COPPE/UFRJ

MINERAÇÃO DE DADOS APLICADA AO ENTENDIMENTO DO COMPORTAMENTO DO CONSUMIDOR PARA DAR SUPORTE AO PROCESSO DE TOMADA DE DECISÕES

Leandro da Silva Carvalho

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro

Fevereiro de 2009

MINERAÇÃO DE DADOS APLICADA AO ENTENDIMENTO DO COMPORTAMENTO DO CONSUMIDOR PARA DAR SUPORTE AO PROCESSO DE TOMADA DE DECISÕES

Leandro da Silva Carvalho

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____

Prof ª. Beatriz de Souza Leite Pires de Lima, D.Sc.

_____

Prof. Elton Fernandes, Ph.D.

RIO DE JANEIRO, RJ – BRASIL

FEVEREIRO DE 2009

Carvalho, Leandro da Silva

Mineração de Dados Aplicada ao Entendimento do Comportamento do Consumidor para dar Suporte ao Processo de Tomada de Decisões / Leandro da Silva Carvalho. – Rio de Janeiro: UFRJ/COPPE, 2009.

XV, 108 p.: il., 29,7cm

Orientador: Nelson Francisco Favilla Ebecken

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Civil, 2008.

Referencias Bibliográficas: p. 107-108.

1. Mineração de Dados. 2. Comportamento do Consumidor. 3. Sistema de Apoio à Tomada de Decisão. 4. Segmentação. 5. Regra de Associação. 6. Árvore de Decisão. I. Ebecken, Nelson Francisco Favilla II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

Agradecimentos

Agradeço ao que permitiu saúde e paz não só a mim, mas em especial aos meus familiares e amigos.

À minha mãe Neiva e ao meu irmão Leonardo por tudo.

À minha Lilian, companheira querida e ponto de equilíbrio, que sobre meus ombros, de forma não intencional, foi a primeira a ler este trabalho.

Ao meu orientador Nelson Ebecken por toda liberdade, confiança e paciência que me foi concedida.

Aos meus familiares e amigos pela compreensão de minha ausência por todo este tempo.

Aos amigos de turma, funcionários e prestadores de serviços da COPPE pelo tempo de convivência.

Ao La Mole por acreditar em meus projetos.

À PETROBRAS por toda oportunidade oferecida.

Felicidade, passei no vestibular
Mas a faculdade é particular
...
Livros tão caros tantas taxas pra pagar
Meu dinheiro muito raro,
Alguém teve que emprestar
...
Morei no subúrbio, andei de trem atrasado
Do trabalho ia pra aula, sem jantar e bem cansado
Mas lá em casa à meia-noite tinha sempre a me esperar
Um punhado de problemas e criança pra criar
...
Mas felizmente eu consegui me formar
Mas da minha formatura, não cheguei participar
Faltou dinheiro pra beca e também pro meu anel
Nem o diretor careca entregou o meu papel
...
E depois de tantos anos,
Só decepções, desenganos
Dizem que sou um burguês muito privilegiado
Mas burgueses são vocês
Eu não passo de um pobre-coitado
E quem quiser ser como eu,
Vai ter é que penar um bocado
Um bom bocado, vai penar um bom bocado

Martinho da Vila – O Pequeno Burguês

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MINERAÇÃO DE DADOS APLICADA AO ENTENDIMENTO DO COMPORTAMENTO DO CONSUMIDOR PARA DAR SUPORTE AO PROCESSO DE TOMADA DE DECISÕES

Leandro da Silva Carvalho

Fevereiro/2009

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Este trabalho descreve uma metodologia de mineração de dados para extrair padrões de forma automática e adaptativa e interagir com os clientes considerando as suas próprias características pessoais. Foi utilizada uma base de dados de um sistema de comércio eletrônico contendo uma grande quantidade de clientes não ativos.

Identificaram-se os atributos que diferenciam um cliente ativo de outro não ativo através da aplicação do método de segmentação de dados *K-means*. Foram analisadas as relações de compra entre os produtos e seus respectivos clientes através do método de regras de associação com o uso do algoritmo *Apriori*. O estudo desenvolveu dois modelos de classificadores, feitos a partir do método de árvore de decisão, com o objetivo de inferir o tipo de cliente e o valor do pedido a ser realizado a partir dos atributos pessoais de um novo cliente.

Ao final é mostrado como utilizar a mineração de dados para gerar o conhecimento que dará o suporte necessário ao processo de tomada de decisão pelos estrategistas de negócio, possibilitando desta forma uma entrega de valor ao consumidor final.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

DATA MINING APPLIED TO THE CONSUMER BEHAVIOR UNDERSTANDING IN ORDER TO SUPPORT DECISION MAKING PROCESS

Leandro da Silva Carvalho

February/2009

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineer

This paper describes a data mining methodology to extract patterns with an automatic and adaptive way to interact with customers based on their own personal characteristics. An e-commerce database with a large number of inactive customers was used.

It was identified the attributes that distinguish an active client from an inactive one using *K-means* data segmentation method. The relationship between products purchase and their customers was analyzed through the association rules using *Apriori* algorithm. The study developed two model's classifiers, by decision tree method, in order to deduce the type of customer and the value of the request by the personal attributes of a new customer.

At the end is shown how to use data mining to generate knowledge that will provide the necessary support to the decision-making by business strategy, thus allowing value delivery to the final consumer.

Índice

Agradecimentos iv

Índice viii

Índice de Figura xi

Índice de Tabela xiv

Índice de Listagem xv

1 Introdução 1

1.1 Motivação 2

1.2 Objetivos 3

1.3 O NETCOMERCE 3

1.4 O La Mole 4

1.5 Problema em Questão 4

1.6 Metodologia e Recursos Tecnológicos 5

1.7 Organização do Trabalho 5

2 Mineração de Dados 7

2.1 Processo de Mineração de Dados 7

2.2 Técnicas de Mineração de Dados 9

2.3 Algoritmos de Mineração de Dados 10

2.3.1 Algoritmo de Segmentação de Dados – *K-Means* 10

2.3.2 Algoritmo de Regra de Associação – *Apriori* 11

2.3.3 Algoritmo de Árvore de Decisão 12

3 Pré-Processamento 14

4 Estatísticas Básicas 19

4.1 Análise do Conjunto de Dados de Clientes 19

4.2 Análise do Conjunto de Dados de Pedidos 29

5 Segmentação dos Clientes 35

5.1 Introdução 35

5.2 Análise e Desenvolvimento 35

5.2.1 Segmentação dos Clientes da Região Zona Norte I 46
5.2.2 Segmentação dos Clientes da Região Zona Sul III 48
5.2.3 Segmentação dos Clientes da Região Zona Sul I 50
5.2.4 Segmentação dos Clientes da Região Barra da Tijuca / Recreio 52
5.2.5 Segmentação dos Clientes da Região Centro 54
5.2.6 Segmentação dos Clientes da Região Zona Norte II 56
5.2.7 Segmentação dos Clientes da Região Ilha do Governador 57
5.2.8 Segmentação dos Clientes da Região Zona Norte III 59
5.2.9 Segmentação dos Clientes da Região Zona Sul II 61
5.2.10 Segmentação dos Clientes da Região Niterói 62

5.3 Conclusão 63

6 Análise de Regras de Associação 65

6.1 Introdução 65

6.2 Análise e Desenvolvimento 65

6.3 Conclusão 81

7 Classificação Supervisionada 83

7.1 Introdução 83

7.2 Análise e Desenvolvimento 83

7.2.1 Modelo Classificador de Clientes 84
7.2.2 Modelo Classificador do Valor do Pedido 94

7.3 Conclusão 97

8 Conclusões 98

Anexo A – Modelo de Dados Analítico 103

Anexo B – Conteúdo das Tabelas de Dados..... 105

Referências 107

Índice de Figura

Figura 4.1: Gráfico de freqüência do atributo idade. 21
Figura 4.2: Gráfico de freqüência do atributo tempo de relacionamento. 21
Figura 4.3: Gráfico de freqüência do atributo sexo. 22
Figura 4.4: Gráfico de freqüência do atributo bairro. 22
Figura 4.5: Gráfico de freqüência do atributo loja. 23
Figura 4.6: Gráfico de freqüência do atributo tipo de cliente. 23
Figura 4.7: Gráfico *blox-pot* das variáveis de entrada. 24
Figura 4.8: Gráfico de freqüência do atributo faixa etária. 25
Figura 4.9: Gráfico de freqüência do atributo região. 25
Figura 4.10: Gráfico *blox-pot* do novo conjunto de variáveis de entrada. 26
Figura 4.11: Gráfico que contém a matriz de correlação. 26
Figura 4.12: Gráfico de projeção dos atributos do conjunto de dados de clientes. 27
Figura 4.13: *Data Image* da distância Euclidiana. 28
Figura 4.14: Gráfico de freqüência do atributo região. 30
Figura 4.15: Gráfico de freqüência do atributo dia da semana. 30
Figura 4.16: Gráfico de freqüência do atributo feriado. 31
Figura 4.17: Gráfico de freqüência do atributo horário do pedido. 31
Figura 4.18: Gráfico de freqüência do atributo região. 32
Figura 4.19: Gráfico de freqüência do atributo valor do pedido. 32
Figura 4.20: Gráfico *blox-pot* das variáveis do conjunto de dados dos pedidos realizados. 32
Figura 4.21: Gráfico da matriz de correlação do conjunto de dados dos pedidos. 33
Figura 4.22: Gráfico de projeção dos atributos do conjunto de dados de pedidos. 34
Figura 5.1: Gráfico de perfis dos clusters encontrados pelo algoritmo *K-means*. 37
Figura 5.2: Diagrama de cluster que mostra o primeiro nível de semelhança entre os clusters encontrados pelo modelo. 38
Figura 5.3: Visualização gráfica dos atributos e valores que distinguem o cluster 1 do cluster 4. 39
Figura 5.4: Diagrama de cluster que mostra um nível de semelhança acima do encontrado pela figura 5.2. 39
Figura 5.5: Atributos e valores que distinguem o cluster 3 do cluster 9. 39

Figura 5.6: Resultado obtido pelo *K-means*, com o valor de parametrização igual a 5. 40
Figura 5.7: Diagrama de cluster com os clusters mais semelhantes da segunda análise.41
Figura 5.8: Atributos e valores que distinguem o cluster 4 do cluster 5. 42
Figura 5.9: Resultado obtido pelo *K-means*, com o valor de parametrização igual a 15. 43
Figura 5.10: Diagrama de cluster com os clusters mais semelhantes da terceira análise. 44
Figura 5.11: Segmentação dos clientes da região Zona Norte I. 46
Figura 5.12: Segmentação dos clientes da região Zona Sul III. 49
Figura 5.13: Segmentação dos clientes da região Zona Sul III. 51
Figura 5.14: Segmentação dos clientes da região Barra / Recreio. 53
Figura 5.15: Segmentação dos clientes da região Centro. 55
Figura 5.16: Segmentação dos clientes da região Zona Norte II. 57
Figura 5.17: Segmentação dos clientes da região Ilha do Governador. 58
Figura 5.18: Segmentação dos clientes da região Zona Norte III. 60
Figura 5.19: Segmentação dos clientes da região Zona Sul II. 61
Figura 5.20: Segmentação dos clientes da região Niterói. 62
Figura 6.1: Diagrama de freqüência da quantidade de itens por pedido 66
Figura 6.2: Regras geradas a partir dos produtos mais vendidos. 67
Figura 6.3: Suporte das regras geradas pelos produtos mais vendidos. 68
Figura 6.4: Diagrama de dependências entre as regras geradas. 69
Figura 6.5: Diagrama de dependências com o nível final das regras geradas. 69
Figura 6.6: Regras geradas a partir das categorias de produtos mais vendidos. 70
Figura 6.7: Regras geradas com o valor mínimo de suporte igual a 1% para as categorias de produtos mais vendidos 70
Figura 6.8: Diagrama de dependências das regras geradas pelas categorias dos produtos. 72
Figura 6.9: Regras geradas com o valor mínimo de suporte igual a 1%, a partir da criação das subcategorias de bebidas. 73
Figura 6.10: Diagrama de dependências das regras geradas a partir da criação das subcategorias de bebidas. 74
Figura 6.11: Regras de associação geradas para os produtos relacionados à cerveja. ... 75
Figura 6.12: Suporte das regras geradas para os produtos relacionados à cerveja. 75

Figura 6.13: Diagrama de dependências das regras geradas para os produtos relacionados à cerveja. 75
Figura 6.14: Sugestão de 5 produtos com maior probabilidade de compra para o produto '*couvert* família'. 76
Figura 6.15: Sugestão de 5 produtos com maior probabilidade de compra em conjunto dos produtos '*couvert* família' '*lasgana* à *bolognesa*'. 77
Figura 6.16: Regras de associação geradas com os atributos dos usuários. 78
Figura 6.17: Regras de associação geradas com os atributos dos usuários sem os produtos mais vendidos. 79
Figura 6.18: Regras de associação geradas com os atributos dos usuários utilizando o atributo categoria. 80
Figura 6.19: Regras de associação com algumas variações de importância negativas. . 81
Figura 6.20: Diagrama de dependências das regras geradas para os produtos relacionados à cerveja. 81
Figura 7.1: Exemplos de clusters do conjunto de dados de clientes da região Z. Sul III. 92
Figura 7.2: Ramificação da árvore de decisão Z-Sul-III. 93
Figura A.1: Modelo de dados entre as tabelas de CEP, Bairro, Região e Loja. 103
Figura A.2: Modelo de dados entre as tabelas de Cliente (Ativos e Não ativos), Sexo e Faixa Etária. 103
Figura A.3: Modelo de dados entre as tabelas de dimensões Tempo, Feriado e Período. 104
Figura A.4: Modelo de dados relacionado aos Pedidos e Produtos..... 104

Índice de Tabela

Tabela 3.1: Dados sobre o número de clientes e pedido. 14
Tabela 4.1: Tabela de correlação entre as variáveis de entrada utilizadas pelo estudo.. 26
Tabela 4.2: Tabela de correlação entre as variáveis utilizadas nos dados dos pedidos.. 33
Tabela 5.1: Conhecimento intrínseco obtido pela segmentação da região Z. Norte I.... 47
Tabela 5.2: Resultado obtido na segmentação dos dados da região Z. Sul III. 49
Tabela 7.1: Atributos utilizados por cada modelo no conjunto de dados de clientes..... 85
Tabela 7.2: Resultados obtidos por cada modelo da tabela 7.1. 86
Tabela 7.3: Resultados das matrizes de confusão da dos diversos modelos de clientes da primeira análise. 87
Tabela 7.4: Atributos utilizados por cada modelo de subconjunto de clientes. 88
Tabela 7.5: Resultados obtidos por cada modelo da tabela 7.3. 88
Tabela 7.6: Resultados das matrizes de confusão dos diversos modelos de clientes da segunda análise. 89
Tabela 7.7: Resultados obtidos pelos modelos da região Z. Norte I. 89
Tabela 7.8: Resultados das matrizes de confusão dos modelos Z. Norte I. 90
Tabela 7.9: Resultados obtidos dos modelos por região. 90
Tabela 7.10: Resultados das matrizes de confusão dos modelos por região. 90
Tabela 7.11: Exemplo de predições para os clientes da Zona Sul III. 93
Tabela 7.12: Resultados da matriz de confusão obtida pela validação cruzada. 95
Tabela 7.13: Resultados dos modelos classificadores de pedidos feitos a partir de diferentes subconjuntos. 96
Tabela B.1: Conteúdo da tabela DW_Faixa_Etaria..... 105
Tabela B.2: Conteúdo da tabela de regiões 105
Tabela B.3: Conteúdo da tabela DW_Feriados 106
Tabela B.4: Conteúdo da tabela DW_Perido_Horario 106

Índice de Listagem

Listagem 6.1: Consulta ao modelo de produtos para o produto ‘*couvert* família’. 76
Listagem 6.2: Consulta ao modelo de produtos para a compra em conjunto dos produtos ‘*couvert* família’ e ‘*lasgana* à *bolognesa*’. 77
Listagem 6.3: Consulta ao modelo de categoria através da linguagem DMX. 80
Listagem 7.1: Exemplo de consulta ao modelo de Z.Sul-III. 93
Listagem 7.2: Consulta DMX que retorna a matriz de confusão obtida pela validação cruzada. 95

1 Introdução

As empresas possuem atualmente uma capacidade sem precedentes de coletar informações a respeito de suas atividades e de seus clientes. Isto se deve em grande parte aos avanços tecnológicos e ao modo pela qual estas tecnologias são utilizadas para estreitar o relacionamento com os seus consumidores e parceiros.

O uso do comércio eletrônico, que hoje representa um dos principais canais de vendas a disposição das empresas, é um exemplo do uso da tecnologia como ferramenta para a produção de informações, já que a Internet representa uma interface automática entre clientes e empresas.

Esta interface permite o fornecimento de produtos e serviços com lucro, a um baixo custo e com uma abrangência maior do que os tradicionais pontos de vendas, o que proporciona às empresas participar de um novo setor de mercado, trazendo novos clientes e parceiros.

Tudo isso aumenta a quantidade de dados existentes e faz com que a aplicação dos recursos tecnológicos, em especial o uso da Internet, torne a coleta de informações em uma tarefa simples. Porém, muitas companhias que usam estes recursos não sabem como promover o acesso a estas informações, visto que seu verdadeiro valor encontra-se na transformação dos dados armazenados em conhecimento.

Contudo, a transformação de todo esse volume de dados em conhecimento ultrapassa a capacidade humana de analisar tais informações através dos instrumentos tradicionais disponíveis.

Por conta disso, um grande esforço aplicado à inteligência computacional tem sido feito para auxiliar na busca por técnicas que ajudem na extração de conhecimentos em grandes bases de dados. Estas técnicas estão relacionadas ao que a literatura denomina por mineração de dados ou *data mining*.

A idéia principal é o desenvolvimento de métodos que sejam capazes de analisar e aprender a partir de um determinado conjunto de informações, tendo como objetivo a transformação de dados “crus” em informações relevantes à geração do conhecimento necessário para se tomar uma determinada decisão [4].

Este suporte à tomada de decisões tem feito com que a mineração de dados se torne uma poderosa e efetiva ferramenta na busca implícita pelo conhecimento, em especial ao entendimento do comportamento do consumidor ou CRM (*Customer Relationship Management*).

O entendimento do comportamento do consumidor, segundo Ngai, Xiu *et al* [11], é definido como uma maneira de ajudar às organizações a melhor discriminar seus clientes, ao possibilitar, de forma mais eficaz, a alocação de recursos para o grupo de clientes mais rentáveis através do ciclo de identificação, atração, retenção e desenvolvimento de seus clientes.

Desta forma, analisar os dados dos clientes, a partir das técnicas de mineração de dados existentes, pode produzir o conhecimento que dará o suporte necessário ao processo de tomada de decisões, o que transforma a tecnologia em um recurso estratégico a serviço das organizações.

1.1 Motivação

Vislumbrando uma oportunidade de negócio, foi criado pelo autor, no ano 2000, um produto de comércio eletrônico denominado NETCOMMERCE. Em meados de 2004, o La Mole, uma tradicional rede de restaurantes carioca, mostrou interesse pelo produto com o objetivo de desafogar sua central de atendimento e possibilitar uma facilidade a mais para os seus clientes, de modo que seus pratos pudessem ser oferecidos pela Internet.

Desde então a empresa vem adquirindo uma repleta e rica base de dados sobre os seus clientes e produtos, assim como a relação entre um e outro, definidos através dos pedidos realizados. Porém, atualmente nenhuma análise sobre estes dados é realizada pela empresa.

Assim, o maior desafio é fazer uma análise minuciosa destas informações com o intuito de se gerar um conhecimento específico que dê o suporte necessário à tomada de decisões futuras.

1.2 Objetivos

O estudo pretende, com o uso da mineração de dados, extrair padrões de forma automática e adaptativa para interagir com os clientes tomando por base as suas próprias características pessoais.

Isto possibilitará uma personalização que ajudará na aquisição de novos consumidores, retenção dos atuais clientes e predição do comportamento que irá melhorar a capacidade de produtos a serem oferecidos.

A idéia é utilizar a tecnologia da informação como uma ferramenta estratégica, ao combinar a integração da Internet aos processos de negócio da companhia, construindo assim uma base de conhecimentos necessários para a vantagem competitiva da empresa.

1.3 O NETCOMERCE

O programa de comércio eletrônico denominado NETCOMMERCE encontra-se patenteado no Instituto Nacional da Propriedade Intelectual – INPI – sob o número do processo de registro 049686. Sua estrutura é dividida em duas partes: módulo administrativo e loja virtual.

O módulo administrativo tem por objetivo gerenciar a loja virtual permitindo o controle das informações necessárias para o seu funcionamento. É também de responsabilidade deste módulo a administração e controle dos pedidos feitos pelos clientes.

Quanto à loja virtual, esta representa o canal de vendas utilizado para atrair o cliente. Além disso, neste módulo é possível não só comprar os produtos oferecidos pela empresa, como também escolher a forma de pagamento e o local de entrega do pedido.

Ambos os módulos foram feitos a partir de páginas no formato ASP (Active Server Pages) em conjunto com um componente desenvolvido em Visual Basic 6. Este é o componente proprietário do sistema e que requer o registro em um servidor de aplicação. Neste caso, isto foi feito através do produto Internet Information Server 6, através do módulo COM+ do sistema operacional Windows 2003.

Em relação ao armazenamento dos dados, vale ressaltar que todas as informações produzidas são registradas em uma base própria, que utilizado o banco de dados MS SQL Server. Inicialmente, o produto foi desenvolvido utilizando a versão 2000; porém, atualmente toda a base de dados já foi migrada para a versão 2005.

1.4 O La Mole

A história do La Mole começa em abril de 1958, quando o italiano Domenico Magliano fundou a pequena Sorveteria e Pizzaria La Mole, na Rua Dias Ferreira, no Leblon. O nome da casa foi uma homenagem à torre Mole Antonelliana, localizada em Turim, na Itália, cidade natal de seu fundador.

Ao passar dos tempos, o restaurante foi se expandindo e diversificando sua cozinha, introduzindo novos conceitos na gastronomia carioca. Com público extremamente fiel, o restaurante do Leblon é o começo da "cultura" La Mole como tradição carioca. Em 1974, o La Mole chegou à Barra da Tijuca, bairro ainda pouco explorado na época, e, desde então, não parou de crescer.

Para aqueles que preferem comer em casa, o La Mole oferece entregas em domicílio. Para este tipo de entrega, os pedidos podem ser feitos pelo site www.lamole.com.br ou pelo telefone das lojas.

Hoje, a rede possui mais de 15 restaurantes localizados pelos bairros do Rio de Janeiro e Niterói, contando com aproximadamente 1.000 colaboradores para atender a todas as lojas da rede.

1.5 Problema em Questão

Foi identificado que uma grande quantidade de clientes cadastrados nunca fez algum tipo de pedido pela loja virtual do La Mole. Este é um comportamento muito estranho, pois o principal objetivo do cadastro é permitir identificar os usuários que efetuam pedidos através da Internet.

Vale ressaltar que este cadastro não é obrigatório para quem deseja navegar pelo site da empresa, seja para buscar o telefone de uma das lojas, conhecer o cardápio ou

utilizar o fale conosco da companhia. Este comportamento cria tanto uma necessidade quanto uma oportunidade para a utilização da mineração de dados.

A necessidade criada diz respeito ao entendimento do comportamento do consumidor que faz seu cadastro e não efetua um pedido. Isto mostra que há uma grande oportunidade para o processo de contribuição de entrega de valor ao consumidor final, visto que uma parte dos usuários cadastrados não está sendo atendida.

Desta forma, entender o que diferencia um usuário ativo de outro não ativo é de grande importância para estreitar o relacionamento com os clientes, pois isto reflete diretamente no sucesso da empresa.

1.6 Metodologia e Recursos Tecnológicos

A metodologia a ser utilizada partirá do método exploratório-descritivo, que tem por objetivo caracterizar o processo existente e identificar as variáveis e fluxos pertinentes a cada processo abordado pelo estudo. Este desenvolvimento será realizado a partir de uma breve revisão literária, que em seguida irá avaliar os problemas identificados e sugerir as possíveis formas de solução para os mesmos.

Quanto à metodologia experimental, esta será utilizada na etapa de construção dos modelos de mineração de dados a serem analisados. Esta tarefa será realizada com os recursos tecnológicos disponíveis pela ferramenta Microsoft SQL Server 2008, versão de avaliação CTP novembro de 2007.

Além desta ferramenta, o estudo fez o uso do programa MatLab, versão 7.0, e do Microsoft Excel, versão 2003, para avaliações dos dados estatísticos.

1.7 Organização do Trabalho

O presente trabalho é apresentado em capítulos a partir da seguinte divisão:

- Capítulo 2 – diz respeito a todo o conteúdo teórico necessário para o entendimento do presente trabalho. Neste capítulo é apresentado um breve

resumo sobre a mineração de dados, as etapas necessárias para sua execução e os algoritmos utilizados pelo estudo;

- Capítulo 3 – descreve a etapa de pré-processamento. Nele é mostrado como os dados do ambiente transacional foram modificados para um ambiente analítico, de modo que fosse possível uma melhora na qualidade dos dados a serem analisados e na redução de sua quantidade;

- Capítulo 4 – expõe o processo de entendimento dos dados, também chamado de estatísticas básicas. O objetivo é entender como estão distribuídos os dados e de que forma eles podem impactar nas avaliações futuras;

- Capítulo 5 – exibe a análise da classificação não supervisionada, onde são avaliadas as características que distinguem os clientes ativos dos não ativos;

- Capítulo 6 – avalia o comportamento de compra a partir dos pedidos realizados e das características de seus respectivos clientes. Esta avaliação é feita pelas regras de associações geradas através da análise dos dados contidos na base de pedidos;

- Capítulo 7 – demonstra o processo de classificação supervisionada a partir do uso do algoritmo de árvore de decisão. O objetivo é criação de modelos capazes de reconhecer a classificação de novos registros a partir das características de seus atributos;

- Capítulo 8 – apresenta a conclusão obtida pelo estudo após o termino de todas as análises realizadas;

- Apêndice A – ilustra o modelo de dados analítico e seus respectivos relacionamentos; e

- Apêndice B – contém a descrição do conteúdo das tabelas criadas na etapa de pré-processamento.

Sendo assim, o estudo finaliza sua apresentação e segue adiante com o desenvolvimento das tarefas necessárias à sua conclusão.

2 Mineração de Dados

A capacidade de reconhecimento e classificação de padrões é uma das mais fundamentais características da inteligência humana. De um modo geral, o reconhecimento de padrões pode ser definido como um processo que procura por estruturas nos dados e os classificam de acordo com o grau de relação entre eles [6].

Entre as várias técnicas de reconhecimentos de padrões existentes podemos citar as técnicas de agrupamentos, regras de associação e modelos classificadores. Todas estas técnicas estão relacionadas à mineração de dados, de modo que suas aplicações podem produzir um melhor resultado na busca pelo conhecimento empresarial.

Estes métodos possibilitam a observação de padrões previamente desconhecidos, mas que estão contidos de forma oculta nas bases de dados existentes. Entretanto, para que estas técnicas consigam obter um bom resultado, é necessário que algumas etapas sejam realizadas, como por exemplo, a definição do objetivo do problema, o pré-tratamento dos dados, o processamento da técnica selecionada, a interpretação do resultado e aplicação do mesmo ao problema selecionado.

Este é um processo iterativo e que deve ser feito de forma constante para que os modelos produzidos consigam sempre capturar os novos padrões, de forma a manter estes modelos atualizados continuamente com as necessidades das empresas.

2.1 Processo de Mineração de Dados

A primeira etapa do processo a ser realizada é entender de que maneira as técnicas de mineração de dados poderão ajudar na busca pelo conhecimento implícito dos dados, visto que estas técnicas são agrupadas em diferentes categorias e que suas utilizações dependem exclusivamente do tipo de conhecimento a ser extraído. Uma breve descrição de cada uma das categorias usadas pelo estudo é apresentada no item 2.2 deste capítulo.

Uma vez definido o tipo de problema a se resolver, a segunda tarefa é identificar qual o algoritmo será aplicado a cada uma das técnicas escolhidas. Isto é necessário

dado ao fato de que, para cada técnica existente, vários algoritmos podem ser encontrados.

Esta escolha é ainda um processo não estruturado, sendo baseado em expertises e julgamentos. Muitas vezes, esta decisão está limitada ao ambiente tecnológico utilizado pelas empresas, como é o caso do presente trabalho. Ao final deste capítulo é possível encontrar um resumo dos algoritmos utilizados pelo estudo.

Outra decisão a ser tomada é a definição dos dados que serão usados nas análises, onde é feita a escolha de utilização da totalidade dos dados ou apenas de uma parte, feitos a partir de uma amostragem dos dados.

Em seguida é realizada a etapa de pré-tratamento dos dados, que por sinal consome bastante tempo no processo de análise. Esta etapa se justifica pelo fato de que muitas vezes os dados analisados se encontram em um nível de detalhe muito específico, incompletos ou com algum tipo de ruído.

É nesta etapa que se faz a limpeza dos dados, removendo as informações que não serão utilizadas, assim como a análise da necessidade de aplicação de alguma técnica de tratamento de valores ausentes. Além disso, em muitas ocasiões, é necessário se fazer uma transformação dos dados, como por exemplo, a criação de uma base analítica. Tudo isso tem por objetivo garantir a qualidade dos dados analisados.

A próxima etapa deste processo é a análise exploratória das informações, tornando possível uma percepção inicial dos dados a fim de se obter um melhor entendimento das variáveis utilizadas. Isto é feito a partir da utilização de ferramentas que possibilitam a visualização dos dados, onde se verifica os histogramas de cada atributo, as características das distribuições das variáveis, suas correlações e a existência de dados com o comportamento diferente do demais (*outliers*).

Na execução desta etapa, muitas vezes, é necessário padronizar os dados com o objetivo de normalizar seus valores, não permitindo que um atributo possua uma influência maior sobre outro apenas por ter uma variação na escala. Além disso, alguns métodos de análise e mineração de dados são influenciados por estas distorções de escalas contidas nos conjunto de dados [5].

O modelo empregado pelo estudo utilizou o método *Z-Score*, que padroniza os dados com base na distribuição de freqüência das variáveis utilizadas. Para isto, o método usa a média e o desvio padrão de cada atributo, o que ajudar a reduzir a diferença da variabilidade dos valores contidos [7].

