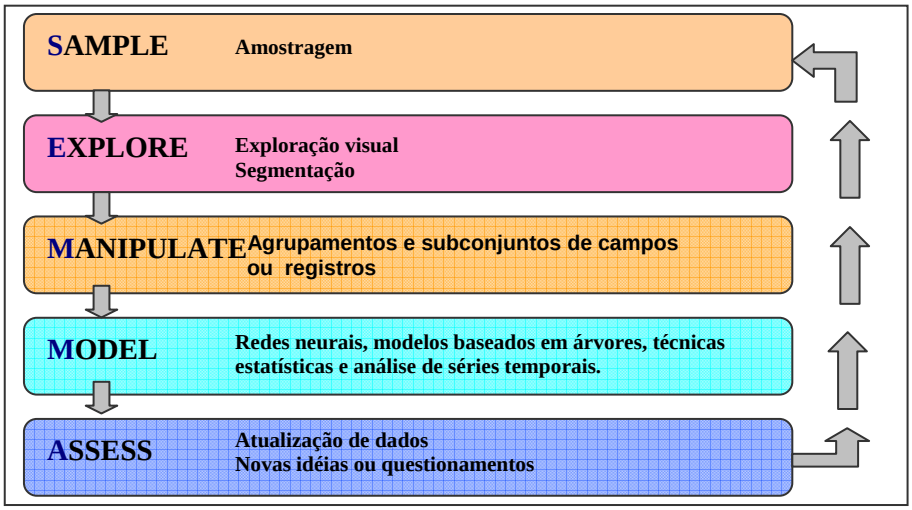

Figura 2.7: Metodologia SEMMA de mineração de dados [34]

As fases da metodologia SEMMA podem ser descritas da seguinte forma [2]:

- **SAMPLE:** envolve a criação de uma ou mais tabelas de dados para a estruturação dos dados, pois a amostra deve contemplar quantidade de dados suficiente para conter informação significativa para realização do processamento.
- **EXPLORE:** corresponde à exploração visual dos dados, a fim de encontrar previamente relacionamentos, tendências e anomalias, para adquirir conhecimento e idéias referentes à base de dados.
- **MANIPULATE:** envolve a criação, seleção e transformação das variáveis com base nas regras e filtros relacionados ao negócio, focando-se o processo de seleção específico que atenda ao modelo.
- **MODEL:** refere-se à utilização de ferramentas estatísticas, tais como regressão, árvores de decisão e redes neurais para encontrar padrões nos dados que facilitem a previsão dos dados de saída.
- **ASSESS:** corresponde à avaliação da utilidade e da veracidade das informações encontradas.

Observa-se que a metodologia SEMMA [34] contempla o processo de mineração, considerando aspectos como seleção da amostra (coleta), prévia exploração dos dados (pré-processamento), manipulação dos dados (transformação), modelagem (extração) e atualização (interpretação e avaliação).

De acordo com [45], o processo de mineração de dados corresponde a um processo cíclico e iterativo definido pelas seguintes etapas (conforme mostrado na figura 2.8):

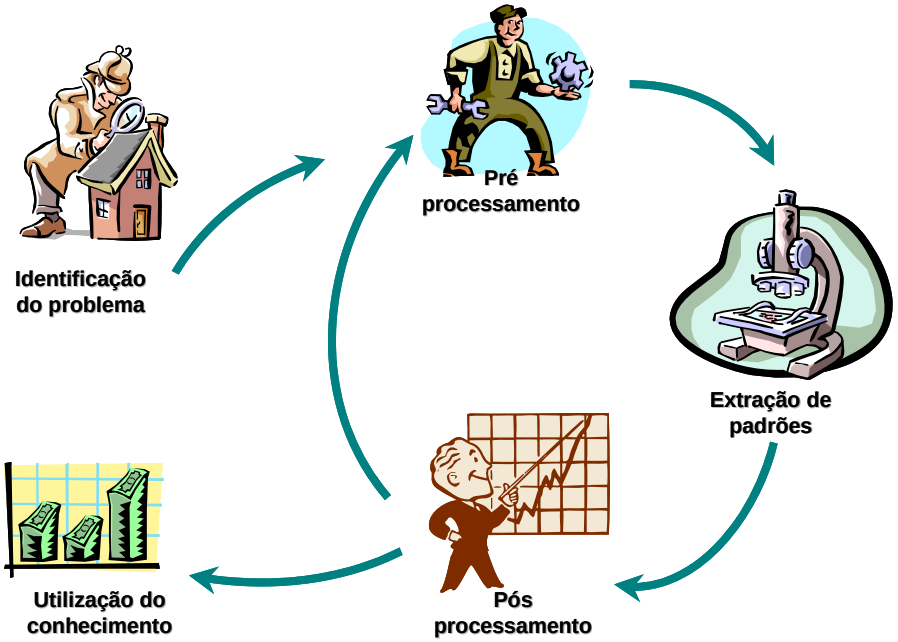

Figura 2.8: Etapas do processo de mineração de dados [45]

1. **Identificação do problema**: compreende as atividades de definição de metas, de objetivos e de restrições, além da obtenção do conhecimento sobre o domínio dos dados, considerado um pré-requisito para a extração de informações úteis para o processo decisório.
2. **Pré-processamento**: contempla as fases de seleção, coleta, integração, transformação, limpeza e redução dos dados.
3. **Extração de padrões**: direciona-se ao cumprimento dos objetivos atingidos na identificação do problema, abrangendo desde a escolha da tarefa (técnica a ser utilizada para mineração de dados), a escolha do algoritmo até a extração dos padrões propriamente dita.
4. **Pós-processamento**: corresponde às atividades de avaliação e interpretação dos resultados, através da medição da compreensão e do interesse gerados pelo conhecimento descoberto, além de outras medidas de desempenho como precisão, erro, confiança, sensitividade,

especificidade, cobertura, suporte, satisfação, velocidade e tempo de aprendizado.

5. **Utilização do conhecimento**: não necessariamente uma atividade, mas parte do processo como resultado positivo da extração de conhecimento, a utilização do mesmo no processo organizacional caracteriza um bom processo de mineração de dados, além de justificar sua existência.

Observa-se que os processos de mineração de dados definidos pelos pesquisadores são muito semelhantes, apresentando apenas pequenas variações, mas que não os diferenciam de forma significativa. Sendo assim, o processo de mineração de dados detalhado anteriormente (com as fases de identificação do problema, seleção e coleta de dados, preparação de dados, extração de padrões e avaliação e interpretação do conhecimento) servirão de base para a análise dos estudos de casos, detalhados a seguir, no capítulo 3.

2.5 Processo de Engenharia de Software na mineração de dados

De acordo com a metodologia de desenvolvimento de sistemas citada por Rezende [46], o processo de Engenharia de Software subdivide-se nas seguintes etapas: estudo preliminar, análise do sistema atual, projeto lógico, projeto físico e projeto de implantação. Destas fases, as duas primeiras podem ser aproveitadas e adaptadas para o processo de mineração de dados:

- Estudo preliminar: corresponde à etapa inicial do projeto, responsável pelo planejamento, contemplando as seguintes subfases:
    - Seleção de equipe;
    - Identificação de diretrizes e necessidades;
    - Detalhamento dos requisitos funcionais;
    - Definição de estratégias do sistema atual;
    - Aprovação do estudo preliminar.
- Análise do sistema atual: corresponde à etapa posterior ao estudo preliminar, responsável pelo mapeamento da situação atual, contemplando as seguintes subfases:
    - Revisão do estudo preliminar;
    - Identificação do ambiente atual;
    - Diagramação do sistema atual;

- Definição da estratégia do projeto lógico;
- Aprovação da análise do sistema atual.

A Engenharia de Software [28] abrange um conjunto de três elementos fundamentais: métodos, ferramentas e procedimentos, possibilitando o controle do processo de desenvolvimento de software de alta qualidade produtivamente, têm-se as seguintes etapas, como mostrado na Figura 2.9.

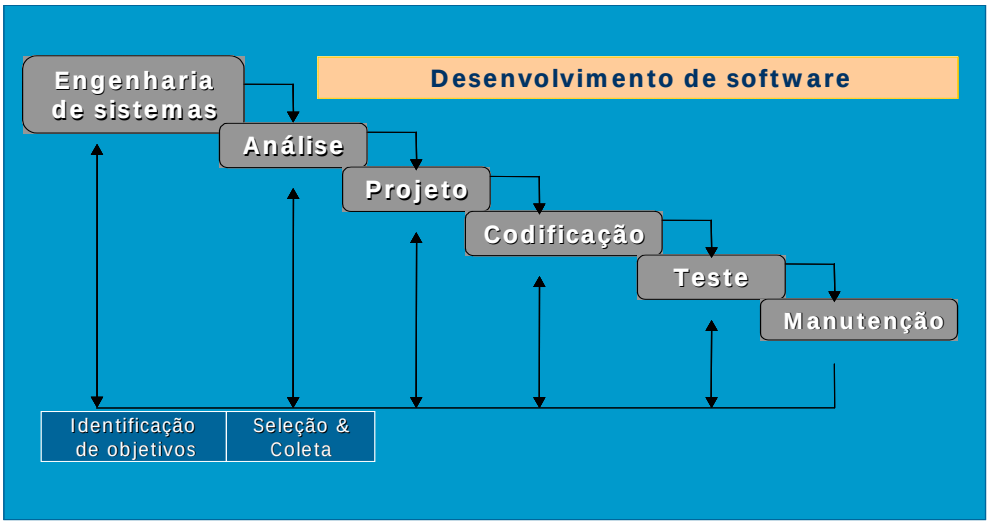

Figura 2.9: Processo de Engenharia de Software [28]

As fases de análise e **engenharia** de sistemas referem-se à coleta dos requisitos em nível de sistema, com uma pequena quantidade de projeto e análise de alto nível.

A **análise** de requisitos de software corresponde ao processo de coleta dos requisitos de forma mais intensificada e concentrada, especificamente no software, devendo compreender o domínio da informação para o software bem como a função, o desempenho e a interface exigidos. Esta fase deve ser documentada e revista com o cliente.

Já o **projeto** de software e, de fato, um processo de múltiplos passos que se concentra em quatro atributos distintos do programa: estrutura de dados, arquitetura de software, detalhes procedimentais e caracterização de interface.

A etapa de **codificação** refere-se à tradução do programa em linguagem legível à máquina.

O processo de realização de **testes** concentra-se nos aspectos lógicos internos do software, garantindo que todas as instruções tenham sido testadas, e concentram-

se também nos aspectos funcionais externos, ou seja, realizando testes para descobrir erros e garantir que a entrada definida produza resultados reais que concordem com os resultados exigidos.

Por fim, a fase de **manutenção** refere-se à realização de alterações sistêmicas, a fim de contemplar mudanças ocorridas ou a correção de erros identificados. Esta etapa reaplica cada uma das etapas precedentes do ciclo de um software existente.

Traçando-se um paralelo comparativo entre a Engenharia de Software e o processo de mineração de dados, observa-se que as fases de identificação de objetivos e seleção e coleta de dados podem contar com o auxilio do processo de Engenharia de Software, através da adequação das fases de analise e engenharia de sistemas e de analise de requisitos de software. As demais fases são especificas do processo de desenvolvimento de sistemas, tornando-se não aplicável à realidade de mineração de dados, conforme destacado na tabela 2.1.

Fase	Engenharia de Software	Mineração de dados
1	Análise e engenharia de sistemas	Identificação dos objetivos
2	Análise de requisitos de software	Seleção & Coleta de dados
3	Projeto	Não aplicável
4	Codificação	Não aplicável
5	Testes	Não aplicável
6	Manutenção	Não aplicável

Tabela 2.1: Quadro comparativo entre desenvolvimento de mineração de dados e Engenharia de Software

Ainda comparando os processos de mineração de dados e de Engenharia de Software, [48] destaca os seguintes aspectos distintos entre ambos, conforme mostra a tabela 2.2.

Vale destacar que, embora as fases iniciais sejam semelhantes, os focos de cada processo são bastante distintos, pois a Engenharia de Software foca na manipulação de informação, através do projeto e da programação do sistema, dedicando 30% do esforço em analise e 70% em programação.

O processo de mineração de dados, por sua vez, foca na escolha dos dados, através da busca e representação do conhecimento, dedicando 70% do esforço na fase de preparação dos dados e apenas 30% em modelagem e testes.

Engenharia de software	Mineração de dados
Foco na manipulação de informação.	Foco na escolha dos dados.
Foco no projeto e na programação do sistema.	Foco na busca e representação do conhecimento.
30% do esforço em análise e projeto e 70% em programação e teste.	70% do esforço na preparação de dados e 30% na geração do modelo e teste.
Prototipagem cara.	Prototipagem barata.
Desenvolvimento pouco iterativo.	Desenvolvimento iterativo.
Manutenção com intervenção humana.	Manutenção exige reaprendizado.

Tabela 2.2: Comparação entre processos de Engenharia de Software e mineração de dados.

Observa-se que a etapa de estudo preliminar revela semelhanças quando comparada à fase de identificação de objetivos do processo de mineração de dados, contemplando ainda a subfase de detalhamento dos requisitos funcionais, podendo ser aproveitada para complementar a metodologia de mineração.

Já a fase de análise do sistema atual pode ser comparada à etapa de seleção e coleta de dados no que se refere ao levantamento dos sistemas e dados envolvidos para a definição da amostra para análise. Aspectos como definição de estratégia do projeto lógico é relativamente específico da Engenharia de Software, porém, em mineração de dados, o projeto lógico pode ser entendido como o fluxo de dados necessário para obtenção dos dados para análise e extração de conhecimento.

Dessa forma, acredita-se que a maior contribuição da Engenharia de Software na mineração de dados corresponda à etapa de identificação de objetivos, devido à falta de uma estrutura formal de etapas a serem seguidas para a realização de um bom planejamento. Dada a maturidade da Engenharia de Software, a utilização das etapas de análise e engenharia de sistemas e análise de requisitos de software, ambas adaptadas à mineração de dados, contribuiria fortemente para o sucesso dos projetos de mineração de dados.

2.6 Plataformas de mineração de dados

Para uma mineração de dados efetiva, devem ser considerados as diversas plataformas de bancos de dados existentes e seus respectivos impactos na construção de um modelo de extração de conhecimento, que é uma das etapas do processo de mineração de dados.

Dentre as plataformas atualmente utilizadas, citam-se os diversos tipos de bancos de dados, armazéns de dados (*data warehouses*) e WEB.

2.6.1 Bancos de dados

De acordo com [3], as estruturas de bancos de dados podem ser classificadas em relacionais, transacionais e avançadas, conforme conceituação a seguir:

2.6.1.1 Relacionais

Entende-se por banco de dados relacional, um conjunto de tabelas, cada qual com nome único, consistindo de um conjunto de atributos (colunas ou campos) e, normalmente, armazenando uma grande quantidade de tuplas (registros ou linhas). Cada tupla em uma tabela relacional representa um objeto, identificado por chave única e descrito por um conjunto de valores de atributos.

Um modelo semântico de dados, assim como Modelo Entidade-Relacionamento, é usado para representar esquemas contendo as entidades e relacionamentos de um negócio. A partir de um esquema construído, derivam-se as tabelas que podem ser acessadas através de consultas escritas em linguagem relacional como o SQL ou com auxílio de interfaces gráficas de usuário (GUIs).

Quando a mineração de dados é aplicada a bancos de dados relacionais, os sistemas podem analisar os dados de clientes e prever o risco de crédito de novos clientes baseado em renda, idade e informações prévias de crédito; podem também detectar desvios, como itens cujas vendas estão longe das expectativas em comparação ao ano anterior, investigando-os posteriormente.

2.6.1.2 Transacionais

Um banco de dados transacional consiste de um arquivo onde cada registro representa uma transação, incluindo um número de identidade única da transação e uma lista de itens relacionados à transação. Pode haver tabelas adicionais associadas, contendo informações relativas à venda, como a data da transação, o

número do cliente, o número do vendedor e da filial. Sendo assim, os sistemas de mineração de dados podem, por exemplo, auxiliar na identificação de conjuntos de itens que são freqüentemente vendidos juntos.

2.6.1.3 Outros

As novas aplicações de bancos de dados, que requerem estruturas de dados eficientes e métodos escaláveis para a manipulação de estruturas de objetos complexas, com registros de tamanhos variáveis e dados semi-estruturados ou até não estruturados, contemplam esquemas de bancos de dados com estruturas complexas e dinâmicas [2, 3, 12], tais como:

- Bancos de dados orientados a objetos: são baseados no paradigma de orientação a objetos, onde cada entidade é considerada com um objeto. Objetos que compartilham um conjunto de características comuns podem ser agrupados em classes de objetos, podendo ser organizadas em hierarquias. Cada objeto é uma instância de sua classe e os objetos têm comportamentos implementados por métodos. Atualmente, há empresas que utilizam bancos de dados orientados a objetos e realizam a extração dos dados deste tipo de estrutura para a realização do processo de mineração de dados.

- Bancos de dados objeto-relacionais: baseados no modelo de dados objeto-relacionais. Este modelo referencia o modelo relacional incrementado por tipos de dados complexos e hierarquias de classes, tornando-se cada vez mais popular em indústrias e em aplicações. Para mineração de dados neste tipo de estrutura, deve ocorrer a extração e a transformação dos dados, respeitando-se os relacionamentos entre os objetos.

- Bancos de dados espaciais: contêm informações relacionadas ao espaço, incluindo bancos de dados geográficos, de projetos de engenharia, de imagem de satélite e de medicina. Possuem mapas n-dimensionais, que podem ser representados em formato de vetor, onde estradas, pontes, edifícios e lagos são representados como uniões de construções geométricas básicas, tais como pontos, linhas e polígonos.

- Bancos de dados temporais e seqüenciais: ambos armazenam dados relacionados ao tempo. Um banco de dados temporal normalmente contempla atributos relacionados ao tempo. Já um banco de dados seqüencial armazena seqüências de valores que mudam com o tempo, como os dados de um estoque. Técnicas de mineração de dados podem ser usadas para encontrar as características de evolução do objeto, ou a tendência de mudança dos objetos no banco de dados.

Estas informações podem ser úteis na tomada de decisões e no planejamento estratégico.

- Bancos de texto: contêm descrições de objetos. Estas descrições não são normalmente palavras-chaves, mas longas sentenças ou parágrafos, tais como especificações de produto, relatórios de erro, mensagens de aviso, notas ou outros documentos. Podem ser altamente desestruturados, como algumas páginas *Web*; semi-estruturados, como mensagens de e-mail e muitas páginas HTML/XML; enquanto outros são bem estruturados, como bibliotecas de bancos de dados, implementados usando sistemas de bancos de dados relacionais. Muitos estudos têm sido feitos recentemente na área de *Text Mining*, como em casos de *sites* de busca, tais como UOL *Miner*.

- Bancos de dados multimídia: armazenam dados de imagem, áudio e vídeo. Eles são usados em aplicações baseadas em fotos, sistemas de correio de voz, sistemas de vídeo por demanda e interfaces de usuário baseadas em reconhecimento de comandos de voz. Devem suportar grandes objetos que podem requerer gigabytes de armazenamento, assim como técnicas de busca e armazenamento especializados, devendo também estar integrados com os métodos padrões de mineração de dados.

2.6.2 Armazéns de dados (*Data Warehouses*)

Segundo [14, 26,29], um *data warehouse*, cuja tradução literal é armazém de dados, é um repositório de informações coletadas de múltiplas fontes, armazenadas sob uma estrutura unificada e que normalmente reside em um *site* único.

Pode ser definido como um banco de dados, destinado a sistemas de apoio à decisão e cujos dados foram armazenados em estruturas lógicas dimensionais, possibilitando o seu processamento analítico por ferramentas especiais (OLAP e data mining) e contemplando na sua construção os processos de limpeza, transformação, integração, carga e atualização periódica dos dados, de forma a facilitar o processo de coleta e extração de dados, tornando-os confiáveis e disponíveis para o processo de mineração de dados, conforme ilustrado na figura 2.10.

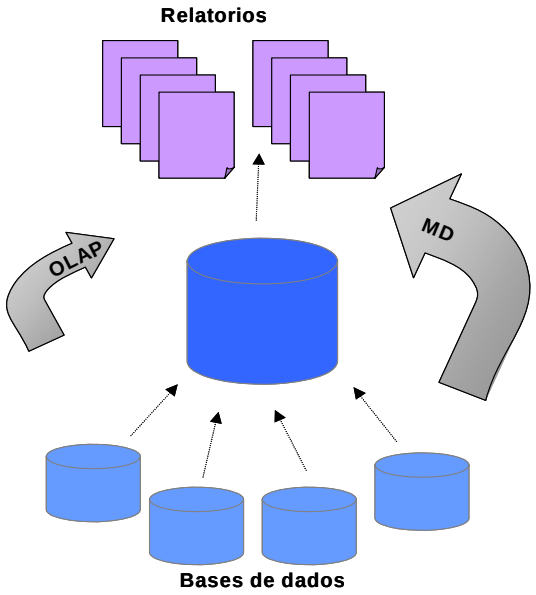

Figura 2.10: Relação entre base de dados, *data warehouse* e mineração de dados (MD) [45]

Normalmente, um armazém de dados é modelado através de uma estrutura multidimensional onde cada dimensão corresponde a um atributo ou conjunto de atributos no esquema e cada célula armazena os valores de algumas medidas agrupadas, assim como quantidade e valor de vendas. A real estrutura de *um data warehouse* pode ser um armazém de dados relacional ou um cubo de dados multidimensional, oferecendo uma visão multidimensional dos dados e permitindo o pré-processamento e acesso rápido a dados sumarizados. Cubos adicionais podem ser usados para armazenar somas agregadas sobre cada dimensão, correspondendo a valores agregados obtidos através do uso de diferentes agrupamentos SQL.

De acordo com [38], um armazém de dados é um sistema complexo que integra muitos componentes: diversos tipos de software e de hardware, redes de computadores, sistemas de comunicação de dados, servidores, mainframes e sistemas de administração de bancos de dados, como também muitas pessoas de diferentes unidades organizacionais, com objetivos diferentes. Porém, para os usuários, não importam os conceitos ou as definições da ferramenta, mas, sim, se ela consegue proporcionar o suporte necessário aos seus processos decisórios.

Como principais utilizações de um armazém de dados, citam-se [38]: armazenamento, extração, transformação, refinamento ou limpeza de dados, repositório de metadados, transferência e replicação e gerenciamento de consultas e relatórios.

Muitos profissionais e pesquisadores defendem a existência de um armazém de dados como condição básica para a obtenção de um processo de mineração de dados aceitável, pois a estrutura integrada de um *data warehouse* já pressupõe a realização de várias etapas do processo de *data mining*, tais como seleção, extração e limpeza dos dados, além de garantir a qualidade, a disponibilidade e a uniformidade dos conceitos das informações dentro da organização.

Como exemplos de ferramentas utilizadas pelas organizações para a implantação de um *data warehouse*, citam-se Business Objects, Powerplay (Cognos), Enterprise Guide (SAS), dentre muitas outras.

2.6.3 WEB

Segundo [3, 7, 11, 17], a Internet está crescendo rapidamente como um importante meio de transações comerciais assim como para a disseminação de informações relacionadas a uma vasta gama de tópicos (por exemplo, negócios, governo, lazer). De acordo com as previsões, a maioria das informações humanas estará disponível na Web em dez anos. Estas imensas quantidades de dados trazem um grande desafio de como tornar a Web o maior centro de utilidades de informações úteis.

Páginas de busca e diretórios como Yahoo! constituem o estado da arte em ferramentas para a obtenção de informações na Web atualmente. Ainda existem alguns problemas nas ferramentas de consulta disponibilizadas na Internet, tais como:

- abundância: o fenômeno de centenas de irrelevantes documentos que retornam em resposta a uma consulta de busca;
- cobertura limitada: de acordo com o domínio de busca de cada site, existe um certo grau de limitação na abrangência de informações;
- interface limitada de consulta: baseada em busca sintática orientada a palavra-chave, havendo necessidade de desenvolvimento de novos tipos de busca, de forma a ampliar os resultados gerados pela interface;
- customização individual limitada aos usuários: considerando-se a heterogeneidade do público que acessa os sites na rede, revela-se condição importante a customização individual desde os usuários mais leigos aos mais sábios no assunto de computação via Web.

