UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Data Mining no Varejo: estudo de caso para loja de materiais de construção

por

ANDRÉ GUSTAVO SCHAEFFER

Trabalho de conclusão submetido à avaliação
como requisito parcial para obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. Luis Otávio Campos Alvares
Orientador

Porto Alegre, abril de 2003

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Schaeffer, André Gustavo

Data Mining no Varejo: estudo de caso para loja de materiais de construção / por André Gustavo Schaeffer. – Porto Alegre: PPGC da UFRGS, 2003.

86f.:il.

Trabalho de Conclusão (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2003. Orientador: Álvares, Luis Otávio Campos.

1. Mineração de dados. 2. Data mining. 3. Data warehousing. I. Alvares, Luis Otávio Campos. II. Título.

Agradecimentos

Agradeço a todos que, direta ou indiretamente, participaram da realização deste trabalho.

Aos proprietários, funcionários e clientes da MK Móveis e Materiais de Construção, ao meu orientador, professor Luiz Otávio Alvares, à minha namorada, Gabriela, ao meu pai Dary, à minha irmã, Tammy, e em especial à minha mãe, Maria Luiza, pela ajuda.

Sumário

Lista de Abreviaturas e Siglas

ABEMD	Associação Brasileira de Marketing Direto
ANSI	American National Standards Institute
CRISP	Cross Industry Standard Process for Data Mining
CRM	Customer Relationship Management
DMA	Direct Marketing Association
DMG	The Data Mining Group
IBM	International Business Machines Corporation
IM	Intelligent Miner
OLAP	On Line Analytic Processing
RBF	Radial Basis Prediction Mining Function
RMS	Root Mean Square
SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language
TXT	Arquivo em formato Texto
WEKA	Waikato Environment for Knowledge Analysis

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho apresenta um estudo de caso de mineração de dados no varejo. O negócio em questão é a comercialização de móveis e materiais de construção. A mineração foi realizada sobre informações geradas das transações de vendas por um período de 8 meses. Informações cadastrais de clientes também foram usadas e cruzadas com informações de venda, visando obter resultados que possam ser convertidos em ações que, por conseqüência, gerem lucro para a empresa. Toda a modelagem, preparação e transformação dos dados, foi feita visando facilitar a aplicação das técnicas de mineração que as ferramentas de mineração de dados proporcionam para a descoberta de conhecimento. O processo foi detalhado para uma melhor compreensão dos resultados obtidos. A metodologia CRISP usada no trabalho também é discutida, levando-se em conta as dificuldades e facilidades que se apresentaram durante as fases do processo de obtenção dos resultados. Também são analisados os pontos positivos e negativos das ferramentas de mineração utilizadas, o IBM Intelligent Miner e o WEKA - Waikato Environment for Knowledge Analysis, bem como de todos os outros softwares necessários para a realização do trabalho. Ao final, os resultados obtidos são apresentados e discutidos, sendo também apresentada a opinião dos proprietários da empresa sobre tais resultados e qual valor cada um deles poderá agregar ao negócio.
Palavras-chave: mineração de dados, data mining, data warehousing.

TITLE: "DATA MINING ON RETAIL MARKET: CASE STUDY FOR A BUILDING MATERIAL SHOP"

Abstract

This work presents a case study of data mining on the retail market. The issue in question is the trade of furniture and building material. The mining was performed over information generated by selling transactions in an eight months time period. Information on customer records were also used and compiled with selling information, in order to obtain results convertible in actions that consequently generate profits to the company. Every modelling, preparation and data transformation was made to ease the application of mining technics provided by the data mining tools for the knowledge discovery. The process was detailed to a better comprehension of the obtained results. The CRISP methodology used in this work is discussed as well, taking into account the difficulties and facilities that showed up during the result compilation steps. The positive and negative points of mining tools are analized as well: the IBM Intelligent Miner and WEKA – Waikato Environment for Knowledge Analysis, as well as all the other softwares needed for this work to be accomplished. At last, the obtained results are presented and discussed, as well as highlighted the company owner's opinion regarding such results and which added value each one of them will present.
Keywords: data mining, data warehousing.

1 Introdução

Tantos dados e tão poucas informações. Esta frase reflete a atual situação de grandes bases de dados existentes atualmente, nas quais o grande volume de dados é explorado somente de forma a atender os requisitos do sistema para o qual ele existe, escondendo valiosas informações que precisam de ferramentas específicas para serem descobertas, e que podem mudar, desde estratégias de negócio dentro de uma empresa, até caminhos para a cura de doenças na medicina.

Quando modelamos a base de dados de algum sistema, preparamos a mesma para nos fornecer ou fornecer ao nosso usuário informações que ele necessitará, como por exemplo, dados de clientes e volume de vendas. Estas são informações que podemos chamar de óbvias, uma vez que o sistema possui esta finalidade. A mineração de dados consiste em obter informações de valor através de uma base de dados, usando seus atributos para extrair informações que não são óbvias, mas evidenciam e mostram como aqueles dados se criam. Em outras palavras, podemos dizer que estas informações são encontradas em dados pessoais de pacientes e seus relacionamentos com diagnósticos médicos, listas com atributos de pessoas para identificar quais consumidores possuem perfil para realizar determinada compra, ou mesmo em dados financeiros que demonstram padrões de risco.

Os usuários de uma aplicação de mineração de dados podem estar interessados em informações como uma totalização dos dados que revele os seus principais padrões, como também podem estar querendo identificar fenômenos na base de dados ignorando casos triviais. Outros podem querer fazer previsões para novos casos, como por exemplo, buscar saber se o novo parceiro de negócio apresenta risco ou não.

As ferramentas de mineração de dados têm por objetivo explorar grandes bases de dados para descobrir informações úteis que estão ocultas dentro delas. Isto é feito através de métodos de classificação, análises preditivas, árvores de decisão, dentre outros. Desta forma, uma empresa que usar uma ferramenta de mineração de dados pode administrar seus dados e usá-los de uma forma mais lucrativa, isto significa aumentar as rendas, baixar os custos e aumentar a produtividade.

Quando optamos por desenvolver um projeto de mineração de dados, precisamos, além da ferramenta de mineração e dos dados a serem minerados, seguir uma metodologia de mineração. Assim seremos conduzidos durante as etapas do processo de forma a garantir resultados mais positivos no final. Basicamente, as metodologias seguem uma linha de trabalho bastante parecida, com fases-macro definidas. Estas fases são:

1. Definição do problema a ser resolvido
2. Preparação dos dados
3. Transformação dos dados
4. Geração de modelos
5. Análise de resultados

Na fase de definição do problema a ser resolvido, precisamos saber exatamente onde queremos chegar e quais condições e restrições devemos respeitar. É nesta fase que devemos responder perguntas como:

Qual problema quero resolver ?
Que dados possuo para resolver meu problema ?
Esses dados serão modificados no decorrer do projeto ?
Para quando devo entregar a resposta do problema ?

Também precisamos projetar os custos e recursos disponíveis.

Logo após, na fase de preparação dos dados, as questões já dizem respeito à quantidade de dados, origem dos dados, e o período ao qual os dados se referem. Também é o momento de levantar o formato no qual os dados se encontram, se eles estão consistentes, quais serão os campos usados e qual será o campo alvo. Na fase de preparação também deve-se verificar a pureza dos dados, criticando a ocorrência de valores incorretos, valores perdidos e delimitadores nos dados. Isto pode levar até mesmo à exclusão de campos e registros do conjunto de dados.

A fase de transformação dos dados existe para que se possa manipular os conjuntos de dados de acordo com o que se quer. Desta forma, podemos fazer uso de várias funções que os programas de mineração de dados nos proporcionam. Podemos usar funções para combinar dois conjuntos de dados distintos, para criar conjuntos de dados pela combinação de colunas de outros dois conjuntos, normalizar valores de determinada coluna, e assim por diante. Também pode ser necessário fazer uso de funções para mudar o conteúdo de um campo baseado em critérios de seleção.

Após estas três fases, o projetista está apto a iniciar a geração de um modelo de mineração de dados. Nesta fase, ele escolhe um modelo dentre os que a ferramenta lhe proporciona. Cada modelo nada mais é do que um algoritmo a ser aplicado sobre o conjunto de dados para a geração dos resultados que serão analisados na fase seguinte. Não é necessário que determinado conjunto de dados de determinado tipo de aplicação esteja associado a um único tipo de modelo (algoritmo de processamento). Pode-se inclusive aplicar um modelo a um conjunto de dados, avaliar os resultados e, se for o caso, optar pelo uso de um outro modelo disponível, até que se tenha os resultados desejados ou para que se possa avaliar qual modelo gerou os melhores resultados.

Finalmente, deve-se analisar os resultados. Nesta fase são geradas regras e/ou predições para o campo ou os campos que se deseja analisar. Em outras palavras, o software destaca de que forma os dados da variável dependente (campo alvo) são gerados a partir dos dados das variáveis independentes. São criados gráficos de classificação, gráficos baseados em regras e, até mesmo, gráficos que destacam o quanto cada variável independente influenciou no resultado. Para cada caso, deve-se considerar a margem de erro existente.

Como foi citado anteriormente, os softwares de mineração de dados seguem, ou sugerem, mais ou menos esta metodologia. Para cada aspecto e para cada fase, uns levam vantagens sobre os outros.

1.1 Objetivos desta dissertação

Este trabalho visa apresentar um estudo de caso de mineração de dados no varejo. O negócio em questão é a comercialização de móveis e materiais de construção. A mineração é feita sobre informações geradas das transações de vendas de produtos por um período de 8 meses. Informações cadastrais de clientes também são usadas e cruzadas com as informações de venda, visando obter resultados que possam ser convertidos em ações que, por conseqüência, gerem lucro para a empresa. Toda a modelagem, preparação e transformação dos dados, é feita visando facilitar a aplicação das técnicas de mineração que a ferramenta utilizada proporciona para a descoberta de conhecimento. O processo é detalhado para uma melhor compreensão dos resultados obtidos. A metodologia de trabalho também é discutida, levando-se em conta as dificuldades e facilidades que se apresentam durante as fases do processo de obtenção dos resultados. Também são discutidos os pontos positivos e negativos das ferramentas de mineração utilizadas, IBM Intelligent Miner e WEKA, bem como de todos os outros softwares que se fazem necessários para a realização do trabalho. Ao final, os resultados obtidos são apresentados e discutidos, sendo também apresentada a opinião dos proprietários da empresa sobre tais resultados e qual o valor que cada um deles pode agregar ao negócio.

Também é objetivo deste trabalho realizar um estudo de caso que contribua para a comunidade acadêmica e profissional por avaliar uma ferramenta de mineração de dados, bem como alguns de seus algoritmos, trazendo de forma clara os resultados obtidos. Destacar as dificuldades encontradas no processo de criação do trabalho é de extrema importância, visto que, hoje em dia, tanto na bibliografia como na internet, é bastante difícil encontrar estudos de caso de mineração de dados.

Finalmente, gerar informações que sejam consideradas úteis para o negócio da empresa onde está sendo aplicado o processo de mineração de dados.

1.2 Organização da dissertação

O texto desta dissertação está organizado em 6 capítulos. Ainda neste capítulo será apresentado o histórico da descoberta de conhecimento em bases de dados e sua conceituação. O processo de descoberta será abordado no capítulo 2, bem como as fases da metologia de trabalho usada - CRISP. Na seqüência, o capítulo 3 apresenta uma breve descrição sobre algumas funções de mineração de dados e suas características. O cenário do estudo de caso é tratado no capítulo 4, seguido pelo capítulo 5 onde são exibidos os resultados e todo o processo de mineração de acordo com os objetivos estabelecidos. Ao final, o capítulo 6 apresenta as conclusões obtidas com o presente trabalho e algumas sugestões para trabalhos futuros. No anexo 1 encontram-se os principais comandos de manipulação de dados usados finalizando com a bibliografia utilizada.

1.3 Histórico

Um dos princípios da mineração de dados está fundamentado na estatística clássica. Ela é a base das tecnologias sobre as quais são implementados os diversos algoritmos usados no processo de mineração de dados. A estatística clássica é formada por conceitos como *intervalos de confiança, análise de conjuntos, análise discriminante, desvio padrão, distribuição padrão e análise de regressão*, que são usados pelos softwares de mineração para o estudo dos dados bem como de seus relacionamentos.

Outro princípio da mineração de dados está fundamentado na Inteligência Artificial e suas construções heurísticas, visando o processamento dos fatos da forma como o cérebro humano faria. Este tipo de comportamento requer um vasto poder de processamento por parte das máquinas e, até o início dos anos 80, o alto custo dos computadores tornava este tipo de pesquisa bastante incomum.

Também podemos citar os conceitos e técnicas de aprendizado de máquina como fundamentos da mineração de dados, já que estas técnicas nada mais são do que combinações de princípios estatísticos e de inteligência artificial. O processo de aprendizado de máquina procura fazer com que programas de computador aprendam com os dados que eles estudam, e possam tomar decisões baseadas na qualidade dos dados estudados.

A evolução da mineração de dados foi impulsionada também pelas próprias necessidades das empresas. Nos anos 60, as empresas procuravam saber, através de seus dados armazenados, o total de suas vendas dos últimos 5 anos. Discos e fitas guardavam tais informações, e demorava-se muito tempo para consegui-las. Já nos anos 80, a introdução de sistemas gerenciadores de bancos de dados (SGBDs) e a linguagem estruturada para consultas em bancos de dados (SQL), possibilitaram às empresas o acesso a informações do tipo: *quanto vendemos do produto X , durante o mês de abril do ano passado, na filial 10?* As empresas foram buscando cada vez mais, e, nos anos 90, este tipo de informação já não era mais suficiente. Com a entrada dos conceitos de *data warehousing* e sistemas de apoio a decisão, o processo de busca de informação ficou mais dinâmico. Surgem então as ferramentas de OLAP (On Line Analytic Processing) e de análises multidimensionais. Finalmente, o tipo de informação buscada pelas empresas passou a ser do tipo: *qual tende a ser o volume de vendas de determinado produto nos próximos 2 meses e por quê ?* Ou ainda: *para que tipo de clientes devo direcionar minha campanha de marketing se quiser ter uma aceitação de, no mínimo, 60% ?* Elas então descobriram que dentro de suas bases de dados poderiam estar escondidas valiosas informações a respeito do próprio negócio com o qual trabalhavam. Isto poderia se refletir em padrões de compras de produtos, perfil de consumidor, tendências, comportamentos, e outras informações desconhecidas.

1.4 Conceituação de mineração de dados

Revisando a bibliografia disponível, podemos definir o processo de descoberta de conhecimento em bases de dados da seguinte forma: *"Descoberta de conhecimento em bases de dados é o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis dos dados."* [FRA 91]. Esta tradicional e

amplamente conhecida definição foi revista cinco anos mais tarde pelos pesquisadores Usama Fayyad, Gregory Piatestsky-Shapiro e Padhraic Smyth na obra *Advances in Knowledge Discovery and Data Mining* [FAY 96], onde cada palavra-chave da definição foi analisada objetivando uma redefinição do processo de descoberta de conhecimento em bases de dados. Devemos entender por ***dados*** todo e qualquer conjunto de fatos ou casos dentro de um banco de dados. ***Padrões*** correspondem a expressões escritas em uma determinada linguagem, e que descrevem fatos que ocorrem dentro de um subconjunto dos dados. O ***processo*** é descrito como um conjunto de etapas que envolve a preparação dos dados, busca por padrões, avaliação de conhecimento e refinamento. Por ***válidos*** devemos entender que os padrões descobertos devem ser caracterizados por algum grau de certeza e segurança. A propriedade ***novos*** sugere uma referência às mudanças pelas quais os dados passam ao longo do tempo, por comparar valores atuais com valores previstos ou esperados. ***Potencialmente úteis*** implica em padrões que conduzam a ações úteis, e que possam ser medidos por alguma função de utilidade. Por fim, os padrões devem ser ***compreensíveis,*** e isto sugere a extração dos mesmos de forma que os seres humanos possam entendê-los. Assim, o conceito de mineração de dados foi separado do conceito de descoberta de conhecimento em bases de dados da seguinte forma:

Mineração de Dados: *é um passo do processo de descoberta de conhecimento em bases de dados, e é consistido pela aplicação de algoritmos de mineração de dados em particular, sob algumas limitações de eficiência computacional aceitáveis, tendo como produto um conjunto de padrões.*

Descoberta de conhecimento em bases de dados: *é o processo de se aplicar métodos ou algoritmos de mineração de dados, para extrair ou identificar o que pode ser considerado como conhecimento de acordo com as medidas e os limites especificados, usando-se uma base de dados que pode ou não ter passado por transformações, pré-processamentos e amostragens.*

Atualmente ainda existem divergências sobre tais definições. Um exemplo seria a conceituação sugerida por [WEI 98] , em que a mineração de dados seria subdividida em duas categorias: predição de conhecimento e descoberta de conhecimento, onde a primeira trabalharia sobre fatos já ocorridos objetivando a projeção de novos casos, e a segunda usaria os dados atuais na busca por padrões ocultos. Por este ângulo, entendemos como mineração de dados não somente um passo do processo de descoberta de conhecimento, mas sim, um processo completo que passa por etapas como preparação, transformação, mineração dos dados e análises.

2 O processo de descoberta de conhecimento

O processo de descoberta de conhecimento em bases de dados é um processo interativo e iterativo, e que envolve numerosos passos e decisões tomadas pelo usuário. Os autores Brachman e Anand [BRA 96] representaram graficamente o processo de descoberta de conhecimento em bases de dados. Na figura 2.1, podemos observar a iteratividade do processo, uma vez que pode-se voltar a qualquer uma das etapas anteriores se isto se fizer necessário. O usuário também precisa interagir com o sistema depois da conclusão de cada uma das etapas.

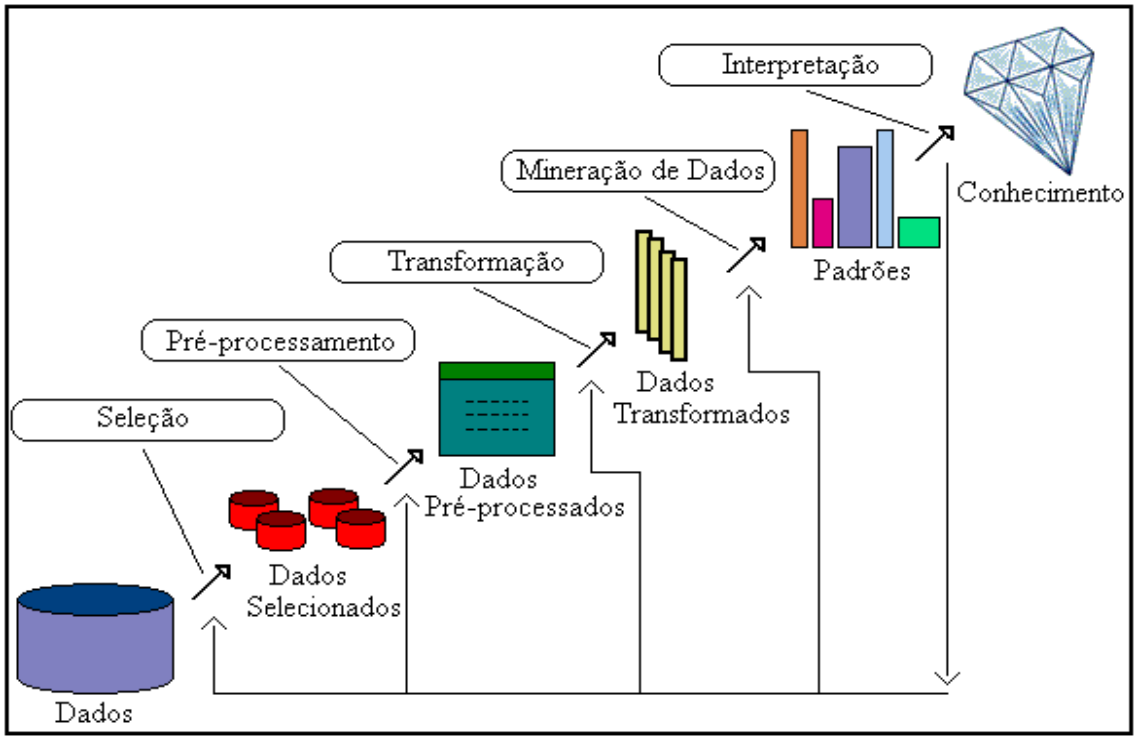

FIGURA 2.1 - O processo de descoberta de conhecimento em bases de dados

Uma visão sistemática do processo e suas fases pode ser descrita da seguinte forma:

1. Desenvolvimento e compreensão do domínio da aplicação. É necessário saber onde se quer chegar. Quais são os objetivos da mineração.
2. Criação do conjunto de dados a serem minerados. Separam-se variáveis ou atributos que podem agregar valor ao processo de mineração a partir dos dados. Esta seleção deve ser orientada pelo tipo de informação que se deseja obter.
3. Limpeza e pré-processamento dos dados, através da remoção de impurezas da base de dados, representadas através de valores nulos ou errados, campos que contém valores inválidos ou fora de intervalo.
4. Redução e projeção de dados. Nesta etapa busca-se separar ou limitar números de variáveis que serão tratadas já dentro da base de dados. Esta etapa é parecida com a etapa de seleção de dados, mas é feita num contexto mais específico, uma vez que trata de valores de atributos e não dos atributos em si.
5. Escolha da tarefa de mineração de dados. Aqui, o usuário deve optar pela geração de classificações, regressões, clusterizações, de acordo com seus objetivos. Dentro da tarefa de mineração, deve-se optar pelo algoritmo de

mineração, ou seja, se a escolha será por redes neurais, árvores de decisão, funções *radial basis*, etc...

6. Minerar os dados. Aqui, buscam-se os padrões de interesse em alguma forma de representação particular ou num conjunto delas. A execução precisa dos passos antecedentes é de fundamental importância para o sucesso desta etapa.
7. Interpretação dos padrões minerados. Trata-se da análise dos resultados obtidos com possibilidade de retroceder a qualquer uma das etapas anteriores. Este desvio para uma etapa anterior pode acontecer em qualquer outro ponto do processo, mas os problemas ou correções são mais facilmente perceptíveis aqui. Portanto, é nesta etapa que a iteração acontece com mais freqüência.
8. Consolidação do conhecimento descoberto. Este trabalho implica na incorporação dos resultados em sistemas que estão em atividade, criação de novos sistemas de informação, ou mesmo na criação de uma documentação para constar o conhecimento descoberto, a fim de apresentá-lo ao usuário ou cliente. Um fato que costuma ocorrer neste ponto é a comparação do conhecimento recentemente obtido com informações previamente existentes, trabalho este que é feito junto ao cliente ou usuário final.

No início dos estudos e pesquisas na área da descoberta de conhecimento em bases de dados, a etapa da mineração de dados recebia uma ênfase muito maior. Isto foi mudando com o passar do tempo, quando percebeu-se a importância das fases precedentes para o sucesso do processo como um todo.

2.1 A Metodologia CRISP

Esta metodologia surgiu em 1996 e foi criada por pesquisadores das empresas Daimler Chrysler , SPSS e NCR. Em meio a um mercado jovem e imaturo de mineração de dados, a Daimler Chrysler, sob a coordenação de Thomas Reinartz e Rüdiger Wirth, já fazia experimentos com mineração sobre seus dados comerciais. A SPSS, criadora do primeiro software comercial para mineração de dados, o Clementine, também participou da concepção desta metodologia através de seus pesquisadores Julian Clinton, Thomas Khabaza e Colin Shearer. Com os mesmos objetivos, a NCR, através de Pete Chapman e Randy Kerber, participou da criação desta metodologia para extração de conhecimento em bases de dados. As fases que compõe o processo de mineração de dados propostas pela metodologia CRISP serão descritas adiante conforme [CHA 00]. O estudo de caso proposto por esta dissertação foi realizado com base nessa metodologia, e os resultados obtidos em cada fase do processo serão apresentados posteriormente. A figura 2.2 demonstra o ciclo de vida de um projeto de mineração de dados.

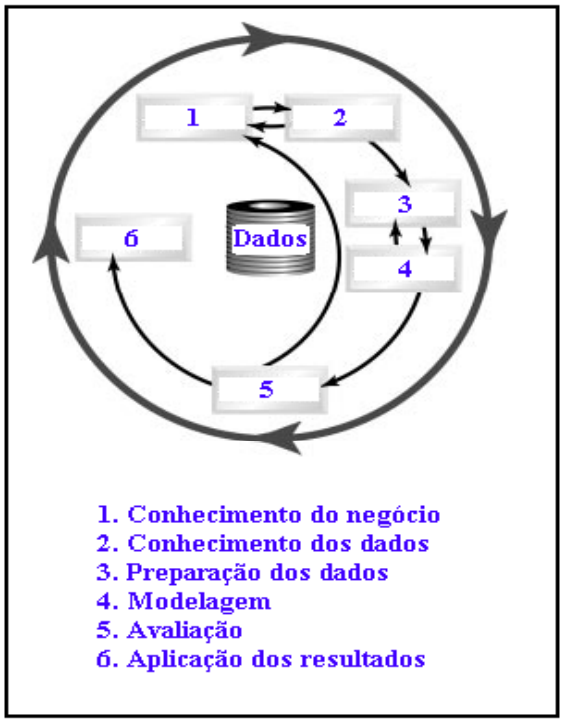

FIGURA 2.2 - Ciclo de vida de um projeto de mineração de dados

Como podemos ver, a metodologia CRISP divide o projeto de mineração de dados em seis fases, que serão descritas a seguir.

2.1.1 Conhecimento do negócio

Nesta fase inicial, o usuário deve estabelecer os objetivos e necessidades, transformando este conhecimento na definição de um problema de mineração de dados bem como criando um plano preliminar de como atingir os objetivos definidos.

2.1.1.1 Determinando os objetivos do negócio

Tarefa: Identificar junto ao cliente o que ele realmente deseja. Seus objetivos e restrições devem ser balanceados. Se esta tarefa não for bem feita, pode-se chegar ao fim do projeto com as respostas certas para as perguntas erradas.
Resultados: ***1) Conhecimento*****:** o conhecimento do negócio e da situação da empresa no início do projeto é então adquirido. ***2) Objetivos do negócio****:* o objetivo primário do cliente é então descrito. ***3) Critério de sucesso do negócio***: descreve o que pode ser considerado como um resultado útil do ponto de vista do negócio.

2.1.1.2 Avaliando a situação

Tarefa: saber em detalhes os fatos a respeito dos recursos, restrições, suposições e outros fatores importantes para atingir os objetivos do projeto.
Resultados: ***1) Inventário de recursos****:* lista os recursos disponíveis para serem usados no projeto, incluindo pessoal especializado no negócio do cliente, analistas de dados, suporte técnico, administradores de banco de dados, analistas de data warehouse, recursos computacionais, dados, ferramentas e softwares relacionados. ***2) Necessidades, suposições e restrições :*** relata a qualidade dos resultados bem como

questões que dizem respeito ao acesso às informações e aos dados. Neste ponto devem estar claras as questões de uso legal dos dados e resultados. Restrições quanto à manipulação dos dados em relação ao tamanho da base também são levantadas aqui. ***3) Riscos e contingências*** **:** demonstram os riscos aos quais o projeto estará sujeito e os planos de contingência associados. ***4) Terminologia :*** trata-se de uma composição de termos que serão usados no projeto para a compreensão do negócio da empresa bem como para a compreensão da terminologia técnica de mineração de dados. ***5) Custos e benefícios :*** demonstra a relação custo-benefício do projeto.

2.1.1.3 Determinando os objetivos da mineração de dados

Tarefa: demonstrar os objetivos do projeto em termos técnicos.
Resultados: ***1) Os objetivos da mineração de dados :*** descreve os resultados pretendidos com o projeto que possibilitam atingir os objetivos do negócio. ***2) Critério de sucesso da mineração de dados :*** define os critérios de sucesso do projeto de mineração de dados em termos técnicos.

2.1.1.4 Produzindo um plano de projeto

Tarefa: especificar os passos a serem executados no decorrer do projeto, incluindo uma seleção de técnicas e ferramentas necessárias.
Resultados: ***1) Plano de projeto :*** lista os estágios pelos quais o projeto irá passar, juntamente com a sua duração, recursos necessários e dependências. ***2) Avaliação inicial de ferramentas e técnicas :*** aqui ocorre a seleção da ferramenta de mineração de dados. Esta avaliação das ferramentas e técnicas pode influenciar o projeto como um todo.

2.1.2 Conhecimento dos dados

É nesta fase que é feita a coleta inicial dos dados, identificação de problemas na qualidade dos mesmos, e também a identificação dos principais subconjuntos de dados. Nesta fase começa a ser estabelecida uma familiarização com as informações.

2.1.2.1 Coletando os dados iniciais

Tarefa: carregar os dados iniciais para dentro da ferramenta selecionada.
Resultados: ***1) Relatório inicial da coleta de dados :*** detalha o conjunto de dados, as formas de extração usadas e os problemas encontrados, bem como as respectivas soluções.

2.1.2.2 Descrevendo os dados

Tarefa: examinar superficialmente o conjunto de dados como um todo e relatar os resultados.
Resultados: ***1) Relatório de descrição dos dados :*** descreve os dados adquiridos incluindo o formato, quantidade, campos de tabelas. Aqui deve ser possível dizer se os dados obtidos servem ou não para realizar o trabalho.

