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

RODRIGO ALVES SILVA

Performance financeira da carteira na avaliação de modelos de análise e concessão de crédito:
uma abordagem baseada em aprendizagem estatística

ORIENTADOR: PROF. DR. EVANDRO
MARCOS SAIDEL RIBEIRO

Ribeirão Preto

2014

Prof. Dr. Marco Antônio Zago
Reitor da Universidade de São Paulo

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

Profa. Dra. Sonia Valle Walter Borges de Oliveira
Chefe do Departamento de Administração

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

RODRIGO ALVES SILVA

Performance financeira da carteira na avaliação de modelos de análise e concessão de crédito:
uma abordagem baseada em aprendizagem estatística

Tese apresentado à Faculdade de Economia,
Administração e Contabilidade de Ribeirão
Preto da Universidade de São Paulo, como
requisito para obtenção do título de Doutor em
Ciências. Versão Corrigida. Original encontra-
se disponível na FEA-RP/USP

ORIENTADOR: PROF. DR. EVANDRO
MARCOS SAIDEL RIBEIRO

Ribeirão Preto

2014

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.’

Silva, Rodrigo Alves

Performance financeira da carteira na avaliação de modelos de análise e concessão de crédito: Uma abordagem baseada em aprendizagem estatística. Ribeirão Preto, 2014.

126 p.: il. ; 30 cm

Tese de Doutorado, apresentada à Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto/USP. Área de concentração: Administração de Organizações.

Orientador: Ribeiro, Evandro Marco Saidel.

1. Risco de crédito. 2. Aprendizagem Estatística. 3. Classificadores. 4. Desempenho financeiro.

Nome: Rodrigo Alves Silva

Título: Performance financeira da carteira na avaliação de modelos de análise e concessão de crédito: Uma abordagem baseada em aprendizagem estatística.

Tese apresentado à Faculdade de Economia,
Administração e Contabilidade de Ribeirão Preto
da Universidade de São Paulo, para obtenção do
título de Doutor Ciências

Área de Concentração: Administração de
Organizações

Aprovada em:

Banca Examinadora

Prof. Dr. _____ Instituição: _____

Julgamento _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento _____ Assinatura: _____

**À minha esposa, Priscila Rodrigues Gomes,
esteio familiar e apoiadora incondicional do
caminho, por mim, trilhado até aqui.**

AGRADECIMENTOS

Agradeço primeiramente à minha mãe Maria J. Alves Silva e ao meu pai Edis de Souza Silva, exemplos de carinho, empenho, dedicação e suporte à minha formação pessoal e profissional, aos quais anseio honrar todos os dias de minha vida.

Às minhas avós Benedita Silva Santos (in memoriam) e Anália de Souza Silva, em nome das quais agradeço aos meus tios, tias e demais familiares, pela eterna dedicação para a constituição de uma família unida e próspera.

Ao meu irmão, Ricardo Alves Silva e à minha Irmã, Renata Alves Silva, pela compreensão de minha falta em momentos importantes de suas vidas e por saberem que sempre estive presente em pensamento.

Ao meu orientador, Prof. Dr. Evandro Marcos Saidel Ribeiro, pela sua brilhante e efetiva atuação como orientador deste projeto, bem como pela sua amizade e compreensão.

Ao meu grande amigo Prof. Dr. Anderson Ara pela sua influência e auxílio decisivo na compreensão e aprofundamento dos temas relacionados à aprendizagem estatística e pela disponibilização de seu algoritmo para cálculo de parte das redes utilizadas neste trabalho.

Ao Prof. Dr. Alberto Borges Matias pelas palavras de sabedoria sobre o tema e pelos inúmeros esclarecimentos de dúvidas em aula.

Aos professores Matheus C. Gerolamo (EESC-USP) e Daisy A. N. Rebelatto (EESC-USP) pela compreensão e apoio durante a execução das disciplinas.

Aos colaboradores do PPGAO Érika L. V. Costa, Matheus C. Costa e Thiago S. C. Sasso pelas orientações e pelo ótimo trabalho de gestão dos aspectos administrativo-operacionais do departamento.

Aos meus amigos de ontem, hoje e sempre pela torcida e pelas palavras de carinho.

À Universidade de São Paulo por possibilitar minha formação e meu crescimento pessoal e profissional.

**“O homem que não tem os olhos abertos
para o misterioso passará pela vida sem ver
nada”**

Albert Einstein

RESUMO

SILVA, R. A. **Performance financeira da carteira na avaliação de modelos de análise e concessão de crédito: uma abordagem baseada em aprendizagem estatística**. 2014. 126 f. Tese (doutorado) – Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2014.

Os modelos de análise e decisão de concessão de crédito buscam associar o perfil do tomador de crédito à probabilidade do não pagamento de obrigações contraídas, identificando assim o risco associado ao tomador e auxiliando a firma a decidir pela aprovação ou negação da solicitação de crédito. Atualmente este campo de pesquisa tem ganhado importância no cenário nacional – pela intensificação da atividade de crédito no país com grande participação dos bancos públicos neste processo – e internacional – pelo aumento das preocupações com potenciais danos à economia derivados de eventos de *default*. Tal quadro fez com que fossem construídos e adaptados diversos modelos e métodos à análise de risco de crédito tanto para consumidores como para empresas. Estes modelos são testados e comparados com base na acurácia de previsão ou de métricas de otimização estatística. Este é um procedimento que pode não se mostrar eficiente do ponto de vista financeiro, ao mesmo tempo em que dificulta a interpretação e tomada de decisão por parte da firma quanto a qual o melhor modelo, gerando uma lacuna pelo desprendimento observado entre a decisão de qual o modelo a ser adotado e o objetivo financeiro da empresa. Tendo em vista que o desempenho financeiro é um dos principais indicadores de qualquer procedimento gerencial, o presente estudo objetivou preencher a esta lacuna analisando o desempenho financeiro de carteiras de crédito formadas por técnicas de aprendizagem estatística utilizadas atualmente na classificação e análise de risco de crédito em pesquisas nacionais e internacionais. A pesquisa selecionou as técnicas: análise discriminante, regressão logística, redes bayesianas Naïve Bayes, *k*dB-1, *k*dB-2, SVC e SVM e aplicou tais técnicas junto à base de dados *German Credit Data Set*. Os resultados foram analisados e comparados inicialmente em termos de acurácia e custos por erro de classificação. Adicionalmente a pesquisa propôs o emprego de quatro métricas financeiras (RFC, PLR, RAROC e IS), encontrando variações quanto aos resultados produzidos por cada técnica. Estes resultados sugerem variações quanto a sequência de eficiência e consequentemente de emprego das técnicas, demonstrando a importância da consideração destas métricas para a análise e decisão de seleção de modelos de classificação ótimos.

Palavras-chave: Risco de crédito. Aprendizagem Estatística. Classificadores. Desempenho Financeiro.

ABSTRACT

SILVA, R. A. **Financial performance portfolio to evaluate and select analyses and credit models: An approach based on Statistical Learning.** 2014. 126 p. Tese (doutorado) – Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2014.

Analysis and decision credit concession models search for relating the borrower's credit profile to the nonpayment probability of their obligations, identifying risks related to borrower and helping decision firm to approve or deny the credit request. Currently this search field has increased in Brazilian scenario – by credit activity intensification into the country with a large public banks sharing - and in the international scenario - by growing concerns about economy potential damages resulting from default events. This position leads the construction and adaptation of several models and methods by credit risk analysis from both consumers and companies. These models have been tested and compared based on prediction of accuracy or other statistical optimization metrics. This proceed is eventually not effective when analyzed by a financial standpoint, in the same time that affects the understanding and decision of the enterprise about the best model, creating a gap in the decision model choice and the firm financial goals. Given that the financial performance is a foremost indicator of any management procedure, this study aimed to address this gap by the financial performance analysis of loan portfolios formed by statistical learning techniques currently used in the classification and credit risk analysis in national and international researches. The selected techniques (discriminant analysis, logistic regression, Bayesian networks Naïve Bayes , 1 - KDB , KDB - 2 , SVC and SVM) were applied to the German Credit Data Set and their results were initially analyzed and compared in terms of accuracy and misclassification costs. Regardless of these metrics the research has proposed to use four financial metrics (RFC, PLR, RAROC and IS), finding variations in the results of each statistical learning techniques. These results suggest variations in the sequence of efficiency and, ultimately, techniques choice, demonstrating the importance of considering these metrics for analysis and selection of decision models of optimal classification.

Keywords: Credit risk. Statistical Learning. Classifiers. Financial Performance.

SUMÁRIO

1 INTRODUÇÃO	14
1.1 Contextualização e situação problema	18
1.2 Premissas	20
1.3 Objetivos.....	21
1.3.1 Objetivo geral	21
1.3.2 Objetivos específicos.....	21
1.4 Justificativa da Pesquisa	21
 2 DEFAULT DE CRÉDITO: GESTÃO E RISCO	 24
2.1 Aspectos gerais da avaliação para a concessão de crédito	25
2.2 Risco de crédito	27
2.3 Distribuição de perdas em portfólio	30
2.4 Alocação de capital econômico para risco de crédito	34
2.5 Classificação e mensuração de risco de crédito	35
2.6 A previsão de <i>default</i> de crédito	39
2.7 Aplicações em modelagem de análise de risco de crédito	40
2.7.1 Pesquisas nacionais	41
2.7.2 Pesquisas internacionais	45
 3 APRENDIZAGEM ESTATÍSTICA	 54
3.1 Análise Discriminante	55
3.2 Regressão Logística	55
3.3 Redes bayesianas	56
3.3.1 Definições e propriedades básicas de uma rede bayesiana.....	60
3.3.2 Estruturas de redes bayesianas	62
3.3.3 Estimação e propriedades <i>markovianas</i> em redes bayesianas.....	63
3.3.4 Aprendizagem em redes bayesianas	67
3.3.5 Alguns classificadores bayesianos.....	68
3.3.5.1 <i>Naïve Bayes</i>	70
3.3.5.2 <i>KDB - k-dependence Bayesian Network</i>	71
3.4 <i>Support Vector Machines</i>.....	73
3.4.1 <i>Support Vector Classifier</i>	74
3.4.2 <i>Support Vector Machine</i>	75
 4 PROCEDIMENTOS METODOLÓGICOS	 78
4.1 Caracterização do setor.....	78
4.2 Tipo de pesquisa e quadro teórico	79
4.3 Uso de ferramentas computacionais e bases de dados	82
4.4 Processo de comparação e seleção de modelos de <i>credit scoring</i>	83
4.4.1 Métricas estatísticas e proposta conceitual de métricas financeiras	87
4.4.2 Procedimentos para levantamento das informações para cálculo das métricas.....	91
 5 RESULTADOS E DISCUSSÕES	 94
5.1 Análise quantitativa: aplicação do método	99
5.2 Análise das técnicas empregadas com base na métrica ACC.....	101
5.3 Análise das técnicas empregadas com base na métrica de custos	102
5.4 Análise das técnicas empregadas com base nas métricas RFC e PLR	104

5.5 Análise das técnicas empregadas com base nas métricas IS e RAROC	107
5.6 Discussões e considerações sobre os resultados	110
6 CONSIDERAÇÕES FINAIS.....	112
REFERÊNCIAS BIBLIOGRÁFICAS	116
APÊNDICE A – Classificação via Naïve Bayes	123
APÊNDICE B – Classificação via kdB.....	124
APÊNDICE C – Classificação via SVC.....	125
APÊNDICE D – Classificação via SVM.....	126

LISTA DE FIGURAS

Figura 1 - Atividade de crédito brasileira frente ao PIB	16
Figura 2 - Participação de bancos públicos na recente expansão do crédito no Brasil	18
Figura 3 - Distribuição das perdas em portfólios de crédito.....	31
Figura 4 - Aplicações no Brasil	44
Figura 5 - Aplicações internacionais	51
Figura 6 - Elementos de um grafo probabilístico	58
Figura 7 - Extensão da estrutura de um grafo probabilístico.....	59
Figura 8 - Exemplo do grafo da CPT	61
Figura 9 - Tipos de estruturas de grafos probabilísticos.....	63
Figura 10 - Tipos de estruturas em “V”	64
Figura 11 - Estruturas em “V” com equivalência de Markov	65
Figura 12 - Propriedade de Markov Blanket	66
Figura 13 - Estrutura da CPT para dados da tabela 1	69
Figura 14 - Estrutura Naïve Bayes	70
Figura 15 - Estruturas kdB	72
Figura 16 - Vetores de suporte	73
Figura 17 - Limitações do máximo hiperplano linear	74
Figura 18 - Comportamento não linear do SVM.....	76
Figura 19 - Integração da pesquisa bibliográfica à análise.....	80
Figura 20 - Passos para análise de modelos via métricas financeiras	84
Figura 21 - Distribuição das variáveis na classificação <i>a priori</i>	95
Figura 22 - Matriz de distribuição das variáveis <i>a priori</i>	96
Figura 23 - Duração e montante sobre <i>default</i>	97
Figura 24 - Matriz de barras de variáveis preditoras	98
Figura 25 - Curva ROC e <i>Cut-off</i> modelo Naïve Bayes	99
Figura 26 - Retornos totais das carteiras	108
Figura 27 - Volatilidade do retorno das carteiras	109

LISTA DE QUADROS

Quadro 1 - Categorias de rating S&P	36
Quadro 2 - Técnicas por autores no Brasil	45
Quadro 3 - Técnicas por autores internacionais	52
Quadro 4 - Detalhamento da pesquisa bibliográfica	79
Quadro 5 - Levantamento dos dados das variáveis	92
Quadro 6 - Variáveis presentes na base de dados.....	94

LISTA DE SIGLAS

ACC – Acurácia
AIC – *Akaike Information Criterion*
BCB – Banco Central do Brasil
CMPCG – Custo médio ponderado de capital de giro
CPT –Tabela de probabilidade condicional
CRA's – Agências de risco de crédito
CVM – Comissão de Valores Mobiliários
DAG – *Direct acyclic graph*
DHMM – *Double Hidden Markov Model*
DP – *Default Probability*
EAD – *Exposure at Default*
EL – *Expected Loss*
IS – Índice de Sharpe.
KDB – *K dependence Bayesian network*
KMV – Kealhofer, McQuown & Vasicek Co
LGD – *Loss Given Default*
MB – Markov Blanket
MCC – Coeficiente de correlação de Matthew
MMC – *Maximal Margin Classifier*
MPE – Micro e Pequenas Empresas
MPL – *Multilayer Perceptron*
PIB – Produto Interno Bruto
PIM – *Poisson Intensity Model*
PLR – Perda líquida com a carteira de recebíveis,
RAROC – *Risk Adjusted Return on Capital*
RFC – Resultado financeiro com a carteira de crédito,
ROC – *Receiver operating characteristic*
RSFS – *Rough sets and scatter search*
S&P – Standard & Poors
SEM – Análise de Sensibilidade
SEM – *Structure Risk Minimization*
SMC – *Soft Margin Classifier*
SPE – Análise de especificidade
SVC – *Support Vector Classifier*
SVM – *Support Vector Machine*
VAR – *Value at Risk*

LISTA DE TABELAS

Tabela 1 - Exemplo de classificação via redes probabilísticas.....	68
Tabela 2 - Estatística descritiva dos dados	95
Tabela 3 - Resultados apurados nos modelos.....	100
Tabela 4 - Resultados dos testes de <i>accuracy</i>	101
Tabela 5 - Resultados dos testes de custo.....	103
Tabela 6 - Dados para cálculo das métricas RFC e PLR.....	104
Tabela 7 Resultados das métricas financeiras RFC e PLR.....	105
Tabela 8 - Comparativo das técnicas pela Receita Financeira com a Carteira de Crédito (RFC).....	106
Tabela 9 - Comparativo das técnicas pela Perda Líquida com a Carteira de Recebíveis (PLR).....	106
Tabela 10 - Resultado da métrica Índice de Sharpe (IS)	107
Tabela 11 - Resultados da métrica RAROC	110
Tabela 12 - Comparativo geral de resultados das métricas estudadas.....	111

1 INTRODUÇÃO

As situações nas quais os gestores tomam as decisões podem ser entendidas segundo os níveis de conhecimento e controle que estes têm sobre as diversas informações (variáveis) que compõem os quadros de decisão, gerando decisões mais ou menos estruturadas. (ABDOU; POINTON, 2011). As informações relevantes servem de bases para a tomada de decisão, configurando-se atualmente como diferencial competitivo para aquelas organizações que conseguem coletá-las e as processar de forma eficiente.

Como abordam Abdou e Pointon (2011), as informações pertinentes à decisão podem ser categorizadas segundo seus níveis de certeza, risco (entendido como um conjunto finito de estados da natureza com probabilidades associadas a cada estado) e incerteza (das quais não se tem controle sobre os possíveis estados da natureza e nem sobre as probabilidades associadas). Ainda segundo os autores, são tarefas vitais para as decisões tomadas no âmbito econômico-financeiro o entendimento das situações de certeza, risco e incerteza envolvidas num cenário de decisão e a quantificação das probabilidades de sucesso de cada uma das alternativas presentes nestes cenários.

Corroborando com tal raciocínio Pascual, Martínéz e Alamillos (2013), entendem que a eficiência das decisões das quais se ocupam os gestores pode ser a diferença entre o sucesso e o fracasso de uma empresa. Ainda segundo o Pascual, Martínéz e Alamillos (2013), as decisões de crédito pertencem a este cenário de necessidade de otimização da decisão, fazendo com que o desenvolvimento e uso de mecanismos eficientes de análise e tratamento das probabilidades de um cliente não honrar seus compromissos, sejam questões capitais para o sucesso das organizações.

O crédito tem seu conceito associado à troca de bens por promessa de pagamentos que deverão ocorrer no futuro (MATIAS, 2007). Para as empresas, o crédito é um instrumento que proporciona alavancagem operacional e consequente otimização dos resultados, independentemente do setor de atividade. Esta otimização é derivada dos aumentos da atividade da firma e da capacidade de compra de seus clientes¹.

No que tange a capacidade de compra dos clientes, o crédito também tem forte papel social ao permitir o acesso à compra de bens e à qualidade de vida, abrindo espaço para, em grande escala, aumentar da atividade econômica e produtiva do país.

¹ A fundamentação sobre esta definição de crédito está no capítulo 2.

Em finanças corporativas a atividade de crédito tem sua importância associada à gestão de perdas e garantia do alcance dos objetivos financeiros das firmas, tanto das conessoras quanto das que requisitam os créditos (BLUHM; OVERBECK; WAGNER, 2003).

A decisão de concessão de crédito é uma tarefa de grande complexidade, envolvendo não apenas elementos associados ao perfil e aos riscos da operação e do tomador, mas também as políticas e objetivos de crédito da firma, o que faz com que essa decisão demande total alinhamento estratégico com a gestão.

As políticas de crédito buscam atender aos objetivos de crédito. Cada empresa terá um objetivo com sua carteira e suas políticas de crédito, sendo que estes objetivos, em geral, estão relacionados ao volume operacional da firma.

Como expresso por Matias (2007), para uma empresa decidir conceder crédito é necessário que estabeleça procedimentos para análise e concessão, monitoramento e cobrança. As políticas de crédito visam justamente estabelecer estes procedimentos incluindo padrões de solicitação (requisição), pré-análise e análise, controle do investimento em recebíveis (volume da carteira de crédito), padrões de crédito, prazos, descontos e políticas de cobrança e requisitos mínimos para a concessão.

Este tema tem se destacado nos últimos anos nacional e internacionalmente, principalmente pela recente evolução do cenário econômico. No âmbito nacional, segundo dados do BCB (2014)², a atividade de crédito, integrando todas as fontes e destinos de crédito, saltou de pouco mais de 31% do PIB – Produto Interno Bruto – brasileiro em dezembro de 2007 para 56,5% em dezembro de 2013, isto é, uma variação de aproximadamente 82,26% num prazo de seis anos. Tal comportamento pode ser observado na Figura 1.

Cabe destacar que as idiosincrasias presentes no Brasil – tais como o volume e complexidade de obrigações tributárias, o elevado risco de *default*, as elevadas taxas de *spread* bancário e de títulos públicos, a política monetária, dentre outras – têm direcionado as atividades de análise e gestão de crédito no país (WANG; MA, 2011).

Por outro lado, o recente crescimento da maturidade econômica do país fez surgir instrumentos econômicos novos e mais modernos, voltados para a gestão e atuação no mercado de crédito, tais como as recentes alterações nas publicações contábeis de companhias e a incorporação, por parte do BCB – Banco Central do Brasil e da CVM – Comissão de Valores Mobiliários, de novos instrumentos para securitização e fomento de um insipiente mercado de derivativos de crédito (ARCOVERDE, 2002).

² BCB – Banco Central do Brasil. Sistema de gerenciamento de séries temporais v2.1

Figura 1 - Atividade de crédito brasileira frente ao PIB

Fonte: Banco Central do Brasil (2014).

Já no âmbito internacional, observa-se também grandes movimentações em modelagem e gestão de crédito, especialmente após a crise derivada dos problemas associados aos títulos *subprime* Norte-Americanos em 2008, que provocaram grande mácula no mercado imobiliário e de crédito daquele país, se espalhando por todo o mercado financeiro mundial (DUAN; SHRESTHA, 2011).

Na área acadêmica, tanto em esfera nacional quanto internacional, estas movimentações se traduziram na busca por modelos mais eficientes para a classificação e previsão de *default*, incorporação de instrumentos analíticos como testes de *stress* e incorporação de intervalos de confiança para as perdas dos portfólios (carteiras), especialmente através do VAR – *Value at Risk*³, bem como no desenvolvimento e aplicação de técnicas de aprendizagem estatística.

Entende-se aprendizagem estatística como sendo a área, dentro das ciências estatísticas, dedicada à produção de ferramentas para o entendimento destas massas de dados, com aplicações em diversas áreas, tais como: medicina, finanças, biologia, robótica, etc. (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Essa é uma área que tem atraído grande interesse e que levou ao desenvolvimento recente de técnicas estatísticas e gerou novas áreas de pesquisas, tais como a estatística computacional (envolvendo técnicas de *data mining*, *machine learning*, etc.).

Como referenciado por Lu et al. (2013), a quantidade de operações e a necessidade de otimização dos créditos para maximização dos resultados da firma fazem com que gerenciar

³ Observar pesquisas internacionais na seção 2.6.2.

carteiras de crédito seja uma atividade fundamental para as organizações. Este é um objetivo fundamental dos modelos de classificação, análise e concessão de crédito e de onde surge a necessidade de emprego de ferramentas de aprendizagem estatística para a produção de modelos de análise e concessão de crédito.

Segundo Bluhm, Overbeck e Wagner (2003), os modelos buscam oferecer um mecanismo de proteção contra perdas provenientes do *default* sem que, para tanto, o volume de negócios seja prejudicado. As decisões de concessão de crédito são normalmente baseadas em modelos quantitativos de previsão e otimização da concessão, o que faz com que a construção destes modelos seja um importante passo para o desenvolvimento de sistemas eficientes para gerenciamento do risco de crédito (LU et al., 2013).

A área de classificação, análise e concessão de crédito também tem atraído atenção pela incidência de mecanismos de regulação da atividade de crédito para empresas bancárias, tanto nacionais quanto internacionais⁴, bem como pela evolução de técnicas estatísticas e computacionais que tem aberto espaço para a produção de novas formas e modelos de gerenciamento de risco de crédito (CROUHY; GALAI; MARK, 2000).

Com base neste raciocínio, o desenvolvimento de políticas, modelos e práticas de quantificação, qualificação e gestão de portfólios de crédito é um dos temas mais importantes em finanças, principalmente para a indústria financeira (já que esta tem a comercialização de produtos e serviços de crédito como uma de suas principais atividades).

No âmbito acadêmico, o desenvolvimento de métodos quantitativos de análise de risco de crédito tem produzido um conjunto grande de modelos dedicados à análise e tratamento desta problemática, com grande quantidade de técnicas extraídas das ciências estatísticas, matemáticas e computacionais (ver seção 2.6).

Segundo Duan e Shrestha (2011), análise discriminante, regressão logística, redes neurais, redes bayesianas e métodos *ensemble* são métodos amplamente empregados na área de análise de crédito. Considerando a classificação apresentada por Nagarajan, Scutari e Lèbre (2013), os métodos de análise discriminante, regressão logística, redes bayesianas e métodos *ensemble* são classificados como métodos de aprendizagem estatística, sendo esta a classificação utilizada neste estudo.

A qualidade destes modelos é testada geralmente pelo nível de acerto de classificações ou de métricas de otimização estatística. Contudo, como destacado por autores como Ferson (2013) e Iscoe et al. (2012), as medidas de desempenho financeiras são ideais para a

⁴ Tais como os acordos da Basileia.

comparação e teste de modelos que tenham como objetivo a maximização do resultado financeiro, como é o caso da modelagem para a área de crédito.

Espera-se que, por meio do modelo de classificação, análise e concessão a organização obtenha uma proteção contra os riscos de crédito ao mesmo tempo que os seus objetivos de crédito sejam atendidos. Desta forma, o presente estudo se posiciona nessa lacuna e emprega métricas financeiras para a análise da eficiência dos métodos baseados em Aprendizagem Estatística no tratamento de problemas de classificação e previsão de *default* de crédito através da análise dos resultados financeiros das carteiras formadas a partir de cada uma das técnicas de classificação estudadas.

1.1 Contextualização e situação problema

Como abordado introdutoriamente o crédito, seus mercados, marcos regulatórios e sua importância para a economia e sociedade tem papel de destaque na gestão de organizações.

No cenário nacional, alinhado às questões internacionais, um movimento de expansão do crédito foi observado principalmente nas últimas duas décadas, provocando também o aumento do interesse por esta área no país. Segundo o relatório da Febraban (2013), intitulado “Panorama do Mercado de Crédito” houve grande expansão no crédito entre 2001 e 2013.

Figura 2 - Participação de bancos públicos na recente expansão do crédito no Brasil
Fonte: Febraban (2013).

Este relatório traz que esta expansão foi particularmente forte entre 2011 e 2013 quando o volume de crédito em relação ao PIB brasileiro saltou de 48,3% para 55,6%. Ainda segundo o relatório, este aumento de aproximadamente de 15,11% em dois anos teve os bancos públicos como os maiores responsáveis, uma vez que a participação destes bancos em relação ao PIB saltou de 20,7% para 28,3% no mesmo período.

Paralelamente, há também uma expansão no desenvolvimento e emprego de modelos de análise e concessão de crédito na medida em que ocorrem avanços nos recursos computacionais disponíveis. Dentre estes avanços, os modelos de classificação de redes causais ou redes bayesianas e modelos *ensemble* ganharam destaque principalmente nas áreas das ciências matemáticas, estatísticas e computacionais.

No Brasil estes modelos ainda são timidamente utilizados em ciências administrativas, sendo que sua difusão para tratamento do problema de crédito é também recente dentro das ciências estatísticas e computacionais, como é possível verificar pela pesquisa das aplicações nacionais e internacionais apresentadas na seção 2.7.

Dentre as produções bibliográficas no cenário brasileiro estes métodos são ainda insipientes, uma vez que, em de pesquisa na base de dados brasileira Scielo, são observadas 15 (quinze) publicações sobre o tema, sendo apenas quatro (04) delas na área de negócios, sem que nenhuma seja na área de análise de risco de crédito⁵.

Essa não é a realidade observada no cenário internacional, no qual as redes bayesianas e os métodos de classificação como o *Support Vector Machine* – SVM, têm se configurado como algumas das mais importantes e promissoras técnicas, especialmente para problemas de classificação, como é o caso dos problemas de *credit scoring* e *credit rating* (PASCUAL; MARTÍNEZ; ALAMILLOS, 2013; DUAN; SHRESTHA, 2011). Estas técnicas têm suas vantagens na flexibilidade e capacidade de otimização (DUAN; SHRESTHA, 2011).

A eficácia dos classificadores tem sido, contudo, testada por meio de métricas de otimização estatística, especialmente na previsão de *default*, o que não necessariamente reflete a otimização financeira objetivada pelas empresas ao adotarem modelos estatísticos de classificação e previsão de *default* de crédito.

Como expresso por Lu et al. (2013), estes modelos estão voltados para a minimização das perdas derivadas do erro. Contudo, esta pesquisa parte do raciocínio de que: **A otimização do investimento em crédito advém de um equilíbrio entre as perdas derivadas do *default* e das perdas derivada de políticas de restrição de crédito.**

⁵ Dados baseados em pesquisa realizada pelo próprio autor, cujo método é apresentado em 2.6.

Desta forma, a presente pesquisa tem como escopo empregar métricas de desempenho financeiro para análise de modelos de aprendizagem estatística aplicados ao problema de classificação, análise e concessão de crédito. Considerando tal escopo busca-se responder ao problema pesquisa definido como:

Como selecionar um modelo para classificação, análise e concessão de crédito que se mostre mais eficaz na geração de resultados financeiros e, conseqüentemente, na maximização do valor do investimento da firma em sua carteira de crédito?

Através da resposta desta questão de pesquisa espera-se contribuir para um entendimento financeiro da eficiência dos classificadores empregados no cenário internacional nas áreas de estatística e ciências da computação, tornando-os úteis para análises de problemas de classificação e risco de crédito, popularizando o uso destes e servindo de base para a ampliação do entendimento gerencial sobre seus métodos.