Estes problemas, por sua vez, podem ser atribuídos às seguintes características da Web:

- a Internet é uma grande, diversa e dinâmica coleção de documentos interligados, mas 99% destes documentos não interessam a 99% das pessoas;
- exceto pelos links, a Web é enormemente desestruturada;
- a maior parte das informações constantes na Internet são em formato HTML, o que dificulta análise e extração de seu conteúdo;
- o conteúdo de muitas fontes na Web estão ocultos em interfaces de busca e, assim, não podem ser indexados.

A questão, portanto, é: como podemos transpor estes e outros obstáculos que impedem o processo de descoberta de recursos na Internet. Felizmente, técnicas novas e sofisticadas estão sendo desenvolvidas na área de mineração de dados, podendo auxiliar na extração de informação útil da Web. Algoritmos de mineração de dados têm sido mostrados para abranger melhor grandes conjuntos de dados e têm sido aplicados com sucesso em diversas áreas como em diagnósticos médicos, previsão do tempo, aprovação de crédito, segmentação de clientes, marketing e detecção de fraudes.

Dentre todas as plataformas citadas, as plataformas de bancos e armazéns de dados serão focadas, dado o escopo relativo ao ambiente corporativo, de forma interna, gerencial e estratégica.

2.7 Técnicas de mineração de dados

Nesta seção, são apresentadas as técnicas mais comumente utilizadas no processo de mineração de dados, buscando conceituá-las para análise dos casos práticos vividos pelas organizações.

2.7.1 Classificação

Segundo [2], classificação (ou abordagem de aprendizado supervisionado) é muito comum nos negócios, capaz de prover um detalhamento dos atributos em grupos específicos, mecanismo natural em mentes humanas que segmenta as coisas em grupos distintos. Este tipo de mineração de dados é chamado de aprendizado supervisionado porque existe o conhecimento das classes que o modelo pretende prever.

De acordo com [4], a classificação é uma das técnicas mais utilizadas na mineração de dados simplesmente porque é uma das tarefas cognitivas humanas mais realizadas no auxílio à compreensão do meio ambiente. A tarefa de classificar normalmente exige a comparação de um objeto ou dado com outros dados ou objetos que supostamente pertençam a classes anteriormente definidas. Para comparar dados ou objetos, utiliza-se uma métrica ou forma de medida de diferença entre eles.

Já segundo [3], a classificação é o processo de encontrar um conjunto de modelos (ou funções) para descrever e distinguir classes e conceitos. Neste caso, diferentemente da caracterização e da discriminação, as classes resultantes são desconhecidas; por exemplo, as faixas de renda de uma população para a confecção de um Modelo de renda presumida são diferentes das faixas de renda de clientes para uma campanha de marketing específica. Estas faixas são descobertas ao longo da análise dos dados.

Vale ressaltar que a classificação realiza-se em torno de variáveis discretas e contínuas, devendo respeitar o foco de cada análise para ser realizada, ou seja, no caso citado anteriormente das faixas de renda, as classes a serem criadas devem seguir o escopo do estudo, se este for um Modelo de renda presumida, a classificação deve corresponder às faixas de risco de crédito.

Como algoritmos típicos para a classificação citam-se as árvores de decisão e as redes neurais. Uma árvore de decisão é uma estrutura gráfica de fluxo de dados onde cada nó revela um teste em um valor de atributo, cada ramo representa um resultado do teste e os galhos da árvore representam as distribuições das classes, podendo ser facilmente convertidas em regras de classificação. Já uma rede neural, quando usada para classificação, é tipicamente um conjunto de neurônios que processam unidades com conexões entre elas.

2.7.1.1 Árvores de decisão

Por definição de [3], uma árvore de decisão é uma estrutura gráfica de fluxo de dados em formato de árvore, onde cada nó corresponde a um teste de valor de atributo, cada ramo representa o resultado do teste e cada galho reflete a distribuição das classes. Árvores de decisão podem ser facilmente convertidas para regras de classificação.

Árvores de decisão têm sido utilizadas em muitas áreas de aplicação desde medicina até teoria de jogos e negócios. Elas baseiam-se em vários sistemas comerciais de indução de regras.

Árvores de decisão expressam uma forma simples de lógica condicional. Um sistema de árvore de decisão, simplesmente, divide uma tabela em tabelas menores pela seleção de subconjuntos baseados em valores de um atributo dado, através de técnicas recursivas de particionamento.

Uma boa ferramenta baseada em árvore de decisão permite que o usuário explore a árvore de acordo com sua vontade, do mesmo modo que ele poderá encontrar grupos alvos que lhe interessem mais, ampliando o dado exato associado ao seu grupo alvo. Os usuários podem, também, selecionar os dados fundamentais em qualquer nó da árvore, movendo-o para dentro de uma planilha ou outra ferramenta para análise posterior.

Os algoritmos de análise de classificação em árvores mais comumente usados são: ID3, C4.5, CHAID e CART [2, 3]. Já como exemplos de ferramentas existentes no mercado, citam-se: Alice Disoft, HyperParallel//Discovery, Business Objects Business Miner, Data Mind, Angoss Knowledge Seeker, Answer Tree [2, 3, 4].

2.7.1.2 Redes Neurais

Redes neurais são extensivamente utilizadas no mundo dos negócios como modelos preditivos. Em particular, a indústria de serviços financeiros utiliza bastante as redes neurais para modelar a fraude em cartões de crédito e transações monetárias.

Estruturalmente, uma rede neural consiste em um número de elementos interconectados (denominados neurônios) organizados em camadas que aprendem pela modificação da conexão conectando as camadas. Geralmente, as redes neurais constroem superfícies equacionais complexas, através de interações repetidas, cada hora ajustando os parâmetros que definem a superfície. Depois de muitas repetições, uma superfície pode ser inteiramente definida que se aproxima muito dos pontos dentro do grupo de dados.

O processo tende a imitar o funcionamento de um neurônio em um cérebro humano. As redes neurais aprendem com a experiência histórica e são úteis na detecção de relacionamentos desconhecidos entre um conjunto de dados de entrada e saída. Como outras abordagens, redes neurais detectam padrões nos dados, generalizam relacionamentos encontrados nos dados e fazem previsões. Elas têm sido especialmente observadas pela habilidade de previsão em processos complexos.

Redes neurais utilizam realimentação e são extremamente populares devido à sua relativa simplicidade e estabilidade. A propagação reversa é uma regra de treinamento usada em redes de realimentação.

Dessa forma, as redes neurais são as técnicas preferidas para a realização de estimativas ou saídas numéricas contínuas populares no mercado financeiro e na indústria. Hoje, alguns dos fornecedores de aplicações que utilizam redes neurais são: IBM, SAS, SPSS, HNC, Angoss, RightPoint, Thinking Machines e Neo Vista. O Falcon, da HNC, é um produto de rede neural utilizado na detecção de fraude no mercado financeiro, de forma que grande parte dos cartões de crédito da América tem sido analisados pela empresa de software HNC [19, 35].

Por outro lado, existem deficiências em redes neurais. Primeiramente, elas têm sido criticadas como sendo úteis para previsão, mas não para o entendimento do modelo, visto que as primeiras implementações de redes neurais foram caracterizadas como “caixas pretas”. Todavia, atualmente, as novas ferramentas têm melhorado neste aspecto.

2.7.2 Associação

Segundo [2], associação refere-se à análise de informações úteis ao negócio que podem ser unidas através de associações de agregação entre diferentes itens vendidos em catálogos ou em lojas de varejo; podem ser físicas ou virtuais. As aplicações que utilizam associação incluem marketing cruzado, armazenamento de layout, projeto de catálogos, análise de perda de liderança, promoção e precificação de produtos, possibilitando a inferência da preferência dos clientes, de acordo com seus padrões de compra.

As atividades típicas da associação referem-se a fatos que ocorrem simultaneamente com probabilidade razoável (co-ocorrência) ou a itens de uma massa de dados que estão presentes juntos com uma certa chance (correlação). Dos números obtidos da aplicação desta técnica, podem-se extrair regras que regem o consumo de alguns itens.

Dessa forma, a associação ou análise de afinidade, também conhecida como técnica de Market Basket Analysis [35], acontece quando as ocorrências estão ligadas num único evento, como no exemplo da rede Wal Mart, onde as cervejas e as fraldas apresentavam correlação significativa para os consumidores homens.

Dessa forma, o processo de mineração de dados utiliza a técnica de associação a fim de maximizar o retorno de ações, tais como oferta de produtos e

serviços, a fim de otimizar os resultados, com base na compreensão de padrões de comportamento dos clientes.

2.7.3 Segmentação / Clusterização

Clusterização ou segmentação é o método de agrupamento de registros de dados que compartilham tendências e padrões semelhantes, ou seja, é o processo de divisão dos conjuntos de dados em grupos distintos.

Conforme [2], clusterização (ou aprendizado não supervisionado) é um método de agrupamento de linhas de dados ou registros que compartilham padrões e tendências, de forma que os membros de cada grupo sejam o mais homogêneo quanto possível e os grupos entre si seja o mais heterogêneo quanto possível. Também conhecido como segmentação, é o processo de divisão de conjuntos de dados em grupos distintos, auxiliando na determinação de quais registros podem estar juntos.

Estudos de segmentação são denominados de aprendizado não supervisionado, pois não se tem conhecimento prévio dos grupos resultantes.

Como exemplo de aplicação de segmentação, citam-se as ações de venda de produtos destinadas a públicos alvos diferenciados, de forma a oferecer produtos específicos para públicos com maior propensão para aquisição. Estas ações derivam de um estudo de segmentação, a fim definir melhor o foco das vendas, reduzindo custo e aumentando a taxa de retorno do investimento.

2.7.4 Estimativa / Previsão

De acordo com conceituação de [4], estimar algum índice é determinar seu valor mais provável diante de dados do passado ou de dados de outros índices semelhantes sobre os quais se tem conhecimento. Sendo assim, a arte de estimar é exatamente esta: determinar da melhor forma possível um valor baseando-se em outros valores de situações idênticas, mas nunca exatamente iguais.

Segundo [2], consiste na utilização de dados históricos, a fim de prever comportamentos futuros, buscando padrões e tendências existentes nos dados, pois

os mesmos tendem a se repetir ao longo do tempo. Também conhecida como Raciocínio Baseado em Memória (em inglês, Memory Based Reasoning) [35], esta técnica direciona-se à coleta de dados para análise da experiência anterior para obtenção de similaridade em um novo registro, utilizando-os para classificar e fazer previsões.

Dentre as técnicas utilizadas para estimar e prever grandezas, podem ser destacados: a estatística, as redes neurais artificiais e os algoritmos genéticos.

2.7.4.1 Estatística

A Estatística [2, 3, 4, 16, 19] tem sido usada por muito tempo para a criação de modelos de conjuntos de dados. Serve de base para todas as outras tecnologias criadas para a mineração de dados. Conceitos como distribuição normal, variância, análise de regressão, desvio simples, análise de conjuntos, análises de discriminantes e de intervalos de confiança são utilizadas para realizar as pesquisas nos dados, bem como para analisar e descobrir relacionamentos entre os mesmos.

Dentre os principais conceitos estatísticos, cita-se o processo de regressão linear, um processo que utiliza probabilidade, análise de dados e inferência estatística, através da avaliação do relacionamento existente entre a variável dependente e as variáveis independentes.

De acordo com [3], como principais conceitos da Estatística utilizados na mineração de dados, observa-se:

- Regressão: métodos utilizados para predizer o valor de resposta de uma variável dependente com relação a uma ou mais variáveis independentes. Existem várias formas de regressão: linear, múltipla, ponderada, polinomial, não-paramétrica e logística.
- Modelo linear generalizado: este modelo permite a uma variável resposta categorizada (como por exemplo, renda alta, média ou baixa) relacionar-se a um conjunto de variáveis independentes de forma similar para a modelagem de uma variável resposta numérica utilizando-se regressão linear. Modelos lineares generalizados contemplam regressão logística e de Poisson.
- Análise de variância: esta técnica analisa dados experimentais para duas ou mais populações descritas por uma variável resposta e uma ou mais variáveis categóricas (classificadas como no caso da renda: alta, média ou baixa, por exemplo). Normalmente, um problema de análise de variância

envolve a comparação de n populações para determinar se no mínimo duas destas são diferentes.

- Análise de fator: este método é usado para determinar quais variáveis são combinadas na geração de um determinado fator. Por exemplo, para muitos psiquiatras não é possível medir um determinado fator de interesse, como a inteligência; porém, freqüentemente é possível medir outras quantidades como uma pontuação de um teste em estudantes, que reflete o fator de interesse. Neste caso, nenhuma das variáveis é definida como dependente.
- Análise discriminante: comumente utilizada em ciências sociais, esta técnica aplica-se na predição de uma variável resposta categorizada. Ao contrário de modelos lineares generalizados, ela assume que as variáveis independentes seguem uma distribuição normal multivariada. Esta análise tenta determinar funções determinantes, através da combinação linear de variáveis independentes, que discriminar entre os grupos definidos pela variável resposta.
- Séries temporais: existem muitas técnicas estatísticas para análise de séries temporais, tais como métodos de auto-regressão, modelagem integrada de média móvel (ARIMA) e modelagem de séries temporais de longa memória.
- Análise de sobrevivência: técnica estatística definida para predizer a probabilidade de ocorrência de determinado resultado, dadas certas características. Por exemplo, a probabilidade de um paciente submetido a um tratamento sobreviver em dado tempo mínimo.
- Analise de desvios: Um banco de dados pode conter objetos de dados que não correspondem com o comportamento geral ou com o modelo dos dados. Estes objetos de dados são denominados *outliers* (ou desvios), segundo [3].

  A maioria dos métodos de mineração de dados descarta ou despreza os desvios como “ruídos” ou exceções. Embora, em algumas aplicações tais como detecção de fraudes, os eventos raros podem ser mais interessantes do que as ocorrências regulares.

  Os desvios podem ser detectados através de testes estatísticos, que assumem um modelo de distribuição ou de probabilidade para os dados, ou de medidas de distância onde objetos que possuem distância significativa de algum outro grupo são considerados desvios. Porém,

melhor ainda do que a utilização de estatística ou de medidas de distância, métodos baseados em desvios identificam-nos através do exame das diferenças nas características principais de um grupo de objetos.

A análise de desvios pode descobrir o uso fraudulento de cartões de crédito através da detecção de compras de valores excessivamente altos para uma determinada conta quando comparadas às compras regularmente realizadas pela mesma. Estes valores também devem ser detectados com respeito à localização, tipo e freqüência.

2.7.4.2 Algoritmos genéticos

Algoritmos genéticos refere-se a um método de otimização combinatória baseado em processos de evolução biológica. A idéia básica é que, ao longo do tempo, a evolução seleciona as espécies mais bem adaptadas. Aplicando esta idéia à mineração de dados usualmente envolve otimização de um modelo de dados usando métodos genéticos para obter os modelos mais bem adaptados.

Segundo [2], hoje, existem muitas pesquisas na área de algoritmos genéticos e, de todas as técnicas de modelagem, esta aparenta ser mais bem entendida. As raízes de algoritmos genéticos começam com o trabalho de Darwin, a *Origem das Espécies*, em 1859. Em 1957, GEP Box escreveu *Operação Evolucionária*, que revela um método de aumento da produtividade industrial que teve influência na relação de algoritmos genéticos com problemas de negócio. Outros trabalhos influentes incluem a *Simulação de sistemas genéticos através da automação de computadores digitais*, de A.S. Fraser e a *Otimização através da evolução e recombinação* de H.J. Bremermann´s.

Algoritmos genéticos têm sido freqüentemente utilizados em conjunto com redes neurais e com a técnica de segmentação para a modelagem de dados. O processo inicia-se com o agrupamento randômico dos dados, criando-se três grupos distintos como um organismo. Os algoritmos genéticos terão o que é denominada função de adaptação que determina se um conjunto de dados corresponde a um dos três grupos. Esta função de adaptação identifica os conjuntos de dados que se adaptam melhor do que os outros. Quando conjuntos de dados são lidos, eles podem ser avaliados pela função de adaptação para ver quão bem eles relacionam os outros elementos de dados em um segmento.

Algoritmos genéticos têm operadores que permitem a cópia ou alteração da descrição de grupos de dados. Estes operadores imitam a função encontrada na natureza onde a vida se reproduz, muda e evolui. Se um registro em um conjunto

adapta-se através da função, então ele sobrevive e é copiado para o segmento. Se, por sua vez, um registro não se adapta, então ele pode ser cruzado com outro conjunto de dados ou, em outras palavras, pode ser unido a outros segmentos para criar uma melhor adaptação, podendo ocorrer alterações para criar adaptação mais otimizadas quando novos conjuntos de dados são lidos.

Sendo assim, algoritmos genéticos solucionam problemas complexos que outras tecnologias têm maior urgência para concretizá-los. A principal característica refere-se às propriedades existentes na função de adaptação que permitem a convergência para erros mínimos. Eles têm sido freqüentemente utilizados em conjunto com redes neurais para alcançar um alto nível de entendimento enquanto as redes neurais gravam grupos de entrada de variáveis que impactam diretamente o banco de dados, proporcionando maiores detalhes de documentação de cada modelo de rede neural. Após a experimentação de vários modelos, um modelo final pode ser construído através da leitura dos modelos atuais de conjuntos de variáveis.

Como técnica mais comumente usada em análise de risco para realização de previsões e estimativas tem-se o conceito de séries temporais que se refere à análise de evolução de dados ao longo de um período de tempo [3]. Esta análise descreve e modela as regularidades ou tendências para os objetos cujas mudanças de comportamento ocorrem ao longo do tempo, de forma a utilizar conjuntamente técnicas de associação, classificação e segmentação de dados relacionados ao tempo, identificando características de determinada análise, incluindo análise de séries temporais, seqüências ou correspondência de padrões dada uma periodicidade.

Neste caso, os eventos estão ligados ao longo do tempo. Por exemplo, na compra de uma nova geladeira, em 45% das vezes um novo fogão será comprado dentro de um mês e um novo refrigerador será comprado em duas semanas, em 60% das vezes.

No ambiente de análise de risco, a previsão é extensivamente utilizada para identificar tendências relativas a clientes inadimplentes, propensos ao cancelamento ou ao abandono na utilização dos produtos, clientes com maior probabilidade de retorno de uma mala direta de venda ou ativação de produto e ainda clientes com maior perfil de rentabilidade.

Existem muitas outras técnicas, porém não foram citadas, pois não correspondem ao escopo da presente pesquisa.

Capítulo 3: Ambiente corporativo e estudos de casos

3.1 Introdução

Este capítulo apresenta aspectos relativos ao ambiente corporativo bem como os casos de estudo, a fim de introduzir as situações práticas vivenciadas no processo de mineração de dados.

Cabe esclarecer que seu desenvolvimento real transcorreu de forma diferente da esperada, como explicado adiante. No entanto, para facilitar o entendimento procurou-se descreve-los na ordem cronológica dos passos da metodologia apresentada no capítulo 2.

3.2 Ambiente Corporativo

3.2.1 Estrutura Organizacional

Dentro desta administradora de cartões de crédito, existem muitos casos práticos de aplicação do processo de mineração de dados, tendo em vista o risco envolvido no negócio de concessão de crédito e a grande quantidade de informações disponíveis nos sistemas produtivos deste tipo de empresa. Porém, vale ressaltar, que outros tipos de negócio como bancos, seguradoras e indústrias também utilizam o processo de mineração de dados da mesma forma. Com isso, este trabalho pode ser aproveitado e aplicado para qualquer outro tipo de negócio, dado seu foco em metodologia de processo e não em área de interesse de descoberta de conhecimento.

A fim de diversificar a experiência obtida através dos casos práticos, foram selecionados cinco estudos de casos, cada qual pertencente a uma fase do ciclo de crédito. Esta diversidade facilita a avaliação, uma vez que os objetivos divergentes, remetem a utilização distinta do processo de mineração de dados, possibilitando ainda testar se a metodologia utilizada varia de caso a caso.

Como administração de cartões de crédito, compreendem-se todas as atividades relacionadas ao controle dos processos, políticas e procedimentos pertinentes ao ciclo de crédito, desde o planejamento do produto até a contabilização das receitas e perdas. Na figura 3.1, tem-se o fluxo do ciclo de crédito com suas principais fases:

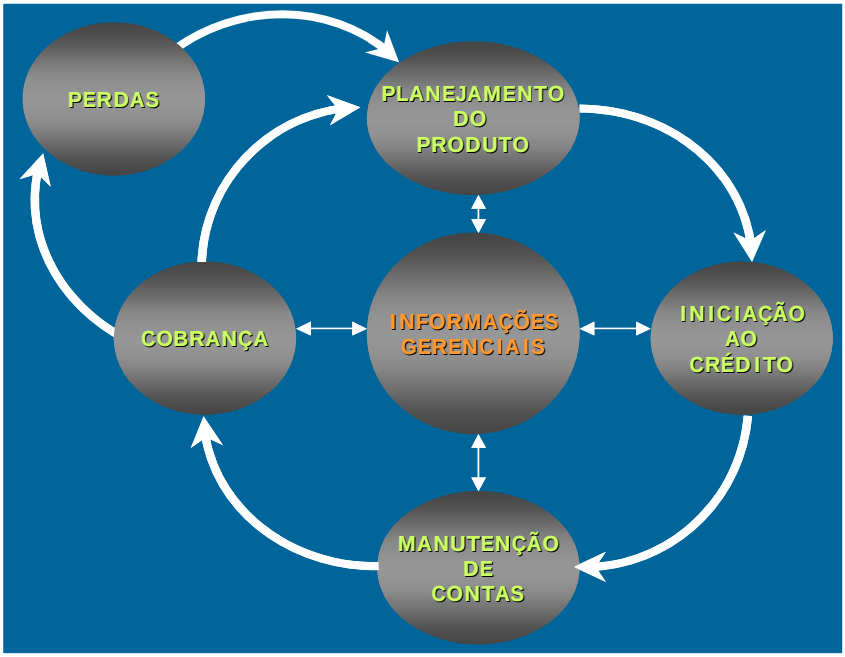

Figura 3.1: Fases do ciclo de crédito

- **Planejamento de produto:** esta etapa refere-se ao processo de definição e desenho dos produtos a serem oferecidos aos clientes. Nesta fase, as principais atividades são:
    - Análise das tendências de mercado, a fim de planejar e lançar produtos de acordo com as necessidades dos clientes e buscando um alinhamento com os produtos da concorrência.
    - Análise comportamental dos clientes, baseada na retroalimentação do ciclo de crédito, visando realizar ajustes e melhorias nos produtos novos ou já existentes.
    - Definição de processos, políticas e procedimentos pertinentes ao produto a ser planejado.
    - Levantamento e especificação de documentação necessária ao lançamento do produto bem como para todas as implantações sistêmicas referentes ao produto.

- **Iniciação ao crédito:** esta etapa contempla todo o processo de aquisição de novos clientes, seja através de envio de malas diretas ou de solicitações por parte do proponente. Nesta fase, as principais atividades são:
    - Análise e prospecção de novos clientes, através da avaliação de listas e cadastros.
    - Controle do processo de análise de crédito, do ponto de vista operacional, através da definição de políticas e procedimentos para a aprovação de novas contas.
    - Utilização de modelos estatísticos para previsão de riscos de inadimplência e fraude, a fim de otimizar o processo de análise de crédito, aumentando a precisão do processo decisório e reduzindo os custos de análise operacional das novas propostas.
    - Interface com áreas de vendas para garantir o alcance de metas com manutenção do risco de crédito, ou seja, possibilitando a venda de novos produtos a clientes efetivamente rentáveis para a empresa.

- **Manutenção de contas:** esta etapa corresponde a todos os processos envolvidos à atividade do cliente, desde sua aprovação até o seu cancelamento. Como principais atividades, citam-se:
    - Definição e controle dos processos de autorização de vendas, atendimento a clientes, manutenção de limites, renovação de contas, dentre outros.
    - Levantamento e estabelecimentos de políticas de autorização de vendas e manutenção de limites, de forma a satisfazer e fidelizar os clientes.
    - Interface com áreas de Atendimento a Clientes e Marketing, a fim de buscar melhorias ao processo.
    - Interface com parceiros (desde as empresas responsáveis pela aceitação dos cartões de crédito nos estabelecimentos, denominadas bandeiras, até as empresas responsáveis pela captura das transações junto aos estabelecimentos e respectiva transmissão aos sistemas internos da administradora, denominadas adquirentes), a fim de garantir o funcionamento e a disponibilidade dos sistemas, visando redução de atrito com os clientes.

