UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DE RIBEIRÃO PRETO
DEPARTAMENTO DE CONTABILIDADE
PROGRAMA DE PÓS-GRADUAÇÃO EM CONTROLADORIA E CONTABILIDADE

VANESSA ANELLI BORGES

CONTRIBUIÇÃO DA SEGMENTAÇÃO DE DADOS PARA A DECISÃO DE CONCESSÃO DE CRÉDITO AO CONSUMIDOR: UMA COMPARAÇÃO DE RESULTADOS

Orientador: Prof. Dr. Fabiano Guasti Lima

RIBEIRÃO PRETO
2011

Reitor da Universidade de São Paulo
Prof. Dr. João Grandino Rodas

Diretor da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto
Prof. Dr. Sigismundo Bialoskorski Neto

Chefe do Departamento de Contabilidade de Ribeirão Preto
Profa. Dra. Adriana Maria Procópio de Oliveira

VANESSA ANELLI BORGES

CONTRIBUIÇÃO DA SEGMENTAÇÃO DE DADOS PARA A DECISÃO DE CONCESSÃO DE CRÉDITO AO CONSUMIDOR: UMA COMPARAÇÃO DE RESULTADOS

Versão corrigida da dissertação apresentada ao Programa de Pós-Graduação em Controladoria e Contabilidade da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo como requisito para obtenção do título de Mestre em Ciências. A original encontra-se disponível no Serviço de Pós-Graduação da FEA-RP/USP.

Orientador: Prof. Dr. Fabiano Guasti Lima

RIBEIRÃO PRETO

2011

Borges, Vanessa Anelli
Contribuição da segmentação de dados para a decisão de concessão de crédito ao consumidor: uma comparação de resultados. Ribeirão Preto, 2011.
126 p. : il.; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto/USP. Área de concentração: Controladoria e Contabilidade.
Orientador: Lima, Fabiano Guasti.

1. Redes Neurais 2. Análise Multivariada de Dados 3. Segmentação de Dados 4. Crédito 5. Inadimplência

FOLHA DE APROVAÇÃO

VANESSA ANELLI BORGES

CONTRIBUIÇÃO DA SEGMENTAÇÃO DE DADOS PARA A DECISÃO DE CONCESSÃO DE CRÉDITO AO CONSUMIDOR: UMA COMPARAÇÃO DE RESULTADOS

Dissertação apresentada ao Departamento de Contabilidade da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, como parte dos requisitos para obtenção do título de Mestre em Ciências.

Área de Concentração: Controladoria e Contabilidade

Aprovado em:

Banca Examinadora

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

DEDICATÓRIA

À MINHA FAMÍLIA

AGRADECIMENTOS

Aos meus pais amados, Célia e Fernando, pelo amor incondicional, pelo apoio, pela dedicação e carinho de sempre.

Ao meu querido irmão, pelo grande exemplo de vida.

Ao professor Dr. Fabiano Guasti Lima, meu orientador, pela confiança de sempre e por fazer valer todos os sentidos da palavra orientação.

À minha amiga Maíra Assaf Andere, pelo incentivo e apoio nas horas em que precisei.

Ao Sr. Antônio Carlos Maçonetto por entender a importância deste Mestrado para mim.

Aos amigos e familiares que sempre me incentivaram e apoiaram: Gustavo, Fabrício, Rodrigo, Leandro, Janaína, Maria Gabriela, Andiara, Patrícia, Ana Paula, Fernanda, Maíra Anelli, Marisa, Margareth, Sílvia, Matilde e Maurícia.

Ao João Jeronimo Berbardi Junior e ao professor Dr. Herbert Kimura pela ajuda na execução desta pesquisa.

À professora Dra. Adriana Maria Procópio de Araújo e ao professor Dr. Evandro Marcos Saidel Ribeiro pelas orientações e contribuições.

Aos meus professores de mestrado em Controladoria e Contabilidade da FEA-RP pelos conhecimentos transmitidos.

Às meninas da Secretaria de Pós-Graduação, Vânia e Érika, pelo carinho e atenção com que atendem os alunos.

RESUMO

BORGES, V. A. **Contribuição da segmentação de dados para a decisão de concessão de crédito ao consumidor:** uma comparação de resultados**.** 2011. 126 f. Dissertação (Mestrado) – Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2011.

Este trabalho explora a contribuição da segmentação de dados, manual e estatística, combinada com análise discriminante e com redes neurais, para a tomada de decisão de concessão de crédito ao consumidor. A grande importância que a decisão de concessão de crédito tem para o mercado varejista e para a área de controladoria de uma empresa dão cenário para o aumento da relevância do gerenciamento do risco de crédito. O mercado necessita, cada vez mais, de modelos capazes de produzir boas expectativas do comportamento dos clientes, com vistas de reduzir perdas com inadimplência. Dado um banco de dados composto por 50 mil clientes de uma importante loja do setor varejista, primeiro aplica-se a análise discriminante, depois as redes neurais, para que se classifique a capacidade preditiva de cada técnica nesta etapa. Posteriormente, os dados são segmentados com base na região à qual a filial de venda pertence e, depois, por meio das análises de *clusters K-Means* e *TwoStep Cluster*. A próxima etapa compreende a aplicação da análise discriminante, depois das redes neurais, para cada um dos grupos formados, tanto pela segregação por região, quanto pela segregação por meio das técnicas de análise de *clusters*. A última etapa abrange a comparação da soma dos acertos dos bons e dos maus pagadores obtida tanto para análise discriminante, quanto para redes neurais, combinadas com a segmentação de dados, com os resultados obtidos na primeira etapa – sem a segmentação dos dados. O modelo híbrido que combina a segmentação manual dos dados com análise discriminante e com redes neurais, formando-se 21 micro-regiões foi o que apresentou maiores porcentagens de acerto de classificação. O modelo híbrido que combina análise discriminante e redes neurais com a análise de *clusters TwoStep Cluster* não apresenta resultados de classificação adequados à proposta deste trabalho, devendo, portanto, ser descartado.

Palavras chaves: redes neurais, análise discriminante, segmentação dos dados, crédito, inadimplência.

ABSTRACT

BORGES, V. A. **Contribution of targeting data to the decision to grant credit to consumers:** a comparison of results. 2011. 126 f. Dissertação (Mestrado) – Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2011.

This paper explores the contribution of data segmentation, and statistical manual, combined with discriminant analysis and neural networks, for making the decision to grant credit to consumers. The great importance that the decision to grant credit is for the retail market and the area of controlling a business scenario to give increasing importance of managing credit risk. The market needs, increasingly, models capable of producing good expectations of customer behavior, in order to reduce losses from default. Given a database consisting of 50 000 customers of a major retail store, the first applies to discriminant analysis, then the neural networks, in order to classify the predictive ability of each technique in this step. Subsequently, the data are segmented based on the region to which the branch belongs to sell and then through the analysis of *cluster*s *K-Means* and *TwoStep Cluster*. The next step involves the application of discriminant analysis, neural networks then, for each of the groups formed by both the segregation by region, by segregation and by the techniques of *cluster* analysis. The last step includes comparing the sum of the hits of the good and bad debtors obtained for both discriminant analysis and neural networks, combined with the segmentation of data, with the results obtained in the first stage - without the segmentation of the data. The hybrid model that combines the manual segmentation of the data with discriminant analysis and neural networks, forming 21 micro-regions showed the highest percentage of correct classification. The hybrid model that combines neural networks and discriminant analysis with *cluster* analysis results *TwoStep Cluster* does not have appropriate rating to the proposal of this work and should therefore be discarded.

Keywords: neural networks, discriminant analysis, segmentation of data, delinquency.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTAS DE ABREVIATURAS E SIGLAS

AIC – *Akaike's Information Criterion* (Critério de Informação de Akaike)

ANOVA – *Analysis of Variance* (Análise da Variância)

BACEN – Banco Central do Brasil

BIC – *Bayesian Information Criterion* (Critério de Informação Bayesiano)

CEP – Código de Endereçamento Postal

COSIF - Plano Contábil das Instituições do Sistema Financeiro Nacional

CURVA ROC - *Receiver Operating Characteristic Curve*

INDECO – Indicadores Econômicos Consolidados

MANN - *Multi-dimensional Analysis of Nearest Neighbors*

MAPE - *Mean Absolute Percentage Error*

MLP – *Multilayer Percepton*

PIB – Produto Interno Bruto

RNA – Rede Neural Artificial

SC – *Schwartz Criterion* (Critério de Schwartz)

SEBRAE – Serviço Brasileiro de Apoio às Micro e Pequenas Empresas

SERASA - Centralização dos Serviços Bancários S/A

SPC – Serviço de Proteção ao Crédito

TSC – *TwoStep Cluster*

LISTAS DE VARIÁVEIS

b_k – Viés

d_{ij} – Distância entre duas observações (i e j)

d_{ij}^2 – Distância quadrática entre duas observações (i e j)

$d_{(ij)k}$ – Distância entre dois grupos (i e j) e (k)

d_{ik} - Distância da variável k referente à observação i

d_{jk} - Distância da variável k referente à observação j

$e_j(n)$ - Sinal de erro na saída do neurônio j

S - Estimativa amostral da matriz de variância-covariância Σ dentro dos agrupamentos

u_k - Saída do combinador linear devido aos sinais de entrada

x_i - Observação i

X_{ik} - Variável independente i para o objeto k

x_{ik} - Valor da variável k referente à observação i

x_{jk} – Valor da variável k para a observação j

x_j - Observação j

x_j - Sinais de entrada da rede

W_i - Peso discriminante para a variável independente i

$w_{ji}(n)$- Peso sináptico

w_{kj} - Pesos sinápticos do neurônio k

y_k – Sinal de saída do neurônio

Z_{jk} - Escore Z discriminante da função discriminante j para o objeto k

α - Intercepto da função discriminante

$\varepsilon\ (n)$ – Soma do valor instantâneo da energia do erro de todos os neurônios da camada de saída

η - Parâmetro da taxa de aprendizagem do algoritmo de retropropagação

$v_j\ (n)$ - Campo local induzido

$y_j(n)$ - Sinal funcional que aparece na saída do neurônio j na interação n

$\varphi(\cdot)$ - Função de ativação

$\delta_j(n)$ - Gradiente local

$\partial e_j(n)$ - Derivada do sinal de erro na saída do neurônio j

$\partial \varepsilon(n)/\partial w_{ji}(n)$ - Derivada parcial

$\partial \varepsilon(n)$- Derivada da soma do valor instantâneo da energia do erro de todos os neurônios da camada de saída

$\partial v_j(n)$ - Derivada do campo local induzido

$\partial w_{ji}(n)$ - Derivada do peso sináptico

$\partial y_j(n)$ - Derivada do sinal funcional que aparece na saída do neurônio j na interação n

$\frac{1}{2}\ e_j^2\ (n)$ - Valor instantâneo da energia do erro

SUMÁRIO

1. INTRODUÇÃO

O efeito catalisador do Comércio Varejista na economia, acelerando o crescimento nacional, é comprovado pelo desempenho deste setor que, desde 2003, cresce mais que o Produto Interno Bruto (PIB) do Brasil. O Gráfico 1 apresenta um comparativo entre o crescimento do PIB e o crescimento do volume de vendas no varejo[1] nos últimos 6 anos:

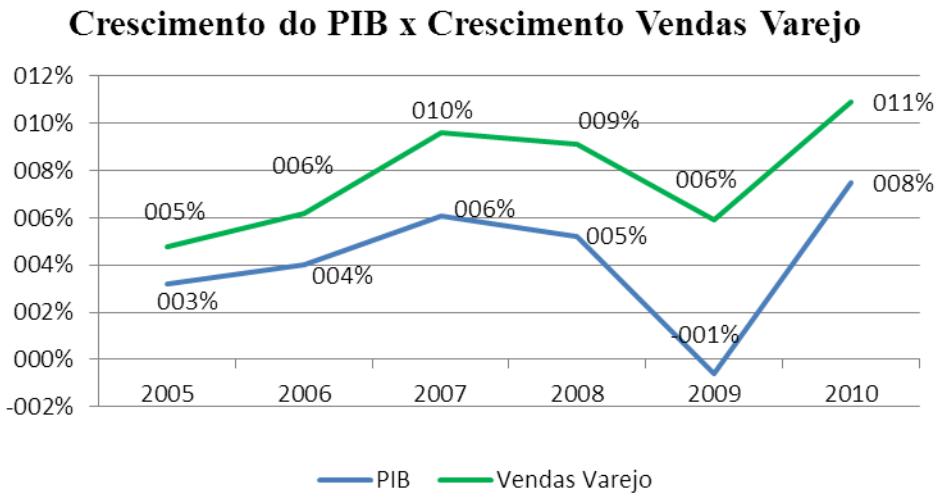

Gráfico 1 - Crescimento do PIB x Crescimento Volume de Vendas Varejo

Fonte: IBGE/2011

Em todos os anos o crescimento do volume de vendas no varejo foi superior ao crescimento do PIB. Dado seu desempenho, o setor varejista além de ser um importante empregador, ainda pode contribuir com a minimização dos impactos causados pelos problemas externos na economia brasileira, como já pode ser observado durante a instabilidade financeira mundial entre 2008 e 2009. Enquanto a produção industrial teve uma queda abrupta, acompanhada pela redução das exportações, o volume de vendas no varejo permaneceu estável neste período. (SEBRAE, 2009, p. 6).

O volume das operações de crédito do sistema financeiro nacional privado atingiu aproximadamente 1,6 bilhões de reais em 2010, conforme tabela a seguir:

[1] Vendas varejo – conceito restrito: não inclui segmentos de veículos e materiais de construção.

Tabela 1- Operações de Crédito do Sistema Financeiro Nacional Privado

Operações de Crédito do Sistema Financeiro Nacional Privado (em milhões)							
Ano	Indústria	Habitação	Rural	Comércio	Pessoas Físicas	Outros Serviços	**Total Setor Privado**
2010	361.055	138.778	123.928	172.642	549.188	292.389	**1.637.979**

Fonte: http://www.bcb.gov.br/?INDECO. Acesso em 24 de Maio de 2011.

As operações de créditos realizadas por pessoas físicas foram as que apresentaram maior volume (cerca de 34% do total). Já o comércio apresentou, aproximadamente, 11% do volume total.

Em 2010 o percentual de operações de crédito com recursos livres, em relação ao PIB, foi de, aproximadamente, 30%, ao passo que o percentual de operações de crédito lastreadas com recursos direcionados atingiu cerca de 14,4% do PIB, conforme Tabela 2:

Tabela 2- Operações de Crédito do Sistema Financeiro - Percentual do PIB

Operações de Crédito do Sistema Financeiro - Percentual do PIB			
Ano	Recursos Livres	Recursos Direcionados	**Total**
2010	30%	14,4%	**44,4%**

Fonte: http://www.bcb.gov.br/?INDECO. Acesso em 24 de Maio de 2011.

O volume de 44,4% das operações de crédito em relação ao PIB denota a importância da eficiência na tomada de decisões quanto a análise e a concessão de crédito.

Os empréstimos com recursos livres são impulsionados pelo consumo das famílias, o que se reflete, principalmente, no crédito pessoal e no financiamento de bens cíclicos e não cíclicos[2], por isso, apresentam maior volume do que as operações com recursos direcionados. (BACEN, 2010, p. 1).

[2] Consumo cíclico – Tecido, vestuário e calçados, utilidades domésticas, automóveis e motocicletas, mídia, hotelaria e restaurantes, lazer , comércio. Consumo não-cíclico - Agropecuária, alimentos processados, bebidas, fumo, produtos de uso pessoal e limpeza, saúde, comércio e distribuição.

O capital de giro representa o total de recursos demandados pela empresa para financiar seu ciclo operacional e a administração desse capital envolve diversas decisões como política de estocagem, compra de matéria-prima, produção, venda de produtos e serviços e prazo de recebimento. Uma empresa que apresenta folga financeira é aquela que possui fontes de financiamento de longo prazo para cobrir os investimentos de longo prazo e, ainda, parte dos investimentos de curto prazo, bem como bons indicadores de liquidez indicam capacidade de cobrir com as obrigações assumidas.

Além disso, para muitas empresas, os investimentos em valores a receber representam uma parte significativa de seus ativos circulantes, exercendo, em consequência, importantes influências em suas rentabilidades (ASSAF NETO, 2009, p. 554).

Desta forma, modelos desenvolvidos para dar suporte à decisão de concessão de crédito trazem para a área de controladoria de uma empresa não só mais uma ferramenta a ser utilizada na administração do capital de giro, como também uma garantia de redução de riscos com inadimplência e possível maximização da rentabilidade.

> Ao comparar os custos e benefícios globais para a empresa, a decisão de oferecer crédito deve ser seguida por uma análise adicional definindo a quais clientes será oferecido e em que termos. Ao tomar essas decisões, as empresas geralmente se baseiam na análise de crédito, com o fim de avaliar o mérito dos clientes isolados. A realização de uma análise de crédito dependerá do volume desse crédito. Como há um custo administrativo associado à realização dessa análise, pode não valer a pena colocá-la em prática quando o crédito oferecido é pequeno ou quando o risco de não pagamento é muito baixo (DAMODARAN, 2004, p. 343).

Em uma análise de crédito comum, os clientes prestam informações sobre suas características, capacidade de pagamento, capital, garantia oferecida pelo crédito, e situação empregatícia. Esses são os chamados cinco Cs[3] do crédito, que representam características que se correlacionam com o merecimento do crédito (ou, alternativamente, com taxas de não

[3] Os cinco Cs do crédito, em inglês, são: *character, capacity to pay, capital, collateral offered for the credit e condition of employment.*

pagamento) no passado. Várias outras variáveis foram historicamente fatores eficazes para avaliação do mérito de crédito. Por exemplo, um cliente que é proprietário de uma casa tem menos probabilidade de ser inadimplente do que um que aluga uma casa. Similarmente, um cliente contratado pelo mesmo empregador por um longo período tem menos possibilidade de ser inadimplente do que alguém que ficou empregado apenas um curto período de tempo. A quantidade de informação exigida aumenta de acordo com o volume de crédito oferecido para o cliente (DAMODARAN, 2004, p. 343).

Basicamente, existem duas modalidades de crédito: o crédito interempresarial (ou crédito mercantil), envolvendo empresas e clientes, e o crédito pessoal (ou crédito direto ao consumidor), administrado por instituições financeiras (ASSAF NETO, 2009, p. 554). O foco deste trabalho está na modalidade de crédito mercantil.

Henley e Hand (1996, p. 77) destacaram que há cerca de 30 anos já se adotavam modelos de *scores* para avaliar viabilidade creditícia. Desde então, diversas técnicas quantitativas têm sido empregadas para melhorar a qualidade dos modelos, como regressão logística, análise discriminante, regra dos *k* vizinhos mais próximos, árvores de decisão e, mais recentemente, foram retomadas as redes neurais.

Seguindo uma linha de análise que utiliza as práticas dos pesquisadores citados e outras inferências como Hair et al. (2005), Fávero et al. (2009), Pestana e Gageiro (2003), Pohlmann (2007), Haykin (2001), entre outros, o presente estudo busca comparar a contribuição obtida com a combinação da segmentação de dados, manual e estatística, com análise discriminante, ou com redes neurais, para a decisão de concessão de crédito. Segmentando a população, trabalha-se com subpopulações mais homogêneas, devido à redução da variância entre os agrupamentos, fato que possibilita uma análise mais eficiente, a partir do modelo adotado.

A metodologia a ser empregada consiste na segmentação de 50 mil clientes de uma importante loja do setor varejista brasileiro, de duas maneiras distintas, sendo a primeira delas manual, realizada de acordo com a região[4] à qual a filial da venda pertence, ou seja, o grupo 1 será formado por todos os casos nos quais as vendas tiverem sido realizadas em filiais que

[4] De acordo com a empresa que concedeu o banco de dados, as regiões são formadas segundo a posição geográfica das filiais.

fazem parte da região 1; o grupo 2 será formado por todos os casos nos quais as vendas tiverem sido realizadas em filiais que fazem parte da região 2 e, assim, sucessivamente; e a segunda; estatística, realizada por meio das análises de *clusters K-Means* e *TwoStep Cluster.*

Resumidamente, a pesquisa consiste em duas etapas: primeiro aplica-se a análise discriminante e as redes neurais, separadamente, para que se identifique as características dos bons e dos maus pagadores que caracterizam a amostra global, bem como se classifica a capacidade preditiva de cada técnica para esta etapa. Na segunda etapa, aplica-se a segmentação dos dados, de acordo com a região à qual cada filial de venda pertence (manual) e por meio da análise de *clusters* (estatística), nos casos, não nas variáveis, com o objetivo de classificar os clientes a partir de suas semelhanças de comportamento. Em seguida, aplica-se a análise discriminante e as redes neurais, separadamente, para cada um dos subgrupos formados e soma-se os resultados dos acertos dos clientes adimplentes e dos clientes inadimplentes, para a segmentação manual e, depois, para a segmentação estatística. Este resultado é comparado com o obtido na primeira etapa com o objetivo de que o papel da segmentação regional e estatística seja testado, e se confirme se o modelo proposto agrega, ou não, melhora na capacidade preditiva de concessão de crédito direto ao consumidor.

A Figura 1 apresenta um resumo da metodologia utilizada neste trabalho:

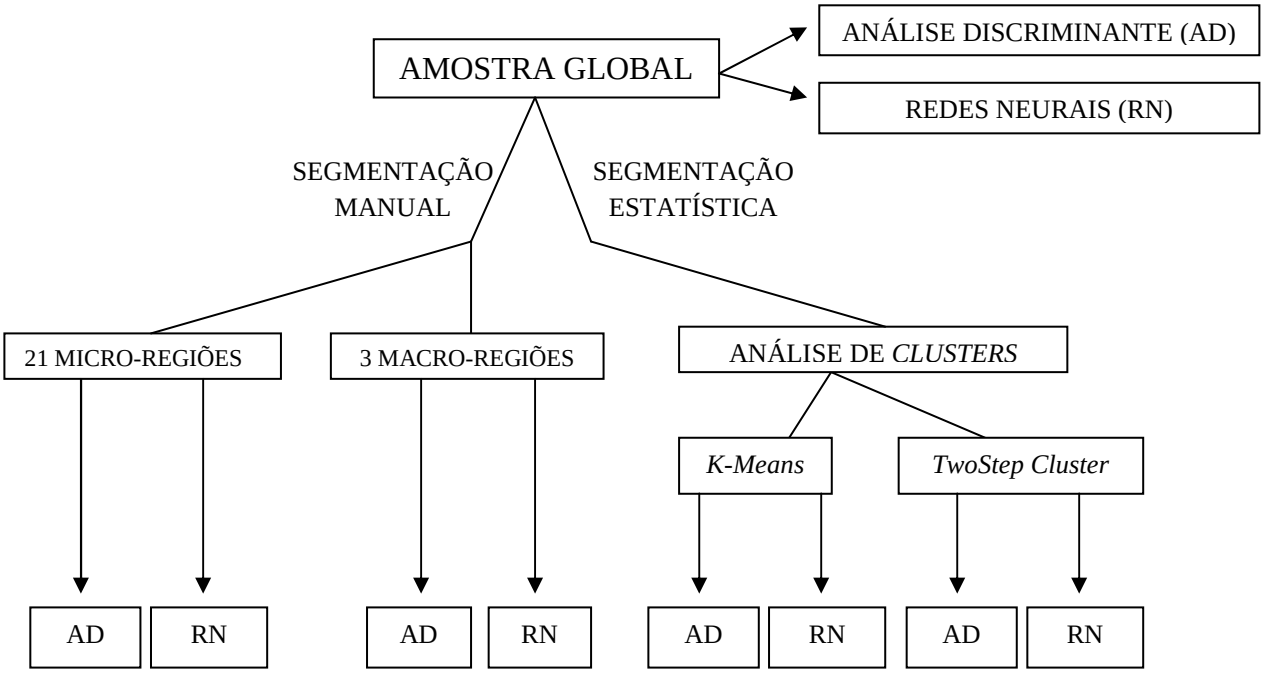

Figura 1 – Resumo da Metodologia

Acredita-se que a soma dos resultados da classificação para cada grupo formado seja superior ao resultado da classificação para a amostra global, tanto para a análise discriminante, quanto para as redes neurais. Além disso, os acertos de classificação poderão ser maiores para as redes neurais, combinadas ou não com a segmentação de dados, manual ou estatística, do que para a análise discriminante, combinada ou não, com esta mesma técnica, com base no comportamento dos dados que compõem a base estudada. Tal fato ocorre porque modelos lineares de análise apresentam melhores resultados quando aplicados a dados de comportamento linear, bem como modelos não lineares o fazem quando aplicados a uma base de dados de comportamento não linear. Outro fator que interfere nos resultados é o tamanho da base estudada.

Para todos os modelos propostos foi rodada a análise discriminante considerando a probabilidade de um cliente ser classificado como adimplente igual à probabilidade de tal cliente ser classificado como inadimplente, ou seja, 50% para cada grupo, bem como a análise discriminante considerando a opção classificação pelo tamanho do grupo (ferramenta *a priori* do SPSS), por meio da qual se verificou que a chance de um cliente ser classificado como adimplente é de 80,38%, ao passo que a chance de que tal cliente seja classificado como inadimplente é de 19,62%, em todos os cenários estudados.

Hand e Henley (1997, p. 523) lembram que os métodos tradicionais de análise de crédito utilizavam pessoas experientes para julgar se o crédito seria ou não concedido para um determinado demandante. Era um trabalho artesanal e meticuloso. Entretanto, as pressões econômicas resultantes da crescente demanda por crédito, aliadas à competição comercial e novas tecnologias computacionais, permitiram o desenvolvimento de sofisticadas técnicas e modelagens estatísticas que vieram abreviar e massificar a decisão de crédito, no âmbito das grandes corporações financeiras e não-financeiras.

É neste contexto que o problema de pesquisa é inserido, ou seja, sob a ótica da combinação de técnicas de modelagem de dados, para diagnóstico e previsão. O diagnóstico é feito para se identificar as variáveis que apresentam maior contribuição ao modelo de suporte à decisão da concessão de crédito. A previsão é dada pelo uso da análise discriminante e das redes neurais, tendo como variável de previsão a classificação do cliente como adimplente ou

inadimplente. Tal variável é representada como variável binária 0 ou 1, sendo classificado como 0 o cliente adimplente e como 1 o cliente inadimplente. A decisão de concessão ou não do crédito se dá com base na variável prevista, ou seja, o crédito deverá ser concedido para clientes classificados como adimplentes e negado para clientes inadimplentes. As variáveis preditoras são indicadas como características, com atributos que podem ser numéricos ou não. Todas as variáveis serão tratadas com maiores detalhes nos próximos capítulos.

Os bancos de dados das empresas estão apresentando, cada vez mais, informações advindas de variáveis criadas e desenvolvidas pelos analistas, com o intuito de refinar a análise para identificar clientes solventes e insolventes. Assim, considerando o aperfeiçoamento do conjunto de dados, pode-se refinar tais variáveis graças à relação mais íntima que tais clientes possam apresentar para um restrito grupo de informações.

Assim, a questão de pesquisa é: qual é a capacidade preditiva da variável binária adimplente/inadimplente utilizada na decisão de concessão de crédito, que se obtém por meio do uso combinado da segmentação de dados com análise discriminante e com redes neurais?

Pela hipótese de pesquisa assumida, a combinação da segmentação de dados; seja com base no agrupamento dos casos de acordo com uma das variáveis do modelo (manual) ou pela utilização de técnicas de análise multivariada, como a análise de *clusters*; com análise discriminante e com redes neurais, contribui para o refinamento da capacidade preditiva da variável binária adimplente/inadimplente, utilizada na decisão de concessão de crédito, por conseguir agrupar indivíduos que tenham características creditícias semelhantes.

> ... a análise de *clusters* é uma das técnicas de análise multivariada cujo propósito primário é reunir objetos, baseando-se nas características dos mesmos. Ela classifica objetos segundo aquilo que cada elemento tem de similar em relação a outros pertencentes a determinado grupo, considerando um critério de seleção predeterminado (POHLMANN, 2007, p. 235).

Trabalhos como o de Hua, Song e Li (2009), bem como o de Lima e Pereira (2010), exploram a aplicação de modelos híbridos que combinam a segmentação de dados com técnicas de análise multivariada ou redes neurais. Os resultados das pesquisas desses autores foram satisfatórios, ou seja, o refinamento dos dados permitiu uma melhora na capacidade

preditiva da variável binária adimplente/inadimplente pelos modelos propostos, fato que sustenta a hipótese defendida neste trabalho.

Busca-se atingir o mesmo objetivo da análise de *clusters* com a segmentação manual dos dados. Ademais, também se deseja testar os vários tipos de análise de *cluster* existentes na literatura e identificar qual a contribuição de cada um deles.

Pohlmann (2007, p. 329) reúne três objetivos da análise de *clusters*: (1) a descrição taxonômica, que apresenta propósitos exploratórios e de formação de uma classificação (taxonomia) de objetos com base empírica; (2) a simplificação de dados, que busca ver as observações como membros de grupos e perfiladas, segundo suas características gerais; e (3) a identificação das relações, que procura uma imagem dos relacionamentos existentes entre as observações, informação que provavelmente não seria possível obter com a análise de observações individuais.

