WAGNER AUGUSTO NASCIMENTO FILHO

APLICAÇÃO DO PROCESSO DE DATA MINING UTILIZANDO OS ALGORITMOS DE ÁRVORE DE DECISÃO E CLUSTERING

Monografia apresentada à Escola Politécnica da Universidade de São Paulo para conclusão do curso de MBA em Engenharia de Software

Área de Concentração:
Engenharia de Software
Orientador: Prof. Dr. Jorge Rady de Almeida Junior

São Paulo
2008

FICHA CATALOGRÁFICA

Nascimento Filho, Wagner Augusto
Aplicação do processo de data mining utilizando os algoritmos de árvore de decisão e clustering / W.A. Nascimento Filho. -- São Paulo, 2008.
74 p.

Monografia (MBA em Engenharia de Software) - Escola Politécnica da Universidade de São Paulo. Programa de Educação Continuada em Engenharia.

1.Mineração de dados 2.Algoritmos I.Universidade de São Paulo. Escola Politécnica. Programa de Educação Continuada em Engenharia II.t.

DEDICATÓRIA

Dedico este trabalho a minha esposa Fabiola.

OK

AGRADECIMENTOS

Ao Prof. Dr. Jorge Rady de Almeida Junior pela orientação e pelo conhecimento transmitido durante todo o trabalho.

Aos profissionais que trabalham comigo na EXECPLAN, com os quais eu aprendo diariamente a evoluir na minha carreira profissional, através da troca de experiências constantes.

Aos amigos que colaboraram direta ou indiretamente na execução desse trabalho.

Aos meus pais que sempre me apoiaram e me incentivaram nos meus estudos.

Tudo vale a pena quando a
alma não é pequena.
(Fernando Pessoa)

RESUMO

A busca por informações valiosas vem se tornando cada vez mais freqüente dentro das organizações, e pode representar um grande diferencial em um mercado cada vez mais competitivo. Neste contexto, é apresentado um estudo sobre o processo de Data Mining realizando a aplicação dos algoritmos de Árvore de Decisão e Clustering, mostrando a execução de cada etapa do processo e os resultados obtidos, podendo ser utilizado como uma poderosa ferramenta na busca por informações valiosas.

Palavras-chave: Data Mining, Árvore de Decisão, Clustering

ABSTRACT

The search for valuable information is becoming increasingly frequent within organizations, and may represent a large gap in an increasingly competitive market. In this context, it presented a study on the process of realizing Data Mining algorithms for the implementation of Decision trees and Clustering, showing the execution of each stage of the process and results, and can be used as a powerful tool in the quest for valuable information.

LISTAS DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DW Data Warehouse

BI Business Intelligence

SUMÁRIO

1 INTRODUÇÃO

Atualmente as empresas possuem um volume de informação enorme, proveniente dos seus bancos de dados transacionais ou gerenciais, além de informações externas que estão disponíveis para consulta. Esse volume de informação muitas vezes acaba atrapalhando na busca pelas informações que realmente importam e que possam otimizar os resultados das empresas.

A crescente competitividade do mundo dos negócios faz com que as empresas não se satisfaçam somente com ferramentas que possibilitam a simples análise dos dados, mesmo que essas ofereçam inúmeras funcionalidades e total flexibilidade na manipulação dos dados.

As empresas estão em busca por ferramentas que, além de possuírem esses recursos, agreguem inteligência ao seu negócio, de forma a direcionarem ou darem respostas a perguntas não respondidas pela simples análise dos dados. Dentre as inúmeras ferramentas que vêm sendo propostas ao mercado encontra-se o processo de Data Mining utilizando o algoritmo de Árvore de Decisão e o algoritmo de Clustering.

1.1 MOTIVAÇÃO

Um dos principais aspectos que motivaram o autor deste trabalho é a necessidade que os clientes da empresa em que trabalha solicitarem, constantemente, ferramentas que de fato direcionem o processo de tomada de decisão dentro das empresas.

A empresa em que o autor trabalha, atua no setor de desenvolvimento de software, especificamente de soluções para o nível tático e estratégico das empresas, soluções de Business Intelligence, e atualmente encontra-se na etapa de pesquisa para criação de soluções utilizando técnicas de Data Mining.

1.2 OBJETIVO

A utilização de ferramentas de simples análise dos dados não supre as necessidades de informações das empresas, muitas vezes perde-se muito tempo na tentativa de encontrar uma informação preciosa, mas por não ter um direcionamento correto de onde está essa informação, não se obtém os resultados esperados.

Nesse contexto, o objetivo deste trabalho é demonstrar a aplicação do processo de Data Mining utilizando o algoritmo de Árvores de Decisão e o algoritmo de Clustering, com a finalidade de obter informações que possam otimizar o resultado das empresas.

Desta forma, apresenta-se um caso real da aplicação do processo de Data Mining utilizando o algoritmo de Árvore de Decisão e o algoritmo de Clustering em uma empresa, mostrando as atividades que foram realizadas em cada etapa do processamento, além dos resultados que foram gerados para auxiliar no processo de tomada de decisão da empresa.

1.3 METODOLOGIA

Primeiramente, foram realizadas diversas pesquisas, consultas e leituras sobre o material encontrado.

As pesquisas foram realizadas sobre os temas:

- Data Mining
- Processo de Data Mining
- Algoritmos de Data Mining
- Análise de resultados dos algoritmos de Data Mining
- Ferramentas que continham algoritmos de Data Mining

Em seguida, as idéias foram organizadas e o desenvolvimento do trabalho foi iniciado na seqüência.

Durante o desenvolvimento do trabalho o autor propôs a um dos clientes da empresa em que trabalha, realizar a aplicação do processo de Data Mining utilizando o algoritmo de Árvore de Decisão, gerando assim informações que poderiam ser utilizadas pela empresa conforme o objetivo definido pela própria empresa.

1.4 ESTRUTURA DO TRABALHO

O Capítulo 2 apresenta o referencial teórico do trabalho, onde é feita a introdução ao processo de Data Mining e detalhada cada uma de suas etapas.

O Capítulo 3 apresenta o referencial teórico do trabalho sobre o algoritmo de Árvore de Decisão.

O Capítulo 4 apresenta o referencial teórico do trabalho sobre o algoritmo de Clustering.

O Capítulo 5 apresenta uma breve introdução sobre o Microsoft SQL Server 2005 e os algoritmos Microsoft Decision Trees e Microsoft Clustering. Esses algoritmos são utilizados na aplicação do processo de Data Mining em um caso real.

O Capítulo 6 apresenta as conclusões finais sobre o estudo de forma geral.

2 DATA MINING

Os constantes avanços na área da Tecnologia da Informação têm viabilizado o armazenamento de grandes e múltiplas bases de dados. Tecnologias como a Internet, sistemas gerenciadores de bancos de dados, leitores de códigos de barra, dispositivos de memória secundária de maior capacidade de armazenamento e de menor custo e sistemas de informação em geral são alguns exemplos de recursos que têm viabilizado a proliferação de inúmeras bases de dados de natureza comercial, administrativa, governamental e científica.

Diante desse cenário, surgem algumas questões como: "O que fazer com todos os dados armazenados?", "Como utilizar esses dados em benefício das instituições?", "Como analisar e utilizar de maneira útil todo o volume de dados disponíveis?", entre outras.

Defini-se Data Mining como o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano. (Carvalho, 2005)

As técnicas de Data Mining fornecem meios de se descobrir relações interessantes, mas para que elas sejam realmente úteis, a instituição deve se dirigir como um todo para seu uso através da aplicação do processo de Data Mining. A simples aplicação de técnicas de Data Mining de forma desorientada pode levar a empresa a resultados que não condizem com a realidade.

O processo de Data Mining está contido dentro de um processo maior denominado KDD (Knowledge Discovering in Databases) que representa uma atividade mais abrangente em relação à análise inteligente dos dados.

2.1 PROCESSO DE DATA MINING

O processo de Data Mining não começa pela análise dos dados, e sim com a definição de um problema a ser resolvido. Uma vez que o objetivo é de gerar resultados (descobertas, idéias, informações e modelos) que deve ajudar a resolver o problema de uma empresa. O processo deve iniciar-se pela definição clara do problema, sendo este traduzido em uma ou mais perguntas que possam ser abordadas pelo Data Mining. (Ballard et al., 2007)

Para que se possa aplicar o processo de Data Mining, é necessário que se defina o problema que se está querendo resolver, ou seja, a definição clara do objetivo a ser alcançado. Todas as etapas que são descritas abaixo sofrem direto impacto, dependendo do objetivo proposto.

Seguem abaixo alguns exemplos dos objetivos que podem ser propostos ao processo de Data Mining:

- Qual o perfil dos clientes para cada uma das suas linhas de produtos?
- Como obter uma melhor lucratividade através da composição de cestas de produtos?
- Como uma empresa pode prever o risco de crédito dos clientes?
- Quais as tendências atuais do mercado, e qual caminho ele está seguindo?
- Como aumentar as vendas da loja virtual através da análise dos hábitos de consumo dos usuários?
- Como uma empresa pode determinar o sucesso da campanha de marketing?
- Como uma empresa pode impedir a entrada de dados má qualidade do sistema?

O processo de Data Mining é caracterizado como um processo composto de várias etapas operacionais, conforme apresentado na **figura 1**. (Goldschmidt; Passos, 2005)

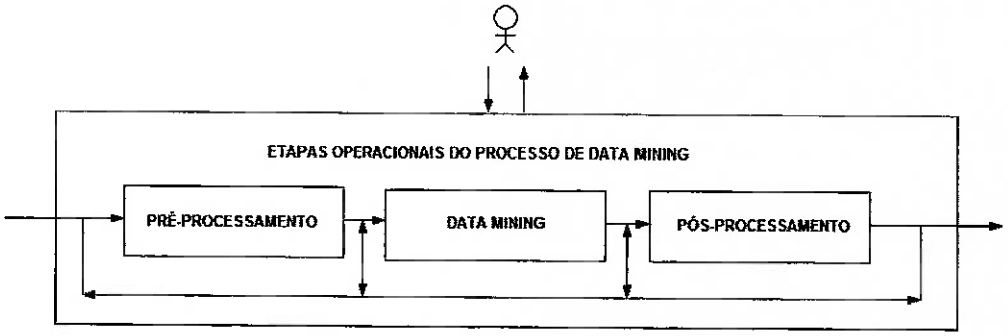

Figura 1. Etapas operacionais do processo de DATA MINING.

- **Pré-Processamento:** compreende as funções relacionadas à captação, à organização e ao tratamento dos dados. Tem como objetivo a preparação dos dados para aplicação dos algoritmos da próxima etapa.

- **DATA MINING:** Durante esta etapa é realizada a busca efetiva por conhecimento no contexto da aplicação.

- **Pós-processamento:** Compreende o tratamento do conhecimento obtido na etapa anterior e tem como objetivo viabilizar a avaliação da utilidade do conhecimento descoberto.

2.1.1 PRÉ-PROCESSAMENTO

Dados incompletos, errados e inconsistentes são comuns nos grandes bancos de dados.

Dados incompletos podem ocorrer por várias razões. Informações importantes para a aplicação podem não estar disponíveis, como as informações dos clientes em transações de vendas. Outros dados podem não estar disponíveis, simplesmente porque não foram considerados importantes na entrada de dados. Os dados também podem não ter sido gravados pelo mau funcionamento da máquina. (Han; Kamber, 2006)

Dados errados pode ser resultado de diversos fatores humanos ou da máquina, através da entrada de dados em campos com o formato errado, erros de transmissão de dados ou mesmo de limitações da tecnologia como o tamanho do espaço de armazenamento. (Han; Kamber, 2006)

O Pré-processamento compreende as funções relacionadas à captação, à organização, ao tratamento e a pré-preparação dos dados para a etapa de DATA MINING. Essa etapa possui fundamental relevância no processo de descoberta do conhecimento. Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de DATA MINING a serem utilizados.

O pré-processamento é caracterizado como um processo composto de várias etapas operacionais, conforme apresentado na ***figura 2***.

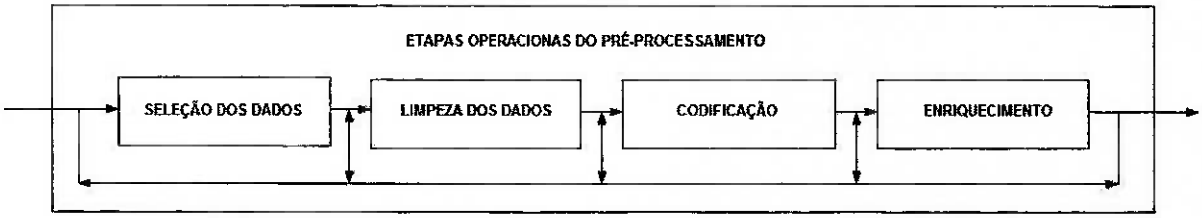

Figura 2. Etapas operacionais do pré-processamento.

2.1.1.1 SELEÇÃO DE DADOS

Essa função compreende, em essência, a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de Data Mining.

Em geral, os dados encontram-se organizados em bases de dados transacionais que sofrem constantes atualizações ao longo do tempo. Assim sendo, recomenda-se que seja feita uma cópia dos dados a fim de que o processo de Data Mining não interfira nas rotinas operacionais eventualmente relacionadas à base de dados. (Goldschmidt; Passos, 2005)

A seleção de dados visa obter uma redução representativa no volume de dados que serão submetidos aos algoritmos de Data Mining, proporcionando um melhor desempenho e resultados analíticos consistentes.

Nos casos em que já exista uma estrutura de Data Warehouse, a mesma deve ser utilizada no processo de Data Mining.

O Data Warehouse proporciona uma sólida e concisa integração dos dados da empresa, para realização de análises gerenciais estratégicas de seus principais processos de negócio. Ele se preocupa em integrar e consolidar as informações de fontes internas, na maioria das vezes heterogêneas, e fontes externas, sumarizando, filtrando e limpando esses dados, preparando-os para análise à decisão. (Machado, 2004)

O Data Warehouse possui um conjunto de características, conforme apresentado a seguir, que o distingue de outros ambientes de sistemas convencionais. (Machado, 2004)

- Extração de dados de fontes heterogêneas (internas ou externas);
- Transformação e integração dos dados antes de sua carga final;
- Normalmente requer máquina e suporte próprio;
- Permite a abstração dos dados em diferentes níveis.
- Serve de base para ferramentas voltadas para o acesso com diferentes níveis de apresentação;
- Dados somente são inseridos, não existindo atualizações.

Independente do fato da empresa possuir ou não um Data Warehouse, a maioria dos algoritmos de Data Mining pressupõe que os dados estejam convenientemente organizados.

Seguem abaixo técnicas que podem ser utilizadas para seleção dos dados para o processo de Data Mining. (Goldschmidt; Passos, 2005),

a) Junção de Dados

Junção Direta – Todos os atributos e registros da base de dados transacional são incluídos em uma única tabela.

Junção Orientada – Uma nova tabela é criada a partir da seleção dos atributos e registros sobre os quais se tenha uma visão clara quanto ao potencial para influir no processo de Data Mining.

b) Redução de Dados Horizontal

Segmentação do Banco de Dados – Deve-se escolher um ou mais atributos para nortear o processo de segmentação. Como por exemplo, a seleção de clientes que moram em residência própria, onde o tipo de residência passa a ser o atributo para nortear o processo de segmentação.

c) Amostragem Aleatória

A operação de Amostragem Aleatória consiste em sortear da base de dados um número pré-estabelecido de registros de forma que o conjunto resultante possua menos registros do que o conjunto original.

d) Agregação de Informações

Esta operação consiste em reunir os dados de forma a reduzir o conjunto de dados original. Na agregação de Informações, dados em um nível maior de detalhamento são consolidados em novas informações com menor detalhe. Como por exemplo, somar o valor de todas as compras de cada cliente.

e) Redução de dados Vertical

Consiste na seleção de atributos que devem ser considerados no processo de Data Mining.

Pode ser implementada pela eliminação ou pela substituição dos atributos de um conjunto de dados, com o objetivo de encontrar um conjunto mínimo de atributos de tal forma que a informação original seja preservada.

f) Redução de Valores

Esta operação consiste em reduzir o número de valores distintos em determinados atributos.

Redução de Valores Discretos - Os valores Discretos possuem um número finito de valores distintos criando possíveis generalizações para os valores de cada atributo, por exemplo, definindo um valor específico para um conjunto de produtos com as mesmas características.