- **Cobrança & Perdas:** nesta etapa, os clientes apresentam características de inadimplência. Dessa forma, as principais atividades são:
    - Definição e implantação de estratégias de cobrança, de modo a recuperar os ativos da companhia.
    - Controle dos processos operacionais, a fim de maximizar a produtividade e a efetividade dos cobradores.
    - Realização de análises de recuperação, visando o mapeamento dos clientes mais e menos propensos ao pagamento para a execução de ações de malas diretas ou vendas de carteiras de crédito em liquidação.

- **Sistemas de informações gerenciais:** esta etapa, embora definida como uma fase separada das demais, refere-se ao processo de elaboração e construção de sistemas de informações gerenciais em todas as fases do ciclo de crédito, a fim de suportar as decisões em cada etapa especificamente e de forma a possibilitar o conhecimento do da carteira de clientes como um todo. Refere-se ao ponto central da corporação, devendo considerar todas as informações, políticas e processos corporativos.

3.2.2 Utilização do Ciclo de Vida do Cliente

Além das fases do ciclo de crédito, outra forma de classificação dos estudos de casos refere-se à fase do ciclo de vida (do ponto de vista do cliente) a que fazem parte. Esta classificação ao longo da vida do cliente dentro da empresa denomina-se ***C**ustomer **L**ife **C**ycle,* sendo comumente utilizada nos conceitos de CRM *(**C**ustomer **R**elationship **M**anagement)*, conforme mencionado em [47]. As etapas do ciclo de vida do cliente são demonstradas na figura 3.2.

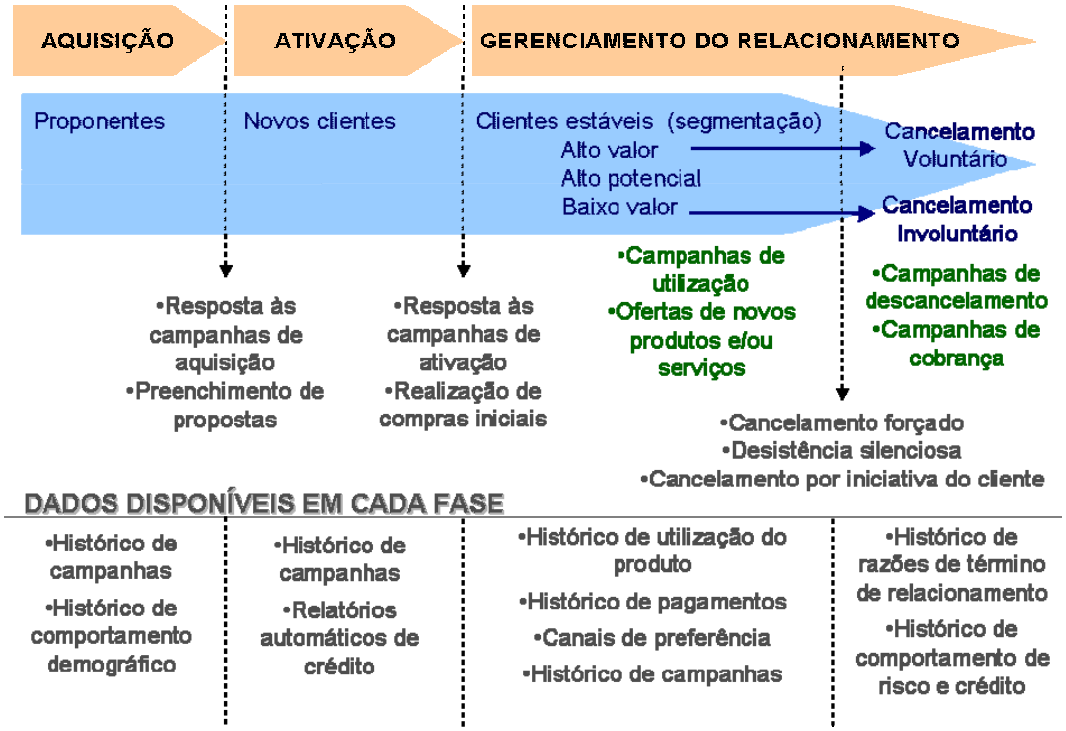

Figura 3.2: Etapas e eventos do ciclo de vida do cliente

A fase de aquisição refere-se à obtenção de novos clientes, através da oferta de produtos e serviços. Nesta etapa, o processo de mineração de dados auxilia na qualificação de potenciais clientes, através de modelos estatísticos baseados em histórico de comportamento de clientes atuais.

A fase de ativação contempla as campanhas de incentivo à utilização do produto (neste caso, cartão de crédito), através de análises de perfil de compra e de resposta a campanhas de ativação anteriores. No ambiente de análise de risco, esta atividade não ocorre, uma vez que a responsabilidade pela ativação dos produtos concentra-se em áreas de Marketing.

Por fim, a fase de gerenciamento do relacionamento (ou ***c****ustomer* ***r****elationship* ***m****anagement*, termo amplamente utilizado na atualidade) corresponde ao processo de gerenciamento do cliente de forma personalizada, através do acompanhamento histórico do relacionamento, visando evitar o cancelamento por parte do cliente (voluntário) e por parte da empresa por inadimplência (involuntário). Nesta etapa, o processo de mineração de dados apresenta grandes contribuições, uma vez que dispões de inúmeras ferramentas e técnicas para análise de dados históricos.

3.2.3 Considerações sobre os casos práticos

Houve algumas dificuldades no detalhamento dos casos práticos, pelo fato dos mesmos não estarem devidamente documentados nem terem seguido uma ordem lógica baseada na metodologia de mineração de dados.

Em um ambiente corporativo desta administradora de cartões de crédito bem como em outros tipos de empresas, muitos projetos são realizados de forma rápida e urgente para atender demandas de curto prazo, não respeitando adequadamente fases de planejamento e documentação, pertinentes a qualquer projeto, seja ele de desenvolvimento sistêmico ou de lançamento de um novo produto. Esta falta de planejamento e documentação resulta em inúmeros retrabalhos bem como dificulta a continuidade dos projetos, no que diz respeito à sua manutenção.

Para detalhamento dos casos práticos constantes neste trabalho, foram realizadas pesquisas diretamente com os profissionais envolvidos em cada estudo, para se obter a descrição das etapas seguidas para a realização do processo de mineração de dados. Neste contexto, observa-se um alto grau de dependência aos profissionais envolvidos em cada caso, o que se reflete diariamente nos projetos da empresa, podendo até inviabiliza-los no caso de desligamento de funcionários, por exemplo.

Muitas vezes, os próprios envolvidos não se recordavam com detalhes das informações importantes para a realização do estudo de caso, tornando o levantamento demorado e até impreciso, devido à dependência direta da memória do profissional.

Mesmo assim, a organização dos casos práticos a seguir, foi feita de acordo com as fases do processo de mineração de dados e sendo assim houve um tratamento prévio dos estudos de casos, a fim de padronizá-los para facilitar a avaliação e o entendimento.

Este aspecto ressalta a importância de um planejamento efetivo e de uma documentação detalhada em todos os estudos detalhados, pois a metodologia de mineração de dados avaliada neste trabalho somente tem sua validade quando utilizada de forma correta e disciplinada, caracterizando o amadurecimento da empresa no que diz respeito à condução de projetos.

3.3 Estudos de Casos

Nesta seção, são apresentados os estudos de casos vivenciados no ambiente de administração de cartões de crédito que servem de base para todo o trabalho de pesquisa. De acordo com a seção anterior, os casos práticos referem-se a projetos de melhoria e desenvolvimento tecnológico e processual, buscando otimizar a análise e o gerenciamento do risco de crédito, focando em aumento de rentabilidade, manutenção dos riscos e custos, dentre várias decisões estratégicas para o progresso do negócio.

Em visitas realizadas em empresas especializadas na prestação de serviços de consultoria em mineração de dados, observou-se que os casos mais solicitados pelas empresas (na maioria, instituições financeiras e seguradoras) referem-se a modelos estatísticos para previsão de cancelamento, de inadimplência e de resposta a malas diretas e catálogos de produtos e serviços.

No negócio de cartões de crédito, os casos práticos também são, em grande parte, de modelos estatísticos para previsão de risco, receita, resposta e cancelamento. Mas existem muitos outros estudos e projetos realizados que podem ser classificados como projetos de mineração de dados, pois utilizam a metodologia para extrair novas estratégias e políticas para a empresa.

A estrutura sistêmica que suporta o ambiente corporativo desta administradora de cartões de crédito é composta basicamente por todos os sistemas produtivos em plataforma alta. As bases de dados são extraídas diretamente do Mainframe, pois não existe nenhum armazém de dados que atenda as demandas de análises específicas para análise de risco.

Sendo assim, todos os estudos de casos apóiam-se numa infra-estrutura tecnológica, conforme mostra a figura 3.3. Esta estrutura compõe-se de vários sistemas produtivos relacionados entre si.

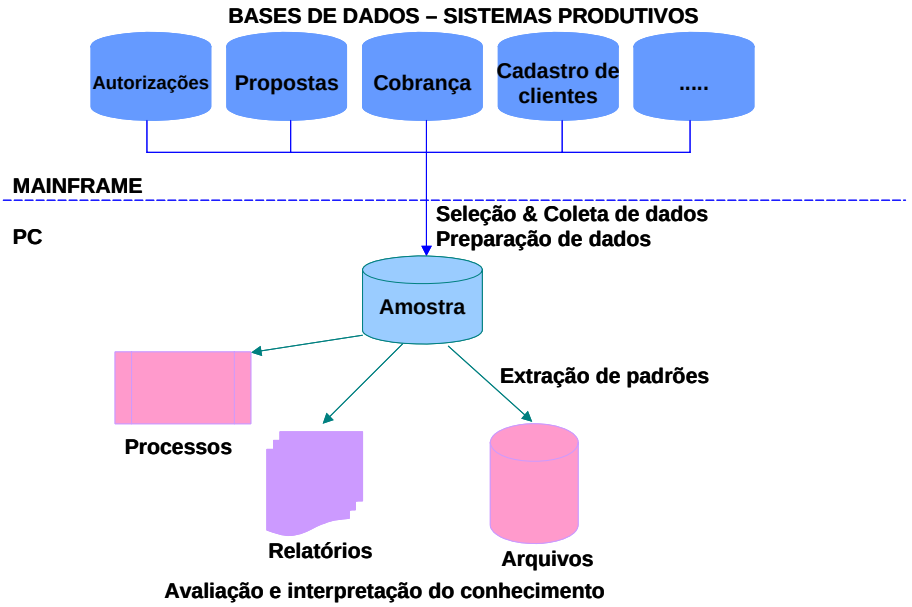

Figura 3.3: Estrutura sistêmica do negócio de cartões de crédito.

Atualmente, existem trabalhos destinados à definição de um extrator único de dados, contemplando todos os tratamentos e cruzamentos realizados entre as tabelas dos sistemas, a fim de minimizar a incidência de informações divergentes entre as diversas áreas da empresa e de automatizar o processo de extração e de tratamentos dos dados utilizados em comum por todas as áreas da empresa.

Os estudos de casos analisados neste trabalho são detalhados de acordo com o processo de mineração de dados, definido no capítulo anterior, a fim de padronizar os conceitos, buscando a existência de divergências e deficiências, através da comparação entre os mesmos.

Foram selecionados cinco estudos de casos, a fim de avaliar as divergências existentes na metodologia de mineração de dados, de acordo com focos de negócio e técnicas aplicadas distintamente. O objetivo deste trabalho também corresponde ao teste desta metodologia teórica sob diversas óticas, para garantir que uma estrutura de processo única garanta um bom projeto de mineração de dados, independentemente da área de atuação da empresa, do foco de negócio, da técnica e das ferramentas utilizadas.

Na prática, os casos não seguiram rigorosamente as fases do processo de mineração de dados, ocorrendo de uma forma mais intuitiva e até desordenada

quando comparada à metodologia teórica. Porém, para facilitar o entendimento e a padronização dos conceitos discutidos sobre o processo de mineração de dados, os casos práticos são descritos sob o foco das fases da metodologia teórica de mineração de dados.

As etapas do processo de mineração de dados utilizadas para detalhamento dos casos práticos são:

1. Identificação dos objetivos;
2. Seleção e coleta de dados;
3. Preparação de dados;
4. Extração de padrões;
5. Interpretação e avaliação do conhecimento.

Dessa forma, os casos práticos são detalhados de acordo com as fases citadas acima, considerando-se as particularidades de cada um, uma vez que as fases são estendidas ou suprimidas, de acordo com as necessidades identificadas ao longo do processo.

Os casos práticos analisados neste trabalho são classificados na tabela 3.1 de acordo com o grau de importância de cada fase para a análise do processo de mineração de dados a que se propõe este trabalho, variando de 1 a 5, de maneira que a fase de maior pontuação em cada caso prático representa a fase de maior contribuição na avaliação do processo de mineração de dados.

Esta classificação tem como objetivo facilitar a leitura dos casos práticos, buscando focar os aspectos mais relevantes de cada um, do ponto de vista da avaliação do processo. A classificação abaixo se baseia no grau de contribuição de cada fase na avaliação do processo realizada por este trabalho.

Estudo de caso	***Identificação de objetivos***	***Seleção e coleta***	***Preparação de dados***	***Extração de padrões***	***Interpretação e avaliação***
Sistema de informações gerenciais de autorização	5	3	4	1	2
Previsão de recebimento de clientes inadimplentes	2	5	4	3	1
Análise de perfil de cadastro	3	1	5	2	4
Modelo de renda presumida	2	3	4	5	1
Sistema de prevenção de fraudes em autorizações	1	2	4	3	5

Tabela 3.1: Grau de importância das fases do processo na avaliação dos estudos de casos

Outro aspecto considerado para o detalhamento dos estudos de casos refere-se às etapas do ciclo de vida do cliente dentro da companhia, conforme descrito na figura 3.2. Como o processo de mineração de dados tem sua atuação principal em CRM (***c****ustomer* ***r****elationship* ***m****anagement*), é importante salientar a utilização deste processo como facilitador na identificação de padrões e manutenção do relacionamento existente entre o cliente e a administradora de cartões de crédito, tendo em vista o diferencial competitivo obtido através de tratamento unificado e personalizado por cliente.

Dessa forma, existem dois casos práticos correspondentes à fase de aquisição e três pertencentes à fase de gerenciamento de relacionamento. Não existem estudos referentes à fase de ativação, pois esta atividade concentra-se em áreas de Marketing e não de análise de risco.

Os casos práticos da fase de aquisição referem-se à etapa de busca de novos clientes no mercado, através de campanhas de vendas de produtos e serviços para públicos pré-analisados (como no caso Análise de Perfil de Cadastro) e de análise de crédito por escores (como no caso Modelo de renda presumida).

Já na fase de Gerenciamento do Relacionamento, existem dois casos correspondentes à manutenção do cliente, através da administração das compras realizadas pelo mesmo (como nos casos Sistemas de Informações Gerenciais de Autorização e Sistema de Prevenção de Fraudes em Autorizações), e, um caso correspondente à administração de carteira de crédito em liquidação, quando o cliente tem seu cartão de crédito cancelado pela administradora devido à inadimplência.

Porém, como o foco do trabalho é o processo de mineração de dados, os casos práticos serão detalhados de acordo com a ordem das fases do processo de mineração de dados.

3.3.1 Sistema de informações gerenciais de autorização

O primeiro estudo de caso refere-se ao projeto de desenvolvimento de um sistema de informações gerenciais para atendimento de demanda interna para tomada de decisões por parte da diretoria da empresa, no que se refere ao processo de autorização de vendas.

Dentro desta administradora de cartões de crédito, o foco de autorização corresponde aos processos que envolvem a utilização do cartão em estabelecimentos comerciais para pagamento via fatura na data de vencimento. Trata-se de um

processo com alto nível de complexidade, pois envolve vários sistemas situados em vários pontos diferentes (estabelecimento, emissor e bandeira), conforme fluxo detalhado na figura 3.4:

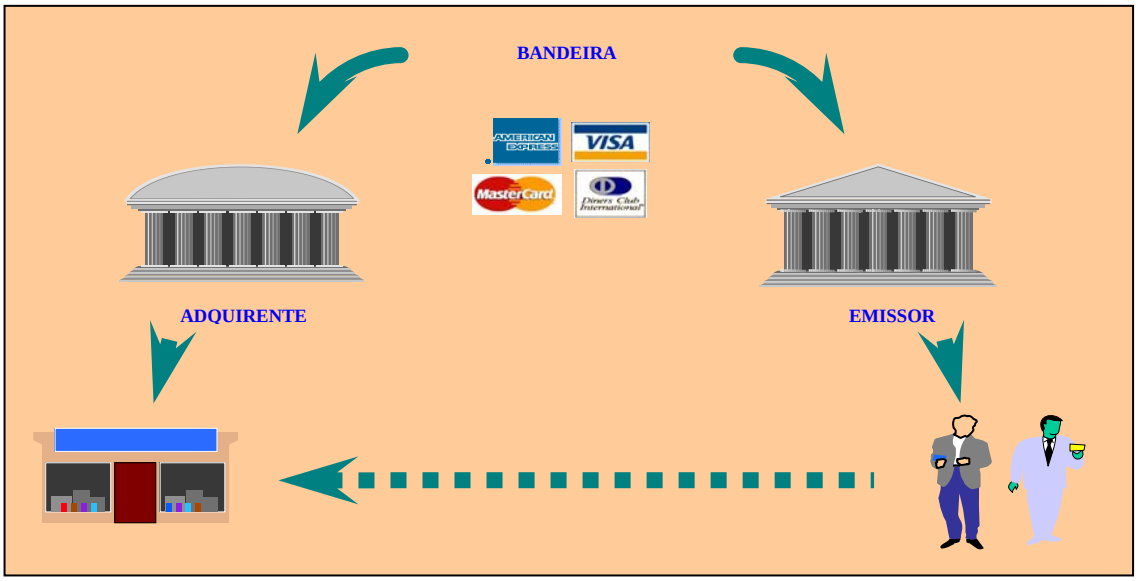

Figura 3.4: Fluxo do processo de autorização

Para desenvolvimento de um sistema de informações gerenciais que atendesse e suportasse o fluxo de dados descrito na figura 3.3, os seguintes passos do processo de mineração de dados foram executados:

3.3.1.1 Identificação dos objetivos:

De acordo com a plataforma de informações existentes, o desenvolvimento de um sistema de informações gerenciais envolve um processo de mineração de dados, principalmente no que diz respeito ao tratamento e à visualização dos dados de forma a facilitar o processo de tomada de decisões.

As atividades realizadas nesta fase são as seguintes:

- Definição do escopo do estudo: refere-se à realização de um sistema de informações gerenciais de autorização para auxiliar a tomada de decisões. Este sistema de informações gerenciais baseia-se numa estrutura de banco de dados em MS Access, sendo que as informações necessárias para a geração de relatórios alocam-se em tabelas.
- Entendimento dos limites do estudo: devido ao grande volume de autorizações recebidas, o estudo limita-se a avaliar o processo de autorização, contemplando ainda análises relacionadas ao comportamento dos clientes. O estudo deve ser realizado de forma sumarizada (não

analítica por autorização) devido à limitação da ferramenta utilizada como banco de dados.

- Escolha de bons estudos a serem realizados: para suprir as necessidades de informação para tomada de decisão, estudos volumétricos das autorizações decididas bem como estudos comportamentais baseados nas autorizações foram definidos. Dentre os bons estudos, citam: análise de clientes que se tornaram inadimplentes, após terem autorizações aprovadas; análise de clientes que cancelaram seus cartões, após terem autorizações negadas; ou ainda, estudos de hábitos de consumo de clientes.
- Determinação dos elementos corretos para o estudo: para viabilizar a realização dos estudos pertinentes ao processo de autorização, os elementos corretos para o estudo são os dados das autorizações e os dados dos clientes.
- Entendimento da amostra: faz-se necessário o conhecimento do domínio das variáveis, a fim de validar a confiabilidade das mesmas para utilização no estudo bem como a adequação do conteúdo das variáveis às regras de negócio. Neste caso, foi extraída uma versão prévia da amostra para a realização de uma análise exploratória dos dados, visando o conhecimento do conteúdo dos mesmos.

3.3.1.2 Seleção e coleta de dados:

Utilizando-se da estrutura sistêmica descrita na figura 3.3, o desenvolvimento de um sistema de informações gerenciais de autorização parte do sistema de autorizações e do cadastro de clientes para a obtenção da amostra.

A fase de seleção e coleta de dados é responsável pela definição dos dados a serem selecionados e coletados para a amostra do estudo, sendo extraídos através de programas de computador localizados no Mainframe ou no micro-computador. Neste estudo de casos, os dados são extraídos e tratados na alta plataforma, através da linguagem Cobol e de bancos de dados DB2.

As atividades realizadas nesta fase são as seguintes:

- Definição das variáveis a serem utilizadas e dos sistemas envolvidos: as variáveis selecionadas para a amostra foram dados cadastrais dos clientes (tipo de pessoa: física ou jurídica, tempo de relacionamento com a administradora, pontuação no escore comportamental, dias de atraso, percentual de excesso ao limite de crédito), dados da autorização (tipo: à

vista, parcelada com juros, parcelada sem juros, saque; decisão, local, ramo de estabelecimento: alimentação, vestuário, etc; valor). Estas variáveis são extraídas do sistema de autorizações e do cadastro de clientes.

- Estabelecimento dos sistemas a serem acessados: para seleção e coleta dos dados necessários para o desenvolvimento do sistema de informações gerenciais de autorização, os sistemas a serem acessados são: o sistema de autorizações e o cadastro de clientes, ambos produtivos no Mainframe.
- Verificação da adequação dos dados na descrição do problema: na análise exploratória dos dados verificou quais dados são relevantes para o desenvolvimento de um sistema de informações gerenciais apresentavam-se adequados para a descrição do problema, conforme descrito acima.
- Validação da consistência e redundância dos dados disponíveis: observou-se que existiam autorizações com valores errados, resultantes de problemas de digitação do valor da autorização pelo estabelecimento. Com isso, definiu-se a alteração do valor de todas as autorizações com valores superiores a US$ 60 mil para US$ 1, pois dessa forma, os valores errados são eliminados, não distorcendo o estudo e não perdendo o registro da autorização.
- Estabelecimento da recência e do período dos dados a serem coletados: como o sistema de informações gerenciais deve ser atualizado diariamente, os dados devem possuir tal recência, sendo que o período dos dados deve ser acumulado dentro do sistema.
- Verificação dos cruzamentos necessários entre diferentes arquivos e/ou tabelas: os cruzamentos necessários referem-se aos sistemas de autorizações e cadastros de clientes, sendo que a chave do relacionamento é o número do cartão do cliente.

3.3.1.3 Preparação de dados:

Visando obter informações consistentes e corretas para a tomada de decisões, a fase de preparação de dados deste estudo, realizada de forma automatizada no Mainframe, através de programas em Cobol, contemplou as seguintes atividades:

- Limpeza dos dados: conforme mencionado na fase anterior, identificou-se a ocorrência de autorizações contendo valores imensamente grandes, decorrentes de erros de digitação. Para eliminar estes desvios, os valores distorcidos (acima de US$ 60 mil) foram substituídos por US$1.

- Integração dos dados: refere-se ao cruzamento das bases de autorizações com o cadastro de clientes, tendo como chave de relacionamento o número do cartão do cliente.
- Transformação dos dados: corresponde ao tratamento de campos para adequação às necessidades de informações, tais como:
  - Classificação dos ramos de estabelecimentos de acordo com oito grandes grupos com características semelhantes (alimentação, vestuário, turismo, artigos para casa e construção, automóveis, companhias aéreas, artigos escolares e lazer).
  - Classificação dos países de origem das autorizações de acordo com os nove grupos com maiores volumes de autorizações nos últimos doze meses, classificando as demais autorizações no grupo Outros.
- Redução dos dados: após o tratamento dos campos selecionados, a redução dos dados consiste na sumarização das autorizações de acordo com os campos definidos e com a totalização das quantidades e dos valores das autorizações classificadas em cada caso. Esta sumarização visa reduzir o tamanho da tabela a ser armazenada no sistema de informações, dada a limitação de 1 Gigabyte de tamanho de banco de dados Access e a melhoria de tempo de processamento.
- Extração dos dados: corresponde à transmissão do Mainframe para o PC das bases de dados previamente definidas, com todos os campos preparados e tratados.