1.2 Premissas

Esta pesquisa parte da premissa de que o desempenho financeiro da carteira de crédito deve ser empregado como critério para seleção de técnicas aplicadas à análise e concessão de crédito. Não obstante, a tese aborda em suas análises e discussões técnicas clássicas em conjunto com novas técnicas, até então, pouco pesquisadas na literatura brasileira de negócios.

Estende-se que é possível definir qual técnica é a mais eficiente para cada situação por meio da análise de elementos como o retorno da carteira, a volatilidade deste retorno e a análise do *trade-off* risco vs retorno proveniente da carteira formada pelas classificações positivas⁶ de cada uma das técnicas.

Extrapolando tal premissa, entende-se que ao comparar os modelos de crédito com base no desempenho financeiro da carteira formada por cada modelo é possível indicar qual deve ser selecionado para utilização.

Dessa forma, esta pesquisa apresenta como produtos principais: i) a apresentação e comparação do poder preditivo de novas técnicas; e ii) a necessidade de comparação destas técnicas por meio de métricas financeiras relacionadas ao desempenho da carteira.

⁶ Nas quais o cliente teria sua solicitação de crédito aprovada.

1.3 Objetivos

1.3.1 Objetivo geral

O objetivo desta pesquisa é empregar métricas para análise da performance financeira da carteira como critério para análise escolha de técnicas e modelos de previsão de *default*.

1.3.2 Objetivos específicos

Com base no objetivo geral estabelecido os objetivos específicos desta pesquisa são:

- a) Identificar e mapear as principais técnicas de Aprendizagem Estatística empregadas para a produção de modelos de análise e concessão de crédito;
- b) Aplicar as técnicas analisando-as pelo critério estatístico de acurácia;
- c) Identificar e aplicar métricas financeiras teoricamente eficazes para comparação das técnicas de aprendizagem estatística aplicada a risco de crédito mapeadas; e
- d) Comparar a eficiência estatística e financeira dessas técnicas.

1.4 Justificativa da Pesquisa

Estudos que tenham como objetivo aprimorar técnicas de análise e gestão de risco de crédito são amplamente desejáveis em pesquisas aplicadas em Ciências Administrativas por sua capacidade de contribuir para a explicação deste problema complexo, contribuindo também para a gestão de valor das instituições e maximização da eficiência econômica dos recursos destinados ao crédito.

Modelos que subsidiem a formação e gestão de portfólio e alocação de capital econômico fornecem uma contribuição interessante no mercado de crédito, especialmente para o brasileiro, tendo em vista a condição econômica de expansão de crédito e as demandas de gestão de crédito já destacadas na introdução da pesquisa.

A previsão de risco de *default* em termos de probabilidade e a classificação dos tomadores de crédito é uma questão central neste sentido. Isso se deve ao fato de que os resultados destas previsões são empregados em um grande conjunto de técnicas de gestão de crédito, tais como modelos de alocação de capital econômico, perdas esperadas, dentre outros (BLUHM; OVERBECK; WAGNER, 2003).

Duan e Shrestha (2011) também entendem que a previsão dos eventos de *default* é um elemento central da gestão de portfólio de crédito. Para os autores, a previsão do *default* é particularmente importante pela extensão dos danos provocados por sua ocorrência, não apenas na empresa que concede o crédito, mas também em seus clientes, fornecedores, funcionários e todos os demais *stakeholders*.

Adicionalmente Duan e Shrestha (2011) salientam que, ainda que a extensão dos danos provocados pelo *default* esteja condicionada à magnitude do *default*, bem como a características econômicas sistêmicas do tomador de crédito em *default*; todas as organizações, mesmo pequenas empresas sem implicações sistêmicas à economia, devem analisar a sua eficiência em concessão de crédito, tendo em vista o potencial dos danos oriundos destes eventos.

Dentro dessas perspectivas, as instituições financeiras que atuam com a análise, concessão e gestão de crédito têm dado ênfase no desenvolvimento de instrumentos para quantificação dos riscos como suporte ao processo de decisão de crédito.

Este interesse tem sua justificativa na necessidade de adoção de métodos robustos para determinação dos riscos aos quais os portfólios de crédito bancários e corporativos estão expostos, bem como para determinação do montante de capital econômico necessário para fazer frente a estes (BLUHM; OVERBECK; WAGNER, 2003).

Recentes desenvolvimentos em mensuração de risco de carteiras de crédito têm produzido vários modelos de classificação de crédito no cenário internacional, baseados em novos e robustos métodos e técnicas estatísticas e computacionais. Esta movimentação é observada não apenas no âmbito acadêmico, mas também nas organizações que empregam tais técnicas para composição de modelos de suporte à decisão.

No que tange o cenário de modelagem brasileiro especificamente, Andrade e Thomas (2007), indicam que problemas relacionados às idiosincrasias institucionais e econômicas fazem com que seja preferido o desenvolvimento de modelos internos de classificação pelas instituições nacionais, dada a realidade observada no país. Dentro deste cenário, emergem técnicas como redes bayesianas, classificadores combinados e *ensembles*, regressões logísticas, análises discriminantes, dentre outras técnicas com diversas aplicações em risco de crédito.

Estudos como Pascual, Martín e Alamillos (2013), Brown e Mues (2012), Danenas, Garsva e Gudas (2011), Duan e Shrestha (2011), dentre outros têm evoluído na classificação, análise e tratamento de risco de crédito através destas técnicas. Todavia, as métricas empregadas para a análise da eficiência destas técnicas são normalmente voltadas para o nível de acurácia do modelo ou outras estritamente estatísticas.

O valor gerado pelo investimento na carteira crédito, o volume e a volatilidade desta carteira passam a ser elementos importantes nas análises dos gestores de crédito. Neste sentido, outros elementos também ganham em importância, tais como a otimização do *trade-off* risco *versus* retorno da carteira de crédito.

No que diz respeito ao escopo desta pesquisa, o qual centra-se também na disseminação de resultados e contribuições de uso de técnicas de Aprendizagem Estatística para a classificação e mensuração do risco de portfólio de crédito, tanto daquelas consideradas tradicionais quanto das técnicas recentemente empregadas na área, o tema mostra-se relevante e contemporâneo. Esta relevância se observa ainda maior considerando o cenário brasileiro e mundial de demandas por pesquisas na área.

Outro ponto que justifica a importância e atualidade do tema é a sua proposta conceitual, qual seja: analisar os resultados dos modelos na ponta da carteira, isto é, analisar os resultados com base no desempenho financeiro da carteira pós-classificação e não apenas analisar o resultado estatístico de capacidade preditiva de cada modelo, sendo esta uma proposta inovadora para a área, uma vez que não foram identificadas teses ou dissertações no âmbito nacional ou mesmo artigos que trouxessem esta proposta conceitual.

Diante do exposto, entende-se que estudos que tenham como objetivo trazer novas técnicas para tratamento da problemática da otimização de portfólios de crédito, especialmente analisando o desempenho financeiro produzido pela aplicação dessas técnicas, são desejáveis em Administração.

2 *DEFAULT* DE CRÉDITO: GESTÃO E RISCO

O crédito pode ser comumente conceituado como o ato de conceder recursos a um tomador mediante compromisso de pagamento futuro. Neste tipo de operação o tomador pode antecipar recurso (sob o formado de valores, bens, etc.) ao qual faz jus para, posteriormente, realizar o pagamento (ou ressarcimento) da importância tomada.

Analisando as origens dos créditos estabelecidas por Matias (2007), os créditos podem ser observados basicamente sobre três formas:

- i. Sob a forma de bens e serviços: neste tipo o vendedor oferece seus bens e serviços em troca de promessas de pagamento futuras ou pelo não recebimento de venda à vista⁷.

Aqui o tomador do crédito recebe o bem ou serviço e contrai, em contrapartida, uma obrigação de pagamento futura referenciada pelo valor do bem ou serviço hora adquirido somada às taxas de remuneração do capital. Este tipo de operação possibilita aumentar o poder de compra dos consumidores e alavancar as operações dos vendedores.

- ii. Sob a forma de valores monetários: neste tipo de crédito o tomador recebe valores monetários contraindo uma obrigação futura.

As operações de crédito sob forma de valores monetários visam suprir necessidades de recursos para finalidades diversas dos tomadores e não a realização de contraprestação a bens e serviços hora adquiridos. Este tipo de operação pode ocorrer de várias formas, tais como o adiantamento a fornecedores feitos por uma empresa, por agentes financeiros credenciados (que cobram taxas de juros como forma de remuneração do capital), etc.

- iii. Sob a forma de valores disponíveis: este conceito está também relacionado ao ato de colocar um dado valor à disposição do tomador, mesmo que este ainda não tenha sido retirado.

⁷ Entende-se o não recebimento de venda à vista como uma operação de venda à vista que, por algum motivo, deixa de ser liquidada, tal como: falha em sistema de cartões, cheques sem fundos, etc.

Comumente conhecida como crédito pré-aprovado, nesta modalidade o tomador poderá fazer uso do crédito caso seja de seu interesse. Tal modalidade é uma forma de gestão amplamente empregada por bancos e empresas para facilitar a operação, reduzindo a demanda por avaliação sempre que o cliente precisar de crédito.

2.1 Aspectos gerais da avaliação para a concessão de crédito

Bluhm, Overbeck e Wagner (2003), expressam que crédito é, por definição, uma operação baseada na expectativa de recebimento futuro dos valores concedidos ao tomador, o que tem um risco de não ocorrer. Este risco é o que torna necessária a avaliação do tomador.

De uma forma geral, a avaliação do crédito consiste em comparar as características ou padrões de comportamento do proponente ao crédito com as características de outros clientes recentemente assistidos pelos sistemas de análise e concessão de crédito da empresa fornecedora, verificando se as características do atual proponente estão mais próximas das características daqueles que honraram ou mais próximas das características daqueles que não honraram seus compromissos dentro do período (ABDOU; POINTON, 2011; BLUHM, OVERBECK; WAGNER, 2003).

Espera-se que, caso um solicitante tenha características similares às daquelas de outros solicitantes que tiveram sua solicitação aprovada e que apresentaram algum *default* ao longo do tempo, a tendência da análise é naturalmente rejeitar a solicitação deste agente, sendo que a recíproca também é verdadeira, isto é, caso as características do solicitante sejam similares às daquelas de solicitantes que tiveram suas solicitações aprovadas e honraram seus compromissos, a tendência é a aprovação da solicitação (ABDOU; POINTON, 2011).

Esta não é uma posição consensual, pois estratégias particulares das organizações poderão atrair clientes com perfil de risco mais alto, desde que este perfil produza retornos (normalmente a título de taxas de juros) superiores (MATIAS, 2007).

Segundo Abdou e Pointon, (2011), há duas formas de se analisar as características dos tomadores de crédito, sendo uma a análise subjetiva – também conhecidas como modelos de julgamento – e a outra através dos modelos objetivos ou de previsão (quantitativos).

Os modelos de julgamento são aqueles nos quais os proponentes ao crédito são julgados individualmente por analistas de crédito, sendo que o sucesso do processo de julgamento depende da habilidade e da competência do analista (BAILEY, 2004). Este procedimento apresenta alta subjetividade dado que, mesmo de forma inconsciente, o analista não consegue

se desvencilhar de seu sistema de preferências individuais que tendem a direcionar suas decisões (BAILEY, 2004).

Como referenciado por Abdou e Pointon (2011), os sistemas de julgamento têm também pontos positivos, tais como a capacidade do analista utilizar seu histórico de experiências em seu favor para a realização das análises, bem como sua capacidade de captura de sinais (muitas vezes psicológicos) sobre o ambiente e sobre o tomador. Ainda segundo estes autores, modelos objetivos são aqueles voltados para a análise quantitativa de dados de crédito, também conhecidos como modelos de *credit scoring*. O objetivo desses modelos é parametrizar variáveis e apresentar resultados sobre aspectos que auxiliem no julgamento da concessão ou não do crédito ao solicitante. Busca-se prever o comportamento de pagamento de compromissos do tomador e classificá-lo em grupos de clientes que entrarão ou não em *default*. Por tais características, estes modelos também são chamados de modelos de previsão ou de classificação.

Os modelos de previsão ou classificação normalmente empregam técnicas estatísticas sobre a experiência histórica das empresas conessoras crédito, buscando gerar uma classificação das solicitações de crédito em aceitáveis ou inaceitáveis do ponto de vista da probabilidade de ocorrência de *default* e dos valores de perdas passíveis de recuperação.

Segundo Abdou e Pointon (2011), estas técnicas são largamente empregadas por reunirem características como a não necessidade de julgamento, uma vez que a classificação é dada pelo próprio modelo através de suas variáveis *preditoras*⁸. Outra vantagem segundo os autores é a rapidez com que se obtêm os resultados das análises, o que garante eficiência ao processo de avaliação, especialmente com os recentes avanços em tecnologia da informação.

A classificação do proponente é indicada pelo próprio sistema por meio de um ponto de corte, isto é, um ponto que indica que, a partir daquela faixa, os clientes têm um risco associado alto e inaceitável pelo sistema.

Contudo, Abdou e Pointon (2011), indicam que normalmente o analista pode individualmente aceitar ou deixar de aceitar a solicitação à revelia do modelo quantitativo de avaliação. Ainda segundo os autores, as maiores críticas sobre este tipo de modelo recaem sobre as técnicas estatísticas empregadas, seus pressupostos e hipóteses.

Outro ponto importante a ser considerado é que as operações de crédito não estão associadas apenas à probabilidade de que as obrigações não sejam honradas por parte do

⁸ Também conhecidas como variáveis explicativas ou independentes. São as variáveis com base nas quais o modelo realizará a previsão ou classificação dos indivíduos no caso da análise de crédito.

tomador de crédito. Como expresso por García, Giménez e Guijarro (2013), existem diversos riscos associados às operações de crédito, dentre os quais os autores destacam:

- Risco de mercado: risco de ocorrência de mudanças não previstas nos preços ou taxas praticadas no mercado e que influenciam o volume de operações do prestador, bem como o valor de sua carteira;
- Risco de crédito: risco de ocorrência de mudanças na qualidade do tomador ou mesmo da carteira de crédito como um todo, provocando uma redução do valor esperado desta carteira;
- Risco de liquidez: risco de ocorrência de perdas ou aumento dos custos financeiros para manutenção da liquidez da firma, ou mesmo redução do acesso ao financiamento de suas atividades;
- Risco operacional: risco associado à ocorrência de fraudes, erros humanos ou falhas de sistema durante a execução dos processos de análise e concessão de crédito; e
- Risco sistemático: risco associados a reações sistêmicas derivadas de crises ou *stress* mercadológico que afeta todo o mercado de crédito.

Especificamente sobre o risco de crédito, as mudanças de que tratam García, Giménez e Guijarro (2013), estão associadas ao *default* dos tomadores de crédito. Este elemento é um dos principais pilares da análise e gestão de crédito e é o foco da presente pesquisa.

2.2 Risco de crédito

Como abordam autores como Iscoe et al. (2012), Bluhm, Overbeck e Wagner (2003) e García, Giménez e Guijarro (2013), o risco de crédito pode ser expresso como a probabilidade (ou esperança) de que o tomador de crédito não cumpra suas obrigações – situação também conhecida como evento de *default*.

Para Bessis (1998), o risco de crédito pode ser subdividido em três principais tipos de riscos, quais sejam:

- a) risco de *default*: associado à probabilidade do tomador não honrar suas obrigações e ficar em situação de insolvência;

- b) risco de exposição: associado à volatilidade do valor esperado da carteira ou da operação de crédito; e
- c) risco de recuperação: associado às incertezas quanto ao volume que será possível recuperar no caso de ocorrência de um evento de *default*.

É importante indicar que existem diferenças substanciais sobre *default* e o não cumprimento de obrigações por um período de tempo. Como abordam Brito e Assaf Neto (2008), a inadimplência é o não cumprimento de obrigações de crédito por um período de tempo. Por outro lado o evento de *default* é o não pagamento das obrigações, o que gera custos e compromete o resultado da carteira de crédito.

Matias (2007), indica que a inadimplência está associada ao atraso no pagamento das obrigações por parte do tomador de crédito enquanto a insolvência é o não pagamento destas em qualquer época.

Caso entenda-se a mudança da qualidade do tomador como *default* (seja esta mudança derivada da inadimplência ou da insolvência), existirão diversas outras situações que poderão ser caracterizadas como um evento de *default*, tais como: início de processo de falência ou concordata, inadimplência de natureza econômica (caracterizada por uma redução brusca no valor econômico dos ativos da firma para níveis abaixo da dívida) ou mesmo descumprimento de cláusulas contratuais de caráter restritivo.

Ainda segundo Matias (2007), a parcela inadimplente não apenas não pode ser considerada como um evento que gera perdas (*default* de crédito) como também é normalmente uma parcela que apresenta maior rentabilidade para a carteira de crédito, tendo em vista a possibilidade de incorporação de taxas de juros e multas pelo não cumprimento das obrigações no período estipulado.

Contudo, autores como Bluhm, Overbeck e Wagner (2003), entendem que perceber um evento de inadimplência como um evento *default* justificasse pelo impacto de tal evento sob as políticas de alocação de capital econômico, bem como para entendimento das probabilidades de ocorrência de tais eventos.

Não obstante, as bases de dados utilizadas para testes de modelos de previsão normalmente são classificadas à priori como *default* e não *default*, não fazendo distinção sobre adimplentes, inadimplentes e insolventes, bem como desconsiderando, no julgamento dos modelos, dados relacionados a valores de recuperação, multas e taxas de juros associadas.

Iscoe et al. (2012), indica que o risco de crédito pode ser entendido tanto como uma função das perdas derivadas do *default* do tomador quanto como uma função da perda da

qualidade de crédito do tomador ou da carteira de crédito. Esta segunda abordagem se embasa em grande parte no efeito observado por modelos de análise e precificação do valor da carteira de crédito das instituições financeiras que entendem que tal comportamento traz maior risco às esperanças de retorno. Ao risco de mudanças na qualidade de uma carteira de crédito dá-se o nome de migração de risco de crédito.

Não obstante, a precificação destes modelos apresenta movimentações em função de migrações da qualidade de crédito, tendo em vista que, ainda que não resulte em perda imediata ou mesmo inadimplência para a instituição, a deterioração da qualidade do crédito do tomador traz consigo um aumento na probabilidade de que um evento de *default* potencialmente danoso venha a ocorrer num futuro próximo (CHAN; WONG; ZHAO, 2012).

De qualquer forma, o risco de crédito, como definido anteriormente, está associado à probabilidade de perda financeira em uma concessão de crédito, o que gera a necessidade de análise de quais são estas perdas e probabilidades.

Segundo Saita (2007), um sistema de gerenciamento de risco de crédito pode considerar vários tipos de perdas, dependendo do método utilizado e dos raciocínios por trás destes métodos. A visão sobre perdas comumente encontrada dentro da literatura está associada à probabilidade do tomador de crédito ser classificado como em “*default*”, ou seja, o tomador ser classificado como um mau pagador.

Como estabelecem Bluhm, Overbeck e Wagner (2003), o princípio empregado na teoria de probabilidade a respeito da esperança, em que uma média dos resultados das probabilidades associadas a cada possível estado da natureza é entendida como sendo a previsão – ou valor esperado – para ocorrência deste evento, também é empregado na gestão de risco de crédito.

O cálculo das probabilidades de *default* é o principal fator motivador da produção de modelos internos de previsão destes eventos.

Segundo Bluhm, Overbeck e Wagner (2003), os métodos mais comuns para esta finalidade são a calibração das probabilidades através de modelos de *rating* internos (modelagem de previsão de *default*) e a calibração por modelos oferecidos por agências externas, tais como: Mood’s Investors Services, Standard & Poor’s e Fitch ou por dados de mercado, através de modelos como KMV®, *Credit Metrics*, *Credit risk+*, dentre outros. Tais métodos são abordados nas sessões 2.3 e 2.5.

No mercado brasileiro, por características econômicas particulares, é comum o desenvolvimento e emprego de modelos internos de mensuração das probabilidades de *default* (ANNIBAL; KOYAMA, 2011).

Matias (2007), oferece um exemplo de calibração através dos resultados da probabilidade de *default* a partir de uma regressão logística. Este método é, segundo Pascual, Martín e Alamillos (2013), o mais difundido dentro das análises e mensurações de risco de crédito no mercado. Outros modelos são oferecidos e discutidos nos âmbitos nacional e internacional na seção 2.7.

Para gerenciamento desse risco, o mecanismo comumente empregado é a instituição de políticas de crédito organizacionais, as quais são instrumentos que visam maximizar receitas derivadas das vendas ao mesmo tempo em que oferecem uma prevenção contra as perdas provenientes do *default* (GETLER; KIYOTAKI, 2010).

As políticas são fundamentais para a maximização do resultado da carteira de crédito tendo em vista que, como expresso por Matias (2007), políticas restritivas de crédito podem provocar danos à organização uma vez que provocam insatisfação em clientes e consequente redução no volume de vendas, enquanto políticas muito permissivas tendem a aumentar o volume das perdas com a insolvência.

Para Bluhm, Overbeck e Wagner (2003), os esforços para criação e gestão de políticas de crédito passam por modelos de classificação e mensuração de risco, os quais vão de encontro com marcos para prevenção de riscos e gestão de coberturas, como os Acordos da Basileia I, II e III.

2.3 Distribuição de perdas em portfólio

A distribuição das perdas de um portfólio é o conjunto das medidas das perdas deste *portfólio*. Conhecer esta distribuição é particularmente importante para se mensurar o risco do *portfólio* de crédito, isto é, para fornecer medidas para as perdas esperadas e não esperadas, possibilitando a mensuração do capital econômico a ser alocado pela instituição.

Entender e calcular a distribuição das perdas de um portfólio é uma tarefa central na gestão do risco do portfólio de crédito.

Segundo Bluhm, Overbeck e Wagner (2003), a distribuição das perdas de um portfólio de crédito geralmente não responde a um comportamento de normalidade, apresentando distribuição assimétrica à direita (positiva com mediana maior que a média e menor que a moda) e caudas longas como demonstrado na Figura 3.

Segundo Bluhm, Overbeck e Wagner (2003), os principais métodos para cálculo da distribuição das probabilidades das perdas de um portfólio de crédito são o uso de simulações e de aproximações.

Figura 3 - Distribuição das perdas em portfólios de crédito

Fonte: Bluhm, Overbeck e Wagner (2003, p. 34).

Nas simulações são construídas distribuições empíricas das perdas da carteira, simulando-se as variáveis de entrada do modelo explicativo utilizado (modelo de risco de crédito empregado pela instituição).

O método de simulação de Monte Carlo é um dos mais empregados para tais fins, contudo podem ser empregados outros métodos, tais como o método de simulação de *Bootstrap* (BLUHM; OVERBECK; WAGNER, 2003). Segundo Iscoe et al. (2012), nestes métodos as perdas não esperadas são geralmente posicionadas como partes superiores de alguma medida *separatriz* da distribuição das perdas (quartis superiores, percentis superiores, etc.).

Nas aproximações a análise assume premissas sobre as distribuições das perdas da carteira, identificando tais distribuições como sendo próprias de distribuições teóricas parametrizadas em função de suas médias e desvios-padrão. Neste caso as distribuições das perdas esperada e não esperada são obtidas a partir da distribuição de probabilidades assumida na aproximação (a priori). Para Iscoe et al. (2012), este método identifica a perda esperada é normalmente distribuída (o que não recebe sustentação da figura anterior) e assume a perda não esperada como um múltiplo do desvio-padrão.

Perda esperada é a probabilidade de perda assumida (esperada) por uma instituição em sua carteira de crédito, a qual é obtida a partir do somatório da multiplicação da probabilidade

de ocorrência de eventos de *default* associados aos tomadores e a perda em caso de insolvência (esta última é a diferença entre o montante da exposição no momento de *default* e a parcela deste montante passiva de recuperação) (SMITHSON, 2003).

Segundo Bluhm, Overbeck e Wagner (2002), o trabalho está em atribuir uma probabilidade de ocorrência de *default* para cada cliente (*DP – Default Probability*), de uma fração de perda no caso de ocorrência deste *default* (*LGD – Loss Given Default*), bem como do volume de recursos exposto por ocasião dos empréstimos e que será perdido em caso de *default* em um determinado período de tempo (*EAD – Exposure at Default*). Dentro desta perspectiva a expectativa de perda (*EL – Expected Loss*) é definida como:

$$EL = E[L] = EAD \times LGD \times L \quad \text{Com: } L = 1D, P(D) = DP \quad (2.1)$$

Em que: D representa a ocorrência do evento de *default* em algum período de tempo dentro da vigência da operação, normalmente tratado em anos; P(D) representa a probabilidade de D; EAD e LGD são constantes e L representa uma variável dicotômica que indica a ocorrência ou não da perda.

Para Garside et al. (1999), as perdas esperadas são, por definição, o custo do negócio em uma instituição financeira, não podendo ser entendido como um risco financeiro, uma vez que são intrínsecas aos negócios de crédito, devendo ser, assim, incluídas nas taxas das operações precificadas.

Para os autores, as características dos riscos esperados, por estarem associadas ao risco do tomador, podem ser otimizadas, o que tem se mostrado um fértil campo para pesquisas na área de crédito, devido às dificuldades para cálculo dos riscos de perda por insolvência (associada a fatores como garantias e instrumentos de recuperação de crédito) e a dificuldade de estimação do risco de *default* dos tomadores dentro dos portfólios ao longo do tempo.

Contudo Garside et al. (1999), apontam para a importância das perdas não esperados para o sucesso da instituição (e para seu risco global), tecendo entendimento de que este tem características próximas ao risco sistêmico. Segundo Bluhm, Overbeck e Wagner (2003), assumindo que o evento de *default* (V) e suas implicações (em termos de severidade) são independentes, a perda inesperada pode ser dada por:

$$UL = EAD \sqrt{V[SEV]DP + DP(1 - DP)LGD^2} \quad (2.2)$$

Considerando que $V(X) = E(X^2) - E(X)^2$ e $V(1D) = DP(1 - DP)$

Esta não é uma proposição amplamente aceita, devido à inconstância da severidade das perdas e da correlação que pode ocorrer entre a perda e a severidade, isto tendo em vista que condições econômicas adversas podem provocar um aumento na frequência do *default* além de reduzir o valor de mercado de coberturas do portfólio. Independentemente, em todos os casos o cálculo das previsões para as perdas não esperadas é preponderante para a definição do capital econômico a ser alocado pela instituição.

Dimensionar estes riscos e incorporá-los a elementos próprios do portfólio da instituição é o objetivo de diversos modelos desenvolvidos por instituições financeiras e empresas de consultoria. Como abordam Bluhm, Overbeck e Wagner (2002), estes modelos buscam mensurar o risco agregado do portfólio e suas abordagens se tornaram bastante difundidas no mercado. Dentre os principais modelos os autores destacam:

- a) Credit Metrics: Criado pelo banco J. P. Morgan é um modelo baseado no valor de mercado e avalia a volatilidade do valor de mercado do portfólio de crédito derivada da ocorrência de eventos de *default* e de alterações no *rating* dos tomadores;
- b) KMV: Modelo criado pela consultoria KMV⁹ que tem suas bases no modelo de estrutura de capital de Merton (1974). Este modelo entende que a empresa entra em *default* se o valor de mercado de seus ativos for inferior ao de seus passivos. Para tanto, calcula e monitora a distância entre o valor da firma e a taxa de *default* – a qual é baseada em um *benchmarking* de várias instituições;
- c) Credit Portfolio View: Este modelo agrupa tomadores em segmentos econômicos, em países e em faixas de *rating*, sendo que para cada segmento é previsto um modelo econométrico que busca explicar sua taxa de *default* a partir de variáveis explicativas macroeconômicas; e
- d) Credit Risk+: Modelo com base nos eventos de *default* (e não nas alterações no valor de mercado do portfólio) que tem por objetivo modelar a distribuição das perdas de portfólio a partir da distribuição das perdas dos *defaults* para posterior definição e alocação do capital econômico para sua cobertura.

⁹ KMV é a sigla para o nome da consultoria Kealhofer, McQuown & Vasicek Co.

No mercado brasileiro o Credit Metrics tem sua aplicação prejudicada por dois fatores principais, levantados por Bluhm, Overbeck e Wagner (2003), como limitações, quais sejam: número reduzido de corporações com ativos negociados em bolsa e ausência de um mercado secundário de crédito desenvolvido, fundamental para obtenção da curva de juros empregada para avaliação da carteira de crédito.

Por ser um modelo baseado em valores de mercado, o KMV também tem sua aplicação restringida no mercado de crédito brasileiro pelos mesmos motivos. Corroborando, Majumder (2006), aponta que a base de dados sobre a frequência de eventos de *default* deve ser grande, o que, para o mercado brasileiro não ocorre, restringindo suas aplicações.

Segundo Bluhm, Overbeck e Wagner (2003), o Credit Portfólio View é um modelo que exige uma série longa de observações sobre taxas de *default* para cada segmento da carteira, bem como dos fatores macroeconômicos, o que hipoteticamente não ocorre no Brasil dada a ausência dos dados para estruturação dos impactos macroeconômicos na carteira de crédito.

Para Lesko, Schlottman e Vorgrimber (2004), o Credit Risk+ é o modelo com maior aderência às diversas condições dos mercados de crédito, contudo, este modelo assume uma hipótese de que as taxas de *default* são baixas, o que encontra resistência em estudos empíricos no Brasil como apontam Brito e Assaf Neto (2008).