Pode-se dizer que os objetivos 1, 2 e 3 estão presentes neste trabalho como objetivos secundários. Procura-se explorar a base de dados classificando os clientes em grupos, de acordo com suas características (relações), o que não deixa de ser uma maneira de simplificar os dados para aprimorar as análises.

Com relação à proposta de filtrar informações do banco de dados, encontrou-se suporte na literatura que já vinha apontando para a necessidade de melhoria nas bases de dados com os trabalhos de (HUA; SONG; LI, 2009, p. 57) e (LI; XU, 2009, p. 428).

A estrutura desta dissertação está dividida em seis capítulos; após este Capítulo Introdutório, o Capítulo 2 apresenta uma revisão bibliográfica dos trabalhos na área de concessão de crédito a consumidores e diversas aplicações possíveis para redes neurais e análise multivariada. No Capítulo 3 é descrita a Metodologia completa desenvolvida nesta pesquisa, permitindo uma visão geral das técnicas adotadas. No Capítulo 4 é realizada a caracterização da amostra e o detalhamento do passo a passo da pesquisa. O Capítulo 5 aborda os resultados obtidos e a comparação entre as técnicas. Por fim, o Capítulo 6 traz as conclusões advindas deste estudo, bem como as recomendações para futuros trabalhos.

2. REVISÃO TEÓRICA

Neste capítulo será apresentada uma revisão bibliográfica dos trabalhos na área de concessão de crédito a consumidores, com ênfase na aplicação de técnicas de análise multivariada (como análise discriminante e análise de *clusters*) e de redes neurais na previsão do *status* (adimplente/ inadimplente) destes consumidores.

2.1 EVOLUÇÃO HISTÓRICA E ANÁLISE DA CONCESSÃO DE CRÉDITO

A sociedade feudal consistia de três classes: sacerdotes, guerreiros e trabalhadores, sendo que o homem que trabalhava produzia para ambas as outras classes, eclesiástica e militar. O clero e a nobreza constituíam as classes governantes. Controlavam a terra e o poder que delas provinha. A Igreja prestava ajuda espiritual, enquanto a nobreza, proteção militar. Em troca exigiam pagamento das classes trabalhadoras, sob a forma de cultivo das terras. Nos primórdios da sociedade feudal, a vida econômica decorria sem muita utilização de capital. Era uma economia de consumo, em que cada aldeia feudal era praticamente auto-suficiente. Já no capitalismo, sistema que substitui o feudal, as mercadorias que o trabalho do operário transformam de matérias-primas em produtos acabados geram lucros, ou seja, o trabalhador recebe um salário menor do que o valor da coisa produzida. O capitalista é dono dos meios de produção: edifícios, máquinas, matéria-prima e compra a força de trabalho. É toda associação dessas duas coisas que decorre a produção capitalista. (HUBERMAN, 1981, p. 5).

No ano de 1.800, a importância e utilidade da invenção da máquina a vapor havia se tornado tão evidente aos ingleses, que ela estava em uso em 30 minas de carvão, 22 minas de cobre, 28 fundições, 17 cervejarias e 8 usinas de algodão. A invenção de máquinas para fazer o trabalho do homem era uma história antiga, mas com a associação da máquina à força do vapor ocorreu uma modificação importante no método de produção. O aparecimento da máquina a vapor foi o nascimento do sistema fabril em grande escala. O sistema fabril, com sua organização eficiente em grande escala e sua divisão de trabalho, representou um aumento vertiginoso na produção. As mercadorias saíam das fábricas num ritmo intenso. Esse

aumento da produção foi em parte provocado pelo capital, abrindo caminho na direção dos lucros. Foi, em parte, uma resposta ao aumento da procura. (HUBERMAN, 1981, p. 155).

Antes da idade capitalista, o capital era acumulado principalmente através do comércio. Uma vez iniciada uma indústria moderna, ela obtém seus lucros e acumula seu capital muito depressa. O acúmulo de capital, que veio do comércio primitivo, somado à existência de uma classe de trabalhadores sem propriedades, prenunciaram o início do capitalismo industrial. O sistema fabril, em si, proporcionou o acúmulo de uma riqueza ainda maior. Os donos dessa nova riqueza, educados na crença de que o Reino dos Céus era deles, se economizassem e reinvestissem suas economias, empregavam novamente seu capital em fábricas. Assim, o sistema moderno, tal como é conhecido, começou a existir. (HUBERMAN, 1981, p. 143).

No ano de 1.920, Henry Ford e A.P. Sloan reconheceram que não era suficiente somente produzir carros, mas que meios de financiar a compra desses carros necessitavam ser desenvolvidos. Tal constatação levou ao desenvolvimento de financeiras como a GE Capital e a GM *Finance*. Alguns anos mais tarde, o advento do cartão de crédito em 1.960 passou a proporcionar aos consumidores a possibilidade de financiar desde a compra de um simples clipes, até a compra de computadores e o pagamento de viagens de finais de semana. Na sequência, vem o crescimento em outras áreas de concessão de crédito como empréstimos pessoais, financiamento de veículos e cartões de lojas. Cada um destes produtos tem suas próprias características, de modo que os mercados financeiros incluem um *mix* de crédito e risco de taxa de juros em um complexo ambiente econômico e financeiro (THOMAS, 2003, p. 1).

> O crédito usado adequadamente, tanto por governos quanto por empresas, como instrumento de gerenciamento do consumo, continua a mostrar vigor notável, graças ao papel sumamente importante desempenhado no cotidiano da humanidade de facilitar as transações de bens e serviços. A despeito das elevadas taxas de juros básicas no Brasil, que se refletem em taxas ainda mais altas para operações de empréstimos ao consumidor, um volume considerável de operações a prazo é realizado por empresas varejistas. Dada a baixa renda média do brasileiro, as empresas, para viabilizarem suas vendas, têm recorrido ao financiamento de seus clientes (LIMA et al., 2009, p. 34).

> Dados do Banco Central do Brasil mostram a evolução do crédito a partir de 1.988, evidenciando que o volume geral de crédito não acompanhou a evolução do Produto Interno Bruto (PIB) ao longo do tempo, colocando mais uma vez em destaque a falta de poupança interna e a dependência de capitais externos para o desenvolvimento econômico do país. No entanto, também se evidencia, a partir de 1.994, um significativo e contínuo crescimento do crédito ao consumidor, mostrando que o Plano Real, além de resgatar os instrumentos que têm permitido o controle da inflação, também alcançou sucesso na distribuição de renda, fato evidenciado pelo expressivo aumento no crédito ao consumidor quando considerado como percentual do PIB. (LIMA et al., 2009, p. 36).

Paralelo ao crescimento da relevância do crédito para a viabilização de compras e o aquecimento da economia, os órgãos reguladores começaram a advertir sobre o aumento da exposição ao risco de crédito. Em particular, o advento do Basileia II chamou a atenção para a necessidade de ser feita uma boa modelagem de risco dos portfólios de crédito ao consumidor, em vez de simplesmente se avaliar o risco de crédito de forma individual e independente. A crise recente, que se iniciou com um deterioramento dos empréstimos imobiliários concedidos no mercado norte-americano a clientes de baixa renda, evidenciou a necessidade de melhores modelos de análise e mecanismos de controle. Se por um lado a concessão de crédito permite o giro dos ativos das empresas, por outro induz à assunção de riscos financeiros que, eventualmente, podem se transformar em perdas advindas da inadimplência. (LIMA et al., 2009, p. 36).

Lin et al. (1995, p. 1) consideram que detectar tendências e padrões em uma série de dados financeiros representa um grande interesse do mundo dos negócios, uma vez que informações dessa natureza dão suporte ao processo de tomada de decisões.

Bruni, Fama e Murray (1998, p. 2) constataram a necessidade do desenvolvimento de novas técnicas que auxiliassem a administração do risco de crédito e realizaram uma avaliação explanatória inicial sobre até que ponto os modelos preditivos desenvolvidos no Brasil ainda poderiam ser capazes de predizer corretamente o futuro de empresas, escolhidas entre aquelas que apresentaram graves dificuldades financeiras no ano de 1997, seguidas de concordata ou falência, e algumas empresas consideradas como tradicionais e solventes.

Ao analisar modelos desenvolvidos por Kanitz, Elizabetsky, Matias, Pereira, Bragança e Altman, de 1976 a 1996, os autores concluíram que os resultados apresentados por estes modelos não são mais satisfatórios, dados mais de dez anos decorridos do seu desenvolvimento. Logo, as hipóteses levantadas são de que, em função da antiguidade, da falta de especificidade no que tange a atividade econômica das empresas para que fossem designados, ou por outras razões quaisquer, os modelos não conseguem atingir níveis satisfatórios de sucesso na previsão da situação da empresa em horizontes de cinco, três e, até mesmo, um ano. Ao finalizarem dizendo que a academia pode contribuir de forma significativa, explorando com mais força o tema, justifica-se a importância deste estudo, ao buscar o desenvolvimento e o aprimoramento de modelos utilizados na tomada de decisão da concessão de crédito.

Hill, O`Conner e Remus (1996) também realizaram um estudo exploratório sobre modelos preditivos, como redes neurais, modelos de regressão e análise discriminante. Na verdade, o objetivo dos autores é comparar o desempenho das redes neurais com o desempenho de outros modelos preditivos disponíveis na literatura. Os resultados indicaram que as redes neurais apresentam desempenho tão bom quanto (e às vezes superior) aos outros modelos estatísticos. Condições como linearidade e volatilidade (variância) dos dados analisados influenciam diretamente o desempenho das redes, afinal, modelos lineares podem apresentar melhores resultados de previsão para dados lineares, bem como as redes neurais fazem com dados não-lineares. Quanto à volatilidade, observa-se que as redes neurais apresentam melhores resultados quando os dados são menos voláteis.

De acordo com Henley e Hand (1996, p. 77), dentre as metodologias tradicionais de análise de crédito destacam-se a análise discriminante, as regressões linear e logística, a programação linear e as árvores de decisão.

O trabalho de Lima e Almeida (2004), que verifica o ganho na previsão de séries temporais financeiras com o uso de *wavelets*; bem como o trabalho de Almeida et al. (2004), que identifica um algoritmo genético que pode ser utilizado na parametrização de redes neurais artificiais para aplicação na elaboração de um orçamento de vendas; o estudo de Bressan (2004), que trata da aplicabilidade de modelos de previsão de séries temporais como ferramenta de decisão de compra e venda de contratos futuros de boi gordo, café e soja, em datas próximas ao vencimento; e a pesquisa de Broomhead e Lowe (1988), que investiga os

pressupostos implícitos quando se emprega um modelo de rede neural de alimentação para frente na solução de problemas complexos, são alguns exemplos da adoção de uma nova geração de metodologias.

Explorar novas metodologias que estão disponíveis na literatura, bem como comparar suas eficiências com as mais antigas, com o objetivo de obter melhora na capacidade preditiva da variável binária, que é utilizada como ferramenta para a análise de concessão de crédito, resume a proposta do presente estudo.

2.2 APLICAÇÃO DAS TÉCNICAS DE ANÁLISE DISCRIMINANTE E REGRESSÃO LOGÍSTICA

Rosenberg e Gleit (1994) discutiram em seu trabalho o uso da análise discriminante, das árvores de decisões e de sistemas especialistas, em decisões estatísticas e o uso de programação dinâmica, programação linear e cadeias de Markov, em modelos de decisão dinâmica. A intenção dos autores foi a de revisar as técnicas matemáticas mais importantes encontradas na literatura e relacionar a teoria com a prática, na solução de problemas envolvidos no processo decisório da concessão de crédito. Em nenhum momento os autores apontam um dos métodos como o melhor, ao invés disso, chamam a atenção para o fato de que geralmente esses modelos são aplicados na tomada de decisões do tipo liberar ou não o crédito para um cliente, e que existem muitas outras decisões importantes que estão relacionadas à concessão de crédito que deveriam ser mais bem exploradas.

Ademais, os autores afirmam que a análise discriminante é a técnica quantitativa mais utilizada na análise da concessão de crédito, de acordo com as respostas observadas em sua pesquisa.

Hand e Henley (1997) também desenvolveram um trabalho para estudar parte das técnicas de análise de crédito disponíveis. Os autores afirmam que, em geral, não existe um método que seja superior aos outros. A escolha do modelo que será utilizado para resolver um problema depende das características apresentadas por este problema.

Quanto à análise discriminante, os autores defendem que o fato de determinado estudo não respeitar o pressuposto da normalidade das variáveis independentes não torna a aplicação de tal análise inadequada, inclusive porque o pressuposto de normalidade determina que a sua violação possa levar a decisões incorretas, principalmente quando as amostras são pequenas, o que não é o caso deste trabalho.

Os mesmos autores acreditam que, se a variável apresenta uma distribuição normal, então a análise discriminante linear é a ideal (ignorando a variação da amostra). No entanto, se a análise discriminante for considerada como um produto da combinação linear das variáveis, que maximiza um critério particular de separação, então ela é perfeitamente aplicável. Esta afirmativa também justifica o fato de as variáveis utilizadas para rodar a análise discriminante neste trabalho não terem sido normalizadas. Ademais, a transformação das variáveis para terem média 0 e desvio padrão 1 nem sempre consiste na melhor estratégia, pois a variabilidade de uma determinada medida pode fornecer informação útil, que será eliminada com a padronização dos dados.

Brito e Assaf Neto (2008) desenvolveram um modelo de classificação de risco de crédito para grandes empresas que atuam no Brasil, utilizando a técnica estatística regressão logística. Os autores afirmam que o modelo proposto estabelece uma relação estatística entre o *default* da empresa e um conjunto de índices econômico-financeiros calculados a partir das demonstrações contábeis. Com base nessa relação, o estudo avaliou se as demonstrações contábeis fornecem informações que permitam aos seus diversos usuários prever a ocorrência de uma insolvência empresarial. Cada instituição financeira adota seu próprio conceito de evento de *default*, estando normalmente relacionado ao atraso no pagamento de um compromisso assumido pelo tomador por períodos como 60 ou 90 dias.

A amostra final utilizada no desenvolvimento do modelo compreendeu 60 empresas, sendo 30 insolventes que se tornaram concordatárias ou falidas entre 1994 e 2004, e 30 solventes que foram emparelhadas com as primeiras. O nível de acerto do modelo desenvolvido foi de 91,7%, tendo sido classificadas corretamente 55 das 60 empresas da amostra. O nível de acerto obtido na validação do modelo foi de 88,3%, tendo sido classificadas incorretamente 7 empresas da amostra (11,7%). Os resultados denotam um excelente desempenho do nível de acerto do modelo de regressão logística desenvolvido.

Conforme os autores, a regressão logística é uma técnica de análise multivariada apropriada para as situações nas quais a variável dependente é categórica e assume um entre dois resultados possíveis (binária), tais como normal ou anormal, cliente ou não cliente e solvente ou insolvente. Além disso, os mesmos afirmam que esta técnica apresenta certas vantagens em relação à análise discriminante linear, principalmente devido às suas suposições iniciais serem menos rígidas. A análise discriminante linear está baseada em uma série de pressupostos bastante restritivos, como a normalidade das variáveis independentes. Essas suposições podem não ser válidas em muitas situações práticas de análise de crédito, principalmente quando há variáveis independentes de natureza não métrica.

2.3 APLICAÇÃO DE MODELOS DE REDES NEURAIS

Lin et al. (1995, p. 2) chamam atenção para o fato de que poucos trabalhos descrevem quais os cuidados devem ser tomados ao determinar os sinais de entrada de uma rede neural, dado que o desempenho das previsões pode ser melhorado com esses cuidados. Os autores afirmam que a análise da autocorrelação pode ser usada para determinar os padrões corretos dos sinais de entrada de uma rede neural utilizada na previsão de séries temporais não-lineares. Fazem, ainda, uma distinção entre previsões curtas e previsões longas. As previsões curtas são representadas por um passo à frente e envolvem termos residuais, enquanto que as previsões longas são importantes para determinar tendências futuras ou qualquer outra quantidade de passos à frente, que não um passo apenas. Além disso, as previsões longas não envolvem termos residuais.

Depois de determinar os padrões dos sinais de entrada, por meio da análise da autocorrelação, os autores determinaram a quantidade de neurônios na camada oculta, por meio da regra de *Baum-Haussler*. Para maiores detalhes ver Lin et al. (1995, p. 3).

As redes neurais propostas pelos autores foram utilizadas para previsões da média semanal do índice da Bolsa de Valores de Hong Kong, de 1980 a 1990. Um total de 500 semanas foi utilizado para treinar a rede e previsões para as próximas 20 a 50 semanas foram realizadas para comparação. A rede utilizada foi a *Three Layered*, com alimentação para frente, e o algoritmo utilizado foi *o QuickProp*. O desempenho das redes neurais foi medido pelo *Mean Absolute Percentage Error* (MAPE).

Para previsões de mais de 60 semanas foram testadas as redes [1,0,1,0], [3,0,1,0], [4,0,1,0], [3,1,1,0], [3,1,1,2]; a rede que apresentou menor MAPE foi a rede [3,1,1,2]. Já para previsões de mais de 20 semanas foram testadas as redes [1,0,1,0], [3,0,1,0], [4,0,1,0], [3,1,1,0]; a rede que apresentou menor MAPE foi a [3,1,1,0]. Os resultados das simulações indicam que a adoção de padrões apropriados para os sinais de entrada podem melhorar o desempenho das previsões.

Correa e Machado (2004, p. 4) criaram um modelo de análise de crédito por meio da utilização de redes neurais, para prever o risco de inadimplência de clientes no produto cheque especial. A amostra utilizada era composta por 2.868 contas de pessoas físicas residentes em um Estado da Paraíba (abertas de março a setembro de 2003), que receberam um limite de cheque especial de R$50,00. Desses 2.868 clientes, 1.774 (61,85%) eram considerados adimplentes e 1.094 (38,15%) inadimplentes. Foram classificados como inadimplentes as contas que, durante os seis primeiros meses de existência, ficaram em atraso por mais de 90 dias, ou seja, permaneceram durante 90 dias com saldo devedor maior do que o limite de crédito concedido.

A rede neural utilizada foi a *Multilayer Percepton* e o algoritmo foi o de retropropagação (*backpropagation*). Foram rodadas sete redes, com arquiteturas distintas, e a comparação e seleção das melhores redes foram feitas com base no percentual de acertos do modelo, considerando uma pontuação de corte igual a 0,5. A melhor rede apresentou acertos gerais de 70,03% sobre a base de teste e 70,57% sobre a base toda e acertos de clientes bons de 84,92% sobre a base de teste e 84,55% sobre a base toda. Os autores chamam atenção para o fato de que mudanças na pontuação de corte entre clientes bons e maus podem ser feitas com a finalidade de liberar ou restringir a concessão de crédito.

Fonseca e Omaki (2004, p. 11) propuseram um modelo de rede neural que pudesse ser utilizado na segmentação de clientes organizacionais, utilizando, para tal, o risco percebido no processo de decisão de compra. Na rede neural proposta, cada neurônio da primeira camada corresponde a um tipo de risco percebido. A primeira camada representa a entrada dos dados, que fornecerá o grau de percepção de cada tipo de risco, sendo obtida por meio do somatório ou multiplicação entre o grau de incerteza deste risco com a consequência de sua ocorrência. O resultado de cada tipo de risco é levado para a camada oculta da rede. Nela é realizado o

somatório ou a multiplicação de todos os riscos, fato que irá permitir identificar a percepção global do risco. A rede neural fornece, então, um resultado que corresponde a classificação do perfil do indivíduo, baseado na sua percepção da existência de risco durante uma situação de compra, seja ela alta, moderada ou baixa (variáveis de saída). Após montada sua estrutura, bastaria apenas inserir os dados na rede e ela forneceria o perfil do cliente.

Os autores afirmam que se realizada de forma adequada, eficaz e cuidadosa, a segmentação pode poupar tempo e dinheiro, no preparo de programas de *marketing* que pudessem ser ineficientes, evitando que sejam despendidos esforços, para atender ao mercado-alvo errado. Além do mais, a segmentação torna possível, por exemplo, compreender melhor o perfil de clientes, adequando assim, da melhor forma possível, a oferta às necessidades dos mesmos.

O cuidado com a escolha dos sinais de entrada de uma rede neural, bem como com a escolha de sua estrutura (arquitetura, quantidade de camadas ocultas e algoritmo de processamento), de maneira que estejam adequadas a cada tipo de problema que se pretende resolver, são cruciais para seu bom desempenho. Qualquer alteração nesses itens pode tanto prejudicar, quanto maximizar os resultados esperados.

Com relação ao presente trabalho, considerando que seu principal objetivo é comparar combinações de técnicas de análise e, ainda, que foi utilizada a ferramenta *Intelligent Problem Solver* para rodar as redes neurais, o foco não foi a estrutura das redes, mas sim qual delas apresentou o melhor desempenho.

2.4 APLICAÇÃO DE MODELOS HÍBRIDOS

A combinação de modelos de análise como, por exemplo, em Li e Xu (2009), que combinam a análise do componente principal com redes neurais na previsão do risco de crédito de 20 empresas clientes de um banco comercial chinês, representa melhores resultados preditivos do que quando os modelos são aplicados separadamente; tem sido uma prática utilizada com frequência.

> ... o crescimento do número de empresas comerciais que detectaram falência com a crise econômica vivida nos últimos tempos tem recebido atenção especial. O desenvolvimento de um sistema eficiente de tomada de decisão, que possibilite previsões confiáveis sobre o comportamento das empresas, tornou-se desejável e justificável. Novas técnicas de previsão, que antecipem advertências e permitam os ajustes necessários para a estabilidade econômica das empresas, têm sido testadas, como é o caso das redes neurais.(LACHTERMACHER; ESPENCHITT, 2001, p. 1)

Os autores realizaram um estudo comparativo entre a análise discriminante e as redes neurais. O espaço amostral foi composto por um conjunto de empresas de construção civil, montagem industrial e de projetos de arquitetura e engenharia, pesquisadas no período de 1.983 a 1.993, do Estado do Rio de Janeiro, prestadoras de serviços à Petrobrás S.A. O tamanho do conjunto de exemplos para modelagem de análise discriminante ficou restrito a um total de 83 observações, sendo 48 observações do grupo com excelentes resultados e 35 do grupo com falência requerida ou decretada nesse período. As variáveis utilizadas no modelo foram índices financeiros. Os resultados das classificações corretas foram 81% para a análise discriminante e 88% para as redes neurais, mais uma vez comprovando a superioridade do desempenho das redes neurais. Além disso, mesmo que as variáveis tenham sido normalizadas, os resultados obtidos corroboram com os resultados obtidos neste trabalho, no qual as variáveis não foram tratadas, ou seja, as redes neurais apresentam desempenho superior à análise discriminante (para 50% dos clientes classificados como adimplentes e 50% como inadimplentes).

Hua, Song e Li (2009) combinaram o método Grey de classificação em *clusters* com redes neurais na previsão de risco de crédito de 26 empresas (e 40 variáveis) de Tecnologia da Informação, das quais 13 apresentavam problemas com crédito e 13 não apresentavam esse tipo de problema (18 empresas foram utilizadas para treinar a rede e 8 empresas foram utilizadas para testá-la). A rede neural utilizada foi a *Three Layer Percepton* e o algoritmo foi o de retropropagação (*backpropagation*).

A classificação correta do risco de crédito das empresas para três anos, antes da crise, foi de 62,5%, sem a combinação da rede neural com a análise de *clusters*, e de 75% com a combinação das duas técnicas. Para dois anos antes da crise, sem a combinação o acerto foi de 37,5% e com a combinação foi de 62,5%. O acerto para um ano antes da crise foi de 75% com

e sem a combinação. Os resultados mostram que a classificação das empresas em *clusters* melhora o desempenho de classificação da rede neural.

Lima e Pereira (2010) combinaram a segmentação por meio de *cluster*s e redes neurais, visando a otimização da análise de crédito ao consumidor, aplicada a uma amostra de clientes de uma grande rede varejista de lojas.

> ...nenhuma promessa de pagamento futuro ou devolução de dinheiro é absolutamente segura. O que existe é uma probabilidade de que o devedor cumpra sua promessa. Para eles, é interessante ressaltar que muitas vezes a quebra da promessa do devedor pode ser involuntária e devida à degradação do conjunto de variáveis (econômicas e financeiras), externas ou internas, vigentes na época do contrato. (LIMA; PEREIRA, 2010).

A amostra tomada de forma probabilística foi composta por 2.500 observações (clientes), das quais 25 foram excluídas por conterem *missing values*. A amostra para análise ficou composta por 2.349 (95%) bons pagadores e 126 (5%) maus pagadores, somando 2.475 indivíduos. A rede utilizada foi a *Multilayer Percepton*. O ganho com a utilização da análise de *clusters* no acerto de bons pagadores foi de 15,8% e no acerto de maus pagadores foi de 8,8%. O ganho geral de acertos estimado foi de 15,47%, independentemente do custo de classificações erradas. Certamente um ganho significativo em modelos preditivos híbridos.

Lemos, Steiner e Nievola (2005, p. 7) comparam redes neurais e árvores de decisão na análise de crédito bancário. Foram utilizados dados históricos de 339 clientes pessoa jurídica de uma agência do Banco do Brasil, em Guarapuava – PR, dos quais 266 são adimplentes e 73 inadimplentes. A rede neural utilizada foi a *Multilayer Percepton* e o algoritmo foi o de retropropagação (*backpropagation*). As árvores de decisão apresentaram um erro de classificação na fase de treinamento igual a 11,49% e um erro de 28,13% na fase de testes. Já as redes neurais apresentaram um erro de classificação na fase de treinamento igual a 4,09%, enquanto que o erro na fase de testes foi 9,96%. Os resultados deixam claro que as redes neurais apresentam melhor desempenho na classificação dos clientes deste banco.

Assim, conforme trabalhos citados, é possível afirmar que a combinação de metodologias maximiza a capacidade preditiva dos modelos, fato que reafirma a importância

deste trabalho, uma vez que, ao se constatar que a combinação da segmentação de dados com outras técnicas maximiza a capacidade preditiva de variáveis utilizadas na decisão de concessão de crédito, cria-se a possibilidade do desenvolvimento de novos modelos, mais eficientes e baratos, que poderão ser utilizados por diversas empresas.

A metodologia adotada para a realização desta pesquisa será apresentada no Capítulo a seguir, no qual se justifica os critérios adotados com base na revisão bibliográfica feita neste capítulo.

3. METODOLOGIA

3.1 ASPECTOS GERAIS

> ... analisar dados significa trabalhar com todo o material obtido durante o processo de investigação, ou seja, com os relatos de observação, as transcrições de entrevistas, as informações dos documentos e outros dados disponíveis. O processo de análise de dados deve ocorrer de forma sistematizada. Inicialmente, é recomendável que o pesquisador encontre meios de organizar o material coletado durante a pesquisa e *a posteriori* analise-os, com maior profundidade, à luz das teorias da metodologia científica. (BEUREN, 2006, p. 136).

Fávero, Belfiore e Silva (2009, p. 3) afirmam que a análise multivariada vem apresentando fundamental importância para a tomada de decisões nos mais variados campos do conhecimento. A quantidade de dados extraída de uma determinada pesquisa pode ser extremamente elevada, dificultando a determinação de como se dão as inter-relações entre as variáveis e, principalmente, a definição do modelo mais apropriado para se chegar às respostas desejadas. Com o crescente desenvolvimento computacional, a análise multivariada passou a ser utilizada para a avaliação do comportamento e de tendências presentes nas mais diferentes áreas do conhecimento.

Segundo os mesmos autores, a elaboração de mapas perceptuais, a criação de modelos de previsão ou a determinação de como um conjunto de variáveis se comporta, quando da alteração de uma ou mais variáveis presentes em outro conjunto, são mecanismos atualmente possíveis graças ao desenvolvimento de *softwares* como SPSS, Stata, Matlab, Minitab, entre outros. A análise multivariada é utilizada para se estudar modelos em que todas as variáveis sejam aleatórias e inter-relacionadas, de modo que seus diferentes efeitos não possam ser interpretados separadamente. Reduzir dados ou simplificar sua estrutura pode ser uma ferramenta muito útil afim de propiciar uma interpretação mais fácil, sem que haja perda considerável de informações. A análise de dados precisa ser entendida como um processo para se atingir um grupo de informações claras e objetivas, voltadas especificamente para uma melhor tomada de decisões.

A seguir são discutidas as técnicas de análise discriminante, redes neurais e segmentação de dados.

3.2 ANÁLISE DISCRIMINANTE

> ... a análise discriminante é uma técnica multivariada aplicada quando a variável dependente é qualitativa e as variáveis independentes são quantitativas. As variáveis dicotômicas, como o sexo, podem ser incluídas no modelo como variáveis explicativas. (PESTANA; GAGEIRO, 2003, p. 655).

Segundo os autores, a análise discriminante permite conhecer as variáveis mais importantes que discriminam os grupos; classificar novos casos por meio da sua inserção na base de dados; escolher um subconjunto alternativo de variáveis com dimensão semelhante à do modelo inicial, que discrimine bem os grupos; identificar grupos similares; identificar casos *outliers*; e validar a análise de *clusters*.