3.3.1.4 Extração de padrões:

Neste estudo de caso, o objetivo principal refere-se à geração de informações gerenciais que orientem o processo de tomada de decisões. Dessa forma, utiliza basicamente recursos de visualização de dados e depende bastante da análise humana das informações geradas. Com isso, as atividades principais desta etapa são:

- Seleção de técnicas: neste estudo de caso, as técnicas selecionadas foram classificação, segmentação e estatística, ambas utilizadas na fase de preparação dos dados (agrupamento de ramos e países e análise de desvios, respectivamente).
- Escolha de ferramentas: para a geração de um sistema de informações gerenciais acessível à Alta Gerência, a escolha das ferramentas considerou o alto nível de portabilidade e compatibilidade do pacote Microsoft Office. Esta escolha teve como conseqüência à necessidade de sumarização das

bases, pois o Access não suporta bancos de dados com tamanho superior a 1 Gigabyte, levando à perda de informações analíticas por autorização e, ainda, identificando-se a necessidade de novos campos, haveria a necessidade de reprocessamento das bases de dados.

- Tratamento dos dados: corresponde à etapa de tratamento dos dados para aplicação de técnicas. Como neste caso, estes tratamentos foram realizados no Mainframe, não foi realizada neste ponto do processo.
- Análise humana: esta etapa refere-se à análise das informações geradas, através de relatórios, para a tomada de decisões. Os relatórios possuem interface gráfica, o que facilita a análise dos resultados obtidos, através do software Microsoft Excel.

3.3.1.5 Interpretação e avaliação do conhecimento:

Como qualquer sistema de informações gerenciais, sua função principal refere-se à análise humana das informações geradas para tomada de decisões, através de ações. As fases que contemplam esta etapa são:

- Avaliação dos resultados: após a geração de relatórios, os resultados obtidos pelo fornecimento de informações gerenciais foram, além da aquisição de conhecimento pertinente ao processo: implantação de nova política de autorização, identificação de problemas no processo de autorização decorrentes de inoperabilidade do sistema, dentre outras melhorias.
- Implantação: com a implantação do sistema de informações gerenciais de autorização, uma nova política de autorização foi implementada assim como novos controles do processo de autorizações referentes ao processo de contingência por falha no sistema produtivo. Dessa forma, o sistema de informações gerenciais possibilita a utilização das informações geradas para a tomada de decisões, através de ações.
- Documentação: para garantir a continuidade do sistema, a documentação foi desenvolvida contendo o detalhamento de todos os conceitos utilizados para o desenvolvimento do sistema de informações gerenciais, tendo em vista a dependência do projeto na experiência dos profissionais envolvidos.
- Acompanhamento: ocorre diariamente com o processamento das informações e com a geração dos relatórios.

3.3.2 Previsão de recebimento de clientes inadimplentes

O segundo estudo de caso corresponde a um projeto de análise de carteira de clientes inadimplentes que teve como principal objetivo identificar o perfil de recuperação dos mesmos, de forma a garantir uma previsão de preço de venda adequada à cessão de direitos.

Historicamente, observa-se que quanto maior o atraso decorrido entre a data do vencimento da fatura, menor é a probabilidade de recebimento do valor por parte do cliente. A carteira a ser vendida já teria passado por todo o processo de cobrança, não tendo seu valor recuperado. Dessa forma, torna-se interessante o processo de cessão, pois garante uma parte do valor e repassa a responsabilidade dos custos estruturais da continuidade da cobrança.

3.3.2.1 Identificação dos objetivos:

Tendo em vista a plataforma de informações existente e a urgência na geração de relatórios de previsão de recebimento de clientes inadimplentes, este estudo de caso envolve o processo de mineração de dados no que diz respeito à identificação de padrões de comportamento e à utilização de técnicas estatísticas para previsão baseada em dados históricos. As atividades realizadas nesta fase são as seguintes:

- Definição do escopo do estudo: refere-se à realização de um estudo de previsão de recebimento de clientes inadimplentes. Este estudo deve ser realizado através da extração de uma base de clientes inadimplentes para análise do recebimento histórico dos mesmos.
- Entendimento dos limites do estudo: corresponde a utilização de premissas para a realização do estudo, como por exemplo, a premissa de que o comportamento histórico dos clientes se repete ao longo do tempo.
- Escolha de bons estudos a serem realizados: para prever a recuperação de valores devidos por clientes inadimplentes, o recebimento histórico deve ser utilizado como parâmetro para análise, realizando-se avaliações de períodos distintos para evitar sazonalidades.
- Determinação dos elementos corretos para o estudo: para viabilizar a realização dos estudos pertinentes à previsão de recebimento de clientes inadimplentes, os elementos corretos são os dados cadastrais dos clientes e os dados dos acordos de pagamentos firmados durante o período de inadimplência do cliente.
- Entendimento da amostra: neste estudo de caso, o entendimento da amostra, realizado através da análise exploratória dos dados (que consiste

na análise do domínio de todos os campos bem como sua adequação à realidade do negócio), revelou muitas inconsistências no conteúdo das variáveis, em decorrência da deterioração do sistema produtivo de clientes em cobrança, construído há mais de vinte anos.

3.3.2.2 Seleção e coleta de dados:

Utilizando-se da estrutura sistêmica descrita na figura 3.3, a previsão de recebimento coleta dados do sistema de cobrança e do cadastro de clientes para obtenção da amostra. As atividades realizadas nesta fase são as seguintes:

- Definição das variáveis a serem utilizadas e dos sistemas envolvidos: as variáveis selecionadas para a amostra foram os dados cadastrais dos clientes (local de residência (unidade federativa), tipo de cartão de crédito (local, internacional, etc), tempo de relacionamento, data de aprovação da conta) e os dados de cobrança (saldo devedor, tempo de atraso, percentual pago após cancelamento).
- Estabelecimento dos sistemas a serem acessados: para seleção e coleta dos dados necessários para o desenvolvimento do estudo, os sistemas a serem acessados são: o sistema de cobrança e o cadastro de clientes, ambos produtivos no Mainframe. O sistema de cobrança é composto por várias tabelas referentes aos clientes inadimplentes: tabela de títulos vencidos, tabela de pagamentos, tabela de acordos e tabela de parcelas de acordos.
- Verificação da adequação dos dados na descrição do problema: neste caso, verificou-se que os dados selecionados eram adequados na descrição do problema, salvo algumas inconsistências descritas abaixo.
- Validação da consistência e redundância dos dados disponíveis: observou-se que campos, tais como saldo devedor não correspondia ao valor total do acordo firmado com o cliente, mesmo não havendo descontos. Inconsistências como estas foram identificadas em todo o estudo, devido à deterioração do sistema (que nem sempre atualiza todos os campos corretamente).
- Estabelecimento da recência e do período dos dados a serem coletados: para realização do estudo de previsão de recebimento de clientes inadimplentes, foram selecionados casos de clientes com atraso superior a 180 dias, sendo analisados 36 meses de aprovação (ou seja, contas aprovadas desde julho de 1996 a junho/1999).

- Verificação dos cruzamentos necessários entre diferentes arquivos e/ou tabelas: os cruzamentos necessários para este estudo referem-se aos sistemas de cobrança e cadastros de clientes, tendo como chave de relacionamento o número do cartão do cliente. Estes cruzamentos foram realizados no PC. O sistema de cobrança, por ter várias tabelas, também tem cruzamentos internos (entre as tabelas).

3.3.2.3 Preparação de dados:

Devido à deterioração do sistema de cobrança, a fase de preparação de dados consumiu grande parte dos esforços deste caso, principalmente na obtenção de variáveis consistentes para a análise. As atividades realizadas nesta fase, através de programas codificados em PC para a realização dos tratamentos necessários na base, são as seguintes:

- Limpeza dos dados: foram realizados filtros para a extração dos casos para análise, tais como atraso superior a 180 dias e saldo maior do que R$ 100,00 (muitos clientes tiveram saldos reduzidos de forma significativa devido às alterações monetárias ocorridas desde 1994).
- Integração dos dados: refere-se ao cruzamento das bases de cobrança com o cadastro de clientes, tendo como chave de relacionamento o número do cartão do cliente.
- Transformação dos dados: corresponde ao tratamento de campos para adequação às necessidades de informações, tais como:
  - Classificação dos clientes por faixa de atraso real, variando de 180 em 180 dias, tendo em vista a mudança de comportamento de pagamento em decorrência do atraso.
  - Classificação dos clientes em tempo de relacionamento no momento do cancelamento do cartão, tendo em vista que clientes que têm seus cartões cancelados por cobrança em menos de um ano apresentam menor probabilidade de recebimento. As faixas de tempo de relacionamento foram definidas da seguinte forma: de 0 a 12 meses, de 13 a 24 meses, de 25 a 36 meses, de 37 a 48 meses e acima de 49 meses.
- Redução dos dados: neste estudo de caso, não foi utilizada redução dos dados, pois os cruzamentos entre as tabelas foram realizados no PC (baixa plataforma).

- Extração dos dados: corresponde à transmissão dos dados do Mainframe para o PC, de forma analítica, ou seja, um registro por cliente, com todos os campos preparados e tratados. Estes tratamentos são realizados através da construção de programas em Cobol.

3.3.2.4 Extração de padrões:

Bastante semelhante ao estudo de caso de desenvolvimento de um sistema de informações gerenciais para o processo de autorização, este estudo de caso representa um estudo esporádico para previsão de recebimento de clientes inadimplentes. Com isso, a análise humana baseada na visualização das informações representa parte importante no processo de mineração de dados, contemplando as seguintes atividades:

- Seleção de técnicas: neste estudo de caso, as técnicas selecionadas foram classificação (faixas de atraso e tempo de relacionamento) e previsão baseada em dados históricos.
- Escolha de ferramentas: para o cruzamento das bases, a ferramenta escolhida foi o banco de dados Microsoft Access. Já para aplicação das técnicas de classificação e previsão foi escolhida a ferramenta SPSS, através da realização de sumarizações de tabelas para obter as melhores combinações de agrupamentos, dado o percentual de recebimento histórico. Dessa forma, identificou-se uma curva média de retorno, capaz de prever o potencial de recebimento nos próximos meses.
- Tratamento dos dados: corresponde à etapa de tratamento dos dados para aplicação das técnicas. Para os testes de classes baseados na previsão de recebimento, foram criadas novas variáveis com o conteúdo da classificação atribuída ao registro (faixa de atraso e tempo de relacionamento).
- Análise humana: esta etapa refere-se à análise das informações geradas em Microsoft Excel, não diferindo entre os estudos de casos e servindo de base para o processo decisório. Neste estudo, a experiência dos analistas de negócio contribui significativamente para a obtenção de melhores resultados baseados na análise histórica das informações, dada a inexistência de documentação pertinente ao processo e aos sistemas envolvidos.

3.3.2.5 Interpretação e avaliação do conhecimento:

Tendo como base à análise humana das informações geradas, a previsão de recebimento de clientes inadimplentes contempla as seguintes fases na etapa de interpretação e avaliação do conhecimento:

- Avaliação dos resultados: este estudo serve de base para ações de cobrança personalizadas para cada tipo de público, dado seu potencial de recebimento. Com isso, foram geradas campanhas de abordagem de clientes inadimplentes, concedendo descontos inversamente proporcionais ao potencial de recebimento previsto.
- Implantação: como se trata de um estudo esporádico, a implantação se dá a cada novo levantamento, através da amostra extraída em cada situação, ou seja, não existe um processo sistêmico produtivo, mas estudos de previsão de recebimento de clientes inadimplentes são bastante freqüentes em análise de risco.
- Documentação: não existe documentação desenvolvida para este estudo, dado a urgência e o pouco tempo disponível para a realização do mesmo. Assim como retardou a realização deste projeto, este aspecto agrava significativamente estudos posteriores, tendo em vista as facilidades de re-execução de um processo, quando o mesmo está devidamente documentado.
- Acompanhamento: os relatórios das campanhas de abordagem dos clientes inadimplentes serviram de acompanhamento para o estudo revelando sua assertividade, tornando-o assim um processo mais contínuo em análise de risco.

3.3.3 Análise de perfil de cadastro

O terceiro estudo de caso corresponde ao projeto de análise de cadastros, através da criação de um modelo estatístico, a fim de identificar clientes potenciais, de acordo com as variáveis fornecidas em cada arquivo.

Comumente conhecido como *Database Marketing*, este processo visa avaliar clientes ou potenciais clientes para futura abordagem de oferta de produtos e/ou serviços. Consiste na atividade responsável pelo tratamento, análise e atribuição de política a cadastros, visando aumento do volume de vendas com manutenção do risco de crédito.

3.3.3.1 Identificação dos objetivos:

Cada vez mais comum, o processo de campanhas de vendas de produtos e serviços, através de análise de bancos de dados, surgiu de forma freqüente e rotineira (uma a cada três dias) em áreas de análise de risco, a fim de classificar potenciais clientes.

Durante o processo, é realizada a carga do cadastro num sistema de banco de dados construído para armazenamento, tratamento e análise das informações contidas no arquivo recebido.

Dessa forma, identificou-se a necessidade de automatizar o processo de análise de perfil dos cadastros recebidos, principalmente no que se refere ao tratamento dos dados. A estrutura sistêmica proposta refere-se a um banco de dados com tabelas relacionadas, cada qual contendo informações pertinente a um cadastro, sendo que todas relacionadas com uma tabela de clientes pré-qualificados (com conhecimento prévio do comportamento).

O objetivo principal deste estudo refere-se à automação do tratamento dos dados recebidos nos cadastros e da tabela principal, a fim de facilitar o relacionamento entre as bases.

As atividades realizadas nesta fase são as seguintes:

- Definição do escopo do estudo: contemplar todos os tratamentos necessários para a qualidade de dados dos cadastros recebidos.
- Entendimento dos limites do estudo: refere-se às freqüentes alterações oriundas de novas tendências (campos, abreviaturas, etc.) vindas nos cadastros.
- Escolha de bons estudos a serem realizados: corresponde ao levantamento de todos os tratamentos realizados até o presente momento, a fim de categorizar a automação, contemplando a maioria das ocorrências de tratamento, dado um histórico de demandas de tratamentos (por exemplo, alteração de prefixos de telefones, nomes de ruas abreviados, etc).
- Determinação dos elementos corretos para o estudo: como não existe uma definição dos campos a serem trazidos em cada cadastro, os elementos corretos para o estudo são campos essenciais para uma análise de perfil, tais como dados cadastrais (estado civil, sexo, idade), dados demográficos (por exemplo, endereço residencial e comercial) e dados profissionais (por exemplo, profissão, tempo de emprego, salário).

- Entendimento da amostra: uma vez definidos os campos para análise, faz-se necessário o conhecimento do domínio das variáveis, a fim de validar a confiabilidade das mesmas para a utilização no estudo, através de uma análise exploratória dos dados. Neste estudo de caso, foi importante o entendimento dos tratamentos das variáveis (por exemplo, abreviaturas, consistências em datas, etc).

3.3.3.2 Seleção e coleta de dados:

Para garantir uma boa preparação dos dados numa estrutura automatizada de tratamento de cadastros de clientes, a fase de seleção e coleta de dados contempla as seguintes atividades:

- Definição das variáveis a serem utilizadas e dos sistemas envolvidos: neste estudo de caso, as variáveis selecionadas são dados cadastrais, dados demográficos e dados profissionais. Os sistemas envolvidos são os cadastros recebidos e o cadastro de clientes.
- Estabelecimento dos sistemas a serem acessados: para seleção e coleta de dados, o único sistema a ser acessado é o cadastro de clientes, em Mainframe, sendo que os cadastros externos recebidos são alocados em bancos de dados Access em baixa plataforma.
- Verificação da adequação dos dados na descrição do problema: uma vez que a análise de perfil dos cadastros já era realizada pelas variáveis disponíveis nestes sistemas, existe a adequação dos dados, pois o estudo pretende apenas automatizar o processo de qualificação de cadastros novos.
- Validação da consistência (domínio consistente e grau de preenchimento adequado dos campos) e redundância dos dados disponíveis: muitas vezes, os dados constantes nos cadastros de clientes são redundantes com os dados recebidos nos cadastros, porém o estudo pretende realizar a consistência destes dados, a fim de qualificar ainda mais os dados constantes no cadastro.
- Estabelecimento da recência e do período dos dados a serem coletados: os dados coletados referem-se a todos os cadastros recebidos nos últimos cinco anos e a recência (nível de atualização) é da chegada do cadastro.
- Verificação dos cruzamentos necessários entre diferentes arquivos e/ou tabelas: para a realização dos cruzamentos necessários entre todos os

cadastros e o cadastro de clientes, a chave de relacionamento é o CPF (número no cadastro de pessoas físicas).

3.3.3.3 Preparação de dados:

Esta etapa corresponde à fase principal de avaliação neste estudo de caso, devido ao conjunto de tratamentos realizados nos dados de forma automática (codificada), dada a diversidade de informações recebidas nos cadastros. As atividades realizadas nesta fase são as seguintes:

- Limpeza dos dados: corresponde à eliminação de registros referentes a Cpfs duplicados ou com códigos identificadores inválidos, com idade fora do perfil de abordagem (abaixo de 18 anos), nomes sem endereço, setor postal genérico, nomes com cheques sem fundo, nomes com solicitação prévia de não recebimento de malas diretas, funcionários da empresa.
- Integração dos dados: refere-se ao cruzamento do cadastro recebido com o cadastro de clientes, tendo como chave de relacionamento o número do CPF e o nome. Este cruzamento utiliza técnicas de atribuição numérica de valor ao nome para obtenção de cruzamento (também conhecida como *match code*). A partir deste cruzamento, são trazidas informações adicionais (comportamentais) para análise de perfil do cliente.
- Transformação dos dados: corresponde ao tratamento de campos para adequação ao padrão dos dados constantes do cadastro de clientes. São eles:
  - Correções de grafia em nomes e endereços, através de abreviaturas;
  - Acertos automáticos de endereços baseados em setores postais;
  - Acertos automáticos de prefixos de telefones;
  - Correções e consistências em datas inválidas;
  - Alteração de grafia de profissões e atribuição de código oficial
- Redução dos dados: nesta tabela, os dados são analíticos por cliente, não havendo redução dos dados.
- Extração dos dados: os dados extraídos correspondem aos dados inseridos no sistema automático para cruzamento com o cadastro de clientes.

3.3.3.4 Extração de padrões:

Uma vez realizado o cruzamento entre o cadastro de clientes e o cadastro recebido, é realizada a classificação dos clientes recebidos no cadastro para possíveis

abordagens via mala direta. Para isso, é necessário aplicar técnicas de classificação baseada em comportamento histórico dos clientes, a fim de inferir o comportamento dos clientes não encontrados. As atividades realizadas nesta fase são as seguintes:

- Seleção de técnicas: para identificação de clientes potenciais, as técnicas utilizadas são a classificação (de acordo com o perfil de risco do cliente) e a segmentação por produtos possíveis de serem ofertados a cada cliente.
- Escolha de ferramentas: para aplicação das técnicas de classificação e segmentação, as ferramentas escolhidas são Access (para cruzamento entre bases), SPSS (para classificação) e Excel (para visualização dos dados).
- Tratamento dos dados: corresponde ao tratamento de dados para a aplicação das técnicas selecionadas. Neste caso, existe apenas a criação das variáveis resultantes de classe de risco e de classe de produtos para oferta.
- Análise humana: neste estudo de caso, a análise humana é muito pequena, pois o resultado da análise de perfil de cadastro é enviado para a área de Vendas para a abordagem dos potenciais clientes. Dessa forma, são gerados arquivos de forma automática para envio às gráficas de mala direta.

3.3.3.5 Interpretação e avaliação do conhecimento:

Tratando-se de um processo automatizado que visa atender outras áreas, esta etapa contempla as seguintes atividades:

- Avaliação dos resultados: após envio do arquivo resultante para as áreas de Vendas, existe ainda a segmentação dos potenciais clientes por canal de venda de abordagem (mala direta ou *telemarketing*) para, finalmente, ofertar novos produtos e/ou serviços aos clientes.
- Implantação: nesta etapa, com a implantação de uma estrutura sistêmica que qualifique automaticamente os cadastros externos, a elaboração de malas diretas ou listas para oferta de novos produtos e/ou serviços funcionará de forma mais efetiva e rápida.
- Documentação: não existe documentação referente a este estudo de caso, inclusive pela confidencialidade deste processo.
- Acompanhamento: também não existem relatórios de acompanhamento do processo de análise de perfil de cadastro.

3.3.4 Modelo de renda presumida

O quarto estudo de caso refere-se ao projeto de construção de um modelo estatístico, baseado nos dados demográficos e profissionais informados na proposta para a obtenção de cartão de crédito, buscando otimizar o processo de aprovação de contas e reduzir o risco de inadimplência gerado pelos clientes aprovados.

A concessão de crédito ao consumidor caracteriza-se pelo alto risco de inadimplência gerado pelo não pagamento do valor emprestado ao cliente. Para minimizar este risco, as administradoras de cartões de crédito, assim como as demais instituições financeiras, utilizam ferramentas de previsão, capazes de antecipar a probabilidade de risco de determinados clientes, dado o comportamento histórico.

3.3.4.1 Identificação dos objetivos:

Visando realizar melhorias no processo de aprovação de novas contas, a fase de identificação dos objetivos para a construção de um modelo de renda presumida contempla as seguintes atividades:

- Definição do escopo do estudo: refere-se à criação de um modelo estatístico, capaz de inferir a renda dos proponentes, com base nas informações demográficas preenchidas na proposta para adesão ao cartão de crédito.
- Entendimento dos limites do estudo: dada a limitação de dados constantes da proposta de adesão ao cartão de crédito, o modelo estatístico deve ser construído tendo como base os dados com grau de preenchimento mínimo, tais como profissão, setor postal residencial, idade e sexo do proponente.
- Escolha de bons estudos a serem realizados: para a realização da inferência de renda dos proponentes, um bom estudo refere-se ao agrupamento dos proponentes, através de uma árvore de decisão, tendo a renda confirmada como variável dependente.
- Determinação dos elementos corretos para o estudo: dentre as variáveis disponíveis no sistema de propostas, os elementos corretos para o estudo foram os dados demográficos e profissionais dos proponentes.
- Entendimento da amostra: uma vez definidos os campos para a amostra, para conhecimento do domínio das variáveis bem como para a avaliação

do grau de preenchimento das mesmas, realiza-se um estudo denominado de análise bivariada, a fim de identificar quais variáveis da amostras são relevantes quando relacionadas com a variável dependente que se deseja prever (neste caso, a renda mensal do proponente).

3.3.4.2 Seleção e coleta de dados:

Utilizando-se da estrutura sistêmica descrita na figura 3.3, o desenvolvimento de um modelo estatístico para inferência de renda de proponentes requer as seguintes atividades:

- Definição das variáveis a serem utilizadas e dos sistemas envolvidos: conforme mencionado na fase de identificação de objetivos, as variáveis a serem utilizadas no estudo são os dados demográficos e profissionais dos proponentes, dados estes preenchidos na proposta de adesão ao cartão de crédito. O sistema envolvido para a seleção destes dados é o sistema de propostas.
- Estabelecimento dos sistemas a serem acessados: para seleção e coleta dos dados, o sistema a ser acessado é o sistema de propostas.
- Verificação da adequação dos dados na descrição do problema: de acordo com a análise exploratória dos dados realizada na fase de entendimento da amostra, verificou-se que as variáveis mais significativas para a inferência de renda dos proponentes eram: código de profissão, idade, sexo e setor postal residencial.
- Validação da consistência e redundância dos dados disponíveis: tendo em vista as variáveis selecionadas para a análise, verificou-se que as mesmas apresentavam consistência, através de um alto grau de preenchimento e da forte relação com a variável que se desejava prever (a renda mensal).
- Estabelecimento da recência e do período dos dados a serem coletados: foram selecionadas 50% das propostas processadas durante o primeiro semestre de 1997 para desenvolvimento do modelo, sendo que os 50% restantes seriam destinados para validação da fórmula resultante da aplicação da técnica de classificação em árvore de decisão.
- Verificação dos cruzamentos necessários entre diferentes arquivos e/ou tabelas: para a seleção e coleta dos dados para este estudo não foram realizados cruzamentos entre bases.