Redução de Valores Contínuos - Os valores Contínuos possuem uma relação de ordenação entre seus valores e geralmente são representados por valores numéricos.

A redução dos valores contínuos consiste no agrupamento dos valores baseado em regras, que podem ser a Mediana, a Média, os limites máximos e mínimos e o arredondamento dos valores.

2.1.1.2 LIMPEZA DOS DADOS

A etapa de limpeza dos dados envolve uma verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores não pertencentes ao domínio. (Goldschmidt; Passos, 2005)

A execução dessa etapa tem como objetivo, corrigir a base de dados, eliminando consultas desnecessárias que poderiam ser executadas pelos algoritmos de Data Mining, afetando assim o seu desempenho.

Seguem abaixo técnicas utilizadas para limpeza dos dados:

a) Limpeza de Informações ausentes

Essa função compreende o tratamento de valores ausentes em um conjunto de dados.

Essa função pode ser aplicada das seguintes formas:

- Exclusão dos registros que possuam informações ausentes;
- Preenchimento manual desses valores;
- Preenchimento com valores globais constantes;
- Preenchimento utilizando técnicas estatísticas;
- Preenchimento utilizando algoritmos de Data Mining.

b) Limpeza de valores Inconsistentes

Essa função compreende o tratamento de valores inconsistentes em um conjunto de dados. As mesmas funções utilizadas para limpeza de informações ausentes podem ser aplicadas para limpeza de valores inconsistentes.

2.1.1.3 CODIFICAÇÃO

A codificação é a operação responsável pela forma como os dados serão representados no processo de Data Mining. Essa codificação tem como objetivo atender a necessidades específicas dos algoritmos de Data Mining, além de obter um melhor desempenho na sua aplicação.

A codificação pode ser aplicada utilizando as seguintes técnicas:

a) Mapeamento Direto

Consiste na substituição dos valores existentes no conjunto de dados pelos valores desejados, conforme o exemplo abaixo.

Sexo:

0 –> M

1 –> F

b) Mapeamento em Intervalos

Consiste em dividir o domínio de uma variável numérica em intervalos.

Como exemplo, considere a quantidade de produtos comprados por um cliente com os seguintes valores, já organizados em ordem crescente: 1000, 1400, 1500, 1700, 2500, 3000, 3700, 4300, 4500 e 5000.

A ***tabela 1*** apresenta um mapeamento em intervalos definidos pelo usuário:

Intervalo	Freqüência (número de valores no intervalo)
1000 a 1600	3
1600 a 4400	5
4400 a 5400	2

Tabela 1. Mapeamento em intervalos definidos pelo usuário.

A ***tabela 2*** apresenta um mapeamento em intervalos com comprimentos iguais:

Intervalo	Freqüência (número de valores no intervalo)
1000 a 2000	4
2000 a 3000	1
3000 a 4000	2
4000 a 5000	3

Tabela 2. Mapeamento em intervalos com comprimentos iguais.

2.1.1.4 ENRIQUECIMENTO

A etapa de enriquecimento consiste em conseguir agregar mais informações ao conjunto de dados existentes para que estes forneçam mais elementos para o processo de Data Mining. (Han; Kamber, 2006)

O enriquecimento do conjunto de dados geralmente é realizado através da consulta a base de dados externas, com o objetivo de complementar as informações já existentes, como informações geográficas dos clientes ou o valor médio do imóvel de clientes de uma determinada região.

2.1.2 DATA MINING

A execução da etapa de Data Mining compreende a aplicação de algoritmos sobre os dados selecionados, procurando abstrair o conhecimento. Estes algoritmos são fundamentados em técnicas que procuram, segundo determinados paradigmas, explorar os dados de forma a produzir modelos de conhecimento baseados no objetivo do processo de Data Mining. (Goldschmidt; Passos, 2005)

Baseados nos diferentes objetivos propostos ao processo de Data Mining, seguem abaixo algumas das tarefas que podem ser realizadas pelos algoritmos de Data mining:

2.1.2.1 CLASSIFICAÇÃO

A classificação é o processo que verifica as propriedades comuns entre um conjunto de objetos em um banco de dados, e então, os classifica em diferentes classes, de acordo com um modelo de classificação. (Chen; Han, 1996)

O objetivo da classificação é inicialmente analisar os dados e em seguida

desenvolver uma descrição precisa ou modelo, para cada classe utilizando os recursos disponíveis nos dados. (Chen; Han, 1996)

A classificação tem como característica a previsão de uma ou mais variáveis discretas, com base nos outros atributos do conjunto de dados. Variáveis discretas são aquelas que possuem um conjunto finito de valores.

Aplicações de classificação incluem diagnóstico médico, a previsão de desempenho, o marketing seletivo, dentre outros.

2.1.2.2 REGRESSÃO

A Regressão é utilizada para previsão de variáveis contínuas, sendo estas as variáveis que possuem um conjunto infinito de valores, geralmente representados por valores numéricos. (Larson, 2006)

Para previsão de variáveis contínuas, a regressão analisa as tendências das variáveis, e a sua freqüência de repetição ao longo do tempo.

A regressão também analisa a relação entre a previsão de um variável continua, e outras variáveis continuas disponíveis nos dados.

Aplicações de regressão incluem previsões como lucros ou prejuízos, ou a previsão de vendas tendo como base os preços.

2.1.2.3 SEGMENTAÇÃO

A segmentação é usada para identificar grupos naturais baseados em um conjunto de atributos. (Tang; MacLennan, 2005)

A segmentação envolve um conjunto de padrões em um grupo, de acordo com alguma medida de similaridade. Padrões pertencentes a um determinado grupo

devem ser mais similares entre si do que em relação a padrões pertencentes a outros grupos. (Lui; Lui, 2005)

O objetivo dessa tarefa é maximizar a similaridade entre os grupos, dividindo os dados em grupos, através da identificação dos itens que têm propriedades semelhantes. (Lui; Lui, 2005)

Uma aplicação da segmentação é a identificação das melhores cestas de produtos para se oferecer a possíveis grupos de clientes.

2.1.2.4 ASSOCIAÇÃO

Esta tarefa consiste em encontrar um conjunto de itens que ocorram simultaneamente e de forma freqüente em um conjunto de dados. (Cheung, et al., 1996)

Muitas vezes a segmentação e a associação são confundidas, a diferença é que no caso da segmentação, o próprio algoritmo cria o agrupamento, com base nas informações que ele determina como características distintivas. Já na associação, se parte de algum tipo de agrupamento existente e os algoritmos de associação fazem determinações sobre elementos prováveis de estarem no grupo pré-existente, ao invés da similaridade que os atributos tenham em comum. (Larson, 2006)

O objetivo desta tarefa é encontrar correlações entre os diferentes atributos em um conjunto de dados. (Cheung, et al., 1996)

Uma aplicação da associação é a identificação dos produtos a serem oferecidos a clientes de determinada região demográfica.

2.1.3 PÓS-PROCESSAMENTO

Essa etapa envolve a visualização, a análise e a interpretação do modelo de conhecimento gerado pela etapa de Data Mining.

Em geral, é nessa etapa que o especialista em Data Mining e o especialista no domínio da aplicação avaliam os resultados obtidos e definem novas alternativas de investigação dos dados.

Os modelos de conhecimento podem ser representados de diversas formas. Árvores, regras, gráficos em duas ou três dimensões, planilhas e tabelas são muito úteis na representação do conhecimento.

Técnicas de visualização dos dados estimulam a percepção e a inteligência humana, aumentando a capacidade de entendimento e associação de novos padrões.

3 MÉTODO DE CLASSIFICAÇÃO POR ÁRVORES DE DECISÃO

O método de classificação em de árvores de decisão é uma das mais populares técnicas de Data Mining, devido à rápida construção e ao fácil entendimento. A construção da árvore de decisão não requer conhecimento detalhamento do domínio ou dos parâmetros de configuração, e, por conseguinte, é adequada para a descoberta exploratória conhecimento.

O método de classificação por árvores de decisão é baseado na construção de árvores de decisão a partir de uma base de dados. Em geral a construção de uma árvore de decisão é realizada segundo uma abordagem recursiva de particionamento da base de dados. (Han; Kamber, 2006)

Uma árvore de decisão consiste em uma estrutura em forma de árvore que auxilia de forma visual a classificação e predição das amostras desconhecidas.

Na ***figura 3*** pode-se visualizar o perfil dos clientes que compram produtos eletrônicos, onde cada nó da árvore representa um atributo que foi particionado em classes. No caso do atributo IDADE foram criadas três classes, clientes com idade inferior a 20 anos, clientes com idade entre 20 e 30 anos e clientes com idade superior a 30 anos.

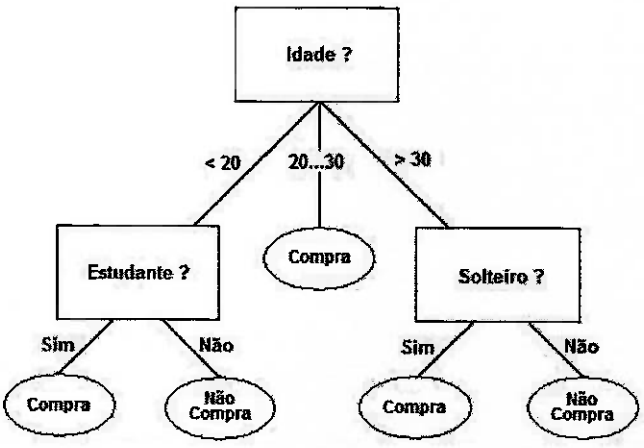

Figura 3. Árvore de Decisão demonstrando o perfil dos clientes que compram produtos eletrônicos.

Uma árvore de decisão é um modelo de conhecimento em que cada nó interno da árvore representa uma decisão sobre um atributo que determina como os dados estão particionados pelos seus nós filhos. Inicialmente, a raiz da árvore contém toda a base de dados com valores que representam todas as classes.

Um predicado, denominado ponto de separação, é escolhido como sendo a condição que melhor separa ou discrimina as classes. Tal predicado envolve exatamente um dos atributos do problema e divide a base de dados em dois ou mais conjuntos, que são associados cada um a um nó filho. Cada novo nó abrange, portanto, uma partição da base de dados que é recursivamente separada até que o conjunto associado a cada nó folha consista inteiramente ou predominantemente de registros de uma mesma classe.

Para construção da Árvore de Decisão o conjunto de dados aplicados ao algoritmo é separado em duas ou mais partições usando restrições sobre o conjunto de valores de cada atributo. O processo é repetido recursivamente até que todos ou a maioria dos dados em cada partição pertençam a uma classe. A árvore gerada abrange todo o conjunto de dados e é construída em largura. Assim, todos os nós em uma determinada altura da árvore devem ser processados antes do nível subseqüente. (Goldschmidt; Passos, 2005)

3.1 ALGORITMO DE ÁRVORES DE DECISÃO

Durante a década de 80 J. Ross Quinlan, pesquisador no assunto desenvolveu um algoritmo de árvores de decisão conhecido como ID3 (iterativos Dichotomiser). Este trabalho foi baseado em trabalhos anteriores sobre conceito de sistemas de aprendizagem, descrito por EB Hunt, J. Marin, e PT Stone. Quinlan posteriormente apresentou o C4.5 (um sucessor de ID3), que se tornou um ponto de referência, servindo como referência os novos algoritmos que estão sendo criados. (Han; Kamber, 2006)

Em 1984, um grupo de estatísticos (L. Breiman, J. Friedman, R. Olshen, e C. Stone) publicou o livro Classificação e Árvore de Decisão (CART), que descreveu a geração

de árvores de decisão binária. ID3 e CART foram criados independentemente um do outro e ao mesmo tempo, ambos seguem uma abordagem semelhante para a construção da árvore de decisão e definição dos pontos de separação. Estes dois algoritmos serviram de referência para muitos trabalhos sobre Árvores de Decisão. (Han; Kamber, 2006)

Nos algoritmos ID3, C4.5 e CART a árvore de decisão é construída de cima para baixo. A maior parte dos algoritmos de árvore de decisão também segue tal abordagem descendente, que começa com um conjunto de tuplas e sua associação às classes rótulos. As classes rótulos representam todos os valores possíveis de um atributo. (Han; Kamber, 2006)

O conjunto de dados é recursivamente particionado em subconjuntos menores conforme a árvore está sendo construída.

A seguir é mostrado um algoritmo genérico para a montagem de uma árvore de decisão. (Han; Kamber, 2006)

Algoritmo: Generate_decision_tree.

Objetivo: Gerar uma árvore de decisão pela classificação do conjunto de dados D.

Entrada:

Conjunto de Dados, D, é um conjunto de tuplas e sua associação as classe rótulos;
Attribute_list, é o conjunto de atributos candidatos;
Attributes_selectin_method, um procedimento para determinar os melhores pontos de divisão nas tuplas em classes individuais. Este critério é composto por um splitting_attribute.

Saída:

Árvore de Decisão.

Método:

(1) Criar um nó N;
(2) **Se** tuplas em D são todos da mesma classe, C **então**
(3) Retorna N como um nó folha nomeado com a classe C;
(4) **Se** attribute_list vazio **então**
(5) Retorna N como um nó folha nomeado com a maioria na classe D; / / votação por maioria
(6) Aplicar **Attribute-selection_method (D, attribute_list)** para encontrar o "melhor" splitting_criterion;
(7) Nomear o nó N com splitting_criterion;
(8) Se splitting_attribute é contínuo e a árvore permite multi-ponto então / / não se limita às árvores binárias
(9) Attribute_list ← attribute_list - splitting_attribute; / / remover splitting_attribute
(10) **Para cada** resultado j do splitting_criterion
/ / separação das tuplas e crescimento da árvore para cada partição
(11) Define Dj com o conjunto de dados(tuplas) em D satisfazendo resultado j; / / uma partição
(12) **Se** Dj vazia **então**
(13) Anexar um nó folha nomeado com a maioria na classe D ao nó N;
(14) **senão** anexar o nó que foi retornado pelo **Generate_decision_tree (Dj, attribute_list)** ao nó N
End Se
End Para cada
(15) Retorna N;

O algoritmo é chamado com os três parâmetros de entrada: D, attribute_list, e Attribute_selection_method. O conjunto de dados D, é o conjunto completo de tuplas e suas classes rótulos. O parâmetro attribute_list é uma lista de atributos que descreve as tuplas e Attribute_selection_method especifica um processo de seleção do atributo que melhor caracteriza as classes.

Este processo emprega um atributo de seleção de medida, como o ganho da informação ou o Índice de Gini. Definir se a árvore é estritamente binária é tarefa realizada pelo atributo de seleção de medida. Alguns atributos de seleção de medidas, como o Índice de Gini, realizam a criação de árvores binárias. Outros, como o ganho da informação, permitem a criação de árvores com multi-pontos de separação.

- A árvore é iniciada com um único nó, N, que representa a formação de tuplas em D (passo 1).

- Se os valores das tuplas em D forem definidos em uma mesma classe, será criado então um nó folha que é tratado pelos próximos passos. (passos 2 e 3). Os passos 4 e 5 se encerram o algoritmo.

- Caso contrário, o algoritmo Attribute_selection_method irá determinar o melhor critério de divisão do nó. O critério de divisão define qual o melhor atributo para ser testado no nó N, determinando a melhor forma de separação ou particionamento das tuplas do conjunto de dados em classes individuais (passo 6). O critério de divisão também define que ramos devem ser criados a partir do nó N. Ou seja, o critério de divisão define qual o atributo que será utilizado para criação de novos nós e também o valor do ponto de separação. O ideal nesse caso é que todas as tuplas do próximo nível pertençam à mesma classe, podendo-se afirmar que os resultados são os mais "puros" quanto possível.

- O nó N é nomeado com o critério de divisão, que serve como um teste ao nó (passo 7). Um ramo é criado a partir de nó N para cada um dos resultados do critério de divisão. As tuplas em D são particionadas em conformidade (passos 10 a 11). Há dois cenários possíveis.

1. Valores Discretos: Neste caso, os resultados do teste no nó N correspondem diretamente aos valores conhecidos na tupla A. Um ramo é criado para cada valor conhecido de A, conforme ***figura 4***. Quando todos os valores de A são representados, na próxima partição não será necessário considerar a tupla A em novos particionamentos, sendo esta removida da attribute_list (passos 8 e 9).