Logo após é iniciada a etapa de criação e avaliação dos modelos. Isto é feito através das diversas parametrizações possíveis para cada algoritmo utilizado, onde o objetivo é encontrar um modelo que seja capaz de produzir um menor erro a partir de uma baixa complexidade.

Escolhido o melhor modelo, este deve ter sua aplicação realizada à solução do problema proposto, de modo que seja avaliada sua eficácia em meio às diversas situações reais do dia a dia das empresas.

2.2 Técnicas de Mineração de Dados

As técnicas de mineração de dados devem ser utilizadas de acordo com o tipo de conhecimento a ser obtido. Por isso, é apresentado a seguir os tipos de técnicas utilizadas pelo estudo.

- Classificação Não Supervisionada – também chamada de agrupamento, clusterização ou segmentação. Os métodos desta categoria buscam a descoberta de padrões contidos em um conjunto de dados que os dividam em grupos que contenham semelhanças entre si. Sua regra principal consiste em encontrar os vários centros de agrupamentos que possam fornecer as características relevantes à identificação de classes similares a partir de um conjunto de dados finito;

- Regra de Associação – técnica introduzida inicialmente por Agrawal [1], obteve uma grande popularidade a partir do desenvolvimento da análise de cestas de mercados. Seu objetivo é identificar padrões de relacionamentos entre os itens de um conjunto de dados do tipo se $A \rightarrow B$, proporcionando um mecanismo útil para descobrir as correlações entre os dados subjacentes; e

- Classificação Supervisionada – tem por objetivo classificar os registros de uma base de dados em diferentes classes, ao buscar a criação de um modelo que seja capaz de inferir a classificação de novos registros.

2.3 Algoritmos de Mineração de Dados

É apresentada logo em seguida uma breve descrição do funcionamento de cada algoritmo usado pelo estudo. Vale ressaltar que o presente trabalho ficou limitado aos algoritmos existentes na ferramenta utilizada.

2.3.1 Algoritmo de Segmentação de Dados – *K-Means*

O algoritmo *K-means* - introduzido inicialmente por McQueen em 1967 [9] - é um método não hierárquico que tem por objetivo particionar um conjunto de dados em *K* grupos.

Sua execução é realizada através de um processo iterativo que trabalha baseado na minimização do erro de uma função de critério. Neste método, um determinado dado só pode pertencer a um único cluster por vez.

A idéia principal é feita a partir da minimização da soma dos quadrados das distâncias entre os dados e os centróides de cada subgrupo, o que normalmente é feito a partir da distância Euclidiana.

Inicialmente, o algoritmo seleciona de forma randômica a posição de cada centro de agrupamento, para logo depois particionar os dados e agrupá-los com base nos seus atributos e características a partir de um critério que é dividido em duas etapas:

Primeiro, associa-se cada dado a um único grupo de acordo com a mensuração de uma função de distância entre os dados e os centros dos clusters; em seguida, atualizam-se os centróides de cada grupo baseado nas atribuições da primeira etapa.

O algoritmo repete estes dois procedimentos de forma iterativa, até que seja interrompido quando um critério de erro específico for acionado, ou após um determinado número de iterações ser alcançado, ou até que os centróides de cada cluster e os dados utilizados na análise não necessitem de mais atualizações.

A principal vantagem do uso do algoritmo K-means é que ele é um método simples e de baixo custo computacional, o que permite ser executado de forma eficiente em grandes volumes de dados.

Entretanto, este método não traz nenhuma garantia quanto ao melhor resultado a ser obtido, já que a quantidade de centros a serem trabalhados pelo modelo deve ser definida previamente. Outro ponto a se considerar é a necessidade de alguma normalização dos dados, já que o efeito de escala pode adulterar o resultado do agrupamento.

De qualquer forma, uma vez determinado os melhores valores para a parametrização inicial, o algoritmo *K-means* é sem dúvida um método muito eficiente na tarefa de agrupar dados devido a sua comprovada escalabilidade e eficiência.

2.3.2 Algoritmo de Regra de Associação – *Apriori*

O algoritmo *Apriori* [2] propõe decompor o problema de encontrar regras de associação em duas etapas: primeiro, encontrar todos os subconjuntos de itens que são freqüentes na base de dados; e segundo, gerar as regras de associação a partir dos itens freqüentes encontrados na primeira etapa.

A primeira parte da solução é um processo iterativo, que contabiliza a freqüência de cada subconjunto de itens, verifica se o valor da freqüência respeita o valor mínimo de suporte estabelecido e separa os novos subconjuntos de itens a serem trabalhados na próxima iteração. Este suporte é calculado pela soma de todas as transações que contenham um determinado item dividido pelo número total de transações.

Já a segunda etapa avalia cada subconjunto de itens aprovado e gera uma regra para cada um destes subconjuntos. Esta é a parte do algoritmo que fica responsável por descobrir as regras que contenham um fator de confiança maior ou igual do que o especificado.

Esta confiança determina se uma regra é interessante ou não, o que é feito a partir do cálculo da probabilidade de que um item B ocorra dado a ocorrência de um item A, em todas as transações que contenham este item A.

A partir do momento em que os valores de suporte e confiança são definidos, a questão fica a cargo de se conseguir uma quantidade razoável de regras e que contenham uma qualidade nas informações.

Valores muito altos podem fazer com que o algoritmo não consiga extrair nenhuma regra interessante. Já o contrário, pode fazer com que uma quantidade exagerada de regras seja encontrada, mas com pouca qualidade.

Assim, podemos considerar que o algoritmo *Apriori* permite determinar e representar a relação entre os valores contidos nos bancos de dados existentes, o que auxilia no processo de análise do histórico de informações transacionais, trazendo uma ajuda no apoio à tomada de decisões.

2.3.3 Algoritmo de Árvore de Decisão

Dentre as muitas técnicas que podem ser encontradas para resolver os problemas de classificação supervisionada, uma que se destaca com bastante notoriedade são as chamadas árvores de decisão. Este é um método de discriminação não linear que providencia uma representação hierárquica das características dos atributos que definem uma determinada classe.

Sua execução tem por objetivo a divisão de um problema de alta complexidade em vários outros menores, porém de maior simplicidade, com o intuito de permitir a predicação de uma variável *Y* a partir de um conjunto de variáveis independentes *X*.

É possível encontrarmos na literatura diversos algoritmos capazes de construir um modelo classificador a partir de uma árvore de decisão. Entretanto, a ferramenta utilizada pelo estudo possui o seu próprio algoritmo, que foi desenvolvido pela equipe de pesquisa da Microsoft.

Por conta disso, o estudo ficou limitado ao descrever o seu funcionamento em detalhes. Sabemos apenas que a construção de uma árvore é efetuada por um processo recursivo, onde cada nó é dividido em novos nós. Este processo faz com que os dados obtidos pelas variáveis dependentes sejam particionados em uma série de nós descendentes até que os critérios de interrupção do crescimento da árvore sejam alcançados.

De início, uma árvore é criada por um nó que é denominado raiz, que logo em seguida é separado em outros nós. Cada um destes novos nós são identificados como nós filhos, que são ligados ao nó raiz por ramos. Os nós que são posteriormente

divididos são denominados de nós pais, enquanto que os nós sem divisão são denominados nós terminais.

Desta forma, os nós terminais representam a maioria dos membros de uma determinada classe, o que permite a predição da classificação de um novo registro a partir da navegação do nó raiz até o nó terminal que melhor se adapta aos dados apresentados à árvore.

Esta navegação gera o que é definido por regras de decisão, o que facilita a interpretação dos modelos e permite que estas regras descobertas sejam apresentadas em um modo lingüístico, de uma maneira mais compreensível e amigável.

3 Pré-Processamento

Esta etapa do trabalho tem por objetivo descrever a criação de um ambiente que possibilite uma melhor qualidade dos dados a serem analisados pelo estudo, ao transformar a base transacional existente em uma base analítica. Esta transformação será feita a partir da criação de um armazém de dados que contenha apenas os dados necessários aos processos de análises futuras.

De início, a primeira tarefa realizada foi exportar toda a base de dados transacional do ambiente de produção. Isto ocorreu através de uma rotina de extração de dados programada para ser executada após o fechamento das vendas do dia 31/12/2007, o que nos dá uma base de dados contendo todas as informações da utilização da loja virtual por um período de três anos. A tabela 3.1 informa a quantidade de registros contidos inicialmente nesta base.

Número total de clientes cadastrados	10.235
Número total de pedidos realizados	16.170

Tabela 3.1: Dados sobre o número de clientes e pedido.

Após a execução da rotina ETL, o próximo passo foi efetuar uma limpeza na base de dados de modo a reduzir a quantidade de informações contidas em cada uma das tabelas existentes. Esta tarefa foi efetuada através da exclusão das colunas das tabelas de dados que não seriam utilizadas no processo de análise, como por exemplo, os campos senha do usuário, número do telefone e informações específicas sobre o endereço.

Por motivos de segurança todas as informações estavam criptografadas. Isto fez com que o estudo realizasse a descriptografia destes dados para tornar possível o seu entendimento.

Um problema identificado inicialmente foi a quantidade de clientes cadastrados e que nunca fizeram algum tipo de pedido. Por isso o estudo separou em duas tabelas distintas os clientes ativos e não ativos, onde os ativos representam os clientes com pedidos (tabela DW_Cliente) e os não ativos os clientes sem pedidos (tabela DW_Cliente_Nao_Ativo).

Esta separação possibilitou a criação de uma tabela com 4.100 registros de clientes ativos e outra com 6.135 registros de clientes não ativos. O objetivo desta separação é facilitar, no momento apropriado, a análise das características dos clientes em um processo de segmentação e classificação dos mesmos.

Para os clientes não ativos, um filtro foi estabelecido para avaliar quem poderia se tornar um cliente ativo. Este filtro consistiu em uma pesquisa para saber se o campo CEP destes usuários estava contido em uma área de entrega válida. Com isso a tabela de clientes não ativos foi reduzida para um total de 4.251 registros.

O próximo passo foi transformar o campo data de nascimento em idade, que logo em seguida foi discretizado em faixas etárias a partir da tabela denominada DW_Faixa_Etária. O objetivo desta tarefa foi substituir os valores contínuos por valores categóricos.

Em seguida foi criado um novo campo, denominado tempo de relacionamento, nas tabelas de clientes com o objetivo de estimar o tempo total (em meses) que um determinado cliente está cadastrado, já que no momento em que o cliente se cadastrava esta informação não era armazenada.

Para a realização desta tarefa foi criado um procedimento para cada tipo de cliente, que obedeceu ao seguinte critério:

- Clientes ativos, feito a partir da identificação da data cadastrada no primeiro pedido realizado pelo cliente; e

- Clientes não ativos, o procedimento considerou a data de cadastro como sendo a data de cadastro do cliente ativo mais próximo e que continha um código de cliente menor que o seu.

Um campo com o este tempo de relacionamento em trimestres também foi adicionado às tabelas.

Em seguida foi feita uma rápida análise que observou uma determinada quantidade de registros com o mesmo número de CPF. Deste modo, foi necessária a realização de um processo de unificação dos dados que continham um mesmo CPF, não permitindo que um usuário tivesse mais de um registro na base existente.

A idéia era evitar que estes registros comprometessem a análise de segmentação dos dados ao considerar mais de uma vez as características de um mesmo usuário. Este processo de unificação levou em consideração a seguinte regra para os clientes ativos: apagar um dos registros duplicados, mantendo os dados do cliente com o maior tempo de relacionamento.

Com isso, os pedidos do registro apagado tiveram o campo código do cliente atualizado com o valor do código do cliente que foi mantido. Já o campo CEP foi atualizado pelo dado encontrado no cliente que possuía o maior número de pedidos.

Para os clientes não ativos, o procedimento manteve apenas o registro que continha o maior valor para o campo código do cliente.

Em seguida, foi criada uma tabela de CEP (DW_Cep) com os respectivos dados: endereço, bairro e cidade. O objetivo era substituir todos os dados de entrada manual por dados automatizados que garantissem a unificação de um mesmo endereço, uma vez que o campo nome da rua poderia ser escrito de várias formas, como por exemplo: Rua Voluntários da Pátria x R. Vol. Pátria.

Esta tabela foi preenchida a partir da busca de todos os CEPs cadastrados nas tabelas de clientes ativos, clientes não ativos e pedidos. A tabela de pedidos foi considerada pelo fato de que um pedido pode ser entregue em um local diferente do qual um determinado cliente mora.

Para se conseguir os dados de todos os CEP foi feito a utilização de um componente da empresa de hospedagem de sites Locaweb [8], que é responsável por hospedar a loja virtual do La Mole com o produto de comércio eletrônico NETCOMMERCE.

Nem todos os CEPs foram encontrados por este componente, por isso o estudo utilizou um *WebService* disponível na Internet [3]. Seu uso só foi realizado após uma validação, que consistiu em comparar os resultados dos endereços dos CEPs encontrados inicialmente pelo componente da Locaweb com os do *WebService*. Esta comparação permitiu verificar que os CEPs encontrados pelo componente eram iguais aos encontrados pelo *WebService*.

Para todos os registros da tabela de CEPs, um campo associativo a loja de entrega responsável por aquele CEP foi relacionado. Esta relação foi feita a partir da tabela DW_Loja.

Em seguida foi criada uma tabela própria para agrupar os nomes dos bairros, denominada DW_Bairro, já que foi verificada a existência de escritas distintas para um mesmo bairro, como por exemplo: Icarai x Icaraí.

Além disso, o estudo criou mais uma tabela (DW_Regiao) para complementar a classificação regional das lojas. O propósito desta tabela é agrupar os bairros em regiões não oficiais para ajudar aos estrategistas de marketing em futuras campanhas comerciais.

A tarefa seguinte foi incluir o campo sexo em cada uma das tabelas de clientes, uma vez que esta informação não era solicitada pelo cadastro. Esta é sem dúvida uma importante informação quando se trata da segmentação das características de um cliente.

Estes procedimentos realizados pelo pré-processamento fizeram com que as tabelas de clientes ativos e não ativos ficassem com um total de 3.851 registros e 3.991 registros respectivamente, reduzindo em mais de 23% o total dos dados contidos na tabela original.

Dando continuidade, o estudo passou a concentrar seus esforços na criação das tabelas de dimensões que iriam compor as tabelas de clientes e pedidos. A primeira tabela criada, denominada no modelo como DW_Feriado, diz respeito aos feriados e datas comemorativas (como é o caso dos dias das mães e dia dos pais).

Outra tabela criada foi a tabela de dimensão DW_Tempo, contendo as características específicas a respeito das datas de cada ano, onde cada registro representa um intervalo de 30 minutos para cada hora de um determinado dia do ano. Além disso, esta tabela contém a informação sobre o dia da semana de uma determinada data e se esta data ocorreu em um dos dias relacionados à tabela de feriados.

Em seguida as tabelas de clientes e pedidos passaram a fazer referência a tabela DW_Tempo, já que ambas possuíam atributos do tipo data em suas tabelas.

Para dar mais suporte a tabela DW_Tempo, foi criada uma tabela com o nome DW_Periodo_Horario, contendo um intervalo de tempo que representa um determinado período do dia.

No que diz respeito à tabela de pedidos, esta teve o atributo valor do pedido discretizado em subgrupos de vinte reais cada, representados pela tabela DW_Valor_Pedido.

Alguns pedidos tiveram que ser excluídos porque seus itens não mais existiam na tabela atual de produtos. Isto ocorreu porque o sistema não permitia uma exclusão lógica dos registros destes produtos, mas apenas uma exclusão física sem forçar uma integridade dos dados.

Desta forma, para evitar uma análise com produtos não mais existentes no cardápio, o estudo optou por apagar estes registros tendo em vista a pequena quantidade de pedidos restritos a esta condição. Ao final desta etapa, a tabela de pedidos ficou com um total de 15.802 registros, contabilizando uma redução de quase 3% do volume total.

Como observado, o processo realizado por esta etapa do estudo permitiu um melhor refinamento das informações a serem utilizadas nas análises futuras. Isto possibilitou a redução do volume de dados, tanto pela identificação de registros com informações redundantes, quanto pela exclusão de informações sem relevância.

Além disso, a criação de novas tabelas tornou possível a inserção de novas informações aos dados já existentes, o que permitiu complementar alguns valores registrados a estes dados.

Esta foi sem dúvida a atividade que mais consumiu tempo. Porém, foi a partir dela que se tornou possível a criação de um conjunto de dados com maior qualidade e com um menor nível de incertezas, o que aumentam as chances de busca por um melhor resultado.

4 Estatísticas Básicas

O processo de análise das estatísticas básicas tem por objetivo fazer uma análise exploratória do conjunto de dados que será utilizado no processo de mineração de dados, o que permitirá um melhor entendimento do conjunto de informações que será utilizado neste processo.

Para esta etapa foram utilizados dois programas específicos: o MatLab versão 7.0 e o Microsoft Excel versão 2003. Com o Excel foram criados os gráficos de barras utilizados para a contagem da freqüência de cada variável, já o MatLab foi utilizado para gerar os gráficos auxiliares à esta análise e que serão descritos mais a frente.

4.1 Análise do Conjunto de Dados de Clientes

O estudo começou esta fase a partir dos conjuntos de dados referentes aos clientes, que como dito anteriormente, foram divididos em duas tabelas distintas, sendo: uma tabela para os clientes e ativos e outra para os não ativos.

Nesta fase, o estudo uniu as duas tabelas em um único conjunto de dados a partir de uma *view* criada diretamente no banco de dados. Isto fez com que todos os clientes fossem analisados de uma mesma maneira e com os mesmos atributos. A única diferença desta *view* em relação às tabelas foi a inclusão de um campo para distinguir se um usuário era ativo ou não.

Inicialmente o estudo levou em consideração os seguintes atributos de entrada:

- X1: Idade;
- X2: Tempo de relacionamento – contados trimestralmente;
- X3: Sexo;
- X4: Bairro onde mora o cliente; e
- X5: Loja responsável por atender o cliente.

Já a variável de saída, que possui o valor com o resultado de um determinado cliente, ou seja, se ele é um cliente ativo ou não, é representada por um valor booleano dentro deste mesmo conjunto de dados (0 se for ativo e 1 se for não ativo). Esta será a variável que iremos identificar sua predição no problema de classificação supervisionada a ser realizado mais a frente pelo estudo.

A primeira parte desta etapa do estudo verifica a existência de valores incompletos no conjunto de dados utilizado, que logo de início observou a existência de valores incompletos para o campo loja relacionado ao cadastro do cliente.

Isto se deve ao fato de um determinado cliente morar em uma região não atendida por uma das lojas. Entretanto, todos os usuários com esta situação estão contidos na base de clientes ativos, o que nos leva a crer que seus pedidos foram solicitados para que fossem entregues em uma região válida pela cobertura de entrega. Por conta disso, e para que o MatLab pudesse ler este atributo, todos os registros que continham um valor nulo (*null*) para o atributo loja foram substituídos pelo valor -1.

Além deste campo, o atributo idade apresentou a existência de algumas anormalidades, já que alguns registros tinham valores considerados muito baixos ou muito altos para quem faz um pedido via internet. Exemplos: Clientes com idade igual a 0 (Zero), 1, 2, 99 e assim por diante.

Isto levou o estudo a supor que estes clientes não informaram corretamente o ano de nascimento, seja por um engano ou por uma questão de vaidade. Assim, o estudo aplicou a estes registros o valor médio das idades obtidas em cada uma das tabelas de clientes, fazendo com que fosse reduzida a probabilidade de erro na análise futura por conta destes valores.

Esta regra foi aplicada para todos os usuários com menos de 15 anos ou acima dos 75. Esta escolha foi feita por uma estimativa através da percepção da quantidade de registros que continham valores maiores que 75 e menores que 15.

O passo seguinte foi analisar a distribuição de freqüência dos valores de cada atributo utilizado. O objetivo é examinar a variação dos valores encontrados e com qual freqüência eles ocorrem.

O gráfico do atributo idade (figura 4.1) mostrou uma vasta quantidade de valores distintos para este determinado campo. Isto já era de se esperar, pois este é um campo com valor contínuo.

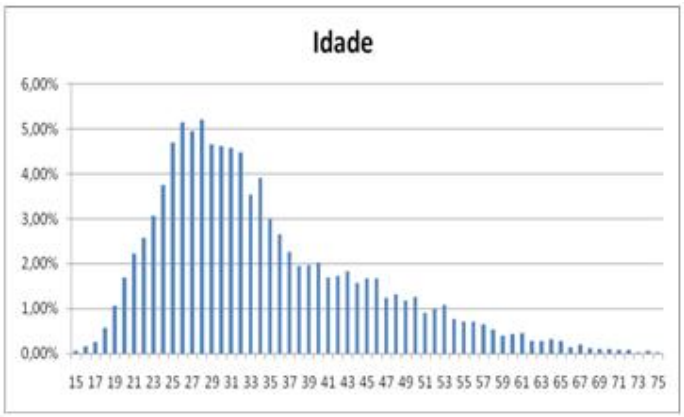


Figura 4.1: Gráfico de freqüência do atributo idade.

Em relação ao gráfico do atributo tempo de relacionamento (figura 4.2), podemos observar que a freqüência do valor cai conforme há o aumento do tempo de relacionamento. A única exceção se deve entre a passagem dos usuários com 1 (um) trimestre de relacionamento em relação aos clientes com 2 (dois) trimestres de relacionamento. Mesmo assim, esta variação é muito pequena, algo em torno de 0,11%. O que indica que a taxa de crescimento da quantidade de clientes é quase sempre positiva.

Figura 4.2: Gráfico de freqüência do atributo tempo de relacionamento.

Quanto ao gráfico do atributo sexo (figura 4.3), temos a quantidade de clientes do sexo feminino um pouco maior do que a quantidade de clientes do sexo masculino.

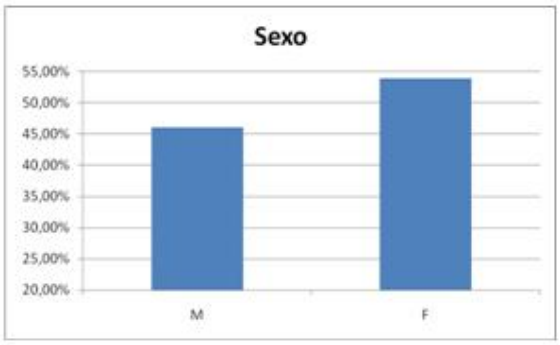


Figura 4.3: Gráfico de freqüência do atributo sexo.

Já o gráfico do atributo bairro (figura 4.4), mostra um alto grau de variabilidade dos seus valores. O que torna difícil a interpretação dos dados a partir de um diagrama de freqüência. Por isso, foi incluído um valor, só para a exibição do gráfico, com o somatório da freqüência de todos os bairros com valor de freqüência menor do que 2%. Este valor é exibido pelo item denominado Outros, que representa um total de 27,67% dos dados.

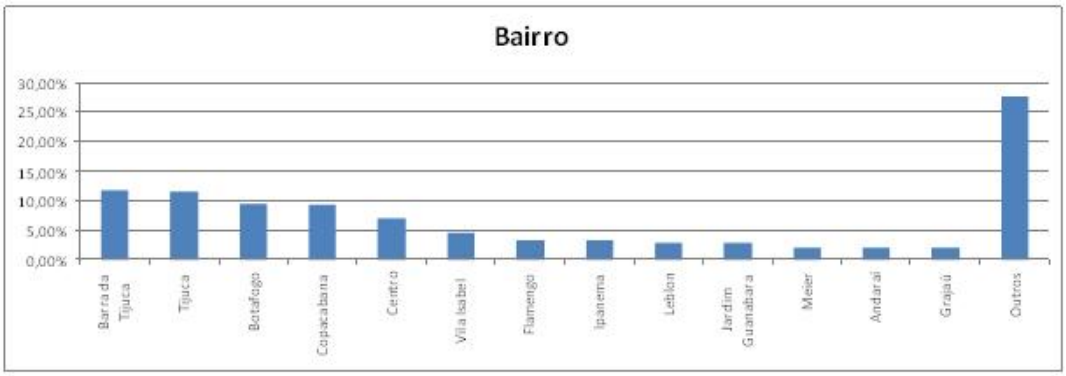


Figura 4.4: Gráfico de freqüência do atributo bairro.

Em seguida, vemos o gráfico do atributo loja (figura 4.5). Este gráfico nos chama atenção pelo fato de que três lojas possuem um total de 42,03% dos clientes cadastrados. Outras cinco possuem seus valores próximos a 6%, o que garante uma boa distribuição dos dados entre elas. Porém o que chamou mais a atenção foi o fato da loja Icaraí ter a menor quantidade de clientes cadastrados.

Após uma profunda análise, o estudo descobriu, junto com analistas de marketing da empresa, que havia uma falha no cadastro dos CEP's que atendiam a esta loja, o que ocasionou esta discrepância em relação aos valores de freqüência das demais lojas.

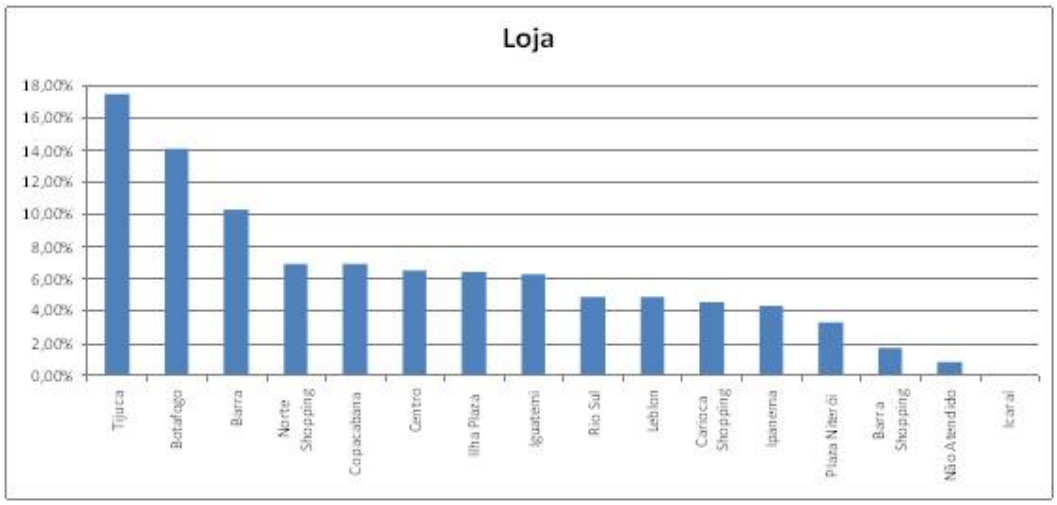


Figura 4.5: Gráfico de freqüência do atributo loja.

Por último, vemos o gráfico que representa o atributo do tipo de cliente (figura 4.6), ou seja, se um cliente é ativo ou não, que como pode ser notado possui um bom balanceamento entre as classes, com um total de 49,11% de clientes ativos e outros 50,89% de clientes não ativos.

Figura 4.6: Gráfico de freqüência do atributo tipo de cliente.

Com isso, observamos a importância desta simples etapa. Pois bastou uma rápida análise dos gráficos de freqüência para o estudo identificar a primeira proposta de melhoria (o acerto do cadastro dos CEPs da loja Icaraí) aos gestores da loja virtual do La Mole.

Após a observação da distribuição de freqüência, o estudo realizou a padronização das variáveis com o objetivo de normalizar seus valores para que a análise dos dados tratasse todas as variáveis de uma forma igual.

O próximo passo realizado pela exploração dos dados foi analisar o gráfico *box-plot*, que tem por objetivo representar os percentis de cada variável e detectar a existência ou não de valores fora da norma padrão, ou seja, valores discrepantes.

Estes valores são comumente chamados de *outliers*, definidos como amostras das quais as mensurações dos parâmetros diferem em comparação com a maioria das outras amostras do conjunto de dados analisado. Normalmente, eles podem representar amostras com características únicas ou como o resultado bruto de um erro analítico [13].

A figura 4.7 mostra o gráfico *box-plot* dos atributos utilizados por esta análise.

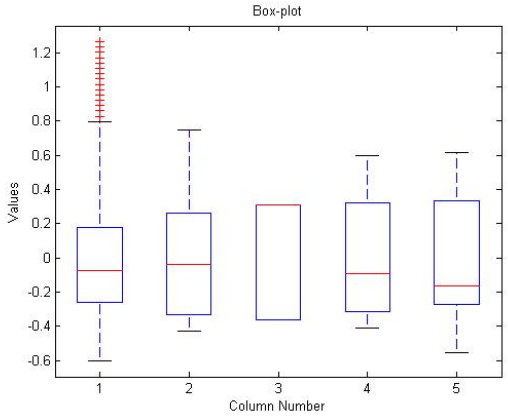


Figura 4.7: Gráfico *blox-pot* das variáveis de entrada.

Podemos notar que o primeiro atributo (idade) contém valores considerados *outliers*, o que não é visto em relação aos demais atributos. Já a variável sexo, por ser uma variável dentro de um intervalo conhecido, ou seja, uma variável discreta com um valor booleano (0 ou 1), não possui dados entre os percentis 25 e 75.

Por conta dos *outliers* encontrados no atributo idade, o estudo substituiu o seu uso pelo atributo faixa etária, com o intuito de reduzir a quantidade de valores aberrantes, uma vez que estes valores podem trazer impactos negativos em relação a qualidade do processo de mineração de dados.

Outra alteração realizada na análise foi a substituição do uso do atributo bairro pelo atributo região. Isto foi feito para que a variação de valores fosse reduzida a um intervalo menor.

Dessa forma, o estudo passou a considerar os seguintes atributos de entrada:

- X1: Faixa etária;

- X2: Tempo de relacionamento trimestral - contados trimestralmente;
- X3: Sexo;
- X4: Região; e
- X5: Loja responsável por atender o cliente.

Estas alterações fizeram com que o estudo voltasse a etapa anterior para analisar o gráfico de freqüências destes novos atributos, apresentados respectivamente pelas figuras 4.8 e 4.9.