> ... a análise discriminante é uma técnica de dependência utilizada quando a única variável dependente é dicotômica (por exemplo, grupo A ou grupo B) ou multicotômica (por exemplo, *cluster* A, *cluster* B ou *cluster* C) e, portanto, não métrica. As variáveis explicativas devem ser métricas ou não métricas transformadas em variáveis binárias. (FÁVERO; BELFIORE; SILVA, 2009, p. 12).

No presente trabalho, transformou-se o *status* adimplente ou inadimplente em variável binária 0 ou 1, respectivamente. Clientes com atrasos maiores que 60 dias são classificados como inadimplentes.

Encontram-se múltiplas aplicações para este tipo de análise em situações nas quais o objetivo principal é identificar o grupo ao qual um objeto pertence. Em cada caso, os objetos recaem em grupos e é desejado que a pertinência a um grupo possa ser prevista ou explicada por um conjunto de variáveis independentes. Portanto, a análise discriminante pode ser considerada como um tipo de análise de perfil ou uma técnica preditiva analítica. Exemplos de aplicações potenciais incluem prever o sucesso ou fracasso de um novo produto, prever se

uma empresa terá sucesso e determinar o risco de crédito de uma pessoa (HAIR et al., 2005, p. 205), como é o caso deste trabalho, no qual a análise discriminante foi utilizada para classificar os clientes de uma grande rede varejista como adimplentes ou inadimplentes.

A análise discriminante envolve determinar uma variável estatística, que represente a combinação linear das variáveis independentes selecionadas por seu poder discriminatório, o qual é usado na previsão de pertinência ao grupo. A discriminação é conseguida estabelecendo-se os pesos da variável estatística para cada variável independente, visando maximizar a variância entre grupos relativa à variância dentro dos grupos. As variáveis com as maiores diferenças entre os grupos são identificadas e um coeficiente discriminante que pondera cada variável para refletir essas diferenças é determinado. Para entender as diferenças de grupos, deve-se discernir o papel de cada variável, bem como definir combinações dessas variáveis, que representam dimensões de discriminação entre grupos. Essas dimensões são os efeitos coletivos de diversas variáveis que trabalham, conjuntamente, para distinguir entre os grupos (HAIR et al., 2005, p. 209).

O valor previsto da função discriminante é o escore Z discriminante, o qual foi calculado para cada um dos 50 mil clientes da base de dados analisada. A função discriminante toma a forma da equação linear:

$$Z_{jk} = \alpha + W_1 X_{1K} + W_2 X_{2K} + \ldots + W_n X_{nK}, \quad (3.1)$$

onde

Z_{jk} = escore Z discriminante da função discriminante j para o objeto k

α = intercepto

W_i = peso discriminante para a variável independente i

X_{ik} = variável independente i para o objeto k.

Segundo Pestana e Gageiro (2003, p. 656) os pressupostos da análise discriminante são:

1. Cada grupo é uma amostra aleatória de uma população normal multivariada. A sua violação pode levar a decisões incorretas, principalmente quando as amostras são pequenas. Quando a violação da normalidade se deve apenas à não simetria da

distribuição, a potência do teste não é afetada, contrariamente ao que acontece se a distribuição não for mesocúrtica, e de forma mais acentuada se for platicúrtica, caso em que se deve optar pela regressão logística;

2. Dentro dos grupos a variabilidade é idêntica, isto é, as matrizes de variâncias e covariâncias são iguais para todos os grupos. A verificação deste pressuposto é feita na própria análise discriminante, por meio do Teste M de Box. Caso seja violado, aumenta a probabilidade dos casos serem classificados no grupo com maior dispersão.

O grau de achatamento de uma distribuição em relação à distribuição normal é chamado de curtose. A distribuição de referência (distribuição normal) é denominada mesocúrtica; quando a distribuição apresenta uma curva de frequência mais fechada (mais aguda em sua parte superior), ela é denominada leptocúrtica; quando a distribuição apresenta uma curva de frequência mais aberta (mais achatada em sua parte superior), ela é denominada platicúrtica (MAGRINI, 2010, p. 3). A adoção do pressuposto 1 será mais bem explorada a seguir. Quanto ao pressuposto 2, o Teste M de Box será analisado no Capítulo 5.

> A análise discriminante é única em uma característica entre as relações de dependência: se há mais de dois grupos na variável dependente, ocorre o cálculo de mais de uma função discriminante (HAIR et al., 2005, p. 209).

Para este estudo foi gerada apenas uma função discriminante em todas as análises discriminantes rodadas, uma vez que existem apenas dois grupos, adimplente e inadimplente. Tais análises serão mais bem detalhadas nos capítulos seguintes.

3.3 REDES NEURAIS

Almeida e Nakao (2007, p. 433) analisam uma das técnicas de tratamento de dados recente que tem despertado grande interesse tanto de pesquisadores da área de tecnologia, como da área de negócios, as redes neurais. Esta técnica torna-se útil quando há a necessidade de se reconhecerem padrões a partir do acúmulo de experiência ou de exemplos cuja representação é complexa. Na computação tradicional as informações são processadas de

forma serial, ou seja, em sequência. Já na computação com redes neurais as informações podem ser processadas em paralelo.

Pesquisas recentes indicam que as redes neurais apresentam relevante poder de classificação e de reconhecimento de capacidades. Este modelo de previsão aprende, por meio de exemplos e captura os relacionamentos entre os dados, mesmo que esses relacionamentos sejam desconhecidos ou de difícil descrição. Além disso, as redes neurais têm a capacidade de fazer inferência sobre a população (parte desconhecida), mesmo quando a amostra apresentar algumas informações viesadas. Tal propriedade se justifica pelo fato de a previsão ser feita por meio de inferências sobre o futuro (desconhecido), que são baseadas em exemplos do passado (ZHANG et al., 1997, p. 35).

Outra característica importante das redes neurais é que elas apresentam formas mais gerais e flexíveis do que outros métodos estatísticos. As redes neurais são não-lineares e podem ser utilizadas sem que se tenha um conhecimento prévio do relacionamento entre as variáveis de entrada e de saída. (ZHANG et al., 1997, p. 36).

O trabalho em redes neurais artificiais tem sido motivado desde o começo pelo reconhecimento de que o cérebro humano processa informações de uma forma inteiramente diferente do computador digital convencional. O cérebro é um computador (sistema de processamento de informação) altamente complexo, não-linear e paralelo. No momento do nascimento, um cérebro tem uma grande estrutura e a habilidade de desenvolver suas próprias regras através do que se denomina experiência. Na verdade, a experiência vai sendo acumulada com o tempo, sendo que o mais dramático desenvolvimento do cérebro humano acontece durante os dois primeiros anos de vida, e continua para muito além desse estágio (HAYKIN, 2001, p. 27).

É evidente que uma rede neural extrai seu poder computacional por meio, primeiro, de sua estrutura maciça paralelamente distribuída e, segundo, de sua habilidade de aprender e de generalizar. A generalização se refere ao fato de a rede neural produzir saídas adequadas para entradas que não estavam presentes durante o treinamento (aprendizagem). Um neurônio em desenvolvimento é sinônimo de um cérebro plástico: a plasticidade permite que o sistema nervoso em desenvolvimento se adapte ao seu meio ambiente.

> Assim como a plasticidade parece ser essencial para o funcionamento dos neurônios como unidades de processamento de informação do cérebro humano, também ela o é com relação às redes neurais construídas com neurônios artificiais. Na sua forma mais geral, uma rede neural é uma máquina projetada para modelar a maneira como o cérebro realiza uma tarefa particular ou função de interesse (HAYKIN, 2001, p. 28).

O sistema nervoso humano pode ser visto como um sistema de três estágios, como mostrado na Figura 2:

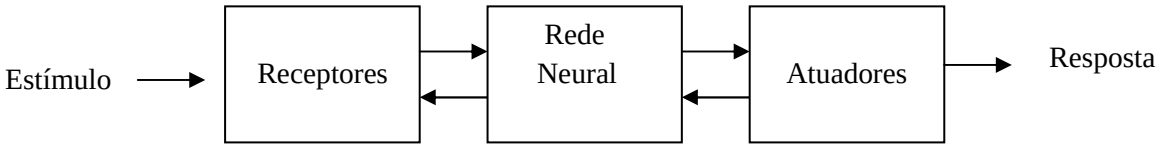

Figura 2 - Representação em diagrama em blocos do sistema nervoso.
Fonte: HAYKIN (2001, p. 32)

> Os receptores convertem estímulos do corpo humano ou do ambiente externo em impulsos elétricos, que transmitem informação para o cérebro. Os atuadores convertem impulsos elétricos gerados pela rede neural em respostas discerníveis como saídas do sistema (HAYKIN, 2001, p. 32).

Analogamente ao que ocorre no cérebro humano, o neurônio da rede neural artificial recebe estímulos, chamados de sinais de entrada, que são multiplicados por pesos sinápticos e somados. Uma função de ativação é acionada, de maneira que os sinais de saída sejam restringidos. O modelo não-linear de um neurônio está representado na Figura 3:

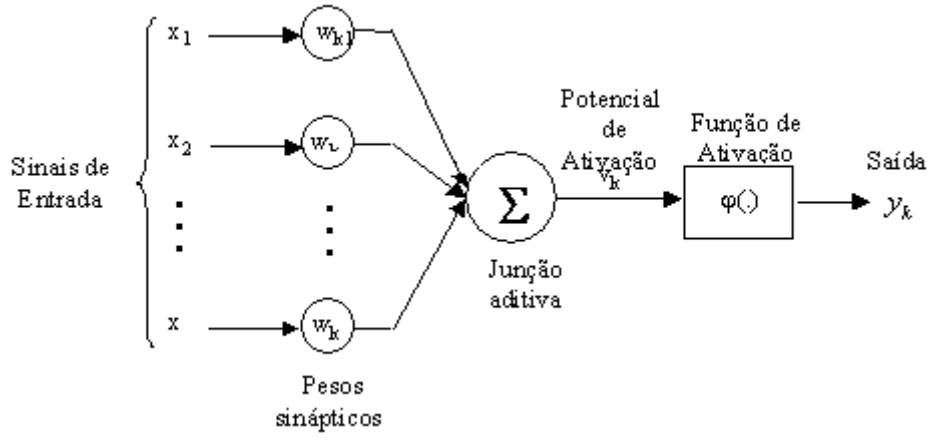

Figura 3 - Modelo não-linear de um neurônio.
Fonte: HAYKIN (2001, p. 36)

Os três elementos básicos do modelo neuronal são: (1) Um conjunto de sinapses ou elos de conexão, cada uma caracterizada por um peso ou força própria. Especialmente, um sinal x_j na entrada da sinapse j conectada ao neurônio k é multiplicado pelo peso sináptico w_{kj}; (2) Um somador para somar os sinais de entrada, ponderados pelas respectivas sinapses do neurônio; (3) Uma função de ativação para restringir a amplitude da saída de um neurônio. Tipicamente, o intervalo normalizado da amplitude da saída de um neurônio é escrito como o intervalo unitário fechado [0,1] ou alternativamente [-1,1]. (HAYKIN, 2001, p. 36).

Na Figura 3, os $x_1, x_2, \ldots, x_m$ são os sinais de entrada da rede; $w_{k1}, w_{k2}, \ldots, w_{km}$ são os pesos sinápticos do neurônio k; u_ké a saída do combinador linear devido aos sinais de entrada; $\varphi(\cdot)$ é a função de ativação e y_k é o sinal de saída do neurônio. Ligando os elementos acima do ponto de vista matemático, o processo de funcionamento da rede pode ser escrito por meio das seguintes equações:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{3.2}$$

$$v_k = u_k + b_k \tag{3.3}$$

$$y_k = \varphi(u_k + b_k) \tag{3.4}$$

As equações acima são descritas a partir da maneira pela qual os neurônios de uma rede neural estão estruturados, que está intimamente ligada com o algoritmo de aprendizagem utilizado para treinar a rede. Em uma rede neural com uma arquitetura específica, a representação do conhecimento do meio ambiente é definida pelos valores assumidos pelos parâmetros livres, pesos sinápticos e bias da rede. A forma dessa representação de conhecimento constitui o verdadeiro projeto da rede neural e, portanto, é a chave para o seu desempenho.

> O processamento realizado em um neurônio artificial pode ser dividido em três passos: os dados que passam ao longo das linhas de entrada para os neurônios são multiplicados pelos pesos; todos esses dados que foram multiplicados pelos pesos são somados dentro do neurônio; o valor total dessa soma é passado através de uma função de transferência, cuja saída representa o valor de saída do neurônio. A função de transferência mais simples é a função linear, adequada para um espaço linearmente separável, a qual iguala a saída à entrada. A utilização de funções de transferência não lineares, apropriadas para um espaço não linearmente separável, é uma das características principais da rede neural (LIMA et al., 2009, p. 38).

Uma rede neural muito utilizada é a *Multilayer Percepton* (*perceptrons* de múltiplas camadas), que consiste de um conjunto de unidades sensoriais (nós de fonte) que constituem a camada de entrada, uma ou mais camadas ocultas de nós computacionais e uma camada de saída de nós computacionais. O sinal de entrada se propaga para frente, através da rede, camada por camada. Os *perceptrons* de múltiplas camadas têm sido aplicados com sucesso para resolver diversos problemas difíceis, através do seu treinamento de forma supervisionada, com um algoritmo conhecido como algoritmo de retropropagação de erro (*error backpropagation*). (HAYKIN, 2001, p. 183).

> A aprendizagem por retropropagação de erro consiste de dois passos através das diferentes camadas da rede: um passo para frente, a propagação, e um passo para trás, a retropropagação. No passo para frente, um padrão de atividade (vetor de entrada) é aplicado aos nós sensoriais da rede e seu efeito se propaga através da rede, camada por camada. Finalmente, um conjunto de saídas é produzido como a resposta real da rede. Durante o passo de propagação, os pesos sinápticos da rede são todos fixos. Durante o passo para trás, por outro lado, os pesos sinápticos são todos ajustados de acordo com uma regra de correção de erro. Especificamente, a resposta real da rede é subtraída de uma resposta desejada (alvo) para produzir um sinal de erro. Este sinal de erro é, então, propagado para trás através da rede, contra a direção das conexões sinápticas – vindo daí o nome de retropropagação do erro. Os pesos sinápticos são ajustados para fazer com que a resposta da rede se mova para mais perto da resposta desejada, em um sentido estatístico (HAYKIN, 2001, p. 183).

A seguir são apresentadas todas as etapas do algoritmo de retropropagação, de acordo com Haykin (2001).

O sinal de erro na saída do neurônio j, na iteração n (isto é, a apresentação do n-ésimo exemplo de treinamento), é definido por

$$e_j(n) = d_j(n) - y_j(n), \text{ o neurônio j é um nó de saída.} \tag{3.5}$$

Define-se o valor instantâneo da energia do erro para um neurônio j como $\frac{1}{2} e_j^2(n)$.

Correspondentemente, o valor instantâneo $\varepsilon(n)$ da energia total do erro é obtido somando-se os termos $\frac{1}{2} e_j^2(n)$ de todos os neurônios da camada de saída; são os únicos neurônios "visíveis" para os quais os sinais de erro podem ser calculados diretamente. Pode-se assim escrever:

$$\varepsilon(n) = \frac{1}{2} \sum_{j \in C} e_j^2(n), \tag{3.6}$$

onde o conjunto C inclui todos os neurônios da camada de saída da rede.

Considere que N represente o número total de padrões (exemplos) contidos no conjunto de treinamento. A energia média do erro quadrado é obtida somando-se $\varepsilon(n)$ para todos os n e, então, normalizando em relação ao tamanho do conjunto N, como mostrado por

$$\varepsilon_{med} = \frac{1}{N} \sum_{n=1}^{N} \varepsilon(n). \tag{3.7}$$

A energia instantânea do erro $\varepsilon(n)$ e, consequentemente, a energia média do erro ε_{med}, são funções de todos os parâmetros livres (isto é, pesos sinápticos e níveis de bias) da rede. Para um dado conjunto de treinamento, ε_{med} representa a função de custo como uma medida do desempenho de aprendizagem. O objetivo do processo é ajustar os parâmetros livres da rede para minimizar ε_{med}. Para fazer esta minimização, considera-se um método simples de treinamento no qual os pesos são atualizados de padrão em padrão até formar uma

época, isto é, uma apresentação completa do conjunto de treinamento inteiro que está sendo processado. Os ajustes dos pesos são realizados de acordo com os respectivos erros calculados para cada padrão apresentado à rede. A média aritmética destas alterações individuais de peso sobre o conjunto de treinamento é, portanto, um estimativa da alteração real que resultaria da modificação dos pesos baseada na minimização da função de custo ε_{med} sobre o conjunto de treinamento inteiro.

Considere um neurônio j sendo alimentado por um conjunto de sinais funcionais produzidos por uma camada de neurônios à sua esquerda. O campo local induzido $v_j(n)$ produzido na entrada da função de ativação associada ao neurônio j é, portanto,

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n), \tag{3.8}$$

onde m é o número total de entradas (excluindo o bias) aplicadas ao neurônio j. O peso sináptico w (correspondente à entrada fixa $y_0 = +1$) é igual ao bias b_j aplicado ao neurônio j. Assim, o sinal funcional $y_j(n)$ que aparece na saída do neurônio j na interação n é

$$y_j(n) = \varphi_j(v_j(n)). \tag{3.9}$$

O algoritmo de retropropagação aplica uma correção $\Delta w_{ji}(n)$ ao peso sináptico $w_{ji}(n)$, que é proporcional à derivada parcial $\partial\varepsilon(n)/\partial w_{ji}(n)$. De acordo com a regra da cadeia do cálculo, pode-se expressar este gradiente como:

$$\frac{\partial \varepsilon(n)}{\partial w_{ji}(n)} = \frac{\partial \varepsilon(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)}. \tag{3.10}$$

A derivada parcial $\partial\varepsilon(n)/\partial w_{ji}(n)$ representa um fator de sensibilidade, determinando a direção de busca no espaço de pesos para o peso sináptico w_{ji}.

Diferenciando ambos os lados da equação (3.10) em relação a $e_j(n)$, obtém-se:

$$\frac{\partial \varepsilon(n)}{\partial e_j(n)} = e_j(n). \tag{3.11}$$

Diferenciando ambos os lados da equação (3.11) em relação a $y_j(n)$, obtém-se:

$$\frac{\partial e_j(n)}{\partial y_j(n)} = -1. \tag{3.12}$$

A seguir, diferenciando a equação (3.12) em relação a $v_j(n)$, obtém-se:

$$\frac{\partial y_j(n)}{\partial v_j(n)} = \varphi_j'\left(v_j(n)\right). \tag{3.13}$$

onde o uso do apóstrofe (no lado direito) significa a diferenciação em relação ao argumento. Finalmente, diferenciar a equação (3.13) em relação a $w_{ji}(n)$ produz

$$\frac{\partial v_j(n)}{\partial w_{ji}(n)} = y_j(n). \tag{3.14}$$

O uso das equações (3.11) a (3.14) em (3.10) produz

$$\frac{\partial \varepsilon(n)}{\partial w_{ji}(n)} = -e_j(n)\varphi_j'\left(v_j(n)\right)y_j(n). \tag{3.15}$$

A correção $\Delta w_{ji}(n)$ aplicada a $w_{ji}(n)$ é definida pela regra delta:

$$\Delta w_{ji}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial w_{ji}(n)}, \tag{3.16}$$

onde η é o parâmetro da taxa de aprendizagem do algoritmo de retropropagação.

O uso do sinal negativo na equação (3.16) indica a descida do gradiente no espaço de pesos, isto é, busca uma direção para a mudança de peso que reduz o valor de $\varepsilon(n)$).

Correspondentemente, o uso da equação (3.15) em (3.16) produz

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_i(n), \tag{3.17}$$

onde o gradiente local $\delta_j(n)$ é definido por

$$\delta_j(n) = -\frac{\partial \varepsilon(n)}{\partial v_j(n)} = -\frac{\partial \varepsilon(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} = e_j(n) \varphi_j'\left(v_j(n)\right). \tag{3.18}$$

O gradiente local aponta para as modificações necessárias nos pesos sinápticos. De acordo com a equação (3.17), o gradiente local $\delta_j(n)$ para o neurônio de saída j é igual ao produto do sinal de erro $e_j(n)$ correspondente para aquele neurônio pela derivada $\varphi_j'(v_j(n))$ da função de ativação associada.

> Modelos de redes neurais não contam com uma base estatística predeterminada, ou seja, não trabalham com variáveis aleatórias que possuem determinada distribuição de probabilidade, apenas com entradas e saídas de informações. A ausência de uma base estatística impede, nesse sentido, a construção de intervalos de confiança para previsões resultantes dos modelos. Consequentemente, as previsões são sempre pontuais, mas passíveis de comparação com outros modelos, por meio de medidas como o Erro Percentual Médio de Previsão (BRESSAN, 2004, p. 13).”

Em geral, podem ser identificadas três diferentes classes de arquiteturas de rede: redes alimentadas adiante com camada única, redes alimentadas diretamente com múltiplas camadas e redes recorrentes. Em uma rede neural em camadas, os neurônios estão organizados na forma de camadas. Na forma mais simples de uma rede em camadas, temos uma camada de entrada de nós de fonte, que se projeta sobre uma camada de saída de neurônios (nós computacionais), mas não vice-versa. Em outras palavras, esta rede é estritamente do tipo alimentada adiante ou acíclica (HAYKIN, 2001, p. 46).

A segunda classe de uma rede neural alimentada adiante se distingue pela presença de uma ou mais camadas ocultas. A função dos neurônios ocultos é intervir entre a entrada

externa e a saída da rede de maneira útil. Adicionando-se uma ou mais camadas ocultas, a rede se torna capaz de extrair estatísticas de ordem elevada. Os nós de fonte da camada de entrada da rede fornecem os respectivos elementos do padrão de ativação (valor de entrada), que constituem os sinais de entrada aplicados aos neurônios (nós computacionais) na segunda camada, isto é, a primeira camada oculta. Os sinais de saída da segunda camada são utilizados como entradas para a terceira camada, e assim por diante, para o resto da rede. Uma rede neural recorrente se distingue de uma rede neural alimentada adiante por ter pelo menos um laço de realimentação (HAYKIN, 2001, p. 46).

Sabe-se, ainda, que o *perceptron* é a forma mais simples de uma rede neural usada para a classificação de padrões ditos linearmente separáveis, isto é, padrões que se encontram em lados opostos de um hiperplano. Basicamente, ele consiste de um único neurônio com pesos sinápticos ajustáveis e bias. O *perceptron* construído em torno de um único neurônio não-linear é limitado a realizar classificação de padrões com apenas duas classes (hipóteses).

A Figura 4 representa o grafo de fluxo de sinal do *perceptron*:

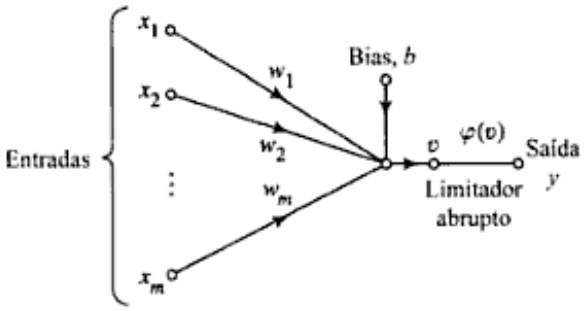

Figura 4 - Grafo de fluxo de sinal do *perceptron*.
Fonte: HAYKIN, 2001, p. 162.

Os pesos sinápticos do *perceptron* são representados por $w_1, w_2, ..., w_m$ e as entradas aplicadas a ele são representadas por $x_1, x_2, ..., x_m$. O bias aplicado externamente é representado por b. A entrada do limitador abrupto ou o campo local induzido do neurônio é:

$$v = \sum_{i=1}^{m} w_i x_i + b \tag{3.19}$$

O objetivo do *perceptron* é classificar corretamente o conjunto de estímulos aplicados externamente $x_1, x_2, \ldots, x_m$ em uma de duas classes C_1 e C_2. A regra de decisão para a classificação é atribuir o ponto representado pelas entradas $x_1, x_2, \ldots, x_m$ à classe C_1, se a saída do *perceptron* y for +1 e à classe C_2 se ela for -1. Os pesos sinápticos $w_1, w_2, \ldots, w_m$ do *perceptron* podem ser adaptados de iteração para iteração (HAYKIN, 2001, p. 162).

Aprender é equivalente a encontrar uma superfície, em um espaço multidimensional, que forneça o melhor ajuste para os dados de treinamento, com o critério melhor ajuste sendo medido em um sentido estatístico. Correspondentemente, generalização é equivalente ao uso desta superfície multidimensional para interpolar os dados de teste. No contexto de uma rede neural, as unidades ocultas fornecem um conjunto de funções que constituem uma base arbitrária para os padrões (vetores) de entrada, quando eles são expandidos sobre o espaço oculto: estas funções são chamadas de funções de base radial. A construção de uma rede de função de base radial, em sua forma mais básica, envolve três camadas com papéis totalmente diferentes. A camada de entrada é constituída por nós de fonte (unidades sensoriais) que conectam a rede ao seu ambiente. A segunda camada, a única camada oculta da rede, aplica uma transformação não-linear do espaço de entrada para o espaço oculto; e, na maioria das aplicações, esse espaço oculto é de alta dimensionalidade. A camada de saída é linear, fornecendo a resposta da rede ao sinal de ativação aplicado à camada de entrada (HAYKIN, 2001, p. 283).

Um *perceptron* de múltiplas camadas tem três características distintivas:

1. O modelo de cada neurônio da rede inclui uma função de ativação não-linear. Uma forma normalmente utilizada de não-linearidade que satisfaz esta exigência é a não-linearidade sigmóide, definida pela função logística:

$$y_j = \frac{1}{1 + \exp(-v_j)}, \tag{3.20}$$

onde v_j é a soma ponderada de todas as entradas sinápticas acrescidas do bias do neurônio j e y_j é a saída do neurônio. A presença de não-linearidade é importante

porque, do contrário, a relação de entrada-saída da rede poderia ser reduzida àquela de um *perceptron* de camada única.

2. A rede contém uma ou mais camadas de neurônios ocultos.
3. A rede exibe um alto grau de conectividade.

A combinação dessas três características, juntamente com a habilidade de aprender da experiência através do treinamento, que o *perceptron* de múltiplas camadas deriva seu poder computacional. Quanto ao desenvolvimento do algoritmo de retropropagação, acredita-se que ele representa um marco na utilização das redes neurais, pois fornece um método computacional eficiente para o treinamento de *perceptron*s de múltiplas camadas. A Figura 5 representa o grafo arquitetural de um *perceptron* de múltiplas camadas com duas camadas ocultas:

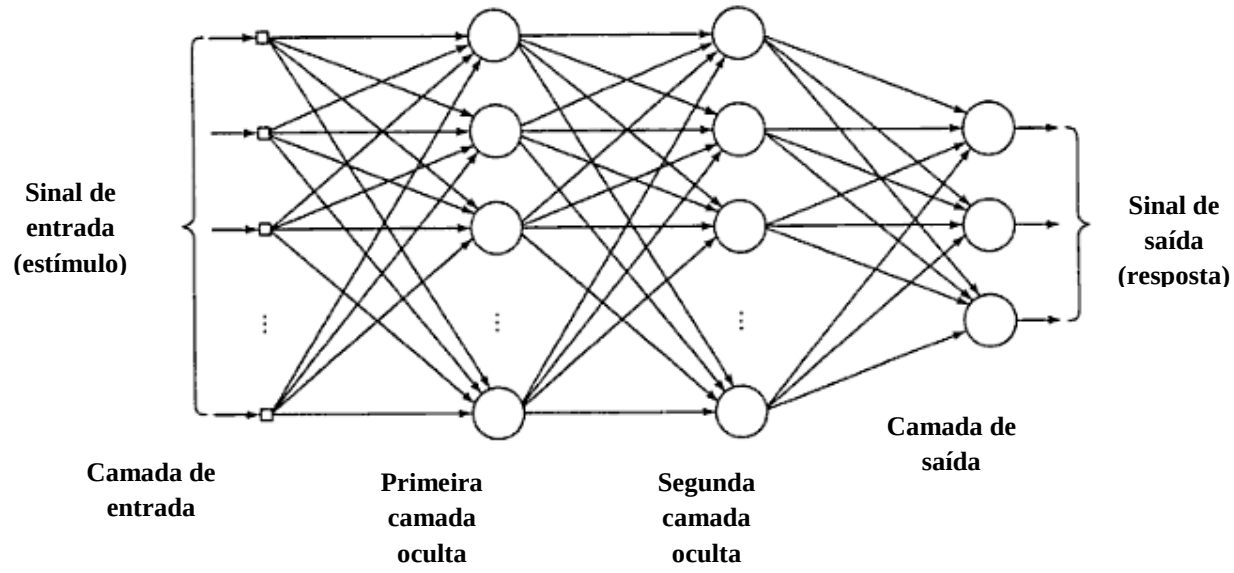

Figura 5 - Grafo arquitetural de um *perceptron* de múltiplas camadas com duas camadas ocultas.
Fonte: HAYKIN, 2001, p. 186.