3.3.4.3 Preparação de dados:

Neste estudo de caso, a amostra não apresentou problemas de consistência nem necessitou de integrações e transformações, devido ao fato de pertencerem a um sistema único, sendo apenas realizada a extração da amostra definida na fase de seleção e coleta. Os tratamentos realizados nas variáveis ocorreram na subfase de tratamento de dados, dentro de Extração de Padrões, uma vez que teve seu foco principal na preparação dos dados para a aplicação da técnica selecionada.

3.3.4.4 Extração de padrões:

Neste estudo de caso, a fase de extração de padrões revelou-se essencial para a construção do modelo estatístico. As atividades desta etapa são as seguintes:

- Seleção de técnicas: a construção de um modelo estatística baseada em classificação, através de árvores de decisão, ocorre devido à facilidade de interpretação e implementação do resultado obtido no formato de árvore, principalmente no que se refere ao agrupamento de registros baseados em uma variável específica (neste caso, a renda mensal).
- Escolha de ferramentas: para a aplicação da técnica de classificação, através de árvore de decisão, as ferramentas escolhidas correspondem aos softwares SPSS (para geração das variáveis agrupadas) e Answer Tree (para a geração da árvore de decisão), pois o mesmo disponibiliza vários algoritmos de classificação, tais como CHAID, CART e EXAUSTIVE CHAID. Para a definição dos grupos, o algoritmo utilizado foi o CHAID, devido à forma de agrupamento baseado na distribuição esperada x observada, denominada de Chi-Square. Dessa forma, a distribuição das variáveis com relação à variável a ser prevista, consideraria a distribuição normal, respeitando os volumes de propostas ingressadas.
- Tratamento dos dados: para aplicação da técnica de classificação, através de árvores de decisão, alguns agrupamentos prévios foram realizados:
  - A variável idade foi agrupada de cinco em 5 anos, totalizando 14 classes, observando-se que 72% das propostas concentravam-se nas faixas de idade entre 21 e 40 anos, idade na qual grande parte da população realmente está exposta ao crédito no mercado financeiro.
  - A variável profissão foi agrupada em 39 grupos profissionais com características semelhantes, sendo que os maiores grupos foram:

funcionários públicos (28,5%), vendedores (14,1%) e serviços gerais (9,7%).

- A variável referente ao setor postal foi agrupada de acordo com as duas primeiras posições do setor postal, a fim de classificar as regiões com características sócio econômicas semelhantes. Dessa forma, a cidade de São Paulo foi dividida em zonas: sul, norte, leste, centro; as regiões do Estado em litoral, interior, norte e oeste. Já os demais estados foram agrupados da seguinte maneira: RJ, MG e Bahia, estados do Norte e Nordeste e estados do Centro-Oeste e Sul. Estes agrupamentos consideraram volume e comportamento da população.

- Análise humana: esta etapa refere-se à análise das informações geradas em cada iteração da árvore de decisão, sendo bastante importante, pois com base na análise humana foram definidos os 22 grupos finais para inferência de renda devido à semelhança existente na distribuição das variáveis com relação à renda mensal do proponente.

3.3.4.5 Interpretação e avaliação do conhecimento:

Para interpretação e avaliação do conhecimento obtido

- Avaliação dos resultados: para avaliar a eficiência dos grupos criados, realizou-se um teste comparativo com os 50% de propostas processadas (não selecionadas previamente na amostra), atribuindo-se uma faixa de renda de acordo com as variáveis independentes (idade, profissão, setor postal). Dessa forma, avaliou-se o quanto à renda presumida atribuída refletia a renda confirmada pelos proponentes, obtendo um nível de aproximadamente 80% de acerto.
- Implantação: corresponde à implantação das regras de classificação dos 22 grupos resultantes no sistema produtivo, a fim de classificar cada proposta, assim de seu ingresso no sistema de propostas. Dessa forma, as propostas classificadas automaticamente em cada um dos 22 grupos ficam isentas da comprovação de renda, através do envio do documento comprobatório, otimizando o processo de decisão de propostas.
- Documentação: esta etapa foi cumprida, através da elaboração de uma documentação que, inclusive, serviu de base para o presente trabalho. Porém, não havia documentações anteriores (dos sistemas e processos envolvidos), o que facilitaria grandemente o entendimento dos mesmos para a realização do projeto.

- Acompanhamento: anualmente, é realizada a validação do modelo que consiste na avaliação da eficácia do mesmo, de acordo com o comportamento atual dos clientes. Para este acompanhamento, são gerados relatórios a partir de um sistema de informações gerenciais automatizado além do reprocessamento do projeto com amostra atualizada de proponentes.

3.3.5 Sistema de prevenção de fraudes em autorizações

O quinto estudo de caso refere-se ao projeto para implantação de regras de detecção de fraudes em compras nos cartões de crédito, baseado no comportamento histórico das ocorrências de fraude.

De acordo com o fluxo de autorização, primeiramente é realizada a validação de critérios de negócio, tais como boletim de proteção e número de cartão. Após os critérios de negócio, o sistema de autorização aplica a política de crédito, a fim de verificar o atraso e o excesso da conta. Por fim, há a aplicação dos critérios do sistema de prevenção de fraudes em autorizações, a fim de verificar características de autorização fraudulenta.

Caso não sejam identificadas estas características, a autorização é aprovada. Por outro lado, a autorização, quando realizada por telefone, é transmitida para uma equipe de atendimento assistido, responsável pela confirmação de dados do cliente, visando certificar que o cliente está presente e é realmente o titular do plástico; quando via equipamento eletrônico do estabelecimento, é gerada uma mensagem no visor, solicitando o contato com o emissor, que está realizando a confirmação positiva.

Esta confirmação positiva de dados do cliente quando ocorre com sucesso, possibilita a utilização do cartão, podendo o estabelecimento solicitar a autorização novamente. Já, quando o cliente não entra em contato ou não confirma seus dados, ocorre o bloqueio temporário do cartão até que o contato ocorra com a devida confirmação de dados.

3.3.5.1 Identificação dos objetivos:

O sistema de prevenção de fraudes em autorizações tem como objetivo prevenir a aprovação de autorizações fraudulentas, através da geração de uma mensagem no painel do equipamento do estabelecimento, solicitando que o mesmo entre em contato com a central de atendimento da administradora para que seja

efetuada a confirmação positiva com o associado no momento da venda para confirmar se a compra está sendo realizada pelo mesmo no estabelecimento.

As atividades realizadas nesta fase são as seguintes:

- Definição do escopo do estudo: refere-se à definição de um processo estruturado que possibilita a definição de regras para detecção de autorizações suspeitas de fraude. Este processo baseia-se na análise histórica das transações fraudulentas, a fim de identificar características semelhantes como pistas para a prevenção de possíveis fraudes, de forma que toda nova autorização com tais características sejam questionadas ao titular do cartão por questão de segurança da administradora, do portador do cartão e do estabelecimento, evitando perda de receita.
- Entendimento dos limites do estudo: entendem-se como limitações do estudo o fato de que são necessárias autorizações fraudulentas anteriores para a detecção de padrões de fraude. Com isso, a análise comparativa de autorizações fraudulentas possibilita criação de regras para evitar novas fraudes.
- Escolha de bons estudos a serem realizados: a partir de autorizações com fraude confirmada pelos clientes, é possível realizar o rastreamento histórico de autorizações com o mesmo perfil.
- Determinação dos elementos corretos para o estudo: tendo como base as variáveis como local, horário, ramo de estabelecimento e valor da autorização torna-se possível à identificação de padrões de comportamento fraudulento.
- Entendimento da amostra: uma vez definidos os campos da amostra do estudo, faz-se necessário o conhecimento do domínio das variáveis, a fim de validar a confiabilidade das mesmas para a utilização no estudo bem como a adequação do conteúdo das variáveis às regras de negócio. Por se tratar de um sistema produtivo de alta criticidade, o sistema de autorizações possui todos os campos com conteúdo preenchido e consistente, cabendo ao analista apenas o domínio dos mesmos.

3.3.5.2 Seleção e coleta de dados:

Utilizando-se da estrutura sistêmica descrita na figura 3.3, o desenvolvimento de regras para detecção de fraude parte do sistema de autorizações e do cadastro de clientes com ocorrências de fraude. Dentre as atividades pertinentes à fase de seleção e coleta de dados, citam-se:

- Definição das variáveis a serem utilizadas e dos sistemas envolvidos: dentre as variáveis selecionadas para o estudo citam-se: tipo de ramo de estabelecimento (alimentação, vestuário, posto de gasolina, etc), país e estado de origem da autorização, tipo de cartão (local, internacional ou gold) e faixa de valor da autorização. Os sistemas envolvidos neste estudo são o sistema de autorização e o cadastro de clientes fraudulentos.
- Estabelecimento dos sistemas a serem acessados: sistema de autorizações e cadastro de clientes com ocorrências de fraude.
- Verificação da adequação dos dados na descrição do problema: de acordo com a análise exploratória dos dados, verificou-se que as variáveis selecionadas na amostra são suficientes para distinguir uma autorização fraudulenta das normais (sem fraude), devido à concentração de fraude em determinados ramos, países e tipos de cartão.
- Validação da consistência e redundância dos dados disponíveis: os campos selecionados na amostra apresentaram consistência relativa aos possíveis domínios (tipos de ramos, países, etc) e alto grau de preenchimento, tendo em vista o alto grau de criticidade do sistema produtivo de autorizações.
- Estabelecimento da recência e do período dos dados a serem coletados: para análise e detecção de autorizações fraudulentas, a recência e o período dos dados para amostra é de quatro meses, dada a dinâmica do processo de fraude.
- Verificação dos cruzamentos necessários entre diferentes arquivos e/ou tabelas: como os sistemas envolvidos neste estudo são o sistema de autorizações e o cadastro de clientes fraudulentos, o cruzamento necessário entre estas bases se faz através da chave de número do cartão do cliente.

3.3.5.3 Preparação de dados:

Para garantir a confiabilidade das informações pertinentes à análise das autorizações com suspeita de fraude, alguns aspectos foram considerados na etapa de preparação de dados, tais como:

- Limpeza dos dados: neste estudo de caso, são utilizadas autorizações contendo todos os campos preenchidos. Dessa forma, as autorizações contendo variáveis de domínios inválidos ou nulos (o que é praticamente inexistente) são excluídas da análise.

- Integração dos dados: refere-se ao cruzamento entre a base de autorizações com o cadastro de clientes fraudulentos, a fim de criar um campo para identificação das autorizações com fraude.
- Transformação dos dados: como os dados devem ser tratados apenas para a aplicação das técnicas, neste momento, não ocorrem transformações nos dados.
- Redução dos dados: após o tratamento dos campos para classificação das autorizações, é realizada a sumarização da tabela de autorizações do período de acordo com todas as combinações geradas pelos conteúdos dos campos agrupados, totalizando quantidade e valor da autorizações e incluindo o campo de identificação de fraude (S/N).
- Extração dos dados: a extração dos dados ocorre através da transmissão do arquivo analítico por autorização do Mainframe para o PC para a realização de todos os tratamentos nos dados e a geração de uma tabela sumarizada. Estes tratamentos são realizados através da execução de um programa codificado em SQL dentro do banco de dados Access.

3.3.5.4 Extração de padrões:

Com o objetivo principal de identificar padrões de comportamento fraudulento em autorizações, esta etapa corresponde à fase mais importante do processo de mineração de dados deste estudo de caso. As atividades relacionadas à extração de padrões são:

- Seleção de técnicas: para análise de tendências de autorizações fraudulentas em cartões de crédito, realiza-se a classificação em fraude x não fraude, a segmentação de grupos com características semelhantes no que diz respeito à probabilidade de fraude e a análise de desvios (uma vez que cada autorização fraudulenta normalmente representa um desvio no comportamento usual do cliente), a fim de identificar padrões de comportamento, ou seja, concentrações de fraudes em ramos e países específicos.
- Escolha de ferramentas: para analisar as distribuições sumarizadas de autorizações são utilizados os softwares Microsoft Access (para cruzamento entre bases de autorização e cadastro de fraude), SPSS (para classificação dos campos de acordo com volumes e semelhanças de fraude) e Microsoft Excel (para análise gráfica dos resultados finais e

definição de pontos de corte para rejeição de autorizações com alto volume de ocorrências de fraude).

- Tratamento dos dados: para sumarização da tabela de autorizações, são utilizados os campos citados na definição da amostra, porém os mesmos passam por um processo de classificação e segmentação, como por exemplo:
  - Origem (país) da autorização: quando o país é Brasil, a autorização é classificada como nacional caso contrário internacional.
  - Ramos de estabelecimento: é realizado o mesmo agrupamento citado no estudo de caso de desenvolvimento de um sistema de informações gerenciais de autorização, utilizando-se oito grandes grupos de ramos.
  - Faixa de valor: são utilizados grupos de valor de autorização, de acordo com análise de volume de autorizações.
- Análise humana: neste estudo de caso, a análise humana refere-se ao processo de avaliação dos resultados gerados pela tabela sumarizada, a fim de identificar agrupamentos com grande incidência de fraude (no mínimo 10%).

3.3.5.5 Interpretação e avaliação do conhecimento:

Por fim, visando transformar a análise em ações para a redução de perdas de fraude da empresa, as seguintes atividades são realizadas:

- Avaliação dos resultados: consiste na definição de novas regras de prevenção de fraudes no processo de autorização, dado os padrões observados através do processo de mineração de dados detalhado acima. Quanto maior a incidência de fraudes em determinados grupos, maior a probabilidade da geração de regras de prevenção no sistema produtivo de autorizações, a fim de reduzir o volume de perdas oriundas de transações fraudulentas.
- Implantação: as novas regras criadas são introduzidas no sistema produtivo de prevenção de fraudes, através de um software denominado Falcon Expert (da empresa HNC). Além de seus recursos baseados em algoritmos de redes neurais, este sistema possibilita a inserção de regras, personalizando o processo de detecção de fraudes, de acordo com a dinâmica e com as tendências de cada mercado.
- Documentação: corresponde à geração de um relatório de eficácia das regras implantada, contendo a comparação do faturamento impactado

pelas autorizações negadas pelas regras x perdas evitadas de fraude. Mesmo assim, a política é revisada a cada quatro meses, dada a dinâmica do processo de fraude.

- Acompanhamento: o monitoramento das regras implantadas se dá a cada quatro meses com a validação do percentual de acerto (baseado nas contas com fraude) e da revisão do estudo.

3.4 Conclusão

Neste capítulo, foram detalhados os estudos de casos vivenciados no negócio de cartões de crédito, de acordo com o processo de mineração de dados definido teoricamente. Com este detalhamento, pode-se conhecer um pouco mais da aplicação do processo de mineração de dados em uma administradora de cartões de crédito no que se refere aos processos de concessão de crédito ao consumidor, autorização de vendas em estabelecimentos, detecção de fraudes e tratamento de cadastros para oferta de produtos e serviços aos clientes.

Utilizou-se a metodologia de mineração de dados definida na Literatura como “pano de fundo”, a fim de facilitar o entendimento, porém os casos práticos não apresentaram a mesma ordem de execução das fases, tal qual descritas neste capítulo. Os estudos de caso foram adaptados e organizados dentro das etapas existentes no processo teórico para viabilizar a análise comparativa do processo teórico com o prático, para melhorar a identificação das falhas e das oportunidades advindas da análise dos casos práticos.

No próximo capítulo, será realizada a avaliação do processo de mineração de dados no negócio de cartões de crédito, tendo como base o processo definido teoricamente, a fim de identificar sugestões de melhoria tanto no processo teórico como na prática organizacional.

Capítulo 4: Avaliação do processo de mineração de dados

4.1 Introdução

Este capítulo descreve e detalha a avaliação do processo de mineração de dados, tendo como base os aspectos teóricos expostos no capítulo 2 e a experiência prática vivenciada nos casos citados no capítulo 3, buscando otimizar e maximizar as potencialidades de cada parte, de forma a eliminar as possíveis falhas ou lacunas existentes entre os dois processos: teórico e prático.

Baseado na análise dos casos práticos do ponto de vista do processo de mineração de dados, a metodologia de mineração de dados é avaliada para a sugestão de melhorias na obtenção dos resultados finais, e, conseqüentemente, no desenvolvimento do negócio como um todo, facilitando o trabalho dos analistas de negócios e dos profissionais responsáveis pela tomada de decisões estratégicas e operacionais.

Para o detalhamento dos casos práticos, a metodologia de mineração de dados teórica foi utilizada como guia para a descrição dos estudos. Porém, observou-se que eles não seguiram a mesma ordem de execução de etapas estabelecidas pela metodologia, sendo adaptados às fases existentes no processo tanto para facilitar o entendimento quanto para uniformizar os conceitos para a realização da análise comparativa da avaliação a que este trabalho se propõe.

Outro aspecto importante a ser ressaltado refere-se à seleção de cinco casos práticos para avaliação do processo de mineração de dados. Embora a quantidade pareça excessiva, a diversidade entre os casos práticos (seja do ponto de vista de negócio ou de técnica a ser aplicada) foi considerada de forma construtiva, pois possibilitou a avaliação da metodologia de mineração de dados de forma mais completa, testando, inclusive, se as particularidades de cada caso, modificariam significativamente o processo de mineração de dados.

Dessa forma, pôde-se chegar a uma metodologia de mineração de dados única e genérica com mais precisão, independente do foco de negócio, da técnica escolhida e do segmento corporativo da empresa.

Ainda como complementação do trabalho de pesquisa, o processo de Engenharia de Software é considerado para possíveis aproveitamentos e adequações para melhorar o processo de mineração de dados, dada a maturidade da sua

existência e sua utilização no mercado, servindo de base para o refinamento da avaliação.

Dessa forma, este capítulo estrutura-se de acordo com cada fase do processo de mineração de dados. Dentro de cada etapa, são destacados os aspectos importantes da fase observados nos estudos de casos, de forma a ressaltar os pontos positivos e negativos da experiência vivida na prática.

4.2 Experiência obtida através dos casos práticos

4.2.1 Identificação de objetivos

Com a experiência obtida com os estudos de casos, verificou-se que a mineração de dados, assim como em diversas áreas, necessita de uma fase de planejamento detalhada sobre os tipos e níveis de informações necessárias, contribuindo significativamente para a redução de custos e tempo nos esforços do projeto bem como para o sucesso do mesmo.

Na maioria dos casos, houve incentivo por parte da alta gerência, pois os projetos suportariam e facilitariam a tomada de decisões. Observa-se que este apoio é muito importante para o sucesso do projeto, porque dessa forma a alta gerência compreende os trabalhos que estão sendo executados, administrando suas agendas e focando os esforços da equipe na realização do mesmo.

O envolvimento de todos os funcionários relacionados ao projeto nesta fase, desde analistas de sistemas até analistas e gerentes de negócio, revela-se ponto-chave para o bom andamento do projeto, pois uma vez planejado com o aval de todos os conhecedores e usuários do projeto se reduzem fortemente à probabilidade de re-execução de fases, devido ao fato de estar de acordo com as normas e conceito de negócio e ao atendimento das necessidades de todos os envolvidos.

Como processo estruturado suficientemente detalhado para esta fase inicial de planejamento, sugere-se a adaptação do processo de análise de requisitos da Engenharia de Software que pode ser utilizado como facilitador no detalhamento do escopo do trabalho bem como documentação das experiências anteriores, tendo em vista o alto nível de retrabalho decorrente da falta de documentação.

A seguir são detalhadas todas as sub fases do processo teórico da fase de identificação de objetivos, contemplando as contribuições obtidas pela avaliação do processo de mineração de dados.

4.2.1.1 Definição do escopo do estudo

Baseado no problema identificado, o escopo do estudo deve ser definido, a fim de atender a demanda levantada. Nos casos práticos, observou-se que a definição do escopo do estudo foi realizada de forma superficial e desordenada, o que causou inúmeros reprocessamentos decorrentes da falta de detalhamento na especificação das variáveis necessárias para a realização do processo de mineração de dados.

Esta característica é bastante comum, uma vez que não existe uma cultura de documentação e de análise de requisitos de forma adequada, tal qual ocorre no processo de Engenharia de Software.

No caso prático da Previsão de Recebimento de Clientes Inadimplentes, por desconhecimento dos sistemas e regras de negócio envolvidos no processo e, principalmente, pela ausência de uma documentação que suportasse o estudo. A definição do escopo do estudo foi realizada várias vezes, tendo em vista erros decorrentes do esquecimento de variáveis importantes para a realização de cálculos e de cruzamentos entres as bases, além da ocorrência de falhas de interpretação dos resultados obtidos.

Por esta razão, acredita-se que a maior contribuição deste trabalho esteja na sugestão da utilização da análise de requisitos na fase de definição do escopo do estudo e do processo de documentação formal em todas as fases do processo de mineração de dados, ambos existentes no processo de desenvolvimento de sistemas.

4.2.1.2 Seleção de equipe

Esta etapa, não existente na metodologia teórica, refere-se à seleção da equipe responsável pelo processo de mineração. Conforme observado nos casos práticos, o sucesso dos projetos esteve bastante vinculado à experiência dos funcionários envolvidos, uma vez que a documentação necessária não supria a necessidade por informações históricas pertinentes aos sistemas e processos relacionados ao escopo do estudo.

Mesmo focando esforços na realização de uma detalhada análise de requisitos e de uma documentação que garanta a continuidade dos trabalhos, a alta gerência

deve decidir os funcionários (analistas de sistemas, analistas de negócio, etc) que estarão envolvidos no projeto, considerando suas habilidades que devem contribuir para a execução precisa dos conceitos de negócio com aderência aos sistemas disponíveis.

Segundo [48], a seleção da equipe deve seguir o fluxo de demanda de informações mostrado na figura 4.1, presente em todos os processos de mineração de dados.

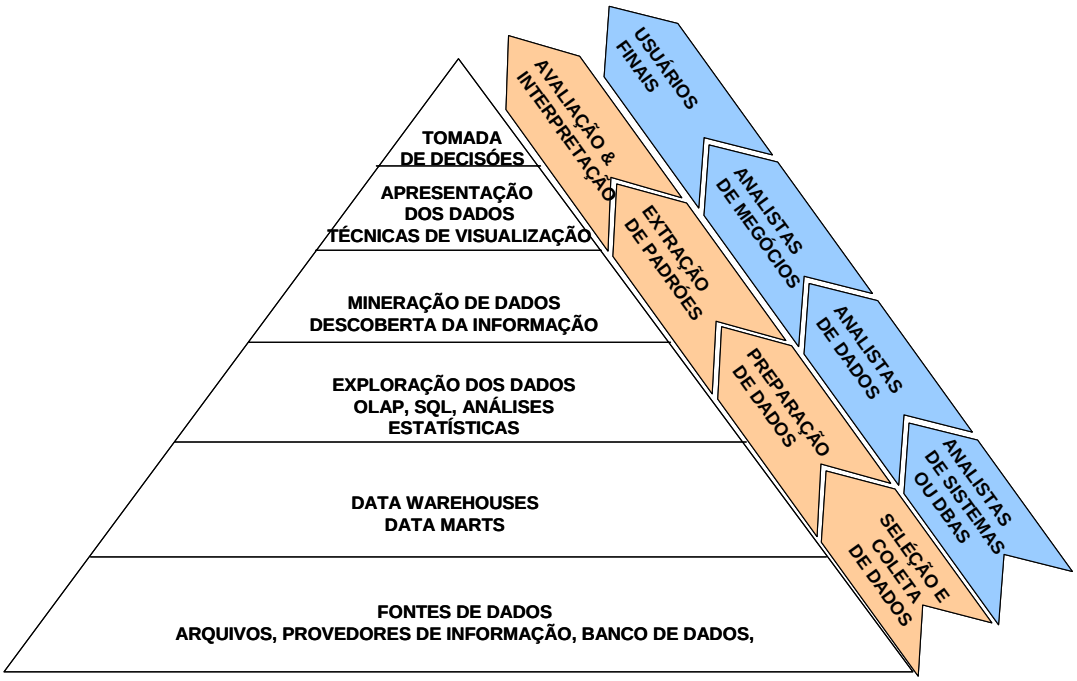

Figura 4.1: Relação entre atividades e fases do processo com a seleção dos profissionais [48]

Observa-se que em cada fase do processo de mineração de dados são necessários profissionais com diversas especialidades. Embora, a participação dos profissionais não se restrinja apenas à fase assinalada na figura 4.1, pois os mesmos participam em diversos graus de todas as fases do projeto, os tipos profissionais foram classificados por fases para facilitar o entendimento do grau de participação em cada fase.