2.4 Alocação de capital econômico para risco de crédito

Por mais apropriada que tenha sido a análise prévia à concessão de crédito, o tomador pode ter uma deterioração de sua situação financeira no decorrer da vigência da operação de crédito contratada, gerando perdas superiores aos valores esperados.

A alocação de capital econômico é o principal mecanismo empregado para proteção contra estas perdas. Como é possível visualizar pela Figura 3, este capital serve para assimilar perdas não esperadas provenientes da carteira de crédito, garantindo a sustentabilidade financeira da instituição.

Como estabelecem Bluhm, Overbeck e Wagner (2003), para uma organização não é suficiente estabelecer uma cobertura para as perdas esperadas pela operação de crédito. Para os autores a organização deve levantar reservas também para cobrir as perdas inesperadas, as quais podem ser definidas como medidas da magnitude das perdas esperadas, isto é, definidas como a porção da perda total que supera a perda esperada, sendo seu cálculo uma estimativa de desvios da variável perda (representada pelo eixo *x* da Figura 3).

Dado que a definição das perdas não esperadas é obtida a partir das medidas de risco do portfólio como um todo, o cálculo do capital econômico exige que o risco de crédito seja obtido a partir do risco portfólio de crédito da instituição (BLUHM; OVERBECK; WAGNER, 2003).

Como apontado por Smithson (2003), o capital econômico pode ser definido como o montante de capital a ser alocado para fazer frente ao risco derivado das perdas não esperadas da instituição, o qual depende do objetivo da instituição no que tange sua taxa de risco de insolvência, podendo ser expresso como:

$$EC = q_{\alpha} - EL_{PF} \quad (2.3)$$

Em que: q_{α} é α -quantil da perda do portfólio dada por:

$q_{\alpha} = \inf\{q > 0 | P[L_{PF} \leq q] \geq \alpha\}$, sendo L_{PF} a perda para o portfólio; e

EL_{PF} a expectativa para esta perda, dada por:

$$EL_{PF} = \sum_{i=1}^m EL_i = \sum_{i=1}^m EAD_i LGD_i DP_i \quad (2.4)$$

Como tratado por Bluhm, Overbeck e Wagner (2003), caso o nível de confiança α seja de 99,98%, a cada 10.000 eventos (expresso pelos autores em anos) o capital econômico será suficiente para cobrir a perda não-esperada em 9.998. Segundo Smithson (2003), esta taxa de cobertura, isto é, o intervalo de confiança que o capital econômico alocado referencia, está diretamente relacionada com a classificação de *rating* desejado pela organização.

2.5 Classificação e mensuração de risco de crédito

Segundo Bluhm, Overbeck e Wagner (2003), as duas principais classes de modelos empregados em risco de crédito são a classe de modelos de classificação e decisão de concessão de crédito – que buscam atribuir medidas de risco ao proponente ou a uma operação de crédito, alocando-o dentro de classes e auxiliando no processo de decisão de concessão – e a classe dos modelos de mensuração de risco – os quais têm como objetivo modelar o comportamento das

variáveis relacionadas aos eventos de *default* atribuindo assim probabilidades de ocorrência deste *default*.

Entender as diferenças entre estas classes de modelo e suas interações com o risco de portfólio se mostra interessante no contexto desta pesquisa.

a) Modelos de classificação de risco

São modelos que atribuem medidas de risco com o objetivo de classificar (atribuir níveis) um tomador ou uma operação de crédito de forma contínua ou discreta sendo, tal classificação, uma função de previsões ou probabilidade de ocorrência de *default*.

Estas classificações podem ser ordinais com valores (0 = elevadíssimo risco de default, 1 = elevado risco de default, ..., n.), ou mesmo categóricas (AAA a D, como no quadro 1 que demonstra o sistema de classificação de *ratings* da agência de classificação de risco Standard & Poors).

Classe	Descrição da Classe
AAA	Melhor qualidade de crédito. Extremamente confiável quanto ao pagamento de suas obrigações;
AA	Qualidade de crédito muito boa. Confiável quanto ao pagamento de suas obrigações;
A	Mais suscetível a condições econômicas, mas ainda com boa qualidade de crédito;
BBB	Classificação mais baixa em grau de investimento;
BB	Cuidado é necessário. Melhor qualidade de crédito <i>subinvestimento</i> ;
B	Vulnerável às mudanças econômicas. Atualmente se mostra capaz de cumprir suas obrigações financeiras;
CCC	Atualmente está vulnerável ao não pagamento de suas obrigações. Dependente de condições econômicas favoráveis;
CC	Altamente vulnerável a um evento de default;
C	Perto ou já falido, contudo os pagamentos das obrigações continuam sendo feitos;
D	Pagamento interrompido em alguma obrigação financeira ocorreu atualmente.

Quadro 1 - Categorias de rating S&P

Fonte: Bluhm, Overbeck e Wagner (2003, p. 20).

Como expresso por Bessis (1998), estes modelos têm suas aplicações concentradas nos processos operacionais de análise e concessão de crédito, fornecendo avaliações de risco que permitem a aprovação ou rejeição de uma solicitação de crédito, seja de forma manual ou mesmo automatizada (reduzindo os custos operacionais de análise de crédito).

Um ponto importante a ser destacado é que tais modelos normalmente não permitem operações aritméticas de suas classes já que as classificações referem-se a categorias, podendo em certos casos assumir variáveis dicotômicas.

Segundo Bluhm, Overbeck e Wagner (2003), os principais modelos desenvolvidos nesta categoria são os modelos de *credit rating*, modelos de *credit scoring* e sistemas especialistas (modelagens computacionais compostas por algoritmos dotados de propriedades capazes de realizar julgamentos sobre o risco de uma operação de crédito).

b) Modelos de mensuração de risco

São empregados especialmente para modelar o comportamento estocástico das variáveis explicativas dos eventos de *default*. Podem ser citadas como exemplos de variáveis explicativas o valor de uma operação para um tomador ou mesmo o comportamento do *default* em uma carteira ao longo do tempo.

Estas estimativas servem tanto para indicar se o proponente ao crédito terá ou não acatada sua proposição, quanto para desenvolver estimativas para coberturas para as carteiras de crédito (alocação de capital econômico) e estimativas de prêmios pelo risco introduzidos nas taxas de retorno exigidas pelo prestador (BLUHM; OVERBECK; WAGNER, 2003).

Como apresentam Bluhm, Overbeck e Wagner (2003), os modelos de mensuração também conhecidos como modelos estocásticos de mensuração de risco de crédito são amplamente empregados por organizações que atuam com a precificação de derivativos de crédito, bem como de alguns títulos de crédito.

Não obstante, esta é uma forma de modelagem que também fornece elementos teóricos interessantes para vários dos principais modelos de risco de portfólio, os quais não tratam elementos de risco das operações individualmente (BLUHM; OVERBECK; WAGNER, 2003). Por definição, os modelos de risco de portfólio buscam quantificar o risco inerente à carteira de crédito e suas aspirações estão relacionadas à composição destas.

O emprego de modelos estocásticos na precificação de derivativos e outros títulos de crédito tem seu fundamento em sua característica de identificação e modelagem das distribuições das perdas ocorridas nas carteiras de crédito, fornecendo elementos para cálculo do VaR – *Value at Risk* destas carteiras, bem como de outras ferramentas de precificação de derivativos – tais como o modelo de Black & Scholes – podendo ainda fornecer informações para análise da taxa de *default* de uma carteira (BLUHM; OVERBECK; WAGNER, 2003).

Gundlach (2004), aponta que os tipos de modelos estocásticos de mensuração de risco de crédito mais difundidos são os modelos conhecidos como estruturais, os modelos de salto-difusão e os modelos de forma reduzida.

Os modelos estruturais tiveram seu desenvolvimento a partir dos estudos de Merton (1974), e tem como elementos principais a estrutura de capital e os riscos de *default*, modelando o comportamento do valor dos ativos da firma.

Tais modelos consideram que o capital próprio dos *shareholders* tem as características de uma opção de compra dos ativos da firma, na qual os acionistas têm um direito, mas não uma obrigação de realizar os ativos da empresa pelo valor de exercício (o qual é igual ao valor do débito da firma).

Merton (1974), entende que, caso o valor da firma seja inferior ao valor de seu débito (valor de exercício da opção), os acionistas não pagaram a dívida (não realizarão a opção), deixando os direitos sobre os ativos para os credores.

Majumder (2006), indica que o modelo de Merton (1974), também é conhecido como processo estocástico de difusão, tendo restrições quanto a sua capacidade de predição de eventos de *default*, uma vez que o valor segue um processo contínuo, assumindo uma baixa probabilidade de ocorrência de quedas repentinas no valor da firma. Sendo assim, o a probabilidade de *default* da firma no curto prazo, assim como os *spreads* de curto prazo tenderiam a zero, o que é contestado por evidências empíricas trazidas por autores como Sarrig e Warga (1989).

Os chamados modelos de salto-difusão buscam contornar este problema ao acrescentar ao modelo de Merton (1974), uma variável de salto que objetiva refletir mudanças bruscas no valor da firma derivadas de alterações nos cenários pela presença de informações relevantes.

Majumder (2006), explica que os modelos de salto-difusão preveem um limite inferior para os resultados da variável de salto, sendo que a ultrapassagem deste valor demonstraria uma situação de insolvência. Como apresentado pelo autor, instrumentos como simulações são importantes para composição de cenários probabilísticos necessários ao uso de modelos de salto-difusão.

Já o modelo de forma reduzida diverge dos dois anteriormente apresentados, uma vez que não buscam modelar a relação entre os eventos de *default* e o valor da firma. Como conceituado por Bluhm, Overbeck e Wagner (2003), estes modelos não trazem em sua formulação premissas teóricas ligadas à estrutura de capital ou às características da firma com algum evento de *default*, bem como não oferecem teoria sobre entendimento da influência do

default sobre o valor da firma. Estes modelos buscam explicar o comportamento e intensidade dos eventos de *default* em grupos específicos de uma carteira de tomadores.

Para Bluhm, Overbeck e Wagner (2003), tais características fazem deste um modelo particularmente interessante para gestão do crédito ao consumidor por sua simplicidade de avaliação e pelo grande número de consumidores que formam a carteira de crédito das instituições financeiras.

Como destacado pelos autores, modelos estruturais exigiriam muito empenho em cálculos, o que poderia inviabilizar o processo, contudo, um único modelo de forma reduzida seria capaz de descrever o comportamento da taxa de *default* da carteira de consumidores, ou mesmo de um segmento específico que se queira testar (por exemplo: idosos, MPE's, etc.).

2.6 A previsão de *default* de crédito

Como visto até o presente momento, a previsão de *default* de crédito é um elemento central na decisão de concessão e gestão de crédito. Esta previsão consiste em realizar a análise do proponente, qualificando-o como pertencente a alguma das classes pré-existent de clientes. De forma básica, as previsões podem classificar os clientes como bons ou ruins, quanto ao pagamento das obrigações contraídas.

Como abordado na seção 2.1 existem duas formas de realizar esta classificação, quais sejam a análise subjetiva e a análise objetiva. A previsão faz parte das análises objetivas e emprega normalmente métodos estatísticos para prever o comportamento de pagamento do cliente, isto é, classifica-lo como bom ou ruim (ANNIBAL; KOYAMA, 2011).

Neste sentido os modelos estatísticos realizam classificações dicotômicas dos indivíduos dentro de uma carteira de crédito. Para tanto, esses modelos selecionam grupos de treinamento e de teste dentro de uma base de dados – ou carteira de clientes – classificando os indivíduos como bons (indivíduos que pagaram suas obrigações) e ruins (indivíduos que entraram em *default* em algum momento) (MATIAS, 2007).

Dentro do grupo de teste é verificada a capacidade de o modelo identificar, com base em um conjunto de variáveis, se o indivíduo será um bom ou mau pagador. Em geral, modelos probabilísticos indicam a probabilidade de o indivíduo pertencer a um dos grupos (bom ou mau pagador) em função das variáveis estudadas.

Segundo Duan e Shrestha (2011), os modelos mais difundidos e tradicionalmente utilizados para previsão de *default* de crédito são os modelos baseados em técnicas estatísticas de regressão discriminante – tais como o modelo Altman Z-score – e as regressões logísticas.

Em seu estudo, Duan e Shrestha (2011), apresentam comparativamente outras quatro técnicas atualmente empregadas para o tratamento do problema de previsão de *default* de crédito, quais sejam: redes neurais artificiais, redes bayesianas, *PIM – Poisson Intensity Model*, e *SVM – Support Vector Machine*.

Para Pascual, Martínéz e Alamillos (2013), nota-se um aumento no uso de redes bayesianas para problemas de previsão e classificação de risco de crédito dada a eficiência destes modelos para tratamento de problemas de classificação.

Brown e Mues (2012), apontam a ampliação do uso de classificadores baseados em técnicas mais modernas de aprendizagem estatística, tais como o já citado *SVM – Support Vector Machine*.

Como forma de aprofundar o conhecimento sobre o tema, em especial entendendo o tratamento dado ao problema de classificação em análise de risco de crédito e selecionando as técnicas de aprendizagem estatística mais empregadas no âmbito da análise de risco de crédito, a pesquisa apresenta na seção 2.7 o estado da arte sobre modelagem aplicada à análise de risco de crédito, contendo pesquisas e aplicações recentes na área.

2.7 Estado da arte sobre em modelagem de análise de risco de crédito

A área de análise de risco de crédito é bastante rica no Brasil, produzindo diversas pesquisas, especialmente em modelagem. Como uma forma de analisar os recentes desenvolvimentos nesta área, a pesquisa realizou buscas em bases de dados para a confecção de um levantamento bibliográfico das produções na área de análise de risco de crédito, englobando artigos e *pappers* produzidos no Brasil e publicados tanto em revistas nacionais quanto internacionais nas áreas de Finanças, Econometria, Informática e Pesquisa Operacional.

As bases de dados selecionadas foram as bases CAPES, Scopus, Science Direct, Web of Science e a base Scielo. Para estas buscas foram utilizadas as chaves: *credit risk*, risco de crédito, *credit scoring* e *credit rating* nos títulos, resumos e palavras-chave dos trabalhos.

As pesquisas incorporaram informações dos últimos 2 anos no caso internacional e dos últimos 15 anos no caso brasileiro em cada uma destas bases, sendo selecionados apenas aqueles estudos cuja contribuição científica está centrada na construção e análise empírica de modelos voltados para o tratamento do problema de classificação de crédito e previsão de *default*, tanto para pessoa física quanto jurídica.

Excluídos os dados repetidos, tendo como foco o cenário de crédito brasileiro, foram encontrados um total de 27 (vinte e sete) artigos dentro do período, sendo que, deste total 14 – catorze – artigos abordaram a problemática da previsão de *default*.

Considerando o objetivo desta pesquisa, estes 14 artigos foram selecionados para compor o quadro teórico de pesquisas sobre modelagem e técnicas de previsão de *default* no Brasil e estão expostos em 2.7.1.

2.7.1 Pesquisas nacionais

As pesquisas abaixo expressam brevemente os recentes avanços na área de análise de risco de crédito no Brasil.

Santos e Famá (2007), apresentaram um modelo baseado no método de regressão múltipla para tratamento do problema de concessão de crédito rotativo a clientes pessoa física que englobou variáveis sistêmicas e não-sistêmicas com o objetivo de minimizar a inadimplência de uma carteira de crédito de um Banco Comercial.

Os resultados indicam que a incorporação de fatores sistêmicos (variáveis sistêmicas) aos modelos de *credit scoring* se mostra particularmente importante por incorporar o impacto do ambiente sobre a renda do requisitante de crédito ao longo do tempo.

Bressan et al. (2011), empregaram um modelo *Logit* através das variáveis do sistema *Pearls* de informações contábeis (reconhecido pelo Conselho Mundial de Cooperativismo de Poupança e Crédito) em cooperativas de Minas Gerais entre janeiro de 1995 e maio de 2008.

O objetivo do estudo foi verificar a eficiência destas variáveis na previsão da insolvência de cooperativas de crédito mineiras. Os resultados demonstraram boa aderência das variáveis na previsão e análise da insolvência do conjunto de cooperativas de crédito estudadas.

Garter, Moreira e Galves (2008), oferecem um modelo no qual empregam o método multicritério AHP para tratamento de problemas de classificação de risco de crédito corporativo setorial. Este estudo teve como objetivo oferecer um mecanismo para classificar, com base em um sistema de julgamentos e pesos, cada segmento de negócio das firmas presentes em um portfólio bancário em cinco classes de riscos (*ratings* de A a E).

O estudo foi particularmente interessante por analisar variáveis endógenas e exógenas à firma, quais sejam: tendências de mercado, mercado externo, estrutura do setor, alavancagem financeira, elasticidade e meio ambiente.

Zambaldi et al. (2011), analisaram as decisões de concessão de crédito de um banco brasileiro para um grupo de mais de 65 mil solicitações de crédito pequenas e médias empresas.

Seu modelo empregou um método probito hierárquico para verificar a adequação do modelo de *credit scoring* à decisão de concessão ou não de crédito a este tipo de empresa, buscando também verificar as principais dificuldades deste tipo de organização para aquisição de crédito.

Os resultados sugerem que as decisões estão correlacionadas com o modelo empregado pelo banco e que elementos como o custo, a dependência de colaterais e restrições derivadas de assimetria de informações são os principais fatores que inibem a concessão de crédito a este tipo de empresa no mercado brasileiro.

Steiner et al. (2007), testaram um modelo de rede neural para previsão de *default* através de simulações de solicitação de crédito como suporte à tomada de decisão de concessão. O objetivo do estudo foi apresentar a eficiência da técnica de extração de regras de classificação *NeuroRule* no processo de classificação de redes neurais treinadas.

De acordo com os autores, a técnica de extração de regras se mostrou eficiente apresentando resultados de acurácia superiores a 80% quanto à decisão de crédito bancário para todas as simulações realizadas.

Brito e Assaf Neto (2008), desenvolveram pesquisa com o objetivo de propor um modelo de classificação (*credit scoring*) para realização de avaliações de risco de crédito corporativo no Brasil. O estudo empregou um método de regressão logística tendo como variáveis índices financeiros extraídos das demonstrações contábeis de empresas de capital aberto entre o período de 1994 e 2004. Os resultados demonstraram, segundo os autores, bom nível de acurácia do modelo indicando a eficiência de índices financeiros para a previsão de solvência e insolvência de empresas brasileiras.

Guimarães e Moreira (2008), também estudaram a eficiência do emprego de um modelo de previsão de insolvência com base em indicadores contábeis extraídos de empresas de capital aberto no mercado brasileiro.

Através de uma análise discriminante os autores indicam que o estudo foi capaz de confirmar o poder discriminatório dos índices financeiros sobre estrutura de ativos, estrutura de capital e geração de caixa na previsão de insolvência, especialmente no que tange o montante de recursos próprios disponíveis, sendo esta a variável identificada no modelo como a detentora do maior poder discriminatório.

Dantas e Souza (2008), propuseram um modelo para classificação e decisão de crédito (*credit scoring*) baseado na estrutura de capital de 241 empresas brasileiras da indústria manufatureira com o objetivo de verificar a eficiência do emprego de parâmetros da estrutura de capital como mecanismo direcionador das decisões de crédito.

Para tanto, o estudo utilizou um modelo logístico sendo as variáveis tratadas por análise fatorial e selecionadas por regressão discriminante pelo método *stepwise*. Os resultados obtidos pelos autores sugerem boa aderência do emprego de parâmetros da estrutura de capital para tratamento de problemas de concessão de crédito corporativo para o perfil de organização analisado no estudo.

Guimarães e Chaves Neto (2002), propuseram um método para análise e classificação de concessão de crédito pessoal para teste de regras de reconhecimento de padrões por meio de uma função linear discriminante e de regressão logística.

Este estudo teve como foco a concessão de limites em cartões de crédito para pessoas físicas, sendo os dois modelos avaliados em sua eficiência pelo método de *Lachenbruch*. Os dados foram retirados de uma administradora de cartões contendo 707 dados dos quais 102 entraram em *default* à priori. O método de extração incorporou maiores resultados ao poder de predição do modelo logístico em relação ao modelo discriminante no que tange à acurácia.

Godói, Yoshino e Oliveira (2008), estudaram mecanismos de mensuração e cobertura de risco de crédito em operações de compra de debêntures no mercado brasileiro. De fato este não é um modelo de previsão de risco de crédito, mas incorpora a característica por empregar um modelo VaR (*Value at Risk*) para mensuração e composição de uma carteira ótima de investimentos, abrindo espaço também para o uso deste método para teste da eficiência de outras carteiras de crédito.

Brito, Assaf Neto e Corrar (2009), propuseram o emprego de uma análise de conglomerado baseada em variáveis contábeis de empresas brasileiras como mecanismos para classificação destas empresas em oito classes distintas (*ratings*).

Para tanto, os autores atribuíram *ratings* para empresas não financeiras listadas em bolsa no período de 1994 a 2006. Além da classificação os autores analisam a matriz de migração de crédito formada pelo portfólio das empresas no período analisado.

Em seus resultados os autores constatarem uma piora nas classificações à priori dos *ratings* das empresas antes do *default*, provocando alterações na matriz de migração de crédito. Tal fato indica, segundo os autores, que as informações contábeis são suficientemente robustas para a análise de risco de crédito corporativo no Brasil.

Minussi, Damacena e Ness Jr (2002), propuseram um modelo de regressão logística para avaliação da concessão de crédito a uma carteira formada por 323 clientes empresariais de um banco brasileiro, sendo todos de um mesmo segmento industrial. O estudo testou um modelo econométrico com 5 variáveis independentes validado pelo método *cross validation* com

amostras equilibradas de 132 indivíduos, sendo os resultados de acurácia bastante elevados, com cerca de 94,85% de acerto.

Os estudos de Louzada-Neto e Ara (2012b) e Louzada-Neto, Ferreira-Silva e Diniz (2012), são os únicos identificados na pesquisa realizada nos indexadores nacionais pelos métodos de seleção empregados a utilizar redes bayesianas em previsão de *default*. Na pesquisa de Louzada-Neto e Ara (2012b), os autores demonstram a aplicabilidade de uma rede KDB – *K dependence Bayesian network* – para um problema de *credit scoring*.

Os autores comparam a técnica Naïve Bayes e KDB (KDB-1, KDB-2 e KDB-3) com outras técnicas, quais sejam: análise discriminante, regressão logística, regressão probito e redes neurais para três conjuntos de dados. As métricas estudadas pelos autores são estatísticas de acurácia – ACC, Sensibilidade – SEM, especificidade – SPE e o coeficiente de correlação de Matthew – MCC. Os resultados demonstraram bom desempenho para as métricas de redes bayesianas estudadas (KDB e Naïve Bayes) frente aos demais modelos estudados.

Louzada-Neto, Ferreira-Silva e Diniz (2012), analisaram o impacto de amostras desproporcionais em técnicas de classificação e concessão de crédito para pessoa física empregando como métricas de comparação a sensibilidade, especificidade, valores preditivos, acurácia, coeficientes de correlação e técnicas de teoria de informação, tais como entropia relativa e informações mútuas.

Os resultados foram estimados através de simulações realizadas com base em um conjunto de dados de um banco brasileiro e sugerem que não há diferenças significativas em termos de capacidade preditiva para amostras desproporcionais para os modelos de regressão logística e modelos naïve logístico testados.

Figura 4 - Aplicações no Brasil

Contudo, os autores indicam que ocorre grande diferença entre estes modelos para amostras balanceadas, onde ocorre uma subestimação das probabilidades quando estimadas pelo modelo naïve logístico, que é, como destacado pelos autores, um modelo de rede probabilística ou rede bayesiana.

Na Figura 4 é possível observar a compilação dos dados sobre as técnicas empregadas no Brasil em âmbito acadêmico para tratamento dos problemas de análise e concessão de crédito. Estes resultados estão compilados no quadro 2 que traz os trabalhos que incorporaram tais técnicas.

Os dados das figuras foram obtidos a partir da observação das técnicas utilizadas em cada uma das pesquisas selecionadas, sendo que, nas ocasiões nas quais houveram comparações de técnicas, todas as técnicas comparadas foram levantadas.

Técnica	Autores
Regressão Logística	Brito e Assaf Neto (2008), Guimarães e Chaves Neto (2002), Minussi, Damacena e Ness Jr (2002), Dantas e Souza (2008), Louzada-Neto e Ara (2012b), Louzada-Neto, Ferreira-Silva e Diniz (2012).
Análise Discriminante	Guimarães e Moreira (2008), Guimarães e Chaves Neto (2002), Louzada-Neto e Ara (2012b).
Modelo Probit/Logit	Bressan et al. (2011), Zambaldi et al. (2011), Louzada-Neto e Ara (2012a).
Regressão Múltipla	Santos e Famá (2007).
Outros: multicritério, conglomerados, VaR	Garter, Moreira e Galves (2008), Godói Yoshino e Oliveira (2008), Brito, Assaf Neto e Corrar (2009).
Redes Neurais	Steiner et al. (2007), Louzada e Ara (2012b).
Redes Bayesianas	Louzada e Ara (2012b), Louzada-Neto, Ferreira-Silva e Diniz (2012).

Quadro 2 - Técnicas por autores no Brasil.

Como é possível observar pelo estudo teórico apresentado, no Brasil, as técnicas comumente empregadas são analisadas quanto a eficiência preditiva para determinação de sua adequação para tratamento do problema de crédito.

2.7.2 Pesquisas internacionais

Para pesquisas internacionais foram selecionados os artigos indexados nas bases de Scopus e Science Direct. Como chaves de busca foram selecionados por abstract, palavras-chave ou título os artigos que contivessem os termos *credit rating* e *credit scoring* para os dois últimos anos em língua inglesa.

Este critério foi selecionado para limitar o número total de artigos publicados na área em escala mundial, bem como cumprir o objetivo de expressar as pesquisas recentes na área. Como resultado a busca retornou um total de 38 artigos indexados nestas duas bases para as áreas *Business and Account Management*, *Economy*, *Econometrics and Finance*. Deste total, foram selecionados 26 artigos para compor esta seção por trazerem discussões a respeito de modelagem para a área de risco de crédito. Estes estudos são apresentados abaixo.

Com base em simulações, Elliott, Siu e Fung (2014), testaram a capacidade discriminatória do modelo *Altman Z-score* com dados gerados a partir de um sistema de simulação – *Double Hidden Markov Model*, DHMM – baseado em dados contábeis de um conjunto de empresas asiáticas e da Oceania. As empresas foram classificadas em sete categorias, sendo estas as mesmas empregadas na classificação da Standard & Poors.

A pesquisa emprega algoritmos de aprendizagem com métodos de filtragem para aumentar a capacidade do modelo e possibilitar uma autocalibração do mesmo. Os resultados demonstram boa adequação do método, contudo com maior sensibilidade do modelo a alterações nas taxas contábeis do que o modelo da Standard & Poors.

Brown e Mues (2012), compararam diversas técnicas de classificação empregando-as em conjuntos desbalanceados de dados de crédito para classificação com o objetivo de verificar o desempenho destas técnicas para esta característica de dados que, segundo os autores, é a característica padrão dos conjuntos de dados de crédito.

Os autores empregaram técnicas de regressão logística, redes neurais, árvore de decisão, análise discriminante, método de impulsão (*boosting*), mínimos quadrados, *support vector machine* e *randon forests*. Todas estas técnicas foram testadas segundo sua performance preditiva pelas métricas ROC – *Receiver operating characteristic*, Estatística de Friedman's e Nemenyi *post hoc*, desbalanceando progressivamente o conjunto de dados a partir da base equilibrada, realizando seleções aleatórias e rodando os testes.

Em seus resultados Brown e Mues (2012), indicam que os métodos classificados como *ensemble* (métodos que misturam outros classificadores com o objetivo de maximizar o resultado preditivo), tais como *boosting* gradiente e *randon forests*, são melhores para dados desbalanceados. Por fim, os autores sugerem que, na medida em que as classes ficam mais desbalanceadas, métodos baseados em árvore de decisão e discriminante se tornam mais ineficientes para o tratamento desta problemática.

Doumpos et al. (2014), desenvolvem um sistema de classificação para predição de *ratings* oferecidos por CRA's¹⁰. Em seu estudo os autores empregaram um sistema de classificação multicritério que combinou dados contábeis com um modelo de equações estruturais para previsão de *default* buscando predições melhoradas, bem como testando o incremento de informação a partir do modelo estrutural.

Os resultados obtidos a partir de um conjunto de dados de empresas europeias no período de 2002 à 2012 indicaram que o modelo estrutural é capaz de incorporar informação significativa aos modelos preditivos. Não obstante, os autores indicam que esta capacidade é consideravelmente pequena quando comparada com a capacidade informativa quando o mercado é também considerado no modelo.

Já Hens e Tiwari (2012), estudaram formas de minimizar o esforço computacional para problemas de *credit scoring* utilizando a técnica de *machine learning SVM – Support Vector Machine*. Para tanto os autores extraíram características de dois conjuntos de dados por meio do modelo e rodaram o teste apenas para uma amostra ao invés de rodar no banco de dados todo, como é comumente realizado.

Segundo os autores a acurácia do modelo não foi significativamente alterada pelo método e o tempo e esforço computacional foi consideravelmente inferior, tornando este um método competitivo para emprego em problemas de *credit scoring*.