> Os neurônios de saída (nós computacionais) constituem a camada de saída da rede. Os neurônios restantes (nós computacionais) constituem as camadas ocultas da rede. Assim, as unidades ocultas não são parte da saída ou da entrada da rede. A primeira camada oculta é alimentada pela camada de entrada, constituída de unidades sensoriais (nós de fonte); as saídas resultantes da primeira camada oculta são por sua vez aplicadas à próxima camada oculta; e assim por diante para o resto da rede (HAYKIN, 2001, p. 187).

Seguem as descrições das três arquiteturas de redes neurais selecionadas para este trabalho, de acordo com dados retirados do *software* utilizado para todos os testes.

A modelagem linear é uma aproximação da função discriminante ou da função de regressão, usando um hiperplano. Pode ser otimizada por meio da utilização de técnicas simples, mas não é adequada para modelar a maior parte dos problemas do mundo real.

A *Radial Basis Function* representa um tipo de rede neural com uma camada oculta de unidades radiais e uma camada de saída de unidades lineares, e pode ser caracterizada pela formação de redes razoavelmente rápidas e compactas.

A *Multilayer Perceptron* (MLP) é composta por arquiteturas de redes neurais para problemas de regressão ou classificação, ou combinação de regressão e classificação. As arquiteturas da MLP podem apresentar 1, 2 ou 3 camadas, e são utilizadas em análises que envolvem variáveis preditoras contínuas e/ou categóricas. Em todas as classificações de melhor desempenho do presente estudo, a MLP ficou em primeiro lugar. Todos os testes realizados com redes neurais serão detalhados no Capítulo 4.

As redes foram testadas no presente estudo. Por meio da ferramenta *Intelligent Problem Solver* (das cinco opções disponíveis no item Tipos – s*oftware* Statistica), optou-se pelas três opções a seguir, por acreditar-se que elas representam as três classes de arquiteturas definidas por Haykin: (1) Linear; (2) *Radial Basis Function*; (3) *Three Layer Percepton.*

3.4 ANÁLISE DE *CLUSTERS*

Análise de *clusters* ou análise de agrupamentos é o nome dado a um grupo de técnicas multivariadas, cuja finalidade primária é agregar objetos com base nas características que eles possuem. A ideia é maximizar a homogeneidade de objetos dentro dos grupos, ao mesmo tempo em que se maximiza a heterogeneidade entre os grupos.

A variável estatística de agrupamento é o conjunto de variáveis que representa as mesmas características usadas para comparar objetos, por isso, afirma-se que ela determina o "caráter" dos objetos.

> A análise de agrupamentos é a única técnica multivariada que não estima a variável estatística empiricamente, mas, ao invés disso, usa a variável estatística como especificada pelo pesquisador. O foco da análise de agrupamentos é a comparação de objetos com base na variável estatística, não na estimação da variável estatística em si. (HAIR et al., 2005, p. 384).

O objetivo principal da análise de *clusters* é definir uma estrutura dos dados, colocando as observações mais parecidas em grupos. Para realizar esta tarefa, três questões básicas devem ser abordadas:

> (1) Como medir similaridade? Necessita-se de um método de comparação simultânea de observações sobre as duas variáveis de agrupamento;
>
> (2) Como formar agrupamentos? Não importa como a similaridade é medida, o procedimento deve agregar as observações que são mais similares em um agrupamento;
>
> (3) Quantos grupos devem ser formados? A tarefa fundamental é avaliar a similaridade "média", entre agrupamentos, de forma que a medida que a média aumenta, os agrupamentos se tornam menos parecidos (HAIR et al., 2005, p. 285).

Pohlmann (2007, p. 325) afirma que a análise de *clusters* classifica objetos segundo aquilo que cada elemento tem de similar em relação a outros pertencentes a determinado grupo. O grupo resultante dessa classificação deve, então, exibir um alto grau de homogeneidade interna (dentro do grupo) e alta heterogeneidade externa (entre os grupos). Se o grupo resultante de fato possuir tais características, ou seja, se a classificação for bem-sucedida, quando se visualizar os agrupamentos organizados em um gráfico, os objetos dentro do grupo devem aparecer juntos e os diferentes grupos distantes uns dos outros. O problema que se pretende resolver é: dada uma amostra de *n* objetos (ou indivíduos), cada um deles medido segundo *p* variáveis, procurar um esquema de classificação que os agrupe em *g* grupos.

Ainda segundo o mesmo autor, o objetivo operacional da análise de *clusters* é vincular novas observações a cada um dos grupos formados, dadas certas características que os diferenciam. O "*cluster*ing" é feito com base em similaridades ou distâncias. Dois objetos são considerados semelhantes se seus perfis são próximos, em termos das variáveis utilizadas. A análise de *clusters* é empregada quando se deseja reduzir o número de objetos, agrupando-os de modo que os objetos que fiquem reunidos em um *cluster* sejam mais parecidos entre si do que os pertencentes a outros *clusters*.

> O conceito de similaridade em análise de *clusters* é de vital importância, uma vez que a identificação de agrupamentos de sujeitos ou variáveis só é possível com a adoção de alguma medida de semelhança que permita a comparação objetiva entre os sujeitos. Na análise de conglomerados, as observações são agrupadas segundo algum tipo de métrica de distância, e as variáveis são agrupadas, conforme medidas de correlação ou associação (FÁVERO; BELFIORE; SILVA., 2009, p. 198).

Como no caso do presente trabalho os *clusters* são definidos por grupos (clientes) e não pelas variáveis (dados sobre os clientes), as medidas de distância serão detalhadas com mais rigor.

As medidas de distância são consideradas medidas de dissimilaridade, pois, quanto maiores os valores, menor é a semelhança entre os objetos, e vice-versa. As principais medidas, segundo Fávero, Belfiore e Silva (2009) são:

a) Distância Euclidiana: a distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j) para todas as p variáveis:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \qquad (3.21)$$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \qquad (3.22)$$

em que x_{ik} é o valor da variável k referente à observação i e x_{jk} representa a variável k para a observação j. Nesta abordagem, quanto menor a distância, mais similares serão as observações.

b) Distância Quadrática Euclidiana: a distância entre duas observações (i e j) corresponde à soma dos quadrados das diferenças entre i e j para todas as p variáveis:

$$d_{ij}^{\,2} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \tag{3.23}$$

c) Distância *Minkowski*: a distância euclidiana é um caso particular de uma distância mais geral, chamada de *Minkowski*, dada pela seguinte expressão:

$$d_{ij} = \left(\sum_{k=1}^{p} \left(|x_{ik} - x_{jk}|\right)^n\right)^{\frac{1}{n}}, \tag{3.24}$$

em que d_{ij} é a distância de *Minkowski* entre as observações i e j, p é o número de variáveis, e n=1, 2, ..., ∞.

Se for aplicado n=2 na formulação *Minkowski*, chega-se à distância euclidiana. Entretanto, para n=1, tem-se uma nova distância, denominada *City-Block* ou *Manhattan Distance*, apresentada na sequência.

d) Distância Absoluta, Bloco, *City-Block* ou *Manhattan*: representa a soma das diferenças absolutas entre os valores das p variáveis para os dois casos:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \tag{3.25}$$

e) *Mahalanobis:* a distância estatística entre dois indivíduos, i e j, considerando a matriz de covariância para o cálculo das distâncias:

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}, \tag{3.26}$$

em que S é a estimativa amostral da matriz de variância-covariância Σ dentro dos agrupamentos.

f) *Chebychev*: diferença absoluta máxima entre todas as p variáveis de duas observações:

$$d_{ij} = max_k \left| x_{ik} - x_{jk} \right| \quad (3.27)$$

O pacote estatístico SPSS utiliza como padrão de distância entre observações a distância quadrática euclidiana.

Fávero, Belfiore e Silva (2009, p. 204) explicam: uma vez selecionadas as variáveis do estudo (porcesso detalhado a seguir) e escolhida a medida de similaridade, é necessário determinar o algoritmo que fará o processo de agrupamento, ou seja, a formação dos grupos decorre do critério de distância entre os vetores de dados e do método de agregação escolhido. Basicamente, há dois métodos de agrupamento: o hierárquico e o não hierárquico. Nas técnicas hierárquicas, distinguem-se dois tipos de procedimentos de agrupamento: os métodos aglomerativos e os divisivos.

Segundo os mesmos autores, no método aglomerativo, cada sujeito começa com seu próprio agrupamento e, a partir deste ponto, novos agrupamentos são realizados por similaridade, ou seja, no início cada indivíduo representa um grupo. Na etapa seguinte, os dois indivíduos mais similares (próximos) são agrupados primeiramente e, nas etapas subsequentes, vão se fundindo com os demais grupos de acordo com a proximidade. Assim, em cada etapa, reduz-se o número de agrupamentos em uma unidade. Ao contrário do método aglomerativo, no método divisivo todas as observações começam em um grande agregado, sendo separadas, primeiramente, as observações mais distantes, até que cada observação se torna um grupo isolado.

> Após a formação do primeiro *cluster*, é preciso definir como a distância entre dois *clusters* será computada. Neste aspecto, há diversos métodos para a formação dos agrupamentos, sendo que o que os diferencia, principalmente, é a maneira como as distâncias são calculadas entre os grupos já formados e os que faltam ser agrupados. Os métodos mais frequentes, são: (FÁVERO; BELFIORES; SILVA, 2009)

a) O método da Ligação Individual ou Menor Distância, baseado na distância mínima entre dois grupos de elementos, buscando agrupar inicialmente os objetos separados pela menor distância. Dados dois grupos (i e j) e (k), a distância entre eles é representada pela distância mínima de qualquer ponto de um grupo até qualquer ponto do outro:

$$d_{(ij)k} = min\ \{d_{ik}, d_{jk}\}. \tag{3.28}$$

b) O método da Maior Distância ou Ligação Completa baseia-se na distância máxima, ao contrário do método da ligação individual. Neste método, a distância entre dois grupos é definida como a distância máxima entre todos os pares de possibilidades de observações nos dois grupos. O método busca agrupar elementos cuja distância entre os mais afastados seja a menor. Dados dois grupos (i e j) e (k), a distância entre eles é representada pela distância máxima de qualquer ponto de um grupo até qualquer ponto de outro:

$$d_{(ij)k} = max\ \{d_{ik}, d_{jk}\}. \tag{3.29}$$

Este método tende a formar grupos mais compactos e compostos de indivíduos muito semelhantes entre si.

c) O método da Distância Média ou Ligação Média trata a distância entre dois grupos como sendo a distância média entre todos os pares de indivíduos dos dois grupos, buscando agrupar os agregados cuja distância média seja menor. Há a vantagem, em relação as outras duas técnicas, de não se precisar de valores extremos e de se utilizar todos os elementos do grupo, em vez de um único par de extremos. Dados dois grupos, (i e j) e (k), a distância entre eles é representada da seguinte maneira:

$$d_{(ij)k} = média\ \{d_{ik}, d_{jk}\}. \tag{3.30}$$

d) O método do Centróide, por sua vez, baseia-se na distância (geralmente euclidiana ou quadrática euclidiana), priorizando a menor distância entre eles. Este método

identifica os dois grupos separados pela menor distância entre os pontos mais próximos e os coloca no mesmo agrupamento.

e) Por fim, o método de *Ward* busca agrupar os agregados que apresentam menor soma dos quadrados entre dois agrupamentos, calculada sobre todas as variáveis. Trata-se de um método que tende a proporcionar agregados com aproximadamente o mesmo número de observações.

Segundo Pestana e Gageiro (2003, p. 555) o SPSS fornece dois métodos para formar *clusters*: a análise de *cluster* hierárquica, que se aplica tanto a casos (agrupamentos das observações, ou seja, entre linhas) quanto a variáveis (agrupamento das colunas), e a análise de *cluster* não hierárquica, aplicada somente a casos. Na análise hierárquica os *clusters* formam-se com base nos pares de casos mais próximos de acordo com uma medida de distância escolhida. Quando dois casos são semelhantes, o valor da medida das distâncias é pequeno e o valor da medida das semelhanças é grande, porque enquanto as distâncias medem o afastamento entre dois casos, as semelhanças medem quão perto estão esses casos entre si. O método é designado hierárquico porque uma vez estando dois casos juntos, eles permanecem assim até o fim das etapas.

Pohlmann (2007, p. 345) esclarece que uma importante característica dos procedimentos hierárquicos é que os resultados de um estágio anterior são sempre incluídos dentro dos resultados dos estágios seguintes, de forma similar a uma árvore. Os cinco algoritmos aglomerativos mais populares usados para desenvolver agrupamentos hierárquicos são: (1) *single linkage (nearest neighbor)*; (2) *complete linkage* (*furthest neighbor)*; (3) *average linkage (between-groups linkage e within-groups linkage)*; (4) *Ward's method*; e (5) *Centroid method*. O primeiro algoritmo encontra os dois objetos separados pela menor distância e os coloca no primeiro grupo. O primeiro grupo é formado pelos dois elementos que possuírem a menor distância entre eles. Então, a próxima menor distância é encontrada e o terceiro objeto é reunido com os dois primeiros para formar um grupo ou um novo grupo de dois membros é formado. O processo continua até que todos os objetos estejam em um grupo.

O segundo procedimento é similar ao anterior, exceto pelo fato de o critério de agrupamento ser baseado na distância máxima. Busca-se agrupar elementos cuja distância entre os mais afastados seja a menor. O terceiro método se inicia da mesma forma que os

demais, mas o critério de agrupamento é a distância média entre todos os pares de indivíduos de dois grupos, buscando-se agrupar os objetos com menor distância média. O método de Ward baseia-se na perda de informação decorrente do agrupamento de objetos em conglomerados, medida pela soma total dos quadrados dos desvios de cada objeto em relação à média do conglomerado no qual o objeto foi inserido. A cada estágio de agrupamento, a soma dos quadrados dos desvios das variáveis em relação a cada objeto é minimizada. No quinto e último algoritmo, a distância entre os grupos é a distância entre seus centróides, ou seja, entre os valores médios das observações sobre as variáveis, priorizando a menor distância.

Agora, tratando de procedimentos não hierárquicos de agrupamento, em contraste com os métodos hierárquicos, Pohlmann (2007, p. 348) afirma que os procedimentos não hierárquicos não envolvem a construção de um processo tipo “árvore”, os resultados são menos suscetíveis a dados suspeitos, à medida de distância usada e à inclusão de variáveis irrelevantes ou inapropriadas. Esses benefícios são obtidos, entretanto, somente com o uso de grupos-sementes escolhidos de forma não aleatória, ou seja, predeterminados especificamente. O principal problema enfrentado pelos procedimentos não hierárquicos diz respeito à seleção dos grupos-sementes.

Ainda de acordo com o mesmo autor, os métodos não hierárquicos são também conhecidos como métodos de partição. Esses métodos procuram diretamente uma partição de n objetos, de modo que satisfaçam às duas premissas básicas: semelhança interna e separação dos grupos. Portanto, eles exigem a prefixação de critérios que produzam medidas sobre qualidade da partição produzida. Os procedimentos não hierárquicos são frequentemente referidos em *K-Means* e usam uma das seguintes abordagens: *sequential threshold*, *parallel threshold* e *optimization.*

A primeira abordagem se inicia pela seleção de um grupo-semente e inclui todos os objetos dentro de uma distância preestabelecida. Quando todos os objetos dentro dessa distância são incluídos, um segundo grupo-semente é selecionado, e todos os objetos dentro da distância preestabelecida são incluídos. Então, um terceiro grupo-semente é selecionado e o processo continua como antes. Quando um objeto é destinado a um grupo-semente, ele não é mais considerado nos grupos origens subsequentes. Pelo método *parallel threshold* vários grupos são selecionados, simultaneamente, no início, e os objetos são distribuídos entre eles,

dentro de uma distância inicial em relação ao grupo-semente mais próximo. O terceiro método é similar aos dois anteriores, exceto pelo fato de permitir a realocação de objetos.

Fávero, Belfiore e Silva (2009, p. 218) afirmam que no método hierárquico o algoritmo estabelece uma relação de hierarquia entre os sujeitos e os grupos. Este fato não ocorre no método não hierárquico, pois, uma vez especificado o número de agrupamentos, o processo é dinâmico e interativo, tendo como objetivo identificar a melhor solução. Os procedimentos não hierárquicos são utilizados para agrupar indivíduos cujo número inicial de *clusters* é definido pelo pesquisador. A probabilidade de acontecerem classificações erradas nos agrupamentos é menor nos métodos não hierárquicos, mas, em contrapartida, há a dificuldade de se estabelecer o número de *clusters* de partida.

Os procedimentos hierárquicos são mais rápidos e, por isso, levam menos tempo para processar os dados do que os procedimentos não hierárquicos. No entanto, como não se realoca combinações anteriores indesejáveis no procedimento hierárquico, os resultados devem ser muito bem avaliados para que as conclusões acerca dos fenômenos estudados não sejam artificiais ou mesmo equivocadas. Os resultados das análises devem ser comparados com as expectativas iniciais do pesquisador, uma vez que grandes variações de tamanho dos grupos ou ainda grupos com um ou com poucos objetos podem indicar a presença de pontos extremos na base de dados.

> No método não hierárquico, pode ser elaborado um comparativo entre a utilização de sementes aleatórias, como resultados obtidos, com o uso de sementes especificadas e, caso haja consistência nos resultados, poder-se-á afirmar com maior segurança sobre a validade da análise, afirmam (FÁVERO; BELFIORE; SILVA, 2009, p. 225).

Outro procedimento não hierárquico é o *TwoStep Cluster* (TSC). De acordo com Brazão et al. (2007) entre as vantagens do TSC estão a possibilidade do uso de variáveis quantitativas e categóricas, além da determinação de um número apropriado de agrupamentos, caso este número não seja fornecido previamente à execução do algoritmo. Não se pode contar com essa possibilidade no método *K-Means.*

Ainda segundo os mesmos autores, a primeira etapa do algoritmo TSC consiste em criar uma árvore, gerando uma coleção de pré-agrupamentos, que é armazenada nos nós folhas da árvore. Na segunda etapa, é utilizado um algoritmo hierárquico aglomerativo para se encontrar os agrupamentos finais. O TSC utiliza a distância log da verossimilhança, acomodando adequadamente variáveis quantitativas e categóricas.

Pela pesquisa de Brazão et al. (2007), comparando o algoritmo *TwoStep Cluster* com outros algoritmos de agrupamento para grandes bases de dados, constatou-se que o TSC teve melhor acurácia quando os grupos tinham diferentes variâncias. Quanto ao tempo de processamento, o gasto é bem maior com o TSC do que com o *K-Means*, entretanto, o percentual de aumento do tempo quando se aumenta o número de registros é menor no TSC do que no *K-Means*. Para eles, a vantagem do TSC aumenta quando se tem poucas variáveis e estas têm variâncias bem diferentes nos agrupamentos.

Neste trabalho os dados foram segmentados de duas maneiras. Na primeira, os clientes foram agrupados com base em uma variável da própria amostra, ou seja, foram formados grupos de clientes com base na região à qual a filial de venda pertencia. Assim, o grupo 1 é formado por todos os casos nos quais as vendas tiverem sido realizadas em filiais que fazem parte da região 1, o grupo 2 será formado por todos os casos nos quais as vendas tiverem sido realizadas em filiais que fazem parte da região 2 e, assim, sucessivamente. Na segunda, os dados foram agrupados por meio da análise de *clusters* não hierárquica, para os métodos *K-Means e TwoStep Cluster*.

Independentemente da forma como os agrupamentos tenham sido formados, o objetivo da segmentação dos dados sempre foi o de maximizar a homogeneidade de objetos dentro dos grupos, ao mesmo tempo em que se maximiza a heterogeneidade entre os grupos, fato que possibilita uma análise mais eficiente a partir do modelo adotado.

4. CARACTERIZAÇÃO DA AMOSTRA E PASSO A PASSO DA PESQUISA

4.1 ASPECTOS GERAIS

Pestana e Gageiro (2003, p. 556) asseguram que a seleção de variáveis que serão incluídas em um modelo é crucial, uma vez que resultados fracos ou enganadores podem provocar a exclusão de variáveis importantes. Pohlmann (2007, p. 330) complementa a afirmação dos autores, ao lembrar que a seleção das variáveis a serem incluídas na análise deve ser feita, tendo em vista, aspectos tanto teóricos e conceituais, quanto práticos.

Com base na utilidade e nas características das técnicas acima, bem como na importância da escolha das variáveis que farão parte dos modelos propostos, a amostra utilizada neste estudo e o passo a passo utilizado para analisar esta amostra serão apresentadas nos próximos itens.

4.2 CARACTERIZAÇÃO DA AMOSTRA

A base de dados utilizada para realizar esta pesquisa foi cedida por uma importante empresa do setor varejista brasileiro, para que fossem realizadas simulações de análise discriminante e redes neurais e simulações de tais técnicas combinadas com a segmentação de dados, manual e estatística, na tentativa de obter *insights* sobre inadimplência. A amostra estudada conta com 50 mil clientes da base de dados da empresa, que realizaram compras a prazo, entre os meses de Novembro de 2.006 a Janeiro de 2.007.

Como variáveis de entrada foram selecionadas características dos clientes, que pudessem ajudar na caracterização e, posterior, previsão do comportamento dos mesmos. Os atributos podem ser numéricos ou não, sendo que variáveis como valor financiado e quantidade de prestações são variáveis numéricas, ao passo que estado civil e sexo são variáveis nominais. As variáveis foram selecionadas em função da disponibilidade na base de dados da empresa e de eventuais critérios de confidencialidade que podem restringir a evidenciação de algumas informações. A variável de saída que se pretende prever possui

característica binária adimplente/inadimplente, assumindo valor 0 (zero) para clientes adimplentes e valor 1 (um) para clientes inadimplentes.

As variáveis independentes que fazem parte do modelo proposto são:

1. **Filial** – filial na qual a venda foi realizada;
2. **Região** – região à qual a filial de venda pertence;
3. **Data da venda**— data da realização da venda pela loja (apenas o dia do mês);
4. **Estado civil** — estado civil informado pelo cliente no momento do seu cadastro;
5. **Sexo** — masculino ou feminino;
6. **Tipo de telefone comercial** – tipo de telefone comercial que o cliente possui;
7. **Tipo de telefone residencial** – tipo de telefone residencial que o cliente possui;
8. **CEP** — CEP da residência do cliente;
9. **Profissão** – profissão informada pelo cliente no momento do seu cadastro;
10. **Idade** — idade do cliente em anos;
11. **Renda** — renda do cliente no momento do seu cadastro;
12. **SPC** – *score* estimado por meio de um modelo próprio da empresa, que utiliza variáveis disponibilizadas pelo SPC[5];
13. **Valor financiado** – valor da compra a prazo;
14. **Quantidade de prestações** – número de vezes em que o valor financiado foi dividido;
15. **Valor da prestação** — valor da prestação a ser paga pelo cliente;
16. **Valor de entrada –** valor da prestação paga no ato da compra, se houver;
17. **Prestações a vencer –** quantidade de prestações que ainda não foram pagas pelo cliente;
18. **Máxima prestação a vencer –** qual foi o número máximo de prestações a serem pagas pelo cliente;
19. **Número de compras –** quantidade de compras liquidadas pelo cliente dentro do período de Novembro de 2006 a Janeiro de 2007;

[5] Não serão divulgados os critérios adotados para calcular o *score* SPC e as variáveis utilizadas neste cálculo, em razão da não autorização da empresa cedente do banco de dados.

20. **Maior atraso –** qual o maior atraso que o cliente já teve para efetuar o pagamento de suas compras a prazo (em dias);
21. **Prestação / renda** – proporção do valor da prestação em relação à renda mensal do cliente.

As variáveis de 1 a 20 foram fornecidas pela empresa e a variável 21, Prestação/ Renda, que representa quanto a parcela mensal consome da renda mensal do cliente, foi criada para esta pesquisa por se tratar de um indicador importante de inadimplência. Acredita-se que quanto maior for o comprometimento da renda, maiores serão as chances de que o cliente se torne inadimplente em momentos adversos, como por exemplo, durante a crise financeira de 2008/ 2009.

As variáveis filial (1), região (2) e CEP (8) foram incluídas no modelo por acreditar-se que o perfil dos clientes varia de região para região, ou seja, o perfil de um cliente adimplente para uma determinada filial, que faz parte de uma região x, provavelmente será diferente do perfil do cliente adimplente de uma filial que pertence à região y.

A variável data da venda (3) faz parte do perfil de adimplência de um cliente porque, por exemplo, se um cliente receber seu pagamento no começo do mês e tiver uma conta para pagar no final do mês, pode ser que, por falta de planejamento, não sobre dinheiro para quitar esta dívida, que vai se acumulando.

O estado civil (4), idade (10), bem como o sexo (5) do cliente podem interferir no seu perfil de adimplência, conforme será detalhado no Capítulo 4, uma vez que clientes casados, mais velhos e do sexo feminino tendem a ser melhor pagadores do que clientes solteiros, mais jovens e do sexo masculino, por exemplo.

As variáveis tipo de telefone comercial (6) e residencial (7) fazem parte do modelo por acreditar-se que possuir telefone comercial e residencial pode indicar maior poder aquisitivo e, portanto, maior capacidade de cumprir com as dívidas assumidas. A mesma justificativa vale para as variáveis profissão (9) e renda (11), já que clientes com profissões mais estáveis e melhor remuneradas, provavelmente, apresentam maior capacidade de cumprir com suas obrigações.

O *score* SPC (12) representa uma pontuação dada ao cliente, de acordo com o risco de inadimplência que ele apresentar. Sua contribuição para o modelo proposto é bastante relevante, conforme apresentado no próximo capítulo.

As características da compra, como valor financiado pelo cliente (13), quantidade de prestações assumidas (14) e valor de cada prestação (15) impactam no seu perfil da seguinte maneira: dependendo da profissão, portanto, da renda, da data da compra, entre outros atributos apresentados por ele, a probabilidade de que a dívida assumida seja paga é, provavelmente, maior para uma compra com valor total financiado menor, com prestações de menor valor, bem como com menos parcelas.

Um cliente que der um valor de entrada (16) pode apresentar maiores condições de pagar as próximas parcelas do que um cliente que não der entrada.

As variáveis prestações a vencer (17), máxima prestação a vencer (18), número de compras (19) e maior atraso (20) indicam o comportamento do cliente em relação às compras que ele fez anteriormente. É muito provável que, em condições semelhantes, este comportamento se repita nas próximas compras.

> Uma solicitação de crédito de um cliente antigo, com o qual a empresa tenha realizado boas relações comerciais, é atendido quase prontamente. É evidente que, se o pedido exceder os limites convencionalmente concedidos, ou a experiência passada tiver sido marcada por problemas de atrasos relevantes ou perdas, ou, ainda, se a conjuntura econômica atual tiver se alterado significativamente em relação a seu comportamento passado, a decisão de crédito deverá exigir outras providências da administração da empresa. Nessa situação, o pedido de crédito é normalmente considerado como se tratasse de uma solicitação inicial, devendo a empresa desencadear o processo de avaliação como se o cliente tivesse entrado pela primeira vez com um pedido (ASSAF NETO, 2009, p. 555).

A escolha das variáveis que compõem o modelo proposto por este trabalho se deu tanto pela contribuição de cada uma delas, como se mostrou nos parágrafos anteriores, quanto por se acreditar que sejam essas as variáveis utilizadas por grande parte das empresas do setor varejista, uma vez que as únicas informações que um vendedor possui são os dados cadastrais

do cliente novo ou dados sobre o comportamento de clientes antigos, e é com base nesses dados que o sistema terá de fornecer informações úteis para a tomada de decisão de concessão do crédito.

> ... a importância da adoção de instrumentais estatísticos pode ser observada em empresas com grande volume de solicitações de crédito, as quais, necessitando descentralizar seu processo de decisão, conferem maiores responsabilidades a escalões inferiores. (ASSAF NETO, 2009, p. 556).

> A empresa deve fixar seus padrões de crédito, ou seja, os requisitos de segurança mínimos que devem ser atendidos pelos seus clientes para que se conceda o crédito, bem como os principais parâmetros de risco que estaria disposta a assumir. O estabelecimento dessas exigências mínimas envolve, geralmente, o agrupamento dos clientes em diversas categorias de risco, as quais visam, normalmente mediante o uso de probabilidades, mensurar o custo das perdas associadas às vendas realizadas a um ou vários clientes de características semelhantes. Assim, para cada classe ou categoria de clientes tem-se um custo (probabilidade) de perdas pelo não recebimento das vendas efetuadas a prazo (ASSAF NETO, 2009, p. 557).

Vasconcellos (2002, p. 7) propõe uma metodologia para análise de concessão de crédito a pessoas físicas a partir do estudo matemático e estatístico de informações sobre créditos concedidos, no passado recente da carteira de crédito em questão. As variáveis utilizadas no modelo proposto foram profissão, valor dos cheques devolvidos pelo motivo 12, saldo médio em conta corrente, saldo médio em aplicações, estado civil, tempo de residência, quantidade de prestações da operação, valor da operação de empréstimo, bloqueio na emissão de cheques, idade da admissão dividida pela idade atual, salário líquido, residência própria e quantidade de dependentes financeiros.