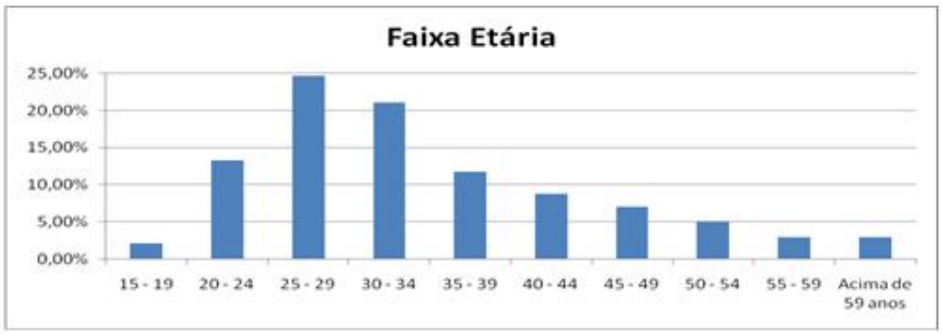


Figura 4.8: Gráfico de freqüência do atributo faixa etária.

Nota-se agora uma melhor visualização dos valores possíveis tanto para o atributo idade, representado de forma discretizada pela faixa etária, quanto para o atributo bairro, representado pela região.

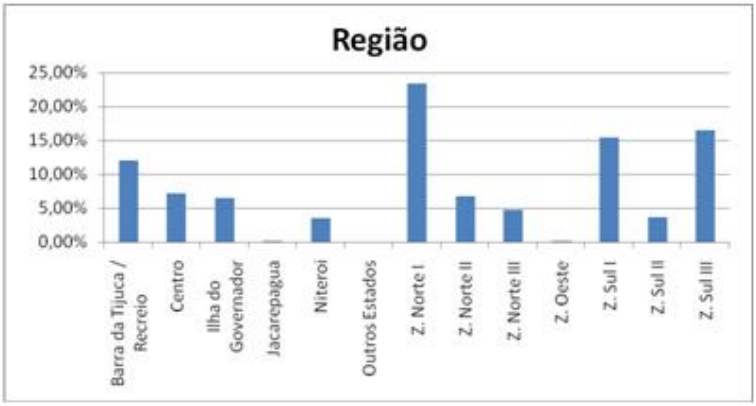


Figura 4.9: Gráfico de freqüência do atributo região.

Além disso, estas mudanças fizeram com o conjunto de dados ficasse sem a existência de *outliers*, como pode ser visto na figura 4.10, o que significa uma base de dados de melhor qualidade.

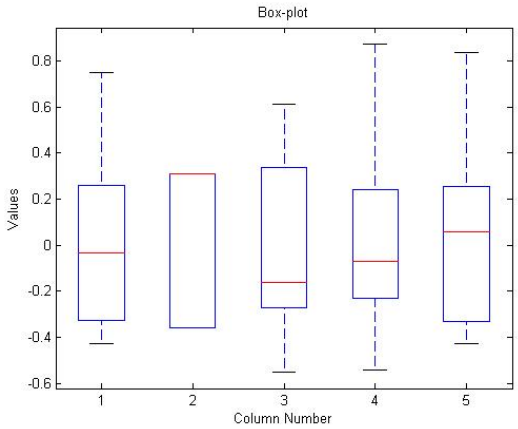


Figura 4.10: Gráfico *blox-pot* do novo conjunto de variáveis de entrada.

O passo seguinte foi verificar o resultado encontrado a partir da matriz de correlação entre todas as variáveis de entradas. Esta matriz pode ser observada através dos valores apresentados na tabela 4.1 e do gráfico exibido na figura 4.11.

	X1	**X2**	**X3**	**X4**	**X5**
X1:	1	-0.0453	-0.0129	0.0554	0.0115
X2:	-0.0453	1	0.0580	-0.0147	0.0360
X3:	-0.0129	0.0580	1	-0.0564	0.2264
X4:	0.0554	-0.0147	-0.0564	1	-0.0115
X5:	0.0115	0.0360	0.2264	-0.0115	1

Tabela 4.1: Tabela de correlação entre as variáveis de entrada utilizadas pelo estudo.

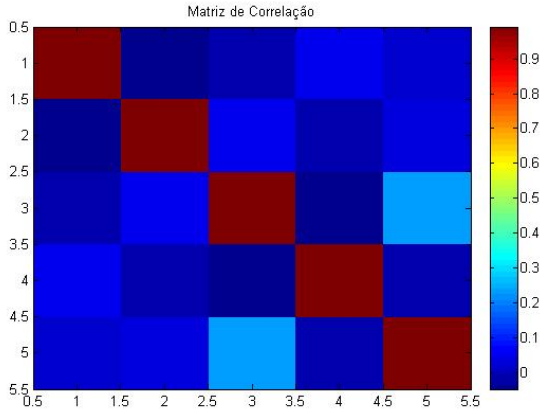


Figura 4.11: Gráfico que contém a matriz de correlação.

Este gráfico funciona como uma forma qualitativa de se mostrar a correlação linear entre as variáveis, que podem ser:

- Positiva, para os casos em que o valor de uma variável aumenta de acordo com o aumento do valor de outra; ou

- Negativa, para os casos em que o valor de uma variável aumenta de acordo com a diminuição do valor de outra.

O valor de correlação entre as variáveis fica compreendido entre um intervalo de -1 a +1, o que significa que quanto mais próximo de -1 maior será o grau de correlação negativa e quanto mais próximo de +1 maior será o grau de correlação positiva.

Dessa forma, tanto pelo modelo gráfico (figura 4.11), quanto pelos valores descritivos (tabela 4.1), podemos notar que o conjunto de dados não possui correlação entre suas variáveis.

Dando continuidade, a figura 4.12 mostra o gráfico de projeção das variáveis descritas anteriormente. Esta modelagem é feita normalmente apenas para variáveis contínuas, mas seu resultado também possui uma boa visualização para variáveis discretas.

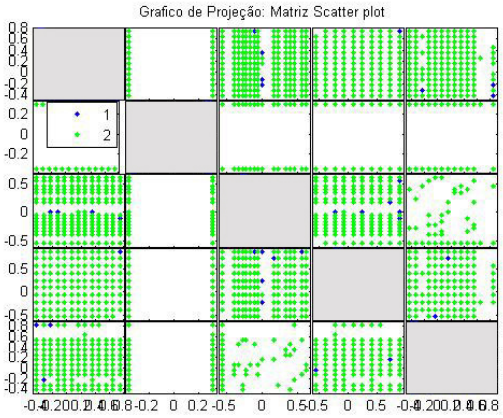


Figura 4.12: Gráfico de projeção dos atributos do conjunto de dados de clientes.

Este gráfico nos ajuda a perceber se algumas das variáveis são linearmente separáveis, o que permite, por exemplo, uma expectativa para a determinação de cada classe em um problema de classificação supervisionada.

A princípio, o resultado obtido mostrou uma grande sobreposição dos dados, não indicando nenhuma variável que possibilite a separação linear entre as classes e nem a existência de um desbalanceamento entre elas. Talvez isto se deva ao volume de dados

utilizado e ao fato de que estamos analisando um gráfico no espaço bidimensional, o que dificulta a interpretação deste gráfico.

Em seguida é avaliado o *data image* (figura 4.13) da distância Euclidiana que contém os registros das variáveis de entrada observados. Esta análise tem por objetivo mostrar a homogeneidade do conjunto de dados, onde cada ponto é a distância entre cada par de registros.

Esta mensuração é feita pela escala de cor apresentada ao lado do gráfico.

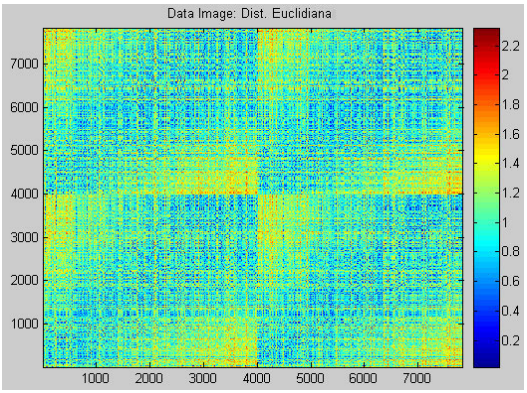


Figura 4.13: *Data Image* da distância Euclidiana.

Podemos observar que alguns pontos são diferentes em relação aos outros, já que não existe a predominância de uma só cor. Há inclusive registros representados pela coloração azul e outros pela coloração vermelha, o que significa a ocorrência de registros em ambos os extremos da escala gráfica.

É possível identificarmos uma pequena dispersão entre os dados, representados na escala gráfica pelos pontos com as cores que variam do amarelo em direção ao vermelho e pelos pontos com as cores que variam do azul claro para o azul escuro.

Entretanto, a ligação entre a escala gráfica de cor amarela com a cor azul claro é feita pela cor verde, que pode ser vista com bastante freqüência no gráfico em questão. Esta relação serve para indicar que, mesmo existindo pontos distintos no conjunto de dados, há uma forte homogeneidade no conjunto de dados.

De uma maneira geral podemos dizer que a análise das estatísticas básicas para o conjunto de dados de clientes foi bastante produtiva, pois permitiu um melhor entendimento dos dados, aumentando ainda mais a qualidade das informações a serem trabalhadas futuramente pelo estudo.

Além disso, foi possível obter uma visão geral do conjunto de dados como um todo, identificando que suas partes não apresentam grandes desigualdades. Isto significa um conjunto de dados mais homogêneo, como descrito acima, o que indica que não será uma fácil tarefa segmentar e criar um modelo classificador para este conjunto de dados. Porém, esta homogeneidade também traz uma vantagem, já que o balanceamento dos dados nos esboça que provavelmente não teremos um modelo de dados tendencioso.

Feito isto, o estudo dá por concluído a análise das estatísticas básicas relacionadas ao conjunto de dados de clientes.

4.2 Análise do Conjunto de Dados de Pedidos

O estudo prossegue com a análise das estatísticas básicas relacionada ao conjunto de dados dos pedidos. Neste primeiro momento, o estudo levou em consideração os seguintes atributos de entrada:

- X1: Loja responsável pelo pedido;
- X2: Dia da semana em que o pedido foi realizado;
- X3: Se o pedido foi realizado na data de um feriado;
- X4: O horário do pedido;
- X5: Região de entrega do pedido realizado.

Quanto ao atributo de saída, este é representado pelo campo que contém o valor do pedido de forma discretizada, o que possibilita a classificação de cada registro do conjunto de pedidos.

Na análise deste conjunto de dados não foi encontrado a existência de valores ausentes. Em seguida o estudo analisou a distribuição de freqüência dos valores de cada atributo.

No gráfico do atributo loja (figura 4.14), podemos ver que a distribuição é bem parecida com a encontrada no gráfico que contém a distribuição de usuários cadastrados por loja. Inclusive, a ordem das três primeiras lojas com a maior quantidade de clientes é exatamente igual a ordem das três primeiras lojas com a maior quantidade de pedidos.

Em relação às demais, vemos uma ligeira troca de posições entre as lojas Copacabana, Centro, Norte Shopping e Ilha Plaza, mas ambas possuem uma semelhança na representação de freqüência dos pedidos em relação a freqüência dos clientes cadastros.

Já a loja Iguatemi, que tinha um número percentual de freqüência de clientes cadastrados semelhante aos destas lojas, teve uma redução na freqüência de pedidos se comparado com a freqüência dos cadastros de clientes. Esta foi a maior mudança de posição entre as lojas quando comparados os gráficos dos clientes com dos pedidos.

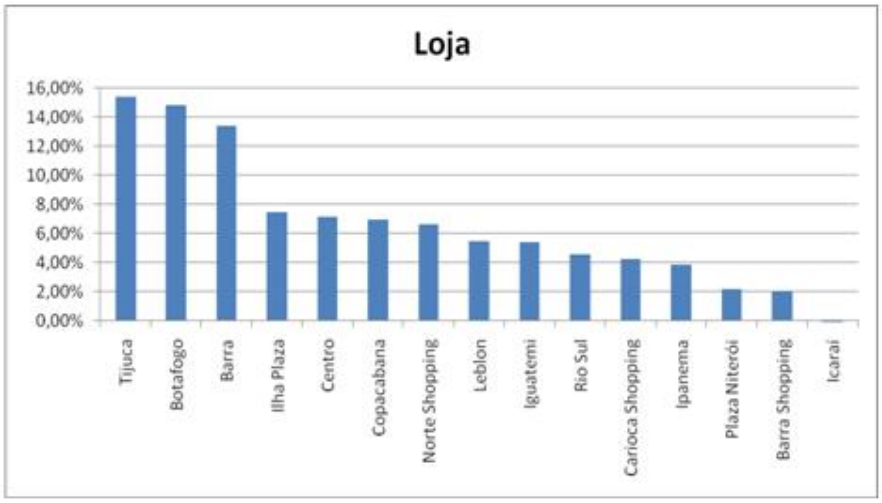


Figura 4.14: Gráfico de freqüência do atributo região.

O gráfico seguinte (figura 4.15) diz respeito ao dia da semana em que o pedido foi realizado. Nele, podemos observar uma boa distribuição dos dados, ressaltando apenas que o domingo é o dia mais utilizado para se fazer pedidos.

Figura 4.15: Gráfico de freqüência do atributo dia da semana.

Em seguida, vemos o gráfico do atributo feriado (figura 4.16), que contém uma freqüência muito maior para os pedidos realizados nos dias comuns do que os realizados nos feriados. Isto já era de se esperar, pois a quantidade dos dias que representam os feriados é muito menor do que a quantidade dos dias comuns.

Figura 4.16: Gráfico de freqüência do atributo feriado.

O próximo gráfico (figura 4.17) representa a distribuição de freqüência do período do horário de realização dos pedidos. Nele, é visualizado que o período mais utilizado é o do almoço, que contém uma freqüência um pouco maior do que a do período do jantar. Já o período da tarde contém uma freqüência bem menor do que a dos outros períodos.

Figura 4.17: Gráfico de freqüência do atributo horário do pedido.

Logo depois é analisado o gráfico de freqüência do atributo região (figura 4.18). Este gráfico é bem semelhante ao encontrado na análise do conjunto de dados dos clientes, o que mostra uma grande variabilidade dos seus valores.

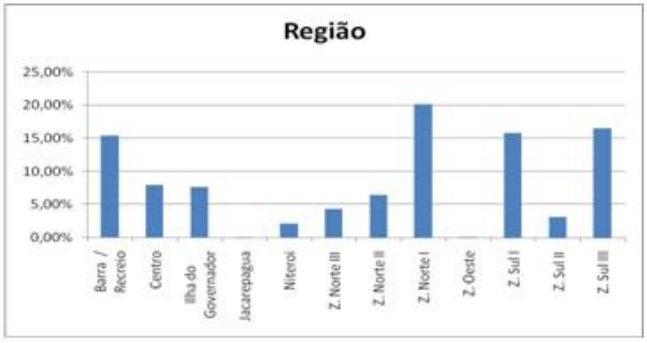


Figura 4.18: Gráfico de freqüência do atributo região.

Por último é analisado o diagrama de freqüência do atributo de saída (figura 4.19). Nele é possível observamos que 31,58% dos pedidos possuem um valor menor ou igual a R$ 20,00 e que outros 45,67% possuem um valor maior do que R$ 20,00 e menor ou igual a R$ 40,00. Os demais valores representam um pouco mais do que 20% dos dados.

Figura 4.19: Gráfico de freqüência do atributo valor do pedido.

O próximo passo desta etapa foi padronizar os valores contidos nos atributos do conjunto de dados de pedidos. A partir desta padronização o estudou gerou o gráfico *box-plot* destes dados, que pode ser visto na figura 4.20.

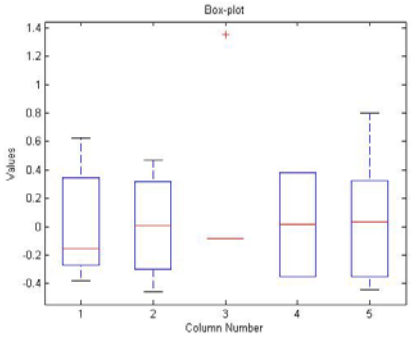


Figura 4.20: Gráfico *blox-pot* das variáveis do conjunto de dados dos pedidos realizados.

Podemos notar que a variável X3 (feriado) é a únicas que possui *outliers*, isto já era de se esperar, pois, mesmo sendo um atributo binário, a quantidade de dias que representam os feriados é muito pequena. O fato de ser um atributo binário justifica a ausência de valores entre os percentis 25 e 75.

Em seguida, o estudo passou a analisar o resultado encontrado pela matriz de correlação entre todas as variáveis utilizadas. Este resultado pode ser observado a partir dos valores contidos na tabela 4.2 e do gráfico representado na figura 4.21.

	X1	**X2**	**X3**	**X4**	**X5**
X1:	1	0.0271	-0.0152	-0.0691	0.1798
X2:	0.0271	1	-0.0697	0.0908	0.0097
X3:	-0.0152	-0.0697	1	-0.0176	-0.0372
X4:	-0.0691	0.0908	-0.0176	1	-0.1599
X5:	0.1798	0.0097	-0.0372	-0.1599	1

Tabela 4.2: Tabela de correlação entre as variáveis utilizadas nos dados dos pedidos.

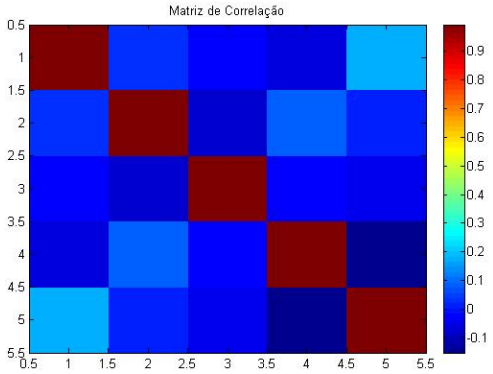


Figura 4.21: Gráfico da matriz de correlação do conjunto de dados dos pedidos.

Observamos tanto pelos valores descritivos da tabela 4.2, quanto pela exibição gráfica da figura 4.21, que não foi possível encontrar alguma correlação significativa entre as variáveis utilizadas.

Em seguida, o próximo passo foi analisar o gráfico de projeção das variáveis dos pedidos realizados, exposto pela figura 4.22.

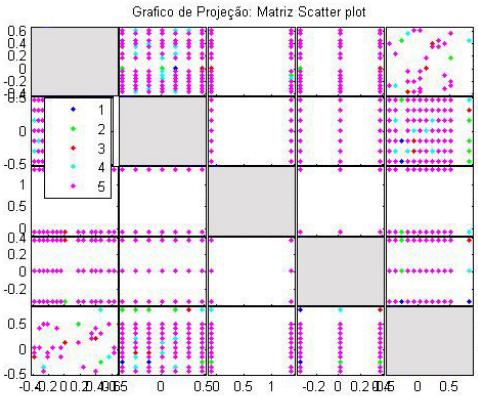


Figura 4.22: Gráfico de projeção dos atributos do conjunto de dados de pedidos.

O resultado obtido mostrou o mesmo problema encontrado na análise de clientes, ou seja, uma grande sobreposição dos dados, o que dificulta a interpretação deste gráfico ao não permitir avaliar se um determinado atributo possui influência em relação a outro.

Para este conjunto de dados o estudo não realizou uma análise do *data image* da distância Euclidiana devido a elevada quantidade de registros.

Como um todo, a análise das estatísticas básicas deste conjunto de dados serviu para termos uma idéia geral da qualidade das informações a serem utilizadas futuramente.

Assim, esta etapa do estudo cumpriu o seu papel ao auxiliar o entendimento das informações contidas na base de dados de pedidos, que em conjunto com a base de clientes, serão utilizados nas próximas etapas do trabalho.

5 Segmentação dos Clientes

5.1 Introdução

Embora a análise das estatísticas básicas tenha nos fornecido uma visão geral a respeito do conjunto de dados de clientes, que como sabemos é dividido entre os clientes ativos e os não ativos, esta análise não define em quantos subgrupos podemos dividir estes dados e nem quais são as características que distinguem cada subgrupo.

Por isso, o presente trabalho utilizou o método de classificação não supervisionada *K-means* para poder entender de que forma é constituído este conjunto de dados. O principal objetivo é poder identificar possíveis padrões de comportamento, que estejam ocultos nos dados, a partir dos valores contidos em seus atributos.

Embora seja relativamente simples e rápido, o algoritmo *K-means* não traz nenhuma garantia quanto ao melhor resultado a ser obtido, já que a quantidade de centros a serem trabalhados pelo modelo deve ser definida previamente. Entretanto, é possível encontrar alguns estudos que procuram estimar de forma eficiente o valor de inicialização da quantidade de grupos, como pode ser visto nos trabalhos realizados por Pelleg e Moore [12] e Su e Dy [14].

De qualquer modo, a ferramenta utilizada não possui nenhum mecanismo para efetuar esta tarefa, o que fez com que o estudo rodasse o algoritmo várias vezes, a partir de distintos valores de parametrizações, a fim de garantir um melhor resultado.

5.2 Análise e Desenvolvimento

Para esta análise o conjunto de dados de clientes foi unificado por uma *view*, de modo que todos os registros de clientes ativos e não ativos fossem analisados em conjunto. Esta *view* contém os seguintes atributos: Se o cliente é ativo ou não (tipo de cliente), faixa etária, loja responsável pelo cliente, região onde mora, sexo e tempo de relacionamento.

Como ponto de partida, a primeira análise foi executada com os valores padrões de parametrização utilizados pelo SQL Server 2008 para o algoritmo *K-means*. A idéia

era ter uma noção primária de como seria a distribuição dos clusters, para em seguida rodar o algoritmo por diversas vezes a fim de se chegar a um melhor modelo.

Por padrão o parâmetro quantidade de clusters é definido com o valor igual a 10 (dez), ou seja, estamos fazendo com que o conjunto de dados de clientes seja dividido em dez clusters de acordo com as semelhanças a serem encontradas pelo algoritmo em questão.

O resultado encontrado pela primeira avaliação realizada pelo estudo pode ser visualizado através da figura 5.1, que representa o perfil de cada cluster encontrado pelo modelo. Nesta figura é possível a visualização de todos os atributos utilizados na análise, assim como a distribuição dos valores de cada um destes atributos em cada cluster gerado.

A primeira coluna desta imagem representa a descrição de cada atributo utilizado. Já a segunda coluna representa a legenda com os valores possíveis para cada atributo, sendo apresentados: os quatro mais populares e um último que representa todos os outros valores possíveis.

Na terceira coluna é possível visualizarmos a representação da distribuição dos dados de cada atributo. Esta coluna basicamente representa a distribuição de freqüência observada nos histogramas produzidos na análise das estatísticas básicas. A partir da quarta coluna, a figura exibe as características de cada cluster gerado.

Junto ao topo de cada coluna é exibida a informação com a quantidade de registros contidos em cada um dos clusters encontrados, o que auxilia no processo de avaliação do resultado obtido.

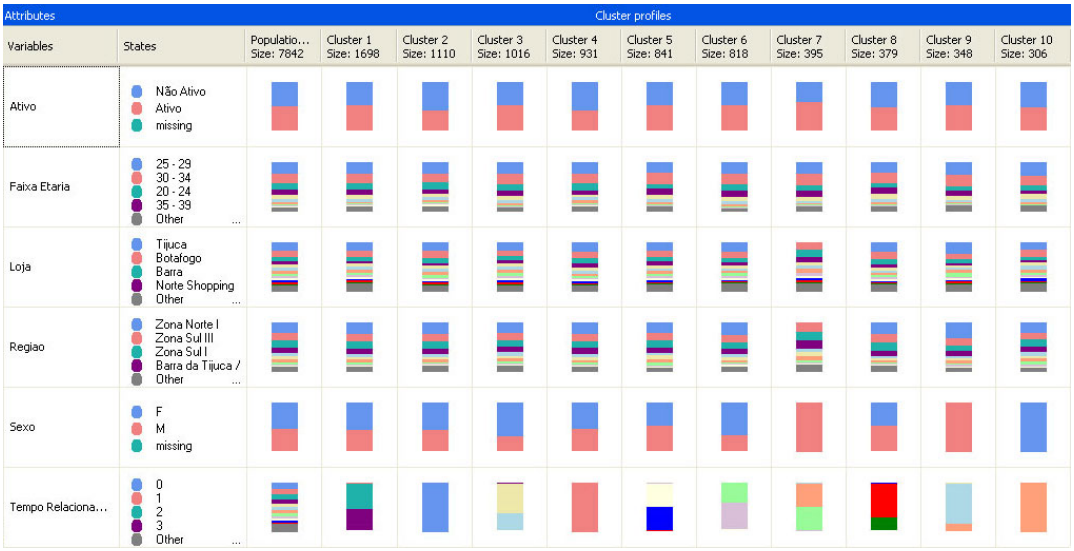


Figura 5.1: Gráfico de perfis dos clusters encontrados pelo algoritmo *K-means*.

Podemos observar que em todos os clusters há um equilíbrio na distribuição dos valores possíveis para os atributos tipo de cliente e faixa etária. Este equilíbrio segue a distribuição encontrada na avaliação de todo o conjunto de dados (terceira coluna da figura).

Esta mesma tendência é encontrada para os atributos loja e região, ressaltando-se apenas que o cluster de número 7 se difere um pouco dos demais, já que em sua distribuição não é possível visualizarmos registros da loja Tijuca e nem da região Zona Norte I. Porém, excluindo-se esta loja e região, a ordem da distribuição dos registros é a mesma encontrada nos demais clusters.

Em relação ao campo sexo, é notado que alguns clusters já começam a ser distinguidos a partir do valor encontrado neste atributo. Este fato pode ser observado nas colunas que representam os clusters de número 7, 9 e 10.

Quanto ao atributo tempo de relacionamento, este foi o mais utilizado pelo algoritmo para a diferenciação dos clusters. É nele que podemos acompanhar a maior variação da distribuição dos valores entre cada um dos clusters encontrados.

Embora seja o principal recurso para a visualização das características que formam cada cluster, o gráfico de perfil não é o único instrumento utilizado para o entendimento do resultado encontrado. Dois outros recursos são muito importantes: Um é o diagrama de cluster, que tem por objetivo exibir a proximidade entre cada cluster e

suas respectivas ligações; já o outro, é o chamado discriminante de cluster, que ajuda a determinar os atributos que diferenciam dois clusters.

A vantagem de se utilizar o diagrama de cluster é que ele possibilita uma navegação gradual entre os diversos níveis de ligações entre cada cluster. Isto é feito pela exibição de links que indicam cada relacionamento existente. Cada nível está relacionado com o grau de semelhança entre um determinado cluster e outro. Este recurso está disponível a partir de uma barra de rolagem contida ao lado do diagrama em questão.

Nesta primeira análise, o nível mais baixo já indica uma forte ligação entre os clusters de número 1 e 4, visualizada a partir da figura 5.2. Estes dois clusters possuem uma semelhança muito grande entre si, como pode ser observado na figura 5.1.

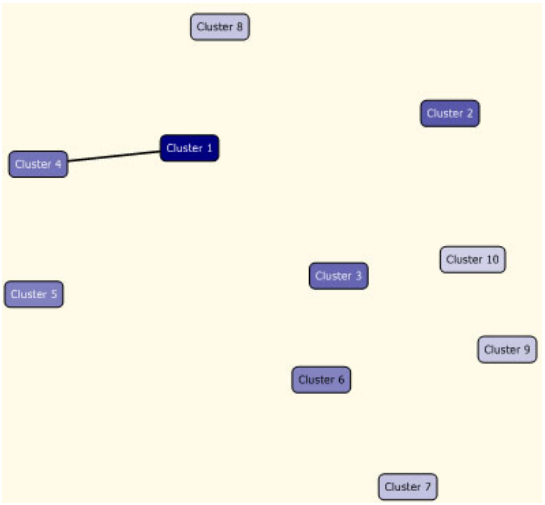


Figura 5.2: Diagrama de cluster que mostra o primeiro nível de semelhança entre os clusters encontrados pelo modelo.

Ambos possuem como diferencial, entre um e outro, os valores contidos nos atributos tempo de relacionamento, conforme demonstra o resultado da figura 5.3, que representa o discriminante entre estes dois clusters.

Discrimination scores for Cluster 1 and Cluster 4

Variables	Values	Favors Cluster 1	Favors Cluster 4
Tempo Relacionamento Trimestral	1		
Tempo Relacionamento Trimestral	2		
Tempo Relacionamento Trimestral	3		

Figura 5.3: Visualização gráfica dos atributos e valores que distinguem o cluster 1 do cluster 4.

Buscando avaliar os diversos níveis de semelhança entre os clusters, logo é possível a visualização de mais dois links: um que liga os clusters 3 e 9 e outro que liga os clusters 6 e 10, como é exibido na figura 5.4.

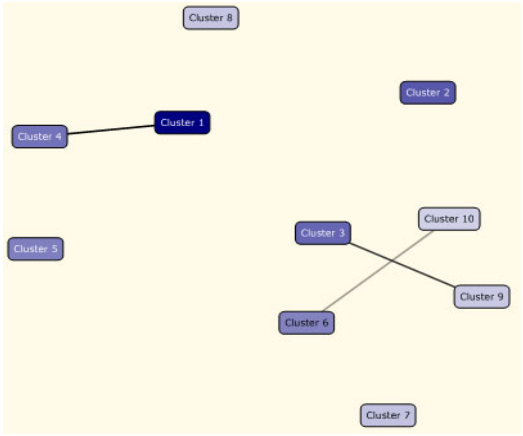


Figura 5.4: Diagrama de cluster que mostra um nível de semelhança acima do encontrado pela figura 5.2.

Podemos notar pela figura 5.1, que para os clusters 3 e 9 o valor do campo sexo possui uma forte distinção entre estes clusters, assim como os valores encontrados a partir do atributo tempo de relacionamento. Para um melhor entendimento sobre os valores e atributos discriminantes destes clusters, recorremos a análise observada na figura 5.5.

Discrimination scores for Cluster 3 and Cluster 9

Variables	Values	Favors Cluster 3	Favors Cluster 9
Sexo	F		
Sexo	M		
Tempo Relacionamento Trimestral	4		
Tempo Relacionamento Trimestral	5		
Tempo Relacionamento Trimestral	6		
Tempo Relacionamento Trimestral	3		

Figura 5.5: Atributos e valores que distinguem o cluster 3 do cluster 9.

Desta forma é possível atentarmos que o atributo sexo com valor igual a feminino é determinante para o favorecimento do cluster 3, assim como o valor igual a masculino é determinante para o cluster 4. Neste caso o valor do atributo sexo é utilizado em conjunto com o valor do atributo tempo de relacionamento para a distinção entre estes dois clusters. Este exemplo ajuda a percebermos a importância de cada recurso oferecido pela ferramenta empregada no estudo.