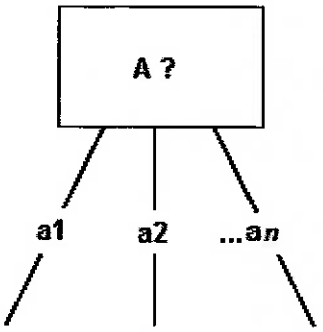

Figura 4. Particionamento de Valores Discretos.

2. Valores Contínuos: Neste caso, o teste no nó N possui dois possíveis resultados, que correspondem às condições A ≤ split_point e A > split_point, respectivamente. O ponto de divisão, muitas vezes é definido como o ponto médio de dois valores adjacentes, por conseguinte, não sendo este um valor pré-existente em A. Dois ramos são criados a partir do nó N, conforme a ***figura 5***.

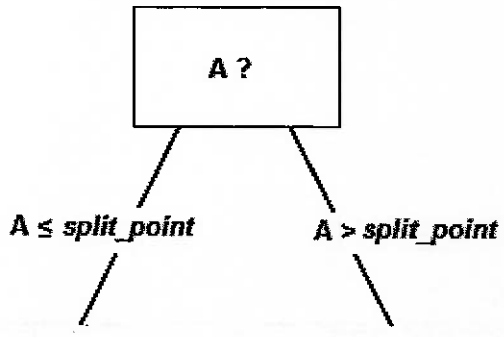

Figura 5. Particionamento de Valores Contínuos.

- O algoritmo utiliza o mesmo processo recursivamente para formar á arvore de decisão para as tuplas em cada partição resultante Dj, de D (passo 14).

- O particionamento recursivo só é finalizado quando qualquer uma das seguintes condições for verdadeira:

1. Todas as tuplas na partição D (representada pelo nó N) pertencerem à mesma classe (passos 2 e 3), ou

2. Não existirem atributos remanescentes em que as tuplas possam ser particionadas (passo 4). Neste caso, uma votação por maioria é empregada (passo 5). Dessa forma, esse nó é convertido em um nó folha e nomeado com o atributo mais comum da classe D.

3. Não existirem tuplas para um determinado ramo, ou seja, uma partição Dj está vazia (passo 12). Neste caso, é criado um nó folha com a maioria na classe D (passo 13).

- Retorna a Arvore de Decisão criada. (passo 15).

A complexidade computacional da execução do algoritmo no conjunto de dados D é O (n x |D| x log (| D |)), onde n é o número de atributos que descreve as tuplas em D e |D|, é o número de tuplas em D. Isso significa que o custo computacional de uma árvore de decisão cresce cada vez mais n x |D| x log (|D|) com |D| tuplas. (Han; Kamber, 2006)

As diferenças nos algoritmos de árvore de decisão incluem o modo como os atributos são selecionados na criação da árvore e os mecanismos utilizados para a poda. (Han; Kamber, 2006)

3.1.1 AVALIAÇÃO DOS PONTOS DE SEPARAÇÃO

Os métodos de avaliação dos pontos de separação representam o melhor critério para separação dos dados em um determinado conjunto. Para dividir um conjunto de dados em partições menores, de acordo com os critérios dos pontos de separação, seria ideal que cada partição fosse pura, ou seja, todas as tuplas de uma partição deveriam pertencer à mesma classe.

A avaliação dos pontos de separação proporciona uma classificação para cada atributo que descreve os dados que formam as tuplas. O atributo que obtiver a melhor pontuação na avaliação dos pontos de separação é escolhido para separar determinada tupla.

Dependendo do atributo em análise, a pontuação mais alta ou a mais baixa é escolhida como a melhor, ou seja, algumas medidas tendem para maximização dos valores enquanto outras tendem para minimização dos valores.

Existem diferentes métodos para avaliação de pontos de separação, os mais utilizados são "Information Gain", "Gain Ratio", "Gini Index", "Bayseian K2" e "Bayesian Dirichlet Equivalent with Uniform prior" (BDEU). (Tang; Maclennan, 2005) (Han; Kamber, 2006)

A seleção do método de avaliação dos pontos de separação é feita baseada no cenário encontrado, podendo variar de acordo com o tipo de atributo que se estiver utilizando, discreto ou contínuo, ou se estiver restrito à criação de árvores binárias.

3.2 CONSTRUÇÃO PRÁTICA DA ÁRVORE DE DECISÃO

Com o objetivo de exemplificar a utilização do algoritmo na construção da Árvore de Decisão o seguinte cenário é proposto: Um sistema de uma escola de línguas envia para um banco, no inicio do mês, um boleto contendo a mensalidade do curso a ser paga pelos alunos. O banco então envia pelo correio a fatura para os alunos e espera os recebimentos. No final do mês, o banco retorna para o sistema da escola

quais alunos pagaram a mensalidade, quais não pagaram e quais alunos pagaram com atraso, dentre outras informações. Com o objetivo de diminuir a quantidade de alunos que pagaram a mensalidade com atraso, o método de classificação por de árvores de decisão é aplicado.

Um pré-processamento dos dados separou as informações dos alunos conforme a ***tabela 3.***

IDADE	SALARIO	SUP_COMPLETO	DEPENDENTES	ATRASOU
< = 30	Alto	Não	Não	Não
< = 30	Alto	Não	Sim	Não
31...40	Alto	Não	Não	Sim
> 40	Médio	Não	Não	Sim
> 40	Baixo	Sim	Não	Sim
> 40	Baixo	Sim	Sim	Não
31...40	Baixo	Sim	Sim	Sim
< = 30	Médio	Não	Não	Não
< = 30	Baixo	Sim	Sim	Sim
> 40	Médio	Sim	Não	Sim
< = 30	Médio	Sim	Sim	Sim
31...40	Médio	Não	Sim	Sim
31...40	Alto	Sim	Não	Sim
> 40	Médio	Não	Sim	Não

Tabela 3. Tabela de informações de atraso na mensalidade.

As colunas da ***tabela 3*** são descritas a seguir:

IDADE: Esse atributo identifica a idade do aluno.
SALARIO: Esse atributo identifica o salário do aluno.
SUP_COMPLETO: Esse atributo Indica se o aluno possui nível superior completo.
DEPENDENTES: Esse atributo indica se o aluno possui dependentes.
ATRASOU: Esta é a coluna que apresenta a classificação das amostras. A classificação indica se o aluno atrasou o pagamento, ATRASOU = SIM, ou se o aluno não atrasou o pagamento, ATRASOU = NÃO.

Inicialmente define-se o nó raiz. Como ainda não existe nenhum nó na árvore, nesse caso basta definir qual o atributo de classificação e calcular a probabilidade de cada valor do atributo.

Esse atributo é definido como o nó raiz, que no nosso caso é representado pela coluna "ATRASOU".

Pode-se notar a probabilidade de o valor estar ausente na ***figura 6*** de 5,88%, embora este valor não exista neste nó. Isto ocorre devido a uma previsão ser considerada no cálculo da probabilidade de cada valor. Normalmente, ela favorece os valores com baixa probabilidade.

Value	Cases	Probability	Histogram
Missing	0	5,88%	
Nao	5	35,29%	
Sim	9	58,82%	

Figura 6. Representação do nó raiz, o número de casos, probabilidade e o gráfico de histograma.

Após a definição do nó raiz é necessário encontrar os nós da árvore que podem ser divididos para novos nós. Novos nós serão identificados entre os demais atributos cuja probabilidade não classifique a amostra totalmente. Classificar a amostra totalmente quer dizer que o nó não deve possuir alguma classe que tenha 100% do número de casos. Caso não exista nenhum nó que pode ser dividido o algoritmo é finalizado.

No próximo passo o algoritmo seleciona, dentre os atributos disponíveis, os nós que podem ser divididos. Como se tem um único nó, o algoritmo irá analisar todos os atributos para verificar aquele que melhor classifica os dados e em seguida define os novos nós da árvore conforme exibido na ***figura 7***.

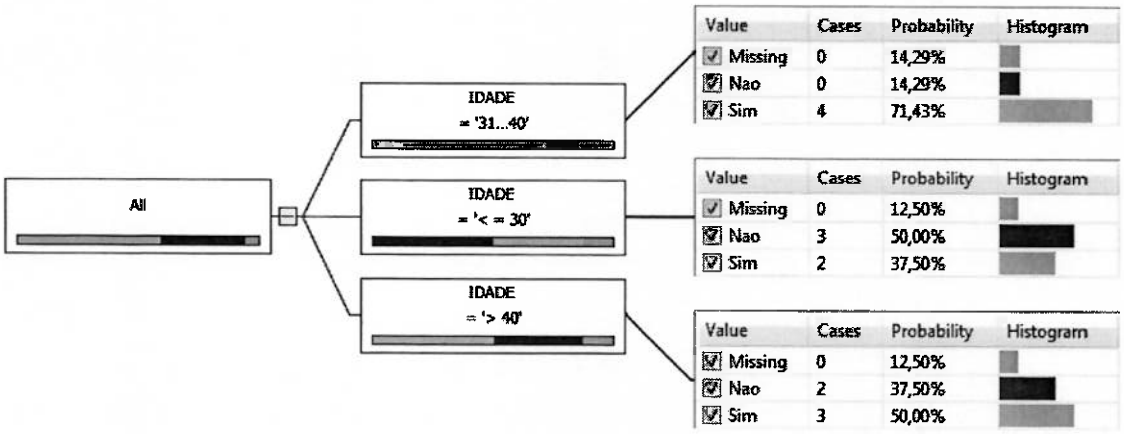

Figura 7. Demonstra os nós criados para o próximo atributo selecionado.

Após a criação desses nós o algoritmo volta para o passo de definição de nós a serem considerados na divisão. Nesse caso, dois nós que podem ser divididos, pois o nó folha gerado pela divisão do valor 31...40 do atributo IDADE não pode ser mais dividido. Em seguida o algoritmo calcula a probabilidade dos atributos SALARIO, SUP_COMPLETO e DEPENDENTES para cada um dos nós que podem ser divididos e assim sucessivamente até não restarem mais nós a serem divididos. A ***figura 8*** mostra o resultado final da aplicação do algoritmo.

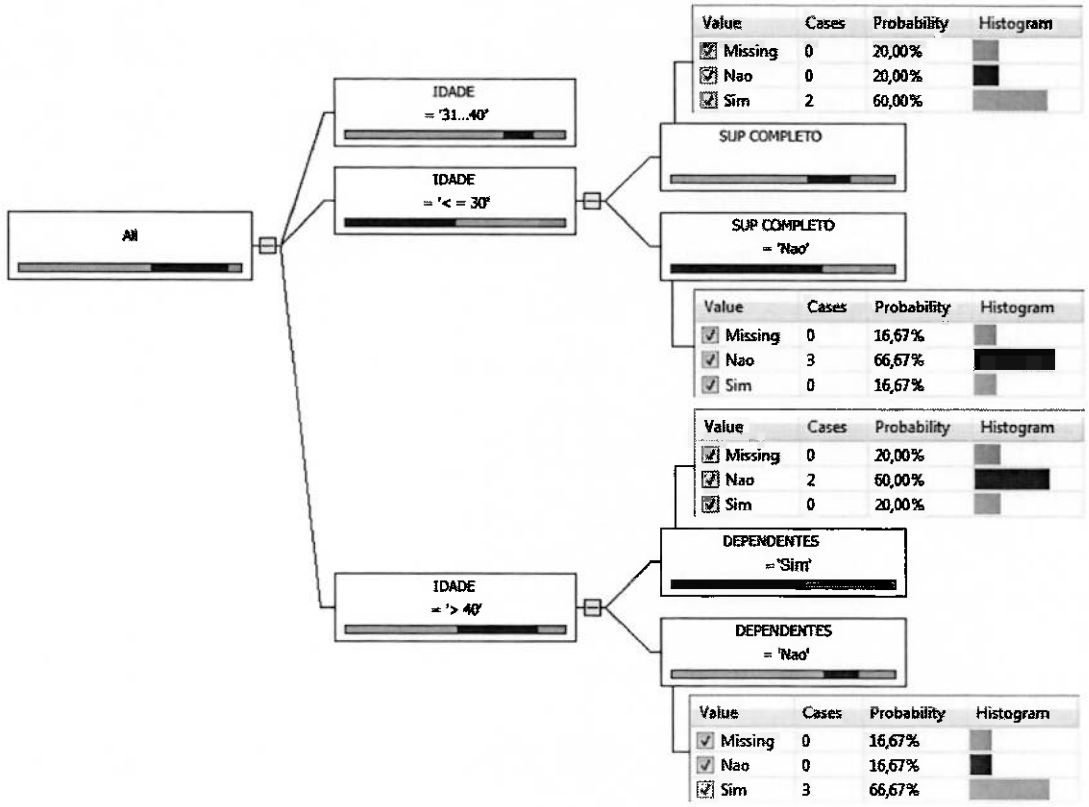

Figura 8. Resultado final da aplicação do algoritmo de Árvore de Decisão.

A árvore de decisão na ***figura 8*** mostra quatro nós folha que classificam os valores dos seus atributos na sua totalidade. O atributo SALARIO não foi utilizado, pois o algoritmo não considerou este atributo relevante para sua classificação.

4 MÉTODO DE SEGMENTAÇÃO POR CLUSTERING

A tarefa de Clustering é uma importante atividade humana. Desde a infância, aprende-se a distinguir entre os gatos e cães, ou entre animais e plantas, sempre buscando a melhor caracterização dos clusters

A tarefa de Clustering é usada para particionar os registros de uma base de dados em subconjuntos ou clusters, de tal forma que elementos em um cluster compartilhem um conjunto de propriedades comuns que o distingam dos elementos de outros clusters. O objetivo desta tarefa é maximizar similaridade intracluster e minimizar similaridade intracluster. Diferente da classificação que tem rótulos predefinidos, o Clustering é também denominado indução não supervisionada. (Goldschmidt; Passos, 2005)

O Clustering é definido como uma das tarefas básicas do Data Mining que auxilia ao usuário a realizar agrupamentos naturais de registros em um conjunto de dados.

A análise de clusters, envolve, portanto, a organização de um conjunto de padrões em clusters, de acordo com alguma medida de similaridade. Intuitivamente, padrões pertencentes a um dado cluster devem ser mais similares entre si, do que em relação a padrões pertencentes a outros clusters.

O processo de Clustering requer que o usuário determine qual o número de grupos a ser considerado. Com base neste número, os registros de dados são então separados nos grupos de forma que registros similares fiquem nos mesmos grupos e registros diferentes em grupos distintos. Uma vez, tendo esses grupos, é possível fazer uma análise dos elementos que compõe cada um deles, identificando as características comuns aos seus elementos e, desta forma, podendo criar um rótulo que represente cada grupo. (Goldschmidt; Passos, 2005)

Existem várias técnicas e métodos de Clustering. Entre os principais algoritmos de Clustering podem ser citados: K-Means, EM (Expectation Maximization), Fuzzy K-Means, K-Modes e K-Medoid. (Goldschmidt; Passos, 2005) (Lui; Lui, 2005)

Conforme (Han; Kamber, 2006), os seguintes requisitos são necessários na tarefa de Clustering no processo de Data Mining:

- **Escalabilidade**: Muitos algoritmos de Clustering funcionam bem em pequenos conjuntos de dados, contendo poucos objetos de dados, no entanto, uma grande base de dados pode conter milhões de objetos. O agrupamento sobre uma amostra de um grande conjunto de dados pode levar a resultados distorcidos.
- **Capacidade de lidar com diferentes tipos de atributos:** Muitos algoritmos são projetados para realizar o Clustering em intervalos numérico de dados. No entanto, em casos reais pode-se exigir agregação de outros tipos de dados, tais como o binário, discreto (nominal) e outros, ou a junção destes tipos de dados.
- **Descoberta de agrupamentos em diferentes formas:** Em algoritmos baseados em distância, as medidas tendem a encontrar agrupamentos esféricos com dimensões e densidades semelhantes. No entanto, um cluster pode ser de qualquer forma. É importante que o algoritmo detecte agrupamentos de forma arbitrária.
- **Requisitos mínimos para o domínio do conhecimento são necessários para definição dos parâmetros de entrada:** Os algoritmos de Clustering exigem que os usuários definam certos parâmetros de entrada, tais como o número de clusters desejado. Os agrupamentos podem ser bastante sensíveis aos parâmetros de entrada.
- **Capacidade para lidar com dados inconsistentes:** A maioria das bases de dados contém inconsistências, ou dados errados. Alguns algoritmos clustering são sensíveis a esses dados, que podem levar a um agrupamento de má qualidade.
- **Agrupamento incremental de novos dados e sensibilidade para definição da ordem de entrada dos registros:** Alguns algoritmos de Clustering não podem incorporar dados inseridos na base em estruturas de Clustering já existentes, dessa forma, é determinado um novo agrupamento. Alguns algoritmos de Clustering são sensíveis à ordem de entrada de dados,

podendo retornar agrupamentos diferentes, dependendo da forma como é definida a entrada dos dados. É importante desenvolver algoritmos incrementais de clustering e algoritmos que são sensíveis à ordem de entrada.