Gonçalves (2005, p. 8) afirma que seu trabalho busca ilustrar os procedimentos a serem adotados por uma empresa para identificar o melhor modelo de concessão de crédito que tenha boa aderência aos seus dados. Segundo ele, a adoção do melhor modelo detectado permite o direcionamento da estratégia da instituição, podendo aumentar a eficiência do seu negócio. As variáveis utilizadas neste estudo foram sexo, estado civil, fone residencial, fone comercial, tempo no emprego atual, salário do cliente, quantidade de parcelas a serem

quitadas, primeira aquisição, tempo na residência atual, valor da parcela, valor total do empréstimo, tipo do crédito, idade, CEP residencial, CEP comercial, código da profissão, nome da profissão, salário do cônjuge e tipo de cliente - bom (máximo 20 dias de atraso) ou mau (acima de 60 dias de atraso).

O trabalho de Karcher (2009, p. 7) compara a técnica proposta em *credit scoring*, Redes Baynesianas, com os resultados obtidos por meio da Regressão Logística. A análise dos modelos ajustados mostrou que as Redes Baynesianas e a Regressão Logística apresentaram desempenho similar, em relação à estatística *Kolmogorov-Smirnov* e ao coeficiente Gini. O classificador *Tree Augmented Naive Bayes* foi escolhido como o melhor modelo, pois apresentou o melhor desempenho nas previsões dos clientes “maus” pagadores e permitiu uma análise dos efeitos de interação entre variáveis. As variáveis utilizadas neste estudo foram bens, salário, poupança do cliente, outros empréstimos, outras dívidas ou garantias, finalidade, histórico do crédito, tempo de trabalho, estado civil, sexo, moradia, emprego, telefone próprio, estrangeiro, duração do empréstimo, valor do empréstimo, taxa de juros em % do valor do empréstimo, tempo de residência, idade, número de dependentes, número de créditos concedidos em seu banco.

Nota-se, por meio das informações extraídas dos trabalhos acima, que grande parte das variáveis utilizadas neste trabalho também são utilizadas em outras pesquisas, fato que valida a escolha dos atributos utilizados para testar o modelo proposto.

A seguir será apresentado o passo a passo adotado para a realização do presente estudo.

4.3 PASSO A PASSO

O primeiro passo para a elaboração deste trabalho foi rodar a análise descritiva da amostra global, de modo a traçar um perfil para os 50 mil clientes que a compõe. As variáveis dicotômicas foram incluídas em todas as análises. A base final ficou composta por 40.178 (80%) clientes adimplentes e 9.822 (20%) inadimplentes.

Posteriormente, foi rodada a análise discriminante para a amostra global, tanto para se conhecer as variáveis mais importantes que discriminam os grupos, quanto para se verificar o resultado da classificação dos clientes como adimplentes ou como inadimplentes.

O próximo passo foi rodar as redes neurais, também para a amostra global, considerando que antes que uma rede neural seja utilizada para qualquer tarefa ela deverá ser treinada para isto. Os pesos sinápticos aplicados a cada sinal de entrada de uma rede são seus principais elementos e, por isso, o treinamento dessa rede se torna muito importante, uma vez que tais pesos serão determinados neste momento. O conhecimento adquirido por uma rede neural é utilizado para determinar os pesos sinápticos e os vieses da fase de teste, que representa a fase de avaliação da capacidade de generalização da rede (ZHANG et al., 1997, p. 38).

No presente trabalho, para todas as redes rodadas, foram utilizados 50% dos dados na fase de treinamento, 25% na fase de validação e 25% na fase de teste.

O número de nós de entrada pode ser definido pela quantidade de variáveis independentes associadas com o problema, para um problema casual de previsão. Determinar a arquitetura da rede, ou seja, determinar o número de camadas, o número de nós em cada camada e o número de arcos que interconectam os nós, bem como determinar o algoritmo, tratamento dos dados e medidas de desempenho, são decisões muito importantes (ZHANG et al., 1997, p. 38).

Neste trabalho, utilizou-se a ferramenta *Intelligent Problem Solver* para escolher qual a melhor rede, entre os tipos Linear, *Radial Basis Function* e *Three layer perceptron*. Em todos os casos a rede *Three layer perceptron,* considerada como uma *Multilayer Perceptron*, foi a que apresentou melhor desempenho. O algoritmo utilizado foi o de retropropagação, já explicado anteriormente. A maior parte dos trabalhos que utilizam redes neurais, já apresentados no Capítulo 2, aplica a rede *Multilayer Perceptron* (*Perceptrons* de Múltiplas Camadas).

Para a segunda parte deste trabalho, composta pela segmentação dos dados para posterior aplicação da análise discriminante e de redes neurais nos grupos formados por esta

segmentação, utilizou-se três tipos diferentes de agrupamento, sendo que dois deles foram realizados manualmente e o terceiro, por meio de ferramentas estatísticas.

No primeiro tipo de agrupamento de dados utilizado, a amostra global foi segmentada em 21 grupos, de acordo com a região à qual a filial de cada venda pertencia. A relação das 21 micro-regiões foi fornecida pela própria empresa cedente do banco de dados, que também informou que o critério adotado para formá-las foi a posição geográfica das filiais[6].

Segue mapa do Brasil com divisão das micro-regiões:

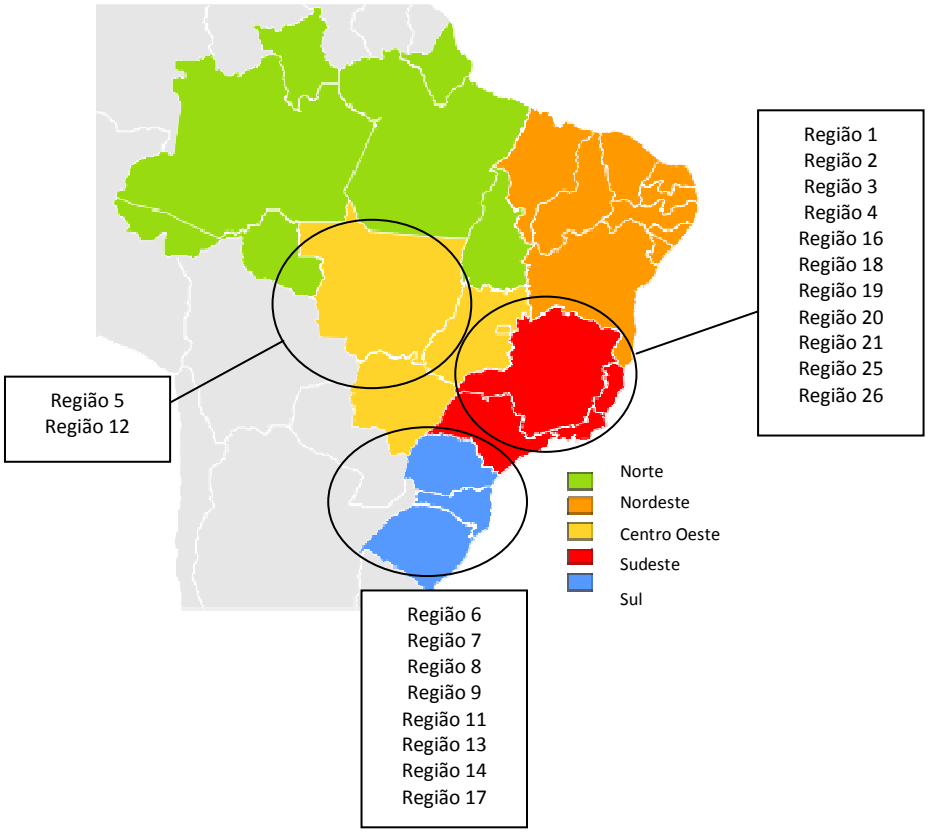

Figura 6 - Mapa das micro-regiões

A numeração das filiais vai de 1 a 26 por se tratar da classificação feita pela empresa. No entanto, como o período de vendas analisado foi de Novembro de 2006 a Janeiro de 2007, não

[6] Não serão divulgados os nomes das regiões em razão da não autorização da empresa cedente do banco de dados.

ocorreram vendas em algumas destas regiões e, portanto, elas não entraram na análise. A saber: regiões 10, 15, 22, 23 e 24.

Para o segundo tipo de agrupamento de dados, a amostra global foi segmentada em 3 macro-regiões, de acordo com a região à qual a filial de cada venda pertencia.

Segue mapa do Brasil com divisão das macro-regiões:

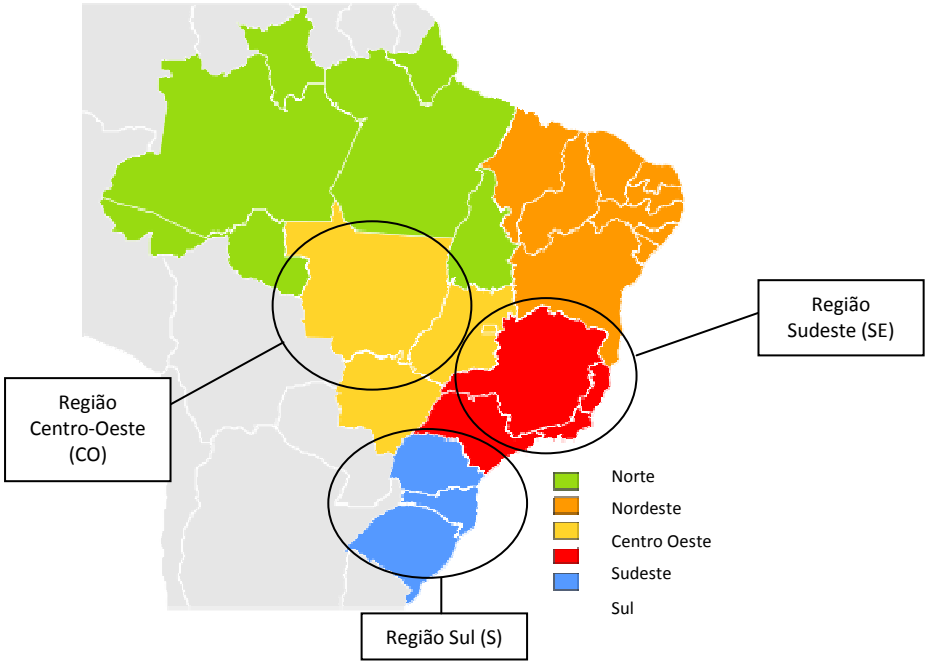

Figura 7 - Mapa das macro-regiões

As regiões Norte (N) e Nordeste (NE) não fazem parte da análise porque não existem filiais da empresa nestas regiões.

No terceiro tipo de agrupamento de dados utilizado, a amostra global foi segmentada por meio da análise de *clusters K-Means* e por meio da análise de *clusters TwoStep Cluster*.

O procedimento utilizado foi o não hierárquico, uma vez que se trata de uma base de dados com elevado número de observações. Tanto a técnica *K-Means*, quanto a técnica *TwoStep Cluster*, foram rodadas para verificar qual das duas apresenta melhores resultados para o fenômeno estudado, de acordo com os trabalhos existentes na literatura.

A utilização da segmentação de dados neste trabalho pode ser justificada por meio de dois objetivos, que se busca alcançar: (1) redução da variância entre os dados, de modo a tornar o modelo proposto mais eficiente (quando da combinação de técnicas) e (2) possibilidade de aplicação do modelo proposto não só para a empresa que forneceu o banco de dados para este estudo, como também para qualquer outra empresa do setor varejista (que possua características estruturais semelhantes a ela); uma vez que ao dividir a amostra por regiões, vieses foram eliminados, ou seja, os grupos tornaram-se mais homogêneos, no que tange ao comportamento e aos atributos apresentados pelos clientes, permitindo a generalização do perfil dos clientes para cada região, independentemente da empresa na qual a venda for realizada.

Como cada região apresenta suas particularidades em termos sociais e econômicos, como por exemplo, profissão e renda mensal, a capacidade de generalização está em considerar esta informação na aplicação do modelo proposto.

O último passo foi rodar a análise discriminante e as redes neurais, separadamente, para cada um dos grupos formados, para cada uma das técnicas de segmentação utilizada, e somar os resultados da classificação dos clientes de cada grupo. Os resultados das somas para cada técnica foram comparados tanto entre si, quanto para as técnicas aplicadas, separadamente, para a amostra global.

A seguir, será apresentada a aplicação prática desta pesquisa, na qual estão detalhados os resultados deste passo a passo.

5. APLICAÇÃO PRÁTICA

5.1 ANÁLISE DESCRITIVA DA AMOSTRA GLOBAL

O primeiro passo da aplicação prática deste trabalho foi analisar a composição descritiva da amostra empregada, de modo a traçar um perfil para os 50 mil clientes que a compõe e, consequentemente, conhecer melhor a qualidade das informações da amostra.

Na tabela 3, mostram-se, comparativamente, a composição de categorias de indivíduos da amostra global, segregada pelo *status* adimplente/ inadimplente.

Tabela 3- Composição da amostra global

		Adimplentes	Inadimplentes
Profissão	**Serviços**	34%	37%
	Outros	35%	35%
	Autônomo	7%	7%
	Aposentado	12%	8%
	Funcionário Público	1%	1%
	Comércio	11%	12%
	Total	**100%**	**100%**
Sexo	**Feminino**	63%	52%
	Masculino	37%	48%
	Total	**100%**	**100%**
Tipo Telefone Comercial	**Fixo**	45%	42%
	Celular	5%	6%
	NI	50%	52%
	Total	**100%**	**100%**
Tipo Telefone Residencial	**Fixo**	65%	53%
	Celular	15%	20%
	NI	20%	27%
	Total	**100%**	**100%**
Estado Civil	**Casado**	59%	47%
	Solteiro	27%	38%
	Divorciado	4%	5%
	Viúvo	6%	5%
	Outros	4%	5%
	Total	**100%**	**100%**

Serviços, outros, autônomo, aposentado, funcionário público e comércio representam as seis categorias de profissão. A maior parte dos clientes pertence às categorias serviços e

outros. Dos 40.178 clientes adimplentes que compõem a base de dados estudada, cerca de 70% estão nessas categorias. A mesma proporção, cerca de 70%, é verificada entre os 9.822 clientes inadimplentes. O restante, cerca de 30%, divide-se entre as categorias autônomo, aposentado, funcionário público ou comércio.

A maior parte dos clientes é do sexo feminino, tanto para o grupo de clientes adimplentes, quanto para o grupo de clientes inadimplentes. Dos 40.178 clientes adimplentes que compõem a base de dados estudada, 63% são do sexo feminino e 37% do sexo masculino. Já dos 9.822 clientes inadimplentes, 52% são do sexo feminino e 48% do sexo masculino.

As três categorias de tipo de telefone comercial são: fixo, celular e não informado (NI). A maior parte dos clientes não informou o tipo de telefone que possui, cerca de 50% tanto para clientes adimplentes, quanto para clientes inadimplentes. Dos clientes que responderam, a maioria possui telefone fixo, tanto para o grupo de clientes adimplentes, quanto para o grupo de clientes inadimplentes. Dos 40.178 clientes adimplentes que compõem a base de dados estudada, 45% possui telefone fixo e 5% possui celular. Já dos 9.822 clientes inadimplentes, 42% possui telefone fixo e 6% celular.

As categorias do tipo de telefone residencial são as mesmas categorias do tipo de telefone comercial. A maior parte dos clientes possui telefone residencial fixo, tanto para clientes adimplentes, quanto para clientes inadimplentes. Dos 40.178 clientes adimplentes que compõem a base de dados estudada, 65% possui telefone fixo e 15% possui celular. Já dos 9.822 clientes inadimplentes, 53% possui telefone fixo e 20% celular.

Os clientes foram classificados em cinco categorias de estado civil: casado, solteiro, divorciado, viúvo e outros. A maior parte dos clientes é casada, tanto para o grupo de clientes adimplentes, quanto para o grupo de clientes inadimplentes. Dos 40.178 clientes adimplentes que compõem a base de dados estudada, 59% são casados, 27% são solteiros e os outros 14% são divorciados, viúvos ou outros. Já dos 9.822 clientes inadimplentes, 47% são casados, 38% são solteiros e os 15% restantes são divorciados, viúvos ou outros.

Assim, traçando um perfil dos clientes, tanto adimplentes quanto inadimplentes, a maioria deles é do sexo feminino, casado, com telefone comercial e residencial fixo e trabalha com serviços ou outra profissão que não o comércio, o funcionalismo público, como

autônomo ou aposentado. Além disso, no geral, os clientes adimplentes são mais velhos, possuem maior renda mensal, financiam valores menores, em quantidade menor de prestações, com parcelas e entrada de maior valor e já liquidaram mais compras do que os clientes inadimplentes. Ademais, a quantidade de prestações a vencer, bem como a quantidade de prestações já assumidas são menores para clientes adimplentes. Os indicadores SPC e parcela sobre a renda são muito próximos para os dois grupos.

O próximo passo será rodar a análise discriminante para a amostra global.

5.2 ANÁLISE DISCRIMINANTE DA AMOSTRA GLOBAL

Pestana e Gageiro (2003, p. 655) sustentam a afirmação de que análise discriminante é uma técnica multivariada aplicada quando a variável dependente é qualitativa e as variáveis independentes são quantitativas. As variáveis dicotômicas como sexo, profissão e tipo de telefone podem ser incluídas no modelo como variáveis explicativas (independentes). Este procedimento tem por objetivo escolher as variáveis que distinguem os grupos, de modo que se conhecendo as características de um novo caso se possa prever a que grupo ele pertence. No caso do presente trabalho, trata-se dos grupos adimplente e inadimplente.

As variáveis sexo, estado civil, tipo de telefone residencial, tipo de telefone comercial e profissão representam as variáveis dicotômicas que foram incluídas como variáveis explicativas no modelo proposto por este trabalho. Para sexo, determinou-se: 1 para feminino e 2 para masculino; para estado civil, determinou-se 1 para casado, 2 para solteiro, 3 para divorciado, 4 para viúvo e 5 para outros; para tipo de telefone, determinou-se 1 para fixo, 2 para celular e 3 para não informado – NI - (tanto para telefone residencial quanto para telefone comercial); para profissão determinou-se: 1 para solteiro, 2 para outros, 3 para autônomo, 4 para aposentado, 5 para funcionário público e 6 para comércio.

Foram utilizados os dados de 49.956 clientes (os outros 44 clientes foram considerados como *missing values*[7]) para rodar a análise discriminante.

[7] Os *missing values* não foram tratados por representarem uma porcentagem muito pequena da base de dados (0,088%).

A Tabela 4 apresenta a média e o desvio padrão de cada variável da amostra global, bem como os testes de igualdade das médias dos grupos da amostra global:

Tabela 4 – Estatística descritiva e testes de igualdade das médias da amostra global

	Adimplentes (40.156)		Inadimplentes (9.800)		Amostra Global		
Variáveis	**Média**	**Desvio Padrão**	**Média**	**Desvio Padrão**	**Lambda de Wilks**	**Estatística F**	**Sig**
Filial	-	-	-	-	0,998	109,397	0,000
Região	-	-	-	-	0,999	74,252	0,000
Data da Venda	-	-	-	-	1,000	5,146	0,023
Estado Civil	-	-	-	-	0,997	136,867	0,000
Sexo	-	-	-	-	0,992	393,951	0,000
Tipo Fone Resedencial	-	-	-	-	0,991	429,772	0,000
Tipo Fone Comercial	-	-	-	-	1,000	17,206	0,000
Profissão	-	-	-	-	0,999	32,678	0,000
CEP	-	-	-	-	1,000	22,796	0,000
Idade	42	14	36	14	0,971	1500,000	0,000
Renda mensal	1.093,49	862,25	959,78	748,87	0,996	199,029	0,000
SPC	0,404388	0,273194	0,407917	0,238961	1,000	1,378	**0,240**
Valor da compra	276,91	324,61	379,55	381,32	0,986	732,883	0,000
Quantidade de parcelas	8	5	11	5	0,932	3644,000	0,000
Valor das parcelas	54,91	91,34	46,19	44,38	0,998	84,344	0,000
Valor da entrada	25,75	83,31	20,96	63,27	0,999	28,300	0,000
Prestações a Pagar	15	20	24	25	0,971	1496,000	0,000
Maior n° prestações a pagar	7	6	11	6	0,957	2253,000	0,000
Compras liquidadas	10	18	5	12	0,986	709,866	0,000
Maior atraso	26	67	35	94	0,998	116,876	0,000
Parcela sobre a renda	6,69%	35,78%	6,41%	16,99%	1,000	0,570	**0,450**

A idade média dos clientes inadimplentes é 6 anos mais baixa do que a idade média dos clientes adimplentes. Assim, podemos afirmar que clientes mais jovens apresentam maior tendência de não honrar com as obrigações financeiras assumidas. Para chegar-se a esta conclusão basta comparar a idade média do grupo de clientes adimplentes e a idade média do grupo de clientes inadimplentes com a idade média do grupo todo. A média dos clientes adimplentes sempre estará acima da média do grupo todo, bem como a média dos clientes inadimplentes sempre estará abaixo dessa média.

A renda média dos clientes inadimplentes é menor, logo, clientes com renda média mensal acima da renda média do grupo tendem a honrar com seus compromissos, ao passo

que pessoas com renda média mensal abaixo da média do grupo terão mais dificuldades de fazê-lo.

O *score* médio é muito próximo entre clientes adimplentes e inadimplentes, mas mesmo assim, apresenta-se maior para clientes inadimplentes. Essas informações indicam que, de acordo com a metodologia de avaliação de risco de inadimplência utilizada pela empresa, que é baseada em variáveis fornecidas pelo SPC, clientes inadimplentes apresentam *scores* maiores que o *score* médio do grupo.

O valor médio financiado pelos clientes inadimplentes é maior, assim, conclui-se que clientes que buscam volumes maiores de crédito (em relação ao volume médio) tendem a não pagar, pelo menos parte de sua dívida.

A quantidade média de parcelas é maior para clientes inadimplentes. Assim, é possível concluir que clientes que precisam dividir sua compra em um maior número de prestações do que o número de prestações médio do grupo, possivelmente, terão dificuldades de pagamento.

O valor médio da parcela é menor para clientes inadimplentes. Logo, os clientes que pagam parcelas menores que a parcela média do grupo apresentarão a tendência de não cumprir com suas obrigações.

O valor médio de entrada é menor para clientes inadimplentes. Assim, clientes que derem uma entrada de valor menor do que o valor médio de entrada do grupo, possivelmente, terão problemas com o pagamento da compra.

A quantidade média de prestações a pagar é maior para clientes inadimplentes. Fica claro, então, que clientes que apresentam maior número de prestações em aberto do que a quantidade média de prestações em aberto do grupo contam com maior dificuldade de pagar suas dívidas.

O número médio de prestações a pagar é maior para clientes inadimplentes. Logo, conclui-se que clientes que já tiveram maior número de prestações em aberto no seu nome do que o número de prestações em aberto médio do grupo tendem a ser maus pagadores.

A quantidade média de compras liquidadas é menor para clientes inadimplentes, portanto, clientes que liquidaram um menor número de compras do que a quantidade média de compras liquidadas do grupo, tendem a não honrar com seus compromissos financeiros.

O valor médio do atraso é maior para clientes inadimplentes, mostrando que clientes que atrasam seus pagamentos mais do que o atraso médio do grupo terão dificuldades de cumprir com suas “novas” obrigações.

A proporção média da prestação sobre a renda é maior para clientes adimplentes, indicando que clientes com maior proporção média da prestação em relação à renda do que a proporção média do grupo tendem a ser bons pagadores.

> ... o lambda de Wilks dá informação sobre as diferenças entre as variáveis e as médias dos grupos. O lambda de Wilks é obtido pela proporção da variação não explicada sobre a variação total. Varia entre 0 e 1, onde os pequenos valores indicam grandes diferenças, enquanto que os valores elevados indicam não haver diferenças. (PESTANA; GAGEIRO, 2003, p. 660).

Em todos os casos acima, nenhuma das variáveis utilizadas no modelo apresenta grande diferença em relação às médias dos grupos, pois estão muito próximos de 1. Esta conclusão corrobora com as observações obtidas por meio da análise do comportamento das médias de cada grupo em relação à média geral, feita anteriormente.

Ainda segundo os mesmos autores, quando uma das variáveis não distinguir entre os grupos e o pesquisador quiser mantê-la, o mesmo deverá recorrer ao método entre grupos e não ao método *stepwise,* informação que valida a escolha do presente trabalho pelo método ente grupos.

Levine et al. (2008, p. 369) asseguram que ao analisar uma variável numérica, com determinadas premissas atendidas, utiliza-se a análise da variância (ANOVA) para comparar as médias aritméticas dos grupos. O procedimento ANOVA, utilizado para o modelo completamente aleatório, é chamado de ANOVA de fator único. Em ANOVA, a variação total é subdividida entre variações que são atribuídas a diferenças entre grupos.

Pestana e Gageiro (2003, p. 663) mostram que a Estatística F é utilizada para descrever os grupos mais parecidos e testar a igualdade das médias dos grupos. Pode ser vista como uma medida de distância entre cada par de grupos. A não rejeição da hipótese de igualdade da média de uma variável nos grupos (significância > 0,05) aumenta a probabilidade de ela ser classificada incorretamente em outro grupo. Por meio da análise da Tabela 4, pode-se afirmar que apenas a significância do SPC e da prestação sobre a renda são maiores que 0,05, ou seja, a hipótese de igualdade da média de uma variável nos grupos adimplente e inadimplente só não é rejeitada para essas duas variáveis.

Por meio da análise do lambda de Wilks procura-se verificar se as médias dos grupos são iguais, uma vez que quanto mais significativamente diferentes, melhor os grupos serão discriminados. A significância do lambda de Wilks para a amostra global é igual a 0,000, ou seja, menor que 0,05, indicando que as médias entre os grupos são diferentes. No entanto, como o lambda de Wilks é alto (0,870), é possível afirmar que a essa diferença é pequena, conforme constatado anteriormente.

Pestana e Gageiro (2003, p. 660) afirmam que os determinantes de log mostram as dispersões existentes nos grupos e que o teste M de Box verifica se as diferentes dispersões observadas são ou não estatisticamente significativas.

Para a amostra global, constatou-se que a dispersão entre clientes adimplentes (91,911) é maior do que a dispersão entre os clientes inadimplentes (87,988). Como a significância do teste M de Box é igual a 0,000, valor inferior a 0,05, pode-se concluir que as diferenças observadas são significantes, ou seja, não há igualdade de dispersão entre os grupos adimplente e inadimplente. O Teste M de Box confirma a conclusão que se tirou analisando os determinantes de log.

> ... os valores próprios representam a razão da variação entre os grupos pela variação dentro dos grupos, ou seja, representam uma medida relativa da diferença entre os grupos na função discriminante. Quanto mais distante de 1, maior será a variação entre os grupos explicada pela função discriminante. (PESTANA; GAGEIRO, 2003, p. 661).

Logo, o valor próprio igual a 0,149 mostra que a variação entre os grupos explicada pela função discriminante é pequena.

Além disto, a função 1 contribui 100% para o total da variância entre os grupos, tendo todo o poder de separação. Quanto à correlação canônica, neste caso igual a 0,361, demonstra o nível de associação entre os *scores* discriminantes e os grupos, ou seja, representa quanto o modelo explica da variável dependente.

Os coeficientes determinam quais variáveis independentes têm maior efeito sobre a variável dependente, logo, as variáveis da amostra global que mais afetam a variável dependente *status* adimplente/inadimplente, tanto considerando os dados normalizados, quanto não normalizados, são estado civil, sexo e tipo de telefone residencial. Para os dados padronizados, as variáveis valor financiado, quantidade de prestações, prestações a vencer, maior número de prestações a vencer e maior atraso também apresentam coeficiente significativo. Para os dados não padronizados a variável SPC apresenta coeficiente significativo. Quanto maior o coeficiente, maior o efeito da variável independente sobre a variável dependente.

As funções dos centróides representam o efeito que cada grupo tem sobre a função discriminante, quando as médias das variáveis estão inseridas na equação discriminante, ou seja, quando se assume os valores médios e não os valores individuais de cada variável para calcular a função discriminante. Para a amostra global, verifica-se que a função dos centróides do grupo inadimplente é maior, então, pode-se afirmar que o grupo inadimplente impacta mais a função discriminante do que o grupo adimplente, quando as médias das variáveis estão inseridas na equação discriminante.

Levine et al. (2008, p. 115) destacam que o coeficiente de correlação mede a força relativa de uma relação linear entre duas variáveis numéricas. Os valores para o coeficiente de correlação vão desde -1, para uma correlação negativa perfeita, até +1, para uma correlação positiva perfeita. Perfeita significa dizer que se os pontos fossem desenhados em um gráfico de dispersão, todos esses pontos poderiam ser ligados por uma linha reta.

Os maiores coeficientes de correlação encontrados para a amaostra global são CEP com filial, valor financiado com valor da prestação e máxima prestação a vencer com

prestações a vencer, sendo o maior deles, máxima prestação a vencer com prestações a vencer.

O impacto da multicolinearidade não foi considerado em nenhuma das análises discriminantes rodadas, uma vez que o foco deste trabalho não está no teste dos modelos com base nas variáveis que o compõem, mas sim, no teste dos modelos aplicados a todas as variáveis que estão na base de dados fornecida pela empresa concedente do banco de dados. Tal decisão se justifica pelo fato de que a empresa utiliza todas as variáveis para analisar a concessão de crédito ao consumidor e não faria sentido, para a proposta deste trabalho, eliminá-las do modelo.