Kruppa et al. (2013), desenvolveram um modelo baseado em *machine learning* para estimar a probabilidade individual para risco de crédito pessoa física. Para tanto, os autores utilizaram um grande banco de dados com aproximadamente 13% de clientes em situação de default. Em seu *framework* os autores estabelecem seis passos para obtenção das probabilidades de *default* de consumidores, iniciando pelo treinamento via diversos algoritmos, *reamostragens* e estimações. Seus resultados são estimados por uma regressão logística, pelo método *ensemble random forests* e por *k-nearest neighbors*. O desempenho foi analisado com base na curva ROC, demonstrando maior desempenho do modelo *ensemble*.

Blanco et al. (2013), empregam um modelo de redes neurais artificiais (MPL – *Multilayer Perceptron*) em uma proposição de modelo para *credit scoring* aplicado a instituições do mercado de micro finanças. Através de uma amostra de 5.500 clientes empresarias de uma instituição peruana os autores aplicam a técnica e comparam os resultados com as técnicas de regressão linear discriminante, regressão discriminante quadrática e regressão logística.

¹⁰ CRA's – Credit Rating Agencies – Agências de risco de crédito.

Segundo os autores, os resultados obtidos a partir dos testes demonstram maior eficiência das redes neurais artificiais em comparação com as técnicas anteriores, tanto em relação à curva ROC quanto em relação à matriz de custos de erros de classificação.

De-La-Veja et al. (2013), também estudaram métodos baseados em algoritmos de classificação para ambientes de micro finanças no Peru através de algoritmos de aprendizagem estatística. Os resultados obtidos corroboram com os resultados de Blanco et al. (2013), indicando que os modelos MLP são eficientes para este propósito.

Marqués, García e Sánchez (2012), analisaram sete classificadores individuais aplicáveis a problemas de *credit scoring* sendo estes utilizados em cinco diferentes classificadores ensembles com o objetivo de determinar quais classificadores deveriam ser empregados para produção de resultados superiores em cada ensemble.

Os resultados apontam para o método conhecido como árvore de decisão C4.5 (algoritmo de classificação de rede neural) como o melhor dentre os métodos ensemble, sendo que os métodos de redes bayesianas Naïve Bayes, regressão logística e método de vizinho mais próximo parecem ser significativamente inferiores em termos de poder preditivo.

Manso (2013), analisou o efeito das classificações de CRA's sobre a situação das empresas classificadas. Em seu estudo o autor aponta três pontos fundamentais sobre o comportamento das agências, quais sejam: i) a *acurácia* dos *ratings* das agências não deve ser o único foco destas, mas também o efeito destes *ratings* sobre a probabilidade de sobrevivência dos agentes classificados; ii) sempre que as agências optam por uma política baseada simplesmente na *acurácia* pode ocorrer rebaixamentos imediatos em resposta a pequenos choques dos fundamentos econômicos e iii) a competição entre as agências de *rating* pode levar a um efeito de rebaixamento dos *ratings*, elevação da frequência de *default* e redução da qualidade dos tomadores.

Alissa et al. (2013), analisaram, por meio de uma regressão múltipla, o efeito da classificação de *ratings* sobre a gestão de lucros e políticas de governança corporativa no período de 1985 a 2010, sendo os dados obtidos a partir de conjuntos de dados da CRA Standard & Poors. Segundo os autores existem evidências de que as firmas usualmente gerenciam seus retornos crescentes ou decrescentes quando entendem que suas classificações de *ratings* estão satisfatórias ou não em relação a suas expectativas.

Não obstante, os autores sinalizam para uma habilidade em gerenciamento de seus *ratings* por parte das empresas neste sentido, indicando que as mesmas obtiveram, em média, êxito nesta gestão para maximização de seus resultados. Este mesmo resultado, através da mesma metodologia, foi observado por Dermitas e Cornaggia (2013).

Bijak e Thomas (2013), analisaram se estratégias de segmentação de mercado são eficientes em todas as situações para aplicação de modelos de *credit scoring*, perseguindo a hipótese de que ao segmentar uma carteira de clientes o poder preditivo de um modelo é maximizado. Para tanto, o estudo empregou duas técnicas de segmentação e analisou os resultados com base em regressão logística.

Os dados foram obtidos a partir de dois conjuntos de dados de bancos do Reino Unido e de um terceiro do European Credit Bureau. O primeiro banco de dados do Reino Unido era composto de 7.835 dados dos quais 6.440 não observaram eventos de *default* e o segundo era composto de 39.858 eventos dos quais em 38.135 não se observaram *defaults*. O conjunto de dados do European Credit Bureau foi composto de 186.574 eventos dos quais 179.544 não eram classificados como em *default* à priori.

Os testes foram realizados com cada tipo de segmentação, bem como sem qualquer segmentação dos dados e indicaram que nem sempre a segmentação dos dados aumenta a performance dos modelos de *credit scoring*. De fato, os autores indicam que em nenhum dos conjuntos de dados os resultados da segmentação foram significativamente superiores aos resultados da não segmentação de clientes.

Segundo Crone e Finlay (2012), existe atualmente uma tendência em empregar amostras de tamanho na casa de 1500 dados como sendo suficientemente grandes para o problema de *credit scoring*, sendo que quando estas são tendenciosas é importante realizar um balanceamento das amostras. Para os autores este é um tema pouco estudado, sendo uma lacuna da área a análise do efeito do tamanho da amostra sobre a acurácia.

Desta forma, em seu estudo, Crone e Finlay (2012), analisam a questão da amostra para a predição do comportamento de pagamento de consumidores pessoa física, avaliando os ganhos em acurácia em métodos de regressão logística, análise discriminante, árvores de decisão e redes neurais em dois conjuntos de dados balanceados. Os autores encontram evidências de que o emprego de amostras maiores (superiores a 1500 observações) promove uma elevação da acurácia dos modelos.

Sohn e Kim (2012), buscaram desenvolver um sistema de *credit scoring* que refletisse o fenômeno das empresas *star-ups*, as quais, segundo os autores, não têm suas características cobertas pelos modelos usualmente empregados e que têm seus créditos condicionados a sistemas de garantias. Neste estudo, os autores desenvolveram um modelo baseado em árvores de decisão indicando que este método é capaz de prover valiosas informações para o analista deste tipo de organização.

Harris (2013), analisou o desempenho do classificador *SVM – Support Vector Machine* – para a predição do comportamento de pagamento para clientes considerados em *default* com atraso inferior e superior a 90 (noventa) dias. O estudo considerou 20 (vinte) variáveis de 250 mil dados de clientes pessoa física entre 1997 e 2012 de uma instituição financeira de Barbados.

Os resultados demonstraram que o desempenho do modelo que utilizou a definição de cliente em *default* como aquele com menos de 90 dias de atraso apresentou resultados superiores significativamente em predição, indicando que o modelo sofre diferenças substanciais quando a definição de *default* considera clientes com atrasos superiores a 90 dias.

Wang et al. (2012), indicaram que uma grande quantidade de informações redundante é observada nos conjuntos de dados de crédito, o que provoca uma redução na acurácia, bem como aumenta a complexidade dos modelos de *credit scoring*. Para resolver esta problemática os autores propõem um método para seleção das categorias necessárias em um conjunto de dados com um grande número de variáveis.

Para testar seu método (baseado em heurística e denominado RSFS – *Rough sets and scatter search* para otimização da informação), os autores empregaram modelos de redes neurais, árvore de decisão e regressão logística. Seus resultados demonstram que o método de extração de características RSFS apresenta performance superior, reduzindo o esforço computacional, bem como melhorando a acurácia dos métodos de classificação.

Bücker, Kampen e Krämer (2013), estudaram a presença de clientes que tiveram créditos negados dentro de um conjunto de dados de um banco Alemão. Segundo os autores este é um tipo de dado que precisa ser considerado pelos modelos, tendo em vista a capacidade de informação dos mesmos, em especial sobre o ambiente ao qual o modelo irá ser empregado.

Os autores compararam o ajuste de um modelo de regressão logística em caso de presença e sem presença de créditos negados no conjunto de dados e desenvolvem um *framework* para tratamento destes casos, indicando que pode-se imaginar tal situação como a existência de taxas de não-resposta em pesquisas clínicas.

Os resultados indicam que o modelo de tratamento de dados adotado pelos autores é interessante por maximizar a capacidade preditiva dos modelos quando em condições de elevada presença de créditos negados.

Nikolic et al. (2013), apresentaram um modelo baseado em regressão logística para clientes *corporate* na Sérvia. Através de um banco de dados com 7590 empresas e 350 indicadores calculados para cada empresa, os autores realizaram uma análise de cluster reduzindo estas variáveis para 24 e aplicaram, após esta redução, a regressão logística. Os

resultados foram analisados segundo a acurácia e performance preditiva do modelo, indicando, segundo os autores, boa aderência e capacidade preditiva.

Pascual, Martín e Alamillos (2013), estudaram a capacidade preditiva das redes bayesianas para dados balanceados e desbalanceados em um conjunto de dados com 19 variáveis e com 1767 registros de clientes de uma empresa espanhola, dos quais 167 apresentaram *default*. Os dados foram testados por 16 diferentes métodos que envolveram redes neurais, redes bayesianas e regressão logística.

Os autores indicam que o emprego das redes bayesianas, desde que empregado o método adequado de seleção de variáveis e tratamento de dados, oferece um excelente resultado para a predição de problemas de *credit scoring*, não apenas por sua capacidade preditiva, mas também por sua capacidade de incorporar informações de especialistas.

Figura 5 - Aplicações internacionais

Na Figura 5 é possível observar a compilação dos dados sobre as técnicas empregadas nos últimos dois anos e publicadas internacionalmente (em língua inglesa) em âmbito acadêmico para tratamento dos problemas de crédito. Estes resultados estão compilados no quadro 3 que traz os trabalhos que compõem a figura.

Como na Figura 4 a Figura 5 também se baseou no uso das ferramentas pelas pesquisas selecionadas, sendo que, quando ocorreram pesquisas de cunho comparativo, todas as técnicas comparadas foram levadas em consideração para a confecção da figura.

Como é possível observar pelo estudo teórico levantado e apresentado, no âmbito internacional, assim como ocorre no Brasil, as técnicas comumente empregadas são analisadas

em termos de sua acurácia para determinação da adequação do modelo para tratamento do problema. Nessa esfera é comum também a observação do uso da matriz de custo como mecanismo de comparação do desempenho econômico dos modelos.

Técnicas	Autores
Árvore de decisão	Brown e Mues (2012), Crone e Finlay (2012), Sohn e Kim (2012), Wang et al. (2012).
Regressão Logística	Brown e Mues (2012), Marqués, García e Sánchez (2012), Bijak e Thomas (2013), Crone e Finlay (2012), Wang et al. (2012), Bucker, Kampen e Krämer (2013), Nikolic et al. (2013), Pascual, Martínéz e Alamillos (2013).
Análise Multicritério	Doumpos et al. (2014).
Análise Discriminante	Elliott, Siu e Fung (2014), Brown e Mues (2012), Marqués, García e Sánchez (2012), Crone e Finlay (2012).
Outros	Manso (2013), Alissa et al. (2013).
Redes Neurais	Brown e Mues (2012), Blanco et al. (2013), De-La-Veja et al. (2013), Marqués, García e Sánchez (2012), Crone e Finlay (2012), Wang et al. (2012), Pascual, Martínéz e Alamillos (2013).
Redes Bayesianas	Brown e Mues (2012), Hens e Tiwari (2012), Kruppa et al. (2013), Marqués, García e Sánchez (2012), Harris (2013), Pascual, Martínéz e Alamillos (2013)

Quadro 3 - Técnicas por autores internacionais

Através de uma comparação das figuras de ambas as pesquisas (nacional e internacional) observa-se grande uso de regressões logísticas como métodos de análise de risco de crédito. Contudo, um comportamento notado e que diverge em essência os tipos de pesquisas levantados nacionalmente e internacionalmente está no empenho da pesquisa internacional em comparar diversas técnicas, sendo as técnicas de regressão discriminante e logística as mais empregadas neste sentido.

Nota-se que a pesquisa internacional levantada demonstra certa preferência às técnicas de aprendizagem, tanto aquelas voltadas para aprendizagem de máquinas, especialmente as técnicas de redes neurais artificiais, quanto as técnicas de aprendizagem estatística (tais como: redes bayesianas e SVM), enquanto a pesquisa nacional está empenhada na discussão teórica, análise e validação das variáveis.

Desta forma, o levantamento teórico realizado sugere que o padrão da pesquisa nacional está focado no fornecimento de modelos voltados para a seleção de variáveis e posterior mensuração da acurácia. Por outro lado, o padrão de pesquisa internacional está na comparação

de diversos modelos em vários conjuntos de dados e situações, buscando apresentar qual se mostra mais eficiente quanto a acurácia.

Ainda que ocorra tal diferença, nota-se certa equivalência de raciocínio quanto à forma de mensurar a eficiência de cada modelo, sendo está baseada normalmente na eficiência preditiva dos mesmos, não considerando variáveis financeiras para tanto.

A presente tese trata, busca contribuir neste aspecto, propondo métricas para análise de modelos de risco de crédito que estejam baseadas na eficiência financeira e não nas tradicionais medidas estatísticas. No capítulo 3 são apresentadas as técnicas estatísticas selecionadas para a composição da pesquisa, as quais são classificadas como técnicas de Aprendizagem Estatística (JAMES et al., 2013).

3 APRENDIZAGEM ESTATÍSTICA

Atualmente uma problemática tem ganhado destaque dentro dos estudos em diversas áreas das ciências, qual seja a análise de massas de dados. Estas massas de dados produzidas, especialmente com os recentes avanços em informática, precisam ser armazenadas, tratadas e analisadas para que organizações, governos, cientistas, etc. tenham condições de gerar conhecimento e tomar decisões eficientes a partir das mesmas.

Como definido introdutoriamente, a aprendizagem estatística é a área das ciências estatísticas dedicada à interpretação de massas de dados, sendo uma de suas aplicações a interpretação, entendimento e análise de dados em finanças (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Segundo James et al. (2013), a aprendizagem estatística envolve ferramentas de dois tipos, quais sejam as ferramentas de aprendizagem supervisionadas e as ferramentas de aprendizagem não supervisionada.

Os modelos de aprendizagem supervisionada são aqueles que buscam prever uma saída (ou conjunto de saídas) do modelo, com base em uma ou mais entrada. Já no modelo de aprendizagem não supervisionada existem as entradas, porém não existem saídas a serem previstas no sistema. Segundo James et al. (2013), dadas as propriedades da aprendizagem supervisionada e as características dos problemas, esta é a mais empregada para problemas nas áreas de negócios, em que se inclui a análise de risco de crédito.

Como abordado por Hastie, Tibshirani e Friedman (2008), em sua maioria, as recentes evoluções das técnicas de aprendizagem estatística são baseadas em técnicas clássicas. Esta posição é corroborada por autores como James et al. (2013), que esclarecem que ainda que o termo aprendizagem estatística seja recente, as ferramentas que compõem esta área são bastante conhecidas dentro de diversas áreas de estudo, tais como as áreas de economia e negócio. Segundo os autores, fazem parte do conjunto destas ferramentas técnicas como: regressões lineares, regressões discriminantes, regressões logísticas, modelos generalizados, dentre outras.

Para a composição do conjunto de técnicas analisadas na presente pesquisa foram estudadas as técnicas clássicas de análise discriminante e regressão logística, bem como as recentes técnicas de aprendizagem estatística redes bayesianas – aqui estudadas em seus modelos Naïve Bayes e KDB – *k-dependence bayesian network*, e os classificadores SVC – *Support Vector Classifier* e SVM – *Support Vector Machine*.

3.1 Análise Discriminante

A análise discriminante é uma técnica tipicamente utilizada para classificação de uma entidade (indivíduo ou objeto) com base em suas características (variáveis) em um número finito (g) de classes, grupos ou categorias, assumindo que a entidade pertence a um destes g grupos (MCLACHLAN, 2001).

Como indicado por Abdou e Pointon (2011), este método foi inicialmente proposto por Fisher como uma forma de discriminar e classificar indivíduos em grupos. Para o presente estudo foi utilizada a equação linear discriminante que tem forma funcional definida por:

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (3.1)$$

Em que z representa a variável predita, α é uma constante e $\beta_1 \dots \beta_n$, representam os coeficientes das variáveis e $x_1 \dots x_n$ são as variáveis discriminantes.

Na área de risco de crédito, modelos como Altman's Z Score são exemplos do emprego de método de classificação que, neste caso específico, empregou taxas contábeis como variáveis *preditoras* da insolvência de empresas (DUAN; SHRESTHA, 2011).

Segundo Abdou e Pointon (2011), diversos autores entendem que dadas as suas características, a análise discriminante é uma das principais técnicas empregadas para classificação de proponentes de crédito.

Um dos principais aspectos que beneficiam o emprego da análise discriminante é a sua semelhança aos modelos de regressão múltipla, o que facilita a interpretação de parâmetros e a análise do peso relativo das variáveis (poder de discriminação das variáveis), bem como facilita a replicação de estudos utilizando esta técnica, servindo também de base de comparação para análise de resultados de outras técnicas.

Suas debilidades, segundo Abdou e Pointon (2011), estão associadas ao uso de modelos lineares ou mesmo quadráticos imperfeitos para os dados, probabilidades *à priori* estabelecidas de forma inapropriada, dentre outros, que provocariam uma baixa acurácia.

3.2 Regressão Logística

A regressão logística é, segundo Mclachlan (2001), uma das técnicas mais utilizadas em situações nas quais a variável dependente assume valores dicotômicos. Especificamente na área

de *credit scoring* Pascual, Martín e Alamillos (2013), apontam que esta técnica é comumente utilizada, oferecendo resultados relevantes em termos de acurácia.

Esta técnica consiste em estimar a probabilidade de ocorrência ou não de um evento qualquer em função de um conjunto de variáveis. Sua forma funcional é dada por:

$$P = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad (3.2)$$

Em que x_n é um vetor formado pelas variáveis independentes, ou covariáveis como também são conhecidas e β é um vetor formado pelos parâmetros do modelo a serem estimados.

Segundo McLachlan (2001), o pressuposto fundamental da abordagem logística é que as variáveis são lineares (ainda que não necessariamente em sua base), o que é equivalente a tomar as chances de ocorrência do evento baseado no logaritmo de razão e sendo sua estimação realizada por método de máxima verossimilhança.

A principal diferença entre a análise de regressão linear e logística é a característica da variável dependente, que no caso da regressão logística poderá assumir valores de uma variável binária, auferido a partir de uma estimação de máxima verossimilhança, podendo incorporar um conjunto de variáveis numéricas e não numéricas.

Seus resultados são gerados em uma escala de 0 a 1 o que pode ser interpretado como a probabilidade da variável dependente ser verdadeira (ou falsa dependendo do tipo de estimação que está sendo proposta no modelo).

No caso da análise de risco de crédito, usualmente são realizadas as previsões de o indivíduo entrar em *default* em algum momento. Desta forma, quando mais se aproximar de 0, maior seria a probabilidade do indivíduo ser classificado *a posteriori* (treinamento) como um cliente ruim.

Abdou e Pointon (2011), indicam que esta propriedade torna a regressão logística mais adequada do que regressões lineares, uma vez que são observadas as duas classes possíveis de crédito (bom ou ruim), o que facilita a interpretação dos resultados destas regressões.

3.3 Redes bayesianas

Atualmente os grafos probabilísticos, também conhecido como redes bayesianas, têm ganhado notoriedade dentro de várias áreas de ciências, especialmente na modelagem estatística

para problemas nos quais se mostra interessante entender as interações entre variáveis aleatórias (COWELL; VERRALL; YOON, 2007).

É possível definir um grafo probabilístico como uma combinação entre elementos da teoria de probabilidade e teoria de grafos. Estes métodos começaram a ser desenvolvidos a partir da década de 1980, através de desenvolvimentos em modelagem estatística e inteligência artificial que geraram algoritmos computacionais voltados para a exploração de interdependências condicionais em modelos gráficos, especialmente para estimação das distribuições marginais à posteriori, e que serviram para estimação de relações causais entre variáveis (YOON, 2003).

Considerando as questões de certeza, risco e incertezas envolvidas em problemas de crédito, Pascual, Martínéz e Alamillos (2013), afirmam que a forma mais eficiente de aplicar técnicas estatísticas para classificação é através de técnicas de estimação de estruturas, sendo assim interessante a utilização das redes bayesianas para esta problemática.

Considerado como um objeto matemático, um grafo pode ser definido como um conjunto de vetores (comumente chamados de nós) que são associados entre eles através de arcos (hastes) tal que $G = (V, E)$, em que V representa o conjunto de nós e E representa os arcos que ligam o conjunto de V nós do grafo G . Dessa forma G representa um grafo, composto de um número finito de V , em que para a maioria dos arcos presentes em G resultam em ligações do tipo simples, isto é, não existem *loops* ou múltiplos arcos que os ligam a outros nós (HØJSGAARD; EDWARDS; LAURITZEN, 2012).

Normalmente os nós ($v \in V$) são representados por meio de círculos ou, quando existe um número muito grande de nós, pontos e os arcos são expressos por linhas ou setas direcionais (uni ou bidirecionais) como demonstrado na Figura 6. À estrutura formada pelos arcos e nós dá-se o nome de Grafo ou rede.

Como objeto estatístico, um grafo normalmente é utilizado para representar modelos no qual os nós denotam as variáveis ou parâmetros do modelo (HØJSGAARD; EDWARDS; LAURITZEN, 2012). Os modelos estatísticos gráficos são interessantes por fornecerem uma estrutura na qual as relações de dependência e independência entre as variáveis podem ser diretamente observadas pelo grafo.

Os arcos que ligam os V 's objetos podem ser entendidos como funções que operam estes objetos. Cada arco pode ser utilizado para definir um par ordenado ou não ordenado de nós, conectando-os em uma relação conhecida como adjacência. Se os nós são adjacentes se diz também que esses são vizinhos.

Segundo Nagarajan, Scutari e Lèbre (2013), existem três tipos de conexões fundamentais em grafos probabilísticos, sendo suas probabilidades influenciadas por estas. Para os autores as relações entre dois nós é a ligação $a = (u, d)$, pode ser: arestas ($u-d$) – em que o sentido da ligação não é relevante resultando em grafos não orientados; ou arcos ($u \leftarrow d$; $u \rightarrow d$) – em que o sentido é relevante e resultam em grafos orientados.

Os elementos mais importantes dentro de um grafo probabilístico segundo Nagarajan, Scutari e Lèbre (2013), podem ser ilustrados na Figura 6 abaixo:

Figura 6 - Elementos de um grafo probabilístico

Fonte: Nagarajan, Scutari e Lèbre (2013, p. 7).

A partir do nó “A” destacado na Figura 6 sua vizinhança é composta pela união de seus nós pais e filhos. Desta forma, o nó “H” não faz parte da vizinhança de A, pois não é seu descendente nem seu predecessor.

Os nós descendentes imediatos de A são seus filhos. Todo o conjunto de nós que tem ligação direta com os filhos de A são descendentes de A. Em termos de probabilidade, isto equivale dizer que os descendentes de A têm probabilidade condicionada por eventos ocorridos em A. De forma análoga, os nós que precedem os pais de A são também predecessores de A, o que equivale dizer que o nó A apresenta uma interdependência condicional com os predecessores de seus pais.

Um grafo probabilístico pode ser discreto ou contínuo. Uma rede bayesiana é considerada discreta quando os nós presentes em sua estrutura podem assumir um número finito de estados da natureza. Desta forma, considerando uma rede com dois nós [A, B], sendo cada nó composto de um vetor que assume um único valor com intervalo definido e finito, diz-se que esta é uma rede bayesiana discreta. Caso os estados da natureza possam assumir intervalos contínuos, esta será uma rede contínua (SANFORD; MOOSA, 2013).

Existem basicamente dois tipos de Grafos empregáveis a redes bayesianas, quais sejam os grafos diretos (ou orientados) e os grafos indiretos (ou grafos não orientados). Diz-se que um grafo é indireto quando este não apresenta uma relação de precedência entre os nós, sendo denotado pelo exemplo $u-v$. Os grafos definidos como diretos são aqueles que expressam alguma relação de precedência, tais como os dos tipos $u \leftarrow d$ e $u \rightarrow d$.

Segundo Nagarajan, Scutari e Lèbre (2013), a notação que caracteriza um grafo direto é $G = (\mathbf{V}, \mathbf{A})$ e a notação que caracteriza um grafo indireto é $G = (\mathbf{V}, \mathbf{E})$, existindo ainda, segundo os autores, grafos que combinam elementos diretos e indiretos $G = (\mathbf{V}, \mathbf{A}, \mathbf{E})$.

O tamanho de um grafo é proporcional à extensão de suas interdependências condicionais, isto é, depende da quantidade de nós e das ligações existentes entre estes nós.

Figura 7 - Extensão da estrutura de um grafo probabilístico

Como é possível observar os pais são nós (variáveis) que precedem alguma outra variável. Desta forma, uma variável pai é condicionalmente correlacionado com seu filho (nó) em uma relação de interdependência condicional.

3.3.1 Definições e propriedades básicas de uma rede bayesiana

Uma rede bayesiana é uma estrutura que busca descrever a relação causal entre um conjunto de variáveis, ou seja, busca verificar a probabilidade de um dado evento dadas as ocorrências observadas em um conjunto de covariáveis (probabilidade condicional).

Pascual, Martínéz e Alamillos (2013), definem redes bayesianas como um método descritivo que tem como objetivo descobrir as relações de dependência e/ou independência entre um conjunto de covariáveis.

De acordo com Nagarajan, Scutari e Lèbre (2013), uma rede bayesiana é uma classe de modelos gráficos que permite construir uma representação de interdependência probabilística entre um dado conjunto de variáveis aleatórias $X = \{X_1, X_2, \dots, X_p\}$ representadas por um grafo direto e acíclico (DAG – *Direct acyclic graph*) $G = (V, A)$ em que cada nó $V_i \in V$ corresponde a uma variável aleatória X_i .

Considerando $P(y)$ como a probabilidade à priori de ocorrência de y ser igual a 1, $P(y/x_1)$ como a probabilidade à posteriori de ocorrência de y dado x_1 e $P(x_1/y)$ como a probabilidade à posteriori de ocorrência de x_1 dado y , pode-se indicar que:

$$P(y \cap x_1) = P(y)P(x_1 | y) = P(x_1)P(y | x_1) \quad (3.3)$$

Desta forma, é possível indicar que:

$$P(y | x_1)_{\text{posteriori}} = P(y)_{\text{apriori}} \frac{P(x_1 | y)}{P(x_1)} \quad (3.4)$$

Para Pascual, Martínéz e Alamillos (2013), seguindo a regra de Bayes sabe-se que $P(y | x_1)_{\text{posteriori}}$ aumenta na medida em que $P(y)$ e $P(y | x_1)$ diminuam. A hipótese de máximo à posteriori (MAP) denota a hipótese mais provável de ocorrência e pode ser definida como:

$$Y_{\text{MAP}} = \arg \max_y P(y | x_1) = \arg \max_y \left[\frac{P(y)P(x_1 | y)}{P(x_1)} \right] \quad (3.5)$$

Adaptando o exemplo oferecido por Nagarajan, Scutari e Lèbre (2013), a obtenção do máximo à posteriori a partir de um grafo DAG ocorre a partir da extração de fatores dadas as

ocorrências à priori. Considerando a probabilidade P das variáveis Y , X e Z dado que Y precede X e Z (ou seja, Y é pai de Z e X) tal que:

$$P(Y_1=0,692); P(Y_2=0,308); P(X_1|Y_1=0,67); P(X_2|Y_1=0,33);$$

$$P(X_1|Y_2=0,5); P(X_2|Y_2=0,5); P(Z_1|Y_1=0,33); P(Z_2|Y_1=0,67);$$

$$P(Z_1|Y_2=0,5); P(Z_2|Y_2=0,5);$$

O grafo desta CPT¹¹ pode ser representado como segue:

Figura 8 - Exemplo do grafo da CPT

Uma CPT é uma tabela que expressa a probabilidade de uma variável condicionada à probabilidade de suas variáveis-pai (HOJSGAARD; EDWARDS; LAURITZEN, 2012). Cada nó terá, desta forma, a sua tabela de probabilidade condicional – CPT. Esta CPT tem suas probabilidades condicionadas às probabilidades de seus pais.

A extração de um fator tal que exprima a probabilidade de ocorrência de Y , X e Z pode ser representada como:

$$P(Y, X, Z) = P(x | y) P(z | y) P(y) \quad (3.6)$$

¹¹ Esta estrutura é denominada por Nagarajan, Scutari e Lèbre (2013) como CPT – *Conditional Probability Table* – Tabela de probabilidade condicional. Maiores detalhes em 3.6.

Assim, para os dados da Figura 8, a probabilidade de ocorrência de Y_1 , X_1 e Z_1 é:

$$P(Y_1, X_1, Z_1) = P(x_1|y_1)P(z_1|y_1)P(y_1) = 0,67 \times 0,33 \times 0,692 = 0,153$$

Nos últimos anos as redes bayesianas têm ganhado espaço dentro das pesquisas em vários campos, especialmente em probabilidade e modelos baseados em grafos aplicados a problemas de reconhecimento de imagem, aprendizagem de máquinas, problemas de classificação, dentre outros (NAGARAJAN; SCUTARI; LÈBRE, 2013).

Alguns elementos das redes bayesianas são importantes para o conhecimento de sua forma operativa, tais como a problemática da estrutura de grafos probabilísticos, da estimação e da aprendizagem em redes.