- **Critérios para definição dos agrupamentos:** Em aplicações reais existe a necessidade de executar a tarefa de Clustering utilizando-se vários tipos de critérios para definição dos agrupamentos
- **Capacidade de interpretação dos resultados e usabilidade:** O resultado do agrupamento deve ser interpretável, compreensível e utilizável. É importante se ter recursos para seleção dos dados e definição dos métodos de agrupamento.

A tarefa de Clustering pode ser utilizada em diferentes aplicações. Como por exemplo, a definição de agrupamentos de clientes, pode ajudar a descobrir grupos distintos de clientes, baseado em padrões de compra, permitindo a uma empresa realizar campanhas de marketing especifica para cada agrupamento.

Afim de demonstrar o funcionamento do método de Clustering, e por ser um dos mais populares algoritmos de Clustering, o algoritmo K-Means é descrito no item 4.1.

4.1 ALGORITMO DE CLUSTERING K-MEANS

O algoritmo K-Means toma um parâmetro de entrada, k, e divide um conjunto de n objetos em k clusters, tal que a similaridade em um cluster é medida em respeito ao valor médio dos objetos neste cluster (centro de gravidade do cluster).

A execução do algoritmo K-Means consiste em, primeiro, selecionar aleatoriamente k objetos, que inicialmente representam, cada um, a média de um cluster. Para cada um dos objetos remanescentes, é feita a atribuição ao cluster ao qual o objeto é mais similar, baseado na distância entre o objeto e a média do cluster. A partir de então, o algoritmo computa as novas médias para cada cluster. Este processo se repete até que uma condição de parada seja atingida. Segue abaixo as principais etapas realizadas pelo algoritmo. (Goldschmidt; Passos, 2005)

- O algoritmo tem início, definindo-se, de forma randomica, k pontos de dados (dados numéricos) como sendo os centróides (elementos centrais) dos clusters, conforme ***figura 9***. O valor de k é especificado inicialmente pelo usuário.

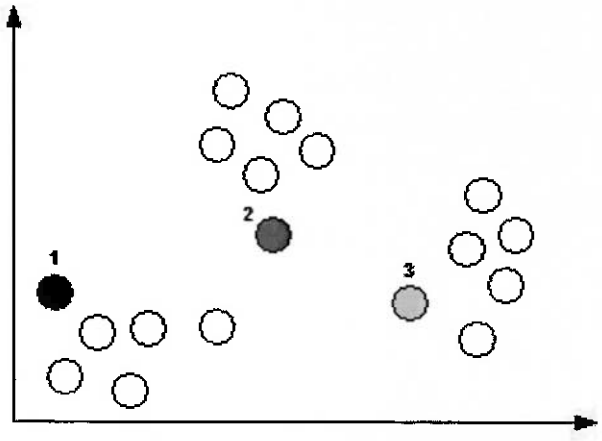

Figura 9. Inicialização das médias.

- Em seguida, conforme a ***figura 10***, cada ponto (ou registro da base de dados) é atribuído ao cluster cuja distância deste ponto em relação ao centróide de cada cluster é a menor dentre todas as distâncias calculadas.

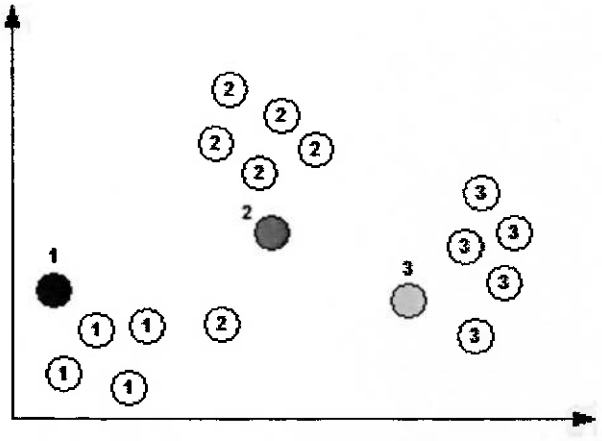

Figura 10. Atribuição dos rótulos aos objetos.

- Um novo centróide para cada cluster é computado pela média dos pontos do cluster, caracterizando a configuração dos clusters para iteração seguinte, conforme a ***figura 11***.

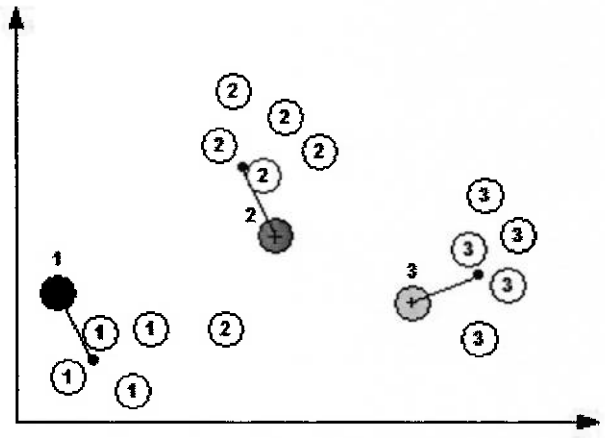

Figura 11. Atualização das médias.

- Após a nova atribuição dos rótulos uma nova iteração é realizada, conforme ***figura 12***. O processo termina quando os centróides dos clusters param de se modificar, ou após um número limitado de iterações que tinha sido especificado pelo usuário.

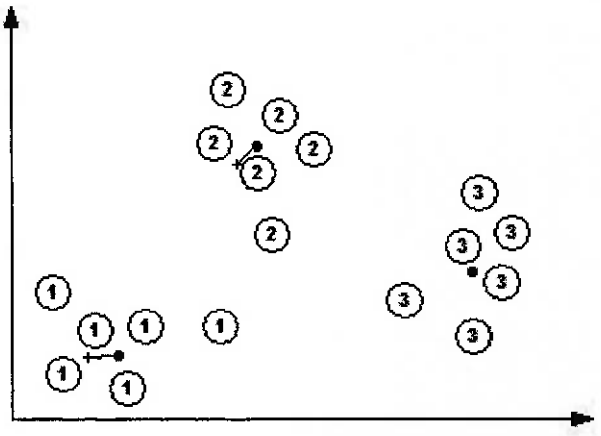

Figura 12. Nova atribuição dos rótulos e atualização das médias.

O processo de execução do algoritmo K-Means encontra-se resumido na ***figura 13***. (Goldschmidt; Passos, 2005)

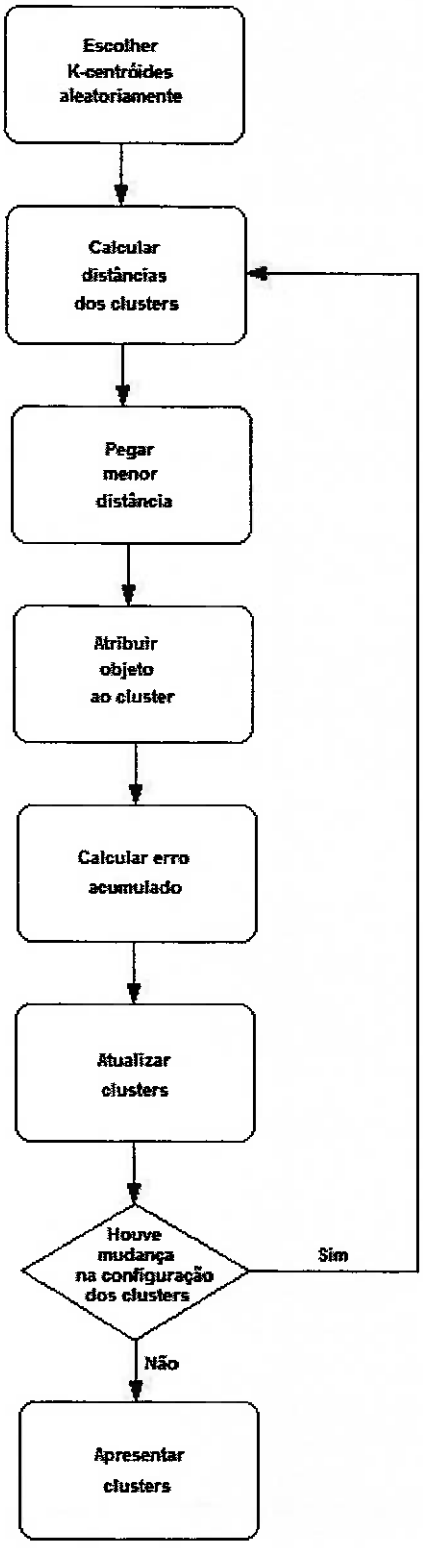

Figura 13. Algoritmo K-Means

5 APLICAÇÃO E ANÁLISE

A fim de realizar a aplicação do processo de Data Mining utilizando o método de classificação de Árvores de Decisão e de Segmentação por Clustering, o autor contatou um dos clientes da empresa em que trabalha, propondo realizar este trabalho em conjunto, com o objetivo de obter resultados reais e que podem ser utilizados pela empresa conforme seu interesse.

Realizou-se uma reunião onde foi explicado todo o processo de Data Mining, o objetivo do trabalho e quais resultados poderiam ser obtidos.

A empresa em questão não permitiu a divulgação dos seus dados reais, sendo assim, foi utilizado o nome "Empresa ABC" sempre que for necessário citar a empresa. Também serão utilizados nomes fictícios para caracterizar os dados da empresa.

É utilizado o processo de Data Mining descrito no Capitulo 2.

5.1 MICROSOFT SQL SERVER 2005

Para a implementação dos algoritmos descritos nos capítulos 3 e 4, foi utilizado o SQL Server 2005, que se constitui em uma plataforma abrangente de banco de dados que fornece recursos de gerenciamento de dados, além de ferramentas de BI (Business Intelligence) integradas. O mecanismo de banco de dados do SQL Server 2005 permite o armazenamento de dados, permitindo a criação e o gerenciamento de aplicativos de dados altamente disponíveis e eficientes para o uso nos negócios.

O mecanismo de dados do SQL Server 2005 é a parte central dessa solução de gerenciamento de dados empresariais. Além disso, o SQL Server 2005 possui recursos de análise, geração de relatórios, integração e notificação. Isso permite a criação e implantação de soluções de BI de baixo custo, e permitindo a distribuição dos dados.

O SQL Server 2005 possui total integração com o Microsoft Visual Studio, o Microsoft Office System, além de possuir um conjunto de ferramentas de desenvolvimento, incluindo o Business Intelligence Development Studio. O objetivo é criar um ambiente que proporciona a criação de soluções e rápida análise dos resultados. (SQL Server 2005 Product Guide, 2005)

O SQL Server 2005 inclui as seguintes ferramentas. (SQL Server 2005 Product Guide, 2005)

- **Banco de dados relacional** - Mecanismo de banco de dados relacional com suporte a dados estruturados e não estruturados (XML).
- **Serviços de Replicação** – Permitem a replicação de dados para aplicativos de processamento de dados distribuídos ou móveis.
- **Serviços de Notificação** - Recursos que permitem a notificação para o desenvolvimento e a implantação de aplicativos escalonáveis que podem fornecer atualizações de informações personalizadas para uma grande variedade de dispositivos conectados e móveis.
- **Serviços de Integração** - Recursos de ETL (extração, transformação e carregamento) de dados para data warehouses e a integração de dados por toda a empresa.
- **Serviços de Análise** - Recursos de OLAP (processamento analítico online) para uma análise rápida de conjuntos de dados extensos e complexos usando o armazenamento multidimensional. Possui um conjunto de algoritmos de Data Mining.
- **Serviços de Relatório** - Uma solução abrangente de criação, gerenciamento e fornecimento de relatórios tradicionais (em papel) e interativos (baseados na Web).
- **Ferramentas de Gerenciamento** - O SQL Server fornece um ambiente integrado para acesso, configuração, gerenciamento e administração de todos os componentes do SQL Server.
- **Ferramentas de Desenvolvimento** - Oferece integração com o Microsoft Visual Studio® 2005. Isto fornece a equipes de desenvolvimento que estão construindo aplicativos orientados a dados

uma maior integração com a plataforma, permitindo desenvolvimentos mais produtivos e colaboráveis da soluções relevantes.

5.1.1 SQL SERVER 2005 DATA MINING

O SQL Server 2005 Data Mining é uma das tecnologias de BI disponíveis no SQL Server 2005 que ajuda a desenvolver modelos analíticos complexos e permite a integração desses modelos a aplicativos comerciais.

A criação dessa plataforma tem como objetivo maximizar a utilização de aplicações de data mining em empresas que não consideravam a utilização desse tipo de solução devido à complexidade na realização desse tipo de projeto, permitindo a fácil utilização dos algoritmos de Data Mining.

Seguem os algoritmos de Data Mining disponíveis no SQL Server 2005. (Microsoft Data Mining Algorithms, 2007)

- Microsoft Decision Trees
- Microsoft Clustering
- Microsoft Naive Bayes
- Microsoft Association
- Microsoft Sequence Clustering
- Microsoft Time Series
- Microsoft Neural Network
- Microsoft Logistic Regression
- Microsoft Linear Regression

Para aplicação e análise propostas no Capitulo 5 desta monografia o próximo tópico aborda as características do Microsoft Decison Trees Algorithm.

5.1.2 ALGORITMO MICROSOFT DECISION TREES

O Microsoft Decision Trees é um algoritmo hibrido desenvolvido através de pesquisas realizadas pela Microsoft. Além da tarefa básica de classificação que todo o algoritmo de árvore de decisão realiza, o Microsoft Decision Trees pode ser usado também para tarefas de regressão e associação. (Tang; MacLennan, 2005)

O motivo pelo qual o algoritmo é chamado "Decision Trees" e não "Decision Tree", é que através de parâmetros de configuração o resultado da árvore de decisão pode ser diferente quanto à separação dos nós, podendo um único modelo conter múltiplas árvores. Estas árvores podem ser conectadas através da rede de dependências que a ferramenta disponibiliza. (Tang; MacLennan, 2005)

Para que o algoritmo seja executado, é necessária a configuração do conjunto de dados que será utilizado, essas configurações direcionam o algoritmo durante a sua execução. (Tang; MacLennan, 2005)

Para cada um dos atributos existentes no conjunto de dados deve ser selecionada uma das seguintes opções:

- **Input** – Representa o atributo que será utilizado pelo algoritmo para definição dos nós da árvore.
- **Predict** – Representa o atributo preditivo utilizado para definição dos pontos de separação da árvore. Esse atributo também é utilizado como input.
- **PredictOnly** - Representa o atributo preditivo utilizado para definição dos pontos de separação da árvore. Não é será considerado como input.
- **Key** – Representa o atributo com o maior nível de detalhamento do conjunto de dados, deve possuir valores únicos.
- **Ignore** – Atributo que será ignorados para criação da árvore.

Além das configurações definidas para o conjunto de dados, é necessária a configuração do algoritmo. Seguem os principais parâmetros de configuração do algoritmo Microsoft Decision Trees. A definição correta desses parâmetros é de

fundamental importância para obtenção dos resultados corretos. (Microsoft Decision Trees Algorithm, 2007)

- **Minimum_Support** - Usado para especificar o tamanho mínimo de cada nó folha da árvore. Quando o número de casos em um nó não atingir esse valor o nó folha será criado.
- **Score_Method** - Usado para definir o método de separação dos valores. Que pode ser: "Gain Information", "Bayseian K2" e "Bayesian Dirichlet Equivalent with Uniform prior" (BDEU). Conforme descrito no item 3.1.2.
- **Split_Method** - Usado para especificar a forma de crescimento da árvore.
    - **1** – Binário
    - **2** – Multi-ponto
    - **3** – Ambos
- **Complexity_Penalty** - Usado para controlar o crescimento da árvore, o valor pode variar entre 0 e 1.
    - De 1 a 9 atributos, o recomendável é 0,5.
    - De 10 a 99 atributos, o recomendável é 0,9.
    - Mais que 100 atributos, o recomendável é 0,99.
- **Maximum_Input_Attributes** - Define o número de atributos de entrada que o algoritmo pode utilizar.
- **Maximum_Output_Attributes** - Define o número de atributos de saída que o algoritmo pode utilizar.