2.1.2.3 Explorando os dados

Tarefa: explorar a base de dados já visualizando, por exemplo, os atributos alvos de predição, relacionamentos entre variáveis e análises estatísticas simples.
Resultados: ***1) Relatório de exploração de dados :*** relata resultados como hipóteses iniciais e seus impactos sobre o resto do projeto.

2.1.2.4 Verificando a qualidade dos dados

Tarefa: examinar a qualidade dos dados, respondendo questões como: *os dados estão completos ?* Ou ainda: *estão corretos ? Existem valores perdidos ou faltando ? Em que freqüência ?*
Resultados: ***1) Relatório de qualidade de dados :*** reporta a verificação da qualidade dos dados. Se existem problemas, este relatório mostra as possíveis soluções.

2.1.3 Preparação dos dados

A fase de preparação dos dados cobre todas as atividades que determinam a construção do conjunto de dados final. Tais atividades poderão ser realizadas múltiplas vezes ou aleatoriamente, de acordo com a necessidade. Seleções de tabelas, registros e atributos, bem como a transformação e limpeza dos dados podem ser citadas como atividades desta fase.

2.1.3.1 Selecionando os dados

Tarefa: escolher os dados que serão usados para análise. Isto inclui a seleção de dados (registros), e as colunas (atributos) de uma tabela.
Resultados: ***1) Análise lógica para inclusão ou exclusão:*** lista dos dados a serem incluídos e excluídos e o motivo da decisão.

2.1.3.2 Limpando os dados

Tarefa: obter dados com qualidade suficiente para que possam ser minerados.
Resultados: ***1) Relatório da limpeza de dados:*** descreve as decisões e ações tomadas para identificar as impurezas nos dados. Também deve relatar quais transformações foram necessárias e o impacto na análise de resultados.

2.1.3.3 Construindo os dados

Tarefa: preparar operações a serem executadas sobre os dados para a produção de atributos derivados, novos registros ou valores transformados a partir de atributos existentes.
Resultados: ***1) Atributos derivados:*** são novos atributos construídos a partir de um ou mais atributos existentes do mesmo registro. Por exemplo, *área = base X altura.* ***2) Registros gerados:*** descreve a geração de novos registros com o propósito de representar fatos dos quais não se tem registro mas que certamente ocorreram.

2.1.3.4 Integrando os dados

Tarefa: combinar várias fontes de dados ou tabelas para gerar novos dados ou valores.
Resultados: ***1) Dados combinados:*** trata-se de uma tabela ou conjunto de dados que não eram encontrados nos dados iniciais mas que são importantes para atingir os objetivos do projeto.

2.1.3.5 Formatando os dados

Tarefa: formatar os dados sem modificar seu significado, visando apenas adaptar os mesmos às necessidades da ferramenta de mineração.
Resultados: ***1) Dados reformatados:*** Dados reformatados, sejam eles ordenados por algum critério ou mesmo alterados randomicamente para serem usados por algum algoritmo de rede neural, ou alterados pela exclusão de vírgulas visando atender uma necessidade específica da ferramenta.

2.1.4 Modelagem

Nesta fase ocorre a aplicação de técnicas de modelagem e também a sintonia das mesmas a nível de acerto de parâmetros. Existem várias técnicas para resolver o mesmo problema de mineração de dados, e, algumas delas, exigem uma reformatação dos dados. Por isso, talvez seja necessário retroceder à fase de preparação de dados.

2.1.4.1 Selecionando a técnica de modelagem

Tarefa: escolher a técnica de modelagem a ser usada de acordo com as necessidades existentes.
Resultados: ***1) Técnica de modelagem:*** Documento reportando as características da técnica de modelagem que será usada. ***2) Suposições da técnica de modelagem escolhida:*** relato de todas as considerações e requisitos necessários sobre os dados para que sejam minerados através da técnica escolhida. Por exemplo: atributos com distribuição uniforme, valores perdidos não permitidos e normalizações.

2.1.4.2 Gerando o projeto-teste

Tarefa: gerar um esquema de validação do modelo.
Resultados: ***1) Projeto-teste:*** deve descrever o plano ou idéia pretendida para executar a validação do modelo. A primeira tarefa é separar o conjunto de dados a ser usado como treino para o modelo, teste e validação do mesmo.

2.1.4.3 Construindo o modelo

Tarefa: utilizar a ferramenta para criar o modelo ou os modelos necessários.
Resultados: ***1) Conjunto de parâmetros:*** relacionar os parâmetros usados bem como os valores escolhidos para os mesmos. ***2) Modelos:*** são os modelos produzidos pela ferramenta. ***3) Descrição do modelo:*** um relatório contendo a interpretação do modelo bem como a descrição das dificuldades encontradas e seus significados.

2.1.4.4 Avaliando o modelo

Tarefa: interpretar os modelos de acordo com os domínios de conhecimento. Isto implica em interpretações a nível técnico e a nível de negócio, com a presença de analistas responsáveis. A avaliação dos modelos deve levar em conta os objetivos do negócio e os critérios de sucesso.
Resultados: ***1) Avaliação do modelo:*** demonstra por ordem de qualidade os modelos avaliados para determinar qual o que trará os melhores resultados. ***2) Revisão do conjunto de parâmetros:*** baseado nas avaliações dos modelos, deve-se gerar uma relação das alterações nos parâmetros que se fizeram necessárias para as próximas execuções.

2.1.5 Avaliação

Na fase de avaliação, os modelos até então considerados bons devem ser analisados. Ao final, deve-se decidir sobre o uso dos resultados da mineração de dados aplicada.

2.1.5.1 Avaliando os resultados

Tarefa: avaliar os resultados gerados com foco nos objetivos estabelecidos, e, no caso de insucesso, especificar o motivo da deficiência encontrada.
Resultados: ***1) Avaliação dos resultados da mineração em relação ao critério de sucesso estabelecido:*** reportar se o projeto de mineração de dados já atingiu os objetivos aos quais se destinava. ***2) Modelos aprovados:*** lista dos modelos usados para atingir os objetivos desejados.

2.1.5.2 Revendo o processo

Tarefa: rever o processo de mineração como um todo objetivando determinar a existência de fatores importantes dos quais não se teve conhecimento.
Resultados: ***1) Revisão do processo:*** descrição das atividades e fatores perdidos que, por exemplo, precisarão ser revistos ou refeitos.

2.1.5.3 Determinando os passos seguintes

Tarefa: avaliar a situação atual do projeto, levando em consideração todos os resultados obtidos, os objetivos que já foram atingidos e o orçamento ainda disponível, para decidir se já pode-se finalizar o trabalho ou deve-se começar algum novo projeto.
Resultados: ***1) Lista de possíveis ações:*** detalhamento das ações que devem ser tomadas juntamente com as razões da existência das mesmas. ***2) Decisão:*** descreve o que foi escolhido como próxima ação.

2.1.6 Aplicação dos resultados

O conhecimento extraído deve ser organizado e apresentado de forma que o cliente possa tirar proveito da melhor forma possível. Esta apresentação pode variar desde a simples exibição de um relatório até o desenvolvimento de uma aplicação a ser

anexada em páginas Web. Em muitos casos, o processo de aplicação dos resultados é conduzido pelo cliente e não pelo analista.

2.1.6.1 Planejando a aplicação dos resultados

Tarefa: desenvolver uma estratégia de aplicação dos resultados encontrados.
Resultados: ***1) Plano de aplicação dos resultados:*** descrição da estratégia de aplicação dos resultados e os passos necessários para executá-la.

2.1.6.2 Planejando a monitoração e a manutenção

Tarefa: determinar um plano para manter o produto da mineração de dados em atividade e gerando resultados.
Resultados: ***1) Plano de manutenção e monitoramento:*** descrição dos passos e tarefas que serão executados com objetivo de monitorar e manter em uso o conhecimento descoberto.

2.1.6.3 Produzindo o relatório final

Tarefa: reportar todos os resultados obtidos bem como as experiências adquiridas com o projeto.
Resultados: ***1) Relatório final:*** listagem organizada dos resultados e experiências adquiridas. ***2) Apresentação final:*** encontro com o cliente e demais envolvidos para uma apresentação verbal dos resultados do projeto.

2.1.6.4 Revendo o projeto

Tarefa: rever o projeto de mineração de dados, atentando para o que foi realizado com sucesso e o que precisa ser melhorado.
Resultados: ***1) Documentação da experiência:*** listagem com uma explicação do aprendizado obtido ao longo da execução do projeto de mineração de dados, para ser usado como base para trabalhos futuros.

3 Funções de Mineração de Dados do IBM Intelligent Miner

No IBM Intelligent Miner, ferramenta usada neste estudo de caso, existem seis tipos de funções de mineração: *Association, Classification, Clustering, Prediction, Sequential Pattern e Similar Sequences.* O método ou algoritmo disponível para uso varia de função para função. Por exemplo, na função de *Clustering* pode-se optar por um Cluster Demográfico ou por um Cluster Neural, enquanto que na função *Prediction* os métodos de mineração são Predição Neural e Predição por função de Base Radial. Todas serão descritas a seguir.

3.1 Associations mining function

O propósito de se descobrir associações é encontrar itens em uma transação que implicam na presença de outros itens na mesma transação. Por exemplo, descobrir que 65% dos clientes que compram cartões postais também compram cosméticos. As afinidades são expressadas na forma de regras do tipo X => Y (leia-se *X implica em Y*). Pode-se gerar restrições às gerações de regras definindo-se que certos campos não devem aparecer nas regras geradas, ou que determinados campos devem aparecer.

3.2 Demographic clustering mining function

O propósito da descoberta de clusters (agrupamentos) é agrupar registros com características similares. Existem dois tipos de campos na entrada de dados para esta função: os campos ativos e os campos suplementares. Os campos ativos são usados pela função para fazer o agrupamento, enquanto que os campos suplementares são usados para fins estatísticos. Os campos em um cluster são ordenados por importância. Algumas vezes os campos suplementares influenciam mais do que os campos ativos no resultado final. Como campos ativos deve-se evitar escolher aqueles que possuem diferentes valores para quase todos os registros, ou com o mesmo valor em todos os registros. O usuário deve também especificar o peso de cada campo pois, caso não o faça, os mesmos pesos são atribuídos para cada campo. Também deve-se atribuir pesos aos valores que podem aparecer no conjunto de dados, uma vez que valores que raramente aparecem podem contribuir mais significativamente para o resultado.

3.3 Neural clustering mining function

Esta função emprega um processo de agrupamento conhecido como Kohonen Feature Map Neural Network, que utiliza uma auto–organização para agrupar registros de entradas similares. A principal tarefa deste processo é encontrar o centro de cada cluster. Este centro também é conhecido como protótipo. Para cada registro nos dados de entrada, o algoritmo Neural Clustering calcula o protótipo que está mais perto do registro. Um valor próximo a zero indica alto grau de similaridade. Quanto maior o valor, maior é o grau de diferenças entre o registro e o protótipo. Como característica das Neural Nets, os dados de entrada devem ser normalizados para intervalos entre 0.0 e 1.0. Valores de categoria devem ser convertidos para números.

3.4 Sequential patterns mining function

O propósito de descoberta de padrões seqüenciais é encontrar padrões previsíveis de comportamento em um período de tempo. Isto significa que um certo comportamento em determinado instante tende a produzir outro comportamento ou uma seqüência de comportamentos em outro instante. Por exemplo, pode-se verificar que 52% dos clientes de contas pré-pagas de telefones celulares recarregarão seus créditos após 60 dias e 35% deles somente após 90 dias passados da primeira carga. Nas consultas a este tipo de função, pode-se fazer restrições, como visualizar somente padrões seqüenciais que incluem determinados itens e possuem uma ocorrência relativa de 5%, e com um determinado tamanho padrão máximo.

3.5 Similar sequences mining function

O objetivo de descoberta de seqüências similares é encontrar todas ocorrências de subseqüências similares em dados de seqüências. Por exemplo, um mercado pode estar interessado em otimizar seu processo de compras e controle de estoque. Após a mineração, descobre-se a similaridade nas vendas de pares de produtos e em qual grau. Pode-se então fazer uma previsão de compras para o próximo ano baseado nas quantidades em estoque atuais. Os seres humanos são capazes de reconhecer com mais facilidade do que as máquinas as similaridades entre desenhos e figuras. A substituição dessa noção intuitiva de similaridade por noções matemáticas é uma tarefa bastante complexa. Encontrar similaridades em um conjunto de dados depende de quão detalhados os dados são. Pode-se encontrar similaridades apenas se dados irrelevantes, como erros, forem desprezados.

3.6 Tree classification mining function

O propósito de predizer uma classificação é criar um modelo baseado em dados conhecidos. Pode-se usar este modelo para analisar porquê certa classificação foi feita, ou para calcular a classificação dos novos dados. Dados históricos geralmente consistem de conjuntos de valores e uma classificação para esses valores. A análise de dados já classificados revela as características que conduziram àquela classificação. O modelo de classificação resultante pode ser usado, então, para predizer as classes de registros contendo os novos valores. Por exemplo, uma empresa de seguros pode usar a mineração de dados através deste modelo para descobrir se o novo cliente não possui o perfil daqueles clientes que pertencem a um grupo de risco. Esta técnica também pode ser usada para detectar fraude em cartão de crédito, identificar defeitos em imagens ou diagnosticar condições de erro. Sempre pode-se também, atribuir o grau de confiança, que pode variar de 0,0 até 1,0, que indica confiança total. Pode-se atribuir pesos às variáveis de entrada mas deve-se observar que a atribuição de pesos influi consideravelmente no resultado final.

3.7 Neural classification mining function

A função de classificação neural emprega o algoritmo *back–propagation neural network* para fazer a classificação dos dados. A classificação é baseada no valor da classe e nos relacionamentos descobertos dos dados por uma mineração previamente

executada. O Intelligent Miner permite habilitar uma opção de *optimize for time* durante o uso de funções que envolvam redes neurais. Isto é particularmente útil uma vez que esses tipos de algoritmos processam os dados mais vezes que os outros tipos. Como acontece nos outros modelos baseados em redes neurais, os dados de entrada devem ser normalizados para valores entre 0,0 e 1,0. Valores de categoria devem ser convertidos para uma representação numérica.

3.8 RBF prediction mining function

O propósito das predições pela função de base radial (Radial-Basis) é descobrir a dependência e a variação de um valor de um campo em relação aos valores dos outros campos do mesmo registro. O modelo gerado pode predizer o valor de determinado campo em um mesmo registro de mesmo formato, baseado nos outros campos. Em marketing, esses modelos podem revelar que, para alguns consumidores, campanhas de incentivo aumentam as vendas, e visitas freqüentes feitas por representantes comerciais geram baixa nas vendas se o consumidor for jovem. O algoritmo *Radial-Basis* trabalha sobre os chamados centros de adaptação (fitting centers). Um centro de adaptação é um vetor no espaço de atributos. Dentro de cada um desses centros, uma função é definida. Parâmetros como número máximo de centros e tamanho máximo de região são particulares desse tipo de algoritmo, e são usados durante a fase de treinamento do modelo.

3.9 Neural Prediction mining function

Possui propósitos iguais aos das outras funções de predição, empregando, porém, o uso das redes neurais. Emprega o algoritmo de *Back-Propagation Neural Network* para fazer predição de valores. Também é possível usar o recurso de *optimize for time* para reduzir o tempo de processamento. Dados de entrada devem ser normalizados e valores de categoria devem ser convertidos para uma representação numérica.

4 Cenário do Estudo de Caso

O presente estudo de caso foi realizado tendo por base as informações geradas das transações de vendas da MK Móveis e Materiais de Construção, estabelecida em Erechim-RS. Os registros usados na mineração de dados foram coletados de maio a dezembro de 2002, representando um total aproximado de 40.100 itens vendidos em 6.422 transações de venda. Também será apresentada neste capítulo a evolução da modelagem dos dados, os processos de refinamento e limpeza dos mesmos, as formas de relacionamento entre tabelas, os modelos de mineração escolhidos e os resultados obtidos. Todas as etapas do processo de mineração de dados serão descritas a seguir, da forma como sugere a metodologia CRISP, que foi apresentada no capítulo 2.

A seguir serão destacadas as etapas do processo de mineração de dados deste estudo de caso.

4.1 Conhecimento do negócio

O primeiro passo do processo é o conhecimento do negócio. É necessário analisar como ocorrem as transações de vendas, os pontos pelos quais um cliente passa até efetivar uma compra, a forma como é prestado o atendimento, a situação atual e as necessidades da empresa. A seguir, estes pontos serão abordados.

A MK Móveis e Materiais de Construção

Com 18 anos de experiência no ramo de móveis e materiais de construção, a empresa conta com uma loja e uma fábrica de móveis, vendendo para todo o estado do Rio Grande do Sul, Santa Catarina, Paraná e Minas Gerais. Os dados usados neste estudo de caso referem-se às vendas efetuadas na loja por clientes finais. Produtos para revenda não foram considerados. A loja conta com 6 funcionários que atendem diretamente o público, incluindo os proprietários. O atendimento ao público ocorre de segunda a sábado das 08:00 às 19:00, com uma circulação média diária de 60 pessoas, entre clientes que compram e clientes que não compram. Podemos destacar basicamente duas formas de atendimento:

1) clientes solicitam ser atendidos por algum funcionário em particular: venda de mercadorias com maior custo são exemplos de situações em que o cliente, na maioria das vezes, solicita atendimento por algum funcionário específico, por possuir mais afinidade ou confiança.
2) clientes são atendidos pelo funcionário que estiver disponível: a abordagem ao cliente é feita assim que o mesmo entra na loja. Se o mesmo preferir, pode solicitar que o atendimento seja feito por outro funcionário. Porém, para compras de menor custo ou mesmo para compras em que o cliente não dispõe de muito tempo, o funcionário que o atendeu primeiro acaba conduzindo a transação de venda até o fim.

Não existe nenhum tipo de bonificação por comissão aplicável sobre as vendas em benefício dos funcionários.

A estrutura organizacional da loja pode ser assim definida:

- Proprietários: Rolfi e Adenor
- Gerente: Inês
- Funcionários: Edgar, Guilherme e Lucinéia

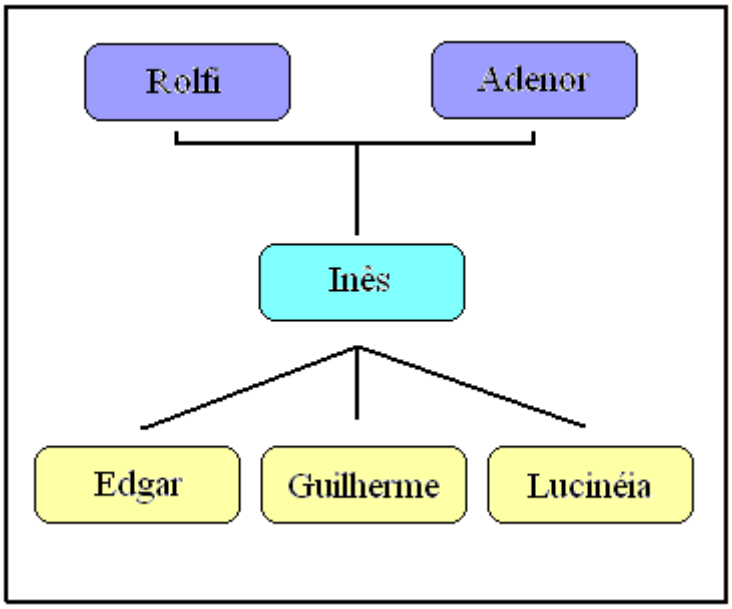

FIGURA 4.1 - A estrutura organizacional da MK Móveis e Materiais de Construção

Quanto ao grau de informatização atual, a empresa dispõe de um sistema em ACCESS responsável pelo registro de vendas, faturamento, controle de estoque, contas a receber e cadastro de clientes. A modelagem da base de dados bem como as alterações que se fizeram necessárias para o projeto, serão vistas mais adiante.

4.1.1 Objetivos do negócio

No início, foi apresentado aos proprietários o conceito de mineração de dados e os benefícios que poderiam resultar caso um trabalho fosse realizado. Alguns exemplos de sucesso tanto no Brasil como em nível mundial foram mostrados destacando a forma como a empresa poderia se beneficiar. As funções mineração de dados (classificações, agrupamentos, predições) foram demonstradas através de exemplos.

O conhecimento que a empresa possui dos seus clientes é trivial. Sabe-se muito bem quem são os clientes que compram com mais freqüência, quais são os clientes que realizam as maiores compras, e, claro, quais são os inadimplentes. Atualmente, a empresa veicula no rádio algumas propagandas da loja, mas não tem costume de anunciar produtos específicos. Na televisão, foram realizados comerciais há muito tempo atrás. Este meio de comunicação, portanto, não é usado como forma de divulgar a empresa e seus produtos. Um canal de marketing usado pela empresa para divulgar seus produtos é a mala-direta. Isto é feito através de correspondências e folders que apresentam produtos em promoção bem como os produtos que dão maior retorno financeiro. Todos os clientes cadastrados costumam receber correspondências, com exceção dos inadimplentes. Existe ainda o grupo de clientes inadimplentes que estão nessa situação apenas por não pagarem em dia suas contas. Porém, são clientes que

pagam e realizam compras com certa periodicidade. Estes não são vistos como maus pagadores e recebem, assim como os bons pagadores, as correspondências e anúncios via mala-direta.

Uma outra forma similar de marketing que a empresa realiza é a contratação temporária de pessoas que distribuem folders em determinados bairros e ruas da cidade. Este tipo de propaganda, apesar de ser apresentado da mesma forma (via folders), é diferente da mala-direta, uma vez que o público-alvo não é composto somente por clientes que já realizaram compras, mas também por clientes que nunca compraram ou nem sabem da existência da empresa. Como desvantagem, podem existir clientes com potencial de compra que residem em determinado ponto da cidade onde não foi distribuído o folder.

Com o conhecimento das vantagens proporcionadas pela aplicação da mineração de dados, foram estabelecidos, juntamente com os proprietários da empresa, os seguintes objetivos:

1. conhecer o perfil do cliente que compra na loja.
2. conhecer o perfil do cliente associado com as compras que o mesmo faz na loja.
3. criar uma lista dos produtos mais rentáveis e verificar quais os clientes com mais tendência a realizar compras desses produtos.
4. através de uma análise de cesta de mercado, conhecer quais produtos estão associados em transações de venda.
5. tentar prever se o cliente que está comprando pela primeira vez na loja tende a não voltar mais.
6. conhecer o perfil dos clientes que compram pela primeira vez na loja.

Sabe-se que as empresas de construção civil pertencem ao grupo dos melhores clientes, mas não se tem conhecimento de outros grupos de bons clientes. O perfil do cliente que compra e não retorna também é interessante, pois é necessário que se saiba qual foi o motivo que fez com que ele não realizasse mais compras na loja. A análise de cesta de mercado também foi vista como uma tarefa importante, uma vez que o número de produtos é bastante grande. Esperava-se descobrir muitas associações de vendas além de cimento, tijolos e areia.

Neste ponto do projeto, é necessário estabelecer alguns pré-requisitos. Juntamente com os proprietários, foi colocado que, para o sucesso do projeto, seria necessário:

1. que todos os funcionários se envolvessem e compreendessem o trabalho que estaria sendo desenvolvido;
2. que todos os funcionários estivessem motivados e colaborassem com as informações que se fizessem necessárias;
3. que o sistema atual de faturamento fosse alterado para contemplar as informações necessárias para o projeto, como alterações no cadastro de clientes e no registro de vendas;

Um ponto bastante importante que foi levantado é que o cliente, em hipótese alguma, deveria se sentir perturbado com as perguntas que fossem a ele dirigidas, tanto no momento do cadastro quanto para o registro de vendas. O sistema foi também

preparado para que o tempo necessário de interação com o cliente fosse o menor possível, visando agilizar a coleta de dados. O objetivo das perguntas sempre foi devidamente destacado para evitar que bons clientes pudessem se ofender ao responder seu endereço ou telefone, pensando que a empresa estivesse questionando sua situação financeira. De um modo geral, não foram encontrados problemas que dissessem respeito a este tipo de problema. Pelo contrário, o estudo era sempre bem visto e os clientes faziam questão de participar.

Avaliando os objetivos, todos eles foram vistos como sendo possivelmente alcançáveis.

Quanto aos critérios de sucesso, ficou definido que qualquer ganho em relação às experiências de marketing anteriores seria visto como uma grande vantagem para o negócio da empresa.

4.1.2 Avaliação da situação atual da empresa

Para avaliar corretamente a situação da empresa e o que ela possui para contribuir com o projeto, serão levantados abaixo todos os recursos disponíveis, desde recursos de hardware até pessoal especializado no negócio da empresa. Também será identificada a plataforma de mineração de dados, ou seja, o hardware e os softwares usados na aplicação dos modelos e extração dos resultados.

4.1.2.1 Recursos de Hardware na empresa

1. Micro computador Pentium III com 64Mb de RAM, HD IDE de 20Gb
2. Impressora HP Desk Jet 692C
3. Impressora matricial Citizen 200Gx para notas fiscais
4. Zip Drive para backup da base de dados

Todos estes recursos de hardware ficaram disponíveis para uso durante todo o projeto, cuidando-se para que a utilização fosse feita fora do horário de expediente da loja.

4.1.2.2 Recursos de Software na empresa

1. Microsoft Windows 98
2. Microsoft Access
3. Microsoft Excel
4. Microsoft Word

4.1.2.3 Recursos de Hardware para a mineração

Todas as etapas do processo de mineração de dados foram realizadas fora da loja, tendo feito uso do seguinte hardware:

1. Micro computador Pentium III com 256Mb de RAM, HD IDE de 40Gb
2. Impressora HP Desk Jet 656C
3. Gravador de CD LG 24x10x40 para backup dos dados

4.1.2.4 Recursos de Software para a mineração

1. Microsoft Windows 2000 Professional
2. Microsoft Excel
3. Microsoft Word
4. Microsoft Access
5. Oracle 9i Database Server
6. Oracle SQL Plus
7. IBM DB2 Intelligent Miner for Data Version 6.1
8. IBM DB2 Universal Database Personal Edition
9. Roxio Easy CD creator
10. Borland C++ 3.1

4.1.2.5 A base de dados da empresa

A fonte atual de informações da empresa estava estruturada sobre o Microsoft Access, e continha informações sobre os clientes, produtos, vendas e estoque. O cruzamento de tabelas da base de dados conseguiria gerar as informações necessárias para atender plenamente os objetivos que estavam sendo propostos, porém, é claro, precisariam sofrer várias transformações. Não foi encontrada nenhuma dificuldade para realizar conversões do formato .MDB do Access para planilhas Excel e para arquivos-texto que pudessem ser lidos pela ferramenta de mineração, ou mesmo para gerar comandos SQL de manipulação de dados na base de mineração (INSERT, UPDATE, DELETE, SELECT). Evidentemente, todas as alterações na base de dados bem como alterações em programas que se fizessem necessárias, teriam que ser feitas pelo autor do atual sistema da empresa.

4.1.3 Objetivos da mineração de dados

A definição dos objetivos da mineração de dados confunde-se um pouco com a definição dos objetivos do negócio. Quando se define os objetivos do negócio, o que se tem em mente é o retorno do investimento para a empresa ou, mais especificamente, o lucro que se vai obter depois da aplicação da estratégia de marketing montada com base nos resultados obtidos. Desta forma, os objetivos da mineração de dados passam a ser a construção de modelos precisos e com o mínimo erro possível, para que possam gerar os resultados mais corretos, de forma ágil e clara. Se os objetivos do negócio forem atingidos, os objetivos da mineração, muito provavelmente, também terão sido atingidos.

4.2 Conhecimento e Preparação dos dados

Nesta etapa do projeto, iniciou-se o contato com a base de dados atual da empresa, visando identificar problemas na qualidade da mesma, subconjuntos de dados e relacionamentos entre tabelas. Os resultados deste estudo serão descritos a seguir.

4.2.1 Coletando os dados iniciais

Não foi feito uso de nenhuma ferramenta específica de visualização ou manipulação de dados. A tarefa neste ponto do projeto foi somente extrair os dados e convertê-los para o formato texto, visando obter um primeiro contato com aquela que seria a forma de integração entre os dados do atual sistema da empresa com a ferramenta de mineração. Na prática, foi realizada uma carga dos dados para dentro de tabelas do banco de dados Oracle e a transformação dos mesmos foi feita via comandos SQL. Isto será relatado mais adiante. Aqui, neste ponto do projeto, somente foi constatada a possibilidade de integrar os dados do sistema da empresa com o IBM Intelligent Miner. Não existia ainda, neste ponto, nenhuma alteração na base de dados para inclusão de novos atributos ou relacionamentos. Mas esta tarefa possibilitou visualizar quais os passos para migrar os dados a serem minerados. Isto parece uma tarefa simples, mas existem inúmeros problemas que podem ocorrer, como por exemplo, problemas na formatação das colunas ou perdas de acentuação. Porém, neste estudo de caso, estes problemas não ocorreram.