3.3.2 Estruturas de redes bayesianas

Como abordado, um grafo tem sua estrutura formada por arcos e nós, podendo ser direto ou indireto. As formas mais comuns de grafos podem ser observadas na Figura 9. Como é possível observar, em estruturas indiretas os arcos não trazem um sentido de precedência entre os nós. Já nos grafos diretos este sentido é oferecido pelo arco (seta).

O sentido de precedência também é conhecido como hierarquia da rede e é um elemento de suma importância para o cálculo das probabilidades condicionais, uma vez que é esta hierarquia quem determina a probabilidade de ocorrência de um evento dada a ocorrência de outros eventos observados, como demonstrado na seção 3.4.

Grafos diretos assumem dois tipos, quais sejam: grafos conectados e não conectados. Grafos diretos do tipo não conectado são aqueles nas quais ao menos um dos nós não está conectado a outro nó por um arco, enquanto nos conectados, todos os nós são conectados.

Os grafos conectados podem ser do tipo cíclico ou acíclico. Nos grafos diretos cíclicos um nó filho é pai do nó que precede seu pai. Como observado na figura, existe uma estrutura em que $B \rightarrow D \rightarrow E \rightarrow B$, isto é, B precede D e é precedido por E, que por sua vez é precedido por D, formando um ciclo.

Já os grafos acíclicos podem assumir duas tipologias, quais sejam: Grafos acíclicos com conexão simples e grafos acíclicos com conexão múltipla. Em grafos acíclicos com conexão múltipla filhos podem ser precedidos de mais de um pai. Os grafos acíclicos com conexão simples podem ser do tipo árvore simples – no qual cada filho é precedido de um único

pai, sendo que o pai pode ser pai de mais de um filho – e do tipo poliárvore, no qual um filho pode ser precedido de mais de um pai.

Figura 9 - Tipos de estruturas de grafos probabilísticos

Fonte: Ara (2011, p. 29).

Contudo, diferentemente do grafo com conexões múltiplas, no grafo poliárvore pelo menos um dos pais não pode ser precedido de um pai pertencente à mesma rede. Como pode ser observado no grafo poliárvore da Figura 9 o nó E é precedido dos nós C e B, contudo o nó C não é precedido por nenhum outro nó da rede.

3.3.3 Estimação e propriedades *markovianas* em redes bayesianas

Um grafo G é considerado um grafo com “mapa independente” – “*Independent map*” ou I-map – em uma estrutura de dependência probabilística, se existir apenas relações do tipo “um para um” entre as variáveis (nós) que compõem o grafo G . Dessa forma, para uma estrutura G de nós B, A e C, em que $\perp_P$ representa uma independência probabilística e $\perp_G$ representa uma ausência de arcos ligando as variáveis (nós):

$$B \perp_P C | A \leftarrow B \perp_G C | A \quad (3.7)$$

De forma análoga, caso existam ligações por arcos dentro de uma estrutura grafo orientado G formada pelos nós A , B e C , esta estrutura será chamada de D-map – “*Dependent map*”.

Como estabelecem Nagarajan, Scutari e Lèbre (2013), as estruturas indicadas na Figura 10 são três tipos de estruturas em V (*V-Structures* CPT), que ferem as condições anteriormente apresentadas, justamente por terem estrutura em V . Para as estruturas deste tipo é necessário considerar a existência de independência condicional abordada pelo critério de d-separações ou separação direcionada.

Diz-se que existe d-separação entre B e C , sendo denotado por: $B \perp\!\!\!\perp_G C|A$ se ao longo da *V-structure* existir uma das duas condições abaixo (NAGARAJAN; SCUTARI; LÈBRE, 2013):

- Se a *V-structure* for convergente (como no item i. da Figura 10), em que nenhum nó precedente de B seja descendente de C (e vice versa);
- Se a *V-structure* está em A e não existe convergência de arcos.

Figura 10 - Tipos de estruturas em “V”
Fonte: Nagarajan, Scutari e Lèbre (2013, pg. 37).

Para cada um dos três tipos, no caso discreto, um dos elementos mais importante é a CPT – *Conditional Probability Table*.

A propriedade *markoviana* das redes bayesianas, considerada a propriedade de d-separação, permite a obtenção da distribuição de probabilidade conjunta das variáveis aleatórias (nós) a partir da fatoração das distribuições de probabilidade condicional. Segundo Nagarajan, Scutari e Lèbre (2013), no caso de variáveis aleatórias discretas, a fatoração para x_i variáveis para obtenção da distribuição de probabilidade conjunta é dada por:

$$p(x_1, x_2, \dots, x_n) = \prod_{j=1}^n p[x_j | PAR_g(x_j)] \quad (3.8)$$

Em que PAR_g representa as variáveis *predecessoras* de x_i (também conhecidas como pais de x_i).

Outro elemento importante diz respeito à fatoração em estruturas de grafos probabilísticos seriais. Se for tomado como referência os elementos presentes na Figura 10, é possível obter, por intermédio de uma fatoração, as probabilidades em ambos os sentidos. A esta equivalência de probabilidades dá-se o nome de equivalência de Markov (NAGARAJAN; SCUTARI; LÈBRE, 2013).

De acordo com Højsgaard, Edwards e Lauritzen (2012), a propriedade de equivalência de Markov pode ser encontrada tanto em DAG's quanto em *Indirect Graphs*, nos seguintes tipos de *V-structure*¹²:

Figura 11 - Estruturas em “V” com equivalência de Markov

Fonte: Højsgaard, Edwards e Lauritzen (2012, p. 254).

Outra propriedade markoviana importante é a conhecida como Markov Blanket - MB.

Por esta propriedade, considerando um nó A e um conjunto de nós pertencentes à vizinhança de A , nomeado λ_A , sendo λ_A composto pelos pais, filhos e dos filhos de A com outros pais, nas condições de estruturas em V apresentadas na Figura 11 anterior, $MB(A)$ é dado por:

¹² Para maiores definições sobre fatoração e propriedade de equivalência de Markov, ver Koller e Friedman (2009, cap. 4).

$$P(A|\lambda A, B) = \Pr(A|\lambda A) \quad (3.9)$$

Como observado na Figura 12 fazem parte da Markov Blanket – MB(A) todos os elementos destacados (pintados). A probabilidade de qualquer nó evidenciado pode ser medida pela probabilidade dos demais nós presentes em MB(A), independentemente do sentido que se tenha para o DAG.

Figura 12 - Propriedade de Markov Blanket
Fonte: Nagarajan, Scutari e Lèbre (2013),

Como definido por Nagarajan, Scutari e Lèbre (2013), esta propriedade implica em duas considerações fundamentais em redes bayesianas:

- i) Os nós não adjacentes, para qualquer estrutura em V (*V-structure*) e presentes em uma MB, são equivalentes a um *Indirect graph*; e
- ii) Independe o sentido da causalidade dos nós definidas pelos arcos dentro de uma MB.

Dessa forma, em uma rede bayesiana, os eventos e probabilidades dos eventos tanto dos pais quanto dos filhos de um nó fornecem informações sobre este nó¹³.

As variáveis (nós) que compõem uma rede podem assumir valores discretos, contínuos, escalares ou mesmo nominais. Contudo em grande parte dos algoritmos, a estimação das redes é direcionada para variáveis discretas (HOJSGAARD; EDWARDS; LAURITZEN, 2012).

¹³ Para maiores definições a propriedade de *Markov Blanket* e processos de moralização de grafos probabilísticos, ver Koller e Friedman (2009, cap. 4)

Tal elemento faz com que variáveis contínuas, escalares e nominais normalmente sejam transformadas em variáveis discretas como uma forma de facilitar o processo de estimação das redes, especialmente através de processos de categorização. De uma forma geral, para o presente trabalho a estimação das redes foi realizada para o caso discreto.

3.3.4 Aprendizagem em redes bayesianas

Aprendizagem em redes bayesianas é um termo normalmente utilizado para definir o trabalho de realizar a modelagem e ajustamento do modelo aos dados (KOLLER; FRIEDMAN, 2009). Este trabalho pode ser realizado em duas etapas, sendo a primeira voltada para a seleção do modelo e a segunda para a seleção das técnicas de estimação dos parâmetros (NAGARAJAN; SCUTARI; LÈBRE, 2013).

Na primeira etapa, também conhecida como aprendizagem da estrutura, o trabalho consiste em identificar a estrutura do grafo que melhor expressa as relações existentes dentro da rede (estrutura de nós e arcos). Este trabalho consiste em identificar a estrutura de grafo que cubra o espaço probabilístico da forma mais eficiente possível.

Segundo Nagarajan, Scutari e Lèbre (2013), este trabalho é conseguido através da minimização do I-map do grafo e de seus nós. Ainda segundo os autores existem muitos tipos de algoritmos de otimização do grafo, sendo estes classificados em três tipos:

- Algoritmos baseados em restrição: que empregam testes de independência condicional, tais como os algoritmos: PC, *Grow-Shirk* (GS), *Incremental Association* (IAMB), *Fast Incremental Association* (Fast-IAMB) e *Interleaved Incremental Association* (Inter-IAMB);
- Algoritmos baseados em *scores*: que empregam técnicas genéricas de heurísticas de otimização, tais como: *hill-climbing*, algoritmos genéticos, dentre outros: e
- Algoritmos híbridos: que combinam elementos dos algoritmos de restrição e de *score*, tais como: *Max-Min Hill-Climbing* (MMHC), dentre outros.

O segundo passo, que consiste na estimação dos parâmetros é também conhecido como aprendizagem dos parâmetros (NAGARAJAN; SCUTARI; LÈBRE, 2013). Este trabalho é realizado pela estimação das distribuições com base na estrutura definida na etapa anterior e gera a distribuição global da rede.

No presente estudo, o algoritmo utilizado para estimação das redes bayesianas testadas foi o *k*dB (*k-dependence Bayesian Network*). No caso do *k*dB-0, a concepção é justamente a não existência de dependência condicional entre as variáveis *preditoras* (ou classes), equivalendo o mesmo ao Naïve Bayes. Estas redes trazem em seu cerne premissas sobre as estruturas das redes, as quais podem ser observadas na seção 3.3.5.

3.3.5 Alguns classificadores bayesianos

No presente estudo, como já indicado, são abordados dois métodos de aprendizagem estatística baseado em redes para classificação, quais sejam: o Naïve Bayes e o *k*dB (*k-dependence Bayesian Network*).

Estes métodos recebem o nome de classificadores por terem como característica a propriedade de definir, com base nas probabilidades associadas a cada estado da natureza, qual a classe, dentre um conjunto de classes, uma determinada entidade é pertencente.

Tabela 1 - Exemplo de classificação via redes probabilísticas.

Bom(C ₀)/Ruim(C ₁)	Idade(A)	Sexo(B)
C ₀	<30	Masculino
C ₀	<30	Feminino
C ₀	≥30	Masculino
C ₀	≥30	Feminino
C ₀	≥30	Masculino
C ₀	≥30	Feminino
C ₁	<30	Feminino
C ₁	≥30	Feminino
C ₁	<30	Masculino
C ₁	<30	Feminino

Nota: Nesta tabela exemplifica-se uma rede composta por três nós, sendo que o nó A representa a idade, podendo assumir valores A₀ (maiores ou iguais a trinta anos) ou A₁ (menores que trinta anos), o nó B representa o sexo que assumem B₀ (masculino) ou B₁ (feminino) e o nó C que assume C₀ (para clientes não *default*) e C₁ (para clientes em *default*).

Esta probabilidade é obtida a partir da aprendizagem dos parâmetros. Como estabelecido por Koller e Friedman (2009), a classificação é o trabalho de determinar o conjunto de probabilidades para as classes de um determinado nó (variável), tal que seja possível associar uma probabilidade deste nó assumir cada classe dados os demais nós que compõem sua rede.

Tomando como exemplo uma estrutura na qual os nós A e B representam idade e sexo de proponentes ao crédito respectivamente, sendo que a Tabela 1 representa as informações de dez observações de proponentes ao crédito presentes em um banco de dados D_G , espera-se

definir a probabilidade de um indivíduo pertencer a cada classe (cliente bom ou ruim), com base em um dado ponto de corte (que sinaliza o risco máximo que a empresa pretende correr).

A representação desta rede pode ser observada na Figura 13.

Figura 13 - Estrutura da CPT para dados da tabela 1

O classificador opera então pela identificação da probabilidade de um indivíduo, dadas as suas características, pertencer ao grupo de clientes bons ou ruins.

No exemplo anterior considera-se A_0 como clientes com idade maior ou igual a 30 anos e clientes A_1 com idade menor que 30 anos, bem como considera-se B_0 como clientes do sexo masculino e B_1 como clientes do sexo feminino.

Caso um novo proponente ao crédito, homem com mais de 30 anos, fosse analisado com base na estrutura estabelecida, o mesmo obteria probabilidade de aproximadamente 0,8462 de pertencer ao grupo de bons clientes e de aproximadamente 0,1538 de pertencer ao grupo de clientes ruins. Isso se explica pelo fato de a probabilidade da CPT ser fatorada em função dos indivíduos pertencentes ao grupo de clientes com esta característica¹⁴.

Caso a empresa entenda que um proponente ao crédito com risco de 0,1538 de ser um mau pagador (entrar em *default*) apresente um risco insuportável, ela declinará a solicitação do proponente. Os algoritmos, após o treinamento e otimização da estrutura interpretam a informação do limite superior de risco como sendo o ponto de corte do modelo e fornecem automaticamente a classificação do proponente.

¹⁴ Observar processo nos exemplos oferecidos na seção 3.4

Como destacado por autores como Pascual, Martínéz e Alamillos (2013) e Brown e Mues (2012), as redes bayesianas têm este nome pela possibilidade de incluir, dentro da estrutura, informações à priori.

Na Tabela 1, por exemplo, caso se queira incluir uma informação à priori das probabilidades relacionadas a uma nova variável como, por exemplo, o nível de formação profissional, a estrutura seria capaz de incorporar esta informação, considerando-a dentro do processo de faturação e obteria novas distribuições, bem como novas probabilidades para proponentes ao crédito.

Na prática, esta possibilidade permite a incorporação de informações de especialistas quanto a eventuais impactos em condições econômicas e indícios de problemas oferecidos pelas análises subjetivas, o que gera grande ganho para os modelos baseados em redes.

Como é possível observar a estrutura da rede é determinante para o resultado das probabilidades. Desta forma, mostra-se interessante observar as premissas sobre as duas redes analisadas neste estudo.

3.3.5.1 Naïve Bayes

Como visto na seção 3.5 existem diversas estruturas de grafos probabilísticos. Autores como Pascual, Martínéz e Alamillos (2013), Brown e Mues (2012), Danenas, Garsva e Gudas (2011) têm estudado as estruturas de redes bayesianas demonstrando a eficiência deste método para o tratamento de problemas de análise de risco de crédito, indicando que esta é uma técnica poderosa para esta área de pesquisa (PASCUAL; MARTINÉZ; ALAMILLOS, 2013).

De acordo com Hastie, Tibshirani e Friedman (2008), a despeito de seu nome, o classificador Naïve Bayes é o tipo de classificador mais utilizado ao longo dos últimos anos, isto dada a sua modelagem mais simples.

Figura 14 - Estrutura Naïve Bayes

Pascual, Martínéz e Alamillos (2013), ressaltam que a principal suposição do classificador Naïve Bayes é a de que as variáveis (nós) são condicionalmente independentes dada a variável predita. Isto implica dizer que no Naïve Bayes um nó filho não poderá ser pai de outro nó filho. Tal comportamento pode ser observado na Figura 14 que retrata a estrutura do Naïve Bayes.

Como é possível observar dada a variável predita y , as variáveis x_i para este caso são independentes. O modelo Naïve Bayes pode ser estimado a partir de:

$$p(x_1, x_2, \dots, x_n) = \prod_{j=1}^n p[x_j | PAR_g(x_j)] \quad (3.10)$$

Em que PAR_g representa as variáveis que precedem x_i (conhecidas como pais de x_i).

Ainda segundo Pascual, Martínéz e Alamillos (2013), a suposição de forte independência entre as variáveis (nós) que compõem a rede é uma suposição que não encontra respaldo em muitas situações, restringindo a aplicação desse tipo de classificador.

Hastie, Tibshirani e Friedman (2008), indicam que esta é uma técnica especialmente recomendada quando a dimensão do espaço probabilístico P é muito grande, o que torna este tipo de estimação atraente.

Segundo os autores, esta propriedade faz com que o classificador, a despeito de estimações de probabilidades grosseiras, tenha boa acurácia para a classificação em alguns casos, especialmente quando os eventos não são equiprováveis e se direciona uma variável de interesse, dado que a região de decisão não se mostra prejudicada.

3.3.5.2 KDB - *k-dependence Bayesian Network*

Como é possível observar, no classificador Naïve Bayes não existe relação de dependência entre os nós filhos. Segundo Pascual, Martínéz e Alamillos (2013), esta é um pressuposto que não encontra sustentação na realidade.

Na rede bayesiana kDB (*k-dependence Bayesian Network*), ao contrário do Naïve Bayes, existe uma relação de precedência na qual, um filho é pai de outro filho. Esta relação pode ser observada através da Figura 15. Na figura 15.(a), não existe qualquer dependência (Naïve Bayes

ou kdB-0). Já na figura 15.(b) um (nó) filho é pai de, pelo menos, outro filho. Nesta estrutura o número 1 (kdB-1) significa que um filho que é pai de um outro filho.

Figura 15 - Estruturas kdB
Fonte: Louzada, Ara (2012a, p.11585).

A estrutura kdB pode assumir então n relações de dependência, o que implica em dizer que há a possibilidade de existência de um kdB- n . Por outro lado, no caso do kdB-0, a ideia é a de inexistência de dependência condicional entre as variáveis preditoras (ou classes), o que implica dizer que o kdB-0 tem suas premissas idênticas às do Naïve Bayes. Como estabelecido por Louzada, Ara (2012a), para o caso estritamente discreto, a probabilidade à posteriori de um evento dado o conjunto de variáveis, pode ser calculado pelo kdB a partir de:

$$P(Y = y_k | x_1, x_2, x_3, \dots, x_p) = \frac{P(Y=y_k) \prod_{i=1}^p P(x_i | \text{pais}(X_i), Y=y_k)}{\sum_j P(Y=y_j) \prod_{i=1}^p P(x_i | \text{pais}(X_i), Y=y_j)} \quad (3.11)$$

Outros classificadores que tem ganhado notoriedade na área de análise de risco de crédito são os classificadores baseados no conceito de MMC – *Maximal Margin Classifier*.

3.4 Support Vector Machines

SVM (Support Vector Machines) é o nome dado a um conjunto de técnicas de classificação e regressão derivadas do conceito de MMC (*Maximal Margin Classifier*). Os dois principais métodos derivados desta abordagem são o SVC (*Support Vector Classifier*) e o classificador de mesmo nome SVM (*Support Vector Machine*).

Em ambos os casos, a classificação está baseada no conceito de hiperplano¹⁵. Considerado um hiperplano linear como um espaço bidimensional, a classificação ou separação de um conjunto de dados se dá por uma reta. Para o problema de análise de risco de crédito, esta abordagem pode ser observada pela Figura 16.

Figura 16- Vetores de suporte
Fonte: Duan e Shrestha (2011, p. 60).

Como é possível observar na Figura 16, neste espaço bidimensional a classificação com base no hiperplano pode ser dada por uma reta entre indivíduos classificados como *default* e *não default*.

Na figura, demonstra-se dois vetores de suporte (*support vector*), sendo estes os elementos que buscam extrapolar o espaço bidimensional para casos de p-dimensionais.

Nesta figura, é possível separar perfeitamente os casos de *default* e *não default* pela observância de um máximo hiperplano, isto é, vetores lineares que separam de forma ótima o

¹⁵ Para definições detalhadas sobre hiperplano em classificadores observar James et al (2013, cap. 9).

hiperplano. Contudo, há muitos problemas de classificação tendo em vista que não é possível separar de forma perfeita um conjunto de dados apenas pela observância do máximo hiperplano linear. Para estas condições, James et al. (2013), apresentam dois classificadores, quais sejam: SVC e SVM.

3.4.1 Support Vector Classifier

Considerando o conceito de hiperplano, na problemática de classificação – como é o caso dos problemas de previsão de *default* – existem muitos casos nos quais não se é possível separar perfeitamente um conjunto de dados em duas (ou mais) classes pela observância de um máximo hiperplano linear. Esta situação pode ser ilustrada pela Figura 17.

Figura 17 - Limitações do máximo hiperplano linear
Fonte: James et al. (2013, p. 349).

O SVC, também conhecido como SMC (*Soft Margin Classifier*), é um método que baseia a classificação no entendimento de que o máximo hiperplano não necessariamente é o melhor vetor de separação ou classificação das entidades existentes no espaço em questão (JAMES et al., 2013).

O método abre mão do máximo hiperplano e determina um hiperplano que seja tão menor quanto necessário para conseguir o máximo em termos de acerto nas classificações, o que é entendido como o ótimo hiperplano.

Assim como nos demais métodos de classificação baseados em um hiperplano o SVC classifica a observação em função do lado do hiperplano ao qual ela se encontra, contudo, classificações erradas são inevitáveis.

A classificação é obtida à partir de um conjunto de dados de treinamento no qual o algoritmo busca fornecer a melhor separação possível dentro do conjunto de dados. James et al. (2013), apontam que, para classificação em duas classes, busca-se a solução para o seguinte problema de otimização:

$$\beta_0, \beta_1, \dots, \beta_p, \epsilon_0, \epsilon_1, \dots, \epsilon_n \quad \text{Maximizar} \quad M \quad (3.12)$$

Sujeito a:

$$\sum_{j=1}^p \beta_j^2 = 1; \quad (3.13)$$

$$y_i(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i), \quad (3.14)$$

$$\epsilon_i \geq 0, \sum_{i=1}^n \epsilon_i \leq C, \quad (3.15)$$

Em que M representa a margem, C é um parâmetro de ajuste não negativo, M é a largura da margem, devendo esta ser tão maior quanto possível, $\epsilon_1 \dots \epsilon_n$ são chamadas de variáveis de folga e têm a função de permitir que uma observação esteja no lado errado do hiperplano (ou da margem).

A classificação é obtida com base no sinal da função $f(x^*) = \beta_0 + \beta_1 x_1^* + \dots + \beta_p x_p^*$. Assim como as demais técnicas apresentadas nesta pesquisa as estimativas do SVC podem ser facilmente obtidas a partir de pacotes implementados no *software* R (JAMES et al., 2013).

3.4.2 Support Vector Machine

O *Support Vector Machine* – SVM, traduzido para o português como vetor de suporte de máquina, pode ser definido como uma extensão do SVC, sendo que sua principal diferença para o SVC é que o SVM permite a incorporação de estimações não lineares em suas classificações (JAMES et al., 2013).

Considerando os dados plotados na Figura 17, observa-se este comportamento de forma bastante clara, tendo em vista que não há um hiperplano que separe linearmente as classes de

uma forma eficiente. Como abordado por James et al. (2013), este é um comportamento bastante comum nos dados reais.

Na Figura 18 os dados da Figura 17 anterior são repetidos, agora para o SVM, considerando um comportamento não linear.

Figura 18 - Comportamento não linear do SVM

Fonte: James et al. (2013, p. 350).

O *Support Vector Machine* – SVM, incorpora ao SVC variáveis do tipo *kernel* (granulares), que são funções capazes de quantificar a similaridade entre duas observações. Considerando o SVC linear para duas observações $(x_i, x_{i'})$, como sendo dado por: $\langle x_i, x_{i'} \rangle = \sum_{j=1}^p x_{ij} x_{i'j}$. O SVC pode ser representado por:

$$f(x) = \beta_0 \sum_{i \in S}^n \alpha_i \langle x_{ij}, x_{i'j} \rangle \quad (3.16)$$

Em que S representa uma coleção de pontos de suporte.

Incorporando a variável *kernel* à função para duas variáveis $(x_i, x_{i'})$, esta poderia ser representada como:

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^p x_{ij} x_{i'j} \right)^d \quad (3.17)$$

Em que d é o grau de um polinômio Kernel, generalizando o SVC.

Esta é uma generalização do SVC para um polinômio *kernel* de grau d . Como exemplificado por James et al. (2013), quando o SVC é combinado com um *kernel* não linear, o resultado é o SVM, um classificador mais flexível e eficaz para diversos problemas. Sua representação pode ser dada por:

$$f(x) = \beta_0 \sum_{i \in S}^n \alpha_i K(x, x_i) \quad (3.18)$$

Segundo Yu et al. (2010), o SVM apresenta diversos benefícios quando empregado em problemas de risco de crédito, como é o caso para a presente pesquisa. Para os autores, os principais benefícios são três a saber:

- i) Requer menos concepções sobre os dados, tais como linearidade e continuidade;
- ii) Possibilita a realização de um mapeamento não linear da estrutura dos dados; e
- iii) É capaz de implementar uma estrutura de minimização de risco (SEM – *Structure Risk Minimization*) a partir de algoritmos que buscam aprender o hiperplano que maximiza as margens.

Dadas suas características de estimação de parâmetros que não são probabilísticos, mas sim medidas de distâncias (lineares ou não lineares), o SVM e o SVC, para cálculo das probabilidades, dependem de uma aproximação das distribuições de suas estimativas para alguma distribuição de probabilidades. Esta aproximação é interessante para cálculo de indicadores que levem em consideração a probabilidade de *default* das carteiras.

4 PROCEDIMENTOS METODOLÓGICOS

Para facilitar o entendimento dos procedimentos metodológicos realizados para a execução desta pesquisa é possível segregá-los em três partes:

- i) Bloco teórico: no qual é apresentado o produto do levantamento teórico-conceitual – composto do capítulo introdutório e dos dois capítulos subsequentes;
- ii) Grupo de métricas empregadas como proposta de análise de desempenho das técnicas de aprendizagem estatística apresentadas no capítulo 3, incluindo procedimentos de aplicação e base de dados (exposto neste capítulo); e
- iii) Aplicação da proposta conceitual, no qual é apresentada a análise das técnicas por meio das métricas estatísticas e financeiras, presentes no capítulo de resultados e discussões.

Antes da apresentação destes três blocos, se mostra interessante, contudo, uma breve exposição do ambiente no qual se descortina o trabalho, isto é, a caracterização do setor no Brasil.

4.1 Caracterização do setor

O mercado de crédito brasileiro remonta a época do Brasil Império, período no qual observaram-se as primeiras casas bancárias e sociedades de poupança e empréstimo.

Segundo Arcoverde (2002), este mercado passou por uma situação de inércia ao longo de sua história, dada a ocorrência de crises econômicas cíclicas e períodos de grande inflação, fazendo com que os bancos tivessem vantagens no financiamento da dívida interna do governo brasileiro - tomador tido como de menor risco de crédito, com substancial liquidez e histórico pagador de elevadas taxas de juros.

A estabilidade econômica recente e a atuação governamental fizeram com que este mercado tivesse as condições necessárias para a expansão observada recentemente, o que reforça o interesse dos bancos na adoção de modelos mais robustos de risco de crédito, bem como de gestão de portfólios de crédito.

Annibal e Koyama (2011), indicam que a conjuntura tem também incentivado ações governamentais no sentido de disponibilizar instrumentos econômicos mais modernos para

fomento do mercado, instituindo novos sistemas de securitização e aprimorando mecanismos executivos por meio de leis voltadas para a promoção do mercado de crédito.

Para os autores, data das últimas duas décadas a insipiente formação de bancos de dados sobre crédito no Brasil, não apenas pela conjuntura econômica nacional, mas também pelo desenvolvimento tecnológico que permitiu a formação e análise de dados mais robustos, contendo séries temporais de *defaults*, recuperações, atrasos e comportamentos de crédito.

Arcoverde (2002), entende que os elementos econômicos, o desenvolvimento de técnicas cada vez mais sofisticadas e o empenho governamental para igualar as práticas nacionais de gestão de crédito às práticas internacionais têm feito com que diversas instituições financeiras do país tenham interesse em adotar métodos e modelos mais sofisticados na gestão de suas carteiras de crédito. Entende-se assim, que a utilização destas técnicas deve ser incentivada por pesquisas para que estas sejam de domínio nacional.

Da mesma forma, entende-se que a análise financeira das técnicas e modelos de risco de crédito é um campo que deve ser explorado, não apenas nacional como internacionalmente, para garantir que os resultados sejam realmente alcançados, não se limitando, desta forma, às práticas de mensuração de métricas estatísticas atualmente empregadas.

4.2 Tipo de pesquisa e quadro teórico

Considerando as conceituações apresentadas por Cooper e Schindler (2003), acerca dos delineamentos de métodos e procedimentos de pesquisa e o seu objetivo, esta é uma pesquisa descritiva, tendo em vista a necessidade de análise e descrição de relações entre dados e variáveis para compreensão da problemática. Estudos exploratórios também são requisitados como forma de maximizar o entendimento sobre os dados empregados na mesma.

Tema	Objetivos
Introdução	Componente que destaca o embasamento sobre o crédito em seus aspectos conceituais, tipos e características, embasando objetivos, premissas, fundamentos e justificativas da pesquisa.
<i>Default</i> de crédito: Gestão e Risco	Componente que destaca os elementos conceituais sobre o risco de crédito, formas de classificação e tratamento do risco de crédito e recentes pesquisas nacionais e internacionais na área.
Aprendizagem Estatística	No qual são apresentadas e conceituadas as principais modelagens de aprendizagem estatística, com foco em técnicas clássicas e modernas voltadas para a problemática de classificação, tais como: análise discriminante, regressão logística, redes bayesianas, SVC e SVM.