5.1.3 ALGORITMO MICROSOFT CLUSTERING

A Microsoft Clustering é um algoritmo de segmentação disponível no Microsoft SQL Server 2005. O algoritmo utiliza técnicas para o agrupamento dos dados que contêm características semelhantes, conforme descrito no capitulo 4.

A configuração do conjunto de dados é semelhante à realizada pelo algoritmo Microsoft Decision Trees, descrito no item 5.1.2.

Além das configurações definidas para o conjunto de dados, é necessária a configuração do algoritmo. Seguem os principais parâmetros de configuração do algoritmo Microsoft Clustering. A definição correta desses parâmetros é de fundamental importância para obtenção dos resultados corretos. (Microsoft Clustering Algorithm, 2007)

- **Clustering_Method** – Define qual o algoritmo utilizado. Os possíveis valores são:
  - 1 – EM
  - 2 – EM sem escalabilidade
  - 3 – K-Means
  - 4 – K-Means sem escalabilidade
- **Cluster_Count** – Define o número de clusters que serão criados pelo algoritmo.
- **Modelling_Cardinality** – Controla quantos modelos são gerados durante a execução do algoritmo. Conforme esse valor é reduzido o desempenho do algoritmo melhora, porém se perde na precisão dos resultados.
- **Cluster_Seed** – É um numero randômico utilizado para inicializar a clusters. Este parâmetro é permitir testar a sensibilidade dos dados. Se o modelo permanecer relativamente estável ao alterar esse valor, é certo que a segmentação dos dados está correta.
- **Minimum_Support** - Usado para especificar o numero mínimo de casos em cada cluster.

5.2 APLICAÇÃO DO MÉTODO DE CLASSIFICAÇÃO POR ÁRVORE DE DECISÃO

5.2.1 PRÉ-PROCESSAMENTO DO ALGORITMO DE ÁRVORE POR DECISÃO

Para inicio do processo foi necessária a definição do objetivo, dessa forma contactou-se o Gerente de Vendas da Empresa ABC, a fim de que uma necessidade real da empresa fosse definida.

O seguinte objetivo foi definido para aplicação do processo de Data Mining: A empresa do ramo de eletrodomésticos quer realizar uma análise de classificação de vendas das suas diferentes marcas de panelas, a fim de realizar uma campanha de marketing na época de Natal. O objetivo da análise é definir, de forma geográfica as possíveis campanhas de marketing e incentivos para os diferentes canais de vendas.

Foi definido pela Empresa ABC um responsável pelo acompanhamento das atividades realizadas com suas informações, essa pessoa é definida como o analista de domínio da aplicação, pois conhece profundamente o negócio da Empresa ABC.

Como premissa para aplicação do algoritmo foi definido pelo analista de domínio da aplicação que todos os clientes corporativos que compram uma das marcas devem ser considerados como um cliente que poderia comprar qualquer uma das demais marcas.

5.2.1.1 SELEÇÃO DOS DADOS

Foi definido pelo analista de domínio da aplicação que a base de dados a ser utilizada para análise de classificação seria a base de dados referente ao mesmo período do ano anterior, ou seja, os dados de vendas dos dias 01/12/2007 a 25/12/2007. Essa definição foi realizada para que a base de dados represente o comportamento de vendas na Empresa ABC na época em que se deseja realizar a campanha de marketing.

A Empresa ABC possui um Data Warehouse gerencial com o foco de fornecer informações que auxiliem no processo de tomada de decisão. Ficou definido que as informações utilizadas para aplicação do processo de Data Mining devem ser extraídas do Data Warehouse uma vez que as mesmas já foram tratadas e modeladas para trabalharem com sistemas de apoio a decisão.

Definiu-se então o conjunto de dados inicial que seria disponibilizado para o inicio da análise.

Através da técnica de seleção de dados "Junção Orientada", foi disponibilizada pela Empresa ABC sua base de dados referente aos dados de venda conforme a tabela 5 abaixo.

CAMPO	DESCRIÇÃO
CAMPO_1	Tipo do Produto
CAMPO_2	Marca - Marca do Produto
CAMPO_3	Família - Família do Produto
CAMPO_4	SKU (Stock Keeping Unit) - Código do Produto
CAMPO_5	Canal - Canal de Vendas
CAMPO_6	Cliente Corporativo - Nome do Cliente Corporativo
CAMPO_7	Estado - Estado do Cliente
CAMPO_8	Cidade – Cidade do Cliente
DATA_1	Data – Data da Transação
VALOR_1	NET Sales - Valor Faturado

Tabela 4 – Conjunto de dados inicial para aplicação do processo de Data Mining.

Com o auxílio do analista de domínio da aplicação foram realizadas diversas consultas ao conjunto de dados inicial, e as seguintes ações foram realizadas com o objetivo de selecionar os dados necessários para aplicação do algoritmo:

- Na tabela estavam disponíveis as informações de todos os tipos de produtos da Empresa ABC. Para a nossa análise só foi necessário os produto do tipo "PANELA", tendo sido aplicada a técnica de redução de dados na horizontal, realizando a seleção dos dados referente aos produtos do tipo "PANELA".

- As informações referentes às vendas estavam com a periodicidade diária fazendo com que o conjunto de dados tivesse 115.890 registros. Uma vez que o conjunto de dados era referente ao período que seria necessário para realizar a análise, não importando o dia exato da transação, foi aplicada a técnica de agregação das informações, onde todos os registros que possuem os mesmos atributos foram agregados pelo CAMPO_6 (Clientes Corporativos) em um único registro independente de data de sua transação. Após a aplicação da técnica o número foi reduzido para 11.388 registros.

- Foi constatado que para o objetivo proposto não seriam necessárias as informações de FAMILIA DE PRODUTO e SKU, uma vez que a classificação será realizada por MARCA, e cada Família de Produto e SKU pertencem a uma só marca. A coluna com as informações do tipo de produto no CAMPO_1, também não é necessária, uma vez que somente os produtos do tipo "PANELA" permanecem na base. Foi aplicada a técnica de redução dos dados na vertical, eliminando essas colunas de forma que informação necessária para análise foi preservada.

- Foram identificadas sete Marcas disponíveis na base, conforme a ***tabela 5*** onde foi aplicada a análise vertical sobre o faturamento de cada marca. Foi aplicada a técnica de redução dos dados na vertical para exclusão das linhas que não representavam um valor significativo no Faturamento Total. Foram excluídas as linha E, F e G.

MARCA	FATURAMENTO	%
MARCA A	R$ 2.438.877,47	31,780
MARCA B	R$ 2.382.986,91	31,051
MARCA C	R$ 1.925.735,39	25,093
MARCA D	R$ 913.504,69	11,903
MARCA E	R$ 9.975,27	0,130
MARCA F	R$ 3.145,35	0,041
MARCA G	R$ 105,70	0,001
TOTAL	R$ 7.674.330,78	100

Tabela 5. Representação do faturamento por Marca.

Na ***tabela 6*** encontram-se o conjunto de dados final para aplicação do algoritmo de Árvore de Decisão.

CAMPO	DESCRIÇÃO
CAMPO_2	Marca - Marca do Produto
CAMPO_5	Canal - Canal de Vendas
CAMPO_6	Cliente Corporativo - Nome do Cliente Corporativo
CAMPO_7	Estado - Estado do Cliente
CAMPO_8	Cidade – Cidade do Cliente
VALOR_1	NET Sales - Valor Faturado

Tabela 6. Conjunto de dados final para aplicação do algoritmo.

5.2.1.2 LIMPEZA DOS DADOS

Como o conjunto de dados foi extraído do Data Warehouse da Empresa ABC, não foram encontrados valores ausentes, inconsistência nas informações ou qualquer outro tipo de problema que necessitasse da aplicação das técnicas de limpeza dos dados.

5.2.1.3 CODIFICAÇÃO

Baseado na premissa definida inicialmente de que todos os clientes corporativos que compram uma das marcas devem ser considerados como um cliente que poderia comprar qualquer uma das demais marcas, foram criadas quatro colunas na tabela, uma para cada marca, para que seja possível a análise individual de cada uma das marcas. Essas colunas foram utilizadas como o atributo preditivo para que o algoritmo possa realizar a classificação para cada uma das marcas. Essas colunas foram preenchidas com os seguintes valores:

SIM – Quando é realizada a transação de venda da marca.
NÃO – Quando não é realizada a transação de venda da marca.

Na ***tabela 7*** encontram-se o conjunto de dados final para aplicação do algoritmo de Árvore de Decisão.

CAMPO	DESCRIÇÃO
CAMPO_2	Marca - Marca do Produto
CAMPO_5	Canal - Canal de Vendas
CAMPO_6	Cliente Corporativo - Nome do Cliente Corporativo
CAMPO_7	Estado - Estado do Cliente
CAMPO_8	Cidade – Cidade do Cliente
VALOR_1	NET Sales - Valor Faturado
MARCA A	Representa a transação realizada com a MARCA A
MARCA B	Representa a transação realizada com a MARCA B
MARCA C	Representa a transação realizada com a MARCA C
MARCA D	Representa a transação realizada com a MARCA D

Tabela 7. Conjunto de dados final após a etapa de codificação.

5.2.1.4 ENRIQUECIMENTO

Devido a todas as informações necessárias para o alcance do objetivo proposto estarem disponíveis não foi necessário o enriquecimento dos dados

5.2.2 APLICAÇÃO DO ALGORITMO POR ÁRVORE DE DECISÃO

Com o conjunto de dados definido, realizou-se a aplicação do algoritmo, sendo utilizada a ferramenta Microsoft SQL Server 2005 descrita no capitulo 4.

O Algoritmo de classificação de Árvore de Decisão foi aplicado individualmente para cada uma das 4 colunas de classificação das marcas, conforme descrito no item 5.1.4.

De forma representativa as ***figuras 14 e 15*** estão exibindo a Árvore de Decisão gerada no primeiro nível de abertura. Com a utilização da ferramenta é possível analisar os próximos níveis.

Com a visualização da árvore de decisão é possível analisar através do gráfico de histograma existente em cada nó da árvore, a probabilidade dos clientes corporativos comprarem uma determinada marca, em um determinado estado.

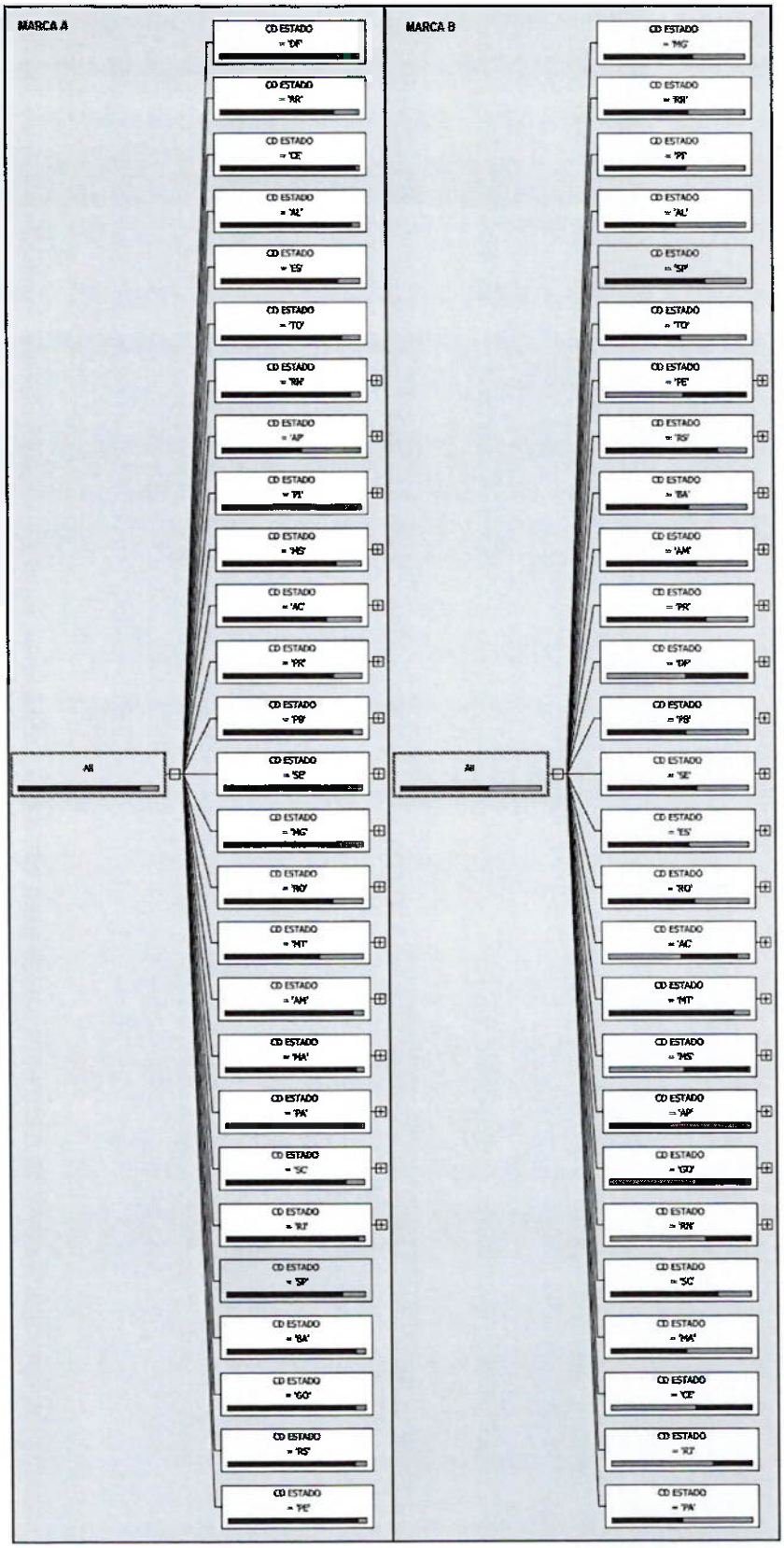

Figura 14. Árvore de Decisão da Marca A e Marca B

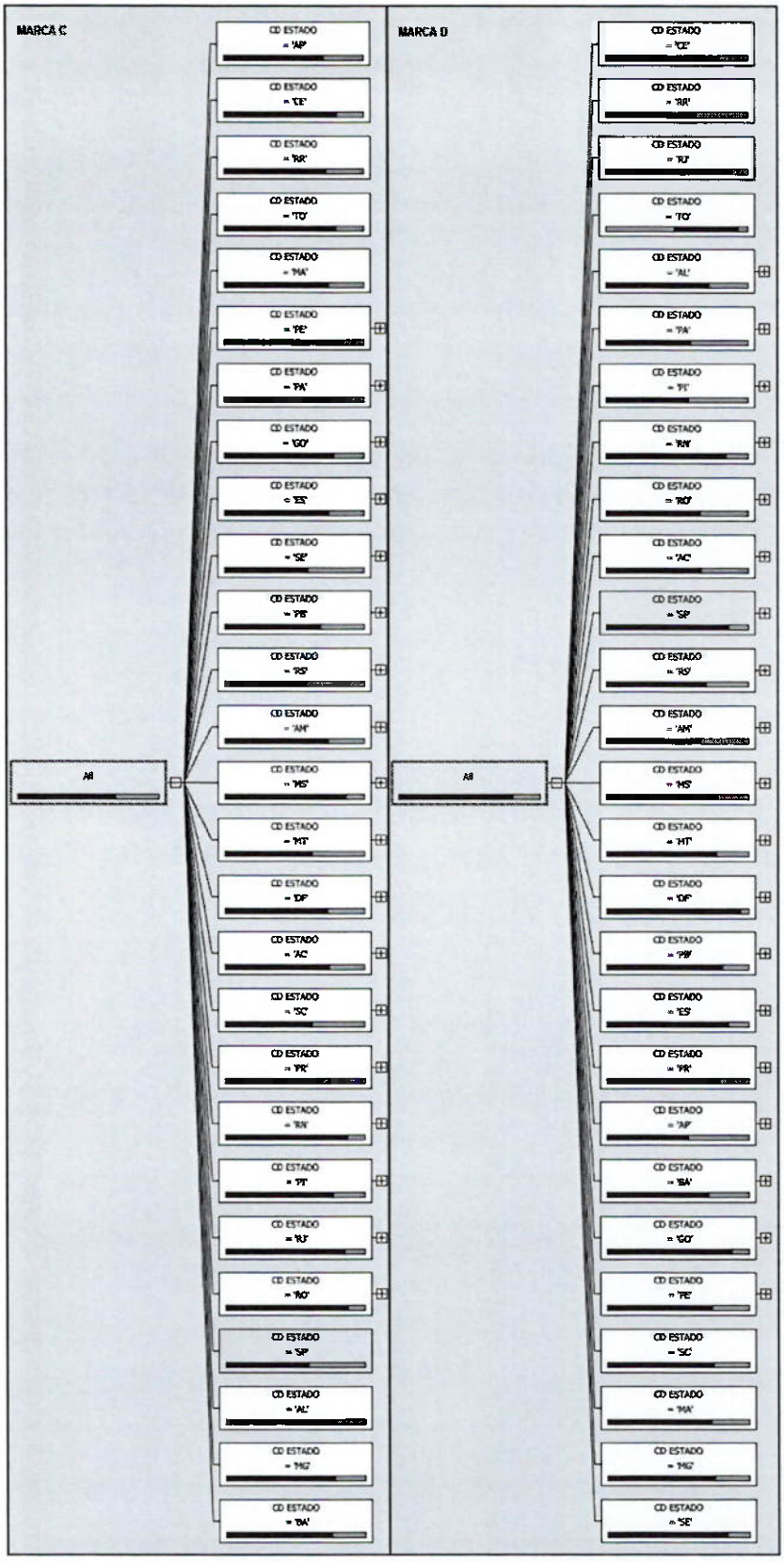

Figura 15. Árvore de Decisão da Marca C e Marca D

Cada um dos nós da árvore de decisão que possui o símbolo "+" na sua frente permite a abertura do nó para o próximo nível.