4.2.2 Descrevendo os dados

A base de dados atualmente usada pela empresa contemplava suas necessidades. Porém, para atender aos objetivos definidos, várias alterações se fizeram necessárias. Trabalhou-se com dois contextos no que diz respeito a dados: as informações cadastrais de clientes, e as informações das transações de vendas. Todo o trabalho foi conduzido deste ponto até a sua conclusão visando gerar estes dois conjuntos de dados. O produto cartesiano destes dois conjuntos de dados geraria uma tabela onde, para cada linha, poderia-se relacionar qualquer informação cadastral de um cliente com qualquer dado de uma transação de venda que aquele cliente teria realizado. As figuras 4.2 e 4.3 descrevem os campos selecionados da base de dados da empresa considerados úteis para a mineração, bem como aqueles que foram agregados ao sistema e implicaram na alteração de programas e da base de dados:

```
SQL> desc cliente
 Nome                                      Nulo?    Tipo
 ----------------------------------------- -------- ------------
 CLIENTE                                   NOT NULL VARCHAR2(44)
 PRIMEIRAVEZ                                        VARCHAR2(7)
 PROFISSAO                                          VARCHAR2(44)
 ESTADOCIVIL                                        VARCHAR2(14)
 TEMPODECASADO                                      VARCHAR2(6)
 NODEFILHOS                                         VARCHAR2(6)
 FAIXAETARIADOSFILHOS                               VARCHAR2(15)
 TELEFONE                                           VARCHAR2(14)
 EMAIL                                              VARCHAR2(34)
 ENDERECO                                           VARCHAR2(44)
 CASAAPTO                                           VARCHAR2(10)
 BAIRRO                                             VARCHAR2(34)
 CIDADE                                             VARCHAR2(24)
 UF                                                 VARCHAR2(6)
 MORAQUANTOTEMPO                                    VARCHAR2(6)
 RESIDENCIAPROPRIAALUGADA                           VARCHAR2(11)
 IDADE                                              VARCHAR2(10)
 PESO                                               VARCHAR2(15)
 ALTURA                                             VARCHAR2(19)
 SEXO                                               VARCHAR2(6)
```

FIGURA 4.2 - Estrutura da tabela de Clientes alterada para o projeto de mineração

```
SQL> desc venda
 Nome                                      Nulo?    Tipo
 ----------------------------------------- -------- ----------------
 CLIENTE                                            VARCHAR2(44)
 IDVENDA                                            VARCHAR2(8)
 CATEGPRODUTO                                       VARCHAR2(34)
 SUBCATEGPRODUTO                                    VARCHAR2(42)
 PRODUTO                                            VARCHAR2(44)
 PRECOUNIT                                          VARCHAR2(12)
 QTDE                                               VARCHAR2(9)
 FUNCIONARIO                                        VARCHAR2(14)
 DECISAODECOMPRA                                    VARCHAR2(6)
 OBJETIVODACOMPRA                                   VARCHAR2(60)
 CONDPAGTO                                          VARCHAR2(10)
 HORA                                               VARCHAR2(6)
 TEMPO                                              VARCHAR2(11)
 TEMPERATURA                                        VARCHAR2(13)
 DIA                                                VARCHAR2(6)
 MES                                                VARCHAR2(6)
 DIADASEMANA                                        VARCHAR2(17)
 ANO                                                VARCHAR2(10)
```

FIGURA 4.3 - Estrutura da tabela de Vendas alterada para o projeto de mineração

4.2.2.1 Justificativas para a escolha dos atributos das tabelas

Tabela *Cliente*

Nome do Cliente: para identificação do cliente.
Primeira Vez: identifica se a compra sendo realizada é a primeira que o cliente faz na loja. Esta informação também é útil para identificar o perfil dos clientes que compraram na loja e não retornaram mais.
Profissão: identifica a atividade profissional do cliente.
Estado Civil: estado civil do cliente.
Tempo de Casado: o tempo em que o cliente está casado, se for o caso. Este valor numérico em anos é válido para o último casamento. Existem situações em que a pessoa casou pela segunda vez ou ficou viúva(o). Isto é importante pois recém-casados têm um perfil de compra bastante significativo.
Número de Filhos: número de filhos do cliente. Busca identificar se esta informação determina algum padrão de compra.
Faixa Etária dos Filhos: dividida em *criança, adolescente e adulto.* O critério de classificação para este atributo é baseado no filho mais novo ou nos dois filhos mais novos. O maior objetivo deste campo é identificar se o cliente ainda possui crianças morando junto em sua residência, e se este fato induz à compra de determinados produtos.
Telefone: telefone de contato do cliente. Também usado como um canal de vendas (telemarketing).
Email: email do cliente. O email é cadastrado com o objetivo de manter um canal de comunicação com o cliente que possibilite realizar uma campanha de marketing de mala-direta com custos bem mais reduzidos.
Endereço: endereço residencial do cliente, incluindo complementos. Também habilita o envio de correspondências e folders com promoções e propagandas de produtos.

Casa/Apto - Tipo de Moradia: indica se o cliente mora em casa ou apartamento. Serve para diferenciar o perfil de compra do cliente em função do tipo de moradia.
Bairro: bairro onde o cliente mora.
Cidade: cidade onde o cliente mora.
UF: unidade federativa da cidade do cliente.
Tempo de Residência: o tempo de residência pode determinar um perfil de compra interessante, uma vez que uma pessoa morando numa casa a muito tempo, provavelmente fará algum tipo de reforma, por exemplo.
Residência Própria/Alugada: para poder verificar se as pessoas que moravam em residência alugada possuíam perfil de compra similar às pessoas com residência própria.
Idade: a idade do cliente também pode contribuir de forma importante para criação de um perfil de compra.
Peso: a informação de peso foi incluída pelo fato de a loja comercializar móveis como camas e cadeiras, e foi levantada pelos proprietários como sendo uma característica importante a ser investigada.
Altura: semelhante ao peso, buscava verificar se a altura do cliente implicava na compra de produtos, como móveis, banheiras ou armários para banheiro.
Sexo: o sexo já é um fator de compra muito popular e não poderia ser deixado de fora.

Tabela *Venda*

Nome do Cliente: no registro de vendas, o nome do cliente foi usado como relacionamento com a tabela de cadastro de clientes. No sistema original, os relacionamentos entre tabelas eram feitos por códigos. Como não interessam códigos para a mineração de dados, esses campos foram eliminados.
ID da Venda: identificador único para cada uma das transações de venda. Importante para ser usado pelos algoritmos de mineração de dados que buscam associar vendas de produtos.
Categoria do Produto: categoria à qual o produto comercializado pertence.
Subcategoria do Produto: subcategoria à qual o produto comercializado pertence.
Produto: o produto que está sendo vendido.
Preço Unitário: preço unitário do produto. Multiplicado pela quantidade vendida usando-se sempre a mesma unidade de medida, retorna o valor total da compra do produto.
Quantidade: quantidade vendida do produto em sua unidade de medida.
Funcionário que Atendeu: a informação do funcionário que atendeu o cliente não existia no sistema da empresa. Não existiam esquemas de bonificação por vendas. Foi incluído no projeto de mineração para tentar identificar afinidades entre os clientes e um funcionário específico. É natural que clientes, principalmente os mais antigos, tenham preferência em ser atendidos por algum funcionário em particular.
Decisão de Compra: na transação de venda, esta informação procura identificar se a decisão de compra foi tomada sozinha ou por influência de algum familiar ou acompanhante.
Objetivo da Compra: procura identificar o tipo de uso para o qual a compra se destina.
Condição de Pagamento: foi dividida em *compras à vista* e *compras à prazo.* As compras a vista são aquelas pagas em dinheiro, cartão ou cheque, no momento da compra, enquanto que as outras envolvem desde compras em prestação até financiamentos.
Hora: a hora em que a venda ocorreu.

Tempo: este atributo buscava identificar se existiam tendências de compra em função da presença de sol, chuva ou tempo nublado.
Temperatura: similar ao tempo, este atributo registrava a temperatura média do dia. A amplitude térmica nunca era muito acentuada.
Dia: dia do mês em que a venda foi realizada.
Mês: mês em que a venda foi realizada.
Dia da Semana: dia da semana em que a venda foi realizada.
Ano: ano em que a venda foi realizada.

4.2.3 Selecionando e Verificando a qualidade dos dados

O processo de verificação da qualidade de dados, neste estudo de caso, foi feito através de validações por comandos SQL. Os dados foram extraídos do sistema, convertidos para arquivos-texto, importados em tabelas da base de dados Oracle na plataforma de mineração, e então manipulados via SQL. As validações procuravam identificar :

1) valores distintos para cada um dos campos
2) conflitos de chaves-primárias
3) registro de vendas de clientes que não tinham cadastro
4) valores nulos
5) agrupamentos de valores que pudessem demonstrar anormalidades na base de dados. Ex.: clientes com mais de 100 anos ou compras com valor superior a R$ 50.000,00.
6) ambigüidades, como produtos que pertenciam a mais de uma categoria ou subcategoria.
7) clientes que residiam em mais de uma cidade ou endereço, por erro de registro ou por falha de integridade.

4.2.4 Limpeza dos dados

Como tarefa resultante do passo anterior, foi necessário executar a limpeza dos dados incompletos ou transformação para outro valor válido, como no tratamento de exceções que poderiam induzir a cálculos e resultados errôneos no processo de mineração. De forma geral, um dos pontos ou passos do projeto que mais teve repetição de tarefas, foi o de preparação dos dados. Neste estudo de caso, todas as tarefas que envolvem as etapas de conhecimento e preparação de dados foram revistas e refeitas inúmeras vezes, chegando ao ponto de, no mesmo dia de trabalho, realizarem-se duas ou mais vezes todas as tarefas destas etapas, devido a correções de valores em registros, exclusão de atributos ou erros na montagem dos conjuntos de dados.

Podemos citar, como exemplo, a conversão de valores das profissões. Alguns clientes estavam cadastrados como *Encanador* e outros como *Instalador Hidráulico*, quando na verdade tinham a mesma função. Porém, o atributo que mais foi motivo de estudo e análise foi o *Produto*. Percebeu-se que existiam produtos vendidos em conjunto e com um único nome. A loja vende um *Móvel para Banheiro* como sendo um único produto, mas na verdade é composto por um Balcão, uma Pia, Aéreo e Torneira. Sentiu-se a necessidade de uma "engenharia de produto", de forma a registrar todas as peças que compõe um produto no momento da compra. Assim, quando um cliente estiver comprando um *quarto completo,* saberemos que ele estava atrás de um armário,

uma cama, uma penteadeira, e assim por diante. É necessário usar o bom senso e conhecer bem os objetivos na hora de fazer este tipo de divisão. Por exemplo, quando um cliente comprar um rack para televisão devem ser incluídos os parafusos usados na montagem do rack ? A princípio não. Mas um parafuso é um produto lucrativo e objeto de análise por parte de algum trabalho de mineração, por exemplo ? Se sim, pode ser necessário destacar o número de parafusos que compõe o produto, senão estará se perdendo informação.

Em conseqüência, começou a se questionar como estavam classificados os produtos. Descobriu-se que existia um grande número de produtos com erros de classificação, que ocorria por falta de um padrão explícito e de conhecimento de todos os funcionários no momento de cadastrar um produto ou até mesmo por engano. Podemos citar, por exemplo, um cano usado para fiação elétrica, classificado como um tubo hidráulico. Ou ainda, uma situação mais complexa, em que um produto recebia, internamente, o mesmo nome técnico, como os joelhos elétricos e os joelhos hidráulicos, que eram chamados de *joelho* mas exigiam uma diferenciação na descrição. O que pode-se observar na bibliografia hoje existente de mineração de dados e de metodologias para extração de conhecimento, é que muito se fala no tempo que é dispensado nas fases iniciais preparando os dados, limpando-os e convertendo-os, e que é realmente uma verdade, mas é dado pouca ênfase na importância de uma correta classificação e identificação de cada produto. Isto gera problemas graves para a extração de conhecimento! Quando se obtiver um resultado com uma confiança de 100% será possível acreditar neste resultado?

Neste estudo de caso, após a identificação destes problemas, revisou-se os produtos, suas classificações, e, baseado nos objetivos, foram feitas algumas correções nos dados para que os resultados pudessem ser confiáveis. Esta revisão cautelosa nas descrições dos produtos, seus códigos e suas classificações, deve ser uma etapa de destaque na execução de qualquer projeto de mineração de dados. Uma conseqüência deste trabalho de revisão é o aumento da credibilidade em relatórios gerenciais como balancetes contábeis e relatórios de faturamento extraídos de data warehouses, uma vez que, se os produtos não estiverem classificados de forma correta, estas informações gerenciais estarão também distorcidas.

Ao final, de posse dos conjuntos de dados de clientes e vendas mais íntegros e com valores considerados válidos para a mineração, ocorreu a integração ou cruzamento dos dados.

4.2.5 Integração dos dados

Um dos conjuntos de dados a serem trabalhados era o produto do cruzamento entre informações de clientes e informações de vendas. Nada mais do que um produto cartesiano entre as duas tabelas. Na base Oracle, as duas tabelas, de vendas e de clientes, já estavam criadas, e o processo de geração da nova tabela foi bastante simples. Essa nova tabela, com todos os dados de clientes cruzados com todos os dados de vendas, passou a chamar-se *Produto_Venda_Cliente.* O passo seguinte seria, então, a geração destas informações contidas na nova tabela gerada para um arquivo-texto, fonte para o processamento da informação pela ferramenta de mineração de dados. A figura

4.4 descreve a estrutura da tabela *Produto_Venda_Cliente,* que juntamente com as tabelas *Cliente* e *Venda*, exibidas anteriormente, foram a principal fonte de dados usada neste projeto de mineração, e que foi resultante de todas as fases executadas do início do projeto até o ponto atual.

```
SQL> desc produto_venda_cliente
 Nome                                      Nulo?    Tipo
 ----------------------------------------- -------- ----------------
 CLIENTE                                            VARCHAR2(44)
 PRIMEIRAVEZ                                        VARCHAR2(7)
 PROFISSAO                                          VARCHAR2(44)
 ESTADOCIVIL                                        VARCHAR2(14)
 TEMPODECASADO                                      VARCHAR2(6)
 NODEFILHOS                                         VARCHAR2(6)
 FAIXAETARIADOSFILHOS                               VARCHAR2(15)
 TELEFONE                                           VARCHAR2(14)
 EMAIL                                              VARCHAR2(34)
 ENDERECO                                           VARCHAR2(44)
 CASAAPTO                                           VARCHAR2(10)
 BAIRRO                                             VARCHAR2(34)
 CIDADE                                             VARCHAR2(24)
 UF                                                 VARCHAR2(6)
 MORAQUANTOTEMPO                                    VARCHAR2(6)
 RESIDENCIAPROPRIAALUGADA                           VARCHAR2(11)
 IDADE                                              VARCHAR2(10)
 PESO                                               VARCHAR2(15)
 ALTURA                                             VARCHAR2(19)
 SEXO                                               VARCHAR2(6)
 IDVENDA                                            VARCHAR2(8)
 CATEGPRODUTO                                       VARCHAR2(34)
 SUBCATEGPRODUTO                                    VARCHAR2(42)
 PRODUTO                                            VARCHAR2(44)
 PRECOUNIT                                          VARCHAR2(12)
 QTDE                                               VARCHAR2(9)
 FUNCIONARIO                                        VARCHAR2(14)
 DECISAODECOMPRA                                    VARCHAR2(6)
 OBJETIVODACOMPRA                                   VARCHAR2(60)
 CONDPAGTO                                          VARCHAR2(10)
 HORA                                               VARCHAR2(6)
 TEMPO                                              VARCHAR2(11)
 TEMPERATURA                                        VARCHAR2(13)
 DIA                                                VARCHAR2(6)
 MES                                                VARCHAR2(6)
 DIADASEMANA                                        VARCHAR2(17)
 ANO                                                VARCHAR2(10)
```

FIGURA 4.4 - Estrutura da tabela de *Produto_Venda_Cliente*

4.2.6 Estatísticas sobre o conjunto de dados

Alguns dados estatísticos sobre o conjunto de dados já podem ser descritos nesta etapa do projeto, e são exibidos na tabela 4.1.

TABELA 4.1 - Dados estatísticos sobre o conjunto de dados

Estatísticas da fonte de dados							
Total de Registros de Vendas	40108,00						
Total de Clientes Cadastrados	1370,00						
Total de Transações de Venda	6422,00						
Média de Compras por Cliente	6,24						
Nome do Campo	**Tipo**	**Tamanho no Banco de Dados**	**Valores Distintos**	**Média**	**Desvio Padrão**	**Menor Valor**	**Maior Valor**
CLIENTE	Texto	44	1370	-	-	-	-
PRIMEIRAVEZ	Texto	7	2	-	-	-	-
PROFISSAO	Texto	44	84	-	-	-	-
ESTADOCIVIL	Texto	14	4	-	-	-	-
TEMPODECASADO	Numérico	6	44	13,19	13,44	0	59
NODEFILHOS	Numérico	6	9	1,44	1,22	0	12
FAIXAETARIADOSFILHOS	Texto	15	3	-	-	-	-
TELEFONE	Texto	14	1115	-	-	-	-
EMAIL	Texto	34	98	-	-	-	-
ENDERECO	Texto	44	1270	-	-	-	-
CASAAPTO	Texto	10	2	-	-	-	-
BAIRRO	Texto	34	40	-	-	-	-
CIDADE	Texto	24	10	-	-	-	-
UF	Texto	6	2	-	-	-	-
MORAQUANTOTEMPO	Numérico	6	39	10,97	11,19	0	44
RESIDENCIAPROPRIAALUGADA	Texto	11	2	-	-	-	-
IDADE	Numérico	10	55	41,88	12,19	15	82
PESO	Texto	15	3	-	-	-	-
ALTURA	Texto	19	3	-	-	-	-
SEXO	Texto	6	2	-	-	-	-
IDVENDA	Numérico	8	6422	-	-	1	6422
CATEGPRODUTO	Texto	34	8	-	-	-	-
SUBCATEGPRODUTO	Texto	42	306	-	-	-	-
PRODUTO	Texto	44	1735	-	-	-	-
PRECOUNIT	Numérico	12	539	25,2	79,57	0,02	1140
QTDE	Numérico	9	97	23,33	273,3	0,3	9000
FUNCIONARIO	Texto	14	5	-	-	-	-
DECISAODECOMPRA	Texto	6	2	-	-	-	-
OBJETIVODACOMPRA	Texto	60	14	-	-	-	-
CONDPAGTO	Texto	10	2	-	-	-	-
HORA	Texto	6	13	-	-	7	20
TEMPO	Texto	11	3	-	-	-	-
TEMPERATURA	Texto	13	5	-	-	-	-
DIA	Texto	6	31	-	-	1	31
MES	Texto	6	8	-	-	5	12
DIADASEMANA	Texto	17	6	-	-	-	-
ANO	Texto	10	1	-	-	2002	2002

5 Modelagem, avaliação e aplicação de resultados

As etapas de modelagem, avaliação e aplicação de resultados serão descritas uma vez para cada objetivo estabelecido, já vez que cada um deles exige um conjunto de dados em particular, uma função de mineração e um modelo distinto.

5.1 Conhecer o perfil do cliente que compra na loja

Para conhecer o perfil do cliente que compra na loja, usaremos a fonte de dados com informações sobre os clientes. Neste estudo de caso, trata-se da tabela *Cliente,* encontrada na base de dados Oracle do projeto de mineração. O Intelligent Miner oferece duas formas de leitura de uma fonte de dados: conexão direta com o banco de dados com execução de consultas SQL, ou leitura de arquivos TXT. Geramos, portanto, um arquivo TXT contendo todas as informações dos clientes. O arquivo TXT foi gerado por comandos do utilitário Oracle SQL-PLUS, e fica perfeitamente formatado facilitando a delimitação de campos dentro do Intelligent Miner, que é o passo seguinte. Os comandos utilizados para tal encontram-se no anexo 1 deste documento.

O problema de se utilizar arquivos TXT como fonte de dados para o Intelligent Miner é delimitar os campos. Esta é uma tarefa extremamente trabalhosa e exige muita atenção, pois a chance de se cometer algum erro é grande. O Intelligent Miner oferece uma estrutura básica para fazer isto, e deixa bastante a desejar neste ponto. Qualquer mudança em algum dos atributos lidos implica em ter que refazer as delimitações de campos tornando-se uma tarefa muito cansativa, principalmente se as tabelas têm muitos campos.

5.1.1 Função de mineração escolhida para geração do modelo

Foi usada a função de mineração *Demographic Clustering*, do Intelligent Miner, que agrupa em clusters clientes com perfil parecido, atendendo ao objetivo proposto. Através do uso de clusters, os clientes podem ser agrupados de acordo com suas características, e os campos mais influentes determinam a formação do grupo.
Este tipo de tarefa deve ser exaustivamente analisado, uma vez que existem várias formas de se aplicar a função de mineração. Os resultados que podem ser obtidos são muitos. Qualquer alteração em um parâmetro de mineração gera um resultado diferente. Quando se fizer uso da função de cluster demográfico, pode-se definir pesos para as variáveis. Dessa forma, baseado em um conhecimento prévio, o analista pode induzir a formação do resultado. Podemos dizer que é possível refinar e aprimorar o conhecimento descoberto, mas fica muito fácil conduzir a função de mineração à geração de resultados errados.

Ao contrário do algoritmo de cluster demográfico, o algoritmo de cluster neural não permite que se atribuam pesos aos campos analisados. Porém, a simples redução ou aumento do limite de número de clusters, que é um parâmetro modificável, faz com que

os resultados sejam facilmente alteráveis. Um número grande de clusters às vezes cria uma série de pequenos conjuntos de dados compostos de menos de 5 registros, ou até de apenas 1 registro. Deve-se perceber e analisar cada um dos casos pois isto pode indicar a presença de exceções nos dados, e que podem induzir ao aumento da margem de erro.

5.1.2 Campos da tabela *Cliente* para a montagem dos clusters

Para a mineração, selecionamos os seguintes campos da tabela *Cliente:*

TABELA 5.1 - Campos da tabela *Cliente* selecionados
na montagem dos clusters demográficos

Tabela *Cliente*		
Nome do Campo	**Tipo**	**Tamanho**
Altura	Texto	19
Bairro	Texto	34
CasaApto	Texto	10
Cidade	Texto	24
EstadoCivil	Texto	14
FaixaEtariadosFilhos	Texto	15
Idade	Numérico	2
MoraQuantoTempo	Numérico	2
NodeFilhos	Numérico	2
Peso	Texto	15
PrimeiraVez	Texto	7
Profissão	Texto	44
ResidenciaPropriaAlugada	Texto	11
Sexo	Texto	1
TempoDeCasado	Numérico	2

5.1.3 Parâmetros usados no processo de mineração

1) Maximum passes: a especificação de um valor alto para o número de passagens de processamento pelos dados aumenta a qualidade dos clusters gerados. Entretanto, um valor mais baixo para este parâmetro representa um tempo de processamento menor. Um número de passagens entre 5 e 10 é geralmente suficiente. O padrão do Intelligent Miner é 2. O valor usado foi 10.

2) ***Maximum clusters***: pode-se limitar também o número máximo de clusters que se deseja criar. Devem ser levadas em conta, antes de limitar o número de clusters, questões relativas à exatidão dos modelos e à geração de clusters muito pequenos. O padrão do Intelligent Miner é 9. O que acontece, na prática, é a necessidade de se verificar os resultados com vários valores para este parâmetro, ou seja, executar a mineração com o número máximo de clusters 3, depois 5, 7, 10 , e assim por diante. Os resultados variam bastante. Não existe uma regra que diga que um número máximo de clusters é o ideal. Neste objetivo, o número de clusters usado foi 5, pois foi entendido como sendo o de melhor resultado gerado.

3) ***Accuracy Improvement***: este parâmetro faz a função de mineração trabalhar da seguinte forma: se o valor especificado para ele for 10, o processo de iteração

termina quando o aumento da qualidade entre duas passagens for menor que 10%. Esta percentagem de aumento é medida a cada passagem sobre os dados. Este parâmetro é, portanto, um critério de parada da função de mineração. Quanto menor for este valor, mais correto é o cluster. O valor *default* do Intelligent Miner para este atributo é 2. O valor usado foi 1.

4) *Similarity Threshold*: este parâmetro limita o enquadramento de um registro dentro de um cluster. Por exemplo, se definimos que o limiar de similaridade é 0,5 , registros com 50% de campos com valores idênticos são prováveis candidatos a fazer parte do cluster. Se quisermos obter um número maior de clusters, devemos aumentar o valor deste parâmetro. O valor *default* do Intelligent Miner é 0,5 sendo este o valor usado.

5) *Outlier treatment*: para os dois tipos de campos numéricos suportados pelo Intelligent Miner (numérico e numérico discreto) a função de cluster reconhece os chamados *outliers*, que nada mais são do que valores que se encontram fora do intervalo "normal". Este intervalo de valores sempre possui um valor mínimo e um valor máximo. Valores menores que um valor mínimo ou maiores que um valor máximo são chamados de *outliers.* Existem três formas de tratá-los:

1- assumindo *outliers* como valores perdidos
2- substituindo o *outlier* pelo valor mínimo ou pelo valor máximo
3- tratando os *outliers* como valores válidos

No primeiro caso, o *outlier* é remapeado para a unidade de entrada correspondente. No segundo caso, qualquer valor menor que o valor mínimo é substituído pelo valor mínimo, e qualquer valor maior que o valor máximo é substituído pelo valor máximo. Já no último caso, um *outlier* é tratado como se pertencesse ao intervalo normal de valores. O padrão do Intelligent Miner é assumir *outliers* como valores perdidos. Neste estudo de caso, estamos tratando *outliers* como valores válidos.

Como foi destacado anteriormente, é muito fácil alterar os resultados através da mudança do valor de um atributo de mineração. O usuário pode e deve experimentar mudanças nos parâmetros de mineração e analisar cada um dos resultados gerados. A ferramenta Intelligent Miner está longe de ser uma ferramenta fácil de usar por pessoas com conhecimento básico em microinformática. Isto seria interessante, uma vez que somente um usuário que vive o dia-a-dia do negócio da empresa consegue separar com precisão resultados importantes de resultados óbvios ou curiosos mas sem importância.
O que acontece, na prática, é a geração de relatórios e telas que são entregues ao usuários para que sejam analisados. O tempo e o custo deste processo são maiores.

6) O peso atribuído a cada campo para o processo de montagem dos clusters influencia e muito o resultado final. Esta é, particularmente, uma vantagem do uso de funções de *clustering* demográfico em relação ao uso de funções neurais, pois com elas não é possível atribuir pesos aos campos. Neste estudo de caso, o conhecimento do negócio era bastante consolidado e aos atributos que mais influenciavam foi atribuído um peso maior. A tabela 5.2 relaciona os atributos e seus respectivos pesos.

TABELA 5.2 - Pesos dos atributos da tabela *Cliente* na montagem dos clusters demográficos

Tabela *Cliente*	
Nome do Campo	**Peso**
Altura	1
Bairro	4
CasaApto	3
Cidade	2
EstadoCivil	5
FaixaEtariadosFilhos	2
Idade	8
MoraQuantoTempo	8
NodeFilhos	4
Peso	1
PrimeiraVez	5
Profissão	7
ResidenciaPropriaAlugada	9
Sexo	5
TempoDeCasado	5

5.1.4 Clusters gerados

Os cinco clusters gerados são descritos a seguir. Como citado anteriormente, foram verificados resultados com vários valores para este parâmetro, ou seja, com o número máximo de clusters 3, depois 5, 7, 10 , e assim por diante. Não existe uma regra que diga que um número máximo de clusters é o ideal. Neste objetivo, o número de clusters usado foi 5, pois foi entendido como sendo o de melhor resultado gerado.

5.1.4.1 Cluster 0

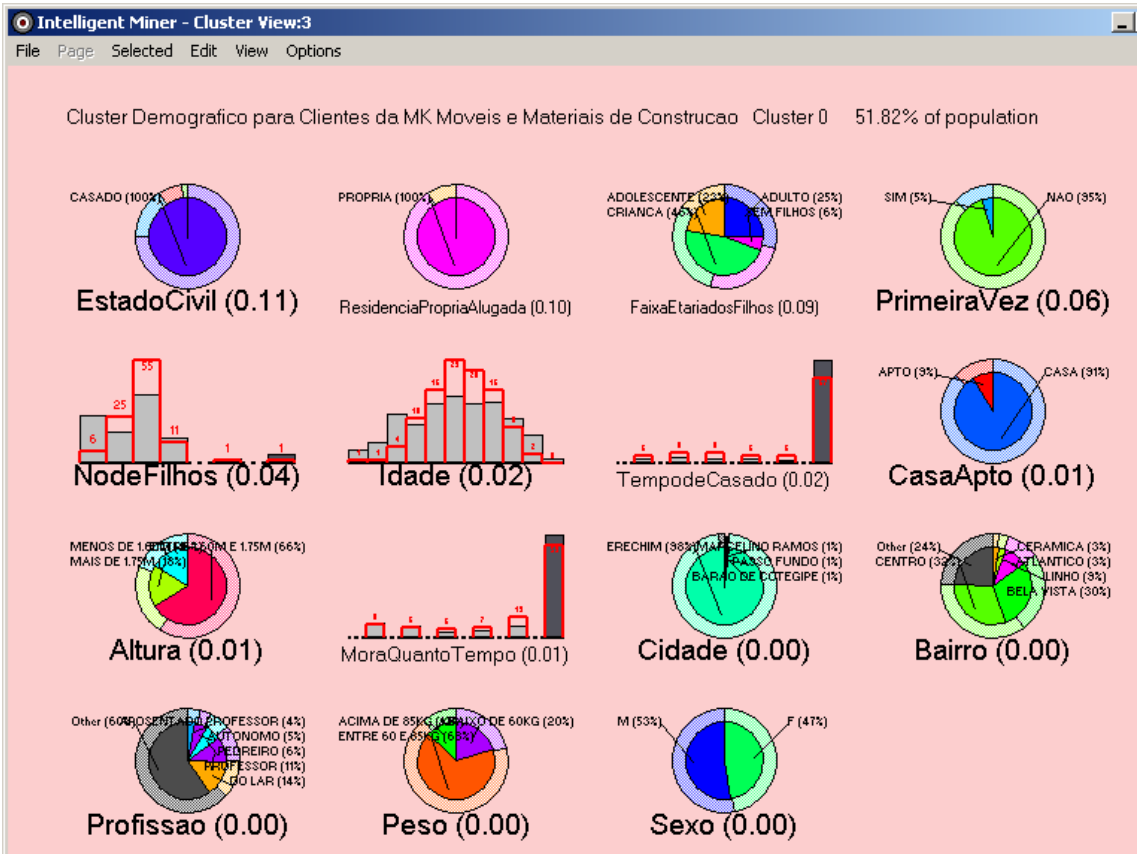

FIGURA 5.1 - Cluster Demográfico de Clientes – 51,82% da população

Este grupo, o maior deles, representado por 51,82% da população, indica o perfil de um consumidor casado, com residência própria, predominantemente em casas contra uma pequena parcela que reside em apartamentos, que já fez mais de uma compra na loja, dividindo-se quase que igualmente entre homens e mulheres.