Apresentados os recursos fornecidos pela ferramenta, em conjunto com a primeira análise realizada, o estudo segue agora com objetivo de encontrar um melhor resultado para a segmentação da base de dados de clientes. Deste modo, o estudo avançará na geração e avaliação de diversos modelos a partir de diferentes valores de parametrização para a variável quantidade de clusters.

O estudo passa agora a analisar o impacto gerado a partir do aumento ou diminuição do valor deste parâmetro. Para isso, foram criadas diversas análises a partir da variação do valor do parâmetro quantidade de clusters, que para o caso em questão, teve sua variação analisada entre os valores de 5 até 15 clusters.

Embora o estudo tenha feito várias análises para verificar o resultado desta variação, apenas os dois resultados mais significantes, que serviram de base para as conclusões deste tipo de análise, são apresentados. A próxima análise demonstra o resultado obtido através do valor igual a 5 (cinco) para o parâmetro quantidade de clusters, como pode ser visto na figura 5.6.

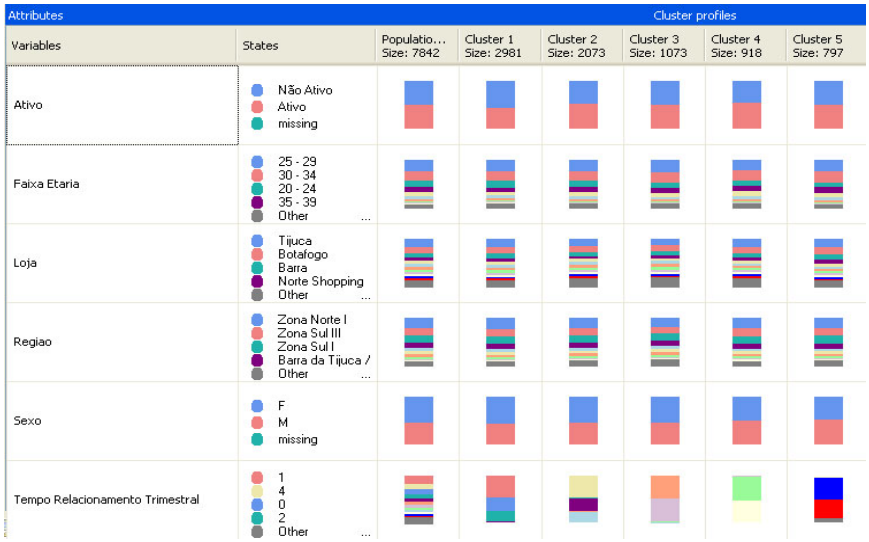


Figura 5.6: Resultado obtido pelo *K-means*, com o valor de parametrização igual a 5.

Foi observado pelo estudo que o resultado obtido ficou ainda mais prejudicado com a redução da quantidade de clusters. Isto já era de se esperar, uma vez que ao juntarmos os registros em uma menor quantidade de subgrupos faz com que estes registros fiquem cada vez mais genéricos em relação a todo o conjunto de dados. Este resultado mostrou como o atributo tempo de relacionamento é o mais utilizado para separar os dados em vários clusters.

Na análise anterior observamos que em alguns casos o atributo sexo já indicava um nível de separação entre os clusters. Porém, nesta análise, nenhum subconjunto de dados foi distinguido por algum outro atributo que não tenha sido o tempo de relacionamento.

Outro ponto relevante é o fato de que as distribuições de todos os atributos, exceto o tempo de relacionamento, seguem a mesma ordem de distribuição do conjunto de dados. Isto indica que não estamos tendo uma boa qualidade no resultado obtido, já que é identificado apenas um padrão referido ao tempo de relacionamento, o que não traz nenhuma informação relevante a respeito das características dos clientes.

No auxílio do entendimento dos clusters gerados, o estudo analisou o diagrama de cluster entre os subconjuntos de dados com as maiores semelhanças entre si. Este diagrama pode ser visualizado na figura 5.7, e demonstra a proximidade entre os clusters de número 4 e 5.

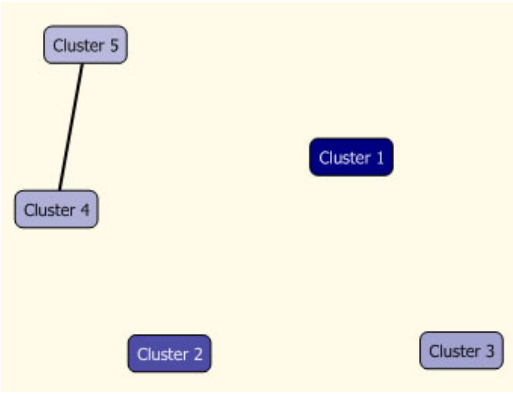


Figura 5.7: Diagrama de cluster com os clusters mais semelhantes da segunda análise.

Apesar de ser explicitamente visualizado na figura 5.6, o estudo analisou o atributo discriminante entre estes dois clusters, que pode ser visto na figura 5.8, que

confirmar o fato de que apenas o atributo tempo de relacionamento é utilizado para diferenciar os clusters.

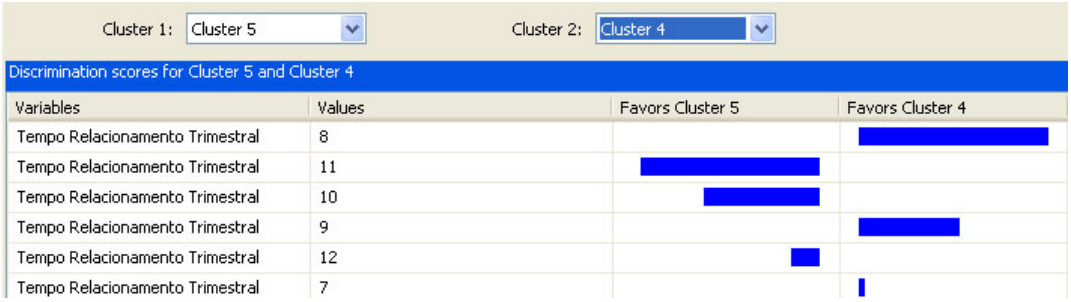


Figura 5.8: Atributos e valores que distinguem o cluster 4 do cluster 5.

Agora o estudo demonstra o resultado encontrado a partir do valor igual a 15 (quinze) para o parâmetro quantidade de clusters, como pode ser acompanhado pela figura 5.9. Com esta análise é possível observar que o aumento na quantidade de clusters proporciona um aumento na qualidade dos mesmos.

O resultado obtido mostra a primeira existência de clusters formados apenas com clientes ativos ou não, como é o caso dos clusters 4, 7 e 15. Também é possível notar que há um aumento da quantidade de clusters formados pela distinção do campo sexo, além de um maior número de clusters formados por um único valor para o atributo tempo de relacionamento.

Ademais, podemos tomar em consideração que todos os clusters apresentam a mesma ordem de distribuição para os atributos faixa etária, loja e região, o que mais uma vez dificulta a interpretação das características que compõem um determinado cluster.

Este caso nos chama a atenção se compararmos o resultado encontrado nesta análise com os da primeira (figura 5.1), onde pode ser observado que o cluster de número 7 não possui registros contidos na loja Tijuca e nem na região Zona Norte I, o que o diferencia se comparado aos demais.

Desta forma, podemos inferir que o conteúdo deste cluster foi dissipado entre os diversos clusters criados na análise atual. Este resultado nos possibilita um melhor entendimento do conjunto de dados, uma vez que nos permite uma segregação maior entre os diversos subconjuntos encontrados.

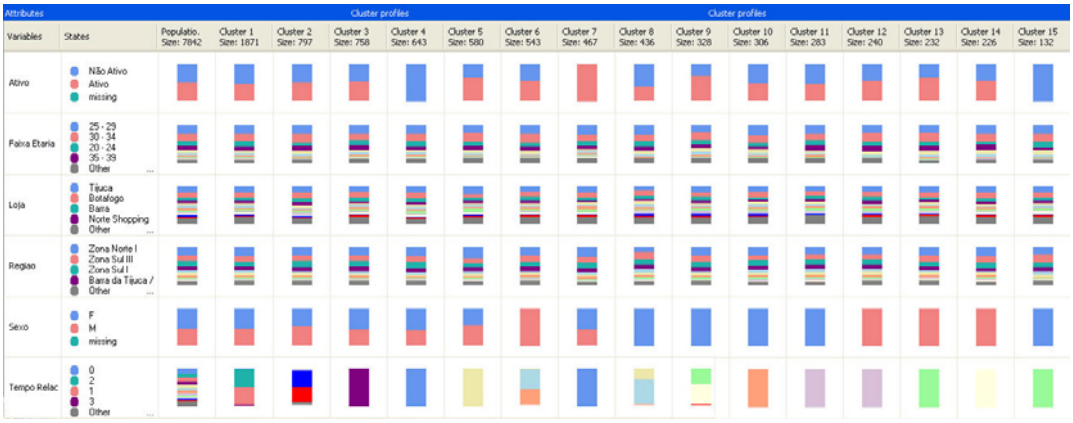


Figura 5.9: Resultado obtido pelo *K-means*, com o valor de parametrização igual a 15.

Como exemplo, podemos citar o cluster de número 4 que contém apenas registros de clientes não ativos e com um tempo de relacionamento trimestral igual a 0. Este cluster levanta a suspeita de que os clientes mais recentes representam um padrão de comportamento para os clientes não ativos.

Porém, esta suspeita é logo descartada ao analisarmos o cluster de número 7, que contém apenas usuários ativos e que possuem um tempo de relacionamento trimestral igual 0, ou seja, o cluster7 praticamente anulou a possibilidade de identificação de um padrão a partir do cluster 4, para o qual esperávamos encontrar um comportamento que distinguisse um cliente não ativo dos demais.

Do cluster 8 ao cluster 11 podemos observar apenas registros do sexo feminino, tendo somente variações no valor do atributo tempo de relacionamento. Em contrapartida, do cluster 12 ao cluster 14 podemos visualizar que estes possuem apenas registros do sexo masculino, variando também apenas o valor contido para o atributo tempo de relacionamento.

Para ajudar no processo de entendimento dos clusters criados, o estudo recorreu mais uma vez a análise do diagrama de cluster, que pode ser visualizado logo em seguida a partir da figura 5.10.

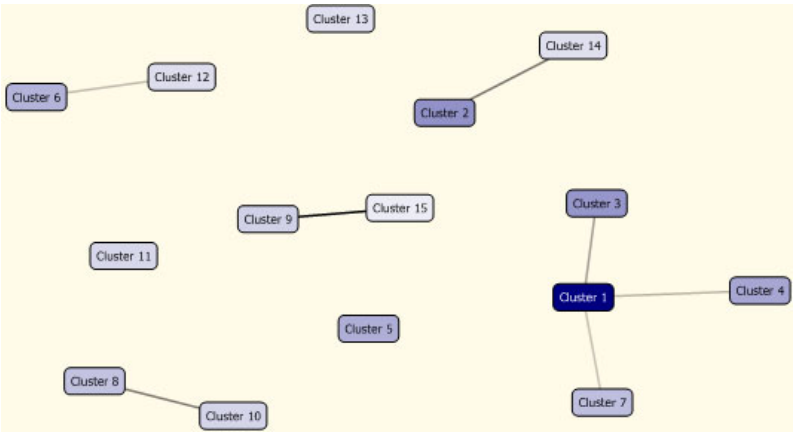


Figura 5.10: Diagrama de cluster com os clusters mais semelhantes da terceira análise.

Neste gráfico podemos visualizar que o cluster 2 é bem semelhante ao cluster 14, embora este último possua apenas registros do sexo masculino e que tenham o valor tempo de relacionamento igual a 9.

A relação entre o subconjunto 8 e 10 e a relação entre o subconjunto 9 e 15 indicam que estes clusters são relacionados pelos registros que contém apenas o atributo sexo igual a feminino, mas com variações distintas para o atributo tempo de relacionamento. Já a relação entre os clusters de número 6 e 12 é feita a partir do valor masculino, mantendo a variação a partir do valor contido no atributo tempo de relacionamento.

Estas três análises realizadas nos mostram que a segmentação dos dados está sendo bastante influenciada pelo atributo tempo de relacionamento, o que vem trazendo um impacto bastante indesejado a análise dos dados. Isto tem gerado resultados de baixa qualidade e que não ajudam na identificação de um comportamento que determine uma característica específica para separar os clientes ativos dos não ativos.

Com isso, é feita uma primeira conclusão a respeito da análise de classificação não supervisionada realizada pelo estudo. A de que o tempo de relacionamento não possui efeito algum no fato de um cliente ser ativo ou não. A razão para termos utilizado este atributo era poder examinar se em um determinado período houve alguma causa especifica para que os clientes se cadastrassem na loja virtual e não realizassem pedidos.

Esta conclusão nos ajuda a responder algumas perguntas do tipo: Será que as pessoas que não compram são os clientes mais antigos? Será que no início do

funcionamento da loja virtual os clientes, por algum motivo, encontravam alguma dificuldade ou falta de segurança para efetuar um pedido? Será que são os clientes mais recentes que não fazem pedidos? Será que há algo hoje que esteja falhando na hora de fechar um pedido. Vale ressaltar que a loja virtual sofreu uma reformulação no terceiro trimestre de 2007.

Todas estas perguntas são respondidas de forma negativa, já que não é encontrado nenhum período específico com uma grande quantidade de clientes ativos ou não ativos.

Como este atributo está causando um forte efeito sobre a análise gerada, o estudo passará a avaliar o conjunto de dados sem este atributo daqui por diante. Deste modo, o estudo continuará a busca por algum padrão que especifique as características relacionadas ao conjunto de dados de clientes e que, de preferência, responda o que leva a um cliente ser ativo ou não.

Para tentar melhorar um pouco mais a qualidade da análise, o estudo passará a examinar o conjunto de clientes separado por cada região, considerando apenas as dez regiões com mais usuários cadastrados, uma vez que as outras regiões somadas representam uma população menor 0,5% dos clientes.

O objetivo de todo esse esforço é tentar identificar as características dos clientes de acordo com uma região e loja, a partir de um maior refinamento das informações, auxiliando o entendimento dos diversos comportamentos que formam cada subgrupo de cliente. Assim, quem sabe o estudo não possa encontrar um perfil de cliente que tanto procura.

Outro ponto importante é que, a partir de agora, as análises realizadas levarão em consideração o atributo bairro, já que a variação deste campo por região é bem menor se comparado a variação encontrada em todo o conjunto de dados.

Como sabemos, para se obter um melhor resultado através do algoritmo K-means, em muitas ocasiões, se faz necessário processar o algoritmo por várias vezes. Embora o estudo assim tenha o feito, ao executar o método em questão com diferentes valores de parametrizações, será apresentado a partir de agora apenas o resultado final determinado pelo estudo para cada conjunto de dados, por região.

5.2.1 Segmentação dos Clientes da Região Zona Norte I

A primeira análise realizada foi efetuada com a região definida por Zona Norte I, que possui o maior número de clientes cadastrados, um pouco mais do que 23% do total de registros, e que abrange uma área próxima ao bairro da Tijuca. Esta região é atendida por duas lojas (Tijuca e Iguatemi) e contém os seguintes bairros além da própria Tijuca: Vila Isabel, Andaraí, Grajaú, Maracanã, Rio Comprido, Praça da Bandeira, Alto da Boa Vista e Mangueira. Há outros bairros contidos nesta região, porém o estudo só levará em conta os que estão cadastrados na base de clientes.

O resultado final obtido para esta região pode ser visualizado na figura 5.11. Este resultado foi gerado com um valor de parametrização da quantidade de clusters a ser encontrado igual a 15 (quinze), porém é verificado que o programa agrupou os dados em apenas 13 clusters.

Isto se deve ao fato de que o modelo chegou a um grau de convergência mínimo especificado pelo programa sem a necessidade de se criar todos os clusters estabelecidos. Este valor é definido pelo atributo *STOPPING_TOLERANCE* no momento em que se configura os parâmetros do método *K-means* pelo SQL Server.

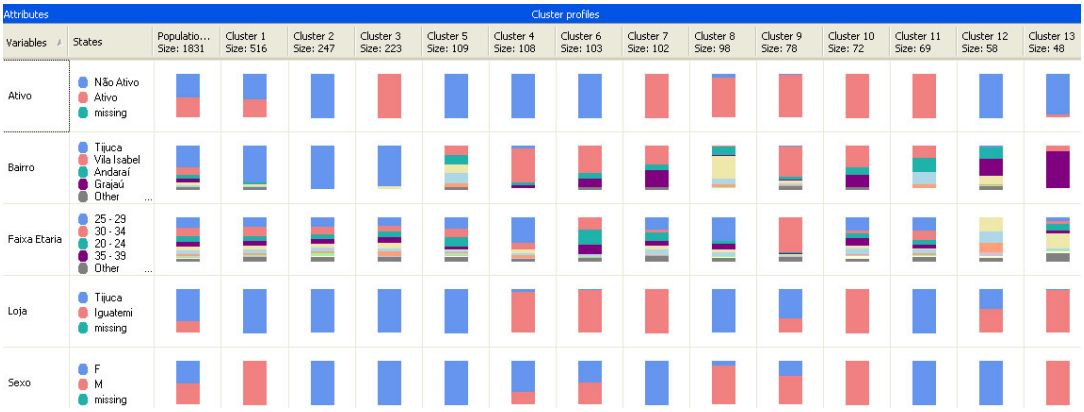


Figura 5.11: Segmentação dos clientes da região Zona Norte I.

A análise realizada para esta região mostrou-se bastante sensível ao atribuo tipo de cliente (ativo ou não), podendo ser visualizado na maioria dos clusters gerados pelo modelo.

O primeiro cluster, o que contém o maior número de registros, é formado especificamente por clientes do sexo masculino, moradores da Tijuca e que são

atendidos pela loja do próprio bairro. Neste cluster, não foi possível encontrar uma definição quanto a faixa etária dos clientes e nem em relação a informação de que um cliente deste cluster é ativo ou não. Isto faz com que este cluster não traga uma informação muito relevante ao estudo.

Em compensação, se comparamos os clusters 2 e 3 veremos que os mesmos são quase idênticos, sendo diferenciados apenas pela informação de que um contém os clientes não ativos e outro os ativos.

Embora tenhamos conseguido identificar plenamente este atributo, a análise entre estes dois clusters não nos traz um conhecimento específico quanto ao comportamento de compra do usuário, já que temos um mesmo perfil para cada tipo de cliente, ou seja, um cluster anula o padrão identificado pelo outro na determinação do tipo de cliente.

Por conta desta análise, o estudo pode inferir que nem todo cluster trará um conhecimento específico para a avaliação da informação que queremos buscar, como é o caso destes três clusters já avaliados. Por isso, o estudo criou na tabela 5.1 uma lista com o conteúdo dos clusters mais significativos em termos de conhecimento das características dos clientes observadas por este modelo.

Cluster	**Tipo de Cliente**	**Bairro**	**Faixa Etária**	**Loja**	**Sexo**
9	Ativo	Vila Isabel	30 – 34	Tijuca	M
8	Ativo	Maracanã	25 – 29	Tijuca	M
4	Não Ativo	Vila Isabel	25 – 29	Iguatemi	F
12 e 13	Não Ativo	Grajaú Andaraí	40 – 44 45 – 59 50 – 54	Iguatemi	M F

Tabela 5.1: Conhecimento intrínseco obtido pela segmentação da região Z. Norte I.

O atributo faixa etária possui um comportamento muito interessante na definição do tipo de usuário quando analisado em conjunto com o bairro, loja de entrega e o sexo. Como exemplo, podemos avaliar os usuários com faixa etária entre 25 – 29 anos, em que os ativos correspondem a: morar no bairro Maracanã, ser do sexo masculino e ser atendido pela loja Tijuca, e os não ativos correspondem a: morar no bairro Vila Isabel, ser atendido pela loja Iguatemi e ser do sexo feminino.

Outros dois clusters trouxeram um comportamento a respeito dos clientes não ativos, como pode ser observado na linha da tabela 5.1 que representa os clusters 12 e 13. Nele é possível observar grupos de clientes formados por pessoas que moram nos bairros Grajaú e Andaraí, e que a medida que possuem um aumento na a faixa etária, aumenta-se também a possibilidade de um usuário ser um cliente não ativo.

A realização desta análise possibilitou a identificação de algumas características que possibilitam um melhor entendimento do conjunto de dados em questão. Para os analistas de marketing, fica uma primeira sugestão levantada pelo estudo, a de que é necessário analisar a possibilidade uma campanha específica para os bairros de Vila Isabel, Grajaú e Andaraí, tentando fazer com que os usuários não ativos, moradores destes bairros, possam se tornar usuários ativos.

5.2.2 Segmentação dos Clientes da Região Zona Sul III

Dando continuidade ao processo de análise proposto por este capítulo, o estudo segue agora na geração de um modelo para a região Zona Sul III, que é composta pelos seguintes bairros: Copacabana, Ipanema, Leblon, Leme e São Conrado. Esta região é composta por bairros que beiram a orla do município do Rio de Janeiro e que estão contidos entre os bairros do Leme e São Conrado. Nesta região há outros bairros cadastrados, porém apenas estes bairros foram encontrados na base de clientes.

As lojas que atendem a esta região são as seguintes: Copacabana, Ipanema, Leblon, Rio Sul e Barra. Esta última loja só aparece nesta região por conta do bairro São Conrado, que é atendido tanto pela loja Leblon, quanto pela loja Barra. O resultado obtido pelo estudo para esta região pode ser visualizado na figura 5.12, que foi encontrado a partir da valoração do parâmetro quantidade de cluster igual a 15 (quinze).

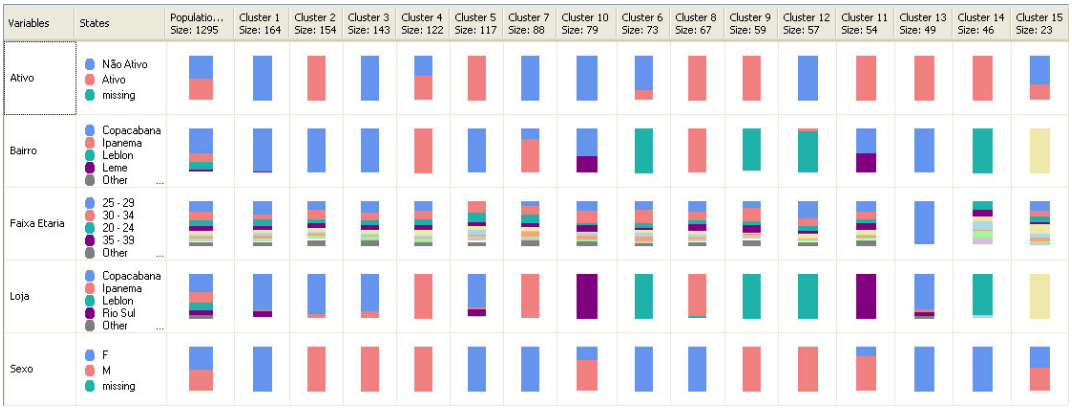


Figura 5.12: Segmentação dos clientes da região Zona Sul III.

O cluster 2 que indica a existência de uma grande quantidade de clientes ativos, moradores do bairro de Copacabana, atendidos em sua maioria pela loja Copacabana e que são do sexo masculino. Em contrapartida, o cluster 3 é formado por clientes não ativos e que possuem o mesmo perfil do cluster 2. O que não nos permite inferir diretamente, pela comparação entre estes dois clusters, a respeito das características que determinam se um usuário é ativo ou não para esta região.

Os clusters de número 9 e 12 nos mostra que é possível diferenciar o tipo de cliente a partir do atributo bairro, já que há uma demonstração no cluster 12 de que os clientes não ativos são os atendidos pela loja Leblon, de sexo masculino e que moram em Ipanema.

Desta forma, o estudo apresenta na tabela 5.2 os resultados mais significantes obtidos a partir da análise dos dados da região Zona Sul III.

Cluster	**Tipo de Cliente**	**Bairro**	**Faixa Etária**	**Loja**	**Sexo**
9 e 12	Não Ativo	Ipanema	-	Leblon	M
13	Ativo	Copacabana	25 – 29	Copacabana	F
15	Não Ativo	São Conrado	-	Barra	-
10 e 11	-	-	-	Rio Sul	M
14	Ativo	Leblon	Não possui clientes na faixa de 25 - 29	Leblon	F

Tabela 5.2: Resultado obtido na segmentação dos dados da região Z. Sul III.

O cluster 13, analisado com o conjunto dos clusters 1 e 5, nos informa que é possível separamos um grupo específico de clientes ativos para os moradores de Copacabana atendidos pela loja do bairro. Este cluster nos informa exatamente a

respeito de grupo constituído de clientes ativos, contidos em uma faixa etária entre 25 – 29 anos e que são do sexo feminino.

Já o cluster 15 nos traz uma informação de que os clientes de São Conrado, atendidos pela loja Barra, são na maioria clientes não ativos. Enquanto que os clientes da loja Rio Sul, identificados pelo cluster 10 e 11, são na maioria clientes do sexo masculino, mas que possuem um equilíbrio de distribuição a respeito do tipo de cliente.

Uma informação interessante é fornecida pelo cluster 14 quando comparado aos clusters de número 6, 12 e 9, que indica um subgrupo de clientes ativos, do sexo feminino, moradores do Leblon, atendidos pela loja do próprio bairro e que não possui clientes na faixa etária entre 25 – 29 anos.

O resultado obtido pela análise desta região leva a uma consideração em relação a faixa etária para as futuras campanhas a serem realizadas, já que para Copacabana é encontrado uma formação de clientes ativos, do sexo feminino, na faixa etária entre 25 – 29 anos, e no Leblon é encontrado uma formação de clientes, também ativos, do mesmo sexo e que não estão dentro desta faixa etária. Isto é muito importante, uma vez que o publico alvo das futuras campanhas podem ser distintos.

Uma ação mais enérgica deve ser considerada no que diz respeito aos clientes de São Conrado atendidos pela loja Barra, já que a maioria destes clientes não é ativa. Este pode ser um local específico para mensurar uma futura campanha que contenha o objetivo de tornar os clientes não ativos em ativos.

Porém, um fato relacionado a região deve ser considerado: será que os clientes não ativos deste bairro não são moradores da comunidade da Rocinha? Caso isto se confirme, os analistas de marketing devem ter em mente que o tipo de campanha a ser realizada neste local é diferente do que se fosse realizada apenas para os moradores da parte da orla deste bairro.

5.2.3 Segmentação dos Clientes da Região Zona Sul I

A próxima análise realizada é feita para a região denominada Zona Sul I, que é composta pelos seguintes bairros: Botafogo, Flamengo, Laranjeiras, Urca, Glória e

Catete. Para estes bairros, as lojas que fazem atendimento são: Botafogo, Rio Sul, Centro e Leblon.

Quanto a esta última loja, vale ressaltar que apenas dois clientes estão cadastrados em sua zona de entrega, o que não trará impacto algum na análise realizada. O resultado encontrado pelo modelo pode ser visualizado na figura 5.13, que foi encontrado a partir da valoração do parâmetro quantidade de clusters igual a 15 (quinze).

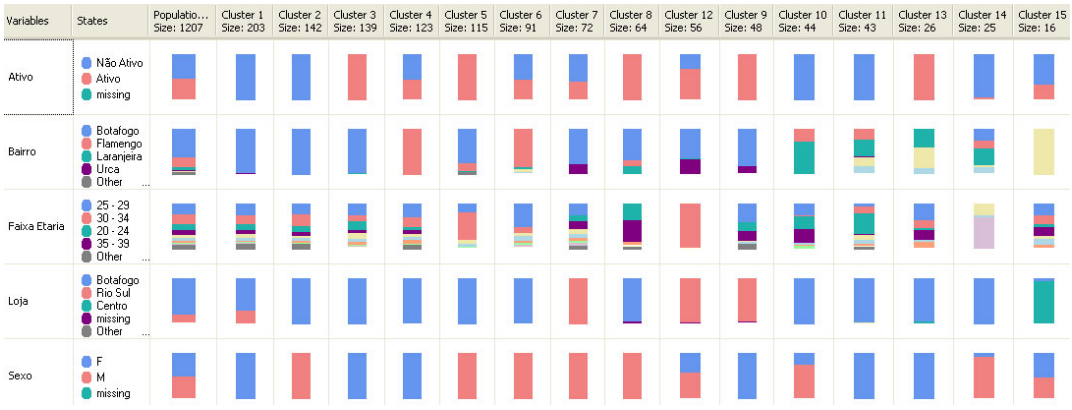


Figura 5.13: Segmentação dos clientes da região Zona Sul III.

A partir deste resultado foi possível identificar alguns perfis associados a quem é ativo ou não ativo:

- Os moradores do bairro Laranjeiras estão mais tendentes para os clientes não ativos (clusters 10, 11 e 14) do que para ativos (clusters 13 e 8);
- Os clientes do bairro Glória, que formam o cluster de número 15, atendidos pela loja Centro, possuem uma tendência maior para os clientes não ativos do que para os ativos;
- A Urca é um bairro que possui mais clientes ativos (clusters 12, 9 e 7) do que não ativos (cluster 7);
- O cluster 14, embora seja um cluster pequeno, mostra que os clientes do sexo masculino, atendidos pela loja Botafogo, com faixa etária entre 40 – 44 e acima de 59 anos, possuem uma maior tendência para os clientes não ativos; e

- Os clusters 4 e 8 mostram que os usuários com o mesmo perfil do item acima, mas com uma menor faixa etária são bem propensos a compra. Em especial os que possuem faixa etária entre 30 – 34 anos, como confirma o cluster de número 12.

Com o que foi produzido nesta análise já dá para termos uma melhor noção dos bairros (Laranjeiras e Gloria) que devem ser atingidos para se tentar tornar um cliente não ativo em ativo. Um bom ponto para esta tarefa poderia ser as estações de metrô do Largo do Machado e dá Glória, já que atendem respectivamente aos usuários destes bairros.

Para o bairro Botafogo, vale uma campanha com um público de maior faixa etária, com a finalidade de tornar os clientes não ativos em clientes ativos. Já uma campanha de marketing, realizada para um público com faixa etária menor que 34 anos, para este mesmo bairro, terá uma maior tendência ao sucesso, uma vez que estes clientes formam a maioria dos clientes ativos.