Conforme a ***figura 11***, com a abertura do próximo nível é possível analisar a probabilidade dos clientes corporativos comprar determinada marca, em um determinado estados, através de cada um dos canais de vendas disponíveis. Sendo que, no gráfico de histograma a barra "verde" representa a probabilidade dos clientes comprarem, enquanto a barra "vermelha" representa a probabilidade dos clientes não comprarem esta marca. A barra "cinza" representa os casos ausentes, conforme descrito no item 3.2 deste trabalho.

Com a análise da ***figura 11*** pode-se identificar que dentre todos os canais de vendas disponíveis no estado de 'SC', a probabilidade das vendas dessa marca é muito maior através do canal "AUTO SERVIÇO" em relação aos demais canais.

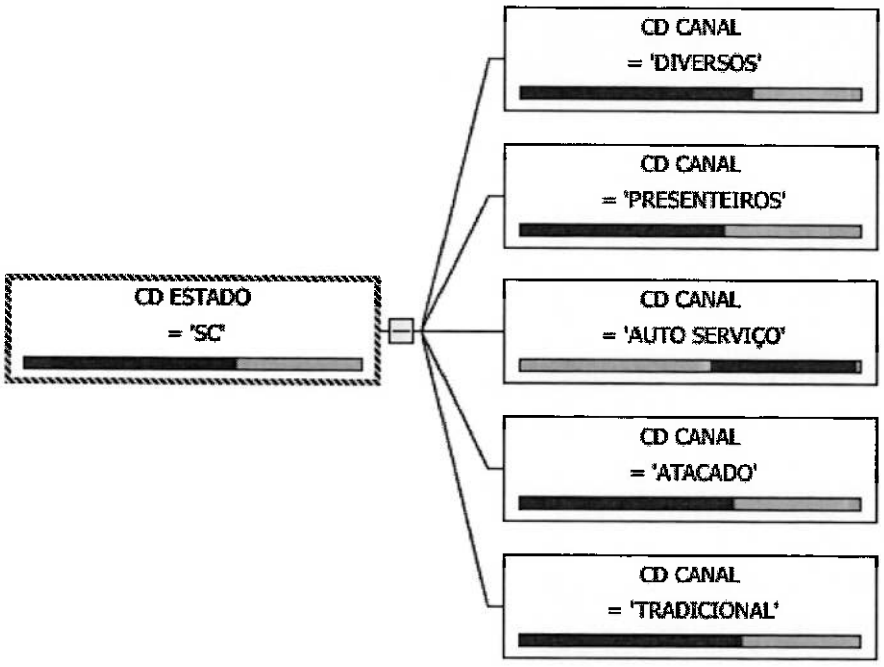

Figura 16. Árvore de decisão da probabilidade de venda por canal de vendas para clientes da marca C no estado de SC.

5.2.3 PÓS-PROCESSAMENTO DO ALGORITMO DE ÁRVORE DE DECISÃO

Após a finalização da etapa de aplicação do algoritmo inicia-se a etapa de pós-processamento. Com a utilização da ferramenta Microsoft SQL Server 2005 é possível uma análise iterativa dos resultados, além da visualização dos resultados conforme um nó é selecionado.

Como não era possível a disponibilização da ferramenta para os Gerentes e Supervisores Regionais de Vendas, bem como o departamento de Marketing da Empresa ABC, decidiu-se criar um relatório e gráficos comparativos para permitir o compartilhamento dos resultados.

Utilizando a ferramenta Microsoft SQL Server 2005 realizou-se a extração dos resultados referentes à aplicação do algoritmo. Na ***tabela 7*** pode ser visualizado o resultado da aplicação do algoritmo de Árvore de decisão. Pode-se visualizar o número de casos e probabilidade de compra de cada uma das marcas para cada um dos estados e seus canais de vendas.

Também foram criados gráficos para realização das análises comparativas da probabilidade de compras entre os estados, conforme demonstrado no ***gráfico 1***, outros gráficos comparativos entre os canais de venda que seguem o modelo do gráfico 1, foram criados para cada um dos estados.

Os resultados obtidos foram disponibilizados para Empresa ABC, sendo utilizado como ferramenta para que ela pudesse alcançar o objetivo proposto inicialmente.

ESTADO / CANAL	MARCA A						MARCA B						MARCA C						MARCA D						Total de CASOS
	CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			
	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	
AC	**8**	**1**	**0**	**75,00%**	**16,67%**	**8,33%**	**4**	**5**	**0**	**41,67%**	**50,00%**	**8,33%**	**8**	**1**	**0**	**75,00%**	**16,67%**	**8,33%**	**7**	**2**	**0**	**66,67%**	**25,00%**	**8,33%**	**9**
DIVERSOS	8	0	0	81,82%	9,09%	9,09%	3	5	0	36,36%	54,55%	9,09%	7	1	0	72,73%	18,18%	9,09%	6	2	0	63,64%	27,27%	9,09%	8
TRADICIONAL	0	1	0	25,00%	50,00%	25,00%	1	0	0	50,00%	25,00%	25,00%	1	0	0	50,00%	25,00%	25,00%	1	0	0	50,00%	25,00%	25,00%	1
AL	**128**	**6**	**0**	**94,16%**	**5,11%**	**0,73%**	**68**	**66**	**0**	**50,36%**	**48,91%**	**0,73%**	**108**	**26**	**0**	**79,56%**	**19,71%**	**0,73%**	**98**	**36**	**0**	**72,26%**	**27,01%**	**0,73%**	**134**
ATACADO																			2	3	0	37,50%	50,00%	12,50%	5
AUTO SERVIÇO																			25	4	0	81,25%	15,63%	3,13%	29
TRADICIONAL																			71	29	0	69,90%	29,13%	0,97%	100
AM	**110**	**7**	**0**	**92,50%**	**6,67%**	**0,83%**	**75**	**42**	**0**	**63,33%**	**35,83%**	**0,83%**	**88**	**29**	**0**	**74,17%**	**25,00%**	**0,83%**	**78**	**39**	**0**	**65,83%**	**33,33%**	**0,83%**	**117**
AUTO SERVIÇO	88	1	0	96,74%	2,17%	1,09%	61	28	0	67,39%	31,52%	1,09%	65	24	0	71,74%	27,17%	1,09%	53	36	0	58,70%	40,22%	1,09%	89
DIVERSOS	2	4	0	33,33%	55,56%	11,11%	4	2	0	55,56%	33,33%	11,11%	6	0	0	77,78%	11,11%	11,11%	6	0	0	77,78%	11,11%	11,11%	6
TRADICIONAL	20	2	0	84,00%	12,00%	4,00%	10	12	0	44,00%	52,00%	4,00%	17	5	0	72,00%	24,00%	4,00%	19	3	0	80,00%	16,00%	4,00%	22
AP	**3**	**1**	**0**	**57,14%**	**28,57%**	**14,29%**	**2**	**2**	**0**	**42,86%**	**42,86%**	**14,29%**	**4**	**0**	**0**	**71,43%**	**14,29%**	**14,29%**	**3**	**1**	**0**	**57,14%**	**28,57%**	**14,29%**	**4**
AUTO SERVIÇO	2	1	0	50,00%	33,33%	16,67%	1	2	0	33,33%	50,00%	16,67%							3	0	0	66,67%	16,67%	16,67%	3
TRADICIONAL	1	0	0	50,00%	25,00%	25,00%	1	0	0	50,00%	25,00%	25,00%							0	1	0	25,00%	50,00%	25,00%	1
BA	**533**	**37**	**0**	**93,19%**	**6,63%**	**0,17%**	**335**	**235**	**0**	**58,64%**	**41,19%**	**0,17%**	**434**	**136**	**0**	**75,92%**	**23,91%**	**0,17%**	**409**	**161**	**0**	**71,55%**	**28,27%**	**0,17%**	**570**
ATACADO							73	43	0	62,18%	36,97%	0,84%							83	33	0	70,59%	28,57%	0,84%	116
AUTO SERVIÇO							122	122	0	49,80%	49,80%	0,40%							183	61	0	74,49%	25,10%	0,40%	244
DIVERSOS							50	39	0	55,43%	43,48%	1,09%							74	15	0	81,52%	17,39%	1,09%	89
TRADICIONAL							90	31	0	73,39%	25,81%	0,81%							69	52	0	56,45%	42,74%	0,81%	121
CE	**370**	**9**	**0**	**97,12%**	**2,62%**	**0,26%**	**156**	**223**	**0**	**41,10%**	**58,64%**	**0,26%**	**307**	**72**	**0**	**80,63%**	**19,11%**	**0,26%**	**304**	**75**	**0**	**79,84%**	**19,90%**	**0,26%**	**379**
ATACADO																									
AUTO SERVIÇO																									
DIVERSOS																									
TRADICIONAL																									
DF	**130**	**22**	**0**	**84,52%**	**14,84%**	**0,65%**	**69**	**83**	**0**	**45,16%**	**54,19%**	**0,65%**	**113**	**39**	**0**	**73,55%**	**25,81%**	**0,65%**	**145**	**7**	**0**	**94,19%**	**5,16%**	**0,65%**	**152**
ATACADO							14	15	0	46,88%	50,00%	3,13%	25	4	0	81,25%	15,63%	3,13%	25	4	0	81,25%	15,63%	3,13%	29
AUTO SERVIÇO							41	24	0	61,76%	36,76%	1,47%	34	31	0	51,47%	47,06%	1,47%	63	2	0	94,12%	4,41%	1,47%	65
DIVERSOS							0	4	0	14,29%	71,43%	14,29%	4	0	0	71,43%	14,29%	14,29%	4	0	0	71,43%	14,29%	14,29%	4
PRESENTEIROS							4	6	0	38,46%	53,85%	7,69%	9	1	0	76,92%	15,38%	7,69%	9	1	0	76,92%	15,38%	7,69%	10
TRADICIONAL							10	34	0	23,40%	74,47%	2,13%	41	3	0	89,36%	8,51%	2,13%	44	0	0	95,74%	2,13%	2,13%	44
ES	**433**	**78**	**0**	**84,44%**	**15,37%**	**0,19%**	**293**	**218**	**0**	**57,20%**	**42,61%**	**0,19%**	**383**	**128**	**0**	**74,71%**	**25,10%**	**0,19%**	**437**	**74**	**0**	**85,21%**	**14,59%**	**0,19%**	**511**
ATACADO							9	7	0	52,63%	42,11%	5,26%	11	5	0	63,16%	31,58%	5,26%	13	3	0	73,68%	21,05%	5,26%	16
AUTO SERVIÇO							143	111	0	56,03%	43,58%	0,39%	192	62	0	75,10%	24,51%	0,39%	221	33	0	86,38%	13,23%	0,39%	254
DIVERSOS							8	22	0	27,27%	69,70%	3,03%	27	3	0	84,85%	12,12%	3,03%	30	0	0	93,94%	3,03%	3,03%	30
PRESENTEIROS							10	1	0	78,57%	14,29%	7,14%	2	9	0	21,43%	71,43%	7,14%	10	1	0	78,57%	14,29%	7,14%	11
TRADICIONAL							123	77	0	61,08%	38,42%	0,49%	151	49	0	74,88%	24,63%	0,49%	163	37	0	80,79%	18,72%	0,49%	200
GO	**288**	**23**	**0**	**92,04%**	**7,64%**	**0,32%**	**126**	**185**	**0**	**40,45%**	**59,24%**	**0,32%**	**245**	**66**	**0**	**78,34%**	**21,34%**	**0,32%**	**274**	**37**	**0**	**87,58%**	**12,10%**	**0,32%**	**311**
ATACADO							8	7	0	50,00%	44,44%	5,56%	10	5	0	61,11%	33,33%	5,56%	13	2	0	77,78%	16,67%	5,56%	15
AUTO SERVIÇO							80	84	0	48,50%	50,90%	0,60%	116	48	0	70,06%	29,34%	0,60%	139	25	0	83,83%	15,57%	0,60%	164
DIVERSOS							10	46	0	18,64%	79,66%	1,69%	54	2	0	93,22%	5,08%	1,69%	55	1	0	94,92%	3,39%	1,69%	56
PRESENTEIROS							25	40	0	38,24%	60,29%	1,47%	55	10	0	82,35%	16,18%	1,47%	56	9	0	83,82%	14,71%	1,47%	65
TRADICIONAL							3	8	0	28,57%	64,29%	7,14%	10	1	0	78,57%	14,29%	7,14%	11	0	0	85,71%	7,14%	7,14%	11
MA	**151**	**7**	**0**	**94,41%**	**4,97%**	**0,62%**	**86**	**72**	**0**	**54,04%**	**45,34%**	**0,62%**	**120**	**38**	**0**	**75,16%**	**24,22%**	**0,62%**	**117**	**41**	**0**	**73,29%**	**26,09%**	**0,62%**	**158**
ATACADO	9	1	0	76,92%	15,38%	7,69%																			10
AUTO SERVIÇO	31	0	0	94,12%	2,94%	2,94%																			31
DIVERSOS	70	5	0	91,03%	7,69%	1,28%																			75
TRADICIONAL	41	1	0	93,33%	4,44%	2,22%																			42
MG	**740**	**150**	**0**	**82,98%**	**16,91%**	**0,11%**	**558**	**332**	**0**	**62,60%**	**37,29%**	**0,11%**	**701**	**189**	**0**	**78,61%**	**21,28%**	**0,11%**	**672**	**218**	**0**	**75,36%**	**24,52%**	**0,11%**	**890**
ATACADO	56	9	0	83,82%	14,71%	1,47%																			65
AUTO SERVIÇO	365	39	0	89,93%	9,83%	0,25%																			404
DIVERSOS	71	11	0	84,71%	14,12%	1,18%																			82
PRESENTEIROS	64	18	0	76,47%	22,35%	1,18%																			82
TRADICIONAL	184	73	0	71,15%	28,46%	0,38%																			257