5.1.4.2 Cluster 1

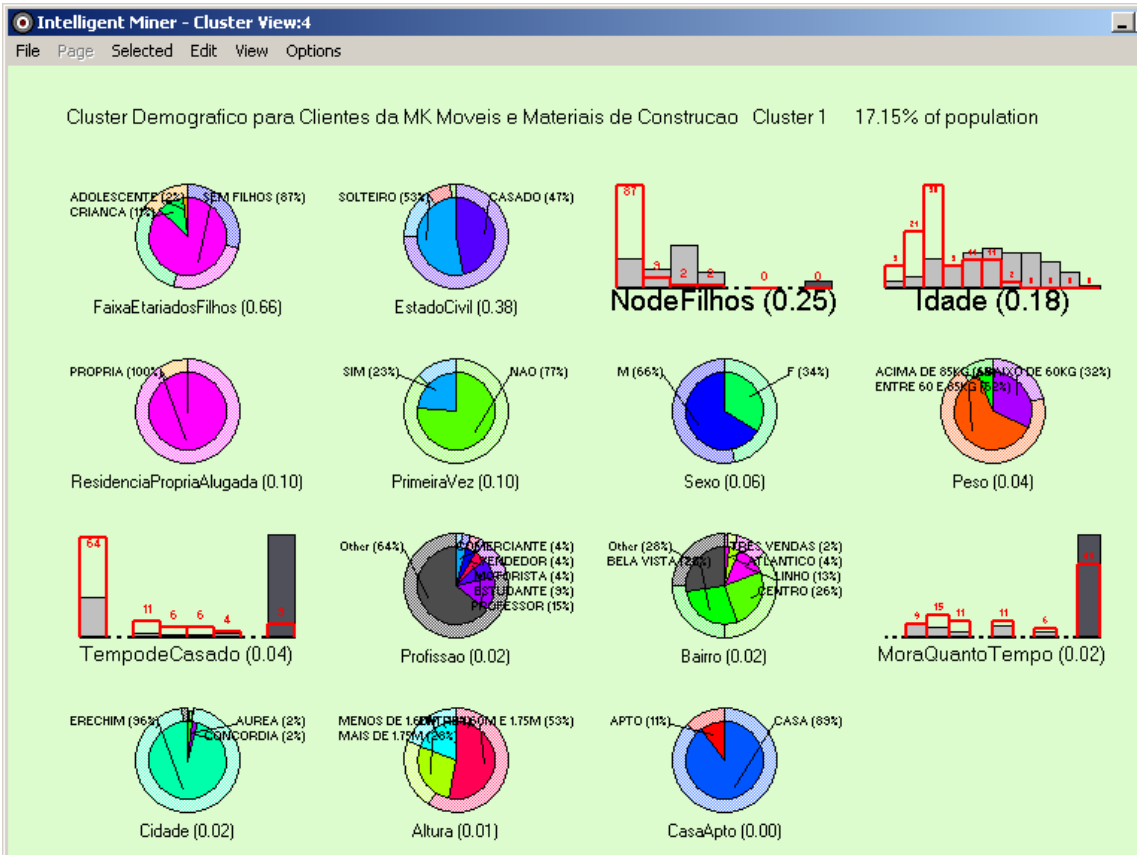

FIGURA 5.2 - Cluster Demográfico de Clientes – 17,15% da população

Este segundo grupo, com 17,15% da população, é caracterizado por clientes sem filhos, solteiros, que residem em casas próprias, sendo a maioria cliente antigo. Outro ponto importante a destacar neste cluster é que os 47% com estado civil *casado* estão, na sua grande maioria, casados a menos de 1 ano.

5.1.4.3 Cluster 3

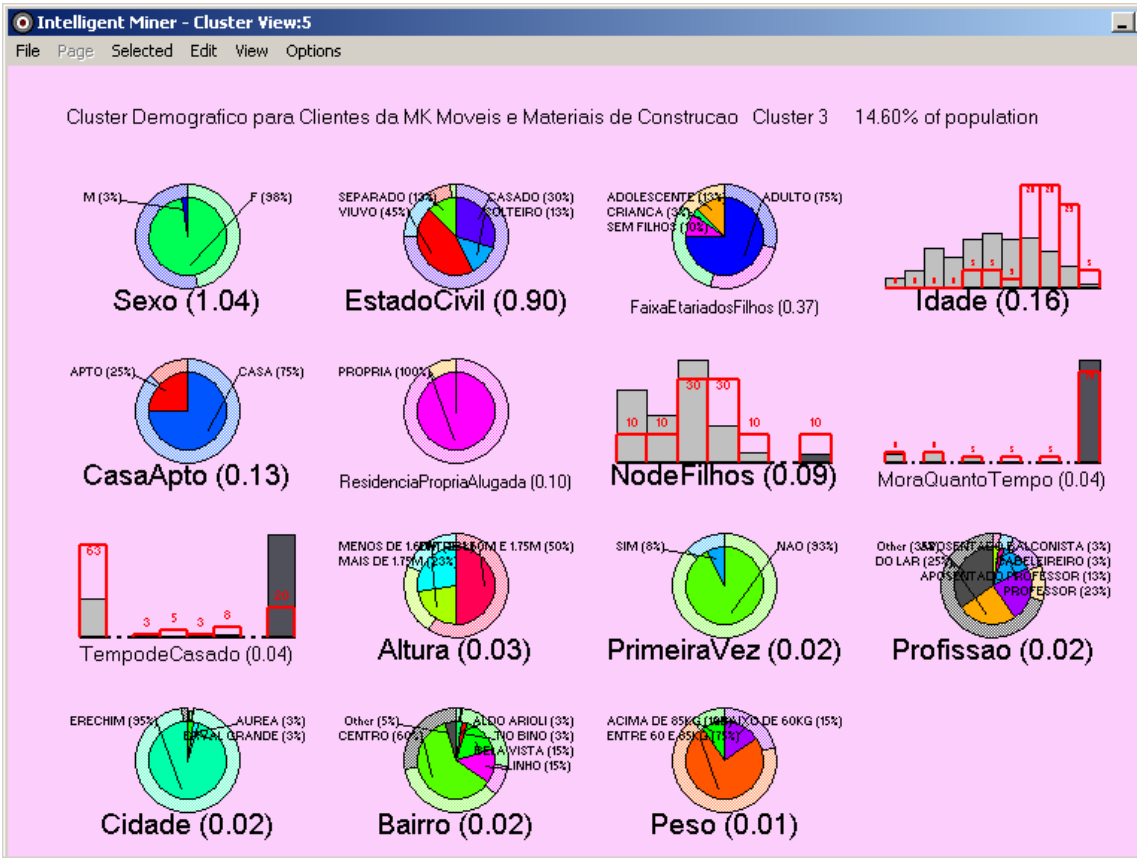

FIGURA 5.3 - Cluster Demográfico de Clientes – 14,60% da população

O terceiro cluster, com 14,60% da população, é formado por clientes do sexo feminino, com filhos adultos, com idade dos 50 aos 65 anos em sua maioria, residindo a maior parte em casas mas com um número relativamente grande morando em apartamentos (25%), todos com residência própria, e que também já compraram mais de uma vez na loja. Um ponto a destacar, neste cluster, é que o bairro, para 60% dos casos, é “CENTRO”.

5.1.4.4 Cluster 4

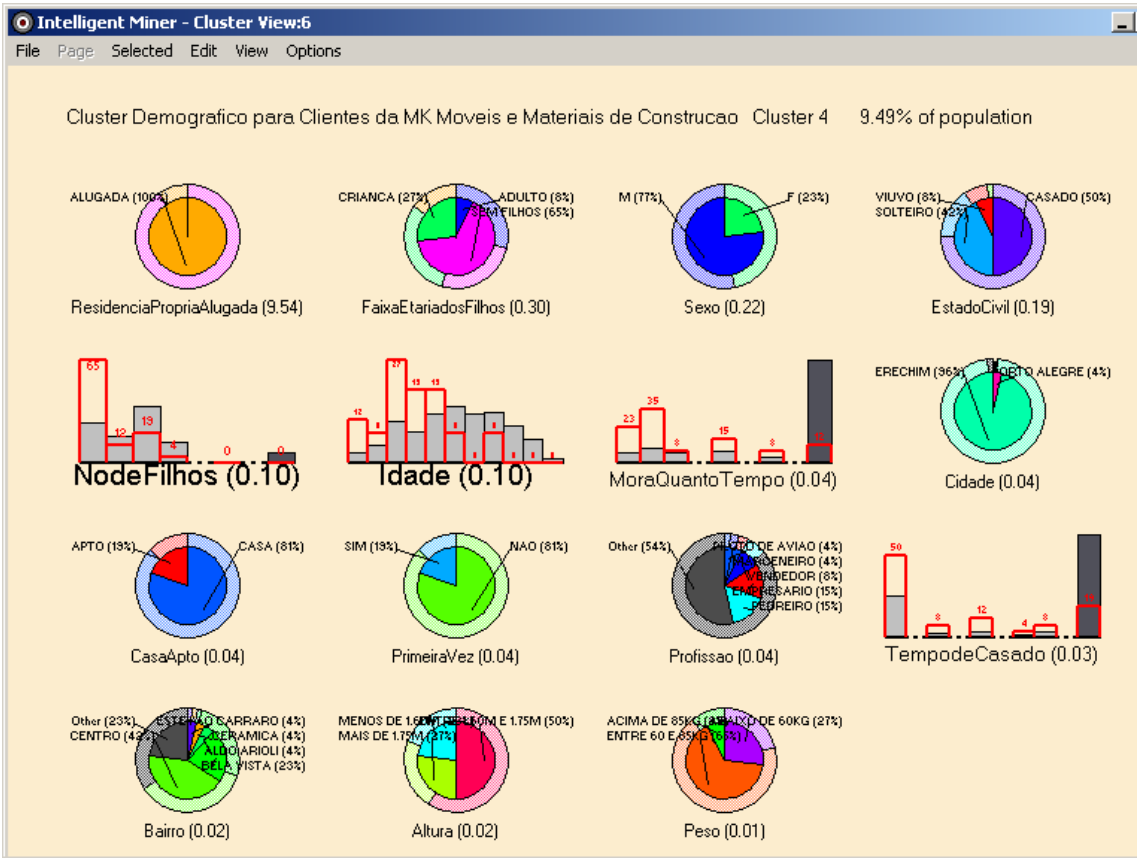

FIGURA 5.4 - Cluster Demográfico de Clientes – 9,49% da população

O quarto cluster, representado por 9,49% dos casos, é formado pelas pessoas que moram em residências alugadas, sem filhos ou com filhos pequenos, a maioria do sexo masculino, morando predominantemente em casas. Dividem-se principalmente entre solteiros e casados, sendo grande parte deles casada a menos de 1 ano.

5.1.4.5 Cluster 2

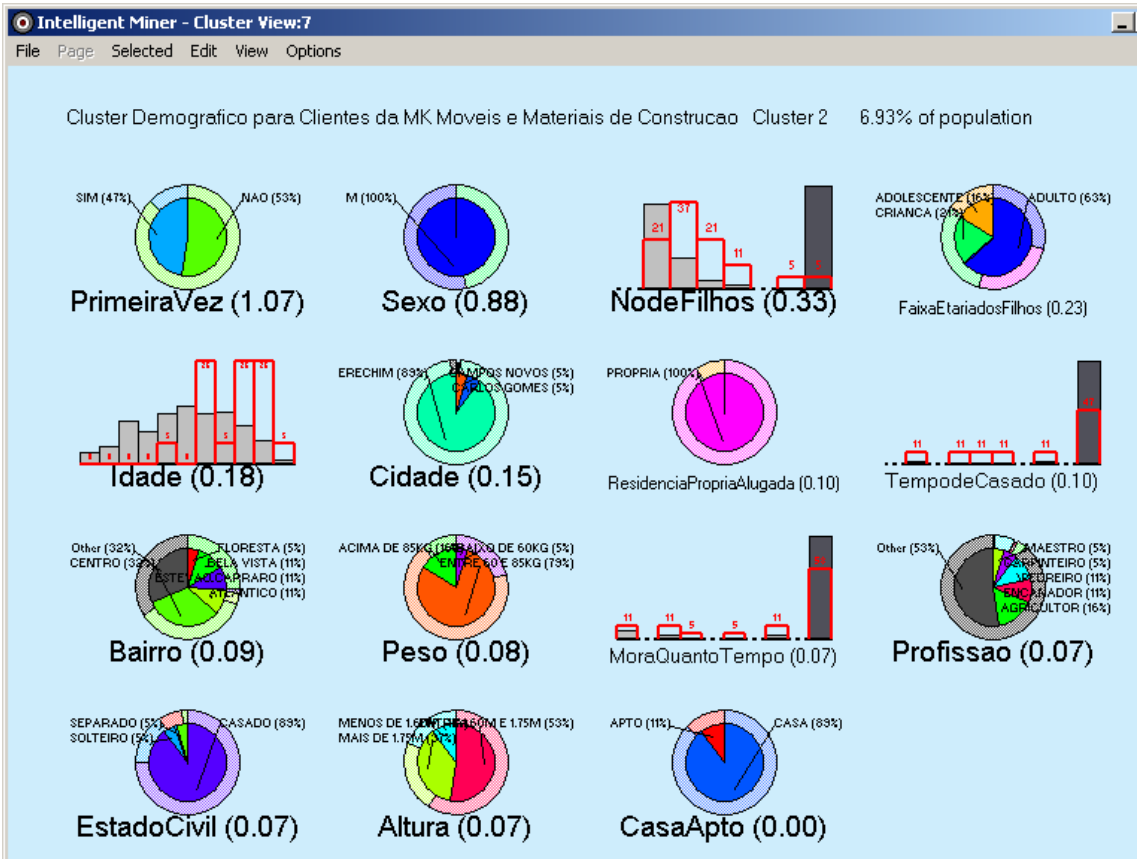

FIGURA 5.5 - Cluster Demográfico de Clientes – 6,93% da população

Este último cluster, com apenas 6,93% da população, mostra um perfil de cliente que vem comprar pela primeira vez na loja. Ainda que a maioria já tenha comprado outras vezes, este grupo apresenta um número relativamente grande para clientes que fazem sua primeira compra na loja (47%). Também são homens adultos e que moram em casas próprias.

O consumidor da loja foi dividido em 5 grupos, resumindo as características da seguinte forma:

TABELA 5.3 - Perfil do consumidor que realiza compras na loja

Perfil do Cliente		
Cluster	**População**	**Descrição**
0	51,82%	Clientes casados, que residem em casas próprias, são fidelizados, homens e mulheres.
1	17,15%	Clientes sem filhos, solteiros, fidelizados, e que moram em casas próprias.
3	14,60%	Mulheres, com filhos adultos, morando em residência própria, grande parte em apartamentos do centro da cidade.
4	9,49%	Clientes que moram em residências alugadas, sem filhos ou com filhos pequenos, em sua maioria homens.
2	6,93%	Clientes que compram pela primeira vez na loja, sendo homens adultos em sua maioria, e que moram em casas próprias.

Para os proprietários, esta tabela demonstra um perfil de cliente um pouco diferente do conhecido até então, e que deve ser considerado a partir de agora na tomada de decisões estratégicas. Dos clusters gerados, o único que demonstrou um perfil previsto foi o cluster 0, o maior deles, enquanto que os outros tratam-se de informações completamente novas.

5.2 Conhecer o perfil do cliente associado com as compras que o mesmo faz na loja.

Para conhecer o perfil do cliente que compra na loja associando com suas compras, usaremos as fontes de dados com informações sobre os clientes e sobre as vendas. Neste estudo de caso, trata-se da tabela *Produto_Venda_Cliente,* encontrada na base de dados Oracle do projeto de mineração. Para esta tarefa, também geramos um arquivo TXT contendo todas as informações dos clientes e suas compras. O arquivo TXT também foi gerado por comandos do utilitário Oracle SQL-PLUS. Os comandos utilizados para isso encontram-se no anexo 1 deste documento.

5.2.1 Função de mineração escolhida para geração do modelo

Neste caso, também foi usada a função de mineração *Demographic Clustering*, do Intelligent Miner, para atender ao objetivo proposto.

5.2.2 Campos da tabela *Produto_Venda_Cliente* usados

Selecionamos os seguintes campos da tabela *Produto_Venda_Cliente:*

TABELA 5.4 - Campos da tabela *Produto_Venda_Cliente* para clusters demográficos

Tabela *Produto_Venda_Cliente*		
Nome do Campo	**Tipo**	**Tamanho**
CasaApto	Texto	10
EstadoCivil	Texto	14
FaixaEtariadosFilhos	Texto	15
Idade	Numérico	2
MoraQuantoTempo	Numérico	2
ObjetivoDaCompra	Texto	60
PrimeiraVez	Texto	7
ResidenciaPropriaAlugada	Texto	11
Sexo	Texto	1
TempoDeCasado	Numérico	2
CategProduto	Texto	34

5.2.3 Parâmetros usados no processo de mineração

1) Maximum passes: o valor usado foi 10.
2) Maximum clusters: o número de clusters usado foi 8.
3) Accuracy Improvement: o valor usado foi 1.

4) Similarity Threshold: o valor usado foi 0,5.
5) Outlier treatment: valores válidos.
6) Peso atribuído a cada campo:

TABELA 5.5 - Pesos dos atributos da tabela *Produto_Venda_Cliente*

Tabela *Produto_Venda_Cliente*	
Nome do Campo	**Peso**
CasaApto	5
EstadoCivil	6
FaixaEtariadosFilhos	1
Idade	4
MoraQuantoTempo	10
ObjetivoDaCompra	15
PrimeiraVez	3
ResidenciaPropriaAlugada	7
Sexo	3
TempoDeCasado	8
CategProduto	20

5.2.4 Clusters gerados

Dos 8 clusters gerados, os dois maiores e mais significativos serão comentados e mostrados a seguir.

5.2.4.1 Cluster 0

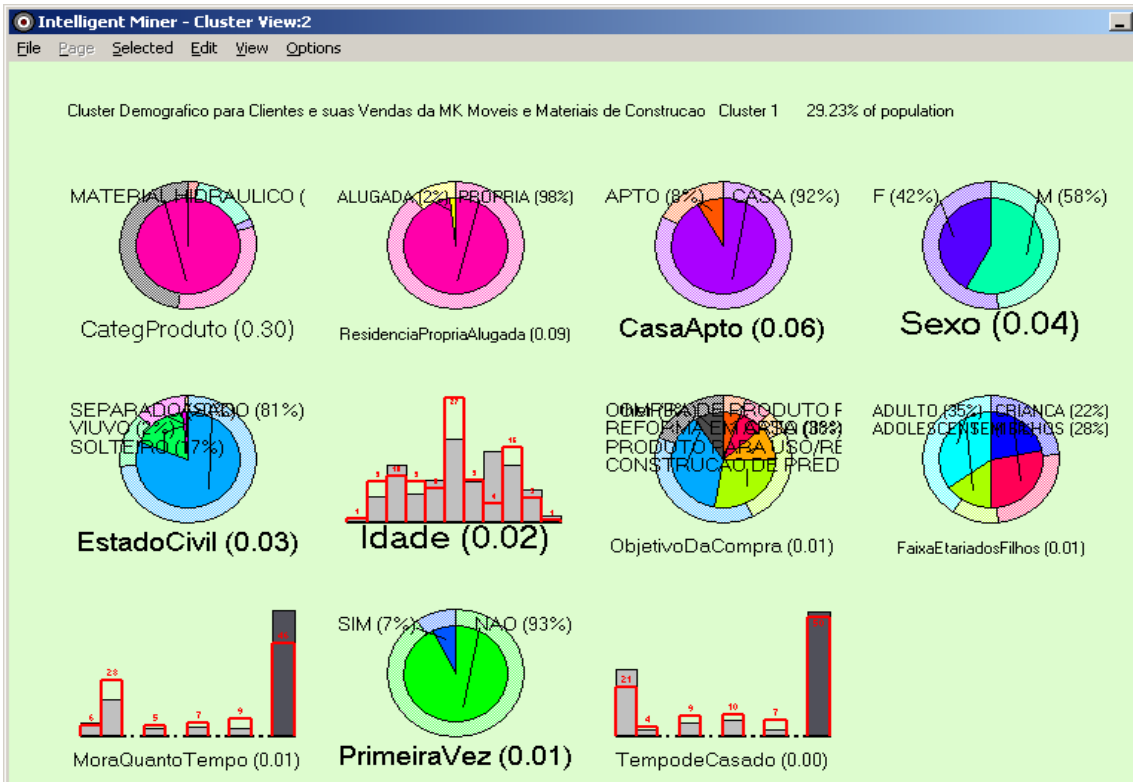

FIGURA 5.6 - Cluster Demográfico de Clientes por categoria de produto comprado

Este primeiro cluster, o maior deles com 29,23% da população, agrupa clientes que compram material hidráulico, residem em casas próprias e são casados. O mais interessante é analisar que o objetivo da compra que predomina é “reforma em casa”, com 38% dos casos. Se associarmos esta informação com o tempo de residência, veremos que existe uma tendência a realizar reformas quando o tempo de residência é de 1 a 2 anos. Podemos ver os valores abaixo.

Field Details:

Field Name: MoraQuantoTempo

Label	Size Cluster %	Size Reference %
0	05.8366	07.0129
1	27.6265	18.4610
2	02.9831	03.3738
3	00.5188	10.0076
4	01.5564	01.7817
5	02.9831	03.5633
6	00.7782	00.7202
7	00.3891	01.0614

Os detalhes em vermelho indicam que 27,6265% dos clientes do cluster se enquadram no grupo de pessoas com tempo de residência entre 1 e 2 anos, e que este intervalo corresponde a 18,461% se considerarmos todos os dados de todos os clusters. O cluster 3, que montou um perfil para clientes que compram materiais de construção, foi similar ao cluster 1 neste ponto. O gráfico da figura 5.7 demonstra a porcentagem de pessoas que realizam reformas e que adquirem materias de construção em função do tempo de residência, predominantemente em casas próprias.

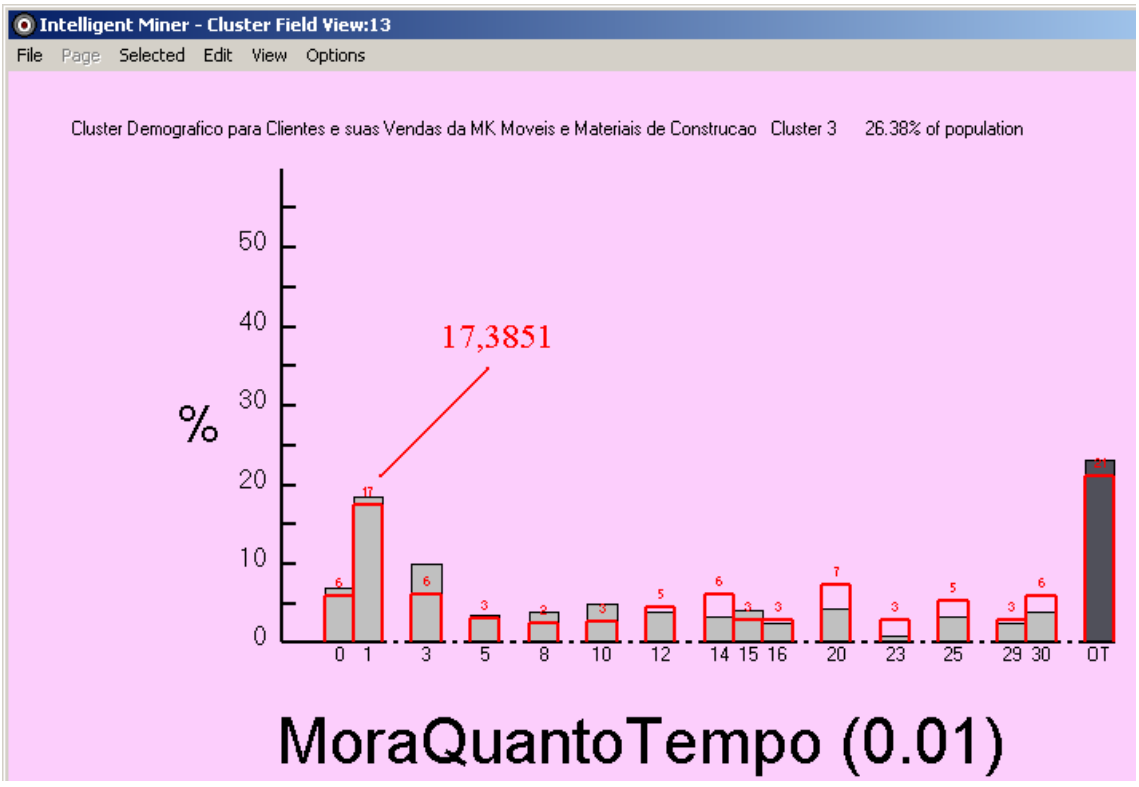

FIGURA 5.7 - Tempo de Residência para a população do Cluster 3

Ainda dentro do objetivo de se descobrir o perfil do cliente associado com as compras que o mesmo faz, buscamos identificar respostas através de regras associadas com as informações de vendas. O Intelligent Miner não gera regras escritas do tipo *SE Atributo1 = x ENTÃO Atributo2=y com confiança = x.x%* . Pode-se conseguir este tipo de regra através das árvores de decisão geradas pelos algoritmos de classificação do Intelligent Miner, mas os resultados obtidos não foram muito satisfatórios, assim como a forma de visualização gráfica oferecida por ele, que é muito ruim de interpretar. Portanto, buscamos completar nosso objetivo pelo uso do software de mineração de dados WEKA - Waikato Environment for Knowledge Analysis, versão 3.2.3, desenvolvido na universidade de Waikato, na Nova Zelândia.

5.2.5 Geração de regras para o perfil do cliente e as compras que o mesmo realiza através do WEKA 3.2.3

A fonte de dados usada foi a mesma, com modificações feitas somente para atender a exigências do WEKA. O software é muito restritivo quanto à formatação do arquivo de entrada. Por exemplo:

1. Os dados, no arquivo de origem, devem estar separados por vírgulas ou por tabulações.
2. Não pode haver valores em branco ou nulos que gerem uma seqüência de duas vírgulas seguidas nos dados (,,).
3. Não pode haver um caracter " antes de uma vírgula delimitadora.
4. Cada valor que existir como um dado deve ser destacado no cabeçalho de atributos.
5. O valor de qualquer campo não pode estar separado por um espaço em branco. Deve-se atualizar cada espaço em branco substituindo por, por exemplo, “_”.
6. Não pode haver espaços em branco ao final de cada linha de dados pois o WEKA interpreta o caracter em branco como sendo parte do valor da última coluna.
7. Não pode haver pontos ‘.’ no valor de um atributo.
8. Não pode haver parênteses no valor de um atributo.

Caso alguma das regras acima for violada, na abertura do arquivo ocorrerá o erro mostrado na figura 5.8.