5.2.4 Segmentação dos Clientes da Região Barra da Tijuca / Recreio

A região em torno da Barra da Tijuca e Recreio é a próxima a ser analisada. Nela estão compreendidos os seguintes bairros: Barra da Tijuca, Itanhangá, Joá e Recreio dos Bandeirantes. Já as lojas que atendem esta região são duas: Barra e Barra Shopping.

Um dado que merece a atenção é o fato de terem sido encontrados nove registros pertencentes a esta área e que estão com o cadastro de loja de atendimento como sendo da loja Iguatemi. Como estes cadastros representam menos de 1% do total de registros desta região, o estudo simplesmente ignorou estes dados na análise e seguiu adiante na geração do modelo, sem se importar com as razões que levaram estes dados a terem estes valores. Quanto ao resultado da análise, este pode ser visto na figura 5.14.

Vale uma ressalva para esta figura: a legenda do atributo dos clientes ativos e não ativos, assim como a legenda do atributo faixa etária, tiveram suas cores alteradas se comparadas com as figuras das análises anteriores. Isto se deve a maneira pela qual o programa utilizado gera estas legendas, já que o estudo não tem controle na definição das mesmas.

Pelo que foi observado, estas legendas são geradas a partir da distribuição dos valores mais populares entre os atributos, nos fazendo notar que, pela primeira vez em uma região, a quantidade de clientes ativos é maior do que a quantidade de clientes não ativos.

O atributo faixa etária também mostrou uma variação na ordem dos valores de sua distribuição. A quantidade de clientes na faixa etária correspondida entre 30 – 34 anos superou a de 25 – 29 anos, assim como a faixa entre 35 – 39 anos superou a de 20 – 24 anos.

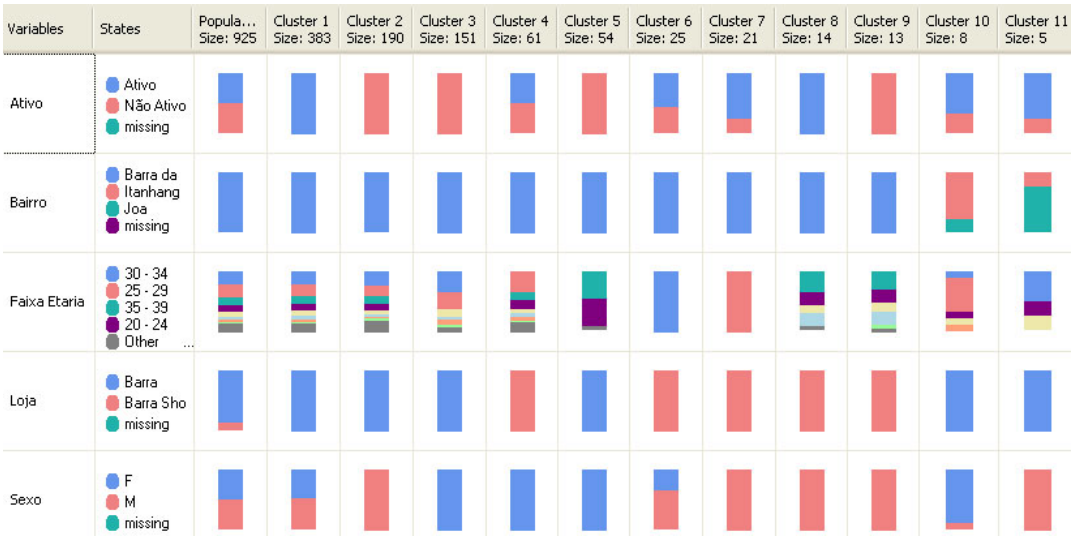


Figura 5.14: Segmentação dos clientes da região Barra / Recreio.

Os bairros do Itanhangá e do Joá representam uma quantidade bem pequena do volume de registros, estes dois bairros são representados pelos clusters 10 e 11 respectivamente. Ambos possuem uma tendência maior aos clientes ativos e são atendidos pela loja Barra.

Quanto ao atributo sexo, o bairro do Itanhangá possui a maioria de seus clientes com o sexo feminino, enquanto que o bairro do Joá possui apenas clientes do sexo masculino.

O cluster 1 (que indica um perfil dos clientes ativos para a loja Barra) quando comparado aos clusters 2 e 3 (que indicam um perfil de clientes não ativos para a loja Barra), não permite uma definição das características que levam a um cliente a ser ativo ou não. O mesmo ocorre entre os clusters 9 (não ativo), 8 (ativo) e 4 (ativos e não ativos) que representam os clientes atendidos pela loja Barra Shopping.

O cluster 5 parece demonstrar um perfil específico para o atributo faixa etária no que diz respeito aos clientes não ativos. Porém, quando considerada a distribuição deste atributo no cluster 3, logo vemos que este é apenas um complemento da distribuição deste atributo entre estes dois clusters.

A análise desta região ficou um pouco prejudicada dada a baixa variação de valores oferecida por cada atributo, o que faz com que os registros de dados não apresentem grandes variações entre si. Este fato sem dúvida gera um impacto que dificulta a classificação dos dados.

Para esta análise, vale a observação da maneira pela qual é constituída a distribuição dos valores do atributo faixa etária, que se diferenciou em relação as demais análises realizadas até o momento.

Como sugestão, fica uma possível avaliação para aumentar o grau de separação destes dados. Isto pode ser feito a partir da subdivisão do bairro Barra da Tijuca em pequenas áreas, uma vez que este bairro compreende uma vasta extensão territorial. Quem sabe assim não seja possível identificar um determinado perfil que torne possibilite a distinção entre os tipos de clientes.

5.2.5 Segmentação dos Clientes da Região Centro

A análise seguinte leva em consideração o conjunto de clientes cadastrados na região Centro, que possui os seguintes bairros cadastrados: Centro, Estácio, Cidade Nova, Saúde, Fátima e Santa Teresa. As lojas que mais atendem esta região são: Centro, Tijuca e Botafogo.

Novamente foi possível encontrar uma pequena quantidade de registros (cinco no total) que não fazem parte desta região, mas que por alguma razão estavam marcados como se fossem. Esta quantidade de dados representa menos de 1% do total de registros, o que fez com que o estudo ignorasse estes dados na análise realizada.

Para evitar a criação de clusters com uma baixa quantidade de registros, o estudo limitou o número mínimo de registros por cluster a 25. Isto foi realizado a partir da valoração do parâmetro suporte mínimo, que tem por objetivo servir de critério de parada na execução do algoritmo *K-means*.

Mesmo com o valor do parâmetro quantidade de clusters igual a 10, a valoração do atributo suporte mínimo fez com que não fosse possível encontrar mais do que cinco clusters na análise desta região, que tem seu resultado apresentado na figura 5.15.

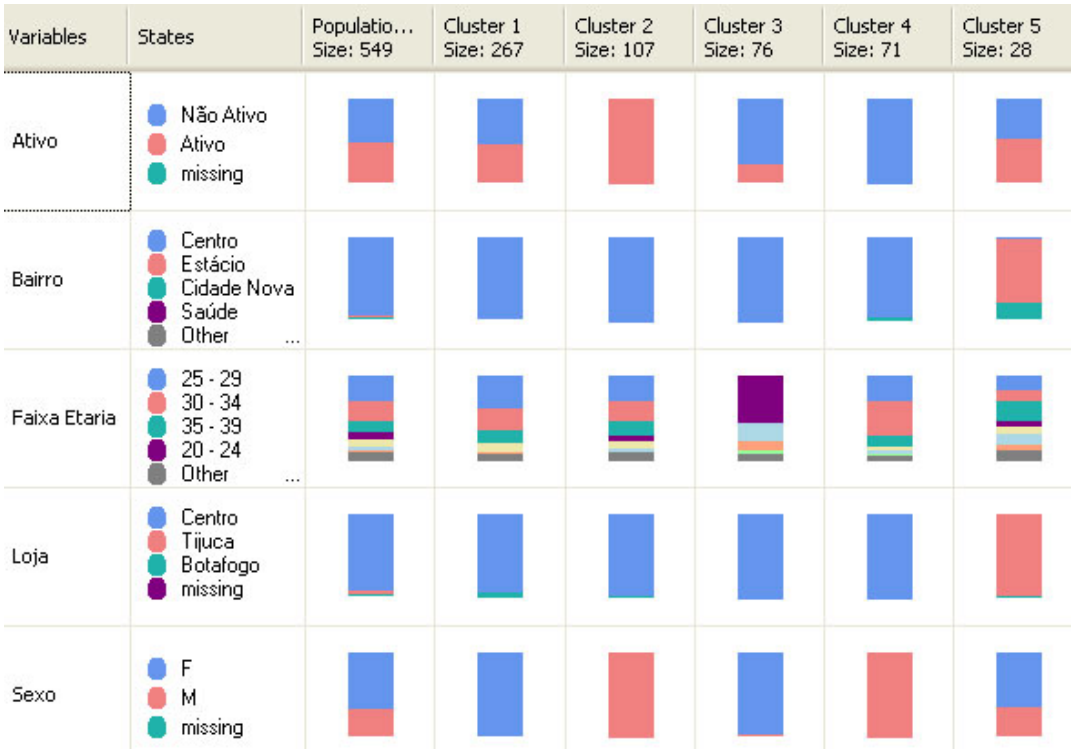


Figura 5.15: Segmentação dos clientes da região Centro.

À medida que se reduz a quantidade de clusters a ser encontrado, menor é o esforço computacional para se obter o resultado. Contudo, isto nos traz uma menor possibilidade de se conseguir um padrão que contenha uma relevante informação ao negócio em questão.

O cluster 3 nos traz uma informação de que os clientes não ativos, registrados no bairro Centro, do sexo feminino, com faixa etária entre 20 – 24 ou acima dos 45 anos, representam um grupo específico de clientes não ativos. Temos uma confirmação para esta informação quando analisamos os clusters 1 e 2, uma vez que estes clusters possuem clientes ativos e que não constam, em sua maioria, entre as faixas etárias do cluster 3.

Outra informação interessante, avaliada nos clusters 1, 2 e 4, é que os clientes do sexo masculino, de uma maneira geral, estão mais propensos a um tipo de cliente não ativo. Assim, é possível a execução de operações de marketing especificas para se tentar

chegar aos clientes não ativos desta região, sendo uma para o público feminino (que deverá levar em consideração o atributo faixa etária) e outra para o público masculino.

5.2.6 Segmentação dos Clientes da Região Zona Norte II

A próxima etapa do estudo trabalhará com os dados relacionados aos clientes cadastrados na região Zona Norte II, que contém os seguintes bairros: Méier, Todos os Santos, Cachambi, Abolição, Engenho de Dentro, Pilares, Engenho Novo, Piedade, Maria da Graca, Lins de Vasconcelos, Del Castilho e Riachuelo.

Apesar de existirem quatro registros com clientes cadastrados na área de entrega coberta pela loja Iguatemi, o estudo levará em consideração apenas os clientes da loja Norte Shopping, devido ao fato de que os clientes desta loja representam mais do que 98% da população do conjunto de dados desta região.

Dentre os vários modelos criados, o estudo optou por ficar com o obtido a partir do valor de parâmetro quantidade de clusters igual a 8 e com a valoração do parâmetro suporte mínimo igual a 25. O resultado alcançado pelo estudo pode ser visualizado na figura 5.16.

Nesta figura é possível perceber que os clusters de número 8 e 7 formam subgrupos, constituídos em sua maioria, de pessoas na faixa etária entre 20 – 24 anos. Porém, se o usuário for do sexo masculino e morar no bairro Cachambi (cluster 8), há uma tendência para um tipo de cliente não ativo; no entanto, se este cliente for do sexo feminino e morar no bairro Todos os Santos (Cluster 7), há uma tendência para um tipo de cliente ativo.

Figura 5.16: Segmentação dos clientes da região Zona Norte II.

O cluster 4 nos informa que os moradores do Méier, de sexo feminino, com faixa etária acima dos 30 anos, possuem uma propensão para os clientes ativos. O cluster 6 também indica que isto ocorre para os outros bairros, mas com clientes do sexo masculino.

No cluster 2 é possível visualizarmos uma grande quantidade de clientes do sexo feminino, na faixa etária dos 25 – 29 anos, mas sem um padrão comportamental a respeito do tipo de cliente. Já o cluster 1 nos informa que as pessoas com este mesmo sexo e fora desta faixa etária possuem um clara tendência a um tipo de cliente não ativo.

Com este resultado é possível chegarmos a uma proposição de campanhas para os diferentes tipos de clientes encontrados. Como por exemplo, uma campanha que contenha os dois perfis encontrados na análise entre os clusters 8 e 7.

Outra possível campanha seria para o bairro Méier, voltada para um público alvo acima dos 30 anos e de sexo feminino, uma vez que o cluster 4 indicou um bom potencial de clientes ativos enquadrados neste perfil, o que indica um maior retorno.

5.2.7 Segmentação dos Clientes da Região Ilha do Governador

A região Ilha do Governador é a próxima a ser analisada pelo estudo, nela encontramos os bairros: Ribeira, Jardim Carioca, Tauá, Praia da Bandeira, Galeão,

Cocotá, Freguesia (Ilha do Governador), Cacuia, Moneró, Pitangueiras, Zumbi e Bancários.

A loja Ilha Plaza é a responsável pela cobertura desta região. Porém, mais uma vez foi encontrado um total de registros (quatro) cadastrados nesta região, mas associados a uma loja que não pertencente a estas localidades. Estes registros foram ignorados pelo estudo, assim como ocorreu nas análises anteriores.

O modelo gerado partiu de um valor igual a 6 para o parâmetro quantidade de clusters e 25 para o parâmetro suporte mínimo. O resultado obtido poder ver visto logo abaixo na figura 5.17.

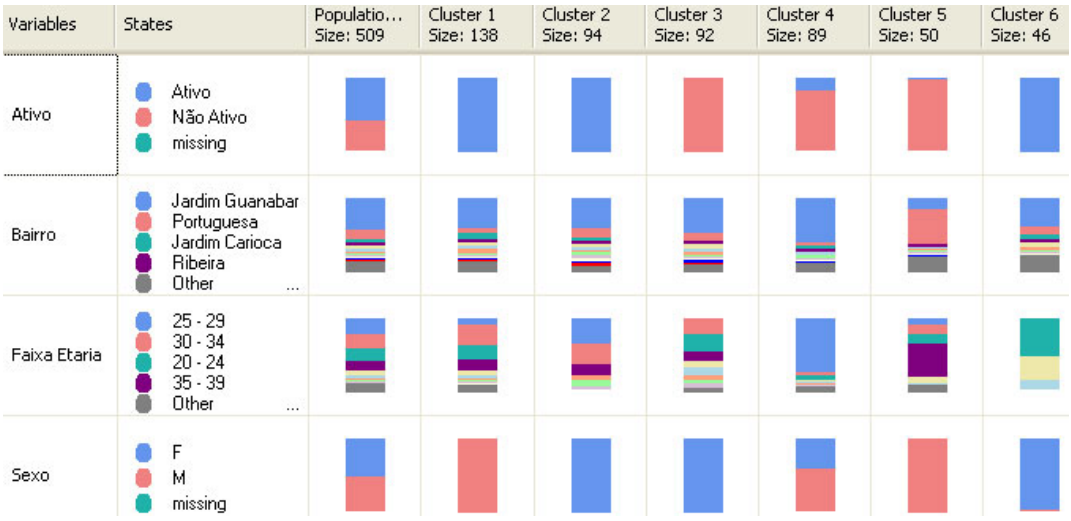


Figura 5.17: Segmentação dos clientes da região Ilha do Governador.

O cluster 5 é o primeiro que nos chama atenção por conta de sua formação. Nele é possível observarmos um perfil gerado pelos usuários moradores do Bairro da Portuguesa, com uma faixa etária entre 35 – 39 anos, do sexo masculino e que não fazem pedidos.

O bairro Jardim Guanabara, pode ser visto com bastante destaque no cluster de número 4. Neste cluster pode-se notar que a maioria dos clientes com faixa etária entre 25 – 29 anos são bem propensos aos clientes não ativos.

Esta informação fica um pouco confusa quando comparada ao cluster 2, que indica que esta faixa etária é mais ativa, quando o sexo do cliente é igual a feminino. Já quando o sexo é masculino, a faixa etária mais ativa para este bairro é a dos 30 – 34

anos, seguida de perto pela faixa etária dos 20 – 24 anos e, por último, as dos 35 – 39 anos, como mostra o cluster de número 1.

Clientes com a faixa etária entre 20 – 24 anos e acima dos 40, valorados com o sexo igual a feminino, como mostra o cluster de número 6, indicam normalmente um tipo de cliente ativo para o bairro Jardim Guanabara.

Tudo isso nos mostra que uma campanha para um público alvo de sexo masculino, moradores do bairro da Portuguesa, com faixa etária entre 35 – 39 anos, terá uma chance maior de alcançar os clientes não ativos.

Já uma campanha voltada para os clientes ativos, moradores do bairro Jardim Guanabara, terá uma maior possibilidade de sucesso se feita para um público com o sexo masculino, faixa etária entre os 20 – 24 e entre 30 – 39 anos. Enquanto que para o público feminino, o melhor é uma campanha levando em consideração a faixa etária entre 25 – 29 anos.

5.2.8 Segmentação dos Clientes da Região Zona Norte III

A próxima etapa do estudo irá avaliar a região Zona Norte III, que compreende os seguintes bairros: Vila da Penha, Irajá, Brás de Pina, Penha, Madureira, Penha Circular, Vista Alegre, Vicente de Carvalho, Tomás Coelho, Olaria, Higienópolis, Vaz Lobo, Ramos, Inhaúma, Colégio, Quintino Bocaiúva, Bonsucesso, Parada de Lucas, Rocha Miranda e Oswaldo Cruz. As lojas que atendem a esta região são: Carioca Shopping e Norte Shopping.

O estudo tentou gerar um modelo a partir do valor igual a 10 para o parâmetro quantidade de clusters. Entretanto, por conter um reduzido volume de registros (apenas 374), o modelo gerou apenas 5 clusters para esta região, fato este ocorrido pelo parâmetros de critério de parada do algoritmo *K-means*, como por exemplo, suporte mínimo e *STOPPING_TOLERANCE*. O resultado pode ser visto na figura 5.18.

Como pode ser visto, vai se tornando cada vez mais difícil encontrar um padrão para se determinar o tipo de cliente a partir dos modelos que contêm um menor número de registros. Conseqüentemente isto faz com que o número de clusters encontrados

diminua. A questão é que com um número cada vez menor de usuários por subgrupo um campanha específica para uma determinada região se torna mais cara.

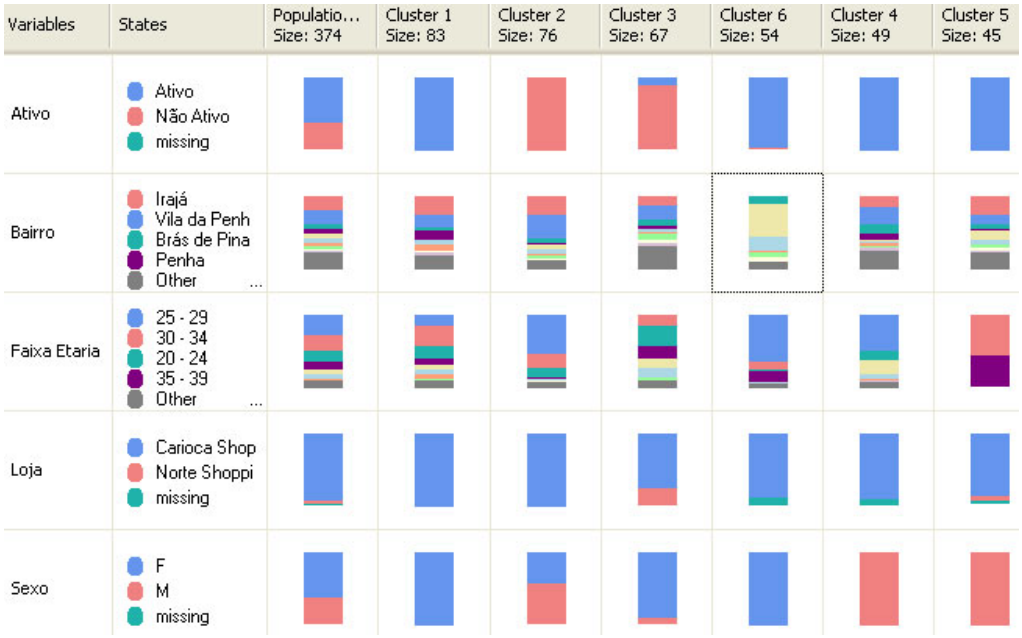


Figura 5.18: Segmentação dos clientes da região Zona Norte III.

Observa-se no cluster 6 que os moradores do bairro Madureira e Penha, correspondem a uma maioria de clientes formados pelo sexo feminino, com faixa etária entre 25 – 29 anos e que são ativos. O cluster 5 indica que os clientes desta região, com faixa etária entre 30 – 39 anos, do sexo masculino, são clientes ativos.

No cluster 2 é possível termos uma idéia de que os clientes com faixa etária entre 20 – 29 anos, moradores dos bairros Irajá e Vila da Penha, têm uma tendência para o tipo de cliente não ativo. Quanto a loja Norte Shopping, o cluster de número 3, indica que os clientes atendidos por esta loja compõe uma parte maior dos dados referentes aos clientes não ativos.

Como um todo, podemos crer que uma campanha para o perfil de clientes extraído a partir do cluster 6 e 5 é sem dúvida uma campanha com uma maior potencial de retorno, já que os clientes destes clusters são, em sua maioria, valorados como clientes ativos. Para uma campanha que tente buscar os clientes não ativos, o melhor seria focar os esforços nos bairros de Irajá e Vila da Penha, para um publico alvo entre 20 – 29 anos.

5.2.9 Segmentação dos Clientes da Região Zona Sul II

A análise seguinte diz respeito ao modelo gerado para a região Zona Sul II, que contém os seguintes bairros: Lagoa, Jardim Botânico, Gávea, Humaitá e Cosme Velho. As lojas que atendem esta região são: Leblon, Botafogo e Ipanema.

O estudo levou em consideração um valor de suporte mínimo igual a 15 para os clusters formados por esta região, já que a quantidade de clientes cadastrados é igual a 286. Para o parâmetro quantidade de clusters, o melhor resultado obtido foi com um valor igual a 10, como pode ser visto na figura 5.19.

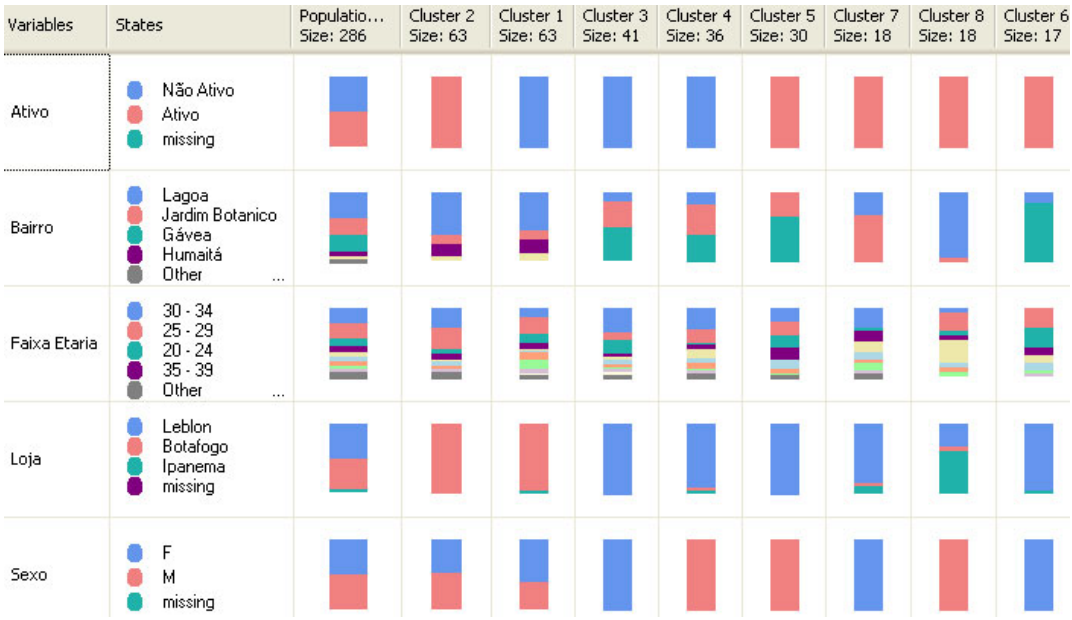


Figura 5.19: Segmentação dos clientes da região Zona Sul II.

Os clusters de número 1 e 2 pouco indicam a respeito de um cliente ativo ou não, já que um é o complemento do outro para se definir o tipo de cliente em questão, ou seja, o mesmo perfil é encontrado para indicar um cliente ativo e não ativo.

Os clusters de número 3 e 4, quando comparados aos cluster de número 5 e 7, possuem o mesmo comportamento citado acima, não nos trazendo nenhuma informação que faça a distinção entre o tipo de cliente.

A loja Ipanema (cluster 8) é a que contém a maior média de clientes ativos se comparada as outras lojas desta região. O perfil deste cluster é formado por clientes com idade entre 25 – 29 e 50 – 54 anos, sexo masculino e moradores da Lagoa. Porém, é

bom lembrar que este cluster possui uma baixa quantidade de registros. De qualquer forma, este perfil indica um público alvo de bom retorno.

Os moradores da Gávea, que são sempre atendidos pela loja Leblon, de sexo feminino, com faixa etária entre 20 – 29 anos, são mais dispostos a serem clientes ativos dos que os que possuem a faixa etária entre 30 – 34 anos. Este pode ser também um bom público alvo de retorno das campanhas.

5.2.10 Segmentação dos Clientes da Região Niterói

Por último, o estudo avalia a região de Niterói, que possui os seguintes bairros: Ingá, Fonseca, Centro, Santa Rosa, Icaraí, São Domingos, Ilha da Conceição, São Lourenço, Pé Pequeno, Boa Viagem, Barreto, Piratininga, Itaipu e Largo do Barradas. As lojas que atendem esta região são: Plaza Niterói e Icaraí.

Como identificado no levantamento das estatísticas básicas, a loja Icaraí possui um baixo número de clientes devido ao incorreto cadastro dos CEP's atendidos por esta loja. Por isso, o estudo não levou em consideração o atributo loja para a geração do modelo representado na figura 5.20. Este modelo foi gerado a partir de um valor igual a 10 para o parâmetro quantidade de clusters e um valor igual a 15 para o parâmetro suporte mínimo.

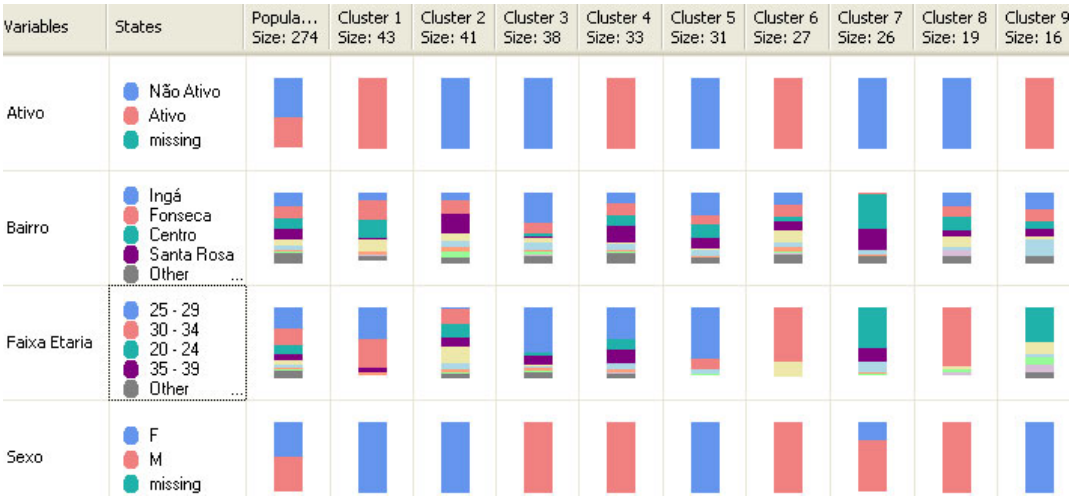


Figura 5.20: Segmentação dos clientes da região Niterói.

O cluster 7 nos traz uma informação de que os clientes com faixa etária entre 20 – 24 e 35 – 39 anos, dos bairros Centro e Santa Rosa, na maioria com o sexo masculino,

fazem parte dos clientes não ativos. Já o cluster 9, indica que clientes na faixa etária entre 20 – 24 anos e acima dos 40 anos, com o sexo feminino, fazem parte do grupo de clientes ativos. Neste cluster, podemos ver representado pela cor verde claro um grupo de clientes na faixa etária entre 15 – 19 anos.

Em relação aos clusters 3 e 4, ambos possuem um semelhante perfil de sexo masculino e com faixa etária entre 25 – 29 em sua grande maioria. O primeiro representa os clientes não ativos, com uma maior distribuição pelo bairro Ingá; Já o segundo, os dos clientes ativos, possui uma maior distribuição para o bairro Santa Rosa.

Os clientes do sexo feminino, com faixa etária entre 25 – 29 e 30 – 34 anos, cadastradas nos bairros Itaipu e Ipiratininga, possuem uma tendência maior de compra, como pode ser visto no cluster 1. Por último, o cluster 2, quando comparado aos demais clusters, mostra que a maioria das mulheres, de faixa etária entre 40 – 44 anos, possuem uma inclinação para os clientes considerados não ativos.

Esta última informação confronta com a adquirida no cluster 9, que nos informa que as mulheres acima dos 40 anos possuem uma inclinação para um tipo de cliente ativo. Isto acontece pelo fato de que estes registros possuem um mesmo conjunto de características.

Para uma futura campanha, os clientes do sexo masculino, com idade entre 25 – 29 anos, terão uma chance maior de retorno quando alcançados no bairro Santa Rosa, ao contrário do bairro Ingá, que possuirá uma chance menor. Para os clientes de faixa etária entre 20 – 24 e 35 – 39 anos, campanhas voltadas para o sexo masculino, realizadas para um público do bairro Centro e Santa Rosa, poderão ter um maior efeito se realizadas com o objetivo de tornar um cliente não ativo em ativo.

5.3 Conclusão

Termina aqui a análise realizada através do método de classificação não supervisionada feita pelo algoritmo *K-means*. De um modo geral, vimos que, mesmo com uma base de informações bem homogênea, é possível extrairmos um padrão de conhecimento que permita um melhor entendimento do comportamento destes dados.