Para a fase de seleção e coleta dos dados, analistas de sistemas ou administradores de bancos de dados são requeridos, tendo em vista a habilidade de manipulação de dados. Para a fase de preparação de dados, os analistas de dados são necessários, a fim de realizar os tratamentos e filtros na base de dados.

Já na fase de extração de padrões, os analistas de negócio são requeridos, uma vez que têm o domínio do negócio para a identificação de tendências e padrões nos dados, tendo como base às regras de negócio.

Por fim, o usuário final, normalmente definido pela Alta Gerência, utiliza as informações geradas, através da tomada de decisões, na fase de avaliação e interpretação dos resultados. O usuário final, embora não destacado na figura 4.1, é responsável também pela fase inicial de identificação dos objetivos, tendo o papel de requisitante na grande maioria dos processos de mineração de dados.

4.2.1.3 Estabelecimento de cronograma

Em nenhum dos casos práticos foi estabelecido um cronograma formal de trabalho, o que prejudicou não só o tempo de entrega e finalização do projeto, como também a distribuição das atividades pelos próprios analistas.

No caso do Sistema de Informações Gerenciais, houve por parte da Gerência um processo lógico de priorização das atividades, baseado na complexidade e no aprendizado gradual da equipe, que permitiu que o projeto alcançasse de forma mais harmoniosa seus objetivos iniciais, não comprometendo o prazo inicial previsto.

Observa-se que o estabelecimento de um cronograma é imprescindível, uma vez que formaliza expectativas entre os usuários e os analistas, a fim de viabilizar o gerenciamento do projeto de forma mais harmônica.

Na Engenharia de Software [28], existem alguns métodos de determinação de cronograma: PERT (avaliação e revisão do programa) e CPM (método do caminho crítico). Ambos desenvolvem uma descrição da rede de tarefas de um projeto, através de uma representação tabular das tarefas, denominada WBS (work breakdown structure), e das disposições, que indica a ordem de execução das tarefas, permitindo a determinação do caminho crítico, o estabelecimento de estimativas de tempo mais prováveis para tarefas individuais (aplicam modelos estatísticos) e o cálculo de limites de tempo (através da definição de “janelas” de tempo para cada tarefa particular).

4.2.1.4 Uso de ferramentas de planejamento

De acordo com Jesus Mena [17], para facilitar a realização desta fase de planejamento, existem várias ferramentas de modelagem e técnicas de construção de um mapa de atividades para projetos de mineração de dados. Estas ferramentas podem alertar sobre possíveis obstáculos, tais como a falta de acesso a certos dados para análise.

Conforme mencionado na etapa de estabelecimento de cronograma, não houve a utilização de ferramentas de planejamento, o que prejudicou fortemente a administração de tempo, principalmente nos estudos de curto prazo como a previsão de recebimento de clientes inadimplentes, como na elaboração de uma documentação do projeto do mesmo estudo.

Outro benefício refere-se à garantia de documentação dos esforços de mineração de dados. A documentação referente às análises do processo de mineração de dados é extremamente necessária durante todo o projeto, tanto no início quanto no final e, principalmente, para o arquivamento de conhecimento histórico, evitando trabalhos concorrentes e erros recorrentes sobre um mesmo tema já investigado, a não ser que se trate de atualização do mesmo, mas mesmo assim, a documentação serve de base para a realização da atualização.

Existem ainda ferramentas de modelagem de processo que também fornecem um *benchmarking* para a medição de desenvolvimento e ainda uma linguagem de comunicação entre a empresa e os componentes da equipe, através da qual o time do projeto consegue trocar documentos, agendar reuniões, realizar cronogramas e até comprometer recursos ao longo do projeto.

Como exemplos, citam-se as seguintes ferramentas: Corporate Modeler, da Casewise Systems; ProCarta, da Domain Knowledge; Aris Toolset, da IDS Scheer; Live Model, da IntelliCorp; Workflow Modeler, da Meta Software; Designer/2000, da Oracle; Aion, da Platinum Technology; ProVision Workbench, da Proforma; e, Visio, da Visio.

4.2.1.5 Utilização da Análise de Requisitos da Engenharia de Software

Conforme avaliação dos casos práticos, a metodologia de desenvolvimento de sistemas pode ser utilizada como ferramenta auxiliar nas fases de identificação de objetivos e nas demais fases no que se refere à documentação. Devido à maturidade do processo de Engenharia de Software, a metodologia de mineração de dados pode ser enriquecida e complementada, utilizando-se a metodologia de análise de requisitos para desenvolvimento de sistemas.

Segundo [28], a análise de requisitos possibilita a especificação da função e do desempenho do software, a interface do software com usuários e outros elementos do sistema e o estabelecimento de restrições do projeto, permitindo a construção de modelos de processo, dos dados e do comportamento geral do sistema.

A tarefa de análise de requisitos é um processo de descoberta, refinamento, modelagem e especificação do software a ser construído. Uma vez que o processo de mineração de dados, na maioria das vezes, necessita da construção de uma estrutura sistêmica (como por exemplo, sistemas de bancos de dados) para a utilização e manipulação dos dados, as atividades da análise de requisitos podem ser aplicadas, garantindo um planejamento detalhado e documentado adequadamente.

Alguns passos da análise de requisitos que não foram utilizados nos casos práticos e que facilitariam o processo de mineração de dados seriam:

- Detalhamento formal de requisitos junto aos usuários: nesta atividade, é realizado o levantamento formal e documentado das necessidades do usuário do ponto de vista de informação resultante do processo de mineração de dados. Por exemplo, no caso do Sistema de Informações Gerenciais de Autorização, os requisitos solicitados pelos usuários seriam os indicadores necessários nos relatórios gerenciais, as dimensões (quebras) de cada relatório, os conceitos de cada cálculo e a forma de disponibilização das informações finais (tabelas, gráficos, fluxos, etc). Como documentação resultante desta atividade, os responsáveis pelo projeto devem gerar atas de reuniões, contendo as solicitações dos usuários e os conceitos discutidos; e, a especificação dos requisitos do usuário, contendo todos os dados pertinentes ao resultado final esperado do processo de mineração de dados.
- Elaboração de dicionário de termos utilizados nos projetos: através da confecção de um glossário de conceitos de negócio utilizados, esta atividade consiste no levantamento de definições e regras de negócio pertinentes ao objeto de estudo do processo de mineração de dados, a fim de padronizar conceitos, pois todas as estruturas a serem criadas os utilizarão como base, sendo importante para consulta e interpretação dos resultados. Por exemplo, no caso de Análise de Perfil de Cadastro, existem variáveis referentes à classificação dos possíveis proponentes dos cadastros como faixa de renda, indicador de telefone (residencial, comercial, etc), nível de risco de crédito. Dessa forma, todas informações constantes nesta estrutura devem ter seus conceitos detalhados e documentados, a fim de facilitar novos projetos que os utilizem e, ainda, reduzir a ocorrência de erros de interpretação errada dos resultados das análises realizadas.
- Elaboração de diagramas relativos ao processo a ser desenvolvido bem como aos fluxos de dados (DFDs, em Engenharia de Software), além da

descrição lógica dos processos envolvidos. Nesta atividade, o processo a ser minerado deve ser previamente descrito, tanto do ponto de vista dos dados, para facilitar a seleção, coleta e preparação dos dados de forma adequada para atingir os resultados esperados, quanto do ponto de vista do processo de negócio, para garantir que os sistemas e dados acessados reflitam as informações necessárias para a mineração de dados alcançar os objetivos identificados. Por fim, como documentação resultante desta atividade, citam-se: descrição lógica dos processos envolvidos, fluxo do processo e dos dados do processo de mineração de dados a ser desenvolvido. Como exemplo desta atividade, tem-se o caso prático de Previsão de Recebimento de Clientes, no qual os dados são extraídos de um banco de dados em Mainframe com várias tabelas, cada qual representando uma unidade (cliente, acordos, pagamentos, etc), tornando-se necessário um detalhamento prévio do processo de cobrança de clientes inadimplentes, a fim de facilitar o entendimento do fluxo sistêmico e dos dados, para, finalmente, construir-se um fluxo do processo a ser desenvolvido, contendo detalhamento dos sistemas a serem acessados e dos dados a serem trabalhados.

- Elaboração de dicionário de dados, através de detalhamento das seguintes estruturas: entidades externas, processos, fluxos e depósitos de dados e/ou seus respectivos objetos, módulos, essências, incluindo a modelagem de dados. Nesta atividade, os dados a serem utilizados no processo de mineração de dados devem ser descritos, a fim de documentar os requisitos necessários para a construção do processo de mineração de dados. Como exemplo, deste detalhamento, cita-se o caso prático de Sistema de Prevenção de Fraude em Autorizações, no qual foram criadas variáveis correspondentes às dimensões de agrupamento para detecção de comportamento fraudulento nas compras realizadas com cartões de crédito. Estas variáveis criadas (calculadas) bem como todos os demais campos utilizados no processo de mineração de dados devem estar descritos no dicionário de dados, contendo nome, tamanho, tipo de dado, fórmula de cálculo (se aplicável), domínios, dentre outras informações necessárias para o entendimento do processo.

No caso prático do desenvolvimento de um sistema de informações gerenciais, foram elaborados os seguintes documentos: dicionário de termos, diagramas relativos

ao processo e ao fluxo de dados, dicionário de dados; além da definição de layout de relatórios contendo os tipos e tamanhos dos campos, obtendo-se maior sucesso e rapidez na conclusão do projeto.

Sendo assim, antes de começar o projeto, é necessário ter claro conhecimento das necessidades e das expectativas a serem alcançadas com o projeto. O importante é o objetivo de negócio, de forma a influenciar a seleção dos dados, das ferramentas e, principalmente, do formato final do projeto a ser implementado, O foco deve estar nas metas de negócio a serem alcançadas, pois a mineração de dados precisa ser vista como um movimento estratégico e competitivo e, como todas as outras iniciativas de negócio, deve ter uma meta mensurável, definida na fase inicial do projeto.

4.2.2 Seleção e Coleta de dados

Nesta etapa, observou-se que no estudo de caso de Previsão de recebimento de clientes inadimplentes, surgiram grandes dificuldades de realização, dada a complexidade e a deterioração dos sistemas envolvidos, pois continham várias tabelas a serem relacionadas contendo alto volume de registros, além da exigência de curto tempo de execução tanto no Mainframe quanto na baixa plataforma.

Mesmo com a realização dos cruzamentos entre as tabelas, os valores dos campos (domínios, tais como saldos devedores de clientes) contidos em cada uma das tabelas não coincidiam entre si (ou seja, uma tabela consolidada continha o saldo devedor do cliente que, em outra tabela, não coincidia com o total de títulos devidos por este mesmo cliente), levando ao reprocessamento por várias vezes.

O maior tempo despendido neste estudo de caso concentrou-se nesta etapa, devido ao alto nível de complexidade da estrutura de dados existente (em diversas tabelas distintas com deterioração de campos) e também ao fato de que a equipe não possuía nenhum profissional com conhecimento profundo destes sistemas.

Outro estudo de caso que apresentou características importantes a serem consideradas na seleção e coleta de dados refere-se às técnicas de amostragem utilizadas no caso do Modelo de Renda Presumida. Para uma análise baseada em amostra, no caso 50% do total da população, é necessário que a mesma tenha sido extraída corretamente, de forma absolutamente aleatória, para não distorcer as conclusões baseadas em dados não aleatórios.

Uma forma de extração de registros para composição de amostra refere-se à extração randômica, através da criação de campo com valor randômico e posterior

seleção de faixas correspondentes ao percentual necessário para análise, e outra forma eficaz de extração de amostra referem-se à seleção baseada em campo já existente na base com características aleatórias, como por exemplo, posições específicas da variável ID ou CPF, desde que esta variável não tenha sido utilizada por alguma estratégia que seleciona os clientes de forma aleatória. No estudo citado, esta seleção foi realizada através da criação de um campo numérico com conteúdo aleatório. Para extração de 50%, utilizou-se os casos pares, deixando os ímpares para amostra de validação do modelo, usada posteriormente.

É muito importante também que a amostra tenha a quantidade de registros suficientes para a análise, nos casos de modelagem baseadas em perfil de risco de clientes, são definidos porcentagens mínimas (aproximadamente 5%) ou valores mínimos (aproximadamente 50 casos) de distribuição de clientes maus e bons ao longo do tempo, visando reduzir sazonalidades e distorções oriundas de baixo volume.

A seguir são detalhadas todas as sub fases do processo teórico da fase de seleção e coleta de dados, contemplando as contribuições obtidas pela avaliação prática do processo de mineração de dados.

4.2.2.1 Definição de variáveis a serem utilizadas

Para a definição das variáveis a serem utilizadas nos casos práticos de mineração de dados, observou-se que a maior dificuldade se concentrou na ausência de uma documentação adequada dos sistemas envolvidos, de forma a suportar a escolha dos campos necessários para a extração do conhecimento identificado pelo problema.

No caso da Previsão de Recebimento de Clientes Inadimplentes, não só o desconhecimento das informações constantes no sistema, mas a deterioração sistêmica, causada pela reutilização de campos, impactaram significativamente o processo de mineração de dados, sendo necessária à extração da amostra inúmeras vezes, até que as variáveis a serem utilizadas fossem extraídas com sucesso.

Com isso, a definição das variáveis para utilização dos estudos foi realizada, através da extração de uma amostra contendo todos os campos disponíveis no sistema e da análise dos valores dos campos na tela produtiva, até que se chegasse num consenso de quais campos continham a informação correta para a análise, demandando tempo de projeto que poderia ser alocado em outras atividades, se houvesse documentação de suporte para a escolha das variáveis importantes para cada estudo.

4.2.2.2 Estabelecimento de formas de acesso aos sistemas

Como não existiam documentações que suportassem o estabelecimento de formas de acessos aos sistemas, os analistas de sistemas envolvidos, que tinham domínio destas informações, realizaram esta atividade, denotando alto grau de dependência dos analistas de sistemas.

Como exemplos de experiências relacionadas a este aspecto, tem-se o caso prático de Previsão de Recebimento de Clientes inadimplentes, em decorrência da deterioração sistêmica causada pela reutilização de campos e o caso de Desenvolvimento de Sistema de Informações Gerenciais de Autorização, em decorrência da alteração não documentada da chave de relacionamento entre os sistemas de autorização e cadastro de associados.

No caso de Análise de Perfil de Cadastro, a etapa de estabelecimento de formas de acesso aos sistemas apresentou grandes dificuldades, uma vez que os cadastros externos são recebidos sob diversos layouts, sendo necessária à definição de um formato padrão de entrada dos dados, contemplando a maior parte dos dados comumente recebidos nestes tipos de arquivos. Com isso, utilizou-se o histórico dos cadastros recebidos para identificar quais seriam as variáveis mais comuns e, principalmente, interessantes para o armazenamento na estrutura construída.

4.2.2.3 Verificação da adequação dos dados

De acordo com os casos práticos, esta atividade acaba sendo efetivamente realizada na etapa de seleção e coleta de dados, através da análise exploratória dos dados que consiste na análise do conteúdo de todos os campos da amostra, a fim de identificar se as variáveis selecionadas correspondem aos dados necessários para resolução do problema.

Dessa forma, sugere-se que esta atividade seja concentrada na fase de seleção e coleta de dados, uma vez que, de posse dos dados, fica mais acessível o processo de verificação da adequação dos dados. E, mesmo existindo uma documentação contendo o domínio das variáveis, é importante realizar a análise exploratória dos dados para se obter um conhecimento prévio da amostra e, principalmente, verificar a existência de campos com baixo grau de preenchimento ou até de campos com desvios (*outliers*).

4.2.3 Preparação de dados

Com a experiência obtida através dos estudos de casos, observou-se que esta etapa varia conforme a documentação disponível e, na ausência desta, a equipe de profissionais com conhecimento das regras de negócio e dos sistemas envolvidos, pois facilitam o processo de avaliação dos dados existentes nos sistemas bem como o tratamento dos mesmos.

No estudo de caso referente à construção de um sistema de informações gerenciais do processo de autorização, esta etapa representou grande parte dos esforços do projeto (aproximadamente 70%), considerando-se ainda que o tempo gasto na preparação dos dados foi bastante reduzido pela participação de um analista de sistemas na equipe que detinha profundos conhecimentos técnicos sobre os sistemas e dados a serem acessados.

Atualmente, existem empresas especializadas na prestação de serviços de preparação de dados, tanto para a construção de plataformas para recebimento de arquivos para *database marketing* como para implementação de *data warehouses*.

A seguir são detalhadas todas as sub fases do processo teórico da fase de preparação de dados, contemplando as contribuições obtidas pela avaliação prática do processo de mineração de dados.

4.2.3.1 Limpeza dos dados

Nos estudos de casos, observou-se a necessidade do conhecimento do conteúdo (domínio) das variáveis para que seja realizada uma análise crítica de seu conceito nos registros da amostra, a fim de decidir a exclusão da variável do estudo ou do tratamento de alguns casos, como no caso dos valores acima de US$ 60 mil nas autorizações.

Através da análise exploratória dos dados, a ser realizada na etapa de verificação de adequação dos dados (em Seleção e Coleta de Dados), já é possível identificar quais dados devem ser eliminados ou alterados na amostra, além de filtros ou exclusões definidas pelas regras de negócio definidas para a realização do estudo de caso.

Como exemplo de limpeza de dados cita-se também o caso de Previsão de Recebimento de Clientes Inadimplentes, os registros referentes a clientes com saldo inferior a R$ 100,00 foram excluídos, pois não faziam parte do escopo da análise.

4.2.3.2 Integração dos dados

No caso prático de previsão de recebimento de clientes inadimplentes houve maior dificuldade, devido à quantidade de tabelas a serem relacionados (dados cadastrais, títulos vencidos, acordos firmados, parcelas de acordos e pagamentos) e ao tempo disponível para a realização destes cruzamentos.

Como os cruzamentos não foram realizados no Mainframe, o tempo de processamento destes dados foi demorado, retardando a entrega do projeto. Já nos estudos de casos de desenvolvimento de um sistema gerencial de autorização e de criação de um sistema de prevenção de fraudes na autorização, todos os cruzamentos entre sistemas foram realizados em alta plataforma, minimizando a incidência de erros e otimizando o processo.

Já no caso prático de Desenvolvimento de um Sistema de Informações Gerenciais de Autorização, a integração de dados foi dificultada pelo desconhecimento e pela falta de documentação decorrente da alteração da chave de relacionamento entre o cadastro de associados e a base de autorização ao longo do tempo. Com isso, o cruzamento entre estes sistemas não se realizava com sucesso, atrasando o processo de mineração de dados até a identificação do problema.

Na Análise de Perfil de Cadastro, a integração dos dados se dá de forma interessante, uma vez que o relacionamento entre as tabelas utiliza uma técnica de relacionamento entre variáveis não numéricas (denominada ***m**atch **c**ode*) que transforma dados alfanuméricos em numéricos, realizando a relação dos registros por similaridade parcial, ou seja, casos de nomes divergentes de um mesmo cliente (em decorrência de mudança de estado civil) podem ser relacionados ao mesmo registro, considerando-se um percentual mínimo de similaridade entre eles.

A integração de dados assim como as demais fases da etapa de preparação de dados são bastante simplificadas quando existe uma plataforma de armazém de dados, pois todos os cruzamentos, transformações e limpezas são definidos e realizados de uma só vez e de forma transparente para o usuário (que pode ser desde um analista de sistemas especializada até um diretor sem profundos conhecimentos

em mineração de dados), não necessitando que o mesmo possua grandes domínios a respeito dos sistemas envolvidos no processo a ser minerado.

4.2.3.3 Transformação dos dados

Dentre os aspectos principais de tratamento de dados, citam-se: análise de desvios, como no caso das autorizações fraudulentas que possuem características divergentes do comportamento usual dos clientes, e tratamento de inconsistências, como no caso das autorizações que, por erro de digitação, possuem valores superiores a US$ 60 mil e que têm esses valores alterados para US$1 para não distorcerem os resultados finais.

Outro tratamento utilizado neste estudo corresponde aos agrupamentos realizados com o intuito de classificar campos de acordo com características semelhantes de acordo com a informação a ser analisada. Como exemplo, citam-se os grupos de ramos e países, criados para geração dos relatórios.

Já com relação ao estudo de caso de Análise de Perfil de Cadastro, a fase de preparação de dados apresentou-se com maior amadurecimento, dada a evolução dos processos de Database Marketing que focam em análise de bancos de dados para realização de vendas e ofertas a clientes ou não clientes. Neste estudo de caso, várias técnicas de preparação de dados foram extensivamente aplicadas para tratamentos de campos, tais como nomes, endereços; além de validações comuns como datas, telefones, profissões, etc.

Assim, observa-se que a transformação de dados corresponde à efetivação das alterações nos dados em resultado da análise exploratória dos dados (realizada na fase de Seleção e Coleta de Dados) e da definição do escopo do estudo (realizada na fase de Identificação de Objetivos), visando adequar os dados para a análise, minimizando desvios resultantes de valores inconsistentes ou de utilização de registros ou campos que não façam parte das regras de negócio pertinentes ao estudo.

4.2.3.4 Redução dos dados

No estudo de caso de desenvolvimento de um sistema de informações gerenciais de autorização, a sumarização (ou geração de cubos) representou condição necessária para a concretização do projeto, tendo em vista o alto volume de

autorizações e, conseqüentemente, o impacto para a construção do sistema em banco de dados MSAccess com limitação de 1GB por arquivo. A geração de cubos facilitou tanto a manipulação de dados na plataforma Microsoft quanto à eficiência para a entrega diária dos relatórios, otimizando e viabilizando o processo decisório nesta área de negócio.

Esta atividade de sumarização também foi utilizada no Sistema de Prevenção de Fraudes em Autorizações, porém com a intenção de agrupar autorizações, classificando-as de acordo com características previamente estabelecidas para detecção de fraudes.

Dessa forma, a redução de dados é comumente utilizada quando existe alto volume de dados a serem manipulados, a fim de melhorar a performance de processamento dos dados e de satisfazer condições e/ou limitações das ferramentas a serem utilizadas no processo de mineração de dados.

Existem muitas tecnologias *OLAP* que armazenam os dados de forma sumarizada (em cubos) para garantir menor tempo de resposta ao usuário. Como exemplo, cita-se a ferramenta Powerplay (Cognos).

4.2.3.5 Uso de *data warehouses*

Nos casos práticos não houve a utilização de *data warehouses*, uma vez que os dados foram extraídos diretamente do sistema produtivo, devido à inexistência de um armazém de dados que satisfizesse as necessidades dos usuários envolvidos nos processos de mineração de dados nos aspectos de disponibilidade (tempo de resposta), de limitação de saída de dados (quantidade máxima de registros) e de definição de conceitos de variáveis (ou seja, as variáveis disponíveis não correspondiam às necessárias para os estudos).

Os *data warehouses,* quando implantados e disponíveis para o processo de mineração de dados, incluem os processos de seleção, coleta e preparação de dados. Estas etapas requerem desenvolvimento de programas em alta ou baixa plataforma, sendo necessário um processo de desenvolvimento de sistemas interno ao processo de mineração de dados que contemple desde as etapas de seleção e coleta de dados até a preparação, incluindo a adaptação dos dados à ferramenta escolhida. Normalmente, a implantação de *data warehouses* é realizada por profissionais de desenvolvimento de sistemas de bancos de dados e não por especialistas em mineração de dados.

Vale ressaltar que a implantação de *data warehouses* reduz significativamente os esforços desta etapa, considerando que os mesmos já foram realizados quando da implantação da plataforma de dados unificados e tratados para a realização contínua e quase automática do processo de mineração de dados.

4.2.4. Extração de padrões

Por se tratar de uma área de administração de cartões de crédito de negócio, os profissionais atuantes em administração de cartões de crédito (na maioria, graduados em Estatística, Matemática e Engenharia) são conhecedores tanto da área de análise de crédito quanto das técnicas e ferramentas que suportam o processo de mineração de dados. Porém, na falta de conhecimento das técnicas, ferramentas e algoritmos, existem empresas especializadas na prestação de serviços de extração de padrões.