ESTADO / CANAL	MARCA A						MARCA B						MARCA C						MARCA D						Total de CASOS
	CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			
	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	
MS	**67**	**13**	**0**	**81,93%**	**16,87%**	**1,20%**	**38**	**42**	**0**	**46,99%**	**51,81%**	**1,20%**	**71**	**9**	**0**	**86,75%**	**12,05%**	**1,20%**	**64**	**16**	**0**	**78,31%**	**20,48%**	**1,20%**	**80**
ATACADO	16	6	0	68,00%	28,00%	4,00%	11	11	0	48,00%	48,00%	4,00%	21	1	0	88,00%	8,00%	4,00%	18	4	0	76,00%	20,00%	4,00%	22
AUTO SERVIÇO	21	3	0	81,48%	14,81%	3,70%	10	14	0	40,74%	55,56%	3,70%	24	0	0	92,59%	3,70%	3,70%	17	7	0	66,67%	29,63%	3,70%	24
PRESENTEIROS	20	0	0	91,30%	4,35%	4,35%	8	12	0	39,13%	56,52%	4,35%	13	7	0	60,87%	34,78%	4,35%	19	1	0	86,96%	8,70%	4,35%	20
TRADICIONAL	10	4	0	64,71%	29,41%	5,88%	9	5	0	58,82%	35,29%	5,88%	13	1	0	82,35%	11,76%	5,88%	10	4	0	64,71%	29,41%	5,88%	14
MT	**60**	**27**	**0**	**67,78%**	**31,11%**	**1,11%**	**78**	**9**	**0**	**87,78%**	**11,11%**	**11,11%**	**55**	**32**	**0**	**62,22%**	**36,67%**	**1,11%**	**68**	**19**	**0**	**76,67%**	**22,22%**	**1,11%**	**87**
ATACADO	8	4	0	60,00%	33,33%	6,67%	9	3	0	66,67%	26,67%	6,67%	8	4	0	60,00%	33,33%	6,67%	11	1	0	80,00%	13,33%	6,67%	12
AUTO SERVIÇO	5	5	0	46,15%	46,15%	7,69%	9	1	0	76,92%	15,38%	7,69%	6	4	0	53,85%	38,46%	7,69%	10	0	0	84,62%	7,69%	7,69%	10
DIVERSOS	5	4	0	50,00%	41,67%	8,33%	5	4	0	50,00%	41,67%	8,33%	9	0	0	83,33%	8,33%	8,33%	8	1	0	75,00%	16,67%	8,33%	9
TRADICIONAL	42	14	0	72,88%	25,42%	1,69%	55	1	0	94,92%	3,39%	1,69%	32	24	0	55,93%	42,37%	1,69%	39	17	0	67,80%	30,51%	1,69%	56
PA	**255**	**5**	**0**	**97,34%**	**2,28%**	**0,38%**	**131**	**129**	**0**	**50,19%**	**49,43%**	**0,38%**	**239**	**21**	**0**	**91,25%**	**8,37%**	**0,38%**	**155**	**105**	**0**	**59,32%**	**40,30%**	**0,38%**	**260**
ATACADO	15	0	0	88,89%	5,56%	5,56%							8	7	0	50,00%	44,44%	5,56%	12	3	0	72,22%	22,22%	5,56%	15
AUTO SERVIÇO	55	0	0	96,55%	1,72%	1,72%							50	5	0	87,93%	10,34%	1,72%	36	19	0	63,79%	34,48%	1,72%	55
DIVERSOS	114	1	0	97,46%	1,69%	0,85%							110	5	0	94,07%	5,08%	0,85%	61	54	0	52,54%	46,61%	0,85%	115
TRADICIONAL	71	4	0	92,31%	6,41%	1,28%							71	4	0	92,31%	6,41%	1,28%	46	29	0	60,26%	38,46%	1,28%	75
PB	**173**	**12**	**0**	**92,55%**	**6,31%**	**0,53%**	**103**	**82**	**0**	**55,32%**	**44,15%**	**0,53%**	**128**	**57**	**0**	**68,62%**	**30,85%**	**0,53%**	**151**	**34**	**0**	**80,85%**	**18,62%**	**0,53%**	**185**
ATACADO	4	0	0	71,43%	14,29%	14,29%	1	3	0	28,57%	57,14%	14,29%	3	1	0	57,14%	28,57%	14,29%	4	0	0	71,43%	14,29%	14,29%	4
AUTO SERVIÇO	55	8	0	84,85%	13,64%	1,52%	45	18	0	69,70%	28,79%	1,52%	29	34	0	45,45%	53,03%	1,52%	60	3	0	92,42%	6,06%	1,52%	63
DIVERSOS	67	3	0	93,15%	5,48%	1,37%	29	41	0	41,10%	57,53%	1,37%	62	8	0	86,30%	12,33%	1,37%	52	18	0	72,60%	26,03%	1,37%	70
TRADICIONAL	47	1	0	94,12%	3,92%	1,96%	28	20	0	56,86%	41,18%	1,96%	34	14	0	68,63%	29,41%	1,96%	35	13	0	70,59%	27,45%	1,96%	48
PE	**389**	**24**	**0**	**93,75%**	**6,01%**	**0,24%**	**191**	**222**	**0**	**46,15%**	**53,61%**	**0,24%**	**355**	**58**	**0**	**85,58%**	**14,18%**	**0,24%**	**309**	**103**	**0**	**73,32%**	**26,44%**	**0,24%**	**413**
ATACADO							53	82	0	39,13%	60,14%	0,72%	125	10	0	91,30%	7,97%	0,72%	95	40	0	69,57%	29,71%	0,72%	135
AUTO SERVIÇO							49	58	0	45,45%	53,64%	0,91%	84	23	0	77,27%	21,82%	0,91%	90	17	0	82,73%	16,36%	0,91%	107
DIVERSOS							10	22	0	31,43%	65,71%	2,86%	32	0	0	94,29%	2,86%	2,86%	26	6	0	77,14%	20,00%	2,86%	32
TRADICIONAL							79	60	0	56,34%	42,96%	0,70%	114	25	0	80,99%	18,31%	0,70%	93	46	0	66,20%	33,10%	0,70%	139
PI	**22**	**1**	**0**	**88,46%**	**7,69%**	**3,85%**	**14**	**9**	**0**	**57,69%**	**38,46%**	**3,85%**	**19**	**4**	**0**	**76,92%**	**19,23%**	**3,85%**	**14**	**9**	**0**	**57,69%**	**38,46%**	**3,85%**	**23**
ATACADO	9	1	0	76,92%	15,38%	7,69%							7	3	0	61,54%	30,77%	7,69%	8	2	0	69,23%	23,08%	7,69%	10
AUTO SERVIÇO	13	0	0	87,50%	6,25%	6,25%							12	1	0	81,25%	12,50%	6,25%	6	7	0	43,75%	50,00%	6,25%	13
PR	**616**	**154**	**0**	**79,82%**	**20,05%**	**0,13%**	**543**	**227**	**0**	**70,38%**	**29,50%**	**0,13%**	**541**	**229**	**0**	**70,12%**	**29,75%**	**0,13%**	**612**	**158**	**0**	**79,30%**	**20,57%**	**0,13%**	**770**
ATACADO	70	14	0	81,61%	17,24%	1,15%	50	34	0	58,62%	40,23%	1,15%	59	25	0	68,97%	29,89%	1,15%	73	11	0	85,06%	13,79%	1,15%	84
AUTO SERVIÇO	186	30	0	85,39%	14,16%	0,46%	196	20	0	89,95%	9,59%	0,46%	136	80	0	62,56%	36,99%	0,46%	130	86	0	59,82%	39,73%	0,46%	216
DIVERSOS	68	8	0	87,34%	11,39%	1,27%	52	24	0	67,09%	31,65%	1,27%	44	32	0	56,96%	41,77%	1,27%	64	12	0	82,28%	16,46%	1,27%	76
PRESENTEIROS	28	15	0	63,04%	34,78%	2,17%	28	15	0	63,04%	34,78%	2,17%	35	8	0	78,26%	19,57%	2,17%	38	5	0	84,78%	13,04%	2,17%	43
TRADICIONAL	264	87	0	74,86%	24,86%	0,28%	217	134	0	61,58%	38,14%	0,28%	267	84	0	75,71%	24,01%	0,28%	307	44	0	87,01%	12,71%	0,28%	351
RJ	**434**	**20**	**0**	**95,19%**	**4,60%**	**0,22%**	**132**	**322**	**0**	**29,10%**	**70,68%**	**0,22%**	**390**	**64**	**0**	**85,56%**	**14,22%**	**0,22%**	**407**	**47**	**0**	**89,28%**	**10,50%**	**0,22%**	**454**
ATACADO	2	1	0	50,00%	33,33%	16,67%							3	0	0	66,67%	16,67%	16,67%							3
AUTO SERVIÇO	89	1	0	96,77%	2,15%	1,08%							76	14	0	82,80%	16,13%	1,08%							90
DIVERSOS	90	1	0	96,81%	2,13%	1,06%							86	5	0	92,55%	6,38%	1,06%							91
PRESENTEIROS	8	0	0	81,82%	9,09%	9,09%							8	0	0	81,82%	9,09%	9,09%							8
TRADICIONAL	245	17	0	92,83%	6,79%	0,38%							217	45	0	82,26%	17,36%	0,38%							262
RN	**173**	**12**	**0**	**92,55%**	**6,91%**	**0,53%**	**62**	**123**	**0**	**33,51%**	**65,96%**	**0,53%**	**163**	**22**	**0**	**87,23%**	**12,23%**	**0,53%**	**157**	**28**	**0**	**84,04%**	**15,43%**	**0,53%**	**185**
ATACADO	48	1	0	94,23%	3,85%	1,92%	9	40	0	19,23%	78,85%	1,92%	47	2	0	92,31%	5,77%	1,92%	43	6	0	84,62%	13,46%	1,92%	49
AUTO SERVIÇO	69	1	0	95,89%	2,74%	1,37%	17	53	0	24,66%	73,97%	1,37%	68	2	0	94,52%	4,11%	1,37%	56	14	0	78,08%	20,55%	1,37%	70
DIVERSOS	22	3	0	82,14%	14,29%	3,57%	14	11	0	53,57%	42,86%	3,57%	21	4	0	78,57%	17,86%	3,57%	18	7	0	67,86%	28,57%	3,57%	25
PRESENTEIROS	9	1	0	76,92%	15,38%	7,69%	9	1	0	76,92%	15,38%	7,69%	3	7	0	30,77%	61,54%	7,69%	9	1	0	76,92%	15,38%	7,69%	10
TRADICIONAL	25	6	0	76,47%	20,59%	2,94%	13	18	0	41,18%	55,88%	2,94%	24	7	0	73,53%	23,53%	2,94%	31	0	0	94,12%	2,94%	2,94%	31
RO	**49**	**12**	**0**	**78,13%**	**20,31%**	**1,56%**	**38**	**23**	**0**	**60,94%**	**37,50%**	**1,56%**	**55**	**6**	**0**	**87,50%**	**10,94%**	**1,56%**	**41**	**20**	**0**	**65,63%**	**32,81%**	**1,56%**	**61**
ATACADO	7	0	0	80,00%	10,00%	10,00%	0	7	0	10,00%	80,00%	10,00%	7	0	0	80,00%	10,00%	10,00%	7	0	0	80,00%	10,00%	10,00%	7
AUTO SERVIÇO	4	0	0	71,43%	14,29%	14,29%	4	0	0	71,43%	14,29%	14,29%	3	1	0	57,14%	28,57%	14,29%	1	3	0	28,57%	57,14%	14,29%	4
PRESENTEIROS	19	6	0	71,43%	25,00%	3,57%	18	7	0	67,86%	28,57%	3,57%	24	1	0	89,29%	7,14%	3,57%	14	11	0	53,57%	42,86%	3,57%	25
TRADICIONAL	19	6	0	71,43%	25,00%	3,57%	16	9	0	60,71%	35,71%	3,57%	21	4	0	78,57%	17,86%	3,57%	19	6	0	71,43%	25,00%	3,57%	25

ESTADO / CANAL	MARCA A						MARCA B						MARCA C						MARCA D						Total de CASOS
	CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			CASOS			PROBABILIDADE			
	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	NAO	SIM	NULL	
RR	**17**	**2**	**0**	**81,82%**	**13,64%**	**4,55%**	**12**	**7**	**0**	**59,09%**	**36,36%**	**4,55%**	**15**	**4**	**0**	**72,73%**	**22,73%**	**4,55%**	**13**	**6**	**0**	**63,64%**	**31,82%**	**4,55%**	**19**
TRADICIONAL																									19
RS	**462**	**40**	**0**	**91,68%**	**8,12%**	**0,20%**	**398**	**104**	**0**	**79,01%**	**20,79%**	**0,20%**	**294**	**208**	**0**	**58,42%**	**41,39%**	**0,20%**	**354**	**148**	**0**	**70,30%**	**29,50%**	**0,20%**	**502**
ATACADO							32	15	0	66,00%	32,00%	2,00%	26	21	0	54,00%	44,00%	2,00%	39	8	0	80,00%	18,00%	2,00%	47
AUTO SERVIÇO							242	38	0	85,87%	13,78%	0,35%	164	116	0	58,30%	41,34%	0,35%	180	100	0	63,96%	35,69%	0,35%	280
DIVERSOS							23	17	0	55,81%	41,86%	2,33%	34	6	0	81,40%	16,28%	2,33%	25	15	0	60,47%	37,21%	2,33%	40
INSTITUCIONAL							0	1	0	25,00%	50,00%	25,00%	1	0	0	50,00%	25,00%	25,00%	1	0	0	50,00%	25,00%	25,00%	1
PRESENTEIROS							1	2	0	33,33%	50,00%	16,67%	3	0	0	66,67%	16,67%	16,67%	2	1	0	50,00%	33,33%	16,67%	3
TRADICIONAL							100	31	0	75,37%	23,88%	0,75%	66	65	0	50,00%	49,25%	0,75%	107	24	0	80,60%	18,66%	0,75%	131
SC	**363**	**55**	**0**	**86,46%**	**13,30%**	**0,24%**	**318**	**100**	**0**	**75,77%**	**23,99%**	**0,24%**	**261**	**157**	**0**	**62,23%**	**37,53%**	**0,24%**	**314**	**104**	**0**	**74,82%**	**24,94%**	**0,24%**	**418**
ATACADO	21	0	0	91,67%	4,17%	4,17%							14	7	0	62,50%	33,33%	4,17%							21
AUTO SERVIÇO	53	6	0	87,10%	11,29%	1,61%							26	33	0	43,55%	54,84%	1,61%							59
DIVERSOS	18	7	0	67,86%	28,57%	3,57%							18	7	0	67,86%	28,57%	3,57%							25
PRESENTEIROS	6	1	0	70,00%	20,00%	10,00%							5	2	0	60,00%	30,00%	10,00%							7
TRADICIONAL	265	41	0	86,08%	13,59%	0,32%							198	108	0	64,40%	35,28%	0,32%							306
SE	**114**	**12**	**0**	**89,15%**	**10,08%**	**0,78%**	**80**	**46**	**0**	**62,79%**	**36,43%**	**0,78%**	**76**	**50**	**0**	**59,69%**	**39,53%**	**0,78%**	**108**	**18**	**0**	**84,50%**	**14,73%**	**0,78%**	**126**
ATACADO	15	0	0	88,89%	5,56%	5,56%	14	1	0	83,33%	11,11%	5,56%	2	13	0	16,67%	77,78%	5,56%							15
AUTO SERVIÇO	91	6	0	92,00%	7,00%	1,00%	56	41	0	57,00%	42,00%	1,00%	61	36	0	62,00%	37,00%	1,00%							97
DIVERSOS	8	6	0	52,94%	41,18%	5,88%	10	4	0	64,71%	29,41%	5,88%	13	1	0	82,35%	11,76%	5,88%							14
SP	**3813**	**745**	**0**	**83,62%**	**16,36%**	**0,02%**	**3227**	**1331**	**0**	**70,77%**	**29,20%**	**0,02%**	**2716**	**1842**	**0**	**59,57%**	**40,41%**	**0,02%**	**3923**	**635**	**0**	**86,03%**	**13,94%**	**0,02%**	**4558**
ATACADO																			304	42	0	87,39%	12,32%	0,29%	346
AUTO SERVIÇO																			1693	401	0	80,78%	19,17%	0,05%	2094
DIVERSOS																			329	21	0	93,48%	6,23%	0,28%	350
PRESENTEIROS																			507	40	0	92,36%	7,45%	0,18%	547
TRADICIONAL																			1090	131	0	89,13%	10,78%	0,08%	1221
TO	**12**	**0**	**0**	**86,67%**	**6,67%**	**6,67%**	**7**	**5**	**0**	**53,33%**	**40,00%**	**6,67%**	**11**	**1**	**0**	**80,00%**	**13,33%**	**6,67%**	**6**	**6**	**0**	**46,67%**	**46,67%**	**6,67%**	**12**
DIVERSOS																									12
TOTAL GERAL	**9.913**	**1.475**	**0**	**87,03%**	**12,96%**	**0,01%**	**7.144**	**4.244**	**0**	**62,72%**	**37,27%**	**0,01%**	**7.900**	**3.488**	**0**	**69,36%**	**30,63%**	**0,01%**	**9.235**	**2.153**	**0**	**81,08%**	**18,91%**	**0,01%**	**11.338**

Tabela 8. Relatório da Probabilidade de Compra Estado/Canal.