Nem sempre esta tarefa pode ser realizada com todo o conjunto de dados de uma única vez. Em certas ocasiões, como ocorreu com o presente estudo, o melhor é subdividirmos o conjunto de dados em outros por menores de acordo com a necessidade do negócio. Isto não só facilita a compreensão dos dados, que ficam de uma forma mais simples, como também exige um menor esforço computacional para se concluir à geração de cada modelo.

Para o caso deste estudo, uma técnica que muito irá ajudar a complementar o trabalho realizado neste capítulo será a técnica de mineração de dados chamada de regras de associação, já discutida anteriormente.

Enquanto o objetivo deste capítulo era buscar distinguir as características que formam cada subgrupo de clientes cadastrados, a regra de associação, aplicada a este estudo, terá por objetivo buscar distinguir os diferentes comportamentos de compra contidos na loja virtual do La Mole.

Desta forma, o estudo conclui a segmentação da base de dados do conjunto de clientes e buscará entender um pouco mais a respeito dos pedidos realizados e seus respectivos produtos associados.

6 Análise de Regras de Associação

6.1 Introdução

Neste capítulo o estudo tem por objetivo entender o comportamento de compra dos clientes. Este entendimento poderá fornecer um conhecimento específico a respeito da maneira pela qual os produtos são vendidos e por qual perfil de cliente é realizada esta compra.

Outra ajuda a ser fornecida será complementar as campanhas a serem realizadas a partir do conhecimento obtido através da segmentação dos clientes, realizada anteriormente pelo estudo. Em muitos casos notou-se que um mesmo perfil foi encontrado tanto para os clientes ativos, quanto para os clientes não ativos.

Este era o tipo de informação que não fornecia qualquer dado relevante à distinção do tipo de usuário. Entretanto, se o estudo utilizar os dados dos pedidos realizados pelos clientes ativos, a fim se fazer uma busca por algum padrão de compra contido nestes dados, poderá ser possível criar campanhas de marketing para empregar estes padrões aos clientes de mesmo perfil e que não são ativos.

Isto aumentaria a diversidade das campanhas a serem desenvolvidas, já que a identificação de um padrão de compra poderá fazer com que sejam realizadas campanhas em conjunto com os fornecedores da empresa.

6.2 Análise e Desenvolvimento

A primeira tarefa realizada foi gerar um levantamento dos dados que serão analisados, o que mostrou a existência de 183 produtos distintos, agrupados em um total de 14 categorias.

Quanto aos pedidos realizados, foi gerado um gráfico contendo o diagrama de freqüência (figura 6.1) com a quantidade de itens por pedido. Neste gráfico é observado que quase a metade possui apenas um item de compra. Desta forma, estes registros não serão levados em consideração na análise a ser realizada.

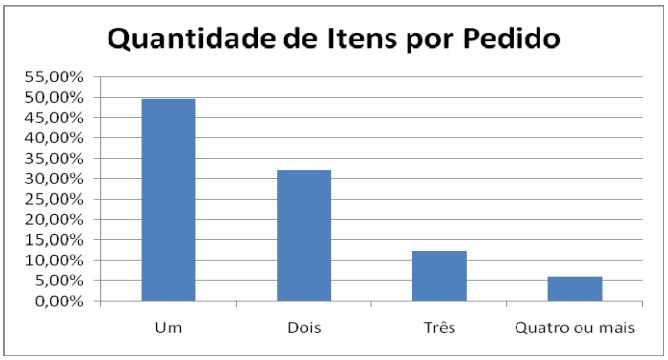


Figura 6.1: Diagrama de freqüência da quantidade de itens por pedido

Embora os parâmetros de confiança e suporte sejam os mais importantes na geração das regras de associação, a ferramenta utilizada pelo estudo possui outros parâmetros que ajudam no processo de execução deste método, que são:

- *MAXIMUM_ITEMSET_COUNT* - Quantidade máxima de regras a serem produzidas;
- *MAXIMUM_ITEMSET_SIZE* - Quantidade máxima de itens por regra;
- *MAXIMUM_SUPPORT* - Valor máximo de suporte para uma regra, ou seja, regras com uma freqüência maior que as definidas para este parâmetro não serão consideradas;
- *MINIMUM_IMPORTANCE* - Valor de mensuração da importância de uma regra;
- *MINIMUM_ITEMSET_SIZE* - Quantidade mínima de itens por regra;
- *MINIMUM_PROBABILITY* - Valor mínimo de confiança para uma regra; e
- *MINIMUM_SUPPORT* - Valor mínimo de suporte para um regra.

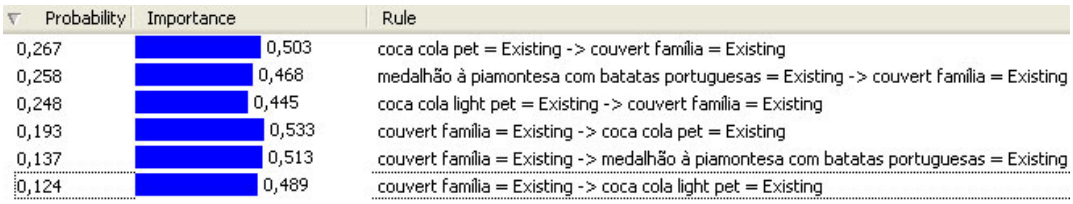
A primeira análise realizada foi feita a partir dos valores padrões de parametrização, que seguindo a ordem acima são respectivamentes valorados em: 200.000, 3, 1.0, -999.999.999, 1, 0.4 e 0.

Vale ressaltar que os valores dos parâmetros de suporte máximo, suporte mínimo e confiança podem ser representados de forma absoluta ou percentual. Isto é definido a partir do valor informado, que quando é menor do que 1 representa um valor

a ser medido de forma percentual; e quando maior ou igual a este valor, representa um valor absoluto.

Desta forma, a primeira análise foi gerada com um valor mínimo de confiança igual a 40% e sem um valor específico para o parâmetro suporte, o que faz com o algoritmo crie regras de associação com qualquer valor para este parâmetro. Esta análise não produziu resultado algum, levando o estudo a iniciar um processo de busca por valores de suporte e confiança que gerassem alguma regra.

O melhor resultado encontrado foi obtido a partir do suporte mínimo igual a 1% e com valor de confiança mínima igual a 10%, como pode ser visualizado na figura 6.2.



Probability	Importance	Rule
0,267	0,503	coca cola pet = Existing -> couvert família = Existing
0,258	0,468	medalhão à piamontesa com batatas portuguesas = Existing -> couvert família = Existing
0,248	0,445	coca cola light pet = Existing -> couvert família = Existing
0,193	0,533	couvert família = Existing -> coca cola pet = Existing
0,137	0,513	couvert família = Existing -> medalhão à piamontesa com batatas portuguesas = Existing
0,124	0,489	couvert família = Existing -> coca cola light pet = Existing

Figura 6.2: Regras geradas a partir dos produtos mais vendidos.

Este resultado mostra todas as regras de associação encontradas pelo modelo, indicando para cada regra seu respectivo valor de confiança e importância. No total apenas seis regras foram criadas, sendo que as três últimas representam o inverso das três primeiras, ou seja, o antecedente encontrado nas três primeiras regras troca de posição com o seu respectivo conseqüente nas três últimas.

As regras relacionadas entre os produtos ‘coca-cola’ e ‘*couvert* família’ não trazem um tipo de conhecimento muito especial ao negócio, uma vez que este tipo de compra já se encontra dentro de um padrão esperado, até porque o produto ‘coca-cola pet’ é atualmente uma indicação de compra do produto ‘*couvert* família’.

Contudo, a regra entre os produtos ‘medalhão à *piamontesa* com batatas portuguesas’ e o ‘*couvert* família’ já demonstra um padrão de compra muito interessante, identificado somente após a realização desta análise.

Na figura analisada é observado que a probabilidade (confiança) da regra entre os produtos ‘medalhão à *piamontesa*’ e ‘*couvert* família’ é bem maior do que a regra inversa, embora o grau de importância desta última seja maior do que o da primeira. Porém, este valor não tem tanta influência para estas regras, já que ambas possuem um valor de correlação positivo e bem próximo um do outro.

Outra forma de avaliar as regras geradas é analisar o conjunto de itens gerados e seus respectivos suportes. Para análise em questão, isto foi feito a partir do resultado obtido por um recurso próprio da ferramenta e que pode ser visto na figura 6.3.

Support	Size	Itemset
190	2	coca cola light pet = Existing, couvert família = Existing
295	2	coca cola pet = Existing, couvert família = Existing
210	2	medalhão à piamontesa com batatas portuguesas = Existing, couvert família = Existing

Figura 6.3: Suporte das regras geradas pelos produtos mais vendidos.

É este resultado que permite a visualização do valor de suporte obtido por cada uma das regras criadas. Para o caso da regra entre o 'medalhão a *piamontesa*' e o '*courver* família', o valor de suporte é definido por um total de 210 pedidos, como mostra a figura acima.

Para validar este resultado, o estudo fez uma busca direta no banco de dados para comprovar este valor e determinar a quantidade de clientes distintos possuidores deste padrão de compra. A idéia era se certificar de que esta regra não tinha sido gerada de forma tendenciosa por alguns poucos usuários da loja, ou seja, usuários muito ativos e que fazem sempre um mesmo tipo de pedido.

Esta validação indicou que 129 usuários distintos foram os responsáveis pelos 210 pedidos realizados com este padrão. Estes usuários fizeram um total de 1.183 pedidos, o que indica a existência de pedidos fora deste padrão e que ajuda a validar ainda mais a regra gerada.

A ferramenta utilizada oferece mais um recurso para examinar as regras de associações criadas. Este recurso proporciona uma visualização das dependências entre os itens contidos nas regras geradas, além de permitir uma navegação entre os diversos níveis de dependência.

Na figura 6.4 é exibido um diagrama de dependência para as regras descritas anteriormente, nela pode-se observar que o produto '*couvert* família' prediz o produto 'medalhão a *piamontesa*', assim como o produto coca-'cola pet'.

Figura 6.4: Diagrama de dependências entre as regras geradas.

Para a regra entre a ‘coca-cola pet’ e o ‘*couvert* família’, pode-se notar que o gráfico exibe uma dependência entre ambos os sentidos, o que indica predição entre um produto e outro. No caso da regra entre os produtos ‘*couvert*’ e ‘medalhão’, este relacionamento só é possível quando exibido a dependência em um nível acima do atual, conforme mostra a figura 6.5.

Figura 6.5: Diagrama de dependências com o nível final das regras geradas.

Como um todo, as regras geradas por esta análise não corresponderam às expectativas do estudo, tendo contribuído para isto a própria dispersão dos dados relacionados às compras e a quantidade de produtos disponíveis para venda.

Deste modo, o estudo seguiu em busca de novas regras de associação, só que agora agrupadas pelas categorias de cada produto. Este modelo poderá ajudar a entender um pouco mais os comportamentos de compra realizados pelos clientes, já que a variabilidade dos produtos vendidos e que estão em um mesmo contexto, como por exemplo, o produto ‘coca-cola pet’ e ‘coca-cola *light* pet’, é uma das causas da baixa quantidade de regras geradas.

A análise realizada para este modelo também partiu dos valores padrões fornecidos pela ferramenta, que desta vez permitiu logo de início um resultado considerável, como pode ser visto na figura 6.6.

Probability	Importance	Rule
0,802	0,554	promoção centro = Existing -> bebidas = Existing
0,500	0,339	pizzas = Existing, carnes = Existing -> bebidas = Existing
0,478	0,333	pizzas = Existing -> bebidas = Existing
0,465	0,309	tortas inteiras = Existing -> bebidas = Existing
0,457	0,399	guarnições = Existing, sobremesas = Existing -> carnes = Existing
0,453	0,319	cardápio executivo = Existing -> bebidas = Existing
0,449	0,392	guarnições = Existing, bebidas = Existing -> carnes = Existing
0,442	0,288	peixes = Existing, carnes = Existing -> bebidas = Existing
0,439	0,285	risotos = Existing, sobremesas = Existing -> bebidas = Existing
0,419	0,359	tortas inteiras = Existing -> carnes = Existing
0,418	0,437	vinhos = Existing -> massas = Existing
0,415	0,263	risotos = Existing -> bebidas = Existing
0,403	0,325	carnes = Existing -> bebidas = Existing

Figura 6.6: Regras geradas a partir das categorias de produtos mais vendidos.

Das regras geradas, a mais notável está compreendida entre as categorias ‘guarnições’ e ‘sobremesas’, que em conjunto predizem a compra de produtos da categoria ‘carnes’. Porém, quando é analisado o valor de suporte para esta regra, percebe-se que apenas 64 pedidos contêm este comportamento de compra, o que representa bem menos de 1% do total de pedidos.

Por conta desta regra ter sido gerada a partir de um baixo valor de suporte, o estudo aponta a necessidade de se gerar modelos com um valor mínimo a ser configurado para este parâmetro. Isto evitará a criação e análise de uma regra com pouca representatividade para o modelo de negócio avaliado.

Desta forma, o estudo criou um novo modelo atribuindo um valor mínimo de suporte igual a 1% e um valor mínimo de confiança igual a 30%. O resultado encontrado é visualizado na figura 6.7.

Probability	Importance	Rule
0,478	0,333	pizzas = Existing -> bebidas = Existing
0,453	0,319	cardápio executivo = Existing -> bebidas = Existing
0,403	0,325	carnes = Existing -> bebidas = Existing
0,399	0,275	aves = Existing -> bebidas = Existing
0,396	0,353	guarnições = Existing -> carnes = Existing
0,374	0,259	entradas e saladas = Existing -> bebidas = Existing
0,360	0,242	massas = Existing -> bebidas = Existing
0,351	0,347	sobremesas = Existing -> carnes = Existing
0,349	0,344	entradas e saladas = Existing -> carnes = Existing
0,336	0,175	sobremesas = Existing, massas = Existing -> bebidas = Existing
0,325	0,354	bebidas = Existing -> carnes = Existing
0,322	0,256	sobremesas = Existing, bebidas = Existing -> carnes = Existing
0,317	0,380	cardápio executivo = Existing -> sobremesas = Existing
0,315	0,164	sobremesas = Existing -> bebidas = Existing
0,313	0,329	sobremesas = Existing, bebidas = Existing -> massas = Existing

Figura 6.7: Regras geradas com o valor mínimo de suporte igual a 1% para as categorias de produtos mais vendidos

A maioria das regras geradas está relacionada aos produtos da categoria ‘bebidas’, o que não fornece um nível de informação muito relevante. Às outras regras geradas, destacadas na figura acima com o fundo azul, merecem uma análise mais apurada, já que algumas categorias são acompanhamentos normais de outras.

A categoria ‘guarnições’ como antecedente e conseqüente valorado com a categoria ‘carnes’ é uma regra que não possui um grande valor para o negócio, já que esta categoria é um complemento tradicional dos pratos vendidos. O mesmo acontece com a regra entre a categoria ‘cardápio executivo’ e ‘sobremesas’.

A única regra que se destaca um pouco além das outras é a que contém a relação entre as categorias ‘entradas e saladas’ e ‘carnes’, isto porque não foi possível observar a existência de outras regras com esta categoria no antecedente, mas com outro tipo de categoria no conseqüente.

Estes são os tipos de regras que, embora tenham todas as características relevantes para a identificação de um comportamento de compra, não traz um conhecimento específico para a empresa. Isto mostra o quanto é relevante o papel de um analista de negócio para avaliar o resultado obtido pelos métodos de aprendizagem de máquina, pois apenas estes especialistas podem diferenciar o que é sabido ou não de antemão.

O resultado gerado poderia ter sido melhor se houvesse um refinamento maior entre os agrupamentos das categorias e produtos, subdividindo-os em grupos menores. Um exemplo disto pode ser observado na categoria ‘bebidas’, encontrada por quase todas as regras geradas, como mostra o diagrama de dependências exibido na figura 6.8.

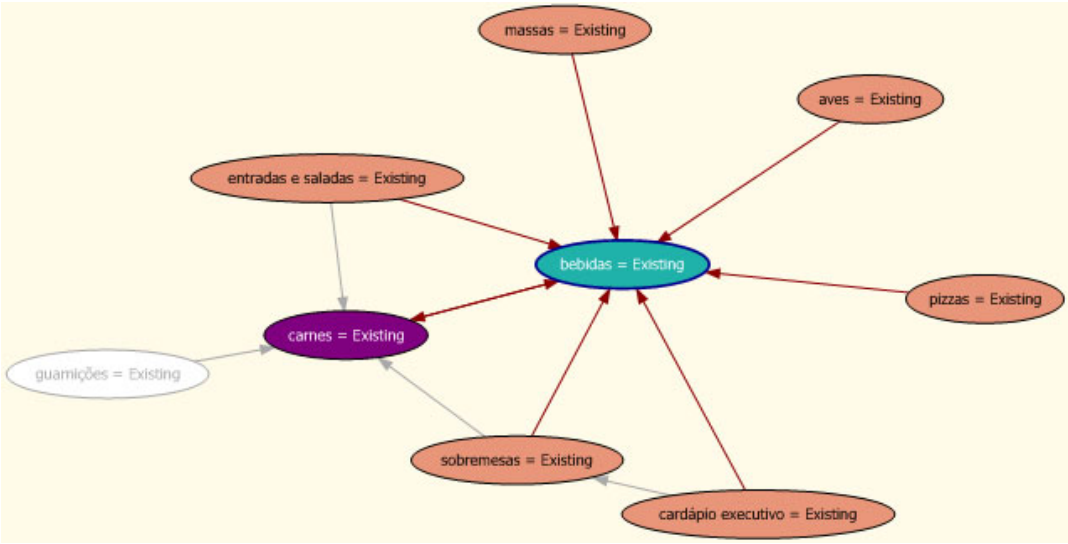


Figura 6.8: Diagrama de dependências das regras geradas pelas categorias dos produtos.

Por conta disso, o estudo resolveu criar subdivisões para a categoria ‘bebidas’, de modo a verificar se os resultados obtidos trarão uma melhor qualidade às informações fornecidas pelas regras a serem criadas. Os novos subgrupos criados são: bebidas naturais (chás e água de coco), refrigerantes em lata, refrigerantes pet e cerveja.

Para a criação deste novo modelo o estudo necessitou reduzir o valor do parâmetro confiança para 25%, mas manteve o valor suporte mínimo em 1%. Isto foi feito pelo fato de que estas subdivisões aumentaram a quantidade de buscas a serem feitas pelo algoritmo.

O resultado obtido pode ser visualizado na figura 6.9, o que já demonstra uma diferenciação se comparado ao resultado obtido anteriormente. Nele é possível observarmos um maior número de regras criadas, fruto da maior quantidade de itens possíveis e da redução do valor do parâmetro suporte mínimo.

Probability	Importance	Rule
0,396	0,353	guarnições = Existing -> carnes = Existing
0,384	0,391	refrigerantes pet = Existing -> carnes = Existing
0,357	0,466	pizzas = Existing -> refrigerantes pet = Existing
0,351	0,347	sobremesas = Existing -> carnes = Existing
0,349	0,344	entradas e saladas = Existing -> carnes = Existing
0,323	0,388	cardápio executivo = Existing -> sobremesas = Existing
0,296	0,397	refrigerantes pet = Existing -> entradas e saladas = Existing
0,293	0,347	sobremesas = Existing -> massas = Existing
0,290	0,312	refrigerantes lata = Existing -> massas = Existing
0,271	0,407	entradas e saladas = Existing -> refrigerantes pet = Existing
0,269	0,171	refrigerantes pet = Existing, entradas e saladas = Existing -> carnes = Existing
0,269	0,356	massas = Existing -> sobremesas = Existing
0,269	0,376	carnes = Existing -> sobremesas = Existing
0,268	0,437	carnes = Existing -> refrigerantes pet = Existing
0,267	0,373	carnes = Existing -> entradas e saladas = Existing
0,261	0,168	refrigerantes lata = Existing -> carnes = Existing

Figura 6.9: Regras geradas com o valor mínimo de suporte igual a 1%, a partir da criação das subcategorias de bebidas.

Para evidenciar que este tipo de análise trouxe um melhor valor de qualidade para as informações criadas, o estudo toma como exemplo as regras que contêm as categorias ‘massas’ ou ‘pizzas’ com uma das subdivisões criadas anteriormente para a categoria ‘bebidas’.

A regra gerada para a categoria ‘pizza’ indica que os produtos desta categoria indicam uma propensão de venda em conjunto com os refrigerantes do tipo pet, enquanto que as ‘massas’ possuem uma tendência maior de compra quando feitas em conjunto com os refrigerantes em lata.

O diagrama de dependências gerado pelo modelo atual (figura 6.10) ilustra de forma gráfica os tipos de comportamentos encontrados. Além destes dois comportamentos citados acima, este diagrama também exibe a dependência entre o tipo de bebida com os produtos relacionados à categoria ‘carnes’.

Como pode ser observada, a predição entre ‘refrigerantes pet’ e a categoria ‘carnes’ pode ser feita tanto de um quanto para o outro, ao passo que a predição entre ‘refrigerantes lata’ e ‘carnes’ é feita apenas da primeira para segunda. Desta maneira, o retorno poderá ser maior caso os itens da categoria ‘carnes’ tenham como sugestão os refrigerantes do tipo pet e não os do tipo lata.

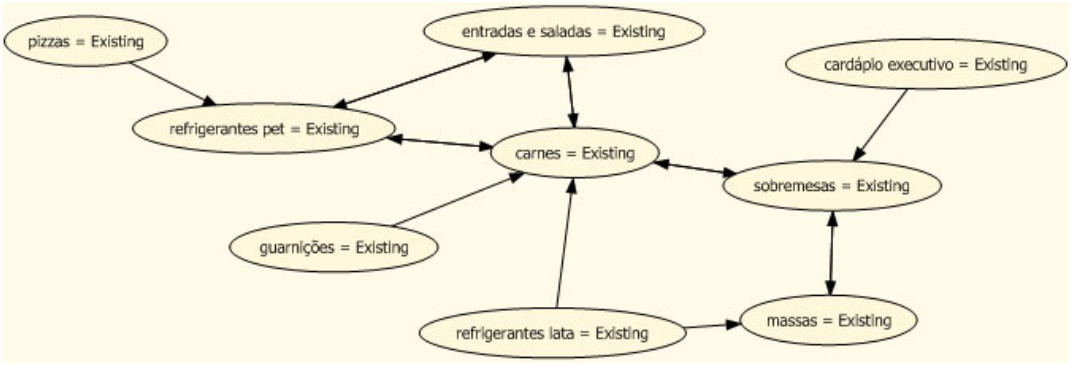


Figura 6.10: Diagrama de dependências das regras geradas a partir da criação das subcategorias de bebidas.

Estes são exemplos que ilustram as diversas maneiras possíveis de se gerar conhecimento a partir das regras de associação, indicando que novas subdivisões podem ser feitas a fim de se tentar extrair conhecimentos específicos a respeito do negócio analisado.

Como sugestão para os analistas de marketing, podemos citar a possibilidade de avaliar o comportamento de compra relacionado ao consumo de produtos do tipo *light*, que é encontrado inclusive nos produtos ligados a categoria ‘bebidas’. Isto já é inclusive outra subdivisão possível para este tipo de categoria.

Outra possibilidade é avaliar os produtos de forma independente para se tentar uma campanha com algum fornecedor. Como exemplo, o estudo fez uma análise com os pedidos que continham pelo menos um item da subcategoria de produtos denominada ‘cerveja’. A idéia era simular uma campanha para um fornecedor deste tipo de produto e que queira testar uma nova cerveja junto aos clientes do La Mole.

O conjunto de dados para esta análise foi limitado para conter apenas os pedidos com mais de um item e que tenham pelo menos um item da subcategoria cerveja, o que fez com que este conjunto de dados ficasse bem reduzido. Então, a partir deste conjunto foi criado um modelo para buscar as regras de associação possíveis para este tipo de produto, que tem seu resultado visualizado nas figuras 6.11, 6.12 e 6.13.

Probability	Importance	Rule
1,000		sobremesas = Existing -> cerveja = Existing
1,000		refrigerantes pet = Existing -> cerveja = Existing
1,000		carnes = Existing -> cerveja = Existing
1,000		refrigerantes lata = Existing -> cerveja = Existing
1,000		massas = Existing -> cerveja = Existing
1,000		entradas e saladas = Existing -> cerveja = Existing
0,405		cerveja = Existing -> entradas e saladas = Existing
0,286		cerveja = Existing -> massas = Existing
0,238		cerveja = Existing -> refrigerantes lata = Existing
0,190		cerveja = Existing -> carnes = Existing
0,190		cerveja = Existing -> refrigerantes pet = Existing
0,190		cerveja = Existing -> sobremesas = Existing

Figura 6.11: Regras de associação geradas para os produtos relacionados à cerveja.

Doze regras foram geradas por esta análise, sendo que as seis primeiras possuem o item cerveja como conseqüente, e a seis últimas possuem a cerveja como antecedente. Este comportamento pode ser visualizado na figura 6.13. Como o principal objetivo é saber quais os itens mais comuns que levam à compra de cerveja, o estudo analisará apenas as seis primeiras.

Support	Size	Itemset
17	2	entradas e saladas = Existing, cerveja = Existing
12	2	massas = Existing, cerveja = Existing
10	2	refrigerantes lata = Existing, cerveja = Existing
8	2	carnes = Existing, cerveja = Existing
8	2	refrigerantes pet = Existing, cerveja = Existing
8	2	sobremesas = Existing, cerveja = Existing

Figura 6.12: Suporte das regras geradas para os produtos relacionados à cerveja.

Como pode ser visto na figura acima, as melhores categorias de produtos para se sugerir a compra de cerveja são: ‘entradas e saldadas’, ‘massas’ e ‘carnes’. As outras regras não trazem um bom valor em suas informações, pois são apenas acompanhamentos dos pedidos realizados.

O exemplo desta análise mostra mais uma sugestão para os diversos tipos de campanhas possíveis para se gerar valor à empresa a partir desta técnica.

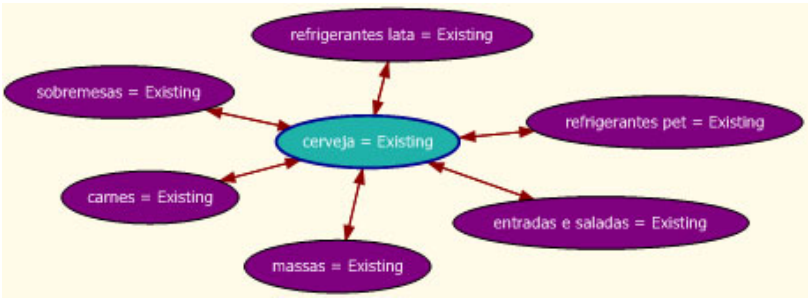


Figura 6.13: Diagrama de dependências das regras geradas para os produtos relacionados à cerveja.

A ferramenta disponibiliza um recurso para se buscar os itens mais prováveis a partir de uma simples consulta ao modelo gerado. Este mecanismo de busca é feito através de uma linguagem denominada *Data Mining Extensions* (DMX) [10] [15], que segue uma série de especificações técnicas para buscar as informações contidas nos modelos de mineração de dados.

O funcionamento deste recurso é explicado logo em seguida pelos comandos contidos na listagem 6.1.

```
select (predict([01-Valor_Minimo].[v DW Itens Pedido Maior
1],INCLUDE_STATISTICS,5)) as [recomendacao] from [01-Valor_Minimo]
natural prediction join
(select ((select 'couvert família' as produto)) as [v DW Itens Pedido
Maior 1]) as modelo
```

Listagem 6.1: Consulta ao modelo de produtos para o produto '*couvert* família'.

A listagem acima representa o comando que solicita uma busca pela predição dos cinco produtos mais indicados para o produto '*couvert* família', com retorno inclusive dos valores estatísticos encontrados para esta consulta. O resultado obtido pode ser visualizado na figura 6.14.

recomendacao

Produto	$SUPPORT	$PROBABILITY	$ADJUSTEDPR..
coca cola pet	1103	0,19281045751...	0,62434397431...
medalhão à piamontesa com batatas portuguesas	813	0,13725490196...	0,62025309002...
coca cola light pet	766	0,12418300653...	0,61546785731...
torta alemã	679	0,04296924439...	0,04205914833...
lasagna à bolognesa	510	0,03227439564...	0,03175913545...

Figura 6.14: Sugestão de 5 produtos com maior probabilidade de compra para o produto '*couvert* família'.

O resultado desta busca retorna quatro campos, que são: produto, suporte, a probabilidade (confiança) e a probabilidade ajustada (confiança ajustada). O primeiro campo representa o nome do produto a ser sugerido, o segundo representa o valor de suporte da indicação deste produto, o terceiro representa o valor da probabilidade da ocorrência de compra deste produto e por último o valor ajustado da probabilidade do produto indicado.

Nem sempre os itens com a maior probabilidade de compra são os melhores a serem sugeridos, como é o caso dos comportamentos de compras já sabidos pelo

especialista de negócio da empresa. Por conta disso, a ferramenta oferece um recurso de ajuste da probabilidade que penaliza os itens de compra mais populares.

Para o exemplo acima, mesmo quando utilizados os valores ajustados, a ordem dos produtos indicados continua a mesma. O modelo deste exemplo, citado na listagem 6.1, pode ser feito para qualquer outro produto, inclusive para a compra em conjunto entre os produtos. O código contido na listagem 6.2 serve de exemplo para um melhor entendimento:

```
select (predict([01-Valor_Minimo].[v DW Itens Pedido Maior
1],INCLUDE_STATISTICS,5)) as [recomendacao] from [01-Valor_Minimo]
natural prediction join
(select ((select 'couvert família' as produto)
union ((select 'lasagna à bolognesa' as produto)))
as [v DW Itens Pedido Maior 1]) as modelo
```

Listagem 6.2: Consulta ao modelo de produtos para a compra em conjunto dos produtos '*couvert* família' e '*lasgana* à *bolognesa*'.

O resultado encontrado pode ser visualizado na figura 6.15 que exibe a lista dos cinco produtos mais indicados, com os seus respectivos valores de suporte e confiança.

recomendacao

Produto	$SUPPORT	$PROBABILITY	$ADJUSTEDPR...
coca cola pet	1103	0,19281045751...	0,62434397431...
medalhão à piamontesa com batatas portuguesas	813	0,13725490196...	0,62025309002...
coca cola light pet	766	0,12418300653...	0,61546785731...
torta alemã	679	0,04296924439...	0,04205914833...
mousse de chocolate	481	0,03043918491...	0,02998057791...