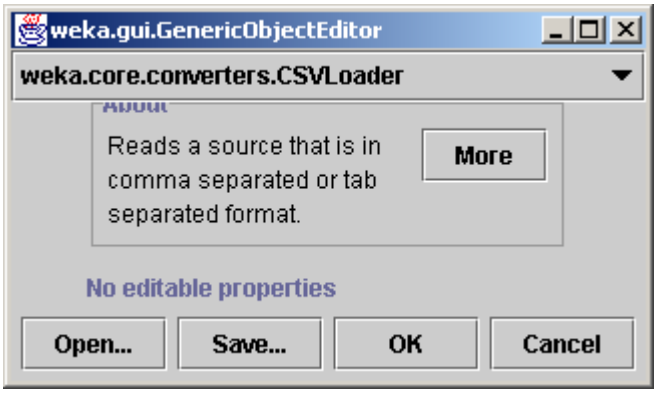

FIGURA 5.8 - Aviso de erro na leitura de arquivo-texto fora do padrão

Um dos arquivo gerados para a mineração pode ser visto abaixo:

```
@relation RegrasClientesVendasMk
@attribute PRIMEIRAVEZ {SIM,NAO}
@attribute PROFISSAO {ADMINISTRADOR_DE_EMPRESAS,AGENTE_DE_OBRAS,...,VENDEDOR}
@attribute ESTADOCIVIL {CASADO,SEPARADO,SOLTEIRO,VIUVO}
@attribute TEMPODECASADO {0,1,10,11,12,13,14,15,...,38,39,4,40,42,44,45,5,59,6,7,8,9}
@attribute NODEFILHOS {0,1,12,2,3,4,5,8,9}
@attribute CASAAPTO {APTO,CASA}
@attribute BAIRRO {AEROPORTO,ALDO_ARIOLI,...,TIO_BINO,TRES_VENDAS,TRIANGULO,VILA_FELIZ}
@attribute MORAQUANTOTEMPO {0,1,10,11,12,13,14,15,16,17,18,...,40,42,44,5,6,7,8,9}
@attribute RESIDENCIAPROPRIAALUGADA {ALUGADA,PROPRIA}
@attribute SEXO {M,F}
@attribute CATEGPRODUTO {FERRAMENTAS,...,MOVEIS,UTILITARIO_RESIDENCIAL}


@attribute DIADASEMANA {QUARTA-FEIRA,...,SEGUNDA-FEIRA,SEXTA-FEIRA,TERCA-FEIRA}
@data
NAO,DO_LAR,CASADO,30,4,CASA,BELA_VISTA,23,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SABADO
NAO,DO_LAR,CASADO,30,4,CASA,BELA_VISTA,23,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SABADO
NAO,COMERCIANTE,SOLTEIRO,0,0,CASA,CENTRO,10,PROPRIA,M,MATERIAL_HIDRAULICO,SABADO
NAO,COMERCIANTE,SOLTEIRO,0,0,CASA,CENTRO,10,PROPRIA,M,MATERIAL_HIDRAULICO,SABADO
NAO,ENCANADOR,CASADO,21,2,CASA,BELA_VISTA,20,PROPRIA,M,MATERIAL_DE_DECORACAO,TERCA-FEIRA
NAO,DO_LAR,CASADO,15,1,CASA,LINHO,15,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SEXTA-FEIRA
NAO,SECRETARIO,CASADO,6,1,CASA,CENTRO,7,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SEXTA-FEIRA
NAO,SECRETARIO,CASADO,6,1,CASA,CENTRO,7,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SEXTA-FEIRA
NAO,ELETRICISTA,SOLTEIRO,0,0,CASA,NAZARE,2,PROPRIA,M,MATERIAL_HIDRAULICO,QUINTA-FEIRA
NAO,ELETRICISTA,SOLTEIRO,0,0,CASA,NAZARE,2,PROPRIA,M,MATERIAL_ELETRICO,SEXTA-FEIRA
NAO,PROFESSOR,VIUVO,0,3,CASA,CENTRO,24,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,TERCA-FEIRA
NAO,METALURGICO,CASADO,12,2,CASA,LINHO,12,PROPRIA,M,MATERIAL_DE_CONSTRUCAO,QUINTA-FEIRA
NAO,METALURGICO,CASADO,12,2,CASA,LINHO,12,PROPRIA,M,MATERIAL_DE_CONSTRUCAO,QUINTA-FEIRA
NAO,VENDEDOR,CASADO,8,2,CASA,COTREL,8,PROPRIA,M,MATERIAL_ELETRICO,QUARTA-FEIRA
NAO,VENDEDOR,CASADO,8,2,CASA,COTREL,8,PROPRIA,M,MATERIAL_ELETRICO,QUINTA-FEIRA
NAO,DO_LAR,VIUVO,15,5,CASA,CENTRO,30,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SEGUNDA-FEIRA
NAO,DO_LAR,VIUVO,15,5,CASA,CENTRO,30,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SEGUNDA-FEIRA
NAO,DO_LAR,VIUVO,15,5,CASA,CENTRO,30,PROPRIA,F,MATERIAL_DE_CONSTRUCAO,SEXTA-FEIRA
NAO,PEDREIRO,CASADO,31,5,CASA,SAO_VICENTE_DE_PAULA,21,PROPRIA,M,MOVEIS,SEXTA-FEIRA
SIM,MAESTRO,CASADO,35,3,APTO,CENTRO,0,PROPRIA,M,MATERIAL_DE_DECORACAO,QUARTA-FEIRA
SIM,MAESTRO,CASADO,35,3,APTO,CENTRO,0,PROPRIA,M,MATERIAL_ELETRICO,QUARTA-FEIRA
SIM,MAESTRO,CASADO,35,3,APTO,CENTRO,0,PROPRIA,M,MATERIAL_ELETRICO,QUARTA-FEIRA
NAO,SOLDADOR,CASADO,10,1,CASA,BELA_VISTA,24,PROPRIA,M,MATERIAL_ELETRICO,QUARTA-FEIRA
NAO,SOLDADOR,CASADO,10,1,CASA,BELA_VISTA,24,PROPRIA,M,MATERIAL_ELETRICO,QUARTA-FEIRA
NAO,SOLDADOR,CASADO,10,1,CASA,BELA_VISTA,24,PROPRIA,M,MATERIAL_ELETRICO,QUARTA-FEIRA
NAO,PINTOR,SOLTEIRO,0,2,CASA,CAMPO_ACEJA,5,PROPRIA,M,FERRAMENTAS,SEGUNDA-FEIRA
NAO,PINTOR,SOLTEIRO,0,2,CASA,CAMPO_ACEJA,5,PROPRIA,M,FERRAMENTAS,SEGUNDA-FEIRA
NAO,AUTONOMO,CASADO,10,2,CASA,PETIT_VILAGGE,6,PROPRIA,M,FERRAMENTAS,QUINTA-FEIRA
...
```

5.2.5.1 Função de mineração escolhida

A função de Associação do WEKA, que implementa o algoritmo Apriori, foi escolhida neste caso. É importante destacar um diferencial entre o WEKA e o Intelligent Miner no que diz respeito ao funcionamento de suas funções de associação. No WEKA, os filtros são mais aprofundados. É possível determinar, por exemplo, o mínimo e o máximo suporte mínimo , confiança mínima e número máximo de regras. Isto dá uma certa flexibilidade para buscar regras interessantes. Uma desvantagem é que não é possível determinar a geração de regras somente para determinado campo, ou seja, especificar um campo para que as regras sejam montadas em função dele. Por outro lado, um funcionamento bastante interessante do algoritmo Apriori é que ele busca, dependendo do valor de uma variável chamada Delta, gerar sempre regras com confiança alta, e, particularmente, essas são as regras mais interessantes.

5.2.5.2 Parâmetros usados no processo de mineração

Tipo de métrica usado: Confiança. Serve para determinar a confiança do resultado e também como critério de ordenação dos resultados, da maior confiança para a menor.

LowerBoundMinSupport: a menor porcentagem permitida para exibir uma regra considerando todo o conjunto de dados. Em outras palavras, uma regra só é gerada se ela ocorrer em uma proporção maior do que o valor especificado por este parâmetro. Como podem existir regras interessantes, com confiança alta, mas que ocorrem pouco freqüentemente, o valor usado para este parâmetro foi 0,001.
MinMetric: a menor confiança aceita. O menor valor usado entre todas as mineração foi 30%.
UpperBoundMinSupport: a maior confiança aceita para exibição de uma regra. Sempre foi de 100% para este estudo de caso.
Delta: a cada iteração, o algoritmo diminui o suporte de confiança pelo valor especificado em delta. Para mineração mais detalhadas, o valor de delta deve ser pequeno. Porém, o tempo de mineração aumenta. O valor usado foi 0,005.
NumRules: o número máximo de regras por mineração. Especificando um valor alto fazemos com que uma mineração só termine se o suporte mínimo tiver sido atingido, o que é melhor, pois se o processo acabar por ter atingido o número de regras limite, alguma regra importante pode ficar de fora. O valor usado foi 1000.

A tela de parâmetros é exibida na figura 5.9.

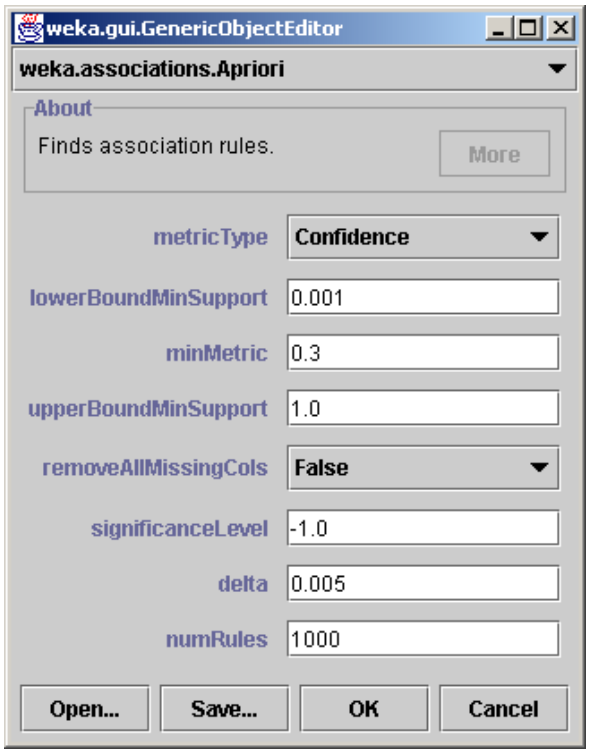

FIGURA 5.9 - Parâmetros para mineração através do algoritmo Apriori do WEKA

5.2.5.3 Regras geradas

De todas as regras geradas em todas variações do conjunto de dados, algumas regras foram destacadas para análise junto aos proprietários da loja, que conhecem bem o negócio da empresa. Uma parte delas é exibida abaixo.

```
  6.CASAAPTO=CASA SEXO=F 1036 ==> RESIDPROPALUG=PROPRIA 1033 conf:(1)
 22.ESTADOCIVIL=CASADO 2330 ==> RESIDPROPALUG=PROPRIA 2243 conf:(0.96)
 39.NODEFILHOS=0 1061 ==> CASAAPTO=CASA 1017 conf:(0.96)
 98.ESTADOCIVIL=CASADO SEXO=M 1390 ==> CASAAPTO=CASA 1279 conf:(0.92)
176.SEXO=F 1369 ==> RESIDPROPALUG=PROPRIA 1164 conf:(0.85)
  5.ESTADOCIVIL=SOLTEIRO BAIRRO=BELA_VISTA RESIDPROPALUG=ALUGADA 282 ==> MORAQTTEMPO=1 282 conf:(1)
 11.ESTADOCIVIL=VIUVO OBJCOMPRA=CONSTR_PREDIO/APTO/CASA_PROPRIA 200 ==> CASAAPTO=APTO 200 conf:(1)
 13.ESTADOCIVIL=VIUVO RESIDPROPALUG=ALUGADA 192 ==> CASAAPTO=APTO 192 conf:(1)
 37.PRIMEIRAVEZ=NAO FAIXETFILHO=SEM_FILHOS SEXO=M 700 ==> CASAAPTO=CASA 697 conf:(1)
 67.BAIRRO=CENTRO SEXO=F 811 ==> PRIMEIRAVEZ=NAO 798 conf:(0.98)
 69.ESTADOCIVIL=CASADO OBJCOMPRA=REFORMA_EM_CASA 840 ==> RESIDPROPALUG=PROPRIA 819 conf:(0.98)
 13.ESTADOCIVIL=CASADO FAIXETFILHO=ADULTO 819 ==> RESIDPROPALUG=PROPRIA 819 conf:(1)
 16.FAIXETFILHO=SEM_FILHOS BAIRRO=BELA_VISTA MORAQTTEMPO=1 282 ==> ESTADOCIVIL=SOLTEIRO 282 conf:(1)
384.ESTADOCIVIL=SOLTEIRO BAIRRO=BELA_VISTA 331 ==> FAIXETFILHO=SEM_FILHOS 327 conf:(0.99)
 79.FAIXETFILHO=ADULTO RESIDPROPALUG=ALUGADA SEXO=F 192 ==> CASAAPTO=APTO 192 conf:(1)
379.OBJCOMPRA=REFORMA_EM_CASA 993 ==> RESIDPROPALUG=PROPRIA 957 conf:(0.96)
 17.SUBCATEGPRODUTO=ROLO_PARA_PINTURA 41 ==> OBJCOMPRA=REFORMA_EM_CASA 22 conf:(0.54)
  9.PRIMEIRAVEZ=SIM SEXO=F 84 ==> FAIXETFILHO=SEM_FILHOS 71 conf:(0.85)
 34.PRIMEIRAVEZ=SIM 343 ==> SEXO=M 259 conf:(0.76)
 35.SUBCATEGPRODUTO=LUVA 96 ==> SEXO=M 72 conf:(0.75)
 41.SUBCATEGPRODUTO=ADAPTADOR 30 ==> SEXO=M 22 conf:(0.73)
 44.FAIXETFILHO=ADULTO SUBCATEGPRODUTO=REJUNTE 28 ==> SEXO=F 20 conf:(0.71)
 46.SUBCATEGPRODUTO=CHUVEIRO 38 ==> SEXO=M 27 conf:(0.71)
 47.SUBCATEGPRODUTO=REDUCAO 27 ==> SEXO=M 19 conf:(0.7)
 52.SEXO=F SUBCATEGPRODUTO=PISO 31 ==> FAIXETFILHO=ADULTO 21
 54.SUBCATEGPRODUTO=TOMADA 70 ==> SEXO=M 47 conf:(0.67)
 56.SUBCATEGPRODUTO=FITA_VEDA_ROSCA 30 ==> SEXO=M 20 conf:(0.67)
 72.FAIXETFILHO=ADULTO 1136 ==> SEXO=F 705 conf:(0.62)
269.SUBCATEGPRODUTO=LUVA 96 ==> SEXO=M 72 conf:(0.75)
344.FAIXETFILHO=CRIANCA 765 ==> SEXO=M 487 conf:(0.64)
386.SUBCATEGPRODUTO=PISO 74 ==> SEXO=M 43 conf:(0.58)
 88.PRIMEIRAVEZ=NAO 3065 ==> SEXO=F 1285 conf:(0.42)
226.SUBCATEGPRODUTO=PISO 74 ==> RESIDPROPALUG=PROPRIA 60 conf:(0.81)
 58.FAIXETFILHO=CRIANCA 765 ==> RESIDPROPALUG=PROPRIA 741 conf:(0.97)
 72.SUBCATEGPRODUTO=TE_HIDRAULICO 101 ==> PRIMEIRAVEZ=NAO 96 conf:(0.95)
 84.SUBCATEGPRODUTO=CAL 36 ==> PRIMEIRAVEZ=NAO 34 conf:(0.94)
  3.BAIRRO=VILA_FELIZ 48 ==> OBJCOMPRA=CONSTR_PREDIO/APTO/CASA_PROPRIA 48 conf:(1)
 18.BAIRRO=ESPERANCA OBJCOMPRA=REFORMA_EM_CASA 20 ==> FUNCIONARIO=EDGAR 20 conf:(1)
 11.BAIRRO=MORRO_DA_CEGONHA 34 ==> FUNCIONARIO=EDGAR 34 conf:(1)
 55.BAIRRO=ESTEVAO_CARRARO 49 ==> FUNCIONARIO=EDGAR 36 conf:(0.73)
 45.BAIRRO=ESTEVAO_CARRARO 49 ==> OBJCOMPRA=REFORMA_EM_APTO 39 conf:(0.8)
126.SUBCATEGPRODUTO=INTERRUPTOR 70 ==> BAIRRO=CENTRO 36 conf:(0.51)
  1.OBJCOMPRA=AMPLIACAO_RESIDENCIAL_-_CASA 53 ==> FUNCIONARIO=EDGAR 50 conf:(0.94)
  2.OBJCOMPRA=COMPRA_DE_PRODUTO_PARA_APTO_NOVO 38 ==> FUNCIONARIO=INES 24 conf:(0.63)
  5.SUBCATEGPRODUTO=ROLO_PARA_PINTURA 41 ==> OBJCOMPRA=REFORMA_EM_CASA 22 conf:(0.54)
  6.SUBCATEGPRODUTO=FLEXIVEL 39 ==> FUNCIONARIO=INES 20 conf:(0.51)
  7.SUBCATEGPRODUTO=REGISTRO 40 ==> FUNCIONARIO=INES 20 conf:(0.5)
 10.SUBCATEGPRODUTO=PISO 74 ==> FUNCIONARIO=INES 35 conf:(0.47)
  9.SUBCATEGPRODUTO=TINTA_PARA_PINTURA 80 ==> OBJCOMPRA=REFORMA_EM_CASA 38 conf:(0.48)
 11.SUBCATEGPRODUTO=TIJOLO 47 ==> OBJCOMPRA=REFORMA_EM_CASA 22 conf:(0.47)
  1.SUBCATEGPRODUTO=FECHADURA 28 ==> PRIMEIRAVEZ=NAO 28 conf:(1)
  2.SUBCATEGPRODUTO=REDUCAO 27 ==> PRIMEIRAVEZ=NAO 27 conf:(1)
  3.SUBCATEGPRODUTO=CAIXA_DAGUA_PARA_VASO_SANITARIO 17 ==> PRIMEIRAVEZ=NAO 17 conf:(1)
  4.SUBCATEGPRODUTO=LIXA 56 ==> PRIMEIRAVEZ=NAO 55 conf:(0.98)
  5.SUBCATEGPRODUTO=REGISTRO 40 ==> PRIMEIRAVEZ=NAO 39 conf:(0.98)
  6.SUBCATEGPRODUTO=MANGUEIRA_PARA_AGUA 32 ==> PRIMEIRAVEZ=NAO 31 conf:(0.97)
  7.SUBCATEGPRODUTO=LAMPADA_INCANDESCENTE 29 ==> PRIMEIRAVEZ=NAO 28 conf:(0.97)
  8.SUBCATEGPRODUTO=TE_HIDRAULICO 101 ==> PRIMEIRAVEZ=NAO 96 conf:(0.95)
  9.SUBCATEGPRODUTO=TAMPA_CEGA 20 ==> PRIMEIRAVEZ=NAO 19 conf:(0.95)
 10.SUBCATEGPRODUTO=RALO 20 ==> PRIMEIRAVEZ=NAO 19 conf:(0.95)
 11.SUBCATEGPRODUTO=JOELHO 238 ==> PRIMEIRAVEZ=NAO 226 conf:(0.95)
 12.SUBCATEGPRODUTO=JUNCAO 19 ==> PRIMEIRAVEZ=NAO 18 conf:(0.95)
 13.SUBCATEGPRODUTO=PREGO 56 ==> PRIMEIRAVEZ=NAO 53 conf:(0.95)

 14.SUBCATEGPRODUTO=CAL 36 ==> PRIMEIRAVEZ=NAO 34 conf:(0.94)
 15.SUBCATEGPRODUTO=SOLVENTE 18 ==> PRIMEIRAVEZ=NAO 17 conf:(0.94)
 16.SUBCATEGPRODUTO=LUVA 96 ==> PRIMEIRAVEZ=NAO 90 conf:(0.94)
```

```
 17.SUBCATEGPRODUTO=CANO_HIDRAULICO 175 ==> PRIMEIRAVEZ=NAO 164 conf:(0.94)
 18.SUBCATEGPRODUTO=ADAPTADOR 30 ==> PRIMEIRAVEZ=NAO 28 conf:(0.93)
 19.SUBCATEGPRODUTO=AREIA 29 ==> PRIMEIRAVEZ=NAO 27 conf:(0.93)
 20.SUBCATEGPRODUTO=VALVULA 41 ==> PRIMEIRAVEZ=NAO 38 conf:(0.93)
 21.SUBCATEGPRODUTO=ASSENTO_SANITARIO 23 ==> PRIMEIRAVEZ=NAO 21 conf:(0.91)
 22.SUBCATEGPRODUTO=COLA/ADESIVO 75 ==> PRIMEIRAVEZ=NAO 68 conf:(0.91)
 23.SUBCATEGPRODUTO=PISO 74 ==> PRIMEIRAVEZ=NAO 67 conf:(0.91)
 75.PRIMEIRAVEZ=NAO CATEGPRODUTO=MATERIAL_HIDRAULICO FUNCIONARIO=EDGAR 299 ==> SEXO=M 231 conf:(0.77)
 77.FUNCIONARIO=EDGAR 943 ==> SEXO=M 695 conf:(0.74)
 82.CATEGPRODUTO=FERRAMENTAS 95 ==> SEXO=M 67 conf:(0.71)
 24.CATEGPRODUTO=MOVEIS FUNCIONARIO=GUILHERME 42 ==> SEXO=F 25 conf:(0.6)
 25.SEXO=F CATEGPRODUTO=UTILITARIO_RESIDENCIAL 148 ==> FUNCIONARIO=INES 88 conf:(0.59)
 31.CATEGPRODUTO=MOVEIS 219 ==> SEXO=F 117 conf:(0.53)
 32.FUNCIONARIO=ROLFI 108 ==> SEXO=F 56 conf:(0.52)
 36.SEXO=F CATEGPRODUTO=MOVEIS 117 ==> FUNCIONARIO=INES 58 conf:(0.5)
 52.CATEGPRODUTO=MATERIAL_DE_CONSTRUCAO 1045 ==> SEXO=F 440 conf:(0.42)
 58.CATEGPRODUTO=FERRAMENTAS 95 ==> FUNCIONARIO=GUILHERME 38 conf:(0.4)
 60.FUNCIONARIO=LUCINEIA 336 ==> SEXO=F 130 conf:(0.39)
 89.DIADASEMANA=QUINTA-FEIRA 554 ==> SEXO=M 384 conf:(0.69)
 93.DIADASEMANA=SABADO 139 ==> SEXO=M 92 conf:(0.66)
 94.CATEGPRODUTO=MATERIAL_DE_CONSTRUCAO DIADASEMANA=QUINTA-FEIRA 168 ==> SEXO=M 111 conf:(0.66)
 97.PROFISSAO=CONSTRUTOR 86 ==> SEXO=M DIADASEMANA=SEGUNDA-FEIRA 55 conf:(0.64)
 98.PROFISSAO=CONSTRUTOR SEXO=M 86 ==> DIADASEMANA=SEGUNDA-FEIRA 55 conf:(0.64)
 99.PROFISSAO=CONSTRUTOR 86 ==> DIADASEMANA=SEGUNDA-FEIRA 55 conf:(0.64)
207.SEXO=M DIADASEMANA=SEXTA-FEIRA 347 ==> CATEGPRODUTO=MATERIAL_DE_CONSTRUCAO 130 conf:(0.37)
208.SEXO=F DIADASEMANA=SEGUNDA-FEIRA 209 ==> CATEGPRODUTO=MATERIAL_DE_CONSTRUCAO 78 conf:(0.37)
 27.PROFISSAO=CONSTRUTOR 86 ==> DIADASEMANA=SEGUNDA-FEIRA 55 conf:(0.64)
117.DIADASEMANA=QUINTA-FEIRA 554 ==> CATEGPRODUTO=MATERIAL_HIDRAULICO 165 conf:(0.3)
 53.BAIRRO=VILA_FELIZ 48 ==> CATEGPRODUTO=MATERIAL_HIDRAULICO 37 conf:(0.77)
 54.CASAAPTO=APTO CATEGPRODUTO=MATERIAL_DE_CONSTRUCAO 121 ==> BAIRRO=CENTRO 93 conf:(0.77)
 55.CASAAPTO=APTO CATEGPRODUTO=MATERIAL_HIDRAULICO 146 ==> BAIRRO=CENTRO 112 conf:(0.77)
 56.CASAAPTO=APTO 462 ==> BAIRRO=CENTRO 350 conf:(0.76)
 57.BAIRRO=CENTRO CATEGPRODUTO=MOVEIS 92 ==> CASAAPTO=CASA 69 conf:(0.75)
 73.BAIRRO=SAO_CRISTOVAO 76 ==> RESIDPROPALUG=PROPRIA 76 conf:(1)
 74.BAIRRO=SAO_CRISTOVAO 76 ==> CASAAPTO=CASA 76 conf:(1)
406.CASAAPTO=CASA RESIDPROPALUG=ALUGADA 434 ==> BAIRRO=BELA_VISTA 293 conf:(0.68)
```

Podemos tirar algumas conclusões sobre este pequeno exemplo de regras que conduzem a estudos bastante detalhados, de onde podem resultar desde pequenas campanhas de marketing ou promoções de produtos, até uma mudança de processo dentro da empresa. Após uma análise junto aos proprietários, destacaram-se as seguintes regras como mais importantes e que serão objeto de estudo:

76% dos clientes que compram na loja e moram em apartamentos, moram no centro da cidade. Um *outdoor* no centro da cidade, então, deveria somente exibir produtos comprados por pessoas que moram em apartamentos, por exemplo.

68% dos que moram em casas alugadas moram no bairro Bela Vista. Também podemos promover somente produtos comprados para casas alugadas em um folder e distribuir no bairro em questão.

Clientes com filhos adultos, morando em residência alugada e do sexo feminino, todos moram em apartamentos e suas compras têm predominantemente como objetivo a construção de casa/apto próprio. Se for possível obter uma lista de endereços de pessoas com este perfil, pode-se oferecer produtos destinados à construção de uma casa ou de um apartamento próprio.

96% dos clientes sem filhos moram em casas. Vai de acordo com o perfil dos clientes de um modo geral, que moram, em sua maioria, em casas. Conforme mostrado pelos clusters do objetivo anterior, provavelmente estas pessoas são jovens casais e que estarão fazendo alguma reforma entre o primeiro e o segundo ano de moradia na nova casa.

92% dos homens casados moram em casas. Similar ao anterior, merece um tratamento parecido.

Todos os clientes do bairro São Cristóvão moram em casas. Folders com produtos para casa poderiam ser distribuídos em lojas e demais estabelecimentos comerciais do bairro.

A venda de materiais de construção ocorre em sua maior parte nas segundas-feiras e nas sextas-feiras, com 37% das vendas do dia, sendo que na segunda-feira as mulheres realizam mais compras que os homens e na sexta-feira ocorre o contrário. Esta estatística demostra ser necessário um controle de estoque mais intenso para não possibilitar a falta de material de construção entre a sexta-feira e a segunda-feira.

30% das vendas feitas nas quintas-feiras são de material hidráulico. Uma outra informação que exige atenção no controle de estoque.

No bairro Vila Feliz, 77% das vendas são de material hidráulico, com objetivo de construção de residência própria, casa ou apartamento. Se 77% das vendas para este bairro são somente de material hidráulico e para construção de casas próprias, onde os clientes compram os outros materiais necessários para se construir uma casa ou um prédio? Existe alguma concorrência no bairro? É necessário um levantamento sobre isto.

47% dos pisos são vendidos pela funcionária Inês. Isto significa que os clientes que compram pisos preferem ser atendidos pela Inês ou é ela quem consegue vender uma quantidade maior de pisos que os outros 3 funcionários? Talvez seja necessário uma troca de experiência no tratamento com o cliente. Cada funcionário tem sua característica e todos precisam compartilhar o segredo que conduz o cliente à compra.

50% dos móveis e 59% de utilitários residenciais são vendidos pela funcionária Inês para clientes do sexo feminino. Esta informação merece o mesmo tratamento que a anterior.

54% das vendas de rolo de pintura e 48% das vendas de tinta para pintura são feitas para reformas em casas. Isto mostra que a venda destes materiais não esta exclusivamente associada à obras como casas e prédios novos, mas sim que existe uma preocupação dos proprietários de casas com a pintura das mesmas maior do que o imaginado.

Somente 6% dos clientes que compram canos hidráulicos estão comprando pela primeira vez na loja. Não estaria faltando uma divulgação maior do tipo de produto comercializado pela loja? Talvez o público em geral ainda associe mais a loja com o seu passado recente, de comercialização somente de móveis. Isto poderia ser resolvido com uma propaganda em rádio ou televisão para o fortalecimento da imagem da loja, associando com todos estes materiais comprados somente por clientes antigos. Apenas para citar outros exemplos, 100% dos clientes que compram caixa d'água para vaso sanitário são clientes antigos assim como 95% dos clientes que compram joelhos hidráulicos.

58% dos pisos vendidos são vendidos para mulheres. Piso é um produto bastante rentável e merece uma boa atenção. Se a maioria dos clientes que compra é do sexo feminino, vale a pena veicular uma propaganda em televisão num horário em que mais mulheres estejam vendo.

74% das vendas do funcionário Edgar são feitas para homens. Pode existir algum tipo de confiança maior de homem para homem e mulher para mulher, pois as mulheres, principalmente acima de 45 anos, preferem comprar com a funcionária Inês, e, da mesma forma, o funcionário Guilherme realiza mais vendas de material elétrico. É necessário um estudo para saber se não seria ideal especializar cada um dos funcionários em cada categoria de produto ou generalizar e compartilhar o conhecimento de cada um. Por quê um funcionário consegue vender mais determinado produto do que outro? Quantas vendas deixaram de ser feitas devido ao cliente não ter sentido segurança ao questionar o funcionário sobre o produto? São questões que precisam ser esclarecidas.