Pestana e Gageiro (2003, p. 661) afirmam que a matriz de estrutura evidencia a contribuição de cada variável para a função discriminante. Quanto maiores os coeficientes, mais a função discriminante detém a informação contida nessas variáveis. Os maiores coeficientes da função discriminante criada para a amostra global são quantidade de prestações, máxima prestação a vencer, prestações a vencer, valor financiado, tipo de telefone residencial, sexo, estado civil, maior atraso, n° da filial e CEP, com destaque para quantidade de parcelas, com 0,699.

Os coeficientes das funções da amostra global determinam qual o efeito das variáveis independentes sobre a variável dependente. Os coeficientes das funções das variáveis são praticamente os mesmos para os dois grupos, fato que indica que as variáveis independentes têm praticamente o mesmo impacto sobre os grupos de clientes adimplentes e clientes inadimplentes.

Os resultados da classificação da amostra global estão na Tabela 5:

Tabela 5 - Resultados da classificação da amostra global para 50% dos clientes classificados como adimplentes e 50% classificados como inadimplentes

Resultado da Classificação - AD sem segmentação			
Am Global	Adimplente	Inadimplente	Total
Adimplente	28176	11980	40156
Inadimplente	2950	6850	9800
Adimplente	**70,17%**	29,80%	100%
Inadimplente	30,10%	**69,90%**	100%

A função discriminante consegue classificar de maneira correta 70,17% dos clientes adimplentes. Para os clientes inadimplentes, 69,90% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 29,8%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes classificados como adimplentes é de 30,10%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com a análise discriminante da amostra global, e constatou-se que 71,39% dos clientes adimplentes foram classificados de maneira correta, bem como 70,23% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto, inclusive porque as porcentagens de acerto do teste são maiores que os resultados obtidos pelo modelo.

Para Brito e Assaf Neto (2008, p. 13), um outro procedimento que pode ser utilizado para avaliar a performance de um modelo é a construção de uma Curva ROC (*Receiver Operating Characteristic*). A curva ROC constitui uma técnica bastante útil para validar modelos de risco de crédito e está baseada nos conceitos da sensitividade e da especificidade. A sensitividade é a proporção de acerto na previsão da ocorrência de um evento nos casos em que ele de fato ocorreu. A especificidade é proporção de acerto na previsão da não ocorrência de um evento nos casos em que ele de fato não ocorreu.

Para a construção da Curva ROC, neste trabalho, utilizou-se o *software* SPSS. A figura a seguir representa a Curva ROC para a amostra global:

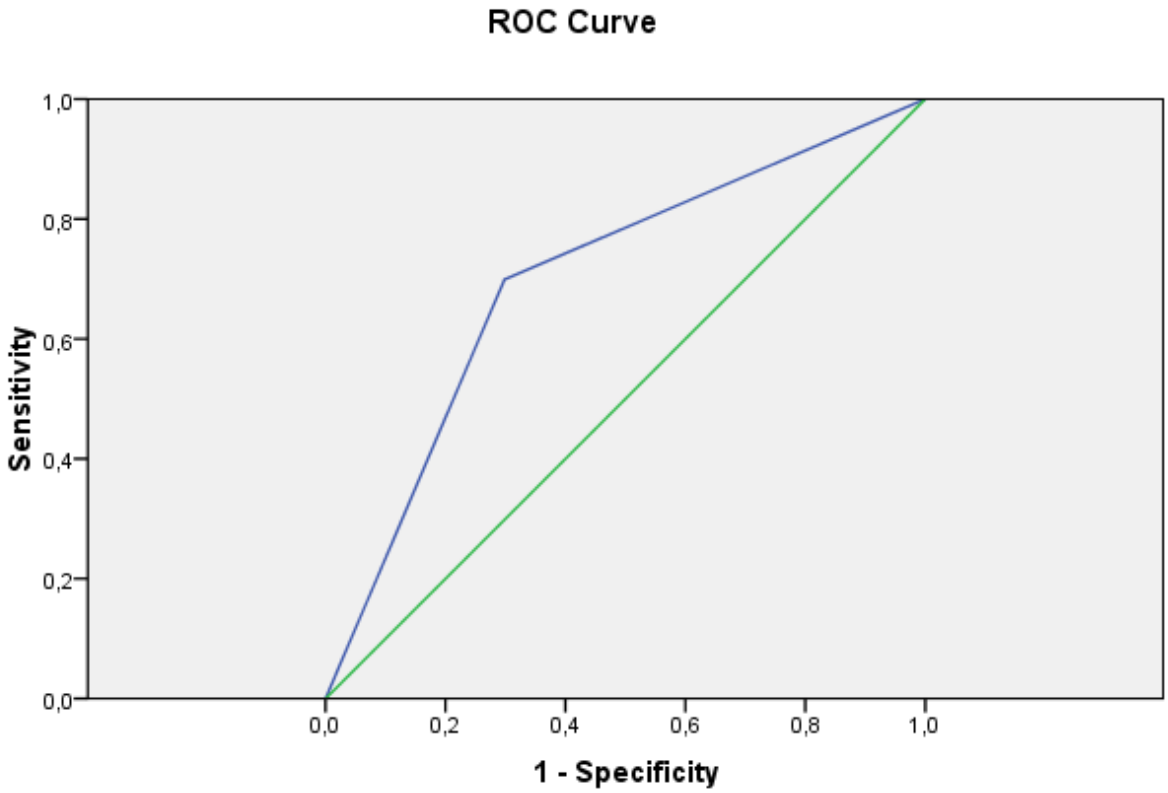

Gráfico 2 – Curva ROC para a amostra global

Brito e Assaf Neto (2008, p. 13) utilizam os seguintes critérios para avaliar o resultado da área sob a Curva ROC: (i) área no intervalo entre 0,7 e 0,8: discriminação aceitável; (ii) área no intervalo entre 0,8 e 0,9: excelente discriminação; (iii) área acima de 0,9: excepcional discriminação.

A Curva ROC do modelo de risco de crédito, representada no gráfico acima, revela que a área sob a curva é de 0,70. Segundo a escala proposta, esse valor indica um poder discriminatório aceitável para o modelo.

Para os resultados apresentados até aqui, considerou-se a probabilidade de um cliente ser classificado como adimplente igual à probabilidade de tal cliente ser classificado como inadimplente, ou seja, 50% para cada grupo. No entanto, para a análise discriminante rodada no SPSS é possível escolher entre as opções todos os grupos iguais ou classificação pelo tamanho do grupo. Assim, as duas formas foram testadas.

Os resultados da opção todos os grupos iguais foram explorados no item 5.2 deste trabalho. Para a opção classificação pelo tamanho do grupo, verificou-se que a chance de um cliente ser classificado como adimplente é de 80,38%, ao passo que a chance de que tal cliente seja classificado como inadimplente é de 19,62%. Tais resultados foram obtidos por meio da análise discriminante rodada no SPSS, considerando a classificação pelo tamanho do grupo.

A função discriminante consegue classificar de maneira correta 96,56% dos clientes adimplentes. Para os clientes inadimplentes, 14,98% são classificados corretamente. Logo, fica evidente que a capacidade de classificar corretamente os clientes adimplentes é maior para a classificação pelo tamanho do grupo, ao passo que a capacidade de classificar os clientes inadimplentes de maneira correta é maior para a opção todos os grupos iguais.

No entanto, considerando as porcentagens de erro obtidas, verifica-se que apesar de o Erro Tipo I, representado pelo clientes adimplentes, classificados como inadimplentes, ter sido de 3,44%; o Erro Tipo II, representado pela porcentagem de clientes inadimplentes classificados como adimplentes foi de 85,02%, fato que anula o ganho de acerto na classificação de clientes adimplentes, que se obteve com esse tipo de teste.

A seguir serão apresentados os resultados da análise discriminante para as 21 micro-regiões formadas para análise.

5.3 ANÁLISE DISCRIMINANTE DAS 21 MICRO-REGIÕES

O próximo passo é rodar a análise discriminante para cada um dos grupos formados com base na região à qual a filial de venda pertence.

A distribuição dos clientes utilizada para rodar a análise discriminante das 21 micro-regiões está na Tabela 6:

Tabela 6 - Distribuição dos clientes por micro-região

Regiões	Qtde de Clientes	%
Região 1	1.492	2,99%
Região 2	4.289	8,59%
Região 3	766	1,53%
Região 4	3.467	6,94%
Região 5	2.884	5,77%
Região 6	2.016	4,04%
Região 7	1.674	3,35%
Região 8	1.960	3,92%
Região 9	1.803	3,61%
Região 11	2.099	4,20%
Região 12	2.804	5,61%
Região 13	74	0,15%
Região 14	3.703	7,41%
Região 16	1.067	2,14%
Região 17	1.596	3,19%
Região 18	1.086	2,17%
Região 19	5.827	11,66%
Região 20	3.900	7,81%
Região 21	784	1,57%
Região 25	3.556	7,12%
Região 26	3.108	6,22%
Total	**49.955**	**100,00%**

Foi rodada a análise discriminante para cada uma das micro-regiões apresentadas na Tabela 6 acima. No entanto, optou-se por não apresentar todas as tabelas geradas, dada a extensa quantidade de regiões estudada e, por isso, os resultados serão apresentados de maneira resumida. Mesmo assim, cabe lembrar que o passo a passo dessas análises foi o mesmo apresentado no item 5.2, para a amostra global.

Foram utilizados os dados de 49.955 clientes (os outros 45 clientes foram considerados como *missing values*), para rodar a análise discriminante.

A idade média dos clientes que compõem o grupo inadimplentes é mais baixa do que a idade média dos clientes que estão no grupo adimplentes, em todas as regiões analisadas. Assim, podemos afirmar que clientes mais jovens apresentam maior tendência de não honrar com as obrigações financeiras assumidas. Para chegar-se a esta conclusão basta comparar a

idade média dos grupos clientes adimplentes e clientes inadimplentes com a idade média do grupo todo. A média dos clientes adimplentes sempre estará acima da média do grupo todo, bem como a média dos clientes inadimplentes sempre estará abaixo dela.

A renda média dos inadimplentes é menor para as regiões 1 a 12 e 14 a 26; já para a Região 13, a renda média dos inadimplentes é maior. O fato de a Região 13 ser composta por apenas 74 clientes, número menor do que o encontrado nas outras regiões, justifica este resultado. Para os outros casos, é possível afirmar que, geralmente, clientes com renda média mensal acima da renda média do grupo tendem a honrar com seus compromissos, ao passo que pessoas com renda média mensal abaixo da média do grupo terão mais dificuldades de fazê-lo.

O *score* médio apresenta-se maior para clientes adimplentes nas regiões 1, 6, 7, 16, 18, 19, 25 e 26; já para as regiões 2, 3, 4, 5, 8, 9, 11, 12, 13, 14, 17, 20, 21 o *score* médio se apresenta maior para clientes inadimplentes. Essas informações indicam que, de acordo com o *score* calculado pela empresa, que é baseado em variáveis fornecidas pelo SPC, o perfil dos clientes muda de região para região. Em parte delas, mesmo apresentando *score* médio maior que o *score* médio do grupo, os clientes pagam suas dívidas. Já nas outras, o *score* médio é maior do que o *score* médio do grupo apenas para clientes possivelmente inadimplentes. Este perfil deve ser analisado antes de se tomar qualquer decisão estratégica.

O valor médio financiado pelos clientes inadimplentes é maior do que o valor médio financiado pelo grupo, em todas as regiões, ou seja, clientes que buscam volumes maiores de crédito tendem a não pagar, pelo menos parte de sua dívida.

A quantidade média de parcelas é maior para clientes inadimplentes nas 21 regiões. Assim, é possível concluir que clientes que precisam dividir sua compra em um maior número de prestações do que o número de prestações médio do grupo, possivelmente, terão dificuldades de pagamento.

O valor médio da parcela é menor para clientes inadimplentes para as regiões 1 a 14 e 17 a 26; já para a Região 16, o valor médio da parcela é maior para clientes inadimplentes. Logo, com exceção da Região 16, em todas as outras os clientes que pagam parcelas menores que a parcela média do grupo apresentarão a tendência de não cumprir com suas obrigações.

Já na Região 16, esta situação é inversa, clientes que pagam parcelas maiores tendem a não cumprir com as obrigações assumidas. No geral, parcelas com valores menores estão vinculadas com maior número de prestações, fato que faz com que a interpretação desta variável corrobore com a interpretação da variável quantidade média de parcelas.

O valor médio de entrada é menor para clientes inadimplentes para as regiões 1 a 18, 20, 25 e 26; já para as Regiões 19 e 21, o valor médio de entrada é maior para clientes inadimplentes. Para as regiões 1 a 18, 20, 25 e 26, clientes que derem uma entrada de valor menor do que o valor médio de entrada do grupo, possivelmente, terão problemas com o pagamento do saldo remanescente. Já para as regiões 19 e 21, os clientes que derem uma entrada de valor maior do que o valor médio de entrada do grupo é que terão esse problema.

A quantidade média de prestações a pagar é maior para clientes inadimplentes em todas as regiões. Assim, fica claro que clientes que apresentam maior número de prestações em aberto do que a quantidade média de prestações em aberto do grupo contam com maior dificuldade de pagar suas dívidas.

O número médio de prestações a pagar é maior para clientes inadimplentes nas 21 regiões. Logo, conclui-se que clientes que já tiveram maior número de prestações em aberto no seu nome do que o número de prestações médio em aberto do grupo tendem a ser maus pagadores.

A quantidade média de compras liquidadas é menor para clientes inadimplentes de todas as regiões, portanto, clientes que liquidaram um menor número de compras do que a quantidade média de compras liquidadas do grupo, tendem a não honrar com seus compromissos financeiros.

O valor médio do atraso é maior para clientes inadimplentes em todas as regiões, mostrando que clientes que atrasam seus pagamentos mais do que o atraso médio do grupo terão dificuldades de cumprir com suas “novas” obrigações.

A proporção média da prestação em relação à renda é maior para clientes inadimplentes para as regiões 1, 11, 16 e 26; já para as regiões 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 17, 18, 19, 20, 21, 25 a proporção média da prestação em relação à renda se apresenta maior

para clientes adimplentes, ou seja, nas regiões 1, 11, 16 e 26, clientes que apresentam maior proporção média da prestação em relação à renda do que a proporção média do grupo tendem a ser maus pagadores. Já para as outras regiões, clientes com maior proporção média da prestação em relação à renda do que a proporção média do grupo, tendem a ser bons pagadores.

Como cada região apresenta suas particularidades em termos sociais e econômicos como, por exemplo, profissão e renda mensal, a capacidade de generalização está em considerar esta informação na aplicação do modelo proposto.

Considerando o perfil dos clientes traçado acima, acredita-se na possibilidade de aplicação do modelo proposto não só para a empresa que forneceu o banco de dados para este estudo, como também para qualquer outra empresa do setor varejista (que possua características estruturais semelhantes a ela); uma vez que ao dividir a amostra por regiões, vieses foram eliminados, ou seja, os grupos tornaram-se mais homogêneos, no que tange ao comportamento e aos atributos apresentados pelos clientes, permitindo a generalização do perfil dos clientes para cada região, independentemente da empresa na qual a venda for realizada.

Para todas as regiões estudadas, os valores do lambda de Wilks (para cada variável) são muito próximos de 1, indicando que nenhuma das variáveis utilizadas no modelo apresenta grande diferença em relação às médias dos grupos.

Por meio da análise do lambda de Wilks (do grupo) procura-se verificar se as médias dos grupos adimplente e inadimplente são iguais, uma vez que quanto mais significativamente diferentes, melhor os grupos serão discriminados. Com exceção da Região 13, os testes para todas as outras regiões apresentam significância igual a 0,000, portanto, menor que 0,05, fato que indica que as médias entre os grupos são diferentes. No entanto, como o lambda de Wilks é alto, é possível afirmar que essa diferença é pequena, conforme constatado anteriormente.

A significância é maior que 0,05 para a Região 13 por se tratar de uma região menor do que as outras, sendo formada por apenas 74 clientes. Assim, explica-se o fato de não haver diferença entre as médias.

Com relação à hipótese de igualdade das médias de grupo podemos elaborar a seguinte tabela:

Tabela 7 - Teste de Igualdade das Médias por micro-região

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R11	R12	R13	R14	R16	R17	R18	R19	R20	R21	R25	R26
Variáveis	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**
Idade	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	0,217	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**
Renda	**0,000**	**0,000**	**0,017**	**0,009**	**0,000**	**0,000**	0,477	**0,009**	**0,020**	0,067	**0,000**	0,769	**0,000**	**0,002**	0,122	0,087	**0,001**	**0,017**	**0,032**	**0,039**	**0,002**
SPC	0,114	**0,006**	0,655	**0,002**	0,418	0,258	0,327	0,959	0,060	0,685	0,700	0,618	0,995	**0,015**	**0,001**	0,397	0,071	0,369	**0,001**	0,050	0,844
Valor_Financiado	**0,000**	**0,000**	0,142	**0,000**	**0,000**	**0,025**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	0,958	**0,000**	**0,000**	0,056	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**
Qtde_Prest	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,017**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**
Valor_Prest	0,094	**0,001**	0,090	**0,027**	**0,005**	**0,005**	0,059	**0,009**	0,888	**0,003**	**0,004**	0,260	**0,011**	0,068	**0,006**	0,153	0,107	**0,044**	0,080	0,141	0,055
Valor_Entrada	**0,012**	0,050	0,492	0,260	0,146	0,071	0,053	0,242	0,444	0,560	0,052	0,478	**0,001**	0,348	0,978	0,517	0,833	0,398	0,847	0,357	0,061
Prest_a_Vencer	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,015**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**
Max_Pret	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,011**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**
N_Compras	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	0,463	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**
Maior_Atraso	0,259	**0,024**	**0,003**	**0,004**	**0,000**	**0,002**	**0,000**	**0,000**	0,405	0,618	**0,000**	0,171	**0,000**	**0,000**	0,741	0,704	**0,000**	**0,013**	**0,036**	**0,015**	**0,000**
Prest_Renda	0,727	0,441	0,218	0,546	0,973	0,050	0,261	0,457	0,684	0,462	0,468	0,322	0,506	**0,001**	0,103	0,497	0,640	0,283	0,616	0,651	0,113

As significâncias das variáveis que estão em negrito são menores que 0,05, fato que indica que a hipótese de igualdade da média de cada variável nos grupos é rejeitada. Assim, é possível concluir, por exemplo, que só não existe diferença significativa entre a média da variável idade no grupo adimplente e a média da mesma variável no grupo inadimplente, para a região 13; bem como só não existe diferença significativa entre a média da variável renda no grupo adimplente e a média da mesma variável no grupo inadimplente, para as regiões 7, 13, 17 e 18; e assim, sucessivamente.

Por meio da análise dos determinantes de log, conclui-se que a dispersão entre as variáveis que caracterizam o grupo dos clientes adimplentes é maior do que a dispersão entre as variáveis que caracterizam o grupo de clientes inadimplentes, para as regiões 1 a 12 e 14 a 25. Já para a Região 26, a dispersão entre as variáveis do grupo de clientes adimplentes é menor do que a dispersão entre as variáveis dos clientes inadimplentes. Para a Região 13, existe dispersão apenas entre as variáveis dos clientes do grupo adimplente, mas o coeficiente dessa dispersão não foi medido devido ao número baixo de casos.

Com exceção da Região 13, para todas as outras regiões, a significância do teste M de Box é igual a 0,000, que por se tratar de valor inferior a 0,05, nos leva a concluir que as

diferenças de dispersão observadas entre os grupos são significantes, ou seja, não há igualdade de dispersão entre as variáveis que caracterizam os clientes dos grupos adimplente e inadimplente e a média desses grupos. O teste M de Box confirma a conclusão que se tirou analisando os determinantes de log.

O teste M de Box não foi realizado para a Região 13, uma vez que não se tem a dispersão entre as variáveis do grupo dos clientes inadimplentes.

O valor próprio calculado para as 21 regiões representa uma medida relativa da diferença entre os grupos na função discriminante, que variando de 0,124 a 0,372, indica que a diferença entre os grupos adimplente e inadimplente na função discriminante é pequena.

Para todas as regiões estudadas, o resultado é de apenas uma função, correspondendo a 100% da variância explicada em termos de diferenças entre grupos. Quanto à correlação canônica, que demonstra o nível de associação entre os escores discriminantes e os grupos, os valores variam de 0,3 a 0,5, representando para cada região quanto o modelo explica da variável dependente.

As variáveis que mais afetam a variável dependente *status* adimplente/inadimplente são estado civil, sexo, tipo de telefone residencial e comercial e quantidade de prestações, tanto normalizadas quanto não normalizadas; a variável SPC não normalizada; e as variáveis valor financiado, prestações a vencer, máxima prestação a vencer e maior atraso, normalizadas.

As funções dos centróides do grupo inadimplente é maior para todas as regiões, então, pode-se afirmar que o grupo inadimplente impacta mais a função discriminante do que o grupo adimplente, quando as médias das variáveis são consideradas para compor a equação discriminante.

Os maiores coeficientes de correlação (entre 0,5 e 0,99) para todas as regiões são apresentados na tabela a seguir:

Tabela 8 - Maiores coeficientes de correlação por micro-região

	Regiões																				
Coeficientes de Correlação	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**11**	**12**	**13**	**14**	**16**	**17**	**18**	**19**	**20**	**21**	**25**	**26**
Max prest a vencer x prest a vencer	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Valor financiado x valor prest	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x
Prest s/ renda x valor prest	x	x	x	x	x	x					x	x	x	x	x	x	x	x	x	x	
Valor financiado x qtde prest									x		x			x		x	x	x	x		
Prest s/ renda x valor financiado				x							x			x				x			
Valor entrada x valor prest												x									x
CEP x filial			x																		
Tipo fone coml x tipo fone res												x									
Valor financiado x valor da entrada												x									
Max prest a vencer x qtde prest												x									
Prest s/ renda x valor entrada												x									
Qtde prest x prest a vencer												x									

Os coeficientes de correlação máxima prestação a vencer com prestações a vencer, valor financiado com valor das prestações e prestações sobre a renda e valor da prestação são os maiores coeficientes com maior frequência.

As variáveis que mais contribuem para a função discriminante são quantidade de prestações, prestações a vencer, máxima prestação a vencer, sexo, valor financiado, estado civil, tipo de telefone residencial, maior atraso, filial e tipo de telefone comercial.

Os coeficientes das funções das variáveis são praticamente os mesmos para os dois grupos, em todas as regiões, fato que indica que as variáveis independentes têm praticamente o mesmo impacto sobre os grupos de clientes adimplentes e clientes inadimplentes.

A seguir é apresentada a soma dos resultados da classificação da análise discriminante para as 21 regiões:

Tabela 9 - Resultados da classificação da análise discriminante para a soma dos resultados das 21 micro-regiões para 50% dos clientes classificados como adimplentes e 50% classificados como inadimplentes

Resultado da Classificação - AD com segmentação			
21 Regiões	Adimplente	Inadimplente	Total
Adimplente	28665	11490	40155
Inadimplente	2910	6890	9800
Adimplente	**71,39%**	28,61%	100%
Inadimplente	29,69%	**70,31%**	100%

A função discriminante consegue classificar de maneira correta 71,39% dos clientes adimplentes. Para os clientes inadimplentes, 70,31% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 28,61%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 29,69%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com a análise discriminante das 21 micro-regiões, e constatou-se que 72,77% dos clientes adimplentes foram classificados de maneira correta, bem como 71,28% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto, inclusive porque as porcentagens de acerto do teste são maiores que os resultados obtidos pelo modelo.

A Curva ROC do modelo de risco de crédito para as 21 micro-regiões revela que a área sob a curva varia de 0,69 a 0,80, da região 1 até a região 21. Segundo a escala anteriormente proposta, esse valor indica um poder discriminatório aceitável para o modelo.

Para a opção classificação pelo tamanho do grupo, verificou-se que a chance de um cliente ser classificado como adimplente é de 80,38%, ao passo que a chance de que tal cliente seja classificado como inadimplente é de 19,62%. Neste caso, a função discriminante consegue classificar de maneira correta 96,27% dos clientes adimplentes. Para os clientes inadimplentes, 19,53% são classificados corretamente. Logo, fica evidente que a capacidade de classificar corretamente os clientes adimplentes é maior para a classificação pelo tamanho do grupo, ao passo que a capacidade de classificar os clientes inadimplentes de maneira correta é maior para a opção todos os grupos iguais.

No entanto, considerando as porcentagens de erro obtidas, verifica-se que apesar de o Erro Tipo I, representado pelo clientes adimplentes, classificados como inadimplentes, ter sido de 3,73%; o Erro Tipo II, representado pela porcentagem de clientes inadimplentes classificados como adimplentes foi de 80,47%, fato que anula o ganho de acerto na classificação de clientes adimplentes, que se obteve com esse tipo de teste.

A seguir serão apresentados os resultados da análise discriminante para as 3 macro-regiões formadas para análise.

5.4 ANÁLISE DISCRIMINANTE DAS 3 MACRO-REGIÕES

A seguir serão apresentados os resultados da análise discriminante para cada um dos grupos formados, de acordo com a macro-região à qual a filial de venda pertence. As regiões consideradas são Sudeste (1), Centro Oeste (2) e Sul (3).

A distribuição dos clientes utilizada para rodar a análise discriminante das 3 regiões está na Tabela 10:

Tabela 10 - Distribuição dos clientes por macro-região

Regiões	Qtde de Clientes	%
Região 1 - Sudeste	29.343	58,74%
Região 2 – Centro Oeste	5.688	11,38%
Região 3 – Sul	14.925	29,88%
Total	**49.956**	**100,00%**

Dada a distribuição dos clientes por região, são apresentados os resultados da análise discriminante. Da mesma forma como ocorreu com os resultados da análise discriminante para as 21 micro-regiões, os resultados desta análise serão apresentados de maneira resumida.

Para rodar a análise discriminante foram utilizados os dados de 49.956 clientes (os outros 44 clientes foram considerados como *missing values*).

Clientes adimplentes apresentam mais idade, maior renda, assumem parcelas de maior valor, dão maiores valores de entrada e apresentam maior número de compras liquidadas do que os clientes inadimplentes. Já os clientes inadimplentes apresentam maiores *scores* do SPC, financiam valores maiores, em mais parcelas, têm maior quantidade de parcelas em aberto no seu nome e apresentam maiores atrasos em pagamentos anteriores do que os clientes adimplentes.

A proporção média da prestação em relação à renda é maior para clientes inadimplentes na região 1; já para as regiões 2 e 3, a proporção média da prestação em relação à renda se apresenta maior para clientes adimplentes, ou seja, na região 1, clientes que apresentam maior proporção média da prestação em relação à renda do que a proporção média do grupo tendem a ser maus pagadores. Já para as outras regiões, clientes com maior proporção média da prestação, em relação à renda do que a proporção média do grupo, tendem a ser bons pagadores.

Os valores do lambda de Wilks (para cada variável) são muito próximos de 1, para todas as regiões estudadas, indicando que nenhuma das variáveis utilizadas no modelo apresenta grande diferença em relação às médias dos grupos.

Por meio da análise do lambda de Wilks (do grupo) procura-se verificar se as médias dos grupos adimplente e inadimplente são iguais, uma vez que quanto mais significativamente diferentes, melhor os grupos serão discriminados. Os testes para todas as regiões apresentam significância igual a 0,000, portanto, menor que 0,05, fato que indica que as médias entre os grupos são diferentes. No entanto, como o lambda de Wilks é alto, é possível afirmar que a essa diferença é pequena, conforme constatado anteriormente.

Com relação à hipótese de igualdade das médias de grupo temos a seguinte tabela:

Tabela 11 - Teste de Igualdade das Médias por macro-região

	Sudeste	**Centro Oeste**	**Sul**
Variáveis	**Sig**	**Sig**	**Sig**
SPC	0,608	0,384	0,237
Prest_Renda	0,442	0,621	0,327
Valor_Entrada	**0,002**	**0,018**	**0,000**
Idade	**0,000**	**0,000**	**0,000**
Renda	**0,000**	**0,000**	**0,000**
Valor_Financiado	**0,000**	**0,000**	**0,000**
Qtde_Prest	**0,000**	**0,000**	**0,000**
Valor_Prest	**0,000**	**0,000**	**0,000**
Prest_a_Vencer	**0,000**	**0,000**	**0,000**
Max_Pret_a_Vencer	**0,000**	**0,000**	**0,000**
N_Compras	**0,000**	**0,000**	**0,000**
Maior_Atraso	**0,000**	**0,000**	**0,000**

As significâncias das variáveis que estão em negrito são menores que 0,05, fato que indica que a hipótese de igualdade da média de cada variável nos grupos é rejeitada. Assim, é possível concluir, por exemplo, que não existe diferença significativa entre a média da variável SPC no grupo adimplente e a média da mesma variável no grupo inadimplente, para as regiões Sudeste, Centro Oeste e Sul; bem como, entre a média da variável prestação sobre a renda no grupo adimplente e a média da mesma variável no grupo inadimplente, para as 3 regiões.