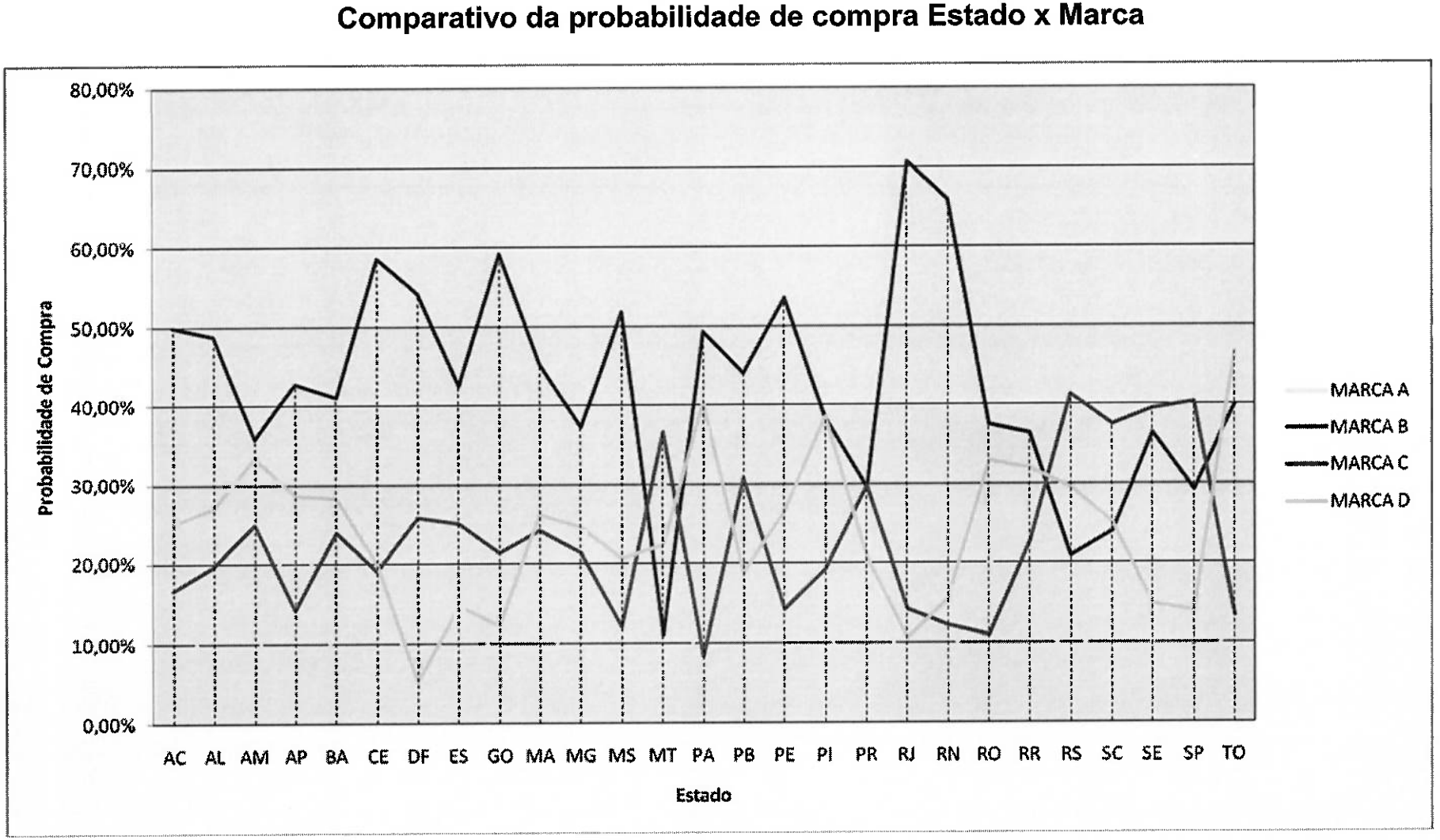

Gráfico 1. Comparativo da probabilidade de compras Estado x Marca

5.3 APLICAÇÃO DO MÉTODO DE SEGMENTAÇÃO POR CLUSTERING

5.3.1 PRÉ-PROCESSAMENTO DO ALGORITMO DE CLUSTERING

O objetivo utilizado para aplicação do método de segmentação por Clustering é baseado no objetivo definido para aplicação do método de classificação por Árvore de decisão descrito no item 5.2.1.

A única alteração é, ao invés de realizar a classificação, busca-se realizar o agrupamento conforme o comportamento de compras dos clientes, a fim de atingir o mesmo objetivo, a realização de uma campanha de marketing na época de Natal.

O mesmo conjunto de dados gerado após a etapa de pré-processamento para algoritmo de Árvore de Decisão é utilizada para aplicação do algoritmo de Clustering.

5.3.2 APLICACÃO DO ALGOLRITMO DE CLUSTERING

Com o conjunto de dados definido, realizou-se a aplicação do algoritmo, sendo utilizada a ferramenta Microsoft SQL Server 2005 descrita no item 5.1.

Foi aplicado o algoritmo de segmentação por Clustering. A ***figura 17*** exibe os resultados obtidos após a aplicação do algoritmo K-Means. Com a utilização da ferramenta de visualização do Microsoft SQL 2005, pode-se visualizar as variáveis (Variables) utilizadas para definição dos agrupamentos e os possíveis estados (States) para cada uma das variáveis. Cada um dos estados de uma variável é identificado por uma cor. Nas próximas colunas pode-se ver todos os clusters criados, permitindo a identificação da composição de cada cluster através das cores de cada estado de uma variável.

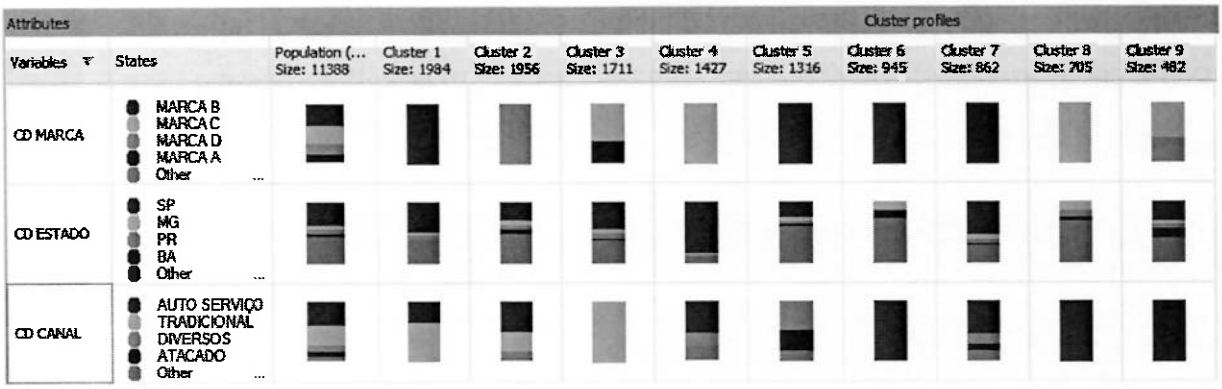

Figura 17. Clusters criados após aplicação do algoritmo.

Para visualização da composição dos casos em cada cluster deve-se selecionar um determinado cluster que a ferramenta de visualização dos resultado do Microsoft SQL 2005 exibe a legenda com os detalhamentos do cluster, sendo exibida a distribuição para cada estado da variável. A ***figura 18*** exibe um exemplo do detalhamento de um cluster.

Color	Meaning	Distribution
	AUTO SERVIÇO	0,506
	TRADICIONAL	0,325
	DIVERSOS	0,126
	ATACADO	0,000
	PRESENTEIROS	0,043
	INSTITUCIONAL	0,000
	missing	0,000

Figura 18. Detalhamento do Cluster

5.3.3 PÓS-PROCESSAMENTO DO ALGORITMO DE CLUSTERING

Após a finalização da etapa de aplicação do algoritmo inicia-se a etapa de pós-processamento. Com a utilização da ferramenta Microsoft SQL Server 2005 é possível uma análise interativa dos resultados, além da visualização dos resultados conforme um cluster é selecionado.

Como não era possível a disponibilização da ferramenta para os Gerentes e Supervisores Regionais de Vendas, bem como o departamento de Marketing da Empresa ABC, decidiu-se criar um relatório de análise dos clusters, para permitir o compartilhamento dos resultados.

Utilizando a ferramenta Microsoft SQL Server 2005 realizou-se a extração dos resultados referentes à aplicação do algoritmo. Na ***tabela 8*** pode ser visualizado o resultado da aplicação do algoritmo de Clustering. Pode-se visualizar para cada cluster a maior representação de cada variável, além da proporção de casos em relação ao total.

Cluster	Proporção de Casos	Maior representação por Marca	Maior representação por Canal	Maior representação por Estado
1	17,42%	MARCA B (100,0%)	TRADICIONAL (64,9%)	SP (49,7%)
2	17,18%	MARCA D (100,0%)	AUTO SERVIÇO (50,6%) TRADICIONAL (32,5%)	SP (30,3%)
3	15,02%	MARCA C (63,8%) MARCA A (36,2%)	TRADICIONAL (100,0%)	SP (46,7%)
4	12,53%	MARCA C (100,0%)	AUTO SERVIÇO (52,1%)	SP (84,1%)
5	11,56%	MARCA B (100,0%)	DIVERSOS (47,1%) ATACADO (34,7%)	SP (26,2%)
6	8,30%	MARCA B (100,0%)	AUTO SERVIÇO (100,0%)	MG (15,2%) BA (12,9%)
7	7,57%	MARCA A (100%)	AUTO SERVIÇO (53,7%)	SP (54,9%)
8	6,19%	MARCA C (100,0%)	AUTO SERVIÇO (100,0%)	RS(16,6%) MG (15,2%) PR (11,3 %)
9	4,23%	MARCA C (59,1%) MARCA D (40,9%)	ATACADO (100,0%)	SP (33,4%)

Tabela 91. Maior representação por cluster.

5.4 COMPARAÇÃO DOS RESULTADOS OBTIDOS

Os resultados obtidos com a aplicação do algoritmo de Árvore de Decisão permitem a classificação dos clientes conforme a sua probabilidade de compra. Através da análise dos resultados é possível identificar os estados e seus respectivos canais de vendas com maiores ou menores probabilidades de comprar determinada marca.

Com os resultados obtidos com a aplicação do algoritmo de Clustering foram definidos os principais agrupamentos de clientes que possuem características em comum.

Os resultados de ambos os algoritmos podem ser utilizados como ferramenta para se alcançar o objetivo proposto inicialmente ao processo de Data Mining. Nesse caso, o conhecimento detalhado do negócio passa a ser fundamental para utilização correta dos resultados. Para realização de uma campanha de marketing, uma outros fatores devem ser levados em consideração, como por exemplo:

- O custo de uma campanha de marketing em cada estado, ou nacional;
- A capacidade de produção dos produtos de cada marca;
- A capacidade de distribuição dos produtos em cada estado;
- O posicionamento dos concorrentes de cada marca;
- O publico alvo de cada marca.

Ao fim do processo de Data Mining, novos objetivos e perguntas podem ser feitas com base nos resultados, e um novo processo pode ser iniciado, podendo ser aplicados outros algoritmos, ou os mesmos algoritmos em novos conjuntos de dados.

6 CONCLUSÕES

Quando aplicado corretamente o processo de Data Mining é uma poderosa ferramenta na busca de informações, podendo trazer ganhos efetivos para as organizações.

Para que um projeto de Data Mining tenha sucesso é fundamental que todas as etapas do processo de Data Mining sejam seguidas rigorosamente, desde a definição clara do objetivo da aplicação do processo, até a análise dos resultados obtidos. Qualquer falha na execução das etapas do processo representa perda na qualidade dos resultados.

A compreensão do objetivo proposto ao processo de Data Mining é de fundamental importância para evitar retrabalhos, diversas decisões que têm que ser tomadas durante a execução do processo são regidas pelo objetivo, de forma que o não entendimento do objetivo irá resultar em ações durante o processo que só serão percebidas em etapas posteriores, e algumas vezes somente com a análise dos resultados.

A etapa mais importante do processo de Data mining é o pré-processamento, a seleção correta do conjunto de dados e a compreensão das informações existentes nas bases dados são de fundamental importância para que os algoritmos de Data Mining não gerem resultados que não representam a realidade do negócio, e pior, podendo servir de base para auxiliar o processo de tomada de decisão dentro das organizações.

Para utilização dos algoritmos de Árvore de Decisão e Clustering é necessário que se tenha conhecimentos básicos sobre o seu funcionamento. Dependendo do objetivo proposto ao processo de Data Mining diferentes parâmetros devem passados para o algoritmo de forma a interferirem nos resultados obtidos.

Ambos os algoritmos podem fornecer respostas para mesma pergunta, dessa forma, o conhecimento do negócio em que os algoritmos estão sendo aplicados é de fundamental importância, gerando um ciclo iterativo do processo de Data Mining, a fim de aferir e detalhar os resultados obtidos. A análise correta dos resultados é que pode otimizar os resultados das empresas.

Prever o futuro é muito difícil, porém a utilização do Data Mining em larga escala nas organizações deve ser considerada como uma opção. Soluções interativas e de fácil utilização, voltadas para usuários finais que permitam a realização do processo de Data Mining, desde o pré-processamento, seleção do algoritmo e análise dos resultados deverão ser criadas.

Como trabalhos futuros são sugeridas a pesquisa e a aplicação em casos reais de outros algoritmos de Data Mining, de forma a aproximar o mundo acadêmico, com o mundo dos negócios, devido ao potencial ganho que o processo de Data Mining como um todo pode trazer para as empresas.

REFERÊNCIAS

1. Ballard, Chuck; Rollins, Jonh; Ramos, Jo; Perkins, Andy; Hale, Richard; Dorneich, Ansgar; Edward, Cas Milner; Chodagan, Jonardhan. Dynamic Wharehousing: Data Mining Made Easy. PoughKeepsie, NY, USA – ibm.com/redbooks. 2007.

2. Carvalho, Luís Carlos Vidal de, DATA MINING: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: Editora Ciência Moderna Ltda., 2005.

3. Chen, Ming-Syan; Han, Jiawei; Yu, Philip S..Data Mining: An Overview from a Database Perspective. IEEE International Professional Communication Conference, 1996.

4. Cheung, David W.; Ng, Vicent T.; Fu, Ada W.; Fu, Yongjian. Efficient Mining of Association Rules in Distributed Databases. IEEE International Professional Communication Conference. 1996.

5. Goldschmidt, Ronaldo e Passos, Emmanuel, DATA MINING: Um guia prático. Rio de Janeiro: Elsevier, 2005.

6. Han, Jiawei e Kamber, Micheline, DATA MINING: Concepts and Techiniques. San Francisco, CA, USA – Elsevier Inc., 2006.

7. LARSON, Brian, Delivering Business Intelligence with Microsoft SQL Server 2005. USA: McGraw-Hill/Osborne, 2006.

8. Liu, Wencai e Lui, Yu. Applications of Clustering Data Mining in Customer Analysis in Department Store. IEEE International Professional Communication Conference Proceedings Computer Society. 2005.

9. Machado, Felipe Nery Rodrigues. Tecnologia e Projeto de Data Warhouse: Uma visão multidimensional. São Paulo. Editora: Érica, 2004.

10. SQL Server 2005 Product Guide (2005). Disponível em http://www.microsoft.com/sql/prodinfo/overview/productguide.mspx. Consulta em 18/04/2007.

11. Microsoft Data Mining Algorithms (2007) - SQL Server 2005 Books On-Line. Disponível em http://msdn.microsoft.com/en-us/library/ms175595.aspx. Consulta em 21/05/2008.

12. Microsoft Decision Trees Algoritm (2007) - SQL Server 2005 Books On-Lin. Disponível em http://msdn.microsoft.com/en-us/library/ms175312.aspx. Consulta em 21/05/2008.

13. Microsoft Clustering Algoritm (2007) - SQL Server 2005 Books On-Line. Disponível em http://msdn.microsoft.com/en-us/library/ms174879.aspx . Consulta em 19/07/2008.

14. Tang, ZhaouHui; MacLennan, Jamie. Data Mining with SQL Server 2005. Indianapólis,USA. Editora: Wiley Publishing Inc. 2005.