Muitas regras geradas não trazem informação alguma por serem inúteis ou redundantes, por se tratarem de exceções, ou mesmo por terem uma confiança muito baixa, sendo então descartadas. Alguns exemplos dentre as dezenas de regras inúteis geradas:

```
  3. ESTCIVIL=SOLTEIRO 728 ==> TEMPODECASADO=0 728    conf:(1)
  4. ESTCIVIL=SOLTEIRO CASAAPTO=CASA 706 ==> TEMPODECASADO=0 706    conf:(1)
496. ESTCIVIL=VIUVO CASAAPTO=APTO MORAQTOTEMPO=3 186 ==> DIADASEMANA=QUARTA-FEIRA 72 conf:(0.39)
412. MORAQUANTOTEMPO=5 185 ==> BAIRRO=LINHO 77    conf:(0.42)
500. NODEFILHOS=4 76 ==> BAIRRO=MORRO_DA_CEGONHA MORAQUANTOTEMPO=3 34    conf:(0.45)
497. TEMPODECASADO=26 70 ==> PROFISSAO=MECANICO 32    conf:(0.46)
479. TEMPODECASADO=40 84 ==> MORAQUANTOTEMPO=22 42    conf:(0.5)
  5. ESTCIVIL=SOLTEIRO NODEFILHOS=0 651 ==> TEMPODECASADO=0 651    conf:(1)
495. ESTCIVIL=VIUVO MORAQTOTEMPO=3 186 ==> CASAAPTO=APTO DIADASEMANA=QUARTA-FEIRA 72 conf:(0.39)
480. TEMPODECASADO=40 84 ==> PROFISSAO=COMERCIARIO 42    conf:(0.5)
280. BAIRRO=JOSE_BONIFACIO 20 ==> MORAQUANTOTEMPO=12 19    conf:(0.95)
281. BAIRRO=JOSE_BONIFACIO 20 ==> TEMPODECASADO=17 19    conf:(0.95)
289. PROFISSAO=AUTONOMO TEMPODECASADO=10 48 ==> BAIRRO=PETIT_VILAGGE 45    conf:(0.94)
 69. NODEFILHOS=4 MORAQTOTEMPO=3 34 ==> TEMPODECASADO=30 BAIRRO=MORRO_DA_CEGONHA 34 conf:(1)
```

5.3 Criar uma lista de produtos mais rentáveis e verificar quais clientes com maior tendência a realizar compras desses produtos

Este objetivo visa definir uma lista dos produtos mais rentáveis para a empresa, reunir informações sobre os clientes que compram esses produtos, e determinar, entre todos os outros clientes, a probabilidade que cada um tem de realizar tais compras. De posse da lista de clientes e da probabilidade que cada um tem de realizar a compra, pode-se fazer uma campanha seja por telemarketing, mala-direta ou outro canal de venda, que ofereça os produtos especialmente para aqueles clientes com maior probabilidade de comprar. Este processo é conhecido como up-selling.

Para eleger um produto, deve-se considerar o quanto ele traz de lucro para a empresa. Em nosso caso, poderíamos escolher ROUPEIRO LUXO, QUARTO DE CASAL, CONJUNTO DE LOUÇAS PARA BANHEIRO, AQUECEDOR, BANHEIRA, dentre outros. Abaixo será descrito o resultado obtido para dois dos produtos acima citados: ROUPEIRO LUXO e CONJUNTO DE LOUÇAS PARA

BANHEIRO. Os produtos, tanto na base de dados original quanto nas tabelas usadas na mineração, pertencem a uma categoria e a uma subcategoria. Neste problema específico, estaremos verificando quais são os clientes que compraram produtos da **subcategoria** *Roupeiro de Luxo* e *Conjunto de Louças para Banheiro*. É necessário verificar a subcategoria do produto uma vez que um cliente pode ter comprado um Roupeiro de Luxo em mogno e outro pode ter comprado em marfim. Ou ainda, um pode ter comprado um roupeiro 6x6 e outro pode ter comprado 4x4. Como os produtos variam, a subcategoria a que eles pertencem é o alvo da mineração.

5.3.1 Função de mineração escolhida para geração do modelo

Vamos utilizar uma função de Cluster Demográfico para agrupar os clientes que compram roupeiros de luxo e conjuntos de louças para banheiro. Também levaremos em conta as outras compras feitas por estes clientes pois o motivo da compra de um roupeiro de luxo pode estar associado à compra de uma cama de casal de luxo, por exemplo.

5.3.2 Campos da tabela *Produto_Venda_Cliente* usados

Selecionamos os seguintes campos da tabela *Produto_Venda_Cliente:*

TABELA 5.6 - Campos selecionados para a verificação de tendências de compra

Tabela Poduto_Venda_Cliente		
Nome do Campo	**Tipo**	**Tamanho**
Bairro	Texto	34
CasaApto	Texto	10
EstadoCivil	Texto	14
FaixaEtariadosFilhos	Texto	15
Funcionario	Texto	14
Idade	Numérico	2
MoraQuantoTempo	Numérico	2
NodeFilhos	Numérico	2
ObjetivoDaCompra	Texto	60
PrimeiraVez	Texto	3
Profissao	Texto	44
ResidenciaPropriaAlugada	Texto	11
Sexo	Texto	1
TempoDeCasado	Numérico	2

5.3.3 Parâmetros usados no processo de mineração

1) Maximum passes: o valor usado foi 10.
2) Maximum clusters: o número de clusters usado foi 5.
3) Accuracy Improvement: o valor usado foi 1.
4) Similarity Threshold: o valor usado foi 0,5.
5) Outlier treatment: valores válidos.
6) Peso atribuído a cada campo:

TABELA 5.7 - Pesos dos atributos da tabela *Produto_Venda_Cliente*

Tabela *Produto_Venda_Cliente*	
Nome do Campo	**Peso**
Bairro	4
CasaApto	3
EstadoCivil	5
Idade	8
MoraQuantoTempo	8
NodeFilhos	4
PrimeiraVez	5
Profissão	7
ResidenciaPropriaAlugada	9
Sexo	5
TempoDeCasado	5
Funcionário	8
ObjetivoDaCompra	5

Para finalizar, precisamos ler da tabela *Produto_Venda_Cliente* somente os registros de venda que representam roupeiros de luxo e, para conjunto de louças para banheiro, somente os registros de venda que os representam. Para isto usamos uma função de filtro de registros do Intelligent Miner. A figura 5.10 demonstra o filtro aplicado na montagem do modelo.

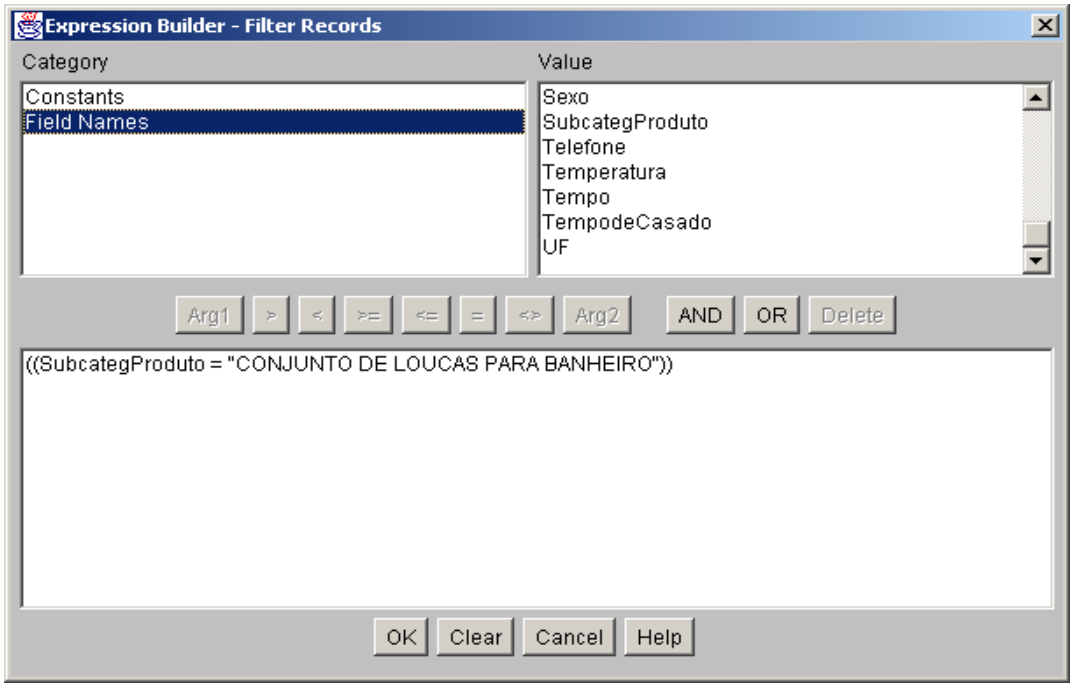

FIGURA 5.10 - Filtro de registros do Intelligent Miner

5.3.4 Resultados gerados

A figura 5.11 mostra os clusters de clientes que compraram roupeiros de luxo.

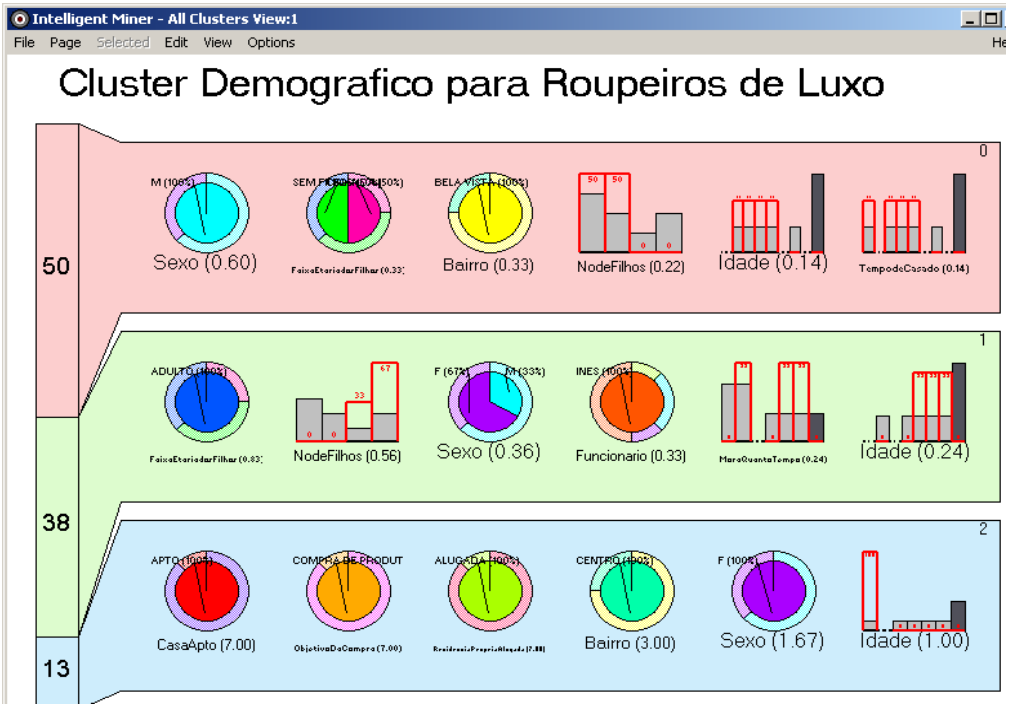

FIGURA 5.11 - Clientes que compraram Roupeiros de Luxo

Para este tipo de objetivo, não interessa diretamente saber a composição de cada um dos clusters. Quando a função de mineração é executada em modo de aplicação, ela varre a fonte de dados e determina a probabilidade que o cliente possui de pertencer a cada um dos clusters. As duas maiores são selecionadas, sendo que, neste caso, usamos apenas a maior delas. Na geração, determinamos uma saída em arquivo-texto para estas informações. Podemos então ordenar os dados gerados em ordem crescente de probabilidade, considerando a confiança, e obter a informação que estamos buscando.

Como exemplo, abaixo será exibida uma tabela similar à tabela gerada.

TABELA 5.8 - Clientes com maior tendência de compra

Cliente	ClusterId	Probabilidade	Confiança
DOUGLAS STIERLE	0	0.696054	0.718094
LUCILA FRANCESCHETTO	1	0.702712	0.883922
EDIRCE CONTE	1	0.707144	0.98318
LUIZ ANTONIO TEFFILI	1	0.719262	0.704714
LUIZA BENASSI	1	0.721606	0.978016
MARIO RONSONI	0	0.732141	0.693702
DIRLEI DE RE	2	0.738703	0.852643
EVERTON ZAIONS	0	0.742446	0.89631
ANDRE CAVAGNI	0	0.752553	0.749539
TANIA FRIEDRICH	2	0.752937	0.978618
CELIA SCARANTO	1	0.772315	0.774198
FERNANDO BORGES	0	0.781043	0.672153

Com base na lista gerada, os proprietários, que já usam a mala-direta como canal de venda, consideram importante direcionar uma campanha de marketing a clientes que tenham uma probabilidade de compra de pelo menos 60%, com uma confiança mínima de 50%. O maior valor obtido para probabilidade foi de 78% com uma confiança de 67%. O retorno de malas-diretas é hoje em torno de 8%, e, segundo os proprietários, se esta taxa de retorno passar para pelo menos 15%, será vista como extremamente satisfatória.

Já para clientes com tendência a comprar conjuntos de louças para banheiro, obteve-se os clusters, exibidos na figura 18.

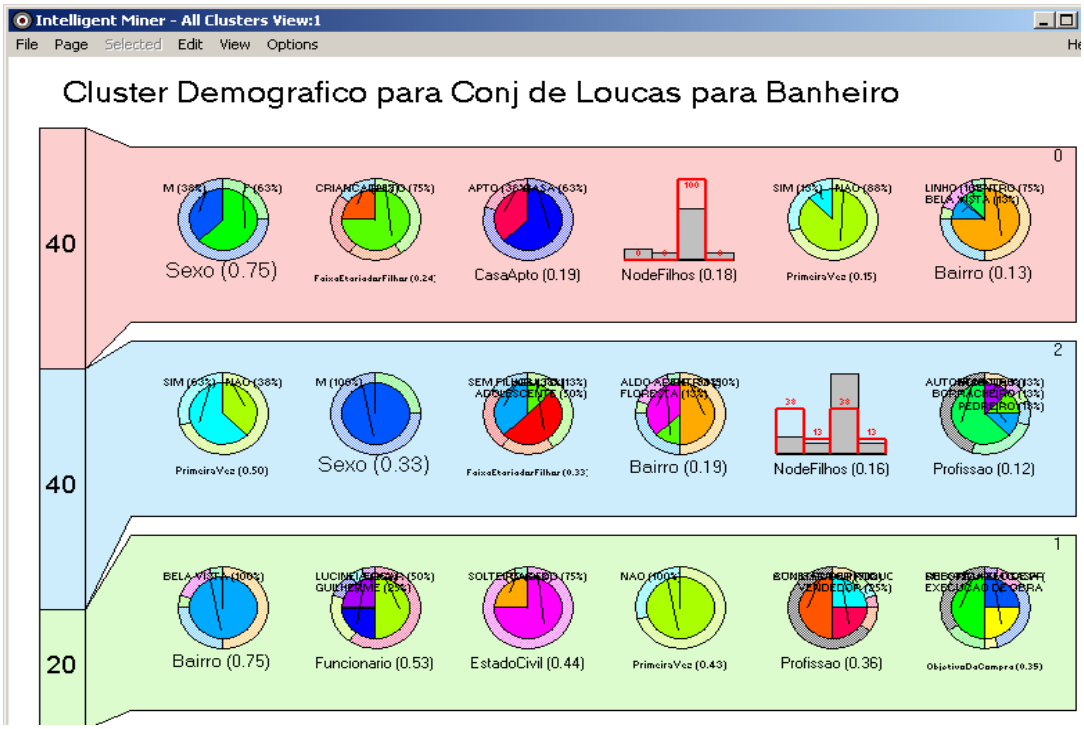

FIGURA 5.12 - Clientes que compraram Conjuntos de Louças para Banheiro

A figura mostra 3 clusters que determinam o perfil dos clientes que compraram conjuntos de louças para banheiro. Da mesma forma, quando a função de mineração em modo aplicação é executada, obtemos uma tabela de probabilidades. Uma parte dessa tabela de clientes que compraram conjuntos de louças para banheiro é exibida abaixo.

TABELA 5.9 - Clientes com maior tendência de compra - Conjunto de Louças para Banheiro

Cliente	ClusterId	Probabilidade	Confiança
EDIRCE CONTE	0	0.597544	0.652891
JOSE MARENGO	2	0.601247	0.598309
SALETE DA ROCHA	0	0.602337	0.662612
IRMA HILGERT	0	0.603523	0.560671
ISIDE B GERALDO	0	0.608259	0.77909
WLADYSLAV HAWRYLUK	1	0.609367	0.590761
SERGIO VASKIEU	1	0.61305	0.622845
EDUARDO SISMOSKI	1	0.614518	0.647631
VOLMAR HLAVAC	0	0.620117	0.532152
ODACIR RODRIGUES SANTOS	2	0.626496	0.706816
DEMETRIO VOLESKI	2	0.627801	0.728644
DARY IVO SCHAEFFER	0	0.628128	0.623238
ROQUE LUIZ GRZYBOSKI	2	0.62835	0.65159
MARIA LUIZA SCHAEFFER	0	0.645734	0.711266
ARGELINDO SKONRA	2	0.66633	0.634299
CELSO MARCOS OELKER	2	0.678826	0.849803

A tabela 5.9 já demonstra probabilidades mais baixas para a aquisição de conjunto de louças para banheiro, sendo a máxima de 68%, com confiança de 85%. Para a empresa, direcionar uma campanha aos clientes com pelo menos 55% de tendência a adquirir este produto é uma boa decisão. Deve-se definir, porém, qual o melhor canal de venda a ser usado.

5.4 Através de uma análise de cesta de mercado, conhecer quais produtos estão associados em transações de venda

Para realizar uma análise de cesta de mercado, não nos interessam dados de clientes. Somente das transações de venda. Porém, não é possível realizar este tipo de mineração se não existir, nos dados minerados, um identificador de transação de venda. Este campo será responsável por distinguir se um produto pertence a uma ou a outra transação e associar um produto com outro de mesmo identificador. Não existe nenhuma outra forma de realizar esta tarefa sem um identificador de transação. É importante que, no início de um projeto, já se saiba do interesse ou não de se fazer este tipo de mineração, uma vez que talvez seja necessário criar um campo de identificação. O que ocorre, na prática, é que os sistemas comerciais já são modelados para conter um identificador de transação, representado, por exemplo, pela chave primária de uma tabela de vendas. Um exemplo para uma chave primária poderia ser a composição dos campos de *Data da Venda*, no formato *Ano Mês Dia*, concatenado com um número seqüencial e incremental da venda no dia. Por exemplo, a quinta venda do dia onze de agosto de 2002, poderia ser representada pelo identificador de transação 200208110005. Neste estudo de caso, o identificador de venda é um simples número seqüencial, e foi herdado do atual sistema de faturamento da empresa.

Outro ponto a ser considerado é que o Intelligent Miner exige que este identificador de venda esteja ordenado na leitura dos dados. Porém, não é necessário se montar uma tabela ou arquivo-texto conforme esta necessidade pois existe um parâmetro que pode ser configurado para fazer com que o software faça esta ordenação automaticamente.

5.4.1 Função de mineração escolhida para geração do modelo

A função de mineração para este tipo de tarefa no Intelligent Miner é a Associations Mining Function.

5.4.2 Campos da tabela *Venda* usados

Para a mineração, selecionamos os seguintes campos da tabela *Venda:*

TABELA 5.10 - Campos da tabela *Venda* selecionados para a técnica de associação

Tabela *Venda*		
Nome do Campo	**Tipo**	**Tamanho**
IDVENDA	Texto	8
SUBCATEGPRODUTO	Texto	42

Não são necessários outros campos além destes. Conforme explicado anteriormente, os produtos, tanto na base de dados original quanto nas tabelas usadas na mineração, pertencem a uma categoria e a uma subcategoria. É necessário verificar a subcategoria do produto uma vez que um cliente pode ter comprado, por exemplo, um *Roupeiro de Luxo em Mogno* e outro pode ter comprado um *Roupeiro de Luxo em Marfim*. Ou ainda, um pode ter comprado um roupeiro 6x6 e outro pode ter comprado 4x4. Como a descrição dos produtos varia, a subcategoria a que eles pertencem é a mesma, e este campo, portanto, deve ser o alvo da mineração.

5.4.3 Parâmetros usados no processo de mineração

1) Minimum support: o suporte mínimo indica a ocorrência relativa da regra de associação detectada dentro do conjunto de dados. É determinado através da divisão do número de transações onde ocorre a regra de associação pelo número total de transações de venda. Ex.: para um conjunto de vendas de 100 transações, se a regra AREIA => CIMENTO (leia-se *a compra de areia implica na compra de cimento*) ocorrer em 10 transações de venda, o suporte mínimo a ser configurado para que esta regra seja gerada é de 10% (10 de 100). Para este estudo de caso, o suporte mínimo usado foi 2%.

2) Minimum confidence: a confiança mínima também indica a ocorrência relativa da regra de associação detectada dentro do conjunto de dados, mas é determinada através da divisão do número de transações onde ocorre a regra de associação pelo número de transações onde ocorre somente o corpo da regra. Ex.: para o mesmo exemplo anterior, se a regra AREIA => CIMENTO ocorrer em 10 transações de venda, mas existirem outras 20 regras do tipo AREIA => *OUTRO PRODUTO QUALQUER ,* então existem 30 das 100 transações onde o corpo da regra contém AREIA. Logo, se

quisermos que a regra AREIA => CIMENTO seja gerada, a confiança mínima definida deve ser de 33% (10 de 30). Para este estudo de caso, a confiança mínima usada foi 60%.

3) Maximum rule length: o tamanho máximo das regras determina o número máximo de itens que podem aparecer numa regra. Por exemplo, se RALO => JOELHO, então o número de itens desta regra é 2; se RALO + FLEXIVEL => JOELHO, então o número de itens desta regra é 3. Para este último caso, se o tamanho máximo de regras for 2, a regra não seria gerada. Para este estudo de caso, não foi limitado o número máximo de regras.

4) Item constraints: pode-se evitar, através da restrição de itens, que um ou mais itens sejam analisados pela função de mineração bastando escrever o nome do item que não se deseja visualizar. Nenhuma restrição foi feita neste estudo de caso.

5.4.4 Resultados gerados

Assim como nas outras funções de mineração, a alteração de parâmetros da função de associação faz variar bastante os resultados obtidos. As figuras 5.13 e 5.14 demonstram a forma como as associações são apresentadas pelo Intelligent Miner, seguidas de algumas conclusões sobre os resultados obtidos.

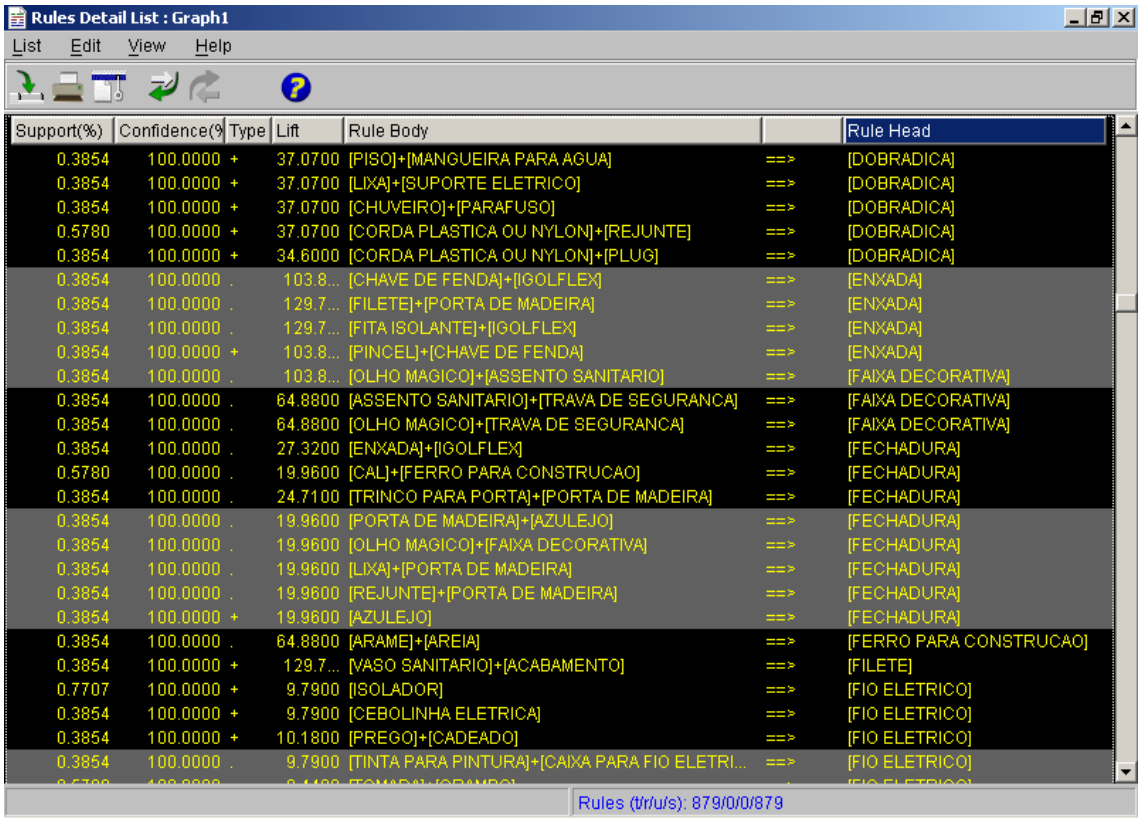

FIGURA 5.13 - Produtos associados em transações de venda – Gráfico 1

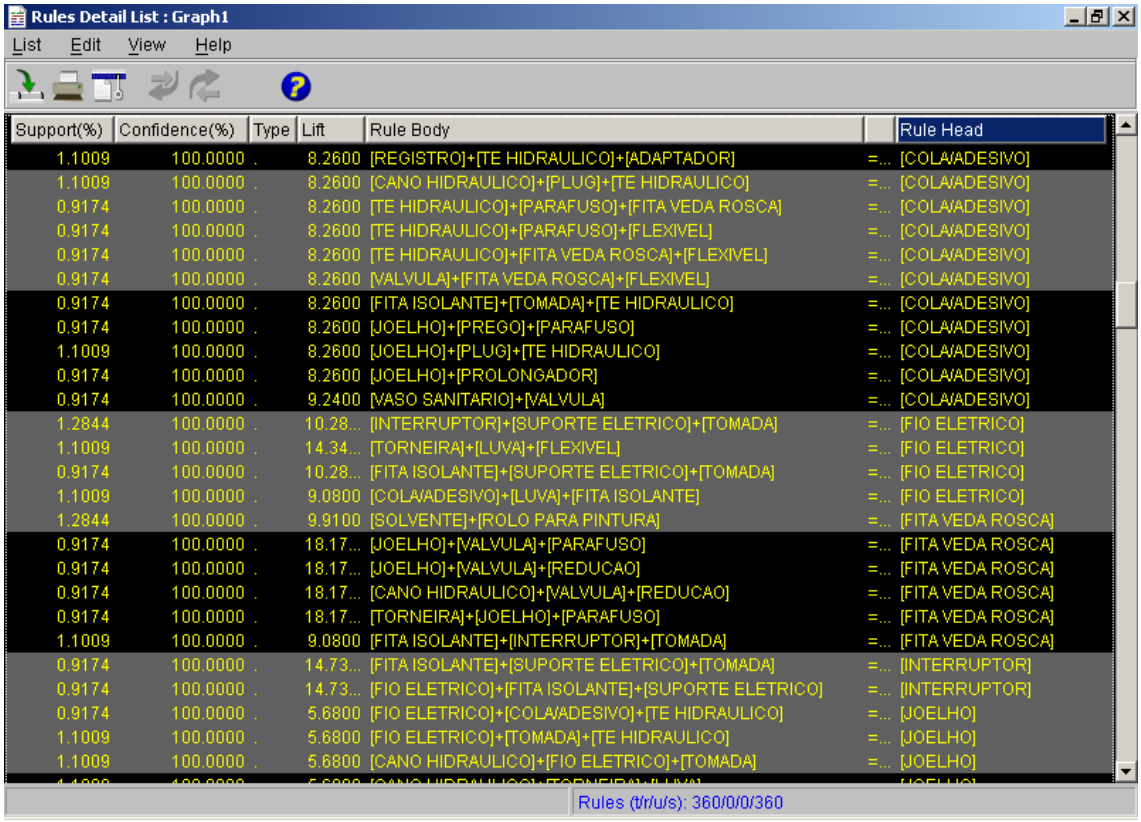

FIGURA 5.14 - Produtos associados em transações de venda – Gráfico 2

As listas de associações demonstradas nas figuras acima estão ordenadas por cabeçalho (Rule Head), ou seja, pela subcategoria de produto que é comprada em função do corpo da regra (que está à esquerda da seta ==>) .

O que se observa, na primeira análise, é o baixo valor para o suporte de cada regra, que na maioria dos casos, não ultrapassa 1%. Isto ocorre devido à grande variedade de produtos vendidos na loja. De fato, uma regra que demonstre uma confiança mínima muito alta ou até mesmo de 100%, independentemente do suporte mínimo, muito provavelmente será uma informação extremamente útil, se não for óbvia.

O suporte baixo não ocorre tanto para materiais hidráulicos, que são mais vendidos. O suporte médio para este tipo de produto é de 8%.

Nos resultados, várias regras foram encontradas e classificadas como regras interessantes. Deve-se considerar que algumas delas tiveram uma confiança de 100% pois foram criadas sobre casos isolados de vendas. Essas regras têm em comum o fato de terem um suporte próximo a zero.