Dessa forma, em todos os estudos de casos, a seleção das técnicas e a escolha dos algoritmos para alcance dos objetivos de mineração de dados previamente definidos foram estabelecidas de forma a atender o objetivo do estudo, valendo ressaltar que esta atividade representa o “coração” do processo de mineração de dados, sendo vital a adequação das técnicas a serem aplicadas para aquisição de conhecimento.

Com isso, é importante a realização de treinamentos relacionados à aplicação das técnicas e à utilização das ferramentas que suportam o processo de mineração de dados para garantir o bom andamento do projeto e, principalmente, a escolha adequada das técnicas e ferramentas na solução de cada problema.

Porém, deve-se salientar que a seleção de técnicas e a escolha de ferramentas a serem utilizadas em cada processo de mineração de dados devem ser realizadas preferencialmente na etapa de identificação de objetivos, após a definição do escopo do estudo, pois a técnica a ser utilizada deve vincular a seleção e o tratamento dos dados até o final do processo de mineração de dados.

Dessa forma, sugere-se que estas duas sub-fases (seleção de técnicas e escolha de ferramentas) sejam deslocadas para a etapa de identificação de objetivos, a fim de otimizar o processo de mineração de dados. Na avaliação dos casos práticos deste capítulo, elas continuarão vinculadas à etapa de extração de padrões, sendo

que a metodologia sugerida no final deste capítulo deve mencionar esta sugestão de alteração do processo teórico.

A seguir são detalhadas todas as sub fases do processo teórico da fase de extração de padrões, contemplando as contribuições obtidas pela avaliação prática do processo de mineração de dados.

4.2.4.1 Seleção de técnicas

Nos casos práticos, a seleção das técnicas se dá a partir da identificação do problema a ser solucionado, ou seja, para classificar dados de acordo com um perfil de comportamento, utiliza-se técnicas de classificação, tais como árvores de decisão, para prever o risco de inadimplência de determinados clientes, utiliza-se técnicas estatísticas, como análise de regressão.

Dessa forma, a seleção de técnicas orienta, desde o início do processo, as atividades a serem realizadas nas demais fases, pois, de acordo com a técnica selecionada para o processo de mineração de dados, os dados necessários devem ser extraídos e trabalhados de formas distintas. Com isso, sugere-se a realização desta fase na etapa de Identificação de Objetivos, conforme mencionado na metodologia sugerida no final deste capítulo.

Destaca-se na tabela 4.1 um breve descritivo das técnicas, dos algoritmos e das ferramentas utilizadas nos estudos de casos.

ESTUDO DE CASO	TÉCNICAS	FERRAMENTAS
Sistema de informações gerenciais de autorização	Classificação, segmentação e estatística (análise de desvios).	Microsoft Access e Microsoft Excel.
Previsão de recebimento de clientes inadimplentes	Classificação e previsão.	Microsoft Access, Microsoft Excel e SPSS.
Análise de Perfil de Cadastro	Classificação e segmentação.	Microsoft Access e SPSS.
Modelo de renda presumida	Classificação (árvores de decisão)	Answer Tree e SPSS.
Sistema de Prevenção de Fraudes em Autorizações	Classificação, segmentação e estatística (análise de desvios).	Microsoft Access, Microsoft Excel e SPSS.

Tabela 4.1: Quadro de extração de padrões baseado nos estudos de casos

Como se pode observar, a Estatística é freqüentemente utilizada para análise de risco de crédito, além de ferramentas simples e conhecimento comum como plataforma Windows.

As decisões referentes às técnicas a serem utilizadas devem ser feitas considerando aspectos, tais como finalidade, eficiência, velocidade e atendimento as necessidades de negócio. Nos casos práticos, observa-se a adequação na escolha das técnicas e, conseqüentemente das ferramentas, uma vez que foram selecionadas aquelas que melhor atendiam o resultado esperado.

A classificação e a segmentação foram utilizadas na maioria dos casos como forma de categorizar as variáveis para fins distintos, tais como agrupar para melhor visualização e interpretação dos resultados (como no sistema de informações gerenciais de autorização), prever recebimento baseado em comportamentos semelhantes (como no caso de previsão de recebimento de clientes inadimplentes), inferir renda mensal de proponentes (como no caso de modelo de renda presumida, onde grupos se diferenciavam entre si ou se assemelhavam em relação à renda). Isto ocorre devido à facilidade de manuseio dos dados para aplicação da técnica, à flexibilidade para parametrização e simulação de novos agrupamentos e ao alto grau de compreensibilidade gerado pelas mesmas.

4.2.4.2 Escolha de ferramentas

Algumas vezes, estas decisões podem impactar em algum destes aspectos conforme observado no sistema de informações gerenciais de autorização, no qual a ferramenta selecionada não suportava inicialmente a quantidade de dados presentes no sistema. Porém, o problema foi contornado pela geração de cubos sumarizados que atendiam a limitação da ferramenta e otimizavam a eficiência de geração dos relatórios.

Através da seleção da tecnologia correta (contendo o algoritmo apropriado), as características e a estrutura dos dados também precisam ser considerados, tais como número de campos com valores contínuos, número de variáveis dependentes, número de campos categóricos, tamanhos e tipos de registros. Com isso, a escolha da ferramenta a ser utilizada no processo de mineração de dados, como totalmente relacionada à técnica selecionada, deve ser realizada na etapa de Identificação de Objetivos, conforme metodologia sugerida no final do capítulo 4.

Embora não utilizadas nos casos práticos, devido a um aspecto cultural de portabilidade e compatibilidade em todos os níveis organizacionais, observa-se na tabela 4.2 a relação existente entre algumas ferramentas disponíveis atualmente no mercado e as respectivas técnicas contempladas pelas mesmas para facilitar a escolha das ferramentas baseada na técnica a ser utilizada.

Ferramenta (fornecedor)	Técnicas disponíveis
Knowledge Studio / Seeker (Angoss Soft Co.)	Segmentação , classificação, árvores de decisão com algoritmos CHAID e CART, redes neurais e associação.
Datamind (Datamind Corp.)	Classificação.
Clementine (SPSS, Inc).	Segmentação, classificação, árvores de decisão, redes neurais, associação e séries temporais.
Enterprise Miner (SAS Inst Inc)	Árvores de decisão, estatística, redes neurais e séries temporais.
Mineset (Silicon Graphics Inc)	Árvores de decisão, segmentação e associação.
Darwin (Oracle Corp)	Árvores de decisão e redes neurais.
PRW Model I (Única Tech.)	Segmentação e classificação.
MIT GmbH (Data Engine)	Redes neurais, segmentação, estatística e classificação.
Intelligent Miner (IBM)	Árvores de decisão, redes neurais, séries temporais, segmentação e associação.
4Thought/Scenario (Cognos Incorporated)	Redes neurais Séries temporais
Database Mining Marksman (HNC Software)	Redes neurais e segmentação.
Decision Series (NeoVista)	Árvores de decisão, redes neurais, segmentação e associação.
Pattern Recognition Workbench (Única Technologies)	Redes neurais, estatística, séries temporais e segmentação.

Tabela 4.2: Ferramentas para mineração de dados

Outros aspectos a serem considerados na seleção das ferramentas são: escalabilidade, precisão, formatos, recursos de pré-processamento (limpeza, integração, transformação e transformação), conectividade, recursos de importação e exportação de dados, gerenciamento de memória, eficiência, tolerância a ruídos e eficiência.

4.2.4.3 Tratamento dos dados

Nos casos práticos, houve alguns casos de reprocessamentos de tratamentos de variáveis causados pela falta de definição formal das transformações necessárias nos dados para aplicação de cada técnica, como, por exemplo, no Modelo de Renda Presumida, no qual houve a carga da base na ferramenta Answer Tree sem os devidos tratamentos de classificação de profissão, idade e setor postal, sem os quais o algoritmo não agrupa os dados de forma adequada, principalmente para variáveis contínuas como renda, os agrupamentos são perdidos.

Com a mudança da realização das fases de seleção de técnicas e escolha de ferramentas para a etapa de Identificação de Objetivos, a fase de tratamento dos dados da etapa de Extração de Padrões tende a ser suprimida, pois todos os tratamentos necessários devem ser realizados na etapa de Preparação dos Dados, pois todo o processo de mineração já será iniciado com foco no objetivo identificado e, conseqüentemente, priorizando as técnicas a serem aplicadas nos dados e evitando reprocessamentos nos dados.

4.2.4.4 Análise humana

Esta etapa de análise humana dos resultados obtidos apresentou diferenças de interpretação decorrentes da experiência dos profissionais envolvidos no processo de mineração de dados, sendo que o grau de dificuldade da interpretação dos dados obtidos do processo de mineração de dados varia de acordo com a técnica utilizada, ou seja, no caso de um modelo estatístico, como no Modelo de Renda Presumida, a análise humana requer um conhecimento do profissional referente aos conceitos estatísticos mais aprofundado do que a interpretação dos resultados expostos num sistema de informações gerenciais, como no caso da Autorização.

Através dos casos práticos, observa-se que a existência de documentações pertinentes aos estudos realizados anteriormente facilita a realização desta etapa, conforme destacado na Análise de Perfil de Cadastro que, embora não tivesse nada documentado, possuía um histórico registrado pelos profissionais, de forma que o tratamento dos dados foi automatizado, tendo como base principal, as transformações realizadas repetidamente nos dados.

Dessa forma, a existência de um processo de documentação consistente facilita, inclusive, a etapa de análise humana, pois minimiza a ocorrência de erros, através da interpretação dos dados baseada na experiência profissional da equipe envolvida, resultando em um processo de mineração de dados mais uniforme e consistente, ou seja, menos subjetivo e suscetível a erros.

4.2.5 Interpretação e avaliação do conhecimento

A avaliação dos casos práticos do processo de mineração de dados reitera a existência desta etapa de avaliação e interpretação dos resultados tal qual definida na literatura, considerando-se seus pontos positivos e já estruturados.

A seguir são detalhadas todas as sub fases do processo teórico da fase de avaliação e interpretação do conhecimento, contemplando as contribuições obtidas pela avaliação prática do processo de mineração de dados.

4.2.5.1 Avaliação dos resultados

Nos casos práticos, esta etapa despertava a necessidade de novas análises. Principalmente na previsão de recebimento de clientes inadimplentes, a técnica de previsão foi aplicada sob diversas óticas, a fim de assegurar que a previsão de recuperação baseada no percentual de recebimento fosse coerente e adequada, considerando o risco envolvido na provisão de tais valores.

Esta fase de avaliação dos resultados garante e define a implantação de novas políticas e a tomada de decisões para a realização de ações de melhoria na análise de risco da carteira de clientes. Com isso, a implantação foi realizada, conforme observado em todos os casos práticos citados neste trabalho. Porém, existem estudos de mineração de dados que não resultam em implantações de novas ações e/ou decisões de negócio, principalmente quando o conhecimento obtido através da mineração de dados contraria as expectativas e as premissas, não gerando novas estratégias.

4.2.5.2 Implantação

Em todos os casos, a fase de implantação referiu-se à automação do processo de mineração de dados, tratando-se da operacionalização do estudo realizado. Esta etapa representa a transformação do estudo realizado em ações estratégicas da empresa, ou seja, através da análise de recebimento de clientes inadimplentes implantou-se um processo de concessão de descontos de acordo com o tempo de

atraso da dívida, aumentando significativamente a recuperação de valores de clientes em atraso.

4.2.5.3 Documentação

Nos casos práticos, as atividades de documentação apresentam falhas por falta de informação necessária para o entendimento total do estudo (como no caso da análise de perfil de cadastro que, por questões de confidencialidade, não detalha todos os tratamentos e técnicas aplicados nos dados) ou não são elaboradas (como no caso da previsão de recebimento de clientes inadimplentes) devido à falta de priorização desta atividade em prol de atividades rotineiras prioritárias (e aparentemente mais importantes) que não possibilitam a realização adequada de tais tarefas.

A falta de documentação dos processos e sistemas relacionados ao negócio e aos projetos de mineração de dados resulta num alto grau de dependência dos profissionais conhecedores dos domínios das informações pertinentes ao negócio, além de inviabilizar a concretização destes projetos em tempo hábil e suficiente para a tomada de decisões de maneira mais ágil e competitiva, pois o tempo gasto na aquisição de conhecimento relacionado aos dados constantes nos processos e sistemas retarda de forma significativa o andamento dos projetos de mineração de dados.

Tendo como base, o processo de desenvolvimento de sistemas, sugere-se na tabela 4.3 a seguinte documentação básica por fase do processo de mineração de dados, a fim de garantir a continuidade e a qualidade das informações resultantes destes projetos. Através da documentação de cada fase obtém-se no final do processo, um documento consolidado do projeto capaz de garantir a continuidade do mesmo e o suporte para novos estudos.

Fase do processo de mineração de dados	Documentação a ser gerada como produto da fase
Identificação de objetivos	• Documento de solicitação do processo de mineração de dados. • Atas de reunião com predefinições e/ou anotações escritas. • Relação de requisitos desejados pelo usuário final. • Definição da equipe selecionada e de respectivas atividades. • Cronograma prévio para a realização do projeto. • Dicionário de termos e demais documentos prévios pertinentes aos processos e sistemas envolvidos no processo a ser estudado. • Fluxo do processo e dos dados do processo a ser desenvolvido. • Dicionário de dados, contendo informações pertinentes à base de dados a ser utilizada: nomes, descrições e domínios de campos, fórmulas de campos calculados, chaves de relacionamentos entre tabelas e arquivos. • Documento contendo informações referentes a filtros, tratamentos e relacionamentos realizados nas bases de dados. • Relação de tratamentos realizados nas variáveis em decorrência da aplicação de técnicas de mineração de dados.
Seleção, Coleta e Preparação de dados	• Relação de programas (códigos) para a extração e tratamento dos dados.
Extração de padrões	• Documento contendo aplicação da técnica de mineração de dados, incluindo as saídas das ferramentas tecnológicas bem como a análise humana sobre os resultados obtidos.
Interpretação e avaliação do conhecimento	• Relação de ações a serem tomadas (alterações de políticas e estratégias, modificações sistemas), além da especificação detalhada para execução das mesmas.

Figura 4.3: Proposta de documentação por fase do processo de mineração de dados [46]

Com isso, observa-se que a documentação é parte integrante de todas as fases do processo de mineração de dados, tendo sua maior participação na etapa de identificação de objetivos, na qual todos os levantamentos tanto de processo sistêmico

como de negócio são realizados, a fim de facilitar o processo de desenvolvimento de uma estrutura automatizada de mineração de dados (por exemplo, um sistema de informações gerenciais, um sistema de tratamento de dados externos ou um estudo de previsão de comportamento de crédito ou de fraude de clientes).

E ainda conforme mencionada na etapa de identificação de problema, a utilização de ferramentas de planejamento que ofereçam recursos automáticos de documentação asseguram que esta atividade seja realizada durante todo o processo de mineração de dados. Estas ferramentas podem ser construídas sob forma de bancos de dados, a fim de organizar o planejamento e documentar o processo de forma padronizada.

4.2.5.4 Acompanhamento

Nos casos práticos, a decisão de acompanhamento, através do monitoramento periódico das políticas ou estratégias implementadas com base em estudos de mineração de dados, deve partir de acordo com a avaliação dos resultados obtidos para definir e garantir a continuidade das melhorias implantadas e, principalmente, o melhor momento para alteração e revisão das mesmas.

4.3 Metodologia sugerida para o processo de mineração de dados

Tendo em vista, a avaliação de cada fase do processo de mineração de dados descrito neste capítulo, a metodologia a ser sugerida contempla o seguinte fluxo, mostrado na figura 4.2.

Esta metodologia reitera o processo de mineração de dados definido teoricamente, contemplando as seguintes sugestões de melhoria:

- Alteração cronológica das fases de seleção de técnicas e de escolha de ferramentas para a etapa de Identificação de Objetivos, de forma a garantir maior coerências das demais fases com a solução do problema identificado na fase inicial do processo. Esta alteração visa a minimização de reprocessamentos de seleção e tratamento de variáveis, além da manutenção do foco do trabalho durante todo o processo de mineração de dados.
- Inclusão das seguintes subfases na etapa de Identificação de Objetivos:

- Seleção de equipe: a inclusão de um processo de seleção de equipe estruturado facilita e ordena a realização de mineração de dados de forma harmônica e produtiva.
- Estabelecimento de cronograma: para fins de controle e organização, esta fase visa a definição de prazos e metas para a equipe envolvida no projeto.

- Utilização do processo de análise de requisitos da Engenharia de Software na etapa de Identificação de Objetivos para melhorar o planejamento de projetos de mineração de dados.
- Inclusão de produtos de documentação em cada fase do processo de mineração de dados, através dos templates também herdados da Engenharia de Software, a fim de garantir a continuidade e a manutenção destes processos.

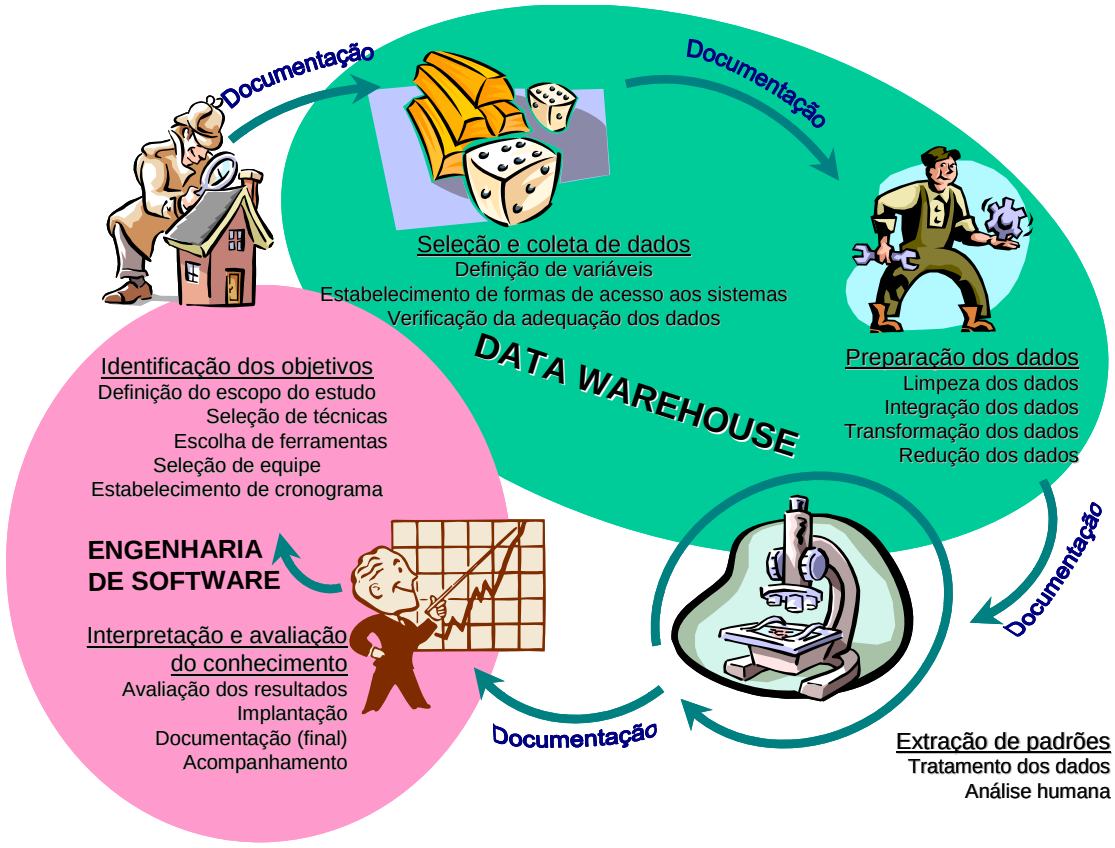

Figura 4.2: Metodologia sugerida de mineração de dados

Outro aspecto importante no processo de mineração de dados refere-se ao uso de armazéns de dados, pois a existência desta plataforma (quando construída de acordo com as necessidades dos usuários finais) reduz significativamente (ou até elimina) as etapas de seleção, coleta e preparação de dados no que se refere à codificação para automação destas etapas na estrutura criada pelo processo de mineração de dados.

Dessa forma, a metodologia sugerida surge como um fruto do aprendizado obtido através do levantamento da teoria sobre mineração de dados e da análise dos casos práticos vivenciados numa administradora de cartões de crédito, possibilitando a identificação de pontos fortes do processo teórico e de melhorias advindas da experiência prática. Com isso, esta metodologia tem como principal objetivo o desenvolvimento do processo de mineração de dados no ambiente corporativo de forma geral, contribuindo para o amadurecimento das empresas no que diz respeito a projetos desta natureza.

4.5 Conclusão

Neste capítulo, os casos práticos detalhados no capítulo 3 foram avaliados de acordo com o processo de mineração de dados definido na teoria exposta no capítulo 2, a fim de identificar os pontos fortes e fracos na metodologia utilizada na prática.

Logo no princípio da análise dos estudos de casos, observou-se que os mesmos não seguiram a mesma ordem das etapas do processo de mineração dedados, mas para facilitar o entendimento e a avaliação do processo, buscou-se padronizar os casos práticos de acordo com a metodologia teórica, mesmo não refletindo exatamente como o mesmo foi conduzido, principalmente, considerando-se os problemas de planejamento e documentação identificados.

Avaliou-se então que a metodologia teórica existente, se aplicada de forma disciplinada e planejada na prática, é bastante consistente e aderente ao processo prático, independentemente do foco de negócio ou da técnica escolhida, devendo sempre ser guiada pelos objetivos identificados na fase inicial do processo.

Para contribuir com o processo de mineração de dados, foram incluídas etapas de seleção de equipe e estabelecimento de cronograma, além da utilização do

processo de análise de requisitos na fase de identificação de objetivos e de documentação, ambos herdados do processo de Engenharia de Software.

No próximo capítulo, será descrito um resumo das conclusões alcançadas por este trabalho bem como as suas contribuições e sugestões para futuros trabalhos.

Capítulo 5: Conclusão

5.1 Resumo

O processo de mineração de dados vem tornando-se cada vez mais parte integrante do processo decisório das organizações, dado o diferencial competitivo e estratégico obtido pelas empresas que fazem uso do conhecimento adquirido através da utilização da informação armazenada nos sistemas produtivos em conjunto com a aplicação de técnicas e o uso de ferramentas de mineração de dados.

Conforme observado no presente trabalho, elaborado através da avaliação de casos práticos vivenciados em uma administradora de cartões de crédito, observou-se que, principalmente em instituições financeiras e seguradoras, o processo de mineração de dados é extensivamente utilizado para definição de estratégias e política obtidas através da análise de informações históricas armazenadas e trabalhadas estatística e computacionalmente para fins de melhoria de negócios e processos e maximização de lucros.

Como o foco do trabalho foi o processo de mineração de dados, realizou-se o levantamento do processo de mineração de dados definido na Literatura e detalhou-se cinco casos práticos vividos numa instituição organizacional. Tendo o processo teórico como pano de fundo, foram identificados pontos fortes e fracos do processo nos casos práticos, a fim de sugerir, detalhar ou ainda complementar o processo teórico, conforme experiência prática.

Foram selecionados cinco estudos de casos, cada qual com foco de negócio e técnico distintos, de forma a possibilitar a avaliação do processo de mineração de dados sob várias óticas, a fim de testar se uma estrutura metodológica única atenderia todos os casos com a mesma eficiência. Observou-se que, independentemente do foco de negócio ou da técnica utilizada, o processo de mineração de dados teórico mostrou-se aderente, podendo concluir que o mesmo se aplica para outros tipos de empresas, tais como indústrias, segurados e bancos.

A princípio, observou-se que os casos práticos não seguiram as etapas do processo de mineração de dados tal qual definido na teoria no que diz respeito à ordem das fases e à utilização disciplina das mesmas, além da identificação de problemas de falta de planejamento e de documentação que suportassem devidamente o detalhamento dos casos práticos.

Optou-se então por organizar a descrição dos casos, de forma a padronizá-los, a fim de facilitar o entendimento e a avaliação do processo prático de acordo com a metodologia teórica.

Através desta análise dos casos práticos já organizados de acordo com as etapas da metodologia teórica verificou-se que o processo prático mostrou-se aderente à teoria, porém a prática revelou-se bastante desorganizada, sendo que os retrabalhos e fracassos no processo de mineração decorreram, na maioria das vezes, da falta de planejamento e de experiência de negócio e de sistemas para a obtenção total dos objetivos identificados no início do projeto.