Quadro 4 - Detalhamento da pesquisa bibliográfica

Como visto no quadro 4, no bloco teórico da pesquisa são explorados três tópicos que configuraram os três capítulos do embasamento.

Este procedimento traz como produto a depuração e embasamento epistemológico do problema, de premissas e de objetivos específicos de pesquisa, bem como garante que os elementos que embasam o raciocínio do problema estejam adequadamente solidificados e discutidos na vanguarda da área, facilitando posteriores avaliações e replicações da pesquisa.

Esquemáticamente o embasamento teórico está integrado aos elementos de análise da pesquisa. Tal integração pode ser observada na Figura 19:

Figura 19 - Integração da pesquisa bibliográfica à análise

Na Figura 19 é possível observar a abordagem de integração buscada na pesquisa para os capítulos de fundamentação teórica, privilegiando a exposição das relações existentes entre as premissas e considerações apresentadas no capítulo de introdução e os embasamentos e discussões fornecidos pelos dois capítulos teóricos que a seguem.

No capítulo de introdução o tema foi apresentado em seu conceito, conjuntura e relevância, com destaque para a problemática da necessidade de utilização das técnicas e a necessidade critérios financeiros para seleção de técnicas e modelos.

Partindo da premissa de que o desempenho financeiro da carteira deve ser o critério que norteia a seleção de técnicas e modelos de classificação e análise de risco de crédito os objetivos foram desenhados para cobrir os dois principais elementos, quais sejam: a) o uso de novos métodos de aprendizagem estatística em classificação de risco de crédito e b) o emprego de critérios financeiros para seleção de modelos e técnicas.

Para dar sustentação à argumentação foram levantados dois capítulos a título de embasamento teórico. O primeiro (capítulo 2) referiu-se ao conceito de *default* de crédito, suas divergências conceituais, os elementos que compõem o risco de crédito e os métodos comumente utilizados para sua análise e gestão.

Ainda dentro deste capítulo foram abordadas pesquisas nacionais e internacionais sobre o tema, dando foco especial nos métodos e técnicas de classificação com o objetivo de levantar o uso destas técnicas e principalmente a forma de análise e validação das mesmas.

Comprovou-se, por estes capítulos, uma hipótese difusa dentro da problemática apresentada no capítulo introdutório, qual seja a de que os métodos e modelos empregados nos estudos, tanto no cenário nacional quanto internacional, são predominantemente comparados por métricas estatísticas, especialmente de acurácia.

Observou-se também diferenças substanciais entre as práticas nacionais e internacionais, sendo que no cenário internacional novas técnicas têm sido utilizadas com grande frequência enquanto no cenário nacional este uso é tímido, ficando as pesquisas concentradas na seleção de variáveis e composição de modelos.

O segundo capítulo teórico (capítulo 3), observando as técnicas mais utilizadas em âmbito internacional e nacional, selecionou e apresentou sete destas técnicas. Este capítulo teve enfoque em duas classes de novos modelos, quais sejam: redes bayesianas e modelos baseados no conceito de hiperplano (*support vectors*) e dois modelos clássicos (análise discriminante e regressão logística).

Como abordado no terceiro capítulo, o avanço de tecnologias de processamento e análise de dados e a grande utilização de técnicas baseadas em aprendizagem estatística na

problemática de classificação e mensuração de risco de crédito são elementos que reforçam a importância do tema. Também, o recente avanço da atividade de crédito no Brasil e sua preponderância no mercado mundial, são elementos que fazem deste um tema de pesquisa ainda mais interessante.

As considerações sobre o cenário de crédito, a preponderância do tema e os avanços tecnológicos que influenciam as formas de gestão e a consideração de que o objetivo do investimento em crédito é a geração de valor pela carteira de crédito, fundamentam a proposta conceitual, uma vez que esta proposta buscou analisar um dos principais instrumentos de gestão de crédito (os modelos e técnicas) através da ótica financeira, ótica essa até então não evidenciada.

4.3 Uso de ferramentas computacionais e bases de dados

Após análise do quadro teórico de pesquisas nacionais e internacionais sobre modelagens de análise e previsão de risco de *default* foram selecionadas sete técnicas para compor o quadro de referência sobre o tema, quais sejam: a análise linear discriminante, a regressão logística, as redes bayesianas Naïve Bayes, kdB-1 e kdB-2 e os classificadores SVC – *Support Vector Classifier* e SVM – *Support Vector Machine*.

A seleção das regressões logísticas e das regressões discriminantes para compor o quadro referencial se deu pela sua importância histórica para o tratamento do problema de análise e previsão de risco de *default* de crédito no cenário brasileiro e mundial.

Já as redes bayesianas e os classificadores SVC e SVM aparecem como técnicas promissoras para análise de classificação no cenário mundial e que, dentro da área das Ciências Administrativas no Brasil, ainda não são difundidas.

Desta forma, a presente pesquisa avança em dois sentidos, sendo o primeiro a incorporação de novas métricas para análise da eficiência dos modelos de previsão de *default* e o segundo o auxílio para difundir junto às Ciências Administrativas, os novos classificadores desenvolvidos dentro da aprendizagem estatística.

Para as análises presentes nos resultados do estudo foi utilizada a base de dados acadêmica *German Credit Dataset*, sendo os testes realizados utilizando o software estatístico R, através dos pacotes MASS e e1071, bem como do algoritmo kdB desenvolvido por Louzada e Ara (2012a).

Para treinamento e teste de todas as técnicas utilizadas na pesquisa foram utilizadas amostras desbalanceadas, obedecendo a integridade dos dados da base de dados *German Credit*

Dataset. O processo de treinamento e teste segue o modelo *K-Fold Cross Validation* com $k=10$, isto é, as amostras de treinamento e teste são de 200 observações.

Para as redes bayesianas a pesquisa emprega *priors* não informativas, as quais são obtidas a partir de amostragens das probabilidades proporcionais das variáveis, isto é, os bancos de dados manterão sua integridade da mesma forma que fora mantida a integridade dos dados no teste realizado sobre o *German Credit Dataset*.

Como expresso por Shevchenko (2011), tal elemento inibe a formação de *priors* informativas, isto é, *priors* que condicionam intencionalmente distribuições e comportamentos das variáveis e dos dados

O uso das ferramentas computacionais está segregado em três estágios. No primeiro momento para obtenção de dados. Para manter a integridade dos dados, considerando as redes bayesianas utilizadas, foi necessária a discretização dos dados.

No segundo instante são computados os diversos métodos de previsão e gerados os resultados das métricas. Esta análise está implementada no software estatístico R e as tabelas de comparação foram feitas por meio do software Microsoft Excel. Para as redes bayesianas Naïve Bayes, kDB e para o SVC e SVM a pesquisa traz em seu apêndice os procedimentos básicos para cálculo em R.

Por fim, o último estágio é a apresentação e análise dos resultados com base nas métricas selecionadas via matriz de confusão.

4.4 Processo de comparação e seleção de modelos de *credit scoring*

Como já indicado, o objetivo da presente tese está centrado na incorporação de novas técnicas de aprendizagem estatística para tratamento do problema de classificação em análise de risco de crédito, bem como na avaliação de modelos de classificação tendo como critério o desempenho financeiro da carteira.

Como forma de facilitar a observação e uso dos conceitos e raciocínios apresentados até aqui pela pesquisa, bem como servir de um primeiro direcionador para a ação de gestores no processo de seleção de modelos para a problemática de classificação, esta pesquisa propõe adicionalmente um conjunto de passos para a execução da avaliação de modelos com base no desempenho financeiro da carteira.

Como referenciado anteriormente, o modelo de classificação adotado é determinante para a formação da carteira. Entende-se que com base no desempenho financeiro da carteira é possível inferir o desempenho do modelo em termos financeiros.

Os passos perseguidos pela pesquisa para a seleção de modelos podem ser observados na Figura 20 a seguir.

Figura 20 - Passos para análise de modelos via métricas financeiras

No primeiro passo do processo mostra-se importante a escolha de uma dentre as carteiras que compõem o montante de investimentos em crédito da empresa.

Neste estudo não foram realizados esforços para segmentações na base de dados para não haver qualquer distorção quanto ao objetivo fundamental de análise dos resultados ou mesmo à integridade dos dados.

Contudo, diferenças substanciais quanto ao conjunto de dados e mesmo a seleção e comportamento de variáveis podem influenciar o desempenho do modelo. Como referenciado por Brito, Assaf Neto e Corrar (2009), a gestão de cada carteira de clientes demanda um modelo ajustado aos dados desta carteira.

Segundo Matias (2007), as políticas de crédito da organização a tornam mais ou menos restritiva em termos de concessão de crédito. Contudo, a segmentação de clientes que compreende o tipo de tomador de recurso, continua a ocorrer.

Elementos como a distribuição dos dados, comportamento das variáveis e critérios de seleção das variáveis que compõem o modelo, podem ser influenciados pela segmentação de clientes que, na prática, formam carteiras (segmentadas) de clientes. Segundo Matias (2007), existem diversas formas de realizar a segmentação, tais como:

- a) Crédito à pessoa física varejo;
- b) Crédito a pessoas físicas de alta renda;
- c) Microcrédito;
- d) Crédito a micro, pequena e média empresa;
- e) Crédito a empresa de médio e grande porte;
- f) Crédito corporativo;

- g) Crédito rural;
- h) Crédito habitacional, etc.

São as políticas da organização quem determinam os perfis de clientes que serão trabalhados, isto é, quais carteiras serão formadas, levando-se em consideração elementos como os retornos e os riscos de mercado, de liquidez, operacional, sistemático e de crédito. Especificamente para o risco de crédito a organização demandará um modelo de classificação, sendo este o recorte da presente pesquisa.

Para reconhecimento do comportamento e da distribuição das variáveis que compõem o conjunto de dados carteira se fez necessária a exploração dos dados.

Após a leitura e o entendimento dos dados que compõem a carteira, se iniciou o segundo passo, no qual foi necessário selecionar um conjunto de técnicas estatísticas para a modelagem. A modelagem aqui é compreendida como a seleção de variáveis e definição da relação funcional das variáveis, ou, no caso de modelos de redes probabilísticas a definição da estrutura de interdependência condicional.

Critérios de ajustamento e de informação, tais como SQR – Soma dos quadrados residuais, AIC – *Akaike Information Criterion*, dentre outros podem ser utilizados para a seleção de variáveis que comporão os modelos.

Para a presente pesquisa optou-se por não realizar trabalhos de inclusão ou exclusão ou mesmo mudança na forma funcional das variáveis, tendo em vista que seu recorte está na apresentação de novas técnicas e análise do desempenho financeiro das carteiras formadas por estas técnicas.

Sendo assim, optou-se por não realizar alterações no número e comportamento das variáveis para que as alterações não se configurassem vieses de pesquisa, demonstrando-se apenas os resultados como ilustração e não aplicação otimizada de cada método.

A fase de análise (terceiro passo) consiste na implementação dos modelos levantados no passo anterior. Esta implementação, para a presente pesquisa, foi realizada com base em grupos de treinamento (no qual o modelo é ajustado aos dados) e teste (no qual o modelo é testado para verificar sua capacidade preditiva).

Com os dados referentes à capacidade preditiva de cada modelo a pesquisa implementou as métricas de avaliação, sendo elas a métrica estatística ACC, e as métricas financeiras: RFC - Resultado financeiro com a carteira de crédito, PLR - perda líquida com a carteira de recebíveis, RAROC – *Risk Adjusted Return on Capital* e IS - Índice de Sharpe.

A apresentação dos resultados desta métrica se deu por meio de matrizes de confusão, na qual são observados comparativamente os resultados de cada modelo para cada uma das métricas testadas. Estes resultados formam o foco das discussões da pesquisa.

Por fim, na quarta fase, entende-se que, com base nos resultados provenientes destas métricas, é possível analisar quais modelos apresentam maior capacidade preditiva, bem como qual modelo é capaz de proporcionar maior retorno para a carteira formada a partir das classificações verdadeiras e falsas, retorno este que pode ser medido em termos gerais ou mesmo em termos relativos (considerando a volatilidade dos retornos).

Este mesmo conjunto de passos pode ser utilizado para avaliar modelos de uma mesma técnica, isto é, para comparar, com base nos critérios financeiros, diversos modelos formados a partir de uma mesma técnica estatística (regressão logística, por exemplo) tendo em vista que, mesmo tendo maior número de classificações corretas, um modelo pode não proporcionar o maior resultado financeiro.

Diversas variáveis devem ser consideradas no último passo para a decisão, tais como:

- Taxa de juros embutida nas vendas: que, aliada ao percentual de recuperação, podem proporcionar vantagem na venda a prazo;
- Custo médio ponderado de capital de giro: que, quanto menor, maior poderá ser o spread da venda, o que proporciona maior alavancagem;
- Perda com a insolvência: que provoca efetiva perda financeira na carteira (pós tentativas de recuperação);
- Ganho com a inadimplência: derivado do *spread* entre custo ponderado de capital e ganhos com taxas de juros por pagamentos atrasados e recuperações;
- Tamanho da carteira de crédito: que é derivado das políticas de crédito e investimentos da empresa;
- Receita com a carteira de crédito: que são dependentes das taxas de juros embutidas nas vendas e das taxas de multa e mora por inadimplência;
- Despesas com a carteira de crédito: derivadas dos gastos para gestão e manutenção da carteira e que corroem a sua rentabilidade;
- Perda esperada com a carteira de crédito: efetivo custo da carteira;
- Perda não esperada: risco ao qual a carteira está exposta e para o qual será necessário alocar um capital de segurança, sendo que, quanto maior o capital

alocado para esta finalidade, menor é a rentabilidade (menor é o capital para o mercado);

- Retorno da carteira de mercado: base de comparação e custo de oportunidade;
- Volume da inadimplência e Desvio padrão da carteira: elementos direcionadores do risco e que fornecem indícios da eficiência da carteira em termos do *trade-off* risco e retorno;

Estas são as principais variáveis que compõem as métricas financeiras trabalhadas nesta pesquisa. Como é possível observar, as perspectivas para diversas variáveis podem influenciar a decisão em termos do modelo de classificação a ser utilizado.

O essencial buscado pela pesquisa é utilizar estas variáveis de decisão para constituição de métricas de comparação das modelagens, permitindo a incorporação de tais métricas no cenário de decisão do modelo, e não apenas nas decisões das políticas.

4.4.1 Métricas estatísticas e proposta conceitual de métricas financeiras

Como métrica estatística foi selecionada apenas a *accuracy* de cada técnica, sendo os dados apresentados sob o formato de uma matriz de confusão, seguindo procedimento de Iscoe et al. (2012). Segundo os autores esta é uma medida comumente utilizada para análise da eficiência de um modelo de crédito. A *accuracy* pode ser medida da seguinte forma:

$$ACC = \frac{Tp + Tn}{Tp + Fp + Tn + Fn} \quad (4.1)$$

Em que Tp é o número de classificações a posteriori verdadeiras positivas, Tn é o número de classificações a posteriori verdadeiras negativas, Fp é o número de classificações a posteriori falsas positivas e Fn é o número de classificações a posteriori falsas negativas.

Como proposta conceitual a presente pesquisa indica a incorporação de quatro métricas financeiras para a análise da viabilidade de métodos e modelos de classificação de risco de crédito.

Serão utilizados dois indicadores de desempenho de carteira de crédito propostos por de Matias (2007), quais sejam: Resultado financeiro com a carteira de crédito e perda líquida com

a carteira de recebíveis, bem como outros dois métodos de análise de investimento, a RAROC – *Risk Adjusted Return on Capital* e o Índice de Sharpe.

O levantamento dos dados das variáveis utilizadas em cada uma das métricas financeiras foi baseado em dados ou em aproximações, quando não observados as informações necessárias dentro do conjunto de dados estudado. O detalhamento destas aproximações é oferecido no capítulo 5.

Entende-se que, considerando o investimento em recebíveis como o saldo dos valores a receber de uma carteira, é possível extrair informações importantes, dentre as quais pode-se destacar o retorno do investimento em recebíveis e a perda líquida com a carteira de recebíveis oferecidas por Matias (2007). Estas métricas são utilizadas aqui como mecanismo de comparação entre os diversos modelos de previsão de *default* empregados neste estudo.

O resultado financeiro com a carteira de recebíveis é uma medida que analisa o ganho com a aplicação de taxa de juros nas vendas a prazo deduzido do custo de captação destes recursos e pode ser expresso como:

$$RFC = (JVp - CMPCG)CR - PC + Gc \quad (4.2)$$

Em que: RFC representa o Resultado financeiro com a carteira de crédito; JVp representa a taxa de juros incluída nas vendas a prazo, CMPCG representa o Custo médio ponderado de capital de giro, Pc representa o montante referente ao percentual de perdas com a insolvência, Gc representa os ganhos com a inadimplência e CR é a carteira de crédito.

Já a perda líquida com a carteira de recebíveis é uma métrica que analisa a perda com a insolvência deduzida dos ganhos derivados das multas por atraso e outras tarifas oriundas da inadimplência. Tal perda pode ser expressa como:

$$PLR = (Pi - Gi)CR \quad (4.3)$$

Em que: PLR é a perda líquida com a carteira de recebíveis (esperada), P_i representa o percentual de perdas de crédito (com a insolvência), G_i representa o percentual de ganhos com a inadimplência e CR representa a carteira de crédito.

O estudo emprega também o índice de Sharpe (FERSON, 2013), que tem como objetivo a mensuração do retorno da carteira em função do risco da mesma, buscando evidenciar a remuneração do agente frente ao risco ao qual está exposto. Este índice é dado por:

$$IS = \frac{E(Ri - Rf)}{\sigma_{Ri}} \quad (4.4)$$

Em que: Ri representa o retorno obtido pela carteira de crédito, Rf representa o retorno obtido a partir de uma carteira de referência de mercado tida como livre de risco¹⁶ e σ_{Ri} representa o desvio padrão ou volatilidade do retorno da carteira.

A última métrica aplicada é a RAROC – *Risk Adjusted Return on Capital* ou retorno ajustado ao risco de capital, que é uma medida de rentabilidade baseada no risco (FERSON, 2013). Por tal métrica, quanto maior for o retorno em relação ao risco maior será o desempenho do negócio.

Como abordado por Ferson (2013), considerando elementos como receita e despesa de um portfólio, sua fórmula pode ser expressa como:

$$RAROC = \frac{r - e - el + ic}{c} \quad (4.5)$$

Em que: r representa as receitas da carteira, e representa as despesas, el representa a perda esperada, ic representa o retorno sobre o capital e c representa o montante de capital investido na carteira de crédito.

Contudo, como abordado por Saita (2007), ocorrem variações sobre o cálculo do RAROC segundo as premissas empregadas.

Prokopzuke et al. (2007), indicam que, considerando um processo estocástico, a métrica pode ser definida como: $RAROC = E[r]/EC$, em que $E[r]$ representa a esperança do retorno e EC o capital econômico alocado. Caso seja assumido que o retorno gerado seja a riqueza medida pelo EVA® e que o capital econômico alocado é o VaR (*Value at Risk*), o RAROC também pode ser medido como sendo: $RAROC = EVA^®/VaR$ (GALVÃO et al., 2007).

¹⁶ Aqui baseada na taxa de retorno dos títulos do governo dos Estados Unidos somada ao risco país Alemão (tendo em vista que a base de dados é daquele país).

Para Saita (2007), o RAROC enquanto métrica de avaliação de desempenho financeiro de um investimento, considerando a exposição do capital ao risco, pode ser expressa como: $RAROC = r/CaR$, em que CaR representa o capital em risco (*Capital at Risk*).

Pela proximidade com os dados trazidos pela base de dados estudada e pela facilidade de aplicação e comparação nos modelos de classificação, a pesquisa utilizou em sua estrutura de análise do desempenho financeiro das carteiras RAROC como descrita por Saita (2007), na qual:

$$RAROC = r/CaR \quad (4.6)$$

Em que r representa o retorno da carteira e CaR representa o capital exposto ao risco. Para a presente pesquisa, CaR pode ser obtido por:

$$CaR = VaR_i * CR_i = (-\bar{\phi}_i + z\sigma_i) * CR_i \quad (4.7)$$

Em que: z é o escore padrão da tabela normal para $\alpha=0,05$ unicaldal (aproximadamente 1,65), σ_i é o desvio padrão dos retornos médios das carteiras formadas por cada técnica.

Já $\bar{\phi}_i$ representa a média da variação lognormal dos retornos das carteiras dada por:

$$\bar{\phi}_i = \frac{1}{n} \sum \ln \left(\frac{r_{in}}{r_{in-1}} \right) \quad (4.8)$$

Em que r_{it} representa o retorno da carteira i para cada uma das n simulações realizadas (com $n=100$) e r_{in-1} representa a simulação imediatamente anterior à simulação n .

Com base nestes indicadores, espera-se oferecer evidências adequadas sobre qual método empregar em uma carteira de crédito para obtenção do máximo retorno.

Cabe ressaltar que, entende-se que as demais formas de mensuração da métrica RAROC aqui apresentadas, bem como outras métricas, podem ser empregadas sem prejuízos para a proposta, que é de avaliação de modelos.

De uma forma geral, a pesquisa entende que ainda que se utilize outras formas de mensuração do RAROC, ou mesmo outras métricas para avaliação da eficiência financeira o

propósito de emprego desta métrica na comparação dos modelos de classificação através do resultado financeiro da carteira continua sendo adequado e necessário.

4.4.2 Procedimentos para levantamento das informações para cálculo das métricas

Todos os dados de classificação foram baseados nas médias dos testes realizados junto ao banco de dados. Tais médias foram apuradas considerando 100 testes (repetições) para cada uma das métricas estudadas. Desta forma, para o caso do ACC por exemplo, os dados das classificações foram obtidos pela média de 100 repetições dos testes de classificação realizadas com amostras aleatórias com reposição.

Como informado no quadro 5, as informações sobre taxa de juros incluída nas vendas foram extraídas da própria base de dados, através da média das taxas praticadas na variável *rates* existente na mesma.

O CMPCG – Custo médio ponderado de capital de giro pode ser obtido pela somatória ponderada do custo de capital dos itens do passivo circulante financeiro e operacional mais o custo do Capital de Giro Líquido (MATIAS, 2007).

Uma vez que não há informações disponíveis sobre o custo médio ponderado de capital de giro do setor ou mesmo informações para seu cálculo na base de dados, a pesquisa empregou como *proxy* o percentual das despesas de captação para pessoa física do mercado brasileiro presente no relatório sobre *spread* bancário da Fipecafi (2004), (no qual, o valor informado é 9,55% ao ano para o saldo do capital emprestado para pessoa física no período estudado pelo referido relatório) sobre à receita média do setor estudado (bancário), o qual foi de 35,94% ao ano. Desta forma, o CMPCG, considerando as taxas praticadas pela base de dados, foi de 0,80%¹⁷ aproximadamente

Para calcular o percentual de perda com a insolvência foi considerado o percentual de recursos emprestados para o conjunto de clientes com erro falso negativo de classificação de cada modelo, deduzida a taxa média de recuperação de crédito do mercado brasileiro pós-*default*. Essa taxa foi baseada no estudo de Oliveira Filho, Almeida e Matsumoto (2014), e publicado pelo Banco Central do Brasil, apenas como uma métrica de comparação, tendo em vista a indisponibilidade destes dados para os clientes selecionados, bem como a indisponibilidade da taxa de recuperação de crédito da base estudada.

¹⁷ Considerando que o custo de captação foi aproximadamente 26,57% dos juros cobrados pelo setor ($9,55\% \div 35,94\%$) e considerando a taxa média da base de dados (de 3,01% anuais), o seu CMPCG seria igual a 0,7998% ao ano.

Métrica	Variável	Levantamento
ACC	Tp	Número total de classificações a posteriori corretas para clientes em não <i>default</i>
	Tn	Número total de classificações a posteriori corretas para clientes em <i>default</i>
	Fp	Número total de classificações a posteriori erradas para clientes em não <i>default</i>
	Fn	Número total de classificações a posteriori erradas para clientes em <i>default</i>
RFC	JVp	Taxa média de juros da variável <i>rates</i> do conjunto de dados;
	$CMPCG$	Aproximação pelo percentual das despesas de captação para pessoa física do mercado brasileiro frente aos juros médios PF
	CR	Montante emprestado a cliente classificados como não- <i>default</i> nos testes realizados
	Pc	Montante solicitado por clientes com erro de classificação de falso negativo deduzida da taxa de recuperação do mercado brasileiro
	Gc	Valor recuperado de clientes com erro de classificação de falso negativo (considerando taxa de recuperação brasileira) multiplicado pela taxa média de juros praticada na variável <i>rates</i> do conjunto de dados e pelo valor emprestado.
PLR	Pi	Pc em termos percentuais (em relação à carteira)
	Gi	Gc em termos percentuais (em relação à carteira)
	CR	Montante emprestado a cliente classificados como não- <i>default</i> nos testes realizados
IS	Ri	Juros praticados na variável <i>rates</i> multiplicado por clientes verdadeiros positivo + recuperação em Gi
	Rf	Juros de títulos americanos + Risco-país alemão
	σ_{Ri}	Desvio padrão médio do retorno da carteira
RAROC	r	Valor de Ri
	CaR	VaR das carteiras simuladas das n=100 classificações realizadas em cada técnica multiplicado pelo montante emprestado a cliente classificados como não- <i>default</i> no conjunto de dados de teste

Quadro 5 - Levantamento dos dados das variáveis

Esta taxa, em 2013 teve média de 27,32%. Desta forma, a pesquisa classificou todos os *defaults* observados em classificações de falso negativo como insolventes com recuperação de 27,32% como forma de observar as comparações.

Para o ganho com inadimplência, não se implementou um valor maior do que a taxa de juros embutida nas vendas e, desta forma, considerou-se como ganho com a inadimplência a taxa de 3,01% sobre a o montante médio de recuperação de crédito.

Para cálculo do retorno obtido na carteira de crédito para a métrica IS a pesquisa, seguindo os padrões até aqui adotados, utilizou-se o montante obtido a partir dos juros

recuperados de empréstimos a clientes classificados como verdadeiro positivo nos testes de cada técnica, somado ao valor de recuperação do ganho com a inadimplência.

A título de comparação o R_f foi baseado nas taxas de retorno dos títulos do governo dos Estados Unidos somada ao risco país alemão¹⁸. Segundo dados publicados por Damodaran em seu *site* no ano de 2014, a taxa de retorno dos títulos do governo dos Estados Unidos em janeiro de 2014 fora de 0,46% e o risco-país alemão para o mesmo período foi de 0,5%. Desta forma, o R_f empregado para a pesquisa foi de 0,96%.

Este procedimento pode ser melhorado caso se tenha dados reais sobre variáveis importantes como: prazo de recuperação, custos e montantes de recuperação de crédito para a base de dados estudada, custo efetivo de captação e de retornos reais de mercado. Em um conjunto de dados real, informações sobre prazos e custo de recuperação, bem como das reais taxas aplicadas sobre o capital ofertado a crédito são de suma importância para que as métricas sejam confiáveis.

Adicionalmente, com base em informações da própria base de dados que indicam que a perda por falso positivo é 1/5 da perda por falso negativo¹⁹ a pesquisa traz ainda uma comparação pela tabela de custos informada pelo banco de dados.

¹⁸ Optou-se pelo risco país alemão por se tratar de uma base de dados alemã. As informações sobre o risco país daquele país foram extraídas do site do autor Aswath Damodaran: <http://pages.stern.nyu.edu/~adamodar/>

¹⁹ Tais dados estão na descrição do banco de dados, disponível em: <http://archive.ics.uci.edu/ml/datasets/>

5 RESULTADOS E DISCUSSÕES

Para demonstração do procedimento de análise do desempenho financeiro cada uma das técnicas de classificação e previsão de *default* trabalhadas na pesquisa foram realizados os procedimentos metodológicos anteriormente definidos na base de dados *German Credit Dataset* (base alemã com dados de pessoa física com 21 covariáveis e 1.000 observações).

Esta base de dados é amplamente empregada em estudos da área de crédito e serve como um *benchmark* para análise dos resultados obtidos pelas modelagens (LOUZADA-NETO; ARA, 2012a). Tal base de dados oferece a seguinte configuração de variáveis:

Variável	Descrição	Tipo de Variável	Nº de Categorias
	Status da conta do solicitante em relação ao		
Checking	salário declarado	Ordinal	4
Duration	Duração do empréstimo requerido	Contínua	-
History	Histórico de créditos anteriores	Ordinal	4
Purpose	Finalidade do crédito requerido	Nominal	11
Amount	Montante do crédito requerido	Contínua	-
Savings	Economias do cliente	Ordinal	5
Employ	Tempo no emprego atual	Ordinal	5
Rates	Juros praticados (%) no empréstimo requerido	Contínua	-
Status	Sexo e estado civil	Nominal	5
debtors	Outras dívidas	Nominal	3
residence	Tempo na residência atual	Contínua	-
property	Propriedades	Nominal	4
Age	Idade	Contínua	-
other_inst	Outros planos de parcelamento	Nominal	3
housing	Tipo de moradia	Nominal	3
exist_cr	Nº de créditos concedidos no banco	Contínua	-
Job	Profissão	Ordinal	4
provider	Número de dependentes	Contínua	-
phone	Telefone próprio	Binária	2
foreign	Estrangeiro	Binária	2
goodbad	Qualificação do cliente como bom ou mal pagador	Binária	2

Quadro 6 - Variáveis presentes na base de dados

Esta base de dados traz um conjunto de dados de crédito de mil clientes pessoas físicas, sendo que destes, um total de 30% dos clientes são classificados como em *default*. Para o presente estudo foram selecionados dados desbalanceados para treinamento dos métodos. Abaixo seguem os dados sumarizados.