Dois trabalhos estão sendo desenvolvidos em função destes resultados. Primeiro, uma nova distribuição visual dos produtos na loja, e o segundo, uma alteração no atual sistema de faturamento para que na montagem de um orçamento ou mesmo na emissão da nota fiscal, seja sugerida a venda de um produto que tenha associação com os outros produtos que estão sendo comprados. Os parâmetros de funcionamento desta rotina de programa serão configurados pelo usuário, que informará um suporte e uma confiança mínima.

Para a redistribuição visual, podemos citar alguns exemplos interessantes.

Lixas são vendidas associadas com vários produtos, com destaque para:

```
0.7707  1       [TOMADA]+[TRINCO PARA PORTA]                ==>     [LIXA]
0.5780  1       [SPOT]+[CADEADO]                            ==>     [LIXA]
0.5780  1       [SOLVENTE]+[PREGO]                          ==>     [LIXA]
0.5780  1       [FECHADURA]+[SOLVENTE]                      ==>     [LIXA]
```

Pincel para pintura possui as seguintes associações:

```
0.3854  1       [ROLO PARA PINTURA]+[ANTI CUPIM]            ==>     [PINCEL]
0.3854  1       [SOLVENTE]+[ANTI CUPIM]                     ==>     [PINCEL]
0.3854  1       [LIXA]+[ANTI CUPIM]                         ==>     [PINCEL]
0.3854  1       [CHAVE DE FENDA]+[ENXADA]                   ==>     [PINCEL]
0.3854  1       [CAL]+[POZOLIT]                             ==>     [PINCEL]
0.3854  1       [TINTA PARA PINTURA]+[ANTI CUPIM]           ==>     [PINCEL]
0.3854  1       [PISO]+[THINNER]                            ==>     [PINCEL]
```

Para o caso das tintas para pintura, podemos ver uma associação mais trivial com o produto Rolo para Pintura, que chega a atingir um suporte de 15%:

```
15.414  1       [SOLVENTE]+[ROLO PARA PINTURA]              ==>     [TINTA PARA PINTURA]
0.7707  1       [SOLVENTE]+[PINCEL]                         ==>     [TINTA PARA PINTURA]
0.5780  1       [SOLVENTE]+[PARAFUSO]                       ==>     [TINTA PARA PINTURA]
0.5780  1       [ARGAMASSA]+[THINNER]                       ==>     [TINTA PARA PINTURA]
0.7707  1       [ROLO PARA PINTURA]+[THINNER]               ==>     [TINTA PARA PINTURA]
0.7707  1       [ROLO PARA PINTURA]+[COLCHAO DE CASAL]      ==>     [TINTA PARA PINTURA]
```

Para assento sanitário, observamos que as compras estão associadas com outros produtos para banheiro, comprados até mesmo em função de uma reforma:

```
0.3854  1       [REJUNTE]+[PORTA PAPEL HIGIENICO]           ==>     [ASSENTO SANITARIO]
0.3854  1       [CUBA]+[GRANITO]                            ==>     [ASSENTO SANITARIO]
```

A compra de fechaduras pode estar sendo feita buscando apenas finalizar a montagem de uma porta nova, mas também pode-se estar querendo reforçar a segurança do imóvel. Temos casos de compras de olho mágico implicando na compra de fechaduras como podemos ver abaixo.

```
0.3854  1       [ENXADA]+[IGOLFLEX]                         ==>     [FECHADURA]
0.5780  1       [CAL]+[FERRO PARA CONSTRUCAO]               ==>     [FECHADURA]
0.3854  1       [TRINCO PARA PORTA]+[PORTA DE MADEIRA]      ==>     [FECHADURA]
0.6090  1       [PORTA DE MADEIRA]+[AZULEJO]                ==>     [FECHADURA]
1.5560  1       [OLHO MAGICO]+[FAIXA DECORATIVA]            ==>     [FECHADURA]
0.9899  1       [LIXA]+[PORTA DE MADEIRA]                   ==>     [FECHADURA]
0.4326  1       [REJUNTE]+[PORTA DE MADEIRA]                ==>     [FECHADURA]
0.9765  1       [AZULEJO]                                   ==>     [FECHADURA]
```

Fitas veda rosca também podem ser vendidas e expostas com produtos aos quais estão associadas:

```
12.844  1       [SOLVENTE]+[ROLO PARA PINTURA]              ==>     [FITA VEDA ROSCA]
0.9174  1       [JOELHO]+[VALVULA]+[PARAFUSO]               ==>     [FITA VEDA ROSCA]
0.9174  1       [JOELHO]+[VALVULA]+[REDUCAO]                ==>     [FITA VEDA ROSCA]
0.9174  1       [CANO HIDRAULICO]+[VALVULA]+[REDUCAO]       ==>     [FITA VEDA ROSCA]
0.9174  1       [TORNEIRA]+[JOELHO]+[PARAFUSO]              ==>     [FITA VEDA ROSCA]
11.009  1       [FITA ISOLANTE]+[INTERRUPTOR]+[TOMADA]      ==>     [FITA VEDA ROSCA]
```

Os joelhos hidráulicos são também bastante vendidos e, por conseqüência, são bastante numerosos nas regras de associação geradas, com destaque para:

```
12.844  1       [TOMADA]+[TE HIDRAULICO]+[FITA VEDA ROSCA]   ==>     [JOELHO]
20.183  1       [CANO HIDRAULICO]+[PARAFUSO]                 ==>     [JOELHO]
16.514  1       [FITA ISOLANTE]+[TE HIDRAULICO]              ==>     [JOELHO]
12.844  1       [TOMADA]+[TE HIDRAULICO]                     ==>     [JOELHO]
14.679  1       [PISO]+[TE HIDRAULICO]+[ARGAMASSA]           ==>     [JOELHO]
38.532  1       [COLA/ADESIVO]+[LUVA]+[PARAFUSO]             ==>     [JOELHO]
11.009  1       [FIO ELETRICO]+[TOMADA]+[TE HIDRAULICO]      ==>     [JOELHO]
```

A lista com todas as regras geradas é bastante extensa. As ferramentas de mineração de dados têm a capacidade de gerar muita informação deste tipo. Para se medir a efetividade dos resultados gerados é necessário verificar se o usuário vai ou não agregar mais produtos à sua "cesta de compras" em função de uma sugestão feita via sistema ou mesmo pela constatação visual de um produto que estava sendo esquecido.

Estas minerações de análise de cesta de mercado foram algumas das minerações consideradas mais interessantes pelos proprietários da loja. As redistribuições visuais já foram feitas considerando o espaço hoje disponível. Um ponto dos mais valiosos e importantes, segundo eles, é o fato de o sistema avisar quando um produto está sendo "esquecido" pelo cliente. É muito fácil e natural que tanto o cliente como o vendedor se esqueçam de comprar ou oferecer um produto que precisará ser usado. Isto faz com que o produto seja comprado em outro lugar ou não seja usado. O sistema, alertando isto, resolveria este tipo de problema.

5.5 Tentar prever se o cliente que está comprando pela primeira vez na loja tende a não voltar mais

Para prever se o cliente que está comprando pela primeira vez na loja tende a não voltar mais, usamos a função de predição do Intelligent Miner através do algoritmo de redes neurais (Neural Prediction Mining Function).

O primeiro ponto a ser analisado é se temos a informação de retorno do cliente na base de dados. Aqui, tanto no banco de dados original quanto nas tabelas preparadas para a mineração, não encontramos este indicador. Precisamos, portanto, incluir um campo que identifique se o cliente que realizou a compra foi um cliente que nunca mais voltou.

Analisando a tabela *Produto_Venda_Cliente,* percebemos que a mesma possui o nome do cliente e o identificador de venda , *Cliente e IdVenda*, respectivamente. Podemos então, gerar uma fonte de dados nova obtendo o nome dos clientes que estão comprando pela primeira vez (primeiravez = 'SIM') e varrendo a base verificando se o nome do cliente aparece de novo em alguma outra transação (com outro *IdVenda*).

As alterações necessárias na base de dados foram feitas via SQL na base Oracle:

- criação da **nova tabela** que contém o indicador de "cliente que retornou ou não":

```
create table produto_venda_cliente_pred storage (initial 1m next 200k pctincrease 0)
as select * from produto_venda_cliente;
alter table produto_venda_cliente_pred add(retornou varchar2(1));
```

- setando valor de ‘1’ para o campo *retornou,* supondo que todos os clientes retornaram:

```
update produto_venda_cliente_pred set retornou='1';
```

- atualizando, agora, o valor ‘0’ para aqueles clientes que compraram somente em uma transação, ou seja, somente uma vez:

```
update produto_venda_cliente_pred set retornou='0' where cliente in
(select distinct t1.cliente from produto_venda_cliente_pred t1 where not exists
 (select null from produto_venda_cliente_pred t2 where t1.idvenda <> t2.idvenda and
  t1.cliente = t2.cliente ) and t1.primeiravez = 'SIM' );
commit;
```

Aqui temos, então, uma tabela que contém o indicador de venda feita por cliente que retornou ou não.

5.5.1 Campos da tabela *Produto_Venda_Cliente_Pred* usados

Selecionamos os seguintes campos da tabela *Produto_Venda_Cliente_Pred*:

TABELA 5.11 - Campos selecionados para a técnica de predição

Tabela Poduto_Venda_Cliente_Pred		
Nome do Campo	**Tipo**	**Tamanho**
Profissao	Texto	44
EstadoCivil	Texto	14
TempodeCasado	Texto	6
NodeFilhos	Numérico	2
FaixaEtariadosFilhos	Texto	15
CasaApto	Texto	10
MoraQuantoTempo	Numérico	2
ResidenciaPropriaAlugada	Texto	11
Idade	Numérico	2
Sexo	Texto	1
Funcionario	Texto	14
ObjetivoDaCompra	Texto	60
Retornou	Numérico	1

5.5.2 Parâmetros usados no processo de mineração

1) In-Sample size e Out-Sample size: o primeiro representa o número de registros consecutivos a serem selecionados a partir dos dados de entrada durante a fase de aprendizado, enquanto o segundo representa o número de registros consecutivos a serem selecionados durante a fase de verificação, visando determinar se os objetivos quanto à exatidão e ao limite de erro foram atingidos. Por exemplo, se usarmos In-Sample = 4 e Out-Sample = 2, durante a fase de aprendizado, 4 registros alternadamente serão lidos e os 2 seguintes serão ignorados. Na fase de verificação, o processo é revertido: alternadamente, 4 registros são ignorados e 2 são usados. O *default* do Intelligent Miner é In-Sample = 2 e Out-Sample = 1. Neste estudo de caso, os valores *default* foram usados.

2) Maximum Number of Passes: limita o número de vezes que a função processa os dados. Se a rede neural atingir os níveis de precisão desejados, o processamento termina. O *default* é 0. O valor usado neste estudo de caso foi 6.

3) Forecast Horizon: representa o relacionamento entre os dados de entrada e o campo de predição. Se for setado para zero, os campos de entrada e o campo de predição serão baseados no mesmo registro. Para prever valores no futuro, deve-se especificar o número de períodos a prever. O valor *default* é zero e este foi o valor usado neste estudo de caso.

4) Window Size: representa o número de registros nos dados de entrada que serão usados para predizer um valor. O padrão é usar 1 registro (Window Size = 1). A combinação com o parâmetro Forecast Horizon determina os tamanhos dos registros lógicos de entrada. O valor usado foi 1.

5) Average Error: representa a percentagem de registros na amostragem de teste usada para determinar se o limite de erro especificado foi atingido. Pode-se especificar um valor para a média de erro RMS (Root Mean Square), que é calculado na verificação de amostragem Out-Sample. Durante a fase de *training*, o erro RMS diminui, devido ao fato de a qualidade das predições aumentar com o aprendizado. Valores para RMS entre 0,25 e 0,01 indicam predições perfeitas. O valor *default* para Average Error é 0,1 e este foi o valor usado.

6) Normalização dos dados: as redes neurais precisam trabalhar com dados normalizados. Para isto, o Intelligent Miner disponibiliza uma função de normalização automática, que, neste caso, precisou ser usada.

5.5.3 Resultados gerados

Para funções de predição, o Intelligent Miner mostra na esquerda de cada grupo o campo analisado destacando seu valor booleano SIM ou NÃO, que neste caso é 1 ou 0, para clientes que retornaram e para clientes que não retornaram, respectivamente. Os dados da tabela *Produto_Venda_Cliente_Pred* foram usados durante a fase de treino do modelo para aperfeiçoar o resultado através do algoritmo de redes neurais. Sobre o modelo treinado, é gerado então o resultado final, mostrado na figura 5.15.

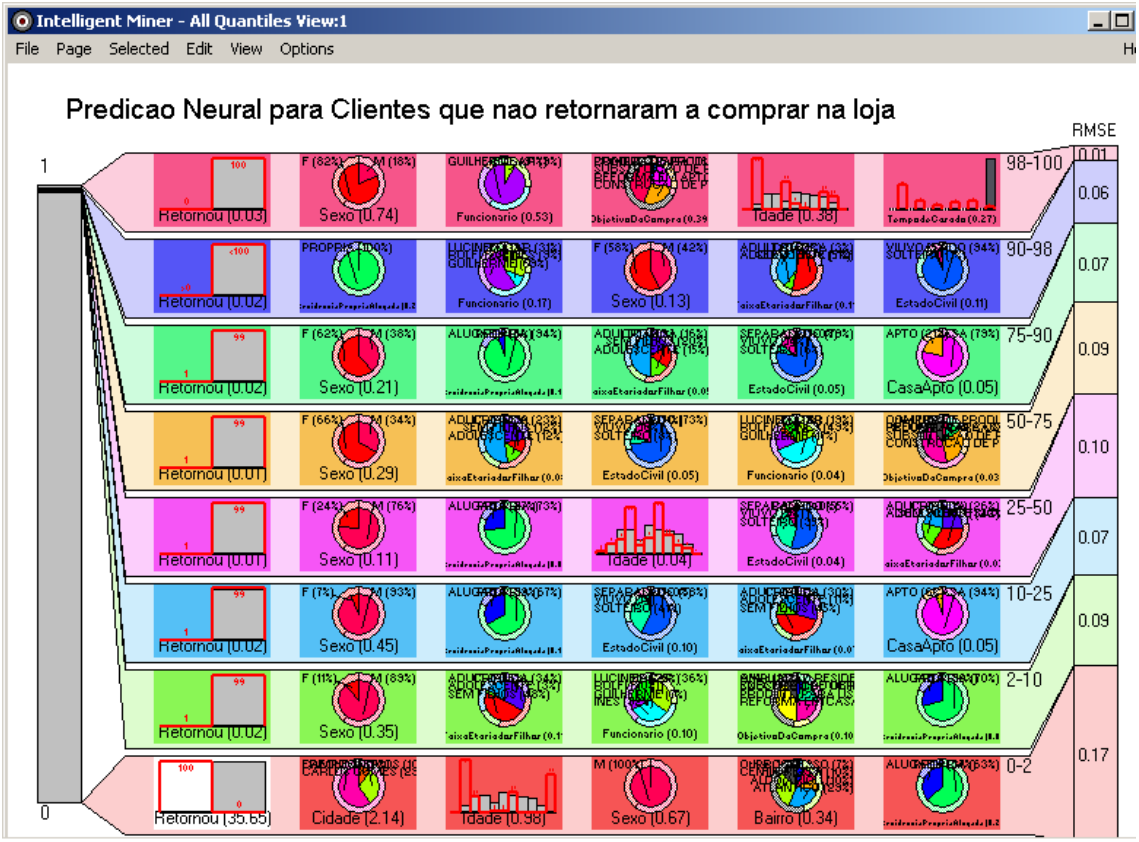

FIGURA 5.15 - Grupos gerados pelo algoritmo de redes neurais de predição

A grande vantagem desta função de mineração, é que para o algoritmo de redes neurais é possível converter o modelo gerado para um código fonte em linguagem C ANSI. Dessa forma, é possível embutir em qualquer programa a lógica gerada pelo modelo para determinar o que se espera, ou seja, neste caso, saber a probabilidade do cliente voltar a comprar na loja. Alguns trechos do código C gerado são mostrados a seguir.

```
/*--------------------------------------------------------------------*/
/*  Program generated by IBM DB2 Intelligent Miner for Data V6.1 */
/*--------------------------------------------------------------------*/
/* Name of model =    Predicao Neural para Clientes que nao retornaram a comprar na loja
** Generated at :     Sun Jan 27 23:20:33 2002
*/
#ifndef IDM_GLOBAL_DEFS

#define IDM_GLOBAL_DEFS
#include <math.h>   /* ? DBL_MIN */
#include <float.h>  /* ? DBL_MIN */

#ifndef DBL_MIN
#define DBL_MIN -0.1243e-30
#endif

#define IDMCHAR  char
#define IDMREAL  double
#define IDM_MISSING_REAL  (DBL_MIN)
#define IDM_MISSING_STRING  ((IDMCHAR*)0)
#endif

...
...
```

```
#define IDM_NFIELDS_NUM 4
struct { IDMREAL inMin, inMean, inMax; }
idm_numstats[IDM_NFIELDS_NUM+1] =  /* +1 entry for predicted */
{  { 15, 41.88174882629108, 82 }      /* Idade */ ,
      { 0, 22, 44 }    /* MoraQuantoTempo */ ,
      { 0, 6, 12 }     /* NodeFilhos */ ,
      { 0, 29.5, 59 }  /* TempodeCasado */ ,
      { 0, 0.5, 1 }};

#define IDM_NVALS1 2
char* idmModelcatvals1[IDM_NVALS1] =  /* CasaApto */
{ "CASA", "APTO" };

#define IDM_NVALS3 4
char* idmModelcatvals3[IDM_NVALS3] =  /* EstadoCivil */
{ "CASADO", "SOLTEIRO", "VIUVO", "SEPARADO" };

#define IDM_NVALS4 4
char* idmModelcatvals4[IDM_NVALS4] =  /* FaixaEtariadosFilhos */
{ "CRIANCA", "SEM FILHOS", "ADOLESCENTE", "ADULTO" };

...
...

#include <stdio.h>  /* printf() */
#include <conio.h>


int main()
{
   struct IDMModelInput inputValues;
   struct IDMModelPredictedValue pred;
   int rc;
   inputValues.Profissao = "DO LAR";
   inputValues.EstadoCivil = "CASADO";
   inputValues.TempodeCasado = 30;
   inputValues.NodeFilhos = 4;
   inputValues.FaixaEtariadosFilhos = "ADULTO";
   inputValues.CasaApto = "CASA";
   inputValues.MoraQuantoTempo = 3;
   inputValues.ResidenciaPropriaAlugada = "PROPRIA";
   inputValues.Idade = 55;
   inputValues.Sexo = "F";
   inputValues.Funcionario = "EDGAR";
   inputValues.ObjetivoDaCompra = "REFORMA EM CASA";
   rc = idmModelPredictValue(&inputValues,&pred);
   if ( rc!=0 ) {
      printf("/* something went wrong */\n");
   } /*endif*/
   clrscr();
   printf("Probabilidade do cliente retornar: '%g'\n",(float)pred.predictedValue);
   getch();
   return 0;
}
```

Os valores destacados em vermelho no código C acima foram as únicas linhas alteradas antes de executar o programa. Para exemplificar, o resultado gerado para a execução acima é mostrado na figura 5.16.

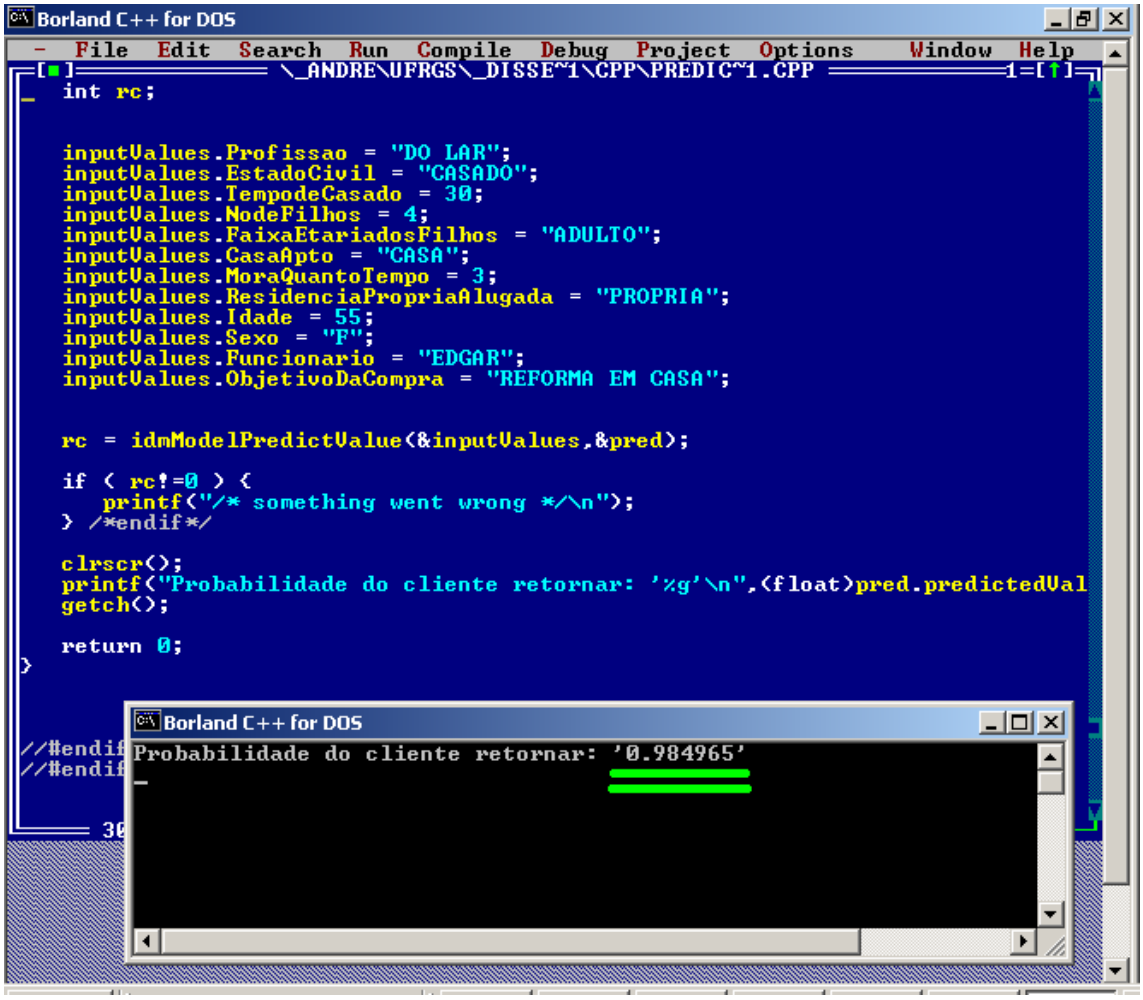

FIGURA 5.16 - Execução do código C++ para predição neural

Outra forma de visualizar os resultados, mas não de uma forma tão dinâmica, é criar listas onde conste o nome do cliente e a probabilidade de retorno do mesmo. Porém, não representa uma forma versátil de obtenção de resultados já que uma predição é algo que se deseja saber sobre algo do qual ainda não se tem informação.

O passo seguinte para a concretização deste objetivo é adaptar o código gerado ao sistema de informação da loja. Deve-se estudar qual a forma de tratamento que terá aquele cliente com tendência a não retornar. Talvez uma promoção para a compra seguinte ou mesmo um desconto especial por ser a primeira compra. Um desconto no frete caso o cliente tenha solicitado entrega a domicílio. Tudo é válido para fazer com que o cliente volte. Para os proprietários da loja, trata-se de uma informação totalmente nova e que antes nunca foi considerada. Segundo eles, existem alguns casos de clientes de outras cidades que entram no grupo das pessoas que não retornam mas pelo simples motivo de residirem numa outra cidade. De qualquer forma, esta informação merece atenção pois o número de clientes que não retorna foi maior do que o previsto.

5.6 Conhecer o perfil do cliente que compra pela primeira vez na loja

Este objetivo visa descobrir o perfil dos clientes que compram pela primeira vez na loja. Para isto, também usamos as fontes de dados com informações sobre os clientes e sobre as vendas. Trata-se da tabela *Produto_Venda_Cliente,* encontrada na base de dados Oracle do projeto de mineração. Para esta tarefa, também geramos um arquivo TXT contendo todas as informações dos clientes e suas compras através de comandos do utilitário Oracle SQL-PLUS. Tais comandos encontram-se no anexo 1 deste documento. Agrupamos os clientes de acordo com suas características com a finalidade de entender quem são as pessoas que estão vindo pela primeira vez na loja, de onde estão vindo, quando, e a procura de que, visando aumentar ainda mais o número de vendas.

A única restrição a ser feita sobre a base de dados é que o campo *PrimeiraVez* deve ter valor *SIM* , indicando que trata-se da primeira compra do cliente na loja.

5.6.1 Função de mineração escolhida para geração do modelo

A função de mineração escolhida foi a *Demographic Clustering*, do Intelligent Miner.

5.6.2 Campos da tabela *Produto_Venda_Cliente* usados

Para a mineração, selecionamos os seguintes campos da tabela *Produto_Venda_Cliente:*

TABELA 5.12 - Campos da tabela *Produto_Venda_Cliente* selecionados na montagem dos clusters de clientes que compram pela primeira vez

Tabela *Produto_Venda_Cliente*		
Nome do Campo	**Tipo**	**Tamanho**
PrimeiraVez	Texto	7
CasaApto	Texto	10
EstadoCivil	Texto	14
FaixaEtariadosFilhos	Texto	15
Idade	Numérico	2
MoraQuantoTempo	Numérico	2
ObjetivoDaCompra	Texto	60
ResidenciaPropriaAlugada	Texto	11
Sexo	Texto	1
TempoDeCasado	Numérico	2
CategProduto	Texto	34

5.6.3 Parâmetros usados no processo de mineração

1) Maximum passes: o valor usado foi 10.

2) Maximum clusters: o número de clusters usado foi 2.
3) Accuracy Improvement: o valor usado foi 1.
4) Similarity Threshold: o valor usado foi 0,5.
5) Outlier treatment: valores válidos.
6) Peso atribuído a cada campo:

TABELA 5.13 - Pesos dos atributos da tabela *Produto_Venda_Cliente* selecionados na montagem do perfil do cliente que compra pela primeira vez

Tabela *Produto_Venda_Cliente*	
Nome do Campo	**Peso**
CasaApto	5
EstadoCivil	6
FaixaEtariadosFilhos	1
Idade	4
MoraQuantoTempo	5
ObjetivoDaCompra	15
ResidenciaPropriaAlugada	7
Sexo	3
TempoDeCasado	15
CategProduto	20

5.6.4 Clusters gerados

Os clientes foram divididos e dois grupos. Podemos vê-los na figura 5.17.

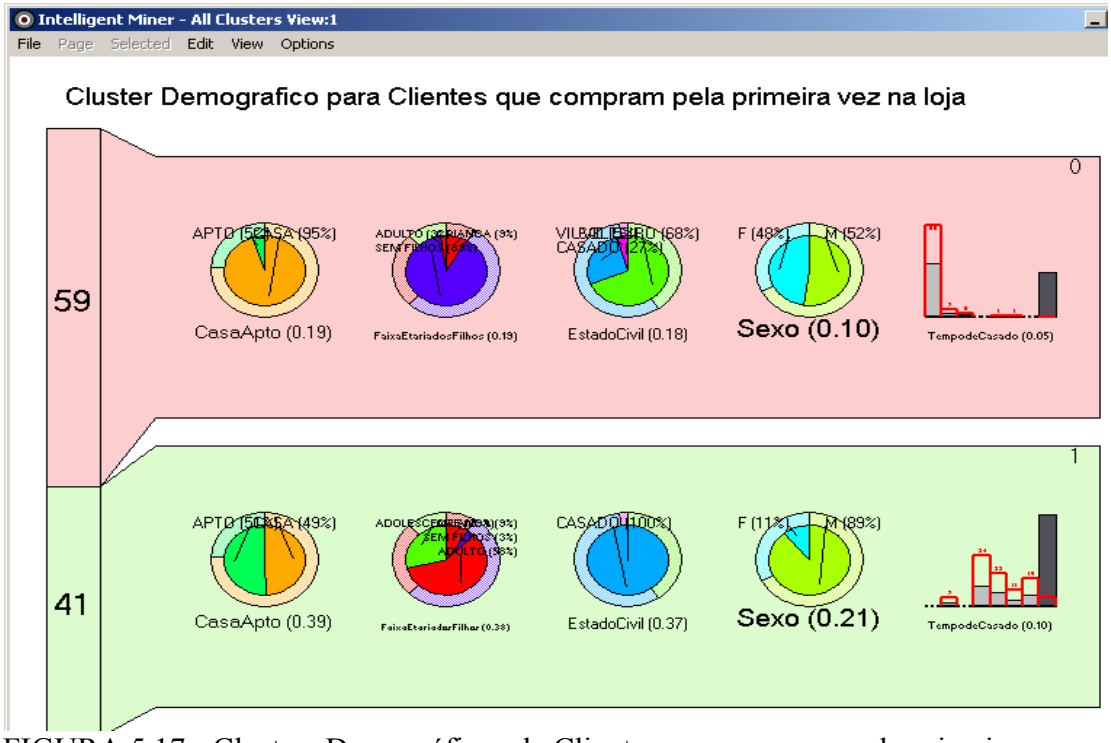

FIGURA 5.17 - Clusters Demográficos de Clientes que compram pela primeira vez na loja

A primeira impressão é de que não existe uma diferença muito significativa do perfil destes clientes em relação aos outros. No cluster 0, com 59% da população, vemos características já conhecidas como clientes que moram em casas, sem filhos, solteiros, sendo a maioria homens. No segundo cluster vemos um destaque para pessoas que moram em apartamentos, casados e com filhos adultos, sendo a grande maioria também homens. Mas o mais interessante é analisar o Tempo de Casado. Para o primeiro cluster (cluster 0), a quantidade de pessoas com tempo de casamento menor de 2 anos representa 85% , enquanto que no segundo cluster (cluster 1), 87% das pessoas estão casadas de 27 a 45 anos. Isto é uma informação bastante interessante. Clientes com tempo de casamento entre 4 e 16 anos têm uma participação insignificante nesta população. É claro que isto não é válido se considerarmos os outros clientes que compram na loja com freqüência.