Conclui-se, por meio da análise dos determinantes de log, que a dispersão entre as variáveis que caracterizam o grupo dos clientes adimplentes é maior do que a dispersão entre as variáveis que caracterizam o grupo de clientes inadimplentes, para as 3 regiões estudadas.

A significância do teste M de Box é igual a 0,000, para todas as regiões, que por se tratar de valor inferior a 0,05, nos leva a concluir que as diferenças de dispersão observadas entre os grupos são significantes, ou seja, não há igualdade de dispersão entre as variáveis que caracterizam os clientes dos grupos adimplente e inadimplente e a média desses grupos. O teste M de Box confirma a conclusão que se tirou analisando os determinantes de log.

O valor próprio calculado para as 3 regiões representa uma medida relativa da diferença entre os grupos na função discriminante, que variando de 0,144 a 0,193, indica que a diferença entre os grupos adimplente e inadimplente na função discriminante é pequena.

Para todas as regiões estudadas, o resultado é de apenas uma função, correspondendo a 100% da variância explicada em termos de diferenças entre grupos. Quanto à correlação canônica, que demonstra o nível de associação entre os escores discriminantes e os grupos, os valores variam de 0,355 a 0,402, representando para cada região quanto o modelo explica da variável dependente.

As variáveis que mais afetam a variável dependente *status* adimplente/inadimplente são estado civil, sexo, tipo de telefone residencial e comercial, SPC e quantidade de prestações, tanto normalizadas quanto não normalizadas; a variável prestação sobre a renda não normalizada; e as variáveis valor financiado, prestações a vencer, máxima prestação a vencer e maior atraso, normalizadas.

As funções dos centróides do grupo inadimplente é maior para todas as regiões, então, pode-se afirmar que o grupo inadimplente impacta mais a função discriminante do que o grupo adimplente, quando as médias das variáveis são consideradas para compor a equação discriminante.

Os maiores coeficientes de correlação (entre 0,5 e 0,99) para todas as regiões são apresentados a seguir:

Tabela 12 - Maiores coeficientes de correlação por macro-região

	Regiões		
Coeficientes de Correlação	**1**	**2**	**3**
Valor prest x valor financiado	x	x	x
Valor prest x Prest s/ renda	x	x	
Max prest a vencer x prest a vencer	x	x	x

Os coeficientes de correlação valor da prestação com valor financiado, valor da prestação com prestação sobre a renda e máxima prestação a vencer com prestações a vencer são os maiores coeficientes, com maior frequência.

As variáveis que mais contribuem para a função discriminante são quantidade de prestações, máxima prestação a vencer, prestações a vencer, valor financiado, maior atraso, tipo de telefone residencial, sexo, estado civil, tipo de telefone comercial, filial e CEP.

Os coeficientes das funções das variáveis são praticamente os mesmos para os dois grupos, em todas as regiões, fato que indica que as variáveis independentes têm praticamente o mesmo impacto sobre os grupos de clientes adimplentes e clientes inadimplentes.

A seguir é apresentada a soma dos resultados da classificação da análise discriminante para as 3 regiões:

Tabela 13 - Resultados da classificação da análise discriminante para a soma dos resultados das 3 macro-regiões para 50% dos clientes classificados como adimplentes e 50% classificados como inadimplentes

Resultado da Classificação - AD com segmentação			
3 regiões	Adimplente	Inadimplente	Total
Adimplente	28295	11861	40155
Inadimplente	2936	6864	9800
Adimplente	**70,46%**	29,54%	100%
Inadimplente	29,96%	**70,04%**	100%

A função discriminante consegue classificar de maneira correta 70,46% dos clientes adimplentes. Para os clientes inadimplentes, 70,04% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 29,54%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 29,96%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com a análise discriminante das 3 macro-regiões, e constatou-se que 71,37% dos clientes adimplentes foram classificados de maneira correta, bem como 69,85% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto.

A Curva ROC do modelo de risco de crédito para as 3 macro-regiões revela que a área sob a curva varia de 0,69 a 0,71, da região 1 até a região 3. Segundo a escala anteriormente proposta, esse valor indica um poder discriminatório aceitável para o modelo.

Para a opção classificação pelo tamanho do grupo, verificou-se que a chance de um cliente ser classificado como adimplente é de 80,38%, ao passo que a chance de que tal cliente seja classificado como inadimplente é de 19,62%. Neste caso, a função discriminante consegue classificar de maneira correta 96,48% dos clientes adimplentes. Para os clientes inadimplentes, 15,70% são classificados corretamente. Logo, fica evidente que a capacidade de classificar corretamente os clientes adimplentes é maior para a classificação pelo tamanho do grupo, ao passo que a capacidade de classificar os clientes inadimplentes de maneira correta é maior para a opção todos os grupos iguais.

No entanto, considerando as porcentagens de erro obtidas, verifica-se que apesar de o Erro Tipo I, representado pelo clientes adimplentes, classificados como inadimplentes, ter sido de 3,52%; o Erro Tipo II, representado pela porcentagem de clientes inadimplentes classificados como adimplentes foi de 84,30%, fato que anula o ganho de acerto na classificação de clientes adimplentes, que se obteve com esse tipo de teste.

A seguir serão apresentados os resultados da análise discriminante dos grupos formados pela análise de *clusters K-Means.*

5.5 ANÁLISE DISCRIMINANTE DOS GRUPOS FORMADOS PELA ANÁLISE *K-MEANS*

Foram utilizadas todas as variáveis elencadas no Capítulo 4 para rodar a análise de *clusters K-Means*. Os casos foram rotulados pela variável dependente adimplente/ inadimplente.

O número de *clusters* determinado para rodar a análise *K-Means* foi dois, com base nos resultados obtidos por meio da análise *TwoStep Cluster*, que gerou dois grupos espontaneamente.

Fávero et al. (2009, p. 220) afirmam que o método de análise de *clusters* escolhido permite que se opte entre interagir e classificar, ou apenas classificar, sendo que, no primeiro caso, o procedimento se encarrega de estimar os centróides interativamente a cada nova

observação designada e de classificar os sujeitos. Na opção classificar apenas, os centróides não são atualizados e é utilizada quando se busca atribuir casos adicionais nos *clusters* já criados. Como neste trabalho nenhum *cluster* foi previamente criado, utiliza-se o método interagir e classificar.

A distribuição dos clientes utilizada para rodar a análise discriminante dos 2 grupos formados está na Tabela 14:

Tabela 14 - Distribuição dos clientes por *cluster* da análise *K-Means*

Clusters	**Qtde de Clientes**	**%**
Cluster 1	3.755	7,52%
Cluster 2	46.201	92,48%
Total	**49.956**	**100,00%**

A seguir serão apresentados os resultados da análise discriminante para cada um dos grupos formados, de acordo com a metodologia de análise de *cluster* não hierárquica *K-Means*.

Foram utilizados os dados de 49.956 clientes (os outros 44 clientes foram considerados como *missing values*) para rodar a análise discriminante.

Clientes adimplentes apresentam mais idade, maior renda, assumem parcelas de maior valor, dão maiores valores de entrada, apresentam maior número de compras liquidadas e maior proporção da prestação sobre a renda do que os clientes inadimplentes. Já os clientes inadimplentes apresentam maiores *scores* do SPC, financiam valores maiores, em mais parcelas, têm maior quantidade de parcelas, em aberto, no seu nome e apresentam maiores atrasos em pagamentos anteriores do que os clientes adimplentes.

Os valores do lambda de Wilks (para cada variável) são muito próximos de 1, para os dois *clusters* estudados, indicando que nenhuma das variáveis utilizadas no modelo apresenta grande diferença em relação às médias dos grupos.

Por meio da análise do lambda de Wilks (do grupo) procura-se verificar se as médias dos grupos adimplente e inadimplente são iguais, uma vez que quanto mais significativamente

diferentes, melhor os grupos serão discriminados. Os testes para os dois grupos apresentam significância igual a 0,000, portanto, menor que 0,05, fato que indica que as médias entre os grupos são diferentes. No entanto, como o lambda de Wilks é alto, é possível afirmar que a essa diferença é pequena, conforme constatado anteriormente.

O teste de igualdade das médias á apresentado na tabela a seguir:

Tabela 15 - Teste de igualdade das médias para cada *cluster* da análise *K-Means*

***Cluster* 1**		***Cluster* 2**	
Variáveis	**Sig**	**Variáveis**	**Sig**
Idade	**0,000**	Idade	**0,000**
Valor_Financiado	**0,000**	Renda	**0,000**
Qtde_Prest	**0,000**	Valor_Financiado	**0,000**
Prest_a_Vencer	**0,000**	Qtde_Prest	**0,000**
Max_Pret_a_Vencer	**0,000**	Valor_Prest	**0,000**
N_Compras	**0,000**	Valor_Entrada	**0,000**
Maior_Atraso	**0,000**	Prest_a_Vencer	**0,000**
Valor_Prest	**0,003**	Max_Pret_a_Vencer	**0,000**
Prest_Renda	**0,003**	N_Compras	**0,000**
SPC	**0,008**	Maior_Atraso	**0,000**
Renda	0,596	Prest_Renda	0,366
Valor_Entrada	0,783	SPC	0,558

As significâncias das variáveis que estão em negrito são menores que 0,05, fato que indica que a hipótese de igualdade da média de cada variável nos grupos é rejeitada. Assim, é possível concluir que não existe diferença significativa entre a média das variáveis renda e valor de entrada no grupo adimplente e a média das mesmas variáveis no grupo inadimplente, para o *Cluster* 1 formado por meio da análise *K-Means*; bem como não existe diferença significativa entre as médias das variáveis prestação sobre a renda e SPC no grupo adimplente e a média da mesma variável no grupo inadimplente, para o *Cluster* 2 formado por esta mesma análise.

Conclui-se, por meio da análise dos determinantes de log, que a dispersão entre as variáveis que caracterizam o grupo dos clientes adimplentes é maior do que a dispersão entre as variáveis que caracterizam o grupo de clientes inadimplentes, para os dois *clusters* estudados.

A significância do teste M de Box foi igual a 0,000 para os dois *clusters* estudados, que por se tratar de valor inferior a 0,05, nos leva a concluir que as diferenças de dispersão observadas entre os grupos são significantes, ou seja, não há igualdade de dispersão entre as variáveis que caracterizam os clientes dos grupos adimplente e inadimplente e a média desses grupos. O teste M de Box confirma a conclusão que se tirou analisando os determinantes de log.

O valor próprio calculado para os dois grupos representa uma medida relativa da diferença entre os grupos na função discriminante, que sendo de 0,166 para o *Cluster* 1 e 0,150 para o *Cluster* 2, indica que a diferença entre os grupos adimplente e inadimplente na função discriminante é pequena.

Para todas as regiões estudadas, o resultado é de apenas uma função, correspondendo a 100% da variância explicada em termos de diferenças entre grupos. Quanto à correlação canônica, que demonstra o nível de associação entre os escores discriminantes e os grupos, os valores são de 0,377 para o *Cluster* 1 e 0,361 para o *Cluster* 2, representando para cada região quanto o modelo explica da variável dependente.

As variáveis que mais afetam a variável dependente *status* adimplente/inadimplente são estado civil, sexo, tipo de telefone residencial e comercial, SPC e quantidade de prestações, tanto normalizadas quanto não normalizadas; e as variáveis valor financiado, prestações a vencer, máxima prestação a vencer e maior atraso, normalizadas.

As funções dos centróides do grupo inadimplente são maiores nos dois *clusters* analisados, então, pode-se afirmar que o grupo inadimplente impacta mais a função discriminante do que o grupo adimplente, quando as médias das variáveis são consideradas para compor a equação discriminante.

A tabela a seguir representa os maiores coeficientes de correlação por *cluster*:

Tabela 16 - Maiores coeficientes de correlação por *cluster* da análise *K-Means*

	Clusters	
Coeficientes de Correlação	**1**	**2**
CEP x filial	x	x
Valor prest x valor financiado	x	x
Max prest a vencer x prest a vencer	x	x
Valor prest x Prest s/ renda	x	
Valor financiado x prest s/ renda	x	

Os coeficientes de correlação CEP com filial, valor da prestação com valor financiado e máxima prestação a vencer com prestações a vencer são os maiores coeficientes, com maior frequência.

As variáveis que mais contribuem para a função discriminante são quantidade de prestações, prestações a vencer, máxima prestação a vencer, valor financiado, tipo de telefone residencial, maior atraso, sexo, estado civil, filial, SPC, tipo de telefone comercial e CEP.

Os coeficientes das funções das variáveis são praticamente os mesmos para os dois grupos, em todas as regiões, fato que indica que as variáveis independentes têm praticamente o mesmo impacto sobre os grupos de clientes adimplentes e clientes inadimplentes.

A seguir é apresentada a soma dos resultados da classificação da análise discriminante para os dois *clusters* da análise *K-Means*:

Tabela 17 - Resultados da classificação da análise discriminante para a soma dos resultados dos *cluster*s da análise *K-Means* para 50% dos clientes classificados como adimplentes e 50% classificados como inadimplentes

Resultado da Classificação - AD com segmentação			
K-Means	Adimplente	Inadimplente	Total
Adimplente	28252	11903	40155
Inadimplente	2960	6840	9800
Adimplente	**70,36%**	29,64%	100%
Inadimplente	30,20%	**69,80%**	100%

A função discriminante consegue classificar, de maneira correta, 70,36% dos clientes adimplentes. Para os clientes inadimplentes, 69,80% são classificados corretamente. O Erro

Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 29,64%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 30,20%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com a análise discriminante dos grupos formados pela análise *K-Means*, e constatou-se que 71,27% dos clientes adimplentes foram classificados de maneira correta, bem como 69,06% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto.

A Curva ROC do modelo de risco de crédito para os 2 *clusters* formados pela análise *K-Means* revela que a área sob a curva varia de 0,70 a 0,72, para os *clusters* 2 e 1, respectivamente. Segundo a escala anteriormente proposta, esse valor indica um poder discriminatório aceitável para o modelo.

Para a opção classificação pelo tamanho do grupo, verificou-se que a chance de um cliente ser classificado como adimplente é de 80,38%, ao passo que a chance de que tal cliente seja classificado como inadimplente é de 19,62%. Neste caso, a função discriminante consegue classificar de maneira correta 96,56% dos clientes adimplentes. Para os clientes inadimplentes, 15,15% são classificados corretamente. Logo, fica evidente que a capacidade de classificar corretamente os clientes adimplentes é maior para a classificação pelo tamanho do grupo, ao passo que a capacidade de classificar os clientes inadimplentes de maneira correta é maior para a opção todos os grupos iguais.

No entanto, considerando as porcentagens de erro obtidas, verifica-se que apesar de o Erro Tipo I, representado pelo clientes adimplentes, classificados como inadimplentes, ter sido de 3,44%; o Erro Tipo II, representado pela porcentagem de clientes inadimplentes classificados como adimplentes foi de 84,85%, fato que anula o ganho de acerto na classificação de clientes adimplentes, que se obteve com esse tipo de teste.

A seguir serão apresentados os resultados da análise discriminante dos grupos formados pela análise de *clusters TwoStep Cluster.*

5.6 ANÁLISE DISCRIMINANTE DOS GRUPOS FORMADOS PELA ANÁLISE *TWOSTEP CLUSTER*

As variáveis categóricas filial, estado civil, sexo, tipo de telefone residencial, tipo de telefone comercial, profissão, CEP e idade; e as variáveis contínuas data da venda, renda, SPC, valor financiado, quantidade de prestações, valor das prestações, valor da entrada, prestações a vencer, máxima prestação a vencer, número de compras, maior atraso e prestação sobre a renda, foram utilizadas para rodar a analise de *cluster TwoStep Cluster* (TSC).

O número de *clusters* gerado automaticamente pela análise TSC foi dois.

A medida de distância *Log-Likelihood* determina uma distribuição de probabilidades para as variáveis. Variáveis contínuas são consideradas normalmente distribuídas, enquanto que as variáveis categóricas são consideradas multinominais. Todas as variáveis são consideradas independentes. A medida de distância Euclidiana só pode ser utilizada quando todas as variáveis são contínuas. Para este trabalho foi utilizada a medida de distância *Log-Likelihood.*

Segundo Pindyck e Rubinfeld (2004, p. 263), em modelos de série temporal pode passar um período de tempo substancial entre o período em que são tomadas decisões de política econômica e o impacto dessas mudanças na variável de política econômica. Se o período entre decisão e resposta é suficientemente longo, variáveis explanatórias defasadas deveriam ser incluídas no modelo.

Os autores afirmam que muitas vezes não se pode saber apenas com base na teoria quantas defasagens incluir na equação. Então é preciso olhar os dados para determinar o número de defasagens “correto”. Uma abordagem é usar o R^2 corrigido, que acrescenta defasagens adicionais até que o R^2 corrigido pare de aumentar, para determinar quantas defasagens devem ser adicionadas. O R^2corrigido mede a porcentagem da variância na variável dependente (diferente da variação) explicada pelas variáveis explanatórias.

Ainda de acordo com os mesmos autores, outra abordagem possível seria o uso do critério de informação de Akaike (AIC), que difere do R^2 corrigido porque penaliza bem mais

a adição de variáveis do lado direito da equação (que reduz o número de graus de liberdade). Em princípio, poderia ser selecionada uma estrutura de defasagem pelo aumento do número de defasagens até o ponto em que o AIC atinja o valor mínimo. Outra estatística relacionada com o AIC é o critério de Schwartz (SC) ou critério de informação Baynesiano (BIC). Este critério igualmente penaliza a adição de variáveis do lado direito mais fortemente do que o R^2 corrigido. As três estatísticas fornecem informações que, combinadas com bom senso, podem ajudar a determinar a especificação de uma estrutura de defasagem.

Sabendo-se que o critério de seleção dos *clusters* determina como será a seleção automática da quantidade de *clusters*, e dadas as definições dos critérios AIC e BIC, entende-se que ambos são muito parecidos e, por isso, optou-se pelo critério BIC, automaticamente selecionado pelo SPSS.

A distribuição dos clientes utilizada para rodar a análise discriminante dos dois grupos formados pela análise TSC está na Tabela 18:

Tabela 18 - Distribuição dos clientes por *cluster* da análise TSC

Clusters	**Qtde de Clientes**	**%**
Cluster 1	25.035	50,11%
Cluster 2	24.921	49,89%
Total	**49.956**	**100,00%**

A seguir serão apresentados os resultados da análise discriminante para cada um dos grupos formados, de acordo com a metodologia de análise de *cluster* não hierárquica *TwoStep Cluster*.

Foram utilizados os dados de 49.956 clientes (os outros 44 clientes foram considerados como *missing values*) para rodar a análise discriminante.

Clientes adimplentes apresentam mais idade, maior renda, assumem parcelas de maior valor, dão maiores valores de entrada e apresentam maior número de compras liquidadas do que os clientes inadimplentes. Já os clientes inadimplentes financiam valores maiores, em

mais parcelas, têm maior quantidade de parcelas, em aberto, no seu nome e apresentam maiores atrasos em pagamentos anteriores do que os clientes adimplentes.

O *score* médio apresenta-se menor para clientes inadimplentes para o *cluster* 1, assim, tal fato indica que, de acordo com o *score* calculado pela empresa, que é baseado em variáveis fornecidas pelo SPC, o *score* médio é menor do que o *score* médio do grupo apenas para clientes possivelmente inadimplentes. No caso do *cluster* 2, o *score* médio é menor do que o *score* médio do grupo apenas para clientes possivelmente adimplentes.

A proporção média da prestação em relação à renda é menor para clientes inadimplentes no *cluster* 1, ou seja, clientes que apresentam menor proporção média da prestação em relação à renda do que a proporção média do grupo tendem a ser maus pagadores. Já para o *cluster* 2, clientes que apresentam maior proporção média da prestação em relação à renda do que a proporção média do grupo tendem a ser maus pagadores.

Para os dois *clusters* estudados, os valores do lambda de Wilks (para cada variável) são muito próximos de 1, indicando que nenhuma das variáveis utilizadas no modelo apresenta grande diferença em relação às médias dos grupos.

Por meio da análise do lambda de Wilks (do grupo) procura-se verificar se as médias dos grupos adimplente e inadimplente são iguais, uma vez que quanto mais significativamente diferentes, melhor os grupos serão discriminados. Os testes para os dois grupos apresentam significância igual a 0,000, portanto, menor que 0,05, fato que indica que as médias entre os grupos são diferentes. No entanto, como o lambda de Wilks é alto, é possível afirmar que a essa diferença é pequena, conforme constatado anteriormente.

O teste de igualdade das médias á apresentado na tabela a seguir:

Tabela 19 - Teste de igualdade das médias para cada *cluster* da análise TSC

***Cluster* 1**		***Cluster* 2**	
Variáveis	**Sig**	**Variáveis**	**Sig**
Idade	**0,000**	Idade	**0,000**
Maior_Atraso	**0,000**	Renda	**0,000**
Max_Pret_a_Vencer	**0,000**	Valor_Financiado	**0,000**
N_Compras	**0,000**	Qtde_Prest	**0,000**
Prest_a_Vencer	**0,000**	Valor_Prest	**0,000**
Prest_Renda	**0,000**	Valor_Entrada	**0,000**
Renda	**0,000**	Prest_a_Vencer	**0,000**
SPC	**0,000**	Max_Pret_a_Vencer	**0,000**
Valor_Financiado	**0,000**	N_Compras	**0,000**
Valor_Prest	**0,000**	Maior_Atraso	**0,000**
Valor_Entrada	0,424	SPC	**0,025**
Qtde_Prest	0,597	Prest_Renda	0,241

Com relação à hipótese de igualdade das médias de grupo, as significâncias das variáveis que estão em negrito são menores que 0,05, fato que indica que a hipótese de igualdade da média de cada variável nos grupos é rejeitada. Assim, é possível concluir que não existe diferença significativa entre a média das variáveis valor de entrada e quantidade de prestações no grupo adimplente, e a média das mesmas variáveis no grupo inadimplente, para o *Cluster* 1 formado por meio da análise TSC; bem como não existe diferença significativa entre a média da variável prestação sobre a renda no grupo adimplente e a média da mesma variável no grupo inadimplente, para o *Cluster* 2 desta mesma análise.

Por meio da análise dos determinantes de log, conclui-se que a dispersão entre as variáveis que caracterizam o grupo dos clientes adimplentes é maior do que a dispersão entre as variáveis que caracterizam o grupo de clientes inadimplentes para os dois *clusters* estudados.

A significância do teste M de Box foi igual a 0,000 para os dois *clusters*, que por se tratar de valor inferior a 0,05, nos leva a concluir que as diferenças de dispersão observadas entre os grupos são significantes, ou seja, não há igualdade de dispersão entre as variáveis que caracterizam os clientes dos grupos adimplente e inadimplente e a média desses grupos. O teste M de Box confirma a conclusão que se tirou analisando os determinantes de log.

O valor próprio calculado para os dois grupos representa uma medida relativa da diferença entre os grupos na função discriminante, que sendo de 0,140 para o *Cluster* 1 e 0,166 para o *Cluster* 2, indica que a diferença entre os grupos adimplente e inadimplente na função discriminante é pequena.

Para todas as regiões estudadas, o resultado é de apenas uma função, correspondendo a 100% da variância explicada em termos de diferenças entre grupos. Quanto à correlação canônica, que demonstra o nível de associação entre os escores discriminantes e os grupos, os valores são de 0,351 para o *Cluster* 1 e 0,377 para o *Cluster* 2, representando para cada região quanto o modelo explica da variável dependente.

As variáveis que mais afetam a variável dependente *status* adimplente/inadimplente são estado civil, sexo, tipo de telefone residencial e comercial, SPC e quantidade de prestações, tanto normalizadas quanto não normalizadas; e as variáveis valor financiado, prestações a vencer, máxima prestação a vencer e maior atraso, normalizadas.

As funções dos centróides do grupo inadimplente é maior nos dois *clusters* analisados, então, pode-se afirmar que o grupo inadimplente impacta mais a função discriminante do que o grupo adimplente, quando as médias das variáveis são consideradas para compor a equação discriminante.

A tabela a seguir representa os maiores coeficientes de correlação por *cluster*:

Tabela 20 - Maiores coeficientes de correlação por *cluster* da análise TSC

	Clusters	
Coeficientes de Correlação	**1**	**2**
CEP x filial	x	x
Valor prest x valor financiado	x	x
Max prest a vencer x prest a vencer	x	x
Valor prest x Prest s/ renda		x

Os coeficientes de correlação CEP com filial, valor da prestação com valor financiado e máxima prestação a vencer com prestações a vencer são os maiores coeficientes, com maior frequência.

As variáveis que mais contribuem para a função discriminante são quantidade de prestações, máxima prestação a vencer, prestações a vencer, valor financiado, tipo de telefone residencial, sexo, estado civil, maior atraso, filial, tipo de telefone comercial e CEP.

Os coeficientes das funções das variáveis são praticamente os mesmos para os dois grupos, nos dois *clusters*, fato que indica que as variáveis independentes têm praticamente o mesmo impacto sobre os grupos de clientes adimplentes e clientes inadimplentes.

A seguir é apresentada a soma dos resultados da classificação da análise discriminante para os dois *clusters* da análise TSC:

Tabela 21- Resultados da classificação da análise discriminante para a soma dos resultados dos *cluster*s da análise TSC para 50% dos clientes classificados como adimplentes e 50% classificados como inadimplentes

Resultado da Classificação - AD com segmentação			
TSC	Adimplente	Inadimplente	Total
Adimplente	28189	11967	40156
Inadimplente	2907	6893	9800
Adimplente	**70,20%**	29,80%	100%
Inadimplente	29,66%	**70,34%**	100%

A função discriminante consegue classificar de maneira correta 70,20% dos clientes adimplentes. Para os clientes inadimplentes, 70,34% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 29,80%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 29,66%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com a análise discriminante dos grupos formados pela análise *TwoStep Cluster*, e constatou-se que 70,54% dos clientes adimplentes foram classificados de maneira correta, bem como 69,82% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto.

A Curva ROC do modelo de risco de crédito para os 2 *clusters* formados pela análise *TwoStep Cluster* revela que a área sob a curva varia de 0,70 a 0,71, para os *clusters* 1 e 2,

respectivamente. Segundo a escala anteriormente proposta, esse valor indica um poder discriminatório aceitável para o modelo.

Para a opção classificação pelo tamanho do grupo, verificou-se que a chance de um cliente ser classificado como adimplente é de 80,38%, ao passo que a chance de que tal cliente seja classificado como inadimplente é de 19,62%. Neste caso, a função discriminante consegue classificar de maneira correta 96,48% dos clientes adimplentes. Para os clientes inadimplentes, 15,76% são classificados corretamente. Logo, fica evidente que a capacidade de classificar corretamente os clientes adimplentes é maior para a classificação pelo tamanho do grupo, ao passo que a capacidade de classificar os clientes inadimplentes de maneira correta é maior para a opção todos os grupos iguais.

No entanto, considerando as porcentagens de erro obtidas, verifica-se que apesar de o Erro Tipo I, representado pelo clientes adimplentes, classificados como inadimplentes, ter sido de 3,52%; o Erro Tipo II, representado pela porcentagem de clientes inadimplentes classificados como adimplentes foi de 84,24%, fato que anula o ganho de acerto na classificação de clientes adimplentes, que se obteve com esse tipo de teste.

O próximo passo será rodar as redes neurais para a amostra global, para os grupos formados pela segmentação em micro-regiões, pela segmentação em macro-regiões, pela segmentação pela análise de *clusters K-Means* e pela *TwoStep Cluster*.

5.7 REDES NEURAIS PARA A AMOSTRA GLOBAL

Foram rodadas redes neurais para a amostra global, para os grupos formados pela segmentação em micro-regiões, pela segmentação em macro-regiões, pela segmentação pela análise de *clusters K-Means* e pela análise *TwoStep Cluster*.

Utilizou-se a ferramenta *Intelligent Problem Solver* para escolher qual a melhor rede, entre os tipos Linear, *Radial Basis Function* e *Three layer perceptron*. Tais tipos de rede já foram explicados anteriormente. Em todos os casos a rede *Three layer perceptron*, considerada como uma *Multilayer Perceptron*, foi a que apresentou melhor desempenho. O algoritmo utilizado foi o de retropropagação, já definido e explicado no Capítulo 3.

Foram utilizados 50% dos dados na fase de treinamento da rede, 25% na fase de validação e 25% na fase de teste.

No caso da amostra global, a melhor rede foi uma MLP com 16 sinais de entrada e 14 neurônios na camada oculta. O acerto na fase de treinamento foi de 77,66%, na fase de validação, foi de 70,52% e na fase de teste de 69,57%.

O resultado da classificação dos clientes para a amostra global é apresentado na tabela 22:

Tabela 22 - Resultados da classificação das redes neurais para a amostra global

Resultado da Classificação - RN sem segmentação			
Am Global	Adimplente	Inadimplente	Total[8]
Adimplente	29582	10595	40177
Inadimplente	2478	7344	9822
Adimplente	**73,63%**	26,37%	100%
Inadimplente	25,23%	**74,77%**	100%

A rede neural consegue classificar de maneira correta 73,63% dos clientes adimplentes. Para os clientes inadimplentes, 74,77% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 26,37%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 25,23%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com as redes neurais rodadas para a amostra global, e constatou-se que 74,40% dos clientes adimplentes foram classificados de maneira correta, bem como 78,89% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto, inclusive porque as porcentagens de acerto do teste são maiores que os resultados obtidos pelo modelo.