Tabela 2 - Estatística descritiva dos dados

	duration	amount	rates	residence	age	exist_cr	provider
Mínimo	4.0	250	1.000	1.000	19.00	1.000	1.000
1º Quartil	12.0	1366	2.000	2.000	27.00	1.000	1.000
Mediana	18.0	2320	3.000	3.000	33.00	1.000	1.000
Média	20.9	3271	2.973	2.845	35.55	1.407	1.155
3º Quartil	24.0	3972	4.000	4.000	42.00	2.000	1.000
Máximo	72.0	18424	4.000	4.000	75.00	4.000	2.000

Nota: As variáveis presentes nesta tabela referem-se apenas às variáveis classificadas como contínuas pela base de dados. As classificações das demais variáveis podem ser observadas no quadro 6²⁰.

Buscando facilitar o entendimento, especialmente do comportamento das variáveis de clientes que entraram ou não em *default*, a pesquisa apresenta tal comportamento para as variáveis contínuas na Figura 21.

Figura 21- Distribuição das variáveis na classificação a priori

²⁰ Na Figura 24 podem ser observados os gráficos de colunas para as variáveis categóricas.

Nesta figura as curvas em vermelho representam as variáveis contínuas dos clientes que não entraram em *default* à priori, enquanto as curvas em verde representam as variáveis contínuas dos clientes em *default*.

Observa-se que créditos em curto prazo apresentam menor probabilidade de *default* do que em longo prazo para este banco de dados, da mesma forma nota-se concentração de *default* na faixa de idade próxima a 20 e 30. Esta classificação expressa que clientes com idades próximas de 20 e menores que 40 anos têm maior risco.

As demais variáveis apresentam certa similaridade em termos de comportamento em relação à variável *default*.

Na Figura 22 são apresentados, sob o formato de uma matriz, gráficos *scatter plot* das variáveis: duração, montante, taxas e idades.

Figura 22 - Matriz de distribuição das variáveis a priori

Os pontos marcados em preto denotam clientes em *default* enquanto os triângulos em vermelho denotam clientes que não entraram em *default*.

As linhas características em vermelho e a reta em preto trazem os cortes destes dois grupos, isto é, são linhas separatrizes dos grupos segundo a concentração, para estas variáveis, de clientes que entraram ou não em *default*.

O comportamento mais característico das variáveis “duração” e “montante” sugere que estas variáveis teoricamente têm maior poder explicativo. Mostra-se correlação positiva em relação à variável “duração” na medida em que a taxa aumenta, o que é perfeitamente compreensível tendo em vista que as prestações se tornam mais caras para o consumidor que, por sua vez, tende a aumentar o prazo para que sua renda tenha menor comprometimento.

Chama, contudo, atenção o fato de não haver correlação positiva entre a taxa praticada e a idade, tendo em vista que, como observado anteriormente pela Figura 22, existe maior risco de *default* para clientes de menor idade do que para clientes de maior idade. Sendo assim, a remuneração do banco para cobrir os riscos desta classe de clientes deveria, teoricamente, ser maior, o que não ocorre no presente conjunto de dados.

Ainda sobre a idade, nota-se que esta apresenta comportamento aleatório quanto ao montante requisitado e a duração dos empréstimos, o que significa que, para o conjunto de dados, a idade não tem qualquer influência sobre o montante e/ou o prazo dos créditos.

As maiores variações nas taxas estão em função da duração e do montante o que também faz sentido, uma vez que maior duração (variável *duration*) tem maior risco de *default* e maior montante apresenta riscos maiores de *default*. Isso pode ser verificado também na Figura 21, na qual nota-se que, na medida em que a variável *duration* cresce, a incidência de *default* cresce também, demonstrando correlação positiva entre a ocorrência e o prazo.

Figura 23 - Duração e montante sobre *default*

Pela Figura 23 é possível observar certa inclinação positiva entre as variáveis duração e montante sobre o *default*, tendo em vista que, para o presente banco de dados, na medida em que se aumenta a duração e/ou o montante aumenta-se a incidência de clientes em *default*, o que segue as indicações da Figura 22.

Esta figura é particularmente interessante por destacar a área na qual este comportamento se dá, indicando a inclinação destas variáveis. Com base apenas nestas variáveis é possível inferir que existe uma probabilidade (presente na área do quadrante no qual os dados das variáveis *preditoras* se enquadram) de o cliente entrar ou não em *default*.

Figura 24 - Matriz de barras de variáveis preditoras

* *Purpose*: A classe A47 – férias – não é observada na base de dados;

** *Personal status and sex*: A classe A95 – Mulher Solteira – não é observada na base de dados.

As demais variáveis foram organizadas em uma matriz de gráficos de barras presente na Figura 24. Nesta figura observa-se o conjunto de classes existentes para cada uma das

variáveis, sendo que a variável “finalidade” é a variável com maior número de categorias (dez no total), em que a maioria das demais tem entre três e cinco classes com concentração em algumas dessas classes.

Como indicado no capítulo anterior, os procedimentos empregados para levantamento dos dados são baseados em simulações e aproximações, tendo em vista que o principal objetivo aqui é apresentar e analisar as técnicas com base no desempenho da carteira.

5.1 Análise quantitativa: aplicação do método

Considerando os dados apresentados pela base de dados teórica *German Credit Dataset*, como abordado anteriormente, esta tese não realiza tratamento nos dados no que tange o seu balanceamento, deixando as amostras fiéis ao banco de dados.

Este procedimento é verificado em pesquisas como a de Chung, Ho e Hsu (2011). Segundo os autores o processo de balanceamento de bancos de dados desbalanceados pode prejudicar a acurácia total de modelos bayesianos. Em seus testes Louzada-Neto, Ferreira-Silva e Diniz (2012), também oferecem indícios deste comportamento. Chung, Ho e Hsu (2011), sugerem adicionalmente que os bancos de dados desbalanceados são mais fiéis à realidade e, desta forma, tais comportamentos devem ser respeitados.

Para teste de todos os modelos foi selecionado conjuntos de 20% (amostras aleatórias). Empregou-se, como ponte de corte, a curva ROC (*Receiver Operating Characteristic*).

Figura 25 – Curva ROC e *Cut-off* modelo Naïve Bayes

Como destacado por Duan e Shrestha (2011), este método apresenta diversas aplicações em *credit rating* e *credit score* para determinação do ponto de corte (*cut-off*) que delimita a classe à qual cada indivíduo pertencerá *a posteriori*.

A curva ROC tem valores variando de 0 a 1 e é obtida através da consideração das proporções de classificação e dos erros e acertos de cada classificação para todos os valores de corte. Com isso pode-se selecionar aquele que promove a maximização das medidas de sensibilidade e especificidade destas classificações (DUAN; SHRESTHA, 2011). Para o presente estudo, este valor foi calculado com auxílio do *software* estatístico R.

Tabela 3 - Resultados apurados nos modelos

Indicadores	Discriminante	Logística	Naïve Bayes	kdB-1
% Médio de requisições aprovadas	51.70%	60.62%	47.03%	61.38%
Montante médio emprestado TP	€\$ 419,393.56	€\$1,088,283.67	€\$ 413,640.44	€\$ 987,575.46
Montante médio emprestado FN	€\$ 91,714.79	€\$ 101,912.95	€\$ 36,610.01	€\$ 161,682.88
Montante médio emprestado (Total)	€\$ 511,108.35	€\$1,190,196.62	€\$ 450,250.45	€\$1,149,258.34
Taxa média (rates)	3.01%	3.01%	3.01%	3.01%
Retorno médio dos empréstimos	€\$ 12,604.69	€\$ 32,707.90	R\$ 12,431.79	€\$ 29,681.15
% <i>Default</i> sobre montante	17.94%	8.56%	8.13%	14.07%
Montante médio recuperado	€\$ 25,221.57	€\$ 28,026.06	€\$ 10,067.75	€\$ 44,462.79
Ganho médio com a inadimplência	€\$ 758.02	€\$ 842.31	€\$ 302.58	€\$ 1,336.31
% de perda sobre montante	13.010%	6.208%	5.895%	10.200%
Desvio padrão do retorno	€\$ 99.31	€\$ 80.49	€\$ 110.97	€\$ 102.99
Desvio padrão da perda	€\$ 720.65	€\$ 674.48	€\$ 1,319.25	€\$ 536.69
Indicadores	kdB-2	SVC	SVM	
% Médio de requisições aprovadas	59.33%	69.82%	72.18%	
Montante médio emprestado TP	€\$ 686,841.90	€\$1,291,530.41	€\$1,444,995.70	
Montante médio emprestado FN	€\$ 213,905.97	€\$ 88,699.85	€\$ 33,504.34	
Montante médio emprestado (Total)	€\$ 900,747.87	€\$1,380,230.26	€\$1,478,500.04	
Taxa média (rates)	3.01%	3.01%	3.01%	
Retorno médio dos empréstimos	€\$ 20,642.74	€\$ 38,816.39	€\$ 43,428.72	
% <i>Default</i> sobre montante	23.75%	6.43%	2.27%	
Montante médio recuperado	€\$ 58,824.14	€\$ 24,392.46	€\$ 9,213.69	
Ganho médio com a inadimplência	€\$ 1,767.93	€\$ 733.10	€\$ 276.91	
% de perda sobre montante	17.217%	4.659%	1.643%	
Desvio padrão do retorno	€\$ 116.23	€\$ 78.44	€\$ 54.35	
Desvio padrão da perda	€\$ 396.12	€\$ 727.80	€\$ 1,161.66	

Nota: Nesta tabela são expostos os indicadores médios necessários no cálculo das métricas financeiras RFC, PRL, IS e RAROC tratadas no capítulo 4. As médias foram apuradas pelos dados de 100 testes (simulações) de classificação para cada técnica de aprendizagem estatística estudada. Adicionalmente outras medidas específicas para cada métrica são apresentadas pertinentemente quando da discussão destas.

Para maximizar a compreensão dos processos realizados, a tabela 3 apresenta os resultados da apuração das informações (indicadores) necessárias para o cálculo de cada métrica

financeira para cada técnica testada. Como é possível observar o método de classificação altera significativamente elementos importantes de uma carteira de crédito, tais como o montante emprestado e, principalmente, o montante das perdas derivadas do *default*. Pela análise das variáveis, considerando elementos norteadores da gestão do crédito como o retorno sobre o montante emprestado e saldado, é possível notar que a técnica SVM apresenta resultados superiores. Contudo, esta técnica é a que forma a maior carteira (considerando-se o montante emprestado por cada técnica).

5.2 Análise das técnicas empregadas com base na métrica ACC

A classificação é realizada através de uma comparação entre a classificação *a posteriori* (obtida a partir de cada um dos modelos separadamente) e a classificação *a priori*, já presente no banco de dados.

Tabela 4 - Resultados dos testes de *accuracy*

Análise Discriminante				Regressão Logística			
Observado	Predito		%	Observado	Predito		%
	ND	D			ND	D	
ND	87.5	52.5	62.5	ND	104.02	35.98	74.3
D	15.9	44.1	73.5	D	17.22	42.78	71.3
ACC		65.8		ACC		73.4	
Naïve Bayes				kdB-1			
Observado	Predito		%	Observado	Predito		%
	ND	D			ND	D	
ND	87.26	52.74	62.3	ND	94.75	33.75	73.74
D	6.79	53.21	88.7	D	28.00	43.50	60.84
ACC		70.24		ACC		69.13	
kdB-2				SVC			
Observado	Predito		%	Observado	Predito		%
	ND	D			ND	D	
ND	82.75	54.10	60.47	ND	124.27	32.64	79.2
D	35.90	27.25	43.15	D	15.36	27.73	64.4
ACC		55		ACC		76	
SVM							
Observado	Predito		%				
	ND	D					
ND	138.45	20.91	86.88				
D	5.91	34.73	85.46				
ACC		86.59					

Nota: Nesta tabela ND Predito vs ND observado representa classificações corretas para clientes não *default* (TP = verdadeiro positivo), D Predito vs ND observado representa erros de classificação para clientes não *default* (falso positivo ou clientes bons classificados como ruins); ND Predito vs D erros de classificação para clientes em *default* (falso negativo ou clientes ruins classificados como bons) e D Predito vs D observado representa acertos de classificação para clientes em *default* (verdadeiro negativo ou clientes ruins classificados como ruins).

Como dito anteriormente este procedimento foi observado em um total de 100 (cem) repetições para cada técnica, gerando os dados médios presentes na tabela 3. Na matriz são apresentados os resultados da métrica ACC – acurácia total, bem como a acurácia (em termos percentuais) de cada um dos tipos de classificação.

Como é possível verificar pelas medidas obtidas e segundo o critério ACC o melhor modelo em acerto seria o SVM com ACC de 86,59%, seguido pelo modelo SVC com 76% e pelo modelo de regressão logística que obteve 73,4%. Os modelos de rede bayesiana Naïve Bayes com 70,2%, kdB-1 com 69,13%, análise discriminante com 64,7% e por último o modelo kdB-2 com 55% formam o restante da análise.

Nota-se, contudo, indícios de desprendimento entre a necessidade gerencial de maximização do ganho e a métrica estatística, tendo em vista que, analisando a tabela 3 nota-se que o Naïve Bayes produziu uma carteira menor que a regressão logística, mas proporcionalmente melhor em termos de montante emprestado para clientes em *default*.

Estes resultados sugerem a necessidade de avaliação das carteiras formadas por cada técnica, com atenção a elementos como retorno e volatilidade do retorno em cada uma.

5.3 Análise das técnicas empregadas com base na métrica de custos

Como contextualizado inicialmente, o objetivo do presente trabalho é verificar o desempenho financeiro de cada um dos modelos para o banco de dados estudado.

Neste sentido, primeiramente, foram levantadas as médias de custos de classificação sugeridas pelo próprio banco de dados no qual os custos dos erros positivos representavam 1/5 dos custos dos erros negativos. Os resultados podem ser observados na tabela 5 na qual é possível verificar uma alteração significativa em relação à interpretação anterior oferecida na tabela 4.

Nesta tabela os acertos não apresentam custos (custo zero), enquanto os erros de classificação falso positivo foram pontuados com custo igual a 1 (um) e os erros de classificação falso negativo foram pontuados com custo 5 (cinco).

A exemplo do ocorrido na tabela 4 com relação à acurácia total, na tabela 5 são trazidos os custos totais de cada técnica, contudo, os valores percentuais na tabela dizem respeito à proporção dos custos de cada erro de classificação (falso positivo e falso negativo) frente aos custos totais.

As classificações de falso negativo dos modelos de análise discriminante e regressão logística, que em termos de acurácia representavam menos de 30% do total dos resultados (15,9

de 68,4 de erro total no caso da análise discriminante e 17,22 de 53,2 de erro total no caso da regressão logística ponderados pela matriz de custos do próprio conjunto de dados representam aproximadamente 60.23% e 70.53% do custo respectivamente.

Tabela 5 - Resultados dos testes de custo

Análise Discriminante				Regressão Logística			
<i>Observado</i>	<i>Predito</i>		% Custo	<i>Observado</i>	<i>Predito</i>		% Custo
	<i>ND</i>	<i>D</i>			<i>ND</i>	<i>D</i>	
<i>ND</i>	0	52.5	39.8%	<i>ND</i>	0	35.98	29.5%
<i>D</i>	79.5	0	60.2%	<i>D</i>	86.1	0	70.5%
<i>Custo</i>		132		<i>Custo</i>		122.08	
Naïve Bayes				kdB-1			
<i>Observado</i>	<i>Predito</i>		% Custo	<i>Observado</i>	<i>Predito</i>		% Custo
	<i>ND</i>	<i>D</i>			<i>ND</i>	<i>D</i>	
<i>ND</i>	0	52.74	60.8%	<i>ND</i>	0	33.75	19.4%
<i>D</i>	33.95	0	39.2%	<i>D</i>	140	0	80.6%
<i>Custo</i>		86.69		<i>Custo</i>		173.75	
kdB-2				SVC			
<i>Observado</i>	<i>Predito</i>		% Custo	<i>Observado</i>	<i>Predito</i>		% Custo
	<i>ND</i>	<i>D</i>			<i>ND</i>	<i>D</i>	
<i>ND</i>	0	54.1	23.2%	<i>ND</i>	0	32.64	29.8%
<i>D</i>	179.5	0	76.8%	<i>D</i>	76.8	0	70.2%
<i>Custo</i>		233.6		<i>Custo</i>		109.44	
SVM							
<i>Observado</i>	<i>Predito</i>		% Custo				
	<i>ND</i>	<i>D</i>					
<i>ND</i>	0	20.91	41.4%				
<i>D</i>	29.55	0	58.6%				
<i>Custo</i>		50.46					

Nota: Nesta tabela ND Predito vs ND observado representa classificações corretas para clientes não *default* (TP = verdadeiro positivo), D Predito vs ND observado representa erros de classificação para clientes não *default* (falso positivo ou clientes bons classificados como ruins); ND Predito vs D erros de classificação para clientes em *default* (falso negativo ou clientes ruins classificados como bons) e D Predito vs D observado representa acertos de classificação para clientes em *default* (verdadeiro negativo ou clientes ruins classificados como ruins).

Os dados da Tabela 5 não representam a realidade dos resultados observados pelo Naïve Bayes. O erro de classificação de falso negativo desta técnica representa pouco menos de 40% dos custos com erros de classificação.

Este resultado é indicativo da importância da ponderação de erros em problemas de classificação como em modelos de *credit scoring*. Em grande parte, tal resultado superior se deve à parametrização do modelo Naïve Bayes que permite a minimizar o custo total através da minimização do erro do tipo falso negativo. Este é um dos principais ganhos das redes bayesianas frente aos demais métodos.

Considerando o custo informado pela base de dados, o modelo Naïve Bayes, que pela métrica ACC ficara na quarta colocação em termos de acurácia de classificação, agora obteve o segundo menor custo dentre todas as técnicas estudadas. Por esta métrica, o SVM continuaria com a melhor estrutura de classificação, seguido dos modelos Naïve Bayes, SVC, Regressão logística, Análise discriminante, kdB-1 e kdB-2.

Dessa forma, há indícios de que a acurácia estatística não pode ser entendida como a principal métrica para a tomada de decisão de um modelo de classificação voltados para a previsão e análise de risco de crédito.

5.4 Análise das técnicas empregadas com base nas métricas RFC e PLR

A tabela 6 apresenta o valor das variáveis financeiras obtidas a partir dos procedimentos descritos nos procedimentos metodológicos e que serviram de base para o cálculo das métricas de Matias (2007), quais sejam a PLR - Perda líquida com a carteira de recebíveis e o RFC - Resultado financeiro com a carteira de crédito.

Tabela 6 - Dados para cálculo das métricas RFC e PLR²¹

Variável	Discriminante	Logística	Naïve Bayes	kdB-1
JVp	3.01%	3.01%	3.01%	3.01%
CR	€\$ 511,108.35	€\$ 1,190,196.62	€\$ 450,250.45	€\$ 1,149,258.34
Pc	€\$ 66,493.22	€\$ 73,886.89	€\$ 26,542.26	€\$ 117,220.09
Pi	13.01%	6.21%	5.89%	10.20%
Gc	€\$ 758.02	€\$ 842.31	€\$ 302.58	€\$ 1,336.31
Gi	0.15%	0.07%	0.07%	0.12%
CMPCG	0.80%	0.80%	0.80%	0.80%
Variável	kdB-2	SVC	SVM	
JVp	3.01%	3.01%	3.01%	
CR	€\$ 686,841.90	€\$ 1,291,530.41	€\$ 1,444,995.70	
Pc	€\$ 155,081.83	€\$ 64,307.39	€\$ 24,290.65	
Pi	17.22%	4.66%	1.64%	
Gc	€\$ 1,767.93	€\$ 733.10	€\$ 276.91	
Gi	0.26%	0.06%	0.02%	
CMPCG	0.80%	0.80%	0.80%	

Nota: Sendo JVp os juros sobre vendas, CR o montante da carteira de crédito, Pc a Perda de crédito, Pi é PC sobre CR, GC o Ganho com a inadimplência, Gi é GC sobre CR e CMPCG o Custo médio Ponderado de Capital de Giro.

Cabe ressaltar que estas métricas foram desenvolvidas por Matias (2007), para medir o desempenho financeiro de uma carteira já formada, servindo de parâmetro para analisar a

²¹ Os dados de Gi e Pi são proporcionais aos valores da carteira.

performance gerencial. Esta pesquisa inova ao empregar estas métricas para analisar a eficiência dos modelos de classificação em termos financeiros.

Tem-se como pressuposto para este emprego o raciocínio de que ao analisar por estas métricas a carteira teórica proveniente de um modelo de classificação, é possível indicar que este modelo será tão eficiente quanto for o desempenho financeiro desta carteira.

Considerando as informações sobre a perda líquida com a carteira de recebíveis e o resultado financeiro com a carteira de crédito foi elaborada a tabela 7 que oferece comparação sobre estas métricas para os modelos discriminante, logístico, Naïve Bayes, kdB-1, kdB-2, SVC e SVM respectivamente.

Tabela 7 Resultados das métricas financeiras RFC e PLR

Variável	Discriminante	Logística	Naïve Bayes	kdB-1
RFC	-€\$ 54,461.95	-€\$ 46,793.04	-€\$ 16,308.74	-€\$ 90,535.20
PLR	-€\$ 65,735.20	-€\$ 73,044.58	-€\$ 26,239.68	-€\$ 115,883.78
Variável	kdB-2	SVC	SVM	
RFC	-€\$ 138,164.59	-€\$ 35,087.69	€\$ 7,857.77	
PLR	-€\$ 116,485.69	-€\$ 59,441.60	-€\$ 23,463.28	

Nota: RFC representa o Resultado Financeiro com a Carteira de Crédito e PLR representa a perda líquida com a carteira de recebíveis.

Cabe ressaltar que, uma vez considerada a taxa de juros da base de dados alemã, utilizando a mesma taxa para o ganho com a inadimplência e aplicando esta taxa sobre o percentual de recuperação de crédito do mercado brasileiro (dada a ausência destes dados para a base) as métricas produziram resultados negativos.

Espera-se que as taxas médias embutidas nas transações, o ganho com a carteira de crédito e o ganho com a inadimplência para o cenário brasileiro sejam maiores que os aqui declarados. Desta forma, as informações da tabela 7 sofrem o viés de a base de dados ser acadêmica de um país e os dados de recuperação serem reais e de um outro país.

Analisando a Tabela 7 observa-se que, em relação aos resultados dos modelos analisados de análise discriminante, regressão logística, kdB 1, kdB-2 e mesmo o classificador SCV, o modelo de rede bayesiana Naïve Bayes proporcionou uma melhora no Resultado financeiro com a carteira de crédito - RFC e Perda líquida com a carteira de recebíveis - PLR. Esse classificador obteve desempenho inferior ao SVM, classificador tido como o melhor em todas as métricas anteriormente analisadas e o único capaz de produzir um resultado positivo para o RFC.

Na tabela 8 são apresentados os resultados comparativos para cada técnica para a métrica RFC. Nesta tabela os resultados negativos representam a variação negativa pelo uso de uma técnica em detrimento a outra.

Tabela 8 - Comparativo das técnicas pela Receita Financeira com a Carteira de Crédito (RFC)

RFC	Discriminante	Logística	Naïve Bayes	kDB-1	kDB-2	SVC	SVM
Discriminante	100%						
Logística	11%	100%					
Naïve Bayes	-60%	-64%	100%				
KDB-1	76%	59%	342%	100%			
KDB-2	77%	59%	344%	1%	100%		
SVC	-10%	-19%	127%	-49%	-49%	100%	
SVM	-64%	-68%	-11%	-80%	-80%	-61%	100%

Nota: Nesta tabela os resultados referem-se à variação % do resultado de RFC caso o analista opte pelo uso das técnicas apresentadas nas colunas em detrimento àquelas apresentadas nas linhas da tabela

Desta forma, caso o analista opte pela análise discriminante em detrimento à regressão logística, o resultado da carteira em termos de RFC será 11% maior. Pelo mesmo raciocínio, a opção pela Análise discriminante em prejuízo ao Naïve Bayes produz um resultado 60% menor.

Como é possível observar pela tabela, o classificador SVM apresenta resultados expressivamente superiores aos demais classificadores, uma vez que a opção por qualquer outro classificador levaria a resultados menores de RFC. Em termos de desempenho desta métrica o SVM é seguido pelo classificador Naïve Bayes, contrariando a métrica ACC. Pela ordem os demais seriam: SVC, análise discriminante, regressão logística, kDB-1 e kDB-2.

Tabela 9 - Comparativo das técnicas pela Perda Líquida com a Carteira de Recebíveis (PLR)

PLR	Discriminante	Logística	Naïve Bayes	kDB-1	kDB-2	SVC	SVM
Discriminante	100%						
Logística	-10%	100%					
Naïve Bayes	151%	178%	100%				
KDB-1	-43%	-37%	-77%	100%			
KDB-2	-44%	-37%	-77%	-1%	100%		
SVC	11%	23%	-56%	95%	96%	100%	
SVM	180%	211%	12%	394%	396%	153%	100%

Nota: Nesta tabela os resultados referem-se à variação % do resultado de RFC caso o analista opte pelo uso das técnicas apresentadas nas colunas em detrimento àquelas apresentadas nas linhas.

O mesmo raciocínio empregado na tabela 8 de comparações da métrica RFC também pode ser empregado para interpretação da tabela 9 de comparações da métrica PLR. Desta forma, qualquer opção em detrimento ao SVM produziria resultados negativos já que, como observado no comparativo com a análise discriminante, caso o analista opte por esta técnica em detrimento ao SVM sua perda líquida cresceria 180%. A mesma sequência de técnicas (classificação das técnicas) observado pela métrica RFC é observada também pela métrica PLR.

5.5 Análise das técnicas empregadas com base nas métricas IS e RAROC

O Índice de Sharpe – IS, como abordado por Ferson (2013), é uma das métricas mais simples e efetivas para avaliação do *trade-off* risco vs retorno. Esta métrica avalia o desempenho de um investimento (carteira de ativos) já descontando o custo de oportunidade de uma taxa livre de risco. Como expresso nos procedimentos metodológicos, a taxa livre de risco para o presente estudo foi calculada em 0,96%.

O retorno sobre o investimento foi calculado a partir do cômputo dos juros praticados na variável *rates* multiplicado pela soma do montante solicitado por clientes verdadeiro positivo e clientes classificados como falso negativos passivos de recuperação. O valor de recuperação representou 27,5% do montante solicitado por clientes classificados com o erro de falso negativo. Considerando tais premissas o IS para cada uma das técnicas está exposto na tabela 10 abaixo:

Tabela 10 - Resultado da métrica Índice de Sharpe (IS)

Variável	Discriminante	Logística	Naïve Bayes	kdB-1
ri	2.61%	2.82%	2.83%	2.70%
rf	0.96%	0.96%	0.96%	0.96%
σ_{Ri}	8.41%	4.74%	8.63%	5.27%
IS	19.67%	39.22%	21.65%	33.00%
Variável	kdB-2	SVC	SVM	
ri	2.49%	2.87%	2.96%	
rf	0.96%	0.96%	0.96%	
σ_{Ri}	6.41%	3.97%	3.62%	
IS	23.84%	48.00%	55.14%	

Nota: Sendo ri o retorno sobre a carteira (medido pela soma do ganho com clientes bons e o ganho da recuperação de clientes ruins), rf é a taxa livre de risco (juros sobre títulos da dívida americana em jan/14 somado ao risco país alemão) e σ_{Ri} a volatilidade média do retorno da carteira (desvio padrão).

Como abordado por Ferson (2013), quanto maior o Índice de Sharp melhor é a relação risco vs retorno. Desta forma, analisando os dados da tabela a técnica SVM novamente aparece

como a melhor dentre as técnicas aqui abordadas. Contudo, contrariando as indicações das métricas RFC e PRL, pela métrica IS o SVC e a regressão logística apresentaram desempenhos superiores ao Naïve Bayes aparecendo como segunda e terceira opções respectivamente, enquanto o classificador Naïve Bayes aparece como quarta melhor opção.

A Figura 26 traz os retornos totais de cada técnica para as n simulações.

Figura 26 - Retornos totais das carteiras

Como é possível observar os maiores retornos foram proporcionados pela técnica SVM, seguida das técnicas SVC, regressão logística e kdB-1. O RAROC é utilizado aqui para ajustar estes retornos aos riscos provenientes das carteiras formadas por cada técnica, risco este proveniente da volatilidade dos retornos destas carteiras. Como indicado nos procedimentos metodológicos para o cálculo do RAROC, o modelo empregado foi: $RAROC = r/CaR$.

Como abordado inicialmente, para cálculo do RAROC a receita da carteira foi calculada como os juros médios praticados na variável *rates* da base de dados, multiplicado pelo montante emprestado para clientes classificados a posteriori como não *default* e clientes recuperados. O CaR – *Capital at Risk* – foi obtido a partir da multiplicação do *VaR* pela carteira de crédito formada por cada modelo.