Figura 6.15: Sugestão de 5 produtos com maior probabilidade de compra em conjunto dos produtos '*couvert* família' '*lasgana* à *bolognesa*'.

Embora as ordens das sugestões feitas pelo modelo tenham se mantido a mesma, observar-se em ambas as consultas que o valor da probabilidade ajustada entre os produtos 'coca cola pet' e 'medalhão à *piamontesa*' ficaram muito mais próximos do que os valores da probabilidade normal.

Isto indica que a penalização feita para o produto com maior probabilidade de compra não surtiu muito efeito para alterar a ordem da indicação. Porém, foi possível notar que alguns produtos ficaram com uma maior probabilidade se comparados as suas respectivas probabilidades normais.

Este recurso pode ser aplicado a vários produtos, permitindo sugestões de compra dos mais variados produtos. Vale destacar que este recurso pode ser aplicado a qualquer modelo gerado, como por exemplo, o da categoria de produtos e o que contém a subdivisão da categoria bebidas.

Terminada a análise de regras de associação entre os produtos e suas respectivas classificações, o estudo segue adiante na busca por outro modelo possível para a utilização desta técnica, que diz respeito aos comportamentos de compras realizados a partir das características de um determinado cliente.

A obtenção das características de cada usuário é fornecida pelo cadastro que o mesmo é obrigado a fazer quando efetua o primeiro pedido, o que torna possível refinar os itens a serem sugeridos para cada cliente a partir dos dados informados. Desta forma, as sugestões são feitas levando em consideração todos os usuários que contenham um mesmo perfil e seus respectivos itens de compra incluídos nos seus pedidos.

O resultado obtido através deste tipo de análise pode ser observado na figura 6.16, que apresenta as regras geradas a partir de um suporte mínimo de 1,5% e um valor de confiança mínima de 20%.

Probability	Importance	Rule
0,311	0,284	Loja = Tijuca, Sexo = F -> couvert família = Existing
0,264	0,207	coca cola pet = Existing -> couvert família = Existing
0,250	0,186	Bairro Moradia = Tijuca, Loja = Tijuca -> couvert família = Existing
0,250	0,202	Loja = Tijuca -> couvert família = Existing
0,247	0,176	Loja = Ilha Plaza -> couvert família = Existing
0,240	0,159	Faixa Etaria = 35 - 39, Sexo = M -> couvert família = Existing
0,233	0,153	Bairro Moradia = Tijuca -> couvert família = Existing
0,231	0,142	Loja = Norte Shopping -> couvert família = Existing
0,220	0,125	Faixa Etaria = 35 - 39 -> couvert família = Existing
0,206	0,089	Loja = Tijuca, Sexo = M -> couvert família = Existing
0,203	0,084	Faixa Etaria = 25 - 29, Sexo = F -> couvert família = Existing

Figura 6.16: Regras de associação geradas com os atributos dos usuários.

Este resultado pouco incrementou ao processo de extração de conhecimento, uma vez que as regras geradas ficaram bastante tendenciosas para o produto '*couvert* família'.

Por este motivo, o estudo criou um novo modelo (com suporte e confiança valorados em 1%) excluindo os produtos mais vendidos, que são: '*couvert* família',

‘medalhão à *piamontesa* com batatas portuguesas’, ‘coca cola pet’ e ‘coca cola light pet’. O resultado pode ser observado na figura 6.17.

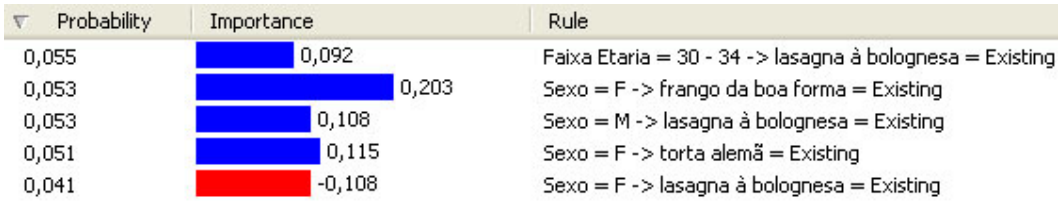


Figura 6.17: Regras de associação geradas com os atributos dos usuários sem os produtos mais vendidos.

Apesar de ter oferecido um conjunto de regras diferentes das análises anteriores, o resultado obtido trouxe um baixo valor de confiança para cada regra, o que prejudica a qualidade das mesmas.

De qualquer forma, esta análise já serviu para visualizarmos uma regra com a correlação negativa, indicada pela cor vermelha na imagem acima. Isto significa dizer que os clientes do sexo feminino não possuem uma tendência de compra para o produto ‘*lasagna* à bolognesa’.

Em contrapartida, é possível visualizar uma regra que indica que usuários deste sexo possuem uma tendência de compra ao produto ‘*lasagna* à boa forma’, porém vale ressaltar que ambas as regra geradas por este modelo possuem um baixo valor de probabilidade de ocorrência.

Por possuírem baixos valores de suporte e confiança, o estudo criou um novo modelo para tentar buscar um melhor resultado a partir dos atributos dos usuários. A única diferença deste novo modelo, se comparado ao utilizado pelo modelo anterior, é que agora o estudo passará a levar em consideração o atributo categoria ao invés do produto.

As regras produzidas podem ser observadas na figura 6.18, porém vale uma ressalva: para melhor visualizar as regras criadas, a análise excluiu deste modelo as compras efetuadas na loja Centro, pois as mesmas estavam criando muitas regras sem que trouxessem um conhecimento específico ao negócio.

Probability	Importance	Rule
0,423	0,347	Bairro Moradia = Centro, Sexo = M -> Categoria = bebidas
0,341	0,195	Bairro Moradia = Leblon -> Categoria = carnes
0,337	0,191	Faixa Etaria = 45 - 49, Sexo = F -> Categoria = carnes
0,322	0,236	Bairro Moradia = Centro -> Categoria = bebidas
0,305	0,146	Bairro Moradia = Copacabana, Sexo = F -> Categoria = carnes
0,302	0,149	Bairro Moradia = Copacabana -> Categoria = carnes
0,298	0,136	Bairro Moradia = Copacabana, Sexo = M -> Categoria = carnes
0,294	0,134	Faixa Etaria = 45 - 49 -> Categoria = carnes
0,280	0,166	Faixa Etaria = 40 - 44, Sexo = F -> Categoria = bebidas
0,269	0,090	Faixa Etaria = 50 - 54 -> Categoria = carnes
0,262	0,077	Bairro Moradia = Botafogo, Sexo = M -> Categoria = carnes
0,262	0,143	Faixa Etaria = 40 - 44 -> Categoria = bebidas
0,257	0,131	Faixa Etaria = 25 - 29, Sexo = M -> Categoria = bebidas
0,255	0,119	Bairro Moradia = Centro, Sexo = F -> Categoria = bebidas
0,250	0,335	Faixa Etaria = 55 - 59 -> Categoria = massas

Figura 6.18: Regras de associação geradas com os atributos dos usuários utilizando o atributo categoria.

O resultado obtido permite identificar melhor os atributos que diferenciam a intenção de compra para cada perfil de usuário, como por exemplo:

- Usuários que moram no Centro e são do sexo masculino, possuem uma tendência de compra à categoria de bebidas; e
- Usuários que moram em Copacabana e são do sexo feminino, possuem uma tendência de compra à categoria de carnes.

Estas predições também podem ser realizadas a partir de consultas DMX, como mostra listagem 6.3.

```
SELECT (PREDICT(Categoria)) From [04-AR-Atributo-Categoria]
NATURAL PREDICTION JOIN
(SELECT 'Centro' AS [Bairro Moradia], 'M' AS [Sexo]) AS t
```

Listagem 6.3: Consulta ao modelo de categoria através da linguagem DMX.

Por último, o estudo buscou a criação de mais um modelo com o objetivo de indicar produtos a serem exibidos na página inicial da loja virtual sem que ao menos se saiba qual é o perfil do cliente.

Para efetuar esta tarefa, o estudo foi em busca de um modelo que faça a predição das categorias de produtos a partir dos atributos dia da semana. A melhor maneira para visualizar este resultado é o diagrama de dependências mostrado na figura 6.20, porém

na figura 6.19, que não exibe todas as regras, é realçado que algumas destas regras possuem valor de importância negativo.

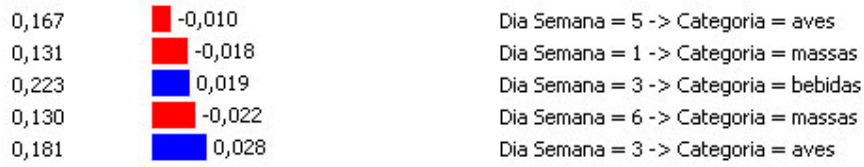


Figura 6.19: Regras de associação com algumas variações de importância negativas.

Esta variação negativa indicada na figura acima demonstra que para certos dias da semana não convém sugerir produtos de determinadas categorias, como é o caso da sexta-feira com a categoria ‘massas’.

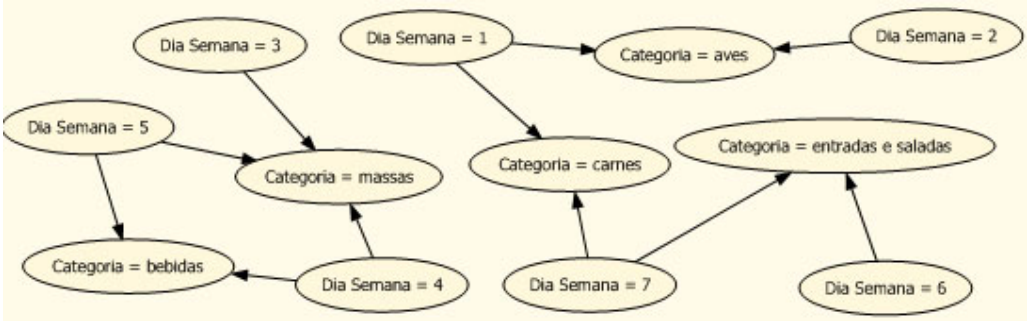


Figura 6.20: Diagrama de dependências das regras geradas para os produtos relacionados à cerveja.

Com o resultado da figura acima fica nítido observar os melhores tipos de categorias a serem indicados por cada dia da semana. Vejamos os exemplos abaixo:

Na segunda-feira é possível observarmos que a categoria ‘aves’ é sem dúvida a melhor a ser indicada, assim como a categoria ‘massas’ é a melhor indicada às terças-feiras. Porém, se for uma quinta-feira o melhor é indicar ‘massas’ com ‘bebidas’.

6.3 Conclusão

O resultado final obtido irá complementar uma função existente na loja virtual e que é feita atualmente de uma maneira manual e fixa, que é a sugestão de produtos aos clientes.

Fica a ressalva de que os resultados gerados por este tipo de análise não dependem apenas dos parâmetros relacionados à execução do método em questão, mais sim do próprio modelo de negócio e da maneira em que estão dispostos os dados.

Desde a primeira análise realizada, quando o estudo buscou encontrar algum padrão de compra relacionado especificamente aos produtos, ficou claro que o resultado depende das várias maneiras de se trabalhar o conjunto de dados em questão.

O estudo mostrou que uma grande quantidade de modelos deve ser criada até se chegar ao mais adequado. Tudo isso demonstra o enorme esforço necessário para se conseguir um bom resultado a partir desta técnica, o que demonstra que não é uma fácil tarefa extrair conhecimentos a partir deste tipo de método.

Além disso, os resultados produzidos devem ser avaliados por alguém que conheça o tipo de negócio, uma vez que o conteúdo de algumas regras pode trazer uma informação já conhecida de antemão ou que não contenha valor significativo ao processo de negócio.

7 Classificação Supervisionada

7.1 Introdução

Depois de termos entendido as características que formam os diferentes conjuntos de clientes, assim como a forma pela qual efetuam seus pedidos, o estudo passa agora a focar na criação de modelos que sejam capazes de inferir a classificação de novos registros a partir dos padrões existentes na base de dados.

O objetivo é a criação de dois modelos, sendo: um para a classificação dos novos clientes (se ativo ou não) e outro para a classificação do valor do pedido a ser feito por estes novos clientes.

O desenvolvimento desta tarefa foi realizado com o uso do algoritmo de árvore de decisão. Para mensurar o resultado de cada modelo o estudo utilizou os valores obtidos pela validação cruzada, já que este tipo de validação permite a utilização de todos os registros contidos na base de dados para o treinamento e teste.

Vale ressaltar que esta forma de validação não garante uma melhora na exatidão dos classificadores, ela apenas providencia uma estimativa mais justa de precisão ao reduzir os riscos de super ajuste no momento de treinamento dos modelos.

7.2 Análise e Desenvolvimento

Criar um modelo que consiga estimar a probabilidade de um novo usuário ser ativo ou não, dado os valores contidos em seus atributos de cadastro, ajudará na segmentação das futuras campanhas a serem oferecidas pelo site logo após o cadastro de um novo cliente.

Como exemplo de estratégia a ser adotado, um bom uso seria considerar a probabilidade de um cliente ser ativo ou não para monitorar o seu comportamento tão logo o mesmo efetue o seu cadastro.

Se o classificador indicar que o usuário cadastrado possui uma tendência a ser um cliente ativo, um alerta seria enviado ao departamento de marketing da empresa assim que o cliente abandonasse o site sem fazer um pedido. Isto daria uma opção de

contato em um momento de intenção de compra do cliente, o que aumentaria a chance de retorno para a empresa.

Desta forma, a equipe de vendas poderia entrar em contato para oferecer uma boa refeição antes que a fome do usuário fosse substituída por outro produto, tanto de um concorrente quanto de um substituto.

Para os casos em que o classificador indicar um novo usuário como um cliente não ativo, um alerta poderia ser enviado ao departamento de relacionamento com os clientes. Neste caso, o objetivo seria contatar o mesmo para entender as razões que o levaram a se cadastrar no site sem que fosse feito um pedido.

Em relação aos pedidos, uma boa utilização seria estimar a faixa de valor do pedido de um novo cliente para poder oferecer os produtos que mais se adéqüem à sua possibilidade de compra.

A utilização deste modelo pode servir de complemento aos resultados obtidos pelas regras de associação, já que possibilita um auxílio no processo de tomada de decisão, pois, ao estimar o valor do pedido, é possível se ter uma idéia da quantidade de produtos que poderá ser oferecida ao cliente.

Assim, aos clientes com uma estimativa de baixo valor seriam oferecidos apenas os produtos principais, enquanto que para os clientes de maior valor poderiam ser oferecidas as refeições principais acompanhadas de bebidas e sobremesas.

7.2.1 Modelo Classificador de Clientes

A primeira análise realizada serviu para avaliar os resultados obtidos a partir da variação dos atributos utilizados por cada modelo construído pelo algoritmo de árvore de decisão.

Esta análise permitiu avaliar quais os atributos de entrada são mais susceptíveis de fornecer as melhores informações em relação ao atributo de saída. Em conjunto, esta análise também avaliou os resultados obtidos a partir de distintos valores de parametrização possíveis. Estes parâmetros são descritos logo em seguida.

- *COMPLEXITY_PENALITY* – controla o crescimento da árvore de decisão;
- *MINIMUN_SUPPORT* – valor mínimo de casos para que seja realizada uma divisão na árvore de decisão;
- *SCORE_METHOD* – método a ser utilizado para calcular a pontuação da divisão; e
- *SPLIT_METHOD* – método utilizado para dividir os nós da rede.

Os atributos utilizados em cada modelo podem ser visualizados logo em seguida na tabela 7.1, enquanto que seus resultados são encontrados na tabela 7.2.

Atributos	**AD-01**	**AD-02**	**AD-03**	**AD-04**	**AD-05**	**AD-06**	**AD-07**	**AD-08**
Bairro	■	■	■	■		■		■
Dia Semana		■	■	■	■		■	
Faixa Etária			■	■	■	■		
Feriado		■	■	■	■		■	
Idade	■	■					■	■
Loja	■	■	■	■	■	■	■	■
Região					■		■	
Sexo				■	■	■	■	■

Tabela 7.1: Atributos utilizados por cada modelo no conjunto de dados de clientes.

A primeira coluna da tabela 7.1 representa os atributos possíveis de utilização para cada modelo, enquanto que as demais colunas representam os modelos gerados por esta primeira análise.

Modelo	% Acertos	Desvio Padrão	RMS	Pontuação Logarítmica
AD-01	49,34%	9,6495	0,4516	-0,8471
AD-02	50,60%	11,258	0,4562	-0,9491
AD-03	52,95%	12,1547	0,4591	-0,7732
AD-04	53,21%	13,3863	0,4575	-0,7875
AD-05	52,89%	12,897	0,4631	-0,7668
AD-06	52,89%	9,6515	0,4566	-0,7294
AD-07	51,40%	13,9159	0,4613	-0,9433
AD-08	50,00%	8,9959	0,4481	-0,8875

Tabela 7.2: Resultados obtidos por cada modelo da tabela 7.1.

Em relação à tabela 7.2, a primeira coluna representa os modelos gerados. Já a segunda coluna representa a taxa percentual de acertos obtida em cada classificador. Em seguida, na terceira coluna, é visualizado o valor do desvio padrão, que representa uma forma de determinar o intervalo de confiança e estimar a variância de cada modelo.

Na quarta coluna é apresentada outra métrica fornecida pela ferramenta, que é o resultado produzido pela raiz quadrada do erro médio quadrático (RMS). Esta métrica possui o seu valor entre zero e infinito, do qual zero corresponde ao valor ideal.

Por último é analisada a variação da probabilidade da pontuação logarítmica, que representa um valor que não pode exceder à zero. Para esta métrica, quanto mais próximo de zero melhor é o resultado do modelo.

Como pode ser visualizado, o resultado obtido, de uma forma geral, foi considerado de pequena representatividade, visto que o melhor modelo encontrado teve apenas uma capacidade de predição de 53,21%. Este é um valor muito baixo para considerarmos como um bom resultado de um modelo classificador.

Quanto aos resultados obtidos a partir da variação dos atributos utilizados por cada modelo, percebe-se somente que há uma pequena diferença nos valores encontrados, não sendo possível identificar um atributo que tenha sido mais favorável a algum modelo.

Estas métricas utilizadas até agora nos dão uma visão geral do resultado de cada classificador, porém elas não nos trazem o quanto de cada classe foi predita corretamente por cada classificador. Por isso é importante avaliar a matriz de confusão

produzida para analisar se o resultado está ou não tendencioso para um determinado valor de predição.

Para os modelos acima, estas matrizes podem ser visualizadas na tabela 7.3.

Modelo AD-01			**Modelo AD-02**		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	23,18%	24,74%	**Ativo**	24,61%	24,90%
Não Ativo	25,93%	26,15%	**Não Ativo**	24,50%	25,99%
Modelo AD-03			**Modelo AD-04**		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	14,14%	12,09%	**Ativo**	14,05%	11,73%
Não Ativo	34,97%	38,80%	**Não Ativo**	35,06%	39,16%
Modelo AD-05			**Modelo AD-06**		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	14,18%	12,18%	**Ativo**	12,97%	10,97%
Não Ativo	34,93%	38,71%	**Não Ativo**	36,14%	39,93%
Modelo AD-07			**Modelo AD-08**		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	24,88%	24,37%	**Ativo**	23,50%	24,39%
Não Ativo	24,23%	26,52%	**Não Ativo**	25,61%	26,50%

Tabela 7.3: Resultados das matrizes de confusão da dos diversos modelos de clientes da primeira análise.

Podemos perceber que alguns modelos possuem um forte desbalanceamento entre as classes, em especial a favor da classe de clientes não ativos. Isto indica que é mais fácil, para estes modelos, acertar os clientes não ativos do que os ativos.

Como ponto de observação vale ressaltar que os modelos com um melhor balanceamento entre as classes utilizaram o atributo idade ao invés do atributo faixa etária, o que nos leva a excluir o atributo faixa etária das demais análises a serem realizadas.

Em busca de um melhor resultado o estudo prosseguiu com uma segunda análise para avaliar a criação de novos modelos com base na separação do conjunto de clientes em diversas características conforme descrito abaixo:

- Masculino – Conjunto de dados com os clientes de sexo masculino;
- Feminino – Conjunto de dados com os clientes de sexo Feminino;
- Dia Útil – Conjunto de dados dos clientes com data de cadastro realizada em um dia útil, ou seja, dia da semana entre segunda-feira e sexta-feira e que não tenha sido feriado; e
- Fim de Semana + Feriado – Conjunto de dados com os clientes cadastrados nos sábados, domingos e feriados.

Os atributos utilizados por estes novos modelos são descritos na tabela 7.4 e seus resultados são descritos na tabela 7.5.

Atributos	Masculino	Feminino	Dias Úteis	Fim de Semana + Feriado
Bairro	■	■	■	■
Dia Semana	■	■		■
Feriado				■
Idade	■	■	■	■
Loja	■	■	■	■
Sexo			■	■

Tabela 7.4: Atributos utilizados por cada modelo de subconjunto de clientes.

Modelo	% Acertos	Desvio Padrão	RMS	Pontuação Logarítmica
Masculino	49,63%	10,3179	0,4626	-0,9693
Feminino	52,28%	6,5872	0,4603	-0,957
Dias Úteis	50,96%	11,8524	0,4585	-0,9846
Fim de Semana Feriado	51,04%	6,7024	0,4561	-0,9691

Tabela 7.5: Resultados obtidos por cada modelo da tabela 7.3.

O resultado encontrado pelo modelo feminino foi o que obteve o maior valor para a taxa de acerto. Quanto aos dias da semana, é possível perceber uma pequena variação entre os resultados a favor do modelo dos fins de semana e feriados, já que contém uma taxa de acerto maior e um desvio padrão menor. Para complementar o entendimento destes resultados, na tabela 7.6 é apresentado as matrizes de confusão destes modelos.

Modelo Masculino			Modelo Feminino		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	30,84%	28,94%	**Ativo**	24,15%	25,48%
Não Ativo	21,44%	18,78%	**Não Ativo**	22,24%	28,13%
Modelo Dias Úteis			**Modelo Fim de Semana + Feriado**		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	25,00%	26,08%	**Ativo**	28,39%	27,01%
Não Ativo	22,97%	25,96%	**Não Ativo**	21,95%	22,65%

Tabela 7.6: Resultados das matrizes de confusão dos diversos modelos de clientes da segunda análise.

O balanceamento da maioria dos modelos ficou mais equilibrado se comparado aos da primeira análise. Ainda assim, podemos perceber que, no caso do modelo masculino o desbalanceamento é a favor dos clientes ativos, enquanto que no modelo feminino o desbalanceamento é a favor aos clientes não ativos.

Em relação aos dias úteis, podemos perceber que as classes ficaram bem balanceadas, enquanto que nos fins de semana e feriados o classificador ficou um pouco mais tendencioso aos clientes ativos.

Por último o estudo tentou obter um melhor resultado a partir da subdivisão do conjunto de clientes por regiões, seguindo o mesmo modelo adotado no capítulo 5. Esta terceira análise foi iniciada com a região Zona Norte I para que fosse possível avaliar os resultados obtidos antes de seguir com as demais regiões.

Assim, mais dois modelos foram criados, tendo o primeiro os seguintes atributos: bairro, loja, sexo e idade; enquanto que o segundo utilizou, além destes atributos, dia da semana e feriado. Os resultados obtidos podem ser visualizados na tabela 7.7, enquanto que as matrizes de confusão podem ser visualizadas na tabela 7.8.

Modelo	**% Acertos**	**Desvio Padrão**	**RMS**	**Pontuação Logarítmica**
AD-Z-Norte-I-a	53,19%	7,1177	0,4497	-0,8789
AD-Z-Norte-I-b	50,85%	6,775	0,4515	-1,0009

Tabela 7.7: Resultados obtidos pelos modelos da região Z. Norte I.

Modelo AD-Z-Norte-I-a			AD-Z-Norte-I-b		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	24,74%	25,18%	**Ativo**	24,80%	27,58%
Não Ativo	21,63%	28,46%	**Não Ativo**	21,57%	26,05%

Tabela 7.8: Resultados das matrizes de confusão dos modelos Z. Norte I.

O primeiro modelo levou vantagem em relação ao segundo, o que nos leva a crer que os melhores atributos a serem utilizados são: bairro, loja, sexo e idade. Por isso o estudo seguiu adiante para criar os últimos modelos que representam cada região criada.

Os resultados destes novos modelos podem ser vistos na tabela 7.9. Porém, apenas os modelos que obtiveram um resultado superior a 50% têm suas matrizes de confusão exibidas na tabela 7.10.

Modelo	% Acertos	Desvio Padrão	RMS	Pontuação Logarítmica
Z-Sul-III	52,05%	4,5864	0,4523	-0,9155
Z-Sul-I	49,46%	5,6562	0,4491	-0,8906
Barra	48,09%	2,8674	0,452	-0,828
Centro	50,08%	2,5111	0,448	-0,8501
Z-Norte-II	40,80%	2,522	0,477	-1,0009
Ilha	56,07%	3,7731	0,4602	-0,9242
Z-Norte-III	57,20%	2,8895	0,4381	-0,9227
Z-Sul-II	44,37%	1,6742	0,4611	-1,0192
Niteroi	46,40%	2,6542	0,4659	-1,0327

Tabela 7.9: Resultados obtidos dos modelos por região.

Z-Sul-III			Centro		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	28,11%	28,18%	**Ativo**	25,66%	27,10%
Não Ativo	19,77%	23,94%	**Não Ativo**	22,82%	24,42%
Ilha			**Zona-Norte-III**		
	Ativo	**Não Ativo**		**Ativo**	**Não Ativo**
Ativo	40,66%	26,62%	**Ativo**	44,40%	25,96%
Não Ativo	17,32%	15,40%	**Não Ativo**	16,84%	12,80%

Tabela 7.10: Resultados das matrizes de confusão dos modelos por região.

Pelos resultados da tabela 7.9 podemos ver que alguns classificadores obtiveram um péssimo desempenho, como é o caso do classificador da região Zona Norte II. Este classificador obteve apenas 40,80% de acerto, o que indica que o seu conteúdo não é suficiente para a criação de um modelo classificador a partir dos seus dados.

O modelo com a maior taxa de acerto é o da região Zona Norte III, que um resultado igual a 57,20%. Porém, ao analisarmos a sua respectiva matriz de confusão, podemos perceber que este classificador está tendencioso aos clientes ativos.

Esta análise nos ajudou a perceber que a variação entre os resultados de cada modelo está diretamente associada com a qualidade das informações contidas em cada subconjunto de dados. É isto que faz com que tenhamos uma diferença nas taxas de acertos obtidas, ou seja, alguns subconjuntos são mais suscetíveis à segregação dos clientes em comparação a outros.

Com esta informação podemos inferir que os modelos criados na primeira análise, a partir de todo o conjunto de dados de clientes, trazem consigo a qualidade dos dados contidos em cada subconjunto de região. Deste modo, podemos dizer que há nestes modelos as certezas relacionadas ao subconjunto de dados da Zona Norte III, assim como as incertezas do subconjunto da Zona Norte II.

No entanto, embora o estudo tenha percebido a importância dos diversos tipos de análises e a avaliações que devem ser feitas ao se criar um modelo classificador, os resultado obtidos, de uma forma geral, ainda são considerados de pequena representatividade.

O estudo já esperava encontrar uma dificuldade na criação de um modelo classificador, pois, como citado nos capítulos 4 e 5, já havíamos visto o quanto é homogêneo o conjunto de dados de clientes. Inclusive, em diversos momentos do capítulo 5, vimos que vários clusters se sobrepunham a outros, como demonstra um exemplo contido na figura 7.1.

Nesta figura é possível observarmos que em alguns casos o perfil do cliente ativo é praticamente igual ao do cliente não ativo, como demonstra o cluster 2 e 3. Em outros casos, como demonstra o detalhe em verde entre os clusters 1 e 5 e os clusters 9 e 12, a diferença é mínima.

Figura 7.1: Exemplos de clusters do conjunto de dados de clientes da região Z. Sul III.

É esta a razão pela qual exista tamanha incerteza nos resultados dos classificadores gerados pelo estudo. Porém, embora o resultado não seja tão representativo, podemos considerar que a utilização dos classificadores, com uma taxa de acerto maior do que 50%, pode ajudar a interagir de forma diferenciada com os clientes da empresa.

Isto por que a predição do tipo de cliente não é algo crítico, como por exemplo, o de uma financeira conceder ou não crédito a uma cliente ou de uma seguradora ao definir o perfil de risco de uma determinada pessoa.

Neste caso, o maior risco é importunar um consumidor que realmente tem a intenção de ser tornar ativo enquanto o mesmo está escolhendo sua refeição, de modo que o cliente se sinta invadido pelo contato da empresa e desista de efetuar seu pedido. Não há problema em relação ao cliente não ativo, pois este cliente é de antemão um cliente "perdido" para a empresa.

Desta forma, a utilização do classificador será bastante útil se, além de predizer o tipo de cliente, informar a probabilidade encontrada pelo modelo para cada tipo de cliente, o que trará uma informação mais valiosa no momento de utilização dos modelos.

Para ilustrar como a predição de novos valores pode ser feita, assim como o valor de probabilidade, o estudo mostra abaixo um exemplo de uma consulta DMX no

modelo gerado para a região Z. Sul III. Em seguida são exibidos, na tabela 7.11, os resultados obtidos a partir da variação de valores passados à mesma consulta DMX.

```
SELECT Predict([Z-Sul-III].[Ativo]) AS
[PREDICAO],PredictProbability([Z-Sul-III].[Ativo]) AS [PROBABILIDADE]
From [Z-Sul-III] natural prediction join
(select '29' as [Idade], 'Rio Sul' as [Loja], 'M' as [Sexo],
'Copacabana' as [Bairro]) AS B
```

Listagem 7.1: Exemplo de consulta ao modelo de Z.Sul-III.

O resultado gerado por esta consulta indica que a predição para um novo cliente com 29 anos de idade, atendida pela loja Rio Sul, com sexo masculino e morador do bairro de Copacabana é de um cliente ativo, com um uma probabilidade de 60%, como demonstra a linha mais escura da tabela abaixo.