[8] As redes neurais foram rodadas no *software* Statistica, portanto, contam com os 50.000 clientes (sem *missing values*).

5.8 REDES NEURAIS PARA AS 21 MICRO-REGIÕES

Para rodar as redes neurais para as 21 micro-regiões foram utilizados 50% dos dados na fase de treinamento da rede, 25% na fase de validação e 25% na fase de teste.

A soma dos resultados da classificação das redes neurais para as 21 regiões é representada na Tabela 23:

Tabela 23 - Resultados da classificação das redes neurais para a soma dos resultados das 21 micro-regiões

Resultado da Classificação - RN com segmentação			
21 regiões	Adimplente	Inadimplente	Total
Adimplente	30715	9463	40178
Inadimplente	2293	7529	9822
Adimplente	**76,45%**	23,55%	100%
Inadimplente	23,35%	**76,65%**	100%

A rede neural consegue classificar de maneira correta 76,45% dos clientes adimplentes. Para os clientes inadimplentes, 76,65% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 23,55%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 23,35%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com as redes neurais rodadas para as 21 micro-regiões, e constatou-se que 78,31% dos clientes adimplentes foram classificados de maneira correta, bem como 78,56% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto, inclusive porque as porcentagens de acerto do teste são maiores que os resultados obtidos pelo modelo.

5.9 REDES NEURAIS PARA AS MACRO-REGIÕES

Foram utilizados 50% dos dados na fase de treinamento da rede, 25% na fase de validação e 25% na fase de teste, para rodar as redes neurais para as 3 macro-regiões.

A soma dos resultados da classificação das redes neurais para as 3 macro-regiões é representada na Tabela 24:

Tabela 24 - Resultados da classificação das redes neurais para a soma dos resultados das 3 macro-regiões

Resultado da Classificação - RN com segmentação			
3 regiões	Adimplente	Inadimplente	Total
Adimplente	29.841	10.336	40.177
Inadimplente	2.573	7.249	9.822
Adimplente	**74,27%**	25,73%	100%
Inadimplente	26,20%	**73,80%**	100%

A rede neural consegue classificar de maneira correta 74,27% dos clientes adimplentes. Para os clientes inadimplentes, 73,80% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 25,73%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 26,20%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com as redes neurais rodadas para as 3 macro-regiões, e constatou-se que 74,08% dos clientes adimplentes foram classificados de maneira correta, bem como 73,44% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto.

5.10 REDES NEURAIS PARA ANÁLISE *K-MEANS*

Para rodar as redes neurais para os dois *clusters* formados pela análise *K-Means* foram utilizados 50% dos dados na fase de treinamento da rede, 25% na fase de validação e 25% na fase de teste.

A soma dos resultados da classificação das redes neurais para os dois *clusters* formados é representada na Tabela 25:

Tabela 25 - Resultados da classificação das redes neurais para a soma dos resultados dos *cluster*s da análise *K-Means*

Resultado da Classificação – RN com segmentação			
K-Means	Adimplente	Inadimplente	Total
Adimplente	30.326	9.829	40.155
Inadimplente	2.536	7.264	9.800
Adimplente	**75,52%**	24,48%	100%
Inadimplente	25,88%	**74,12%**	100%

A rede neural consegue classificar de maneira correta 75,52% dos clientes adimplentes. Para os clientes inadimplentes, 74,12% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 24,48%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 25,88%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com as redes neurais rodadas para análise *K-Means*, e constatou-se que 75,84% dos clientes adimplentes foram classificados de maneira correta, bem como 73,98% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto.

5.11 REDES NEURAIS PARA ANÁLISE *TWOSTEP CLUSTER*

Para rodar as redes neurais para os dois *clusters* formados pela análise *TwoStep Cluster* foram utilizados 50% dos dados na fase de treinamento da rede, 25% na fase de validação e 25% na fase de teste.

A soma dos resultados da classificação das redes neurais para os dois *clusters* é representada na Tabela 26:

Tabela 26 - Resultados da classificação das redes neurais para a soma dos resultados dos *cluster*s da análise *TwoStep Cluster* (TSC)

Resultado da Classificação - RN com segmentação			
TSC	Adimplente	Inadimplente	Total
Adimplente	28.295	11.860	40.155
Inadimplente	2.910	6.890	9.800
Adimplente	**70,46%**	29,54%	100%
Inadimplente	29,69%	**70,31%**	100%

A rede neural consegue classificar de maneira correta 70,46% dos clientes adimplentes. Para os clientes inadimplentes, 70,31% são classificados corretamente. O Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de 29,54%; ao passo que, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes, classificados como adimplentes é de 29,69%.

Testou-se, aleatoriamente, por meio do Excel, 5% (últimos 2.500 clientes) dos resultados obtidos com as redes neurais rodadas para análise *TwoStep Cluster*, e constatou-se que 71,88% dos clientes adimplentes foram classificados de maneira correta, bem como 73,10% dos clientes inadimplentes, fato que comprova a eficiência do modelo proposto.

5.12 RESUMO

A seguir será apresentada uma tabela com os resultados da análise discriminante e das redes neurais para a amostra global, as 21 micro-regiões, as 3 macro-regiões, a análise de *clusters K-Means* e a TSC.

Tabela 27 - Resultados da classificação da análise discriminante e das redes neurais

Resultado da Classificação - AD sem segmentação			
Am Global	Adimplente	Inadimplente	Total
Adimplente	28176	11980	40156
Inadimplente	2950	6850	9800
Adimplente	**70,17%**	29,80%	100%
Inadimplente	30,10%	**69,90%**	100%

Resultado da Classificação - RN sem segmentação			
Am Global	Adimplente	Inadimplente	Total
Adimplente	29582	10595	40177
Inadimplente	2478	7344	9822
Adimplente	**73,63%**	26,37%	100%
Inadimplente	25,23%	**74,77%**	100%

Resultado da Classificação - AD com segmentação			
21 regiões	Adimplente	Inadimplente	Total
Adimplente	28665	11490	40155
Inadimplente	2910	6890	9800
Adimplente	**71,39%**	28,61%	100%
Inadimplente	29,69%	**70,31%**	100%

Resultado da Classificação - RN com segmentação			
21 regiões	Adimplente	Inadimplente	Total
Adimplente	30715	9463	40178
Inadimplente	2293	7529	9822
Adimplente	**76,45%**	23,55%	100%
Inadimplente	23,35%	**76,65%**	100%

Resultado da Classificação - AD com segmentação			
3 regiões	Adimplente	Inadimplente	Total
Adimplente	28295	11861	40155
Inadimplente	2936	6864	9800
Adimplente	**70,46%**	29,54%	100%
Inadimplente	29,96%	**70,04%**	100%

Resultado da Classificação - RN com segmentação			
3 regiões	Adimplente	Inadimplente	Total
Adimplente	29841	10336	40177
Inadimplente	2573	7249	9822
Adimplente	**74,27%**	25,73%	100%
Inadimplente	26,20%	**73,80%**	100%

Resultado da Classificação - AD com segmentação			
K-Means	Adimplente	Inadimplente	Total
Adimplente	28252	11903	40155
Inadimplente	2960	6840	9800
Adimplente	**70,36%**	29,64%	100%
Inadimplente	30,20%	**69,80%**	100%

Resultado da Classificação – RN com segmentação			
K-Means	Adimplente	Inadimplente	Total
Adimplente	30326	9829	40155
Inadimplente	2536	7264	9800
Adimplente	**75,52%**	24,48%	100%
Inadimplente	25,88%	**74,12%**	100%

Resultado da Classificação - AD com segmentação			
TSC	Adimplente	Inadimplente	Total
Adimplente	28189	11967	40156
Inadimplente	2907	6893	9800
Adimplente	**70,20%**	29,80%	100%
Inadimplente	29,66%	**70,34%**	100%

Resultado da Classificação - RN com segmentação			
TSC	Adimplente	Inadimplente	Total
Adimplente	28295	11860	40155
Inadimplente	2910	6890	9800
Adimplente	**70,46%**	29,54%	100%
Inadimplente	29,69%	**70,31%**	100%

Com exceção da classificação dos clientes inadimplentes para a análise de *clusters K-Means*, combinada com a análise discriminante, em todas as outras análises o resultado da classificação da análise discriminante foi melhor após a segregação dos dados, com destaque para a segmentação das 21 regiões, que apresentou maior acerto total.

Em todos os casos, o Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, foi de aproximadamente 30%, bem como o Erro Tipo II.

Com exceção da classificação dos clientes adimplentes para a análise de *clusters Two Step Cluster,* combinada com redes neurais, em todas as outras análises o resultado da

classificação dos clientes adimplentes pelas redes neurais foi melhor após a segregação dos dados, com destaque para a segmentação das 21 regiões, que apresentou maior acerto total. Já em com relação aos clientes inadimplentes, apenas a segmentação de dados pelas 21 regiões apresentou maiores resultados do que a classificação da amostra global.

Em todos os casos, o Erro Tipo I, representado pela porcentagem de clientes adimplentes, classificados como inadimplentes, ficou entre 23% e 30%, bem como o Erro Tipo II.

Em termos de acerto total (soma dos acertos de clientes adimplentes e inadimplentes), a classificação pela análise discriminante, para a amostra global, foi de 70,11%, ao passo que o acerto para as 21 regiões foi de 71,17%, para as 3 regiões 70,38%, para a análise *K-Means* 70,25% e para a *TwoStep Cluster* 70,23%.

Também em termos de acerto total, agora para as redes neurais, para a amostra global, foi de 73,85%, ao passo que o acerto para as 21 regiões foi de 76,49%, para as 3 regiões 74,18%, para a análise *K-Means* 75,25% e para a *TwoStep Cluster* 70,43%.

Todas as porcentagens de acerto total foram maiores para a análise discriminante combinada com a segmentação de dados, quando comparadas com as porcentagens de acerto da análise discriminante para a amostra global.

Com exceção da rede neural combinada com a análise *TwoStep Cluster*, todas as outras porcentagens de acerto total foram maiores para as redes neurais combinadas com a segmentação de dados.

Comparando-se os dois modelos, é possível afirmar que em todos os casos o acerto total de classificação foi maior para as redes neurais, combinadas ou não com a segmentação de dados, em relação à análise discriminante.

Os maiores percentuais de acerto total na classificação de clientes como adimplentes ou inadimplentes foram aqueles obtidos por meio da análise discriminante ou das redes neurais, combinadas com a segmentação de dados manual, pela segmentação dos dados com base na posição geográfica das filias de vendas, formando-se 21 regiões. Os piores

percentuais de acerto de classificação obtidos foram para a análise de *clusters TwoStep Cluster*, combinada com análise discriminante, ou com redes neurais. Logo, é possível afirmar que este modelo não se adequada à proposta deste trabalho, devendo, portanto, ser descartado.

O melhor desempenho do modelo que combina a segregação dos dados em 21 regiões, tanto para a análise discriminante, quanto para as redes neurais, pode ser explicado pelo fato de que quanto maior for o número de grupos formados, menor será a variância entre seus componentes, permitindo resultados mais eficientes. Como a análise por macro-regiões formou apenas 3 grupos e as análises por meio de técnicas de análise multivariada apenas 2, a redução da variância foi pequena, implicando menores ganhos de porcentagens de acerto de classificação dos clientes.

Pela análise discriminante, para a opção classificação pelo tamanho do grupo, verificou-se que, em todos os casos, a chance de um cliente ser classificado como adimplente é de 80,38%, ao passo que a chance de que tal cliente seja classificado como inadimplente é de 19,62%. No entanto, considerando as porcentagens de erro obtidas, verifica-se que apesar de o Erro Tipo I, representado pelo clientes adimplentes, classificados como inadimplentes, ter se mantido pequeno, o Erro Tipo II, representado pela porcentagem de clientes inadimplentes classificados como adimplentes foi sempre muito grande, fato que anulou o ganho de acerto na classificação de clientes adimplentes, que se obteve com esse tipo de teste.

6. CONCLUSÕES

O perfil dos clientes que compõem a base de dados analisada neste trabalho, tanto de adimplentes, quanto de inadimplentes, é composto em sua maioria por pessoas do sexo feminino, casadas, com telefone comercial e residencial fixo, que trabalham com serviços ou outra profissão que não o comércio, o funcionalismo público, como autônomo ou aposentado.

Ademais, clientes adimplentes apresentam mais idade, maior renda, assumem parcelas de maior valor, dão maiores valores de entrada e apresentam maior número de compras liquidadas, do que os clientes inadimplentes. Já os clientes inadimplentes financiam valores maiores, em mais parcelas, têm ou já tiveram maior quantidade de parcelas em aberto no seu nome e apresentam maiores atrasos em pagamentos anteriores do que os clientes adimplentes.

Dentre todos esses atributos, as variáveis que mais contribuem para a função discriminante são estado civil, sexo, tipo de telefone residencial e comercial, SPC e quantidade de prestações.

Com relação ao desempenho dos modelos propostos, foram encontrados os seguintes resultados:

1. Com exceção da classificação dos clientes inadimplentes para a análise de *clusters K-Means*, em todas as outras análises o resultado da classificação da análise discriminante foi melhor após a segregação dos dados;
2. Em todas as análises propostas o **acerto total** da classificação da análise discriminante foi melhor após a segregação dos dados, com destaque para a segmentação das 21 regiões;
3. Com exceção da classificação dos clientes adimplentes para a análise de *clusters Two Step Cluster*, combinada com redes neurais, em todas as outras análises o resultado da classificação dos clientes adimplentes pelas redes neurais foi melhor após a segregação dos dados. Já em com relação aos clientes inadimplentes, apenas a segmentação de dados pelas 21 regiões apresentou maiores resultados do que a classificação da amostra global.

4. Com exceção da rede neural combinada com a análise *TwoStep Cluster*, todas as outras porcentagens de **acerto total** foram maiores para as redes neurais combinadas com a segmentação de dados.
5. Comparando-se os dois modelos, é possível afirmar que em todos os casos o **acerto total** de classificação foi maior para as redes neurais, combinadas ou não com a segmentação de dados, em relação à análise discriminante.

Em todos os cenários analisados, a dispersão entre os atributos dos clientes adimplentes mostrou-se maior do que a dispersão entre os atributos dos clientes inadimplentes. Assim, o ganho na classificação correta de clientes inadimplentes, que se espera obter com a utilização de modelos híbridos, deverá ser menor que o ganho na classificação correta de clientes adimplentes. Considerando que a proposta do presente estudo é aplicar a segmentação de dados para reduzir a variância dentro dos grupos, grupos com menores variâncias apresentarão menores reduções e, portanto, menores ganhos na classificação correta dos clientes.

A diferença entre a classificação correta dos clientes inadimplentes pela análise *K-Means* foi 0,1% menor que a classificação correta desses clientes para a amostra global, assim, considerando esta diferença como irrelevante, não nos atentaremos em analisá-la mais profundamente. Além disso, a porcentagem de acerto total foi maior para a análise *K-Means* (70,25%) do que para a amostra global (70,11%).

O acerto total de classificação foi maior para as redes neurais combinadas com a segmentação de dados para as 21 regiões, para as 3 regiões e para a análise *K-Means.* Os piores percentuais de acerto de classificação obtidos foram para a análise de *clusters TwoStep Cluster*, combinada com análise discriminante, ou com redes neurais. Logo, é possível afirmar que este modelo não se adequada à proposta deste trabalho, devendo, portanto, ser descartado.

Em todos os casos, os acertos totais obtidos com as redes neurais foram superiores aos acertos obtidos com a análise discriminante.

Dentre os quatro modelos propostos, o que apresentou melhores resultados de classificação foi o modelo híbrido que combina a segmentação manual dos dados com análise discriminante e com redes neurais, formando-se 21 micro-regiões. Tal fato se explica pela considerável redução da variância dentro dos grupos que a formação de um número maior de agrupamentos pode oferecer.

Considera-se os resultados desta pesquisa satisfatórios, com base nos trechos a seguir.

O trabalho de Hill, O`Conner e Remus (1996) compara o desempenho das redes neurais com o desempenho de outros modelos preditivos disponíveis na literatura. Os resultados indicaram que as redes neurais apresentam desempenho tão bom quanto, e às vezes superior, aos outros modelos estatísticos. Condições como linearidade e volatilidade (variância) dos dados analisados influenciam, diretamente, o desempenho das redes, afinal, modelos lineares podem apresentar melhores resultados de previsão para dados lineares, bem como as redes neurais fazem com dados não-lineares. Quanto à volatilidade, observa-se que as redes neurais apresentam melhores resultados quando os dados são menos voláteis.

Já Hand e Henley (1997) desenvolveram um trabalho para estudar parte das técnicas de análise de crédito disponíveis. Os autores afirmam que, em geral, não existe um método que seja superior aos outros. A escolha do modelo que será utilizado para resolver um problema depende das características apresentadas por este problema. Os autores realizaram um estudo comparativo entre a análise discriminante e as redes neurais. Os resultados das classificações corretas foram 81% para a análise discriminante e 88% para as redes neurais, comprovando a superioridade do desempenho das redes neurais.

Picinini et al. (2003) consideram enfim que modelos de *credit scoring*, com taxas de acerto acima de 65%, são considerados bons por especialistas. Em todos os casos estudados, as porcentagens de classificação estão muito próximas ou acima de 70%. Logo, tanto o Erro Tipo I, quanto o Erro Tipo II, aparecem sempre próximos de 30%, fato que explica o fato de algumas vezes a análise ter sido feita com base no acerto total dos modelos, ou seja, tanto isolados como acumulados estes erros estão próximos do mesmo valor, não ocasionando viés na análise.

Justifica-se, assim, tanto o desempenho dos modelos propostos na classificação de clientes, quanto os ganhos obtidos nesta classificação, por meio da aplicação de modelos híbridos.

O modelo proposto por este estudo tem caráter preditivo, ou seja, com base nos atributos dos clientes que foram disponibilizados para a análise, pretende-se prever o comportamento de cada cliente, como adimplente ou inadimplente. Por meio desta previsão, toma-se a decisão de conceder ou não o crédito.

Dentre as possíveis explicações para a existência de vendas a prazo estão as restrições de acesso ao mercado de financeiro, uma vez que o custo do financiamento e a quantidade de recursos disponíveis poderiam, de alguma forma, limitar o consumo; e a utilização da concessão de crédito como uma estratégia, já que quanto maior sua oferta, maior o consumo por impulso.

Assim, dado o aumento considerável do volume de crédito oferecido aos consumidores nos últimos anos, bem como o aumento dos riscos assumidos pelas empresas concedentes, entre eles, o risco de inadimplência, o desenvolvimentos de modelos como os propostos por este trabalho são de fundamental importância para a sobrevivência e manutenção da saúde financeira destas empresas. Entre as vantagens apresentadas pela adoção de modelos como estes estão o baixo custo de implantação e manutenção e a necessidade de uma equipe reduzida para operar os *softwares.*

A empresa concedente do banco de dados forneceu também a classificação real do cliente, como adimplente ou inadimplente. Os resultados obtidos por este estudo foram comparados com esta informação, de maneira que as porcentagens de acerto de classificação do modelo aqui proposto pudessem ser verificadas. A partir da validação do modelo, novos casos podem ser incluídos e, para cada um deles, será feita a previsão do comportamento do cliente, ferramenta que servirá para o processo decisório de concessão de crédito.

Entre as contribuições desta pesquisa estão a comparação de quatro modelos que combinam segmentação de dados com uma técnica de análise multivariada, a análise discriminante, e com uma técnica não linear, as redes neurais, com o objetivo de obter ganhos na classificação de clientes como adimplentes ou inadimplentes, informação útil para a

decisão de concessão de crédito ao consumidor. A proposta que se faz é a de reduzir a variância, dentro dos grupos analisados, maximizando as classificações corretas que os modelos realizam. Assim, além de se ter confirmado que é necessário reduzir a variância para se obter melhores resultados, ainda se constatou que a formação de poucos grupos, como dois ou três, não é eficiente para grandes amostras como a que foi utilizada. Por fim, foi possível comparar os resultados de dois métodos diferentes de análise de *clusters*, os métodos não-hierárquicos *K-Means* e *TwoStep Cluster*, bem como dois métodos distintos de análise discriminante.

Fica, para pesquisas futuras, a sugestão de rodar os modelos propostos para uma base de dados menor, bem como trabalhar com a inclusão/subtração das variáveis estudadas e rodar a análise de *clusters K-Means* para um número maior de *clusters*.

BIBLIOGRAFIA

ALMEIDA, F. C. de; NAKAO, S. H. Redes neurais. In: CORRAR, L. J. ; PAULO, E. ; DIAS FILHO, J. M. **Análise multivariada para os cursos de administração, ciências contábeis e economia.** São Paulo: Atlas, 2007.p. 432-58.

ALMEIDA, P. H. et al. Utilização de algoritmo genético na parametrização de redes neurais artificiais para aplicação na elaboração de orçamento de vendas. In: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 28., 2004, Curitiba. **Anais**...Paraná: ANPAD, 2004. p. 1-15.

ASSAF NETO, A. **Finanças corporativas e valor**. 4. ed. São Paulo: Atlas, 2009.

BACEN. Banco Central do Brasil. Disponível em: < http://www.bcb.gov.br/>. Acesso em 24 de Maio de 2011.

BEUREN, I. M. **Como elaborar trabalhos monográficos em contabilidade**: teoria e prática. São Paulo: Atlas, 2006.

BRAZÃO, R. L.; BARBETTA, P. A.; ANDRADE, D. F. *TwoStep Cluster*: análise comparativa do algoritmo e proposta de melhoramento da medida de verossimilhança. In: WORKSHOP EM ALGORITMOS E APLICAÇÕES DE MINERAÇÃO DE DADOS, 3., 2007, João Pessoa. **Anais**...Paraíba: WAAMD, 2007. p. 1-8.

BRESSAN, A. A. Tomada de decisão em mercados futuros agropecuários utilizando modelos de previsão de séries temporais. **RAE eletrônica**, São Paulo, v. 3, n. 1, p. 1-20, 2004.

BRITO, G. A. S.; NETO, A. A. Modelo de classificação de risco de crédito de grandes empresas. **R. Cont. Fin. USP**, São Paulo, v. 19, n. 46, p. 18 – 29, 2008.

BROOMHEAD, D.S.; LOWE, D. Multivariable functional interpolation and adaptive networks. **Complex Systems**, Champaign, v. 2, n.3, p.321-55, 1988.

BRUNI, A. L.; FAMA, R.; MURRAY, A. D. Modelos brasileiros preditivos de risco de crédito: um estudo exploratório atual sobre as suas eficácias. **Periódico Tema**, n. 32, p. 148-67, 1998.

CORRÊA, M. F.; MACHADO, M. A. S. Construção de um modelo de credit scoring em redes neurais para previsão de inadimplência na concessão de microcrédito. In: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 28., 2004, Curitiba. **Anais**...Paraná: ANPAD, 2004.

COSIF. Plano Contábil das Instituições Financeiras do Sistema Financeiro Nacional. Disponível em: < http://www.cosif.com.br/default.asp>. Acesso em 24 de Maio de 2011.

DAMODARAN, A**. Finanças corporativas**: teoria e prática. Tradução de Jorge Ritter. 2. ed. Porto Alegre: Bookman, 2004.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L. da. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Campus, 2009.

FONSECA, F. R.; OMAKI, E. T. Redes neurais artificiais e segmentação psicográfica em marketing: um ensaio sobre a aplicação de RNAs para segmentar os clientes do mercado industrial baseado no risco percebido da compra. In: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓSGRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 28., 2004, Curitiba. **Anais**... Paraná: ANPAD, 2004.

GONÇALVES, E. B. **Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéticos.** 2005. 105 f. Dissertação (Mestrado em Administração) – Faculdade de Administração, Economia e Contabilidade, Universidade de São Paulo, São Paulo, 2005.

HAIR, J. et al. **Análise multivariada de dados**. Tradução de Lene Belon Ribeiro. 5. ed. Porto Alegre: Bookman, 2005.

HAND D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. **Journal of the Royal Statistical Society Series A (Statistics in Society),** Reino Unido, v. 160, n. 3, p. 523-41, 1997.

HAYKIN, S. **Redes neurais:** princípios e prática. Tradução de Paulo Martins Engel. 2. ed. Porto Alegre: Bookman, 2001.

HENLEY, W. E.; HAND, D. J. A k-nearest neighbour classifier for assessing consumer credit risk. **The Statistician**, Reino Unido, v. 45, n. 1, p. 77-95, set. 1996.

HILL, T.; O'CONNOR, M.; REMUS, W. Neural network models for time series forecasts. **Management Science**, Maryland, v. 42, n.7, p.1082-92, jul. 1996.

HUA, L.; SONG, Z.; LI, Y. Choosing credit risk index for listed companies based ond grey *cluster*ing method. In: PROCEEDINGS INTERNATIONAL SYMPOSIUM ON WEB INFORMATION SYSTEMS AND APPLICATIONS, 2., 2009, Nanchang. **Anais**…Jiangxi: WISA, 2009, p. 56-9.

HUBERMAN, L. **História da riqueza do homem**. 17. ed. Rio de Janeiro: Zahar Editores, 1981.

IBGE. Instituto Brasileiro de Geografia e Estatística. Disponível em: <http://www.ibge.gov.br/home/>. Acesso em 24 de Maio de 2011.

KARCHER, C. **Redes baynesianas aplicadas à análise do risco de crédito.** 2009. 103f. Dissertação (Mestrado em Engenharia) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2009.

LACHTERMACHER, G.; ESPENCHITT, D. G. Previsão de falência de empresas: estudo de generalização de redes neurais. In: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 25., 2001, Campinas. **Anais**... São Paulo: ANAPAD, 2001.

LEMOS, E. P.; STEINER, M. T. A.; NIEVOLA, J. C. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. **Revista de Administração da USP**, São Paulo, v. 40, n. 3, p. 225-34, jul./set. 2005.

LEVINE, D. M. et al. **Estatística**: teoria e aplicações. Rio de Janeiro: Editora LTC, 2008.

LI, B.; XU, C. An application of the combination of principal component analysis and BP neural network to credit assessment in the commercial banks. In: PROCEEDINGS INTERNATIONAL FORUM OF INFORMATION TECHNOLOGY AND APPLICATIONS, 1., 2009, Chengdu. **Anais**….Sichuan: IFITA, 2009, p. 426-29.

LIMA, F. G. et al. Aplicação de redes neurais na análise e na concessão de crédito ao consumidor. **Revista de Administração da USP**, São Paulo, v. 44, n.1, p. 34-45, mar. 2009.

_____ ; ALMEIDA, F. C. Previsão de séries temporais financeiras com o uso das wavelets. In: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 28., 2004, Curitiba. **Anais.**.. Paraná: ANPAD, 2004.

_____ ; PERERA, L. C. J. Técnicas de segmentação e redes neurais otimizando a análise de crédito ao consumidor. In: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 34., 2010, Rio de Janeiro. **Anais**... Rio de Janeiro: ANPAD, 2010.

LIN, F. et al. Time series forecasting with neural networks. **Complexity International**, Australia, v. 2, p. 1-12, abr. 1995.

PESTANA, M. H.; GAGEIRO, J. N. **Análise de dados para ciências sociais**: a complementaridade do SPSS. Lisboa: Sílabo, 2003.

PICININI, R. ; OLIVEIRA, G. M. B.; MONTEIRO, L. H. A. Mineração de critério de *credit scoring* utilizando algoritmos genéticos. In: SIMPÓSIO BRASILEIRO DE AUTOMAÇÃO INTELIGENTE, 6., 2003, Bauru. **Anais**...São Paulo: SBAI, 2003.

PINDYCK, R. S.; RUBINFELD, D. L. **Econometria:** modelos e previsões. 4. ed. Rio de Janeiro: Elsevier, 2004.

POHLMANN, M. C. Análise de conglomerados. In: CORRAR, L. J. ; PAULO, E. ; DIAS FILHO, J. M. **Análise multivariada para os cursos de administração, ciências contábeis e economia.** São Paulo: Atlas, 2007. p. 324-87.

ROSENBERG, E.; GLEIT, A. Quantitative methods in credit management: a survey. **Operations Research**, Oxford, v. 42, n. 4, p. 589-613, jul./ago. 1994.

SEBRAE. Serviço Brasileiro de Apoio às Micro e Pequenas Empresas. Disponível em: <http://www.sebrae.com.br/>. Acesso em 24 de Maio de 2011.

THOMAS, L.C. Consumer credit modeling: context and current issues. In: BANFF CREDIT RISK CONFERENCE, 1., 2003, Banff. **Anais**...Canadá: BANFF, 2003.

VASCONCELLOS, M. S. de. P**roposta de método para análise de concessão de crédito a pessoas físicas**. 2002. 142 f. Dissertação (Mestrado em Economia) - Faculdade de Administração, Economia e Contabilidade, Universidade de São Paulo, São Paulo, 2002.

ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks: the state of the art. **International Journal of Forecasting**, Kent (Ohio), v. 14, n. 1, p. 35-62, mar. 1998.