Na Figura 27 são apresentadas as volatilidades percentuais de cada técnica. Pela figura nota-se que as maiores volatilidades nos retornos estão na análise discriminante e no classificador Naïve Bayes, bem como as menores estão nos classificadores SVM e SVC.

Figura 27 - Volatilidade do retorno das carteiras

Na tabela 11 são apresentados os dados do RAROC para cada uma das técnicas. Desta forma, considerando que os maiores retornos estão no SVM e SVC (como visto nas figuras 26 e 27) e que seus retornos apresentam as menores volatilidades, espera-se que o desempenho na métrica RAROC do SVM e do SVC sejam os melhores dentre as técnicas estudadas.

Como pode ser visualizado pela tabela 11 o SVM e o SVC apresentam os maiores índices RAROC. Desta forma, são, por esta métrica, os melhores classificadores para esta base de dados.

Tabela 11 - Resultados da métrica RAROC

Variável	Discriminante	Logística	Naïve Bayes	kdB-1
<i>r</i>	2.61%	2.82%	2.83%	2.70%
<i>VaR</i>	-17.98%	-11.03%	-19.16%	-12.29%
<i>CR</i>	€\$ 511,108.35	€\$ 1,190,196.62	€\$ 450,250.45	€\$ 1,149,258.34
<i>CaR</i>	€\$ 91,908.19	€\$ 131,338.09	€\$ 86,272.60	€\$ 141,250.11
<i>RAROC</i>	14.5%	25.5%	14.8%	22.0%
Variável	kdB-2	SVC	SVM	
<i>r</i>	2.49%	2.87%	2.96%	
<i>VaR</i>	-14.55%	-9.43%	-8.52%	
<i>CR</i>	€\$ 686,841.90	€\$ 1,291,530.41	€\$ 1,444,995.70	
<i>CaR</i>	€\$ 99,969.16	€\$ 121,795.88	€\$ 123,120.96	
<i>RAROC</i>	17.1%	30.4%	34.7%	

Nota: Sendo *r* o retorno sobre a carteira (medido pela soma do ganho com clientes bons e o ganho da recuperação de clientes ruins), *VaR* o montante de perda máxima com intervalo de confiança de 95%, *CR* a carteira de crédito e *CaR* o montante de capital em risco (*VaR* multiplicado pela *CR*).

Assim como ocorrido no índice de Sharpe, no RAROC a regressão logística também apresenta resultados superiores ao Naïve Bayes, indicando que, ao ponderar o retorno pela volatilidade e pelo tamanho da carteira de crédito formada, a regressão logística seria mais interessante do que o classificador Naïve Bayes para a situação analisada.

5.6 Discussões e considerações sobre os resultados

Na tabela 12 são apresentados os resultados agrupados das métricas selecionadas para avaliação do desempenho de classificação e predição de *default* de cada uma das técnicas de aprendizagem estatística estudadas pela pesquisa. Nesta tabela é possível observar que diferentes métricas apresentam diferentes resultados em termos de escolha da melhor técnica de classificação. Dessa forma espera-se que modelos diferentes tendem também a apresentar resultados diferentes em cada métrica.

Chama a atenção a divergência entre a técnica estatística ACC e o critério de custo do erro de classificação, bem como destes para as métricas financeiras estudadas. Considerando o critério ACC o melhor resultado seria do SVM e SVC com a regressão logística superando a análise discriminante e as redes bayesianas Naïve Bayes, kdB-1 e kdB-2. Já analisando a matriz de custo ofertada pela base de dados, a rede bayesiana Naïve Bayes demonstrou um resultado superior à regressão logística.

O mesmo resultado foi demonstrado nas métricas RFC e PLR, sendo que para estas métricas e para a métrica “custos” o resultado do Naïve Bayes foi superior inclusive ao classificador SVC.

Tabela 12 - Comparativo geral de resultados das métricas estudadas

	Discriminante	Logística	Naïve Bayes	kdB-1
ACC	65.80%	73.40%	70.24%	69.13%
Custo	132	122.08	86.69	173.75
RFC	-€\$ 54,461.95	-€\$ 46,793.04	-€\$ 16,308.74	-€\$ 90,535.20
PLR	-€\$ 65,735.20	-€\$ 73,044.58	-€\$ 26,239.68	-€\$ 115,883.78
IS	19.67%	39.22%	21.65%	33.00%
RAROC	14.54%	25.54%	14.76%	21.96%
	kdB-2	SVC	SVM	
ACC	55.00%	76.00%	86.59%	
Custo	233.6	109.44	50.46	
RFC	-€\$ 138,164.59	-€\$ 35,087.69	€\$ 7,857.77	
PLR	-€\$ 116,485.69	-€\$ 59,441.60	-€\$ 23,463.28	
IS	23.84%	48.00%	55.14%	
RAROC	17.09%	30.39%	34.69%	

Nota: Nesta tabela são reunidos os resultados de cada métrica financeira de comparação estudada para cada uma das sete técnicas de aprendizagem estatística.

Considerando as métricas RAROC e IS o resultado do SVC e da regressão logística foram superiores ao Naïve Bayes, indicando que, para métricas que ponderam o *trade-off* risco vs retorno os resultados podem divergir das demais métricas estudadas (custos, RFC e PLR por exemplo).

Estas divergências apontam para a necessidade de ponderação, via políticas de crédito, das métricas para a definição de quais destas deverão ser utilizadas e quais terão maior participação na tomada de decisão de seleção de modelos em termos de peso relativo.

Mostra-se também fundamental considerar a carteira gerada em sua volatilidade, tamanho e composição no trabalho de análise e seleção do modelo a ser adotado, pois, como observado pela pesquisa, o modelo muda o perfil da carteira, sendo assim um instrumento que deve ser alinhado às políticas de crédito da firma.

Para todas as métricas estudadas o classificador SVM se mostrou o classificador mais eficiente, indicando que para este banco de dados e para os métodos empregados, este classificador foi superior.

Considerando a premissa de que o objetivo do crédito é alavancar as operações da empresa tendo no modelo de classificação um mecanismo capaz de oferecer proteção contra perdas derivadas do *default*, o SVM produziu uma carteira 11,88% maior do que a SVC (técnica segunda colocada neste quesito), sendo que a perda relativa do SVC foi de 184,15% maior que a do SVM (1,64% contra 4,66%). Desta forma, entende-se que pela capacidade do SVM gerar a maior carteira com os menores índices relativos de *default*, bem como com a menor volatilidade de retornos previstos, esta técnica apresentou o melhor desempenho.

6 CONSIDERAÇÕES FINAIS

A medida da eficiência e eficácia das decisões em administração está fortemente ligada ao poder que estas decisões têm de produzir resultados econômico-financeiros igualmente eficientes e eficazes. Desta forma, pesquisas que tenham como objetivo produzir mecanismos e modelos de decisão voltados para a otimização do desempenho da firma são amplamente desejáveis em Ciências Administrativas.

Considerando diversos estudos sobre modelagem e análise de risco de crédito, o presente estudo identificou um desprendimento entre a premissa da necessidade de geração de resultados e valor e as decisões de seleção de modelos de análise de risco de crédito. Adicionalmente, a pesquisa identificou significativa carência em Ciências Administrativas no cenário brasileiro no que tange a adoção de técnicas de classificação e aprendizagem estatística mais modernas e amplamente estudadas no âmbito internacional.

Essas considerações foram possíveis a partir do levantamento teórico realizado, o qual sugere que o padrão de pesquisa nacional está focado na seleção de variáveis explicativas e construção de modelos (a partir desta seleção de variáveis); enquanto no âmbito internacional o foco é a comparação entre os modelos. Em ambos os casos não se observa empenho das pesquisas em definir a eficiência do modelo a partir do provável resultado financeiro da carteira formada por cada modelo.

Tendo como premissa a possibilidade de emprego de métricas financeiras de desempenho para contribuir com métodos voltados para a solução dessa lacuna, que prejudica a decisão administrativa de seleção do modelo ótimo, e buscando demonstrar o uso e a eficiência preditiva de novas técnicas utilizadas para tratamento do problema de classificação, análise e decisão de concessão de crédito via previsão de *default* a pesquisa buscou comparar estas técnicas através da análise do desempenho das carteiras montadas por cada modelo.

Como mostrado pela revisão bibliográfica, a previsão de *default* de crédito é de suma importância para a decisão de concessão de crédito. Tal previsão consiste em analisar as características do proponente ao crédito, classificando-o segundo o esquema de classificação adotado por cada empresa e decidido via políticas de crédito. As previsões classificam os clientes segundo o risco de este honrar ou não as obrigações ora contraídas.

Entendeu-se que, através da classificação do cliente como bom ou ruim a decisão automática de concessão de crédito produz uma carteira de crédito com perfil próprio da modelagem utilizada.

O objetivo declarado desta pesquisa foi o de empregar métricas baseadas na performance financeira da carteira como critério para análise de técnicas de previsão de *default* de crédito. Com isso, buscou-se verificar qual técnica produziria os resultados mais satisfatórios em termos financeiros.

Não obstante a pesquisa utilizou também a métrica estatística ACC e as declarações de custos pelo erro de classificação presentes na própria base de dados como mecanismos de comparação dos resultados das carteiras formadas por cada técnica.

Esse conjunto de objetivos foi perseguido por meio de um estudo comparativo de sete técnicas de aprendizagem estatística, quais sejam: i) análise discriminante; ii) regressão logística; iii) rede bayesiana Naïve Bayes; iv) rede bayesiana kdB-1; v) rede bayesiana kdB-2; vi) Support Vector Classifier – SVC; e vii) Support Vector Machine.

Todos os modelos foram testados na mesma base de dados, qual seja a *German Credit Data Set* – conjunto de dados de crédito pessoa física alemão com mil observações e vinte e uma covariáveis. Para cada uma das técnicas foram realizadas cem classificações validadas via *K-Fold Cross Validation*, sendo que os resultados apresentados na pesquisa referem-se aos resultados médios obtidos por cada um dos sete classificadores estudados.

As análises foram realizadas através do *software* estatístico R, com algoritmos dos pacotes MASS e e1071, bem como do algoritmo kdB desenvolvido por Louzada e Ara (2012a). Outros pacotes de apoio são observados no Apêndice.

Como declarado nos procedimentos metodológicos, não foram realizados tratamentos objetivando otimização das técnicas, tendo em vista que o objetivo da pesquisa era demonstrar a importância e utilidade das técnicas e das métricas financeiras. Entendeu-se que, ao realizar tais tratamentos a pesquisa incorreria em um viés que poderia beneficiar uma técnica em detrimento às outras.

Entendeu-se o trabalho junto às variáveis como uma questão secundária para a presente pesquisa, pois seu objetivo foi atuar junto às lacunas apresentadas e não produzir modelos otimizados de análise de risco de crédito. Ainda que se objetivasse a otimização do modelo, os resultados não refletiriam a acurácia necessária já que a base de dados utilizada é acadêmica (e alemã) e que parte dos dados empregados como *proxy* refletem a dados reais brasileiros.

Esta postura adotada pela pesquisa não restringe a necessidade de atuar junto às variáveis selecionando sua forma funcional e sua capacidade de informação e explicação da variabilidade da variável predita. Entende-se que ao avançar rumo a aplicações práticas ganha-se em aderência e efetividade do estudo, ficando como proposta para futuras pesquisas o emprego dos procedimentos aqui apresentados em bases de dados reais.

Os resultados comprovam a importância o emprego de métricas financeiras na seleção e comparação de modelos de análise de risco de crédito, tendo em vista que mesmo a seleção por custos do erro não é capaz de cobrir de forma precisa todo o contingente de variáveis financeiras que influenciam o desempenho do modelo e da carteira.

Como é possível observar, o emprego de métricas puramente estatísticas ou de um custo de erro para escolha de um método de classificação em análise de risco de crédito pode levar a equívocos caso sejam analisadas apenas estas métricas, tendo em vista que o principal objetivo do tratamento do problema é a maximização do resultado financeiro da firma e não somente das estatísticas alcançadas.

Esta consideração é possível a partir da análise dos resultados dos modelos testados, que demonstrou valores proporcionalmente distintos em cada uma das métricas empregadas.

As análises e considerações observadas na seção 5.6 demonstram variações importantes nos resultados das técnicas na segunda e terceira colocação (considerando o *ranqueamento* destas técnicas) em termos de eficiência preditiva e composição de carteira, onde as técnicas SVC, Regressão Logística e Naïve Bayes apresentam alternância em relação à colocação.

Pelas métricas RAROC e IS, por exemplo, os resultados do SVC e da regressão logística foram melhores que os resultados apresentados pelo Naïve Bayes, o que não ocorreu nas métricas RFC e PLR. Para estas duas últimas métricas os resultados do Naïve Bayes foram superiores (Seção 5.6).

Em parte esta alternância se explica pelo tamanho de cada carteira e pela volatilidade de retornos observada em cada carteira. O modelo Naïve Bayes, tendo um volume de aprovações menor produziu uma carteira menor. Mesmo assim, este modelo não foi capaz de produzir uma baixa volatilidade de retornos (em termos proporcionais), o que produziu desempenho inferior nas métricas RAROC e IS.

Com isto, é possível indicar que as variações sobre o volume e a volatilidade da carteira formada são elementos importantes para a otimização e consequentemente para a escolha do modelo a ser empregado.

Em todas as métricas o classificador SVM produziu os melhores resultados dentre as técnicas de classificação estudadas, indicando que, para o banco de dados estudado e para os métodos empregados pela pesquisa, este classificador se mostrou superior aos demais, uma vez que produziu resultados financeiros superiores.

Dessa forma, pode-se concluir que a pesquisa atende a seus objetivos, demonstrando resultados que explicitam os benefícios do emprego de técnicas para a seleção de modelos de classificação baseadas em comparações financeiras.

Especificamente sobre a presente pesquisa entende-se que o emprego de métricas financeiras propostas é capaz de oferecer informações que auxiliam de forma mais precisa a formulação e gestão da política de crédito das firmas, ao prover informações sobre o montante emprestado, seu retorno e a volatilidade à qual o capital estaria exposto em cada situação, bem como o resultado produzido por cada modelo.

Com base nesses resultados, novos estudos podem ser delineados, tais como, por exemplo, estudos que incorporem novos classificadores, realizem otimizações quanto às variáveis e incorporem novas métricas financeiras para a análise de modelos de classificação de crédito são importantes para consolidação do principal pressuposto desta pesquisa, qual seja o de que decisões em risco de crédito devem ser tomadas com base no resultado financeiro produzido pelo modelo.

Para liderar com uma maior diversidade de métricas seria importante explorar técnicas focadas na decisão multicritério. Essas técnicas provêm suporte teórico e computacional para auxiliar na escolha de soluções (respostas) mais eficientes quando considerados, no cenário de decisão, duas ou mais métricas de comparação.

No que tange estritamente o resultado dos modelos testados, é possível verificar a eficiência do classificador SVM, sendo assim um método que deve ser levado em consideração no processo de seleção de modelos para tratamento de problemas de previsão de risco de crédito, dada a sua flexibilidade e expressivo potencial de geração de resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDOU, H; POINTON, J. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. **Intelligent Systems in Accounting, Finance and Management**, v. 18, p. 59-88, 2011.

ALISSA, W.; BONSALL IV, S. B.; KOHARKI, K.; PENN JR, M. W. Firms' use of accounting discretion to influence their credit ratings. **Journal of Accounting and Economics**, v. 55, p. 129–147, 2013.

ANDRADE; F. W. M.; THOMAS, L. Structural models in consumer credit. **European Journal of Operational Research**, v. 183, n. 3, p. 1569–1581, 2007.

ANNIBAL; C. A.; KOYAMA, S. M. Pesquisa trimestral de condições de crédito no Brasil. **Trabalhos para discussão**, Brasília, n. 245, p. 1-62, Jul. 2011.

ARA, A. **Redes probabilísticas de k-dependências para problemas de classificação binária**. Dissertação (Mestrado), Programa de Pós-Graduação em Estatística – Departamento de Estatística. Universidade de São Carlos – Ufscar. 2011, 128. p.

ARCOVERDE, G. L. O mercado de crédito no Brasil. **Notas técnicas do Banco Central**, Brasília, n. 15, p. 1-15, 2011.

BRITO, G. A. S.; ASSAF NETO, A.; CORRAR, L. J. Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil. **Revista Contabilidade & Finanças**, USP, v. 20, n. 51, p. 28-43, 2009.

BAILEY, M. **Consumer credit quality**: underwriting, scoring, fraud prevention and collections. Kingswood, Bristol: White Box Publishing, 2004.

BANCO CENTRAL DO BRASIL. **Série histórica do sistema financeiro nacional**: Operações de crédito do sistema financeiro. 2014. Disponível em: <http://www.bcb.gov.br/?SERIESFN>. Acesso em: 10/02/2014.

BESSIS, J. **Risk management in banking**. Chichester: John Wiley & Sons, 1998.

BIJAK, K.; THOMAS, L. C. Does segmentation always improve model performance in credit scoring? **Expert Systems with Applications**, v. 39, p. 2433–2442, 2012.

BLANCO, A.; PINO-MEJÍAS, R.; LARA, J.; RAYO, S. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. **Expert Systems with Applications**, v. 40, p. 356–364, 2013.

BLUHM, C.; OVERBECK, L.; WAGNER, C. **An introduction to credit risk modeling**. London: Chapman & Hall, 2003.

BRITO, G. A.; ASSAF NETO, A. Modelo de classificação de risco de crédito em empresas. **Revista de Contabilidade e Finanças**. v. 19, n. 46, p. 18-29, 2008.

BRESSAN, V. G. F.; BRAGA, M. J.; BRESSAN, A. A.; RESENDE FILHO, M. A. Avaliação de insolvência em cooperativas de crédito: uma aplicação do sistema pearls. **RAM – Revista de Administração Mackenzie**, v. 12, n. 2, p. 113-144, 2011.

BROWN, I.; MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. **Expert Systems with Applications**, v. 39, p. 3446–3453, 2012.

BÜCKER, M.; KAMPEN, M.; KRÄMER, W. Reject inference in consumer credit scoring with nonignorable missing data. **Journal of Banking & Finance**, v. 37, p. 1040–1045, 2013.

CHAN, H.; WONG, H. Y.; ZHAO, J. Structural model of credit migration. **Computational Statistics & Data Analysis**, v. 56, n. 11, p. 3477-3490, 2012.

CHUNG, H. Y.; HO, C. H.; HSU, C. C. Support vector machines using Bayesian-based approach in the issue of unbalanced classifications. **Expert Systems with Applications**, v. 38, n.9, p. 11447–11452, 2011.

COOPER, D.; SCHINDLER, P. **Método de pesquisa em administração**. 8ª Ed. Porto Alegre: Bookman, 2003.

CRONE, S. F.; FINLAY, S. Instance sampling in credit scoring: An empirical study of sample size and balancing. **International Journal of Forecasting**, v. 28, p. 224–238, 2012.

CROUHY, M.; GALAI, D.; MARK, R. A comparative analysis of current credit risk models. **Journal of Banking & Finance**, v. 24, n. 1–2, p. 59–117, 2000.

DANENAS, P.; GARSVA, G.; SAULIUS, G. Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. **Procedia Computer Science**, v. 4, p. 1699–1707, 2011.

DANTAS, R. F.; SOUZA, S. A. Modelo de risco e decisão de crédito baseado em estrutura de capital com informação assimétrica. **Pesquisa Operacional**, v.28, n.2, p.263-284, 2008.

DE-LA-VEJA, M. D, C.; BLANCO-OLIVER, A.; PINO-MEJÍAS, R.; LARA-RUBIO, J. Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques. **Expert Systems with Applications**, v. 40, p. 6910–6917, 2013.

DERMITAS, K. O.; CORNAGGIA, K. R. Initial credit ratings and earnings management. **Review of Financial Economics**, v. 22, p. 135–145, 2013.

DOUMPOS, M.; NIKLIS, D.; ZOPOUNIDIS, C.; ADRIOSPOULOS, K. Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms. **Journal of Banking & Finance**, 2014. No Prelo.

DUAN, J. C.; SHRESTHA, K. Statistical Credit Rating Metholds. **Global Credit Review**, v. 1, p. 43-64, 2011.

ELLIOTT, R. J.; SIU, T. K.; FUNG, E. S. A Double HMM approach to Altman Z-scores and credit ratings. **Expert Systems with Applications**, v. 41, p. 1553–1560, 2014.

FEBRABAN, **Relatório Anual**. 2013. Disponível em: [www. febraban.org.br/](http://www.febraban.org.br/). Acesso em: 05/12/2012.

FERSON, WAYNE E. **Investment Performance: A Review and Synthesis**. In. Handbook of the Economics of Finance, vol. 2, parte B, 2013.

FIPECAFI. **Estudo sobre a estrutura da taxa de juros no Brasil**: apuração do spread da indústria bancária. 2004. Disponível em: www.febraban.org.br/. Acesso em: 10/02/2014.

GALVÃO, A. et al. Medidas de risco e avaliação. In. GALVÃO, A. et al. **Finanças Corporativas**. 2ª ed. Rio de Janeiro: Elsevier, 2007.

GARCÍA, F.; GIMÉNEZ, V.; GUIJARRO, F. Credit risk management: A multicriteria approach to assess creditworthiness. **Mathematical and Computer Modelling**, v. 57, p. 2009-2015, 2013.

GARSIDE, T. et al. Credit portfolio management. **Oliver, Wyman & Company**, Dec. 1999. Disponível em: <<http://www.erisk.com>>. Acesso em: 12/02/2014.

GARTER, I. R.; MOREIRA, T. B. S.; GALVES, M. Análise do risco setorial como instrumento de controle gerencial em instituições financeiras. **RAM – Revista de Administração Mackenzie**, v. 10, n. 5, p. 107-129, 2009.

GERTLER, M.; KIYOTAKI, N. **Financial Intermediation and Credit Policy in Business Cycle Analysis**. IN: BENJAMIN, M.; FRIEDMAN, M.; WOODFORD, M. Handbook of Monetary Economics, v. 3, San Diego: Elsevier, 2010.

GODÓI, A.; YOSHINO, J. A.; OLIVEIRA, R. D. Risco de Crédito e Alocação Ótima para uma Carteira de Debêntures. **Estudos econômicos**, v. 38, n. 2, p. 349-372, 2008.

GUIMARÃES, I. A.; CHAVES NETO, A. Reconhecimento de padrões: metodologias estatísticas em crédito ao consumidor. **RAE-eletrônica**, v. 1, n. 2, 2002.

GUIMARÃES, A.; MOREIRA, T. B. S. Previsão de insolvência: um modelo baseado em índices contábeis com utilização da análise discriminante. **Revista de Economia contemporânea**, vol. 12, n. 1, p. 151-178, 2008

GUNDLACH, M. **Basics of CreditRisk⁺**. In GUNDLACH, M.; LEHRBASS, F. CreditRisk⁺ in the Banking Industry. Berlin: Springer, 2004.

HARRIS, T. Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. **Expert Systems with Applications**, v. 40 p. 4404–4413, 2013.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. Springer, 2008. Disponível em: <http://www.stanford.edu/~hastie/local.ftp/Springer/ESLII_print5.pdf>. Acesso em: 21/10/2013

HØJSGAARD, S.; EDWARDS, D.; LAURITZEN, S. **Graphical Models with R**. Berlin: Springer, 2012.

ISCOE, I.; KREININ, A.; MAUSSER, H.; ROMANKO, O. Portfolio credit-risk optimization. **Journal of Banking & Finance**, v. 36, n. 6, p. 1604–1615, 2012.

KRUPPA, J.; SCHWARZ, A.; ARMINGER, G.; ZIEGLER, A. Consumer credit risk: Individual probability estimates using machine learning. **Expert Systems with Applications**, v. 40, p. 5125–5131, 2013.

KELLY, D.; SMITH, C. **Bayesian inference for probabilistic risk assessment**. London: Springer, 2011.

KOLLER, D. FRIEDMAN, N. **Probabilistic Graphical Models: Principles and Techniques**. Cambridge: MIT Press, 2009.

LESKO, M.; SCHLOTTMAN, F.; VORGRIMBER, S. **Estimation of sector weights from real-world data**. In GUNDLACH, M.; LEHRBASS, F. CreditRisk+ in the Banking Industry. Berlin: Springer, 2004.

LOUZADA-NETO, F.; ARA, A. Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool. **Expert Systems with Applications**, v. 39, n. 14, p. 11583–11592, 2012a.

LOUZADA-NETO, F.; ARA, A. Modelagem de Crédito via Redes Probabilísticas KDB. **Revista tecnologia em crédito**, n. 80, p. 20–29, 2012b.

LOUZADA-NETO, F.; FERREIRA-SILVA, P. H.; DINIZ, C. A. R. On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data **Expert Systems with Applications**, v. 39, p. 8071–8078, 2012.

LU, F. Q.; HUANG, M.; CHING, W. K.; SIU, T. K. Credit portfolio management using two-level particle swarm optimization. **Information Sciences**, v. 237, n. 10, p. 162–175, 2013.

MAJUMDER, D. Inefficient markets and credit risk modeling: Why Merton's model failed? **Journal of Policy Modeling**, v. 28, n. 3, p. 307–318, 2006.

MANSO, G. Feedback effects of credit ratings. **Journal of Financial Economics**, v. 109, p. 535–548, 2013.

MARQUÉS, A. I.; GARCÍA, V.; SÁNCHEZ, J. S. Two-level classifier ensembles for credit risk assessment. **Expert Systems with Applications**, v. 39, p. 10916–10922, 2012.

MATIAS, A. B. **Finanças corporativas de curto prazo**. São Paulo: Atlas, 2007.

MCLACHLAN, G. Multivariate Analysis: Classification and Discrimination. In. SMELSER, N. J.; BALTES, P. B. **International Encyclopedia of the Social & Behavioral Sciences**, Nova York: Elsevier, 2001.

MERTON, R. C. On the pricing of corporate debt: the risk structure of interest rates. **The Journal of Finance**. Berkeley, v. 29, n. 2, p.449-470, 1974.

MINUSSI, J. A.; DAMACENA, C.; NESS JR, W. L. Um Modelo de Previsão de Solvência Utilizando Regressão Logística. **RAC – Revista de Administração Contemporânea**, v. 6, n. 3, p. 109-128, 2002.

NAGARAJAN, R.; SCUTARI, M.; LÈBRE, S. **Bayesian Networks in R with Applications in Systems Biology**. Nova York: Springer, 2013.

NIKOLIC, N.; ZARKIC-JOKSIMOVIC, N.; STOJANOVSKI, D.; JOKSIMOVIC, I. The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. **Expert Systems with Applications**, v. 40, p. 5932–5944, 2013.

OLIVEIRA FILHO, L. M.; ALMEIDA, A. M. L.; MATSUMOTO, A. S. **Gerenciamento de risco**: um estudo de caso de um banco comercial na região nordeste sobre recuperação de crédito e seu impacto na constituição da PCLD. 2014. Disponível em: <www.bcb.gov.br>

PASCUAL, M. B.; MARTÍNEZ, A. M.; ALAMILLOS, A. M. Redes bayesianas aplicadas a problemas de credit scoring. Una aplicación práctica. **Cuadernos de Economía**, v.37, n. 104, p. 73-86, 2014.

PROKOPCZUKE, M.; RACHEVA, S. T.; SCHINDLMAYRD, G.; TRÜCKC, S. Quantifying risk in the electricity business: A RAROC-based approach. **Energy Economics Review**, v. 29, n. 5, p. 1033–1049, 2017.

SAITA, F. **Value at Risk and Bank Capital Management**: Risk adjusted performance, capital management and capital allocation decision making. San Diego: Elsevier, 2007.

SARIG, O.; WARGA, A. Some Empirical Estimates of the Risk Structure of Interest Rates. **Journal of Finance**, v. 44, p. 1351-1360, 1989.

SANTOS, J. O.; FAMÁ, R. Avaliação da aplicabilidade de um modelo de *credit scoring* com variáveis sistêmicas e não-sistêmicas em carteiras de crédito bancário rotativo de pessoas físicas. **Revista de Contabilidade e Finanças – USP**, n. 44, p. 105 – 117, 2007.

SHEVCHENKO, P. V. **Modeling operational risk using Bayesian inference**. London: Springer, 2011.

SMITHSON, Charles. **Credit portfolio management**. Hoboken: John Wiley & Sons, 2003.

SOHN, S. Y.; KIM, J. W. Decision tree-based technology credit scoring for start-up firms: Korean case. **Expert Systems with Applications**, v. 39, p. 4007–4012, 2012.

STEINER, M. T. A.; NIEVOLA, J. C.; SOMA, N. Y.; SHIMIZU, T.; STEINER NETO, P. J. Extração de regras de classificação a partir de redes neurais para auxílio à tomada de decisão na concessão de crédito bancário. **Pesquisa Operacional**, v. 27, n. 3, p. 407-426, 2007

WANG, J.; HEDAR, A. H.; WANG, S.; MA, J. Rough set and scatter search metaheuristic based feature selection for credit scoring. **Expert Systems with Applications**, v. 39, p. 6123–6128, 2012.

ZAMBALDI, F.; ARANHA, F.; LOPES, H.; POLITI, R. Credit granting to small firms: A Brazilian case. **Journal of Business Research**, v. 64, p. 309–315, 2011.

APÊNDICE A – Classificação via Naïve Bayes

```
# Requisitando pacotes
```