A etapa de planejamento, representada em mineração de dados pela fase de identificação de objetivos, revelou lacunas de aplicação, tendo em vista que a maioria dos estudos de casos não realizou de forma suficientemente detalhada e formal esta etapa, implicando numa fraca definição de requisitos, aspectos necessários para solução eficaz do problema identificado. Neste contexto, existem contribuições a serem aplicadas nesta etapa, considerando-se a análise de requisitos da Engenharia de Software como base metodológica.

A utilização dos conceitos de análise de requisitos, além de garantir a especificação detalhada do processo de mineração de dados, assegura a documentação do projeto em todas as suas fases, através das ferramentas auxiliares da Engenharia de Software.

Com isso, a utilização de uma metodologia de trabalho de mineração de dados conjugada à metodologia de desenvolvimento de sistemas, no que se refere à análise de requisitos e documentação, contribui significativamente para um bom processo de mineração de dados, pois detalha e documenta aspectos importantes para a condução bem sucedida do mesmo.

Outro aspecto importante refere-se à utilização dos objetivos identificados como guia em todas as demais fases do processo, ou seja, de acordo com os resultados a serem alcançados, deve-se selecionar a técnica capaz de atingi-los. De acordo com a técnica selecionada, as ferramentas devem ser escolhidas tanto para aplicação da técnica como para manipulação dos dados. Este aspecto justifica a alteração das fases de seleção de técnicas e de escolha de ferramentas para a etapa de Identificação de Objetivos, garantindo a condução do processo de forma mais focada aos problemas identificados, evitando maiores transtornos causados por retrabalhos e reprocessamentos.

Dessa forma, ressalta-se a importância da etapa de identificação de objetivos e justifica-se o investimento do presente trabalho no desenvolvimento desta fase,

através da utilização de conceitos de Engenharia de Software, a fim de aprimorá-la com a inclusão da análise de requisitos e de documentações previamente definidas para o desenvolvimento de sistemas.

Observou-se ainda que muitos retrabalhos foram decorrentes da má seleção da equipe envolvida no projeto, devido à falta de conhecimento sistêmico e das regras de negócio, o que dificultou o processo de extração de conhecimento, uma vez que os integrantes da equipe nem sempre apresentavam total domínio do conteúdo a ser explorado. É necessário que os participantes do processo sejam profissionais especialistas no domínio de negócio em que se pretende realizar a mineração de dados, embora a existência de uma documentação adequada minimize o impacto da seleção da equipe.

Por esta razão, a metodologia sugerida neste trabalho inclui os conceitos de análise de requisitos, etapas de seleção de equipe e definição de cronograma na fase de identificação de objetivos; ressalta e questiona a utilização de *data warehouses* com substitutos das etapas de seleção, coleta e preparação de dados; além de definir produtos finais de documentação por todo o processo de mineração de dados.

Considerando todos estes aspectos, acredita-se no desenvolvimento de um processo de mineração de dados mais eficiente e produtivo, totalmente focado em planejamento e bem mais documentado, garantindo a continuidade dos processos desenvolvidos e facilitando os futuros estudos.

Por fim, a experiência vivida e salientada neste trabalho adverte sobre as dificuldades enfrentadas para o sucesso do projeto, pois organiza, planeja e documenta o projeto, garantindo um aprendizado pertinente ao processo de mineração de dados de forma completa, abrangendo desde o gerenciamento do projeto, o detalhamento dos processos, sistemas e regras de negócio e, por fim, sua utilização efetiva através da aplicação de técnicas e uso de ferramentas para melhorar o processo decisório, garantindo, principalmente, o amadurecimento das empresas quanto ao processo de mineração de dados, tal qual observado em Engenharia de Software pelo CMM.

5.2 Contribuições

A principal contribuição deste trabalho refere-se à avaliação do processo de mineração de dados no ambiente corporativo de administração de cartões de crédito, validando sua adequação à realidade vivida na prática.

Mesmo tratando-se de um ambiente específico, os casos práticos, bastante diversificados, permitem que as análises e, principalmente, as conclusões possam ser generalizadas e adequadas a outros negócios.

Com isso, a avaliação dos casos práticos de acordo com a metodologia teórica de mineração de dados enfoca a utilização correta da teoria existente, tendo em vista o dinamismo do ambiente corporativo. Este trabalho visa contribuir na formação de uma cultura organizacional de planejamento e documentação de projetos, propiciando, inclusive, o amadurecimento das empresas no que diz respeito aos projetos de mineração de dados.

Outro resultado importante deste trabalho refere-se à análise do processo de Engenharia de Software, visando aproveitá-lo e adequá-lo como parte do processo de mineração de dados, uma vez que a metodologia de desenvolvimento de sistemas apresenta maior maturidade, podendo ter suas fases e características utilizadas de forma bem sucedida em extração de conhecimento.

Dessa forma, as seguintes contribuições podem ser destacadas:

- Avaliação de casos reais quanto à aderência à metodologia apresentada na Literatura.
- Descrição de casos vivenciados em ambientes reais, visando disseminar o aprendizado obtido com a experiência prática e retroalimentar o processo teórico, através das necessidades observadas na avaliação do processo.
- Sugestões de melhorias no processo prático, através da utilização de forma mais disciplinada e planejada da metodologia teórica, e da inclusão de novas etapas identificadas nos processos práticos, tais como seleção de equipe, estabelecimento de cronograma, análise de requisitos e documentação.
- Sugestão de utilização do processo de Engenharia de Software, a fim de suprir lacunas na fase de planejamento do processo de mineração de dados, através da incorporação e adaptação da análise de requisitos e

da utilização da documentação proposta para desenvolvimento de sistemas.

Por fim, acredita-se que outra contribuição trazida por este trabalho corresponde ao amadurecimento quanto aos processos de mineração de dados, reduzindo o tempo e o desgaste nestes projetos, principalmente no que se refere às etapas de planejamento e documentação, assegurando a continuidade destes processos bem como os frutos rentáveis de uma metodologia de mineração de dados bem utilizada.

5.3 Sugestões para futuros trabalhos

Considerando que o presente trabalho não consegue contemplar todos os aspectos pertinentes ao processo de mineração de dados, dado seu foco no processo, existem algumas sugestões para futuros trabalhos:

- Aprofundamento dos estudos relacionados à utilização do processo de análise de requisitos de Engenharia de Software ao processo de mineração de dados, uma vez que existem muitas semelhanças entre estes processos e que a Engenharia de Software apresenta-se bastante madura quando comparada à mineração de dados, do ponto de vista de metodologia, podendo contribuir ainda em vários aspectos da mineração de dados.
- Criação de um processo de automatização da documentação dos processos de mineração de dados bem como para a organização da etapa de planejamento.
- Avaliação do processo de mineração de dados proposto neste trabalho em outros ambientes organizacionais, a fim de validar adequação e aderência.

Referências bibliográficas

[1] GUIZZO, ÉRICO. **Guia do presidente digital - Como o computador pode ajudar a vender melhor**. Revista Exame Negócios, São Paulo, n. 748, p. 28-29, setembro, 2001.

[2] GROTH, ROBERT. **Data mining – building competitive advantage**. New Jersey, Prentice Hall Books, 253 p, 1999.

[3] HAN, J.; KAMBER, M. **Data mining: concepts and techniques.** Canada, Morgan Kaufmann Publishers, 547 p, 1999.

[4] CARVALHO, L. A V. DE. **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. São Paulo, Editora Érica Ltda, 234 p, 1ª. ed, 2001.

[5] WU, XINDONG. Data Mining: Updates in Technologies. **Data Warehousing and Data Mining Conference**, Singapore, Janeiro 1999.

[6] CHAUDHURI, S. Data Mining and Database Systems: Where is the Intersection? **IEEE Computer Society Technical Committee on Data Engineering**, p. 4-8, 1998.

[7] KRANAKIS, E.; KRIZANC, D.; PELC, A.; PELEG, D. The Complexity of Data Mining on the Web. **ACM Inc**, Philadelphia, p. 153, 1996.

[8] HOLSHEIMER, M. Data Mining by Business Users: Integrating Data Mining in Business Processes. **KDD-99**, San Diego, p. 266-291, 1999.

[9] ROSSET, S.; MURAD, U.; NEUMANN, E.; IDAN, Y.; PINKAS, G. Discovery of Fraud Rules for Telecommunications – Challenges and Solutions. **KDD-99**, San Diego, p. 409-413, 1999.

[10] HAN, JIAWEI. Data Mining Techniques. **SIGMOD 96**, Canada, p. 545, 1996, db.cs.sfu.ca/dbminer.

[11] GAROFALAKIS, M.; RASTOGI, R.; SESHADRI, S.; SHIM, K. Data Mining and the Web: Past, Present and Future. **WIDM 99**, Kansas, p. 43-47, 1999.

[12] FAYYAD, U.; UTHURUSAMY, R.; GUEST EDITORS. Data Mining and Knowledge Discovery in Databases. **ACM Inc**, Vol. 39, No. 11 p. 24-26, 1996.

[13] CHUNG, H. M.; GEY, F. C. Data Mining, Knowledge Discovery and Information Retrieval. **24th Hawaii International Conference on System Sciences**, Hawaii, p. 1, 2001.

[14] INMON, W. H. The Data Warehouse and Data Mining. **ACM Inc**, Vol. 39, No. 11 p. 49-50, 1996.

[15] HOFFMAN, P.; GRINSTEIN, G.; MARX, K.; GROSSE, I.; STANLEY, E. DNA Visual and Analytic Data Mining. **IEEE Computer Society Technical Committee on Data Engineering**, p. 437-441, 1997.

[16] JOST, A. Data Mining. In: MAYS, E. **Credit Risk Modeling – Design and Application**. Chicago, Amacon, 1998, Cap. 8, p. 129-154.

[17] MENA, J. **Data Mining Your Website**. USA: Digital Press, 1999.

[18] AMARAL, F. C. N. **Data Mining: Técnicas e Aplicações para o Marketing Direto.** São Paulo: Editora Berkeley, 2001.

[19] NORUKIS, M. J. **SPSS for Windows Professional Statistics Release 6.0**. Chicago, SPSS Inc., 1993.

[20] POLITO, M. **Data Mining**. Setembro/1997. Disponível em: http://www.infolink.com.br/~mpolito/mining/mining.htm. Pesquisa realizada em 11/10/2001.

[21] http://www.cesar.org.br/amalise/n_32/n_32.html. Pesquisa realizada em 11/10/2001.

[22] GUROVITZ, H. O que cerveja tem a ver com fraldas? Revista Exame - Mundo Digital, Abril/1997. http://www2.uol.com.br/exame/33mdig.html. Pesquisa realizada em 04/09/2001.

[23] MACHADO, C. **O abc da mineração de dados**. Revista Info Exame, Janeiro/1999. http://www2.uol.com.br/info/ie154/jan99hd.shl. Pesquisa realizada em 31/08/2001.

[24] http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm. **An introduction to Data Mining – Discovering hidden value in your data warehouse**. Pesquisa realizada em 29/08/2001.

[25] http://studentes.fct.unl.pt/users/nach/dmdw/prologo/prologo.htm. Data Mining / Data Warehousing. Pesquisa realizada em 29/08/2001

[26] SECO, A; SOUZA, C; SILVA, Edilberto; ARAÚJO, José; SOUSA; P.T.C. **Data Warehouse, Data Mart, Data Mining – Estudo de Caso.** Universidade Católica de Brasília, Mestrado em Informática, Junho/2000.

[27] FAYYAD, U. Mining databases: towards algorithms for knowledge discovery. **IEEE**, p. 39-49, 1998.

[28] PRESSMAN, R.S. Engenharia de Software. Makron Books, 1998.

[29] BARBIERI, C. **BI – Business Intelligence – Modelagem & Tecnologia**. Rio de Janeiro, Axcel Books do Brasil Editora, 2001.

[30] FAYYAD, U.; HAUSSLER, D.; STOLORZ.,P. Mining scientific data. **ACM Inc**, Vol. 39, No. 11 p. 51-55, 1996.

[31] BRACHMAN, R. J.; KHABAZA, T.; KLOESGEN, W.; PIATETSKY-SHAPIRO, G.; SIMOUDIS, E. Mining business databases. **ACM Inc**, Vol. 39, No. 11 p. 42-48, 1996.

[32] IMIELINSKI, T.; MANNILA, H. A database perspective on knowledge discovery. **ACM Inc**, Vol. 39, No. 11 p. 58-64, 1996.

[33] PIATETSKY-SHAPIRO, G.; MASAND, B. Estimating campaign benefits and modeling lift. **ACM Inc**, USA, p. 185-193, 1999.

[34] HELD, G. **From data to business advantage: data mining, the SEMMA methodology and SAS software**. SAS Institute Inc, 1998.

[35] RODRIGUES FILHO, J. A F. **Data Mining: conceitos, técnicas e aplicação**. Universidade de São Paulo, Escola Politécnica, Mestrado em Engenharia, 2001.

[36] DUTRA, R. G. **Aplicação de métodos de inteligência artificial em inteligência de negócios**. Universidade de São Paulo, Escola Politécnica, Mestrado em Engenharia, 2001.

[37] WESTPHAL, C. R. **Data Mining solutions: methods and tools for solving real-world problems**. EUA, Wiley Computer Publishing, 617 p, 1998.

[38] BISPO, C.A.F. **Uma análise de nova geração de sistemas de apoio à decisão**. Universidade de São Paulo, Escola de Engenharia de São Carlos, Mestrado em Engenharia de Produção, 1998.

[39]SHEPARD, D. **Database marketing: o novo marketing direto**. Tradução: Kátia Aparecida Roque. São Paulo, Makron Books, 1993.

[40] CARVALHO, A.; BRAGA, A.; MARTINELI, S.; LUDEMIR, T. **Understanding credit card users behaviour: a data mining approach.** Idea Group Publishing, 2002, p. 240-261.

[41] STOREY, A; COHEN, M. **Exploiting Response Models – Optimizing cross-sell and up-sell opportunities in banking**. KDD – 2002, Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Canada, July 23-26, 2002, p.325-331.

[42] ROSSET, S; NEUMMAN, E; EICK, U; VATNIK, N; Ida, Y. **Customer lifetime value modeling and its use for customer retention planning**. KDD – 2002, Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Canada, July 23-26, 2002, p.332-340.

[43] JULISCH, K; DACIER, M. **Mining Intrusion Detection Alarms for Actionable Knowledge**. KDD – 2002, Canada, July 23-26, 2002, p.366-375.

[44] TEIXEIRA, S. **A mina de ouro debaixo dos bits.** Revista Exame, São Paulo, n. 708, fevereiro, 2000.

[45] REZENDE, S. **Sistemas inteligentes – fundamentos e aplicações**. Editora Manole, São Paulo, 2002.

[46] REZENDE, D. **Engenharia de software e sistemas de informação.** Editora Brasport, Rio de Janeiro, 2002.

[47] BERRY, M; LINOFF, G. **Mastering data mining – The art and science of customer relationship management.** Wiley Computer Publishing, New York, 2000.

[48] PELLEGRINI, G. F; COLLAZO, K. **Extração de conhecimento a partir de sistemas de informação.** Universidade Federal de Santa Catarina, 2000.

GLOSSÁRIO

Termo	Significado
Algoritmos genéticos	Técnicas de otimização que utilizam processos como a combinação genética, a mutação e a seleção natural, baseados nos conceitos da evolução natural.
Amostra	Processo pelo qual somente uma fração de todos os dados disponíveis é utilizada para a construção de um modelo ou para a realização de uma análise explanatória. Amostras podem gerar bons modelos com menor volume de despesas computacionais do que com o uso do banco de dados inteiro.
Análise de desvios	Um tipo de análise de dados que busca determinar e relevar nos registros de banco de dados, informações que são significativamente diferentes dos demais registros. A técnica é usada para limpeza de dados, descoberta de tendências e reconhecimento de comportamentos não usuais (boas ou más).
Análise de fator	Técnica estatística que busca reduzir o número de variáveis totais para alguns fatores que tem predominância de impacto na variável de saída.
Análise de séries temporais	Análise de uma seqüência de medidas feitas em intervalos específicos de tempo. Tempo é normalmente a dimensão dominante entre os dados.
Análise exploratória de dados	Refere-se ao uso de técnicas estatísticas descritivas e gráficas para o aprendizado da estrutura de um conjunto de dados.
Aprendizado de máquina	Um campo da ciência e tecnologia concebida com a construção de máquinas que aprendem. Em geral, difere da Inteligência Artificial, na qual o aprendizado é considerado um dos caminhos para a criação de uma inteligência artificial.
Aprendizado não supervisionado	Técnica de análise de dados através da qual um modelo é construído sem um campo de previsão pré-definido ou meta particular. Os sistemas são utilizados para organização e exploração dos dados. Segmentação é um exemplo de técnica de aprendizado não supervisionado.
Aprendizado supervisionado	Uma classe de aplicações de mineração de dados e de aprendizado de máquinas nas quais o sistema constrói um modelo baseado na previsão de um campo pré-definido. É o contraste do aprendizado não supervisionado, no qual não há meta particular para a detecção de padrões.
Armazém de dados	Sistema de armazenamento e entrega de grandes quantidades de dados que facilita a integração de sistemas organizacionais para fins de mineração de dados.
Árvore de decisão	Estrutura semelhante a uma árvore que representa um conjunto de decisões. Estas decisões geram regras para a classificação de um conjunto de dados.
Banco de dados multidimensional	Banco de dados definido para processamento analítico on line e estruturado como um cubo multidimensional com um eixo por dimensão.
Campo	Componente estrutural de um banco de dados, comum a todos os registros no banco. Campos têm valores e também podem ser chamados de características, atributos, variáveis,

	colunas de tabelas, dimensões.
CART	Classificação e regressão em árvores de decisão. Uma técnica de árvore de decisão utilizada para classificação de um conjunto de dados. Apresenta como resultado um conjunto de regras que podem ser aplicadas a um novo conjunto de dados para prever quais registros terão determinadas saídas. Segmenta um conjunto de dados, através da criação de dois segmentos, requerendo menor nível de preparação de dados que CHAID.
CHAID	Uma técnica de árvore de decisão usada para classificar um conjunto de dados. Fornece um conjunto de regras que podem ser aplicadas para um novo conjunto de dados para prever quais registros terão determinadas saídas. Segmenta um conjunto de dados, através de testes de explicação da amostra (R2) para criar vários segmentos. Implica e necessita de maior nível de preparação de dados do que CART.
Classificação	Processo de divisão de um conjunto de dados em grupos mutuamente exclusivos de forma que os membros de cada grupo são mais homogêneos quanto possível e os grupos entre si são mais heterogêneos quanto possíveis, onde a distância é medida de acordo com variáveis específicas para a predição. Por exemplo, um problema típico de classificação consiste na divisão dos bancos de dados das empresas em grupos com alto nível de homogeneidade no que diz respeito a caráter crédito, podendo ser classificado em bom e mau.
Conhecimento	Produto resultante de análises efetuadas em bases de dados, através de métodos e técnicas conhecidas como mineração de dados ou descoberta de conhecimentos, podendo representar novos padrões ou validar hipóteses previamente estabelecidas, considerando que o produto objetivo é algo não óbvio e deveria ser representado como algo mais simples do que o próprio dado analisado, além de ser expresso em linguagem de alto nível e ser considerado interessante para a aplicação para a qual se destina.
Dado	Representa uma variedade de domínios de campos colecionados ao longo do tempo de existência das organizações e armazenados eletronicamente, devido à grande quantidade e capacidade de armazenamento.
Dados anômalos	Dados resultantes de erros (por exemplo, erros de entrada de dados chaves) ou que representam eventos não usuais. Dados anômalos devem ser analisados cuidadosamente, pois podem ter informações importantes.
Dados demográficos	Dados relacionados às características pessoais, tais como idade, local de residência, quantidade de dependentes, etc.
Descoberta de conhecimento em bases de dados	Processo que descobre o conhecimento, envolvendo as seguintes tarefas: obtenção, aquisição, integração, verificação, limpeza de dados, desenvolvimento de hipótese e modelo e mineração de dados.
Desvio	Item de dado cujo valor que difere dos demais valores correspondentes na amostra, podendo indicar dados anômalos e devendo ser examinado cuidadosamente, pois podem trazer informações importantes.
Dimensão	Num banco de dados relacional, cada campo em um registro

	representa uma dimensão. Já em um banco de dados multidimensional, uma dimensão é um conjunto de entidades similares, como por exemplo, um banco de dados multidimensional de vendas provavelmente inclui as dimensões de produto, tempo e cidade.
Entropia	Medida freqüentemente utilizada em algoritmos de mineração de dados para medir a desordem em conjunto de dados.
Indução de regras	A extração de regras de condição úteis, através dos dados, baseada em significância estatística.
Informação	Dados organizados de forma significativa para quem os recebe.
Inteligência artificial	Ramo científico baseado na criação de comportamento inteligentes em máquinas.
Inteligência competitiva	Processo de descoberta de decisões estratégicas de competidores ou características da área de negócios, usando técnicas de análises quantitativas aplicadas aos dados e informações, obtidas por meios legais.
Inteligência de negócios	Ferramentas que buscam usar o conjunto de dados da organização para produzir melhores decisões nos negócios, permitindo que usuários finais e de retaguarda acessem e analisem informações armazenadas em bases de dados transacionais, dados de mercados e dados armazenados em data warehouses. Refere-se a um termo geral que cobre todos os processos, técnicas e ferramentas que suportam tomadas de decisões baseadas em decisões tecnológicas.
Limpeza de dados	O processo de certificação de que todos os valores em um banco de dados estão consistentes e corretamente arquivados.
Mineração de dados	Extração de informação e conhecimento (ocultos) em grandes bancos de dados.
Modelo	Descrição que adequadamente explica e prevê dados relevantes, geralmente é muito menor do que os próprios dados.
Modelo analítico	Estrutura e processo para a análise de um conjunto de dados. Por exemplo, uma árvore de decisão é um modelo para a classificação de um conjunto de dados.
Modelo linear	Modelo analítico que assume relacionamentos lineares nos coeficientes das variáveis estudadas.
Modelo não linear	Modelo analítico que não assume relacionamentos lineares nos coeficientes das variáveis estudadas.
Modelo preditivo	Estrutura e processo para a predição de valores de variáveis específicas em um conjunto de dados.
OLAP	Processamento analítico on line. Refere-se a aplicações de bancos de dados orientadas à matriz que permitem aos usuários recursos como visão, navegação, manipulação e análise de bancos de dados multidimensionais.
Propagação reversa	Um dos mais comuns algoritmos de aprendizado para redes neurais de treinamento.
Redes neurais artificiais	Modelos de predição não linear que aprendem através de treinamento e contemplam redes neurais biológicas em suas estruturas.
Registro	Estrutura de dados fundamental, utilizada para análise dos

	dados. Também chamado de linha de tabela ou exemplo. Um registro típico poderia ser a estrutura que contém todas as informações relevantes pertinentes a um clientes particular ou conta.
Regra de associação	Regra sobre o formato condicional (se... então... senão) que associa eventos em um banco de dados. Por exemplo, a associação entre itens comprados no supermercado.
Regressão	Técnica de analise de dados utilizada em Estatística na construção de modelos preditivos. A técnica define automaticamente a equação matemática que minimiza algumas medidas de erro entre a previsão do modelo de regressão e dos dados reais.
Regressão linear	Técnica estatística utilizada para encontrar o melhor relacionamento entre a variável dependente e as variáveis independentes.
Regressão logística	Regressão linear que prevê a proporção de uma variável categorizada, tais como tipo de cliente em uma população.
Resposta	Campo de previsão binária que indica a resposta ou não resposta a uma variedade de intervenções de marketing. O termo é geralmente usado referindo-se a modelos de resposta ou ao campo de resposta propriamente dito.
Segmentação	Processo de divisão de um conjunto de dados em grupos mutuamente exclusivos de forma que os membros de cada grupo são mais homogêneos quanto possível e os grupos entre si possuem máxima heterogeneidade, onde a distância é medida de acordo com todas as variáveis disponíveis.
Sistemas especialistas	Sistemas de suporte à decisão, baseados em especialistas humanos, na forma de regras, tendo sua interpretação realizada por computador.
Visualização de dados	Interpretação visual de relacionamentos complexos em dados multidimensionais.