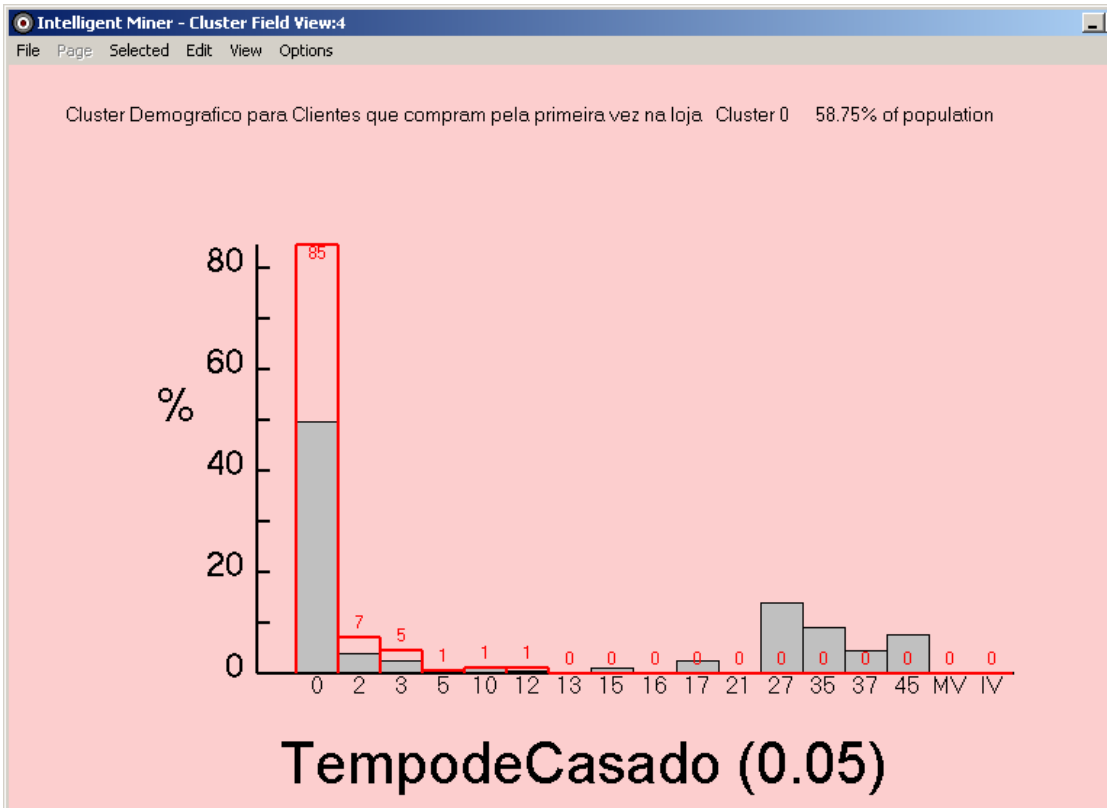

FIGURA 5.18 - Tempo de Casado de clientes que compram pela primeira vez na loja – Cluster 0

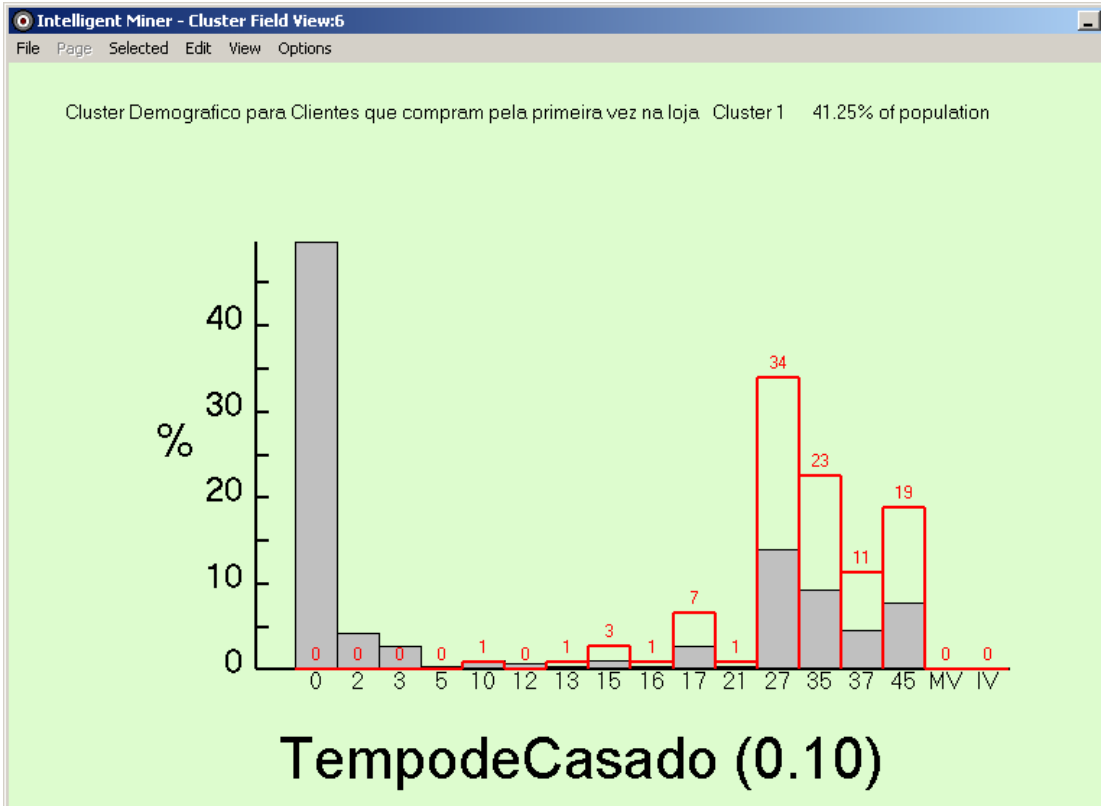

FIGURA 5.19 - Tempo de Casado de clientes que compram pela primeira vez na loja – Cluster 1

Mas o que determina esta divisão? O que recém-casados e casados a muito tempo buscam na loja para realizarem uma compra? Segundo os proprietários, recém-casados geralmente têm mais despesas na construção e reforma da nova residência. Este estudo de caso também nos mostrou que o tempo de residência acompanha o tempo de casamento, indicando que estas pessoas gastam mais no início do casamento. Estariam a procura de um preço mais baixo em função do total de gastos? E as pessoas casadas a bastante tempo, estariam também a procura de preços mais baixos ou consideram mais a qualidade dos produtos? Para os proprietários, esta informação precisa ser trabalhada para que se possa identificar uma forma de atrair mais novos clientes, e o tempo de casamento parece ser um excelente ponto a ser investigado.

5.7 Determinando os passos seguintes

De posse de todos os resultados obtidos, serão escolhidas as melhores formas de fazer uso dos mesmos. O número de relatórios, gráficos e estatísticas é bastante grande. Um resumo do principal foi exibido nesta dissertação, bem como as formas de se apresentar os resultados. Mas o que acontece, na prática, é a obtenção de um vasto material a ser analisado. Para citar um exemplo, mineramos a base de dados verificando a tendência de compra de vários produtos, mas exibimos apenas dois neste texto. Criamos clusters para outros grupos de usuários além dos clientes ou dos clientes que compram pela primeira vez. O que não muda é o processo e a forma como os resultados são exibidos. Feita uma análise detalhada sobre todas essas informações, deve-se fazer uso dos canais de venda hoje existentes para transformar estas informações em

resultados. É claro que tudo isto depende do orçamento disponível, mas este trabalho pode ser feito gradualmente. Os proprietários também trabalham com a idéia da contratação de uma empresa especializada neste tipo de marketing para definir como deve funcionar o processo. Os resultados da mineração já são conhecidos. Até a metade de 2003, o objetivo é poder avaliar qual o verdadeiro lucro que estes resultados trouxeram.

5.8 Planejando a monitoração e a manutenção

Os resultados precisam ser atualizados com o passar do tempo. O ideal mesmo é que a mineração fosse um processo *online*. Que os modelos fossem se alterando na medida em que mais informações fossem agregadas ao banco de dados. Não é o que acontece em nosso caso. Portanto, precisaremos atualizar todos os dados e em conseqüência os resultados obtidos a fim de que os mesmos reflitam a posição mais atual possível das informações. É indiscutível que um segundo trabalho de mineração será orientado de acordo com todo o aprendizado obtido por este. Não serão apenas alterados os dados minerados. Muitas variáveis serão incluídas. Outras serão retiradas, implicando na reconstrução dos modelos de mineração.

6 Conclusão e sugestão para trabalhos futuros

O presente trabalho foi útil para a identificação de pontos importantes a serem levados em conta em um trabalho de mineração de dados. Começando pela metodologia – CRISP – que foi adotada para conduzir o processo. É um guia fundamental para qualquer projeto nesta área. Sem uma metodologia, a necessidade de iteração para fazer correções sem dúvida será maior. As ferramentas usadas – IBM Intelligent Miner e WEKA – mostraram-se bastante produtivas. Ambas merecem destaque quanto ao processo de instalação pela facilidade. O WEKA é muito rigoroso no tratamento de arquivos-texto. Qualquer caracter fora do lugar impede a leitura do arquivo de dados. Exige uma formatação especial que atrapalha um pouco o trabalho. O Intelligent Miner leva vantagem sobre o WEKA neste ponto, mas a delimitação de campos no caso de fontes de dados em arquivo-texto também exige muita atenção. Por outro lado, o WEKA possui uma bibliografia de fácil acesso e *help online* através de uma interface Java, que é um ponto bastante positivo. O Intelligent Miner destaca-se positivamente pelo seus *Wizards* desde a montagem das fontes de dados até a execução das minerações. Os resultados são exibidos, em sua maioria, de forma gráfica. Para um usuário iniciante, isto dificulta um pouco a compreensão. As árvores de decisão são bastante difíceis de entender. Árvores grandes dão muito trabalho ao usuário para identificar regras. Já as regras de associação são fáceis de gerar, mas quando se tem um valor muito grande de regras geradas, a exibição das mesmas é demorada. O WEKA destaca-se na geração de regras de associação pela facilidade de manipulação dos parâmetros de mineração, coisa que é bastante fechada no Intelligent Miner, limitando bastante a interação. O código fonte em C gerado para predições em algoritmos de redes neurais, no Intelligent Miner, é perfeito. Não foi encontrado nenhum problema de compilação ou "linkagem" no momento de execução. Podemos dizer, então, de forma geral, que as duas ferramentas são muito boas para realizar o trabalho para o qual elas foram criadas, mas que podem ser melhoradas em alguns aspectos. Um ponto que poderia ser trabalhado nestas e em outras ferramentas de mineração de dados é a construção das mesmas visando o usuário final, que conhece o negócio da empresa. Hoje isto é impossível. São necessários bons conhecimentos de informática, se possível bons conhecimentos de Windows, banco de dados, seja Oracle, DB2 ou mesmo Access, e um conhecimento de planilhas eletrônicas, além de, é claro, conhecer o negócio da empresa e ter boas noções de marketing. Isto implica no envolvimento de pelo menos 3 especialistas. Neste estudo de caso não seria possível que os usuários finais, que conhecem o negócio da empresa, manipulassem uma ferramenta de mineração alterando parâmetros, retirando campos e gerando novos modelos para obter mais resultados. Existem muitas mensagens geradas pelos programas que exigem conhecimento técnico para compreendê-las. Mas a interação total com a ferramenta e a obtenção dos melhores resultados somente seria possível se o próprio usuário final, especialista no negócio, pudesse manipular os dados e os algoritmos oferecidos pela ferramenta de mineração, coisa que hoje ainda não é possível.

Analisando um pouco os dados usados, foi sentida uma dificuldade muito grande com a classificação dos produtos. Este é um ponto a ser destacado seja qual for a metodologia usada. Revisar a classificação dos produtos é trabalhoso mas de extrema importância. Deve-se tomar cuidado com generalizações. Produtos para diferentes fins que pertencem à mesma categoria geram resultados errados nas minerações. Produtos com venda casada, como por exemplo uma Cozinha Completa, poderá ser vista pelo

sistema como um produto único, quando na verdade é composta de vários outros produtos. Um levantamento dos clientes que compram mesas de cozinha já não incluiria aquele cliente que comprou uma cozinha completa quando na verdade deveria incluir. Isto é um problema a ser tratado pelo analista de mineração de dados ou pelo programador do sistema de vendas. A correção destes tipos de problemas, uma vez que os dados já tenham sido colhidos, é uma das mais demoradas tarefas de um projeto de mineração de dados.

Se estes tópicos forem levados em conta, a chance de se obter informações de valor ao final do projeto é muito grande. As ferramentas através de seus algoritmos têm uma capacidade muito grande de gerar resultados que possam se transformar em lucro para a empresa se forem bem trabalhados.

Comercialmente falando, muitas informações de valor foram obtidas. Para os proprietários, numa visão geral deste trabalho, um novo conceito de relacionamento com o cliente será desenvolvido usando como base as novas informações. As regras de associação geradas tanto no Intelligent Miner como no WEKA foram, em particular, merecedoras de uma maior atenção por parte dos mesmos. A necessidade de padronização na descrição e classificação dos itens de venda foi entendida pelos proprietários como um fator determinante de sucesso, tanto para trabalhos deste tipo quanto para a própria informação gerencial e organizacional da loja.

Por fim, a disseminação do uso de ferramentas de mineração de dados em pequenas e médias empresas pode representar a diferença num mercado competitivo onde é cada vez mais difícil ganhar novos clientes ou manter um cliente fidelizado.

Anexo

Principais comandos SQL usados

Este anexo exibe os principais comandos SQL usados no desenvolvimento e manipulação de dados do estudo de caso.

>>>>>>>>>>>>>>>>>>>>>>>>>>>> Tabela de CLIENTES

```
-- drop table cliente;
create table cliente
(
 Cliente  varchar2(44), PrimeiraVez varchar2(7), Profissao varchar2(44),
 EstadoCivil varchar2(14), TempodeCasado varchar2(6), NodeFilhos varchar2(6),
 FaixaEtariadosFilhos varchar2(15), Telefone varchar2(14), Email varchar2(34),
 Endereco varchar2(44), CasaApto varchar2(10), Bairro varchar2(34), Cidade varchar2(24),
 UF varchar2(6), MoraQuantoTempo varchar2(6), ResidenciaPropriaAlugada varchar2(11),
 Idade varchar2(10), Peso varchar2(15), Altura varchar2(19), Sexo varchar2(6)
) storage (initial 1m next 200k pctincrease 0);

alter table cliente add ( constraint CLIENTE_PK primary key (cliente)
using index storage (initial 500k next 100k pctincrease 0));

truncate table cliente;
SPOOL c:\temp\c.log
@c:\temp\c.txt
spool off

spool c:\temp\cliente.txt
set head off
set lines 388
set pages 10000
select * from cliente;
spool off

-- drop table weka_cliente;
create table weka_cliente storage (initial 1m next 200k pctincrease 0)
as select * from cliente where rownum < 1;

SPOOL c:\temp\wc.log
@c:\temp\wc.txt
spool off
```

>>>>>>>>>>>>>>>>>>>>>>>>>>>> Tabela de VENDAS

```
-- drop table venda;
create table venda
(
 Cliente varchar2(44), IdVenda varchar2(8), CategProduto varchar2(34),
 SubcategProduto varchar2(42), Produto varchar2(44), PrecoUnit varchar2(12),
 Qtde varchar2(9), Funcionario varchar2(14), DecisaoDeCompra varchar2(6),
 ObjetivoDaCompra varchar2(60), CondPagto varchar2(10), Hora varchar2(6),
 Tempo varchar2(11), Temperatura varchar2(13), Dia varchar2(6), Mes varchar2(6),
 DiaDaSemana varchar2(17), Ano  varchar2(10)
) storage (initial 1m next 200k pctincrease 0);

truncate table venda;
SPOOL c:\temp\v.log
@c:\temp\v.txt
spool off
select count(*) from venda;

spool c:\temp\venda.txt
set head off
set feed off
set lines 370
set pages 10000
select * from venda;
spool off

-- drop table weka_venda;
create table weka_venda storage (initial 1m next 200k pctincrease 0)
as select * from venda where rownum < 1;

SPOOL c:\temp\wv.log
@c:\temp\wv.txt
```

```
spool off
```

>>>>>>>>>>>>>>>>>>>>>>>>>> Tabela de PRODUTO_VENDA_CLIENTE

```
-- drop table PRODUTO_VENDA_CLIENTE;
create table PRODUTO_VENDA_CLIENTE storage (initial 1m next 200k pctincrease 0) as
select c.*, v.IdVenda , v.CategProduto , v.SubcategProduto , v.Produto , v.PrecoUnit ,
v.Qtde , v.Funcionario , v.DecisaoDeCompra, v.ObjetivoDaCompra , v.CondPagto , v.Hora ,
v.Tempo , v.Temperatura , v.Dia , v.Mes , v.DiaDaSemana ,  v.Ano
from cliente c, venda v
where c.cliente=v.cliente;

spool c:\temp\PRODUTO_VENDA_CLIENTE.txt
set head off
set feed on
set lines 714
set pages 10000
select * from PRODUTO_VENDA_CLIENTE;
spool off
```

>>>>>>>>>>>>>>> Objetivo: Prever se um cliente novo nao retornara

```
create table produto_venda_cliente_pred storage (initial 1m next 200k pctincrease 0)
as select * from produto_venda_cliente;

set pages 1000
select distinct t1.cliente,t1.idvenda
from produto_venda_cliente_pred t1 where not exists (select null
         from  produto_venda_cliente_pred t2
         where t1.idvenda <> t2.idvenda and
   t1.cliente = t2.cliente )
and t1.primeiravez = 'SIM'
order by 1;

-- Incluir uma coluna "NAO RETORNOU" na tabela produto_venda_cliente_pred:

alter table produto_venda_cliente_pred add(retornou varchar2(1));
update produto_venda_cliente_pred set retornou='1';
update produto_venda_cliente_pred set retornou='0'
where cliente in
(
 select distinct t1.cliente
 from produto_venda_cliente_pred t1 where not exists (select null
           from produto_venda_cliente_pred t2
           where t1.idvenda <> t2.idvenda and
    t1.cliente = t2.cliente )
 and t1.primeiravez = 'SIM'
);

spool c:\temp\produto_venda_cliente_pred.txt
set head off
set feed on
set lines 715
set pages 10000
select * from produto_venda_cliente_pred;
spool off
```

>>>>>>>>>>>>>>> Objetivo: Analise de cesta de mercado

```
set pages 0
select distinct idvenda from produto_venda_cliente where idvenda in
(
select distinct to_number(idvenda)
from produto_venda_cliente P
where subcategproduto = upper('&u')
and exists (select null from produto_venda_cliente p2 where p.idvenda = p2.idvenda and
     subcategproduto = upper('&u') and
     exists(select null from produto_venda_cliente p3
            where p2.idvenda=p3.idvenda and subcategproduto = upper('&u')))
) order by 1;
```

>>>>>>>>>>>>>>> Objetivo: Regras por Arvore de Decisao - WEKA

```
-- drop table WEKA_PRODUTO_VENDA_CLIENTE;
create table WEKA_PRODUTO_VENDA_CLIENTE storage (initial 1m next 200k pctincrease 0) as
select c.CLIENTE,c.PRIMEIRAVEZ,c.PROFISSAO,c.ESTADOCIVIL,c.TEMPODECASADO,c.NODEFILHOS,
c.FAIXAETARIADOSFILHOS,c.TELEFONE,c.EMAIL,c.ENDERECO,c.CASAAPTO,c.BAIRRO,c.CIDADE,
c.UF,c.MORAQUANTOTEMPO,c.RESIDENCIAPROPRIAALUGADA,c.IDADE,c.PESO,c.ALTURA,c.SEXO,
```

```
v.IdVenda  ,  v.CategProduto  ,  v.SubcategProduto  ,  v.Produto  ,    v.Funcionario  ,
v.DecisaoDeCompra, v.ObjetivoDaCompra , v.CondPagto , v.Hora , v.Tempo ,v.Temperatura ,
v.Dia , v.Mes , v.DiaDaSemana ,      v.Ano
from weka_cliente c, weka_venda v
where c.cliente=v.cliente;

--spool c:\temp\WEKA_PRODUTO_VENDA_CLIENTE.txt
spool C:\Arquiv~1\Weka-3-2\_WEKA_PRODUTO_VENDA_CLIENTE.txt
set head on
set feed off
set lines 230
set pages 10000
select
CLIENTE                    ||','||PRIMEIRAVEZ            ||','|| PROFISSAO        ||','||
ESTADOCIVIL                ||','||TEMPODECASADO          ||','|| NODEFILHOS       ||','||
FAIXAETARIADOSFILHOS       ||','||TELEFONE               ||','|| EMAIL            ||','||
ENDERECO                   ||','||CASAAPTO               ||','|| BAIRRO           ||','||
CIDADE                     ||','||UF                     ||','|| MORAQUANTOTEMPO  ||','||
RESIDENCIAPROPRIAALUGADA   ||','||IDADE                  ||','|| PESO             ||','||
ALTURA                     ||','||SEXO                   ||','|| IDVENDA          ||','||
CATEGPRODUTO               ||','||SUBCATEGPRODUTO        ||','|| PRODUTO          ||','||
FUNCIONARIO                ||','||DECISAODECOMPRA        ||','|| OBJETIVODACOMPRA ||','||
CONDPAGTO                  ||','||HORA                   ||','|| TEMPO            ||','||
TEMPERATURA                ||','||DIA                    ||','|| MES              ||','||
DIADASEMANA                ||','||ANO                    ||','||
from WEKA_PRODUTO_VENDA_CLIENTE;
spool off

-- Substituir no arquivo TXT todos os ,, por ,Null,
-- O Weka nao aceita o caracter " antes de uma virgula delimitadora
-- Colocar os valores distintos ao lado dos campos
-- substituir insert_into_venda por insert into venda
-- substituir insert_into_cliente por insert into cliente e _values_ por  values
-- apagar spool off dos arquivos

select distinct &u||','  from weka_produto_venda_cliente;

@relation RegrasClientesVendasMk
@attribute PRIMEIRAVEZ {SIM,NAO}
@attribute PROFISSAO {ADMINISTRADOR_DE_EMPRESAS,...,VENDEDOR}
@attribute ESTADOCIVIL {CASADO,SEPARADO,SOLTEIRO,VIUVO}
@attribute TEMPODECASADO {0,...,50}
@attribute NODEFILHOS {0,1,12,2,3,4,5,8,9}
@attribute FAIXAETARIADOSFILHOS {ADOLESCENTE,ADULTO,CRIANCA,SEM_FILHOS}
@attribute CASAAPTO {APTO,CASA}
@attribute BAIRRO {AEROPORTO,...,VILA_FELIZ}
@attribute MORAQUANTOTEMPO {0,...,44}
@attribute RESIDENCIAPROPRIAALUGADA {ALUGADA,PROPRIA}
@attribute IDADE {15,...,82}
@attribute PESO {ABAIXO_DE_60KG,ACIMA_DE_85KG,ENTRE_60_E_85KG}
@attribute ALTURA {ENTRE_1.60M_E_1.75M,MAIS_DE_1.75M,MENOS_DE_1.60M}
@attribute SEXO {M,F}
@attribute CATEGPRODUTO {FERRAMENTAS,...,UTILITARIO_RESIDENCIAL}
@attribute SUBCATEGPRODUTO {ABAJOUR,...,VERNIZ}
@attribute FUNCIONARIO {EDGAR,GUILHERME,INES,LUCINEIA,ROLFI}
@attribute OBJETIVODACOMPRA {AMPLIACAO_RESIDENCIAL...,SUBSTITUICAO_DE_PRODUTO...}
@attribute DIADASEMANA {SEGUNDA-FEIRA,...,SABADO}

@data
-- validar a base de dados:
-- clientes que existem no registro de vendas mas nao estao cadastrados:
select distinct ':'||cliente||':' from venda where cliente not in(select cliente from
cliente) order by 1;
-- clientes que nao fizeram nenhuma compra mas estao cadastrados:
select distinct cliente from cliente where cliente not in (select cliente from venda)
order by 1;
-- listar por nomes:
select distinct ':'||cliente||':' from clientes order by 1;
select distinct ':'||cliente||':' from vendas order by 1;
-- truncar TODAS as tabelas do esquema !
spool c:\_andre\temp\drops.sql
select 'truncate table '||table_name||';' from user_tables;
spool off
@c:\_andre\temp\drops.sql
```

Referências

[BRA 96] BRACHMAN, R. J.; ANAND, T. The Process of Knowledge Discovery in Databases: a Human-Centered Approach. In: FAYYAD, U. M. et al. (Ed.). **Advances in Knowledge Discovery and Data Mining.** Menlo Park: AAAI, 1996. p. 45 – 51.

[CHA 00] CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0:** Step by step data mining guide. Menlo Park: AAAI, 2000. p. 74 – 75.

[FAY 96] FAYYAD, U. From Data Mining to Knowledge Discovery: an overview. In: FAYYAD, U. M. et al. (Ed.). **Advances in Knowledge Discovery and Data Mining.** Menlo Park: AAAI, 1996. p. 30 – 32.

[FRA 91] FRAWLEY, W.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. Knowledge Discovery in Databases: An Overview. In: FRAWLEY, W. et al. (Ed.). **Knowledge Discovery in Databases.** Cambridge: AAAI/MIT Press, 1991. p. 81 – 88.

[WEI 98] WEISS, S. M.; INDURKHYA, N. **Predictive Data Mining:** a Practical Guide. San Mateo, USA: Morgan Kaufmann, 1998.

Obras consultadas

ABEMD – ASSOCIAÇÃO BRASILEIRA DE MARKETING DIRETO. **Web Site.** Disponível em: <http://www.abemd.org.br>. Acesso em: set. 2002.

BARAGOIN, C.; ANDERSEN, C.; BAYERL, S.; BENT, G.; LEE, J.; SCHOMMER, C. **Mining your own business in Banking - Using DB2 Intelligent Miner for Data.** San Jose: IBM Press, 2001.

BARAGOIN, C.; ANDERSEN, C.; BAYERL, S.; BENT, G.; LEE, J.; SCHOMMER, C. **Mining your own business in Retail - Using DB2 Intelligent Miner for Data.** San Jose: IBM Press, 2001.

BARAGOIN, C.; ANDERSEN, C.; BAYERL, S.; BENT, G.; LEE, J.; SCHOMMER, C. **Mining your own business in Telecoms - Using DB2 Intelligent Miner for Data.** San Jose: IBM Press, 2001.

BERRY, M.; LINOFF, G. **Data Mining Techniques :** for marketing, sales and customer support. New York: John Wiley, 1997.

DIRECT MARKETING ASSOCIATION. **Web Site**. Disponível em: <http://www.the-dma.org>. Acesso em: set. 2002.

IBM SOFTWARE DATABASE AND DATA MANAGEMENT - IBM DB2 INTELLIGENT MINER FOR DATA. **Web Site.** Disponível em: <http://www-4.ibm.com/software/data/iminer/fordata>. Acesso em: set. 2002.

MICROSOFT AND LEADING DATA MINING VENDORS LINE UP OF DATA MINING. **Web Site.** Disponível em: <http://www.microsoft.com/PressPass/press/2000/Mar00/DataMining>. Acesso em: set. 2002.

MICROSOFT SERVERS – DATA MINING. **Web Site.** Disponível em: <http://www.microsoft.com/SQL/productinfo/datamine.htm>. Acesso em: set. 2002.

NATIONAL CENTER FOR DATA MINING AND DATA MINING RESEARCH. **Web Site.** Disponível em: <http://www.ncdm.uic.edu>. Acesso em: set. 2002.

ORACLE DATA MINING SUITE – DATA WAREHOUSE – ORACLE INTERNET PLATFORM. **Web Site.** Disponível em: <http://www.oracle.com/ip/analyze/warehouse/datamining>. Acesso em: set. 2002.

PANORAMA BRASIL. **Web Site.** Disponível em: <http://www.panoramabrasil.com.br>**.** Acesso em: set. 2002.

THE DATA MINING GROUP. **Web Site.** Disponível em: <http://www.dmg.org>. Acesso em: set. 2002.

WITTEN, I.; FRANK, E. **WEKA - Machine Learning Algorithms in Java.** Hamilton: Morgan Kaufmann, 2000.