Idade	Loja	Sexo	Bairro	Predição	%
29	-	-	-	Não Ativo	56,00
29	Ipanema	-	-	Não Ativo	56,25
29	Rio Sul	-	-	Não Ativo	53,84
29	Rio Sul	M	-	Ativo	55,55
29	Rio Sul	F	-	Não Ativo	71,42
29	Rio Sul	M	Ipanema	Ativo	42,85
29	Rio Sul	M	Copacabana	Ativo	60,00
29	Ipanema	M	Ipanema	Não Ativo	66,66
29	Ipanema	M	Copacabana	Não Ativo	60,00

Tabela 7.11: Exemplo de predições para os clientes da Zona Sul III.

A figura 7.2 mostra um exemplo da ramificação da árvore de decisão que representa parte dos itens da consulta acima.

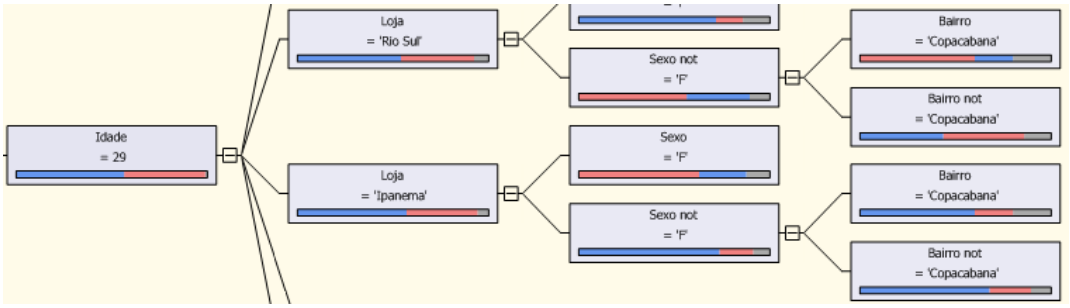


Figura 7.2: Ramificação da árvore de decisão Z-Sul-III.

Estes resultados ilustram como a probabilidade da predição pode ajudar no momento de contatar um cliente, pois, em casos onde múltiplos cadastros ocorram, o valor obtido por esta probabilidade pode permitir a priorização de contato.

7.2.2 Modelo Classificador do Valor do Pedido

Neste tópico o estudo tenta buscar um modelo classificador que seja capaz de inferir o valor do pedido a ser feito por um novo cliente. Para este tópico, o estudo mostrará apenas o resultado final obtido, já que a maneira pela qual é feita a construção de um modelo classificador, através da ferramenta utilizada, foi ilustrada no tópico anterior.

De início o estudo tentou a criação de um modelo classificador a partir de todos os dados contidos na tabela de pedidos. Esta busca seguiu as mesmas características da análise anterior, onde inicialmente foram testados os melhores atributos a serem utilizados.

Neste caso, a primeira análise realizada levou em consideração os seguintes atributos: bairro de moradia, loja, sexo, idade, dia da semana, feriado e horário do pedido. O resultado obtido por este classificador teve um taxa de acerto de apenas 49%.

Por conta disso, o estudo seguiu em busca de melhores resultados a partir da construção de diversos classificadores com as distintas combinações possíveis de atributos.

Assim como na análise de clientes, os resultados mais representativos foram conseguidos pela combinação dos atributos sexo, idade, loja e bairro de moradia. Esta combinação apresentou uma taxa de acerto de 53,84%, com um desvio padrão igual a 11,67, pontuação logarítmica igual a -1,2014 e RMS igual a 0,5464.

Ao gerar a matriz de confusão para este classificador o estudo pode perceber uma limitação da ferramenta. Esta limitação diz respeito à construção da matriz de confusão a partir da utilização dos resultados obtidos pela validação cruzada, já que a ferramenta não possui uma maneira automatizada de geração da matriz a partir deste tipo de validação.

O fato é que para se construir uma matriz de confusão a partir da validação cruzada se faz necessário uma consulta DMX, como demonstra o exemplo da listagem abaixo.

```
CALL SystemGetCrossValidationResults ([PedidoFull], [AD-01], 10, 0,
'Valor Pedido', '0-20')
```

Listagem 7.2: Consulta DMX que retorna a matriz de confusão obtida pela validação cruzada.

Esta listagem traz como retorno o valor da quantidade de registros classificadas corretamente para a classe pretendida, que neste caso representa a classe dos valores ‘0 - 20’. O problema é que os resultados das demais classes são agrupados, o que não torna possível saber o quanto se acertou e se errou em cada classe que não seja a pretendida.

Na análise dos clientes isto não era um problema porque tínhamos apenas 2 classes. Porém, como no conjunto de pedidos temos um total de 5 classes isto se torna um problema, já que se faz necessário repetir esta consulta pela quantidade de classes existentes.

Esta análise não possibilitou saber quais classes foram classificadas como sendo de outra classe específica e nem em qual quantidade. De qualquer forma, de um total de 15.802 registros, sabemos que 8.509 foram classificados corretamente, enquanto que 7.293 registros foram classificados incorretamente. A tabela 7.12 ilustra a quantidade de registros classificados corretamente por cada classe e percentual de acerto de cada uma.

Classe	% acertos	Quantidade de Registros
0 - 20	18,02	2848
20 - 40	31,50	4979
40 - 60	3,52	557
60 - 80	0,29	46
A partir de 80	0,49	79

Tabela 7.12: Resultados da matriz de confusão obtida pela validação cruzada.

Em seguida o estudo passou a avaliar a construção de diversos modelos a partir da subdivisão do conjunto de dados em distintas características. Em todos os modelos avaliados o estudo analisou o melhor resultado a partir da combinação de distintos atributos.

Porém, em todos os modelos, as taxas de acertos mais significantes foram conseguidas a partir do uso dos atributos sexo, idade, loja e bairro de moradia. Os resultados da tabela 7.13 exibem os valores obtidos por cada modelo gerado a partir destes atributos.

Modelo	% Acertos	Desvio Padrão	RMS	Pontuação Logarítmica
Dia da Útil	52,10	8,1078	0,4843	-1,1062
Fim de Semana + Feriados	49,80	5,6799	0,5317	-1,3069
Horário do Almoço	52,80	10,6357	0,5261	-1,1935
Horário da Tarde	48,37	7,9037	0,5706	-1,3343
Horário do Jantar	51,50	0,3002	0,4854	-1,1985
Z-Norte-I	52,45	5,2353	0,5028	-1,2147
Z-Sul-III	53,58	6,0617	0,5185	-1,1646
Z-Sul-I	48,79	5,9036	0,5262	-1,2087
Barra	52,41	3,4742	0,4911	-1,1669
Centro	54,03	2,2334	0,4782	-1,0984
Z-Norte-II	56,18	3,9115	0,5454	-1,1707
Ilha	51,74	5,4609	0,5282	-1,1858
Z-Norte-III	53,90	2,7702	0,5889	-1,2603
Z-Sul-II	45,49	2,4967	0,5929	-1,317
Niterói	56,22	3,1812	0,577	-1,2247
Barra + Dia Útil + Horário do Jantar	55,16	3,1018	0,5647	-1,2497
Centro + Dia Útil + Horário do Jantar	53,76	1,3737	0,59	-1,2759
Z-Norte-II + Dia Útil + Horário do Jantar	54,29	1,7074	0,5593	-1,3036

Tabela 7.13: Resultados dos modelos classificadores de pedidos feitos a partir de diferentes subconjuntos.

Podemos perceber que a separação dos dados ajudou, em alguns casos, a obtenção de taxas de acertos com valores acima de 50%. Já em outros casos, percebemos que os resultados ficaram abaixo de 50%, o que não é um resultado aceitável.

Ainda assim, podemos considerar que os resultados não foram muito expressivos. Porém, como o uso deste classificador será feito apenas para inferir o valor do pedido de um novo cliente para compor uma cesta de produtos adequada, o que não traz nenhum risco ao negócio da companhia, podemos dizer que não há problema na utilização dos classificadores que alcançaram uma taxa de acerto acima de 50%.

7.3 Conclusão

Quanto aos resultados dos classificadores, podemos notar que o produto obtido é conseqüente da qualidade do conjunto de dados analisado. Para o caso em questão, não foi possível a identificação de um atributo ou valor que permitisse a separação linear das mesmas, o que aumentaria as chances por um melhor resultado.

Este fato demonstra o quanto que o conjunto de dados analisado é homogêneo, o que indica que os clientes ativos possuem um perfil muito próximo dos não ativos, trazendo assim um elevado grau de incerteza para cada modelo. Além disso, a baixa quantidade de atributos possíveis à análise foi determinante para que o estudo não encontrasse bons classificadores.

Vale ressaltar que o resultado alcançado é dependente do modo pelo qual foram distribuídos os dados em cada uma das partições da validação cruzada. Esta distribuição é feita pela ferramenta utilizada de forma automática e independente, uma vez que não há parâmetros de configuração para esta funcionalidade. Sabemos apenas que os dados são distribuídos de forma a garantir que cada partição contenha um histograma similar para os estados de cada categoria analisada.

Ao término destas duas análises podemos concluir que é possível utilizar este tipo de técnica para criar novas operações de marketing a serem oferecidas pela empresa.

Isto se deve basicamente a sugestão de mudança do modo pelo qual a companhia se relaciona com os seus clientes, já que é possível fazer com que os dados obtidos pela Internet sejam utilizados para criar uma campanha de marketing ativo, o que faz com que a empresa tome a iniciativa de contato com os seus clientes.

8 Conclusões

Ao término deste estudo podemos concluir que é possível, através da utilização das técnicas de mineração de dados, extrair padrões de forma automática e adaptativa, a partir das informações existentes nas bases de dados das empresas, para se gerar o conhecimento necessário que dará suporte ao processo de tomada de decisões.

No caso em questão, observamos que a utilização destas técnicas, aplicadas aos dados contidos em uma base de dados transacional, pôde produzir uma vasta quantidade de informações que até então eram desconhecidas, o que possibilitou a geração do conhecimento necessário para se criar novas operações de marketing.

Isto só foi possível a partir de um longo e trabalhoso processo que se iniciou na extração dos dados e na sua conseqüente transformação em um formato analítico, o que fez com que os dados possuíssem uma melhor qualidade no momento da análise. Em seguida a definição do problema, visto que foi identificada uma grande quantidade de clientes não ativos, foi primordial para que o estudo pudesse ter uma seqüência de trabalho.

A partir daí o estudo segmentou a base de dados a tal ponto que foi possível compreender as características que formam os distintos tipos de clientes. O estudo só conseguiu alcançar esta compreensão a partir da subdivisão do conjunto de dados em regiões não oficiais, pois até então nenhum resultado produzido diferenciava os clientes ativos dos não ativos.

Nesta etapa do processo pôde-se perceber o quanto homogêneo era o conjunto de dados, já que em muitas análises as informações obtidas por um cluster de um determinado tipo de cliente era logo confrontada por outro cluster, de mesmas características, mas que determinava um tipo de cliente justamente contrário ao anterior.

Identificar os clusters sem contradição aos demais foi a tarefa mais difícil e trabalhosa desta etapa. Em muitos casos, o conhecimento específico só foi possível diante da interpretação de uma pequena variação do formato de que cada cluster, ou seja, uma pequena característica de um cluster determinava a diferenciação dos outros, de forma que o cluster como um todo não trazia a especificação do comportamento de um determinado tipo cliente sem que esta informação não fosse contraditória a uma

informação produzida por outro cluster. Este comportamento é ilustrado com muito detalhe nas figuras 5.12 e 7.1, quando são comparados os clusters 9 e 12 na análise realizada com o subconjunto da região Zona Sul III.

De qualquer forma, podemos dizer que o resultado obtido foi bem satisfatório dado às condições do conjunto de dados analisado. Aqui a estratégia de dividir para conquistar foi muito bem empregada, uma vez que o conjunto de dados como um todo não foi capaz de produzir a informação que se esperava.

Após a segmentação o estudo analisou o comportamento de compra dos clientes através da análise de regras de associação, o que nos permitiu avaliar a maneira pela qual os produtos são vendidos e por qual perfil de cliente são realizadas estas compras.

O resultado obtido por esta análise nos permite, por exemplo, complementar as campanhas a serem realizadas a partir do conhecimento obtido através da segmentação dos clientes, já que o seu resultado possibilita o oferecimento de produtos customizados as características pessoais de cada cliente.

Ao contrário do que aconteceu na segmentação, foi observado por esta análise que agrupar os registros, no caso os itens de produtos em categorias e subcategorias, pode gerar um significado mais consistente ao resultado produzido. Além disso, o estudo ilustrou como utilizar os comandos de consultas para prover a predição a partir dos modelos gerados nesta análise.

No final desta análise o estudo apresentou uma forma de utilização das regras de associação para criar um modelo com o objetivo de indicar produtos, a serem exibidos na página inicial da loja virtual, sem que ao menos se saiba o perfil do cliente. Isto foi feito a partir de um modelo que levou em consideração os dias da semana. Como um todo, podemos dizer que esta etapa do estudo produziu um bom resultado.

Por último o estudo focou na criação de dois modelos classificadores que fossem capazes de inferir a predição do tipo de cliente e o valor do pedido de acordo com as suas características pessoais. Esta etapa do processo foi a mais difícil de todas, pois, como já observado nos capítulos 4 e 5, o conjunto de dados analisado era bastante homogêneo, o que fez com que o resultado obtido não fosse tão satisfatório.

Na verdade, este resultado ficou no limite mínimo de aceitação, já que as taxas de acertos obtidas ficaram um pouco acima de 50%. Em muitos casos nem se quer este valor foi alcançado, o que fez com o estudo desconsiderasse estes modelos.

Esta etapa também utilizou a estratégia de divisão do conjunto de dados em vários outros subconjuntos para tentar buscar um melhor retorno. Porém, diferente do que aconteceu na segmentação, nem todos os modelos classificadores gerados com base nestes subconjuntos obtiveram um melhor resultado.

Isto indicou que o modelo que representava o conjunto de dados como um todo trazia consigo tanto as qualidades de cada subconjunto, que permitiam a separação dos dados, quanto às incertezas, que não permitiam uma separação linear entre os dados.

Assim foi difícil observar uma característica ou conjunto de atributos que pudessem separar os dados de modo que fosse possível produzir um bom classificador. Todavia, como a utilização do classificador não é algo que traga risco à empresa, acreditamos que é válida a tentativa de utilização dos classificadores que obtiveram uma taxa de retorno acima de 50%, já que sua suas aplicações podem gerar uma nova forma de interação com os clientes, o que cria para empresa uma nova estratégia de negócio.

De um modo geral podemos dizer que o estudo obteve êxito na sua execução, pois, a partir de seu desenvolvimento, foi possível extrair um conhecimento que até então era oculto para a organização.

Todo este conhecimento possibilitou um entendimento sobre o comportamento dos consumidores atendidos pelo La Mole, o que gerou para a empresa um suporte necessário a qualquer processo de tomada de decisão. Além disso, o estudo indicou possíveis utilizações estratégicas para cada técnica aplicada e seus resultados, oferecendo assim várias oportunidades de estratégias de marketing aos gestores da empresa.

Quanto à ferramenta utilizada, o estudo conclui que seu funcionamento vai de encontro aos principais padrões de mercado. Além disso, ao incorporar o pacote de mineração de dados ao seu principal gerenciador de banco de dados, a ferramenta facilita a integração das bases transacionais existentes aos novos modelos analíticos a serem gerados. Isto faz com que o gerenciamento de dados e a mineração de dados

sejam visto como um mesmo recurso, o que auxilia e facilita a implementação das técnicas de mineração de dados no dia-a-dia das organizações.

Outra vantagem da ferramenta é a vasta quantidade de material disponível, seja através das fontes de informações oferecidas na Internet, quanto que nos livros encontrados na literatura existente. A facilidade do uso de seus recursos, assim como a fácil instalação do produto, complementam as vantagens de utilização desta ferramenta.

Por fim, embora não tenha sido utilizada no desenvolvimento deste trabalho, a ferramenta possui um poderoso recurso de integração dos modelos de mineração ao programa Microsoft Excel. Isto faz com que o uso destes modelos fique disponível a um número maior de pessoas, já que sua utilização não fica dependente a um especialista do MS SQL Server.

Este é sem dúvida alguma um forte diferencial da ferramenta, o que auxiliará na popularização da utilização da mineração de dados, visto que a maioria dos estrategistas de marketing não possui a destreza necessária para operar um gerenciador de banco de dados.

Entretanto o estudo encontrou algumas limitações na ferramenta, como no caso da validação cruzada, que tem o seu resultado armazenado em memória volátil, ou seja, ao abrir e fechar a validação cruzada todo o processo deve ser refeito. Além disso, como descrito no capítulo 7, a matriz de confusão produzida por esta validação deve ser feita de forma manual, através de uma consulta DMX. Isto é muito ruim, já que eleva o tempo do desenvolvimento e de avaliação dos classificadores.

Por fim, embora alguns métodos disponíveis pela ferramenta sejam conhecidos e explicitados pela literatura, pouco se sabe a respeito do funcionamento interno destes algoritmos na ferramenta utilizada, já que não existe a possibilidade de acesso ao código-fonte do produto ou da aplicação destes algoritmos na ferramenta.

Este estudo nos ajudou a reconhecer que, embora a ferramenta nos auxilie na execução das tarefas relacionadas a mineração de dados, nem sempre o resultado obtido produz uma resposta direta ao conhecimento da empresa, o que torna necessário a interpretação destes resultados por um especialista de negócio.

Isto demonstra o quanto é relevante o papel de um especialista neste processo. Enganam-se os que acham que os métodos de aprendizagem de máquina podem substituir o elemento humano de suas atividades, pois a tecnologia é apenas um recurso neste processo.

Como dito no capítulo 2, a busca pelo conhecimento através dos métodos de mineração de dados é um processo iterativo e que deve ser reavaliado constantemente. Por isso, como trabalho futuro, o estudo sugere que seja feita uma avaliação da aplicação destes modelos e das estratégias apresentadas assim que todo o conteúdo desenvolvido pelo estudo esteja no ambiente de produção da empresa.

Além disso, o estudo pôde perceber que as análises ficaram limitadas a pequena quantidade de atributos existentes nos conjunto de dados analisado. Assim, como mais uma sugestão, o estudo indica que se faz necessário um esforço a fim de se obter uma maior quantidade de informações a respeito dos clientes, como por exemplo, o estado civil do cliente, profissão, se possui microondas e etc.

Por último, o estudo deixa duas sugestões a respeito dos clientes não ativos: a primeira, a criação de uma campanha específica junto a estes clientes, de modo que se possível entender as necessidades que não foram atendidas logo após o momento do cadastramento; a segunda, diz respeito a avaliação das áreas de entregas, da qual seja possível avaliar se é vantajoso para a empresa expandir algumas áreas com o objetivo de se alcançar estes clientes não ativos.

Anexo A – Modelo de Dados Analítico

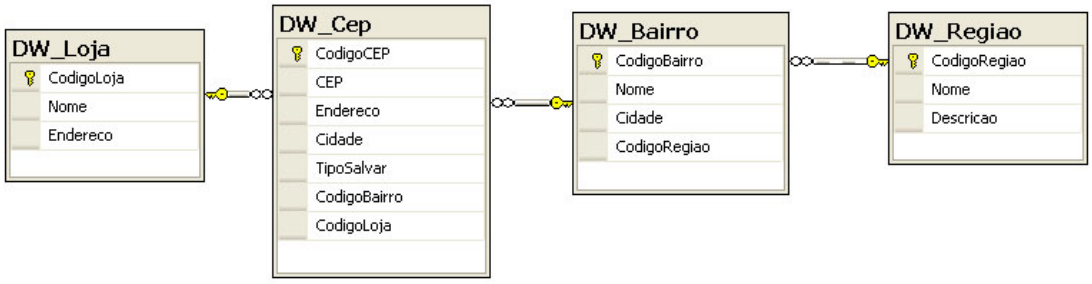


Figura A.1: Modelo de dados entre as tabelas de CEP, Bairro, Região e Loja.

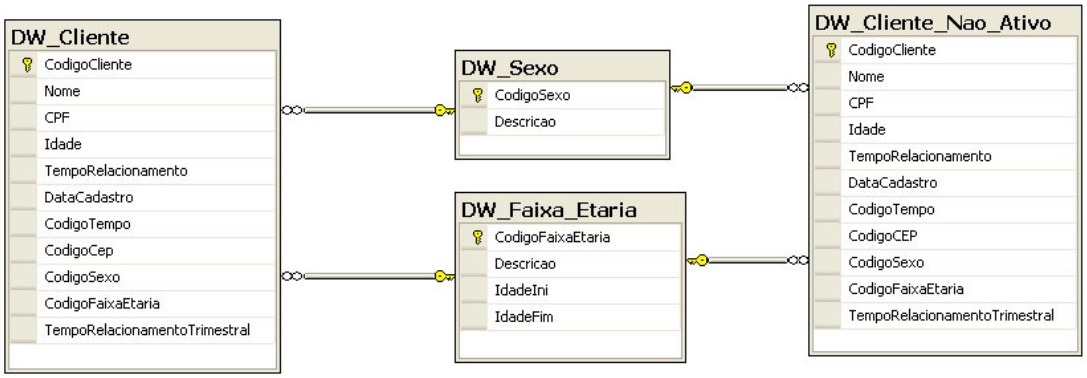


Figura A.2: Modelo de dados entre as tabelas de Cliente (Ativos e Não ativos), Sexo e Faixa Etária.

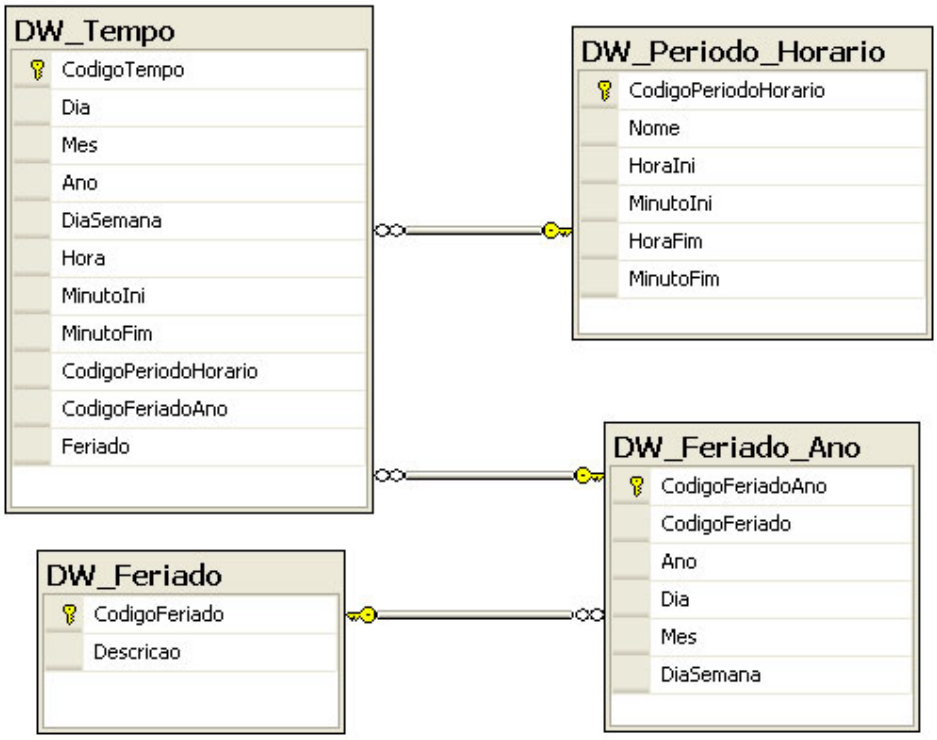


Figura A.3: Modelo de dados entre as tabelas de dimensões Tempo, Feriado e Período.

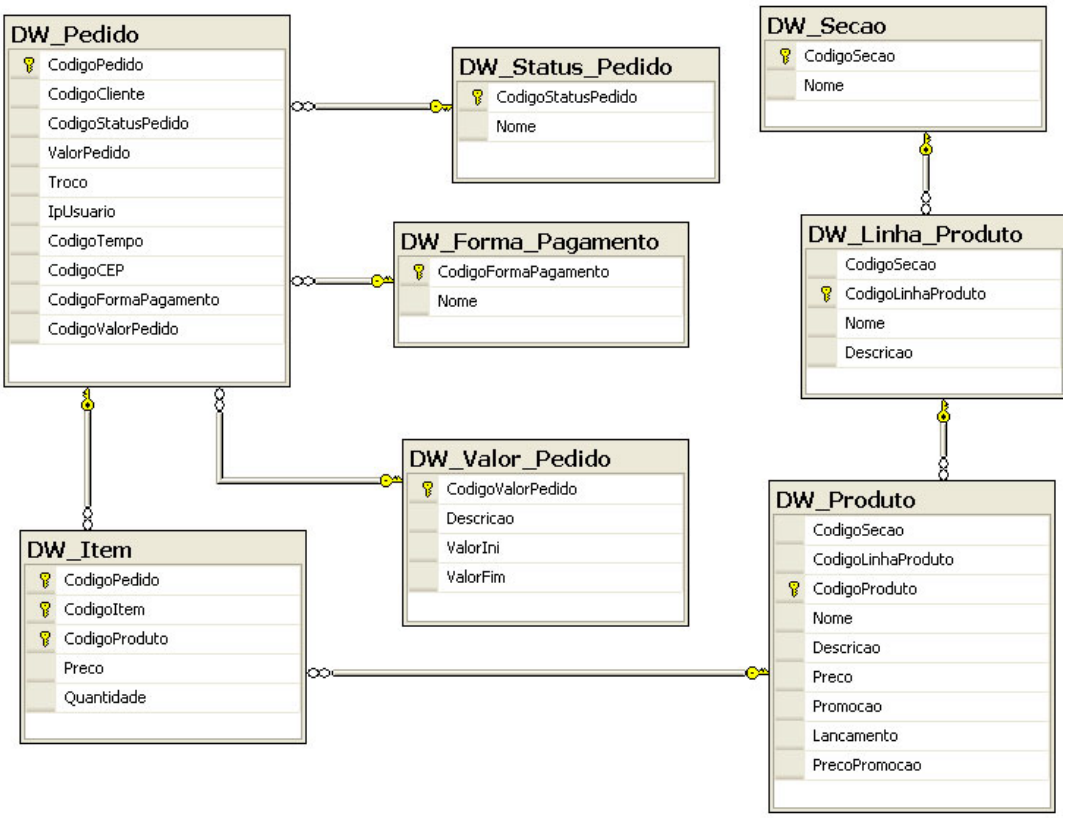


Figura A.4: Modelo de dados relacionado aos Pedidos e Produtos.

Anexo B – Conteúdo das Tabelas de Dados

Código	Descrição da Faixa Etária	Idade Início	Idade Fim
1	15 - 19	15	19
2	20 - 24	20	24
3	25 - 29	25	29
4	30 - 34	30	34
5	35 - 39	35	39
6	40 - 44	40	44
7	45 - 49	45	49
8	50 - 54	50	54
9	55 - 59	55	59
10	Acima de 59 anos	60	150

Tabela B.1: Conteúdo da tabela DW_Faixa_Etaria

Código	Nome da Região	Descrição
1	Zona Norte I	Tijuca
2	Ilha do Governador	
3	Jacarepaguá	
4	Zona Sul I	Orla do Centro até Urca
5	Zona Sul II	Zona Sul sem orla
6	Zona Sul III	Orla do Leme até São Conrado
7	Niterói	
8	Zona Norte II	Méier
9	Zona Norte III	Leopoldina
10	Barra da Tijuca / Recreio	
11	Centro	
12	Outros Estados	
13	Baixada Fluminense	
14	Zona Oeste	

Tabela B.2: Conteúdo da tabela de regiões

Codigo	Nome do Feriado
1	Confraternização Universal
2	Tiradentes
3	Dia do Trabalhador
4	Independência
5	Nossa Senhora da Aparecida
6	Finados
7	Proclamação da República
8	Natal
9	Feriado de São Jorge
10	Feriado de Zumbi dos Palmares
11	Segunda-Feira de Carnaval
12	Terça-Feira de Carnaval
13	Cinzas
14	Paixão de Cristo
15	Páscoa
16	Corpus Chirsti
17	Dia das Mães
18	Dia dos Pais

Tabela B.3: Conteúdo da tabela DW_Feriados

Código	Nome	Hora Inicio	Minuto Inicio	Hora Fim	Minuto Fim
1	Almoço	11	30	14	30
2	Tarde	14	30	18	0
3	Jantar	18	0	22	30
4	Não Disponível	22	30	11	30

Tabela B.4: Conteúdo da tabela DW_Perido_Horario

Referências

[1] Agrawal, R., T. Imielinski, *et al.* Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993. 207-216 p.

[2] Agrawal, R. e R. Srikant. Fast Algorithms for Mining Association Rules. Proc. 20th Int. Conf. Very Large Data Bases, VLDB: Morgan Kaufmann, 1994. 487-499 p.

[3] Bronze & Business Informática LTDA. Disponível em: http://www.bronzebusiness.com.br/webservices/wscep.asmx. Acesso em 25 de março de 2008.

[4] Duda, R., P. Hart, *et al.* Pattern Classification (2nd Edition): Wiley-Interscience. 2000

[5] Glenn, J. M. Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining: Wiley-Interscience. 2006

[6] Klir, G. J. e B. Yuan. Fuzzy sets and fuzzy logic : theory and applications. Upper Saddle River, N.J.: Prentice Hall PTR. 1995. xv, 574 p. p.

[7] Larose, D. Data Mining Methods and Models: Wiley-IEEE Press. 2006

[8] LocaWeb Ltda. Disponível em: http://www.locaweb.com.br. Acesso em: 20 de março de 2008.

[9] Macqueen, J. B. Some methods of classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. 281-297 p.

[10] Microsoft Technet. Resources for IT Professionals. Disponível em: http://technet.microsoft.com/en-us/library/ms132058.aspx. Acesso em 27 de julho de 2008.

[11] Ngai, E., L. Xiu, *et al.* Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, v.36, n.2, p.2592-2602. 2009.

[12] Pelleg, D. e A. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning: Morgan Kaufmann Publishers Inc., 2000. 727-734 p.

[13] Stanimirova, I. e B. Walczak. Classification of data with missing elements and outliers. Talanta, v.76, n.3, p.602-609. 2008.

[14] Su, T. e J. Dy. A deterministic method for initializing K-means clustering. Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, 2004. 784-786 p.

[15] Wikipedia. Disponível em: http://en.wikipedia.org/wiki/Data_Mining_Extensions. Acesso em 27 de julho de 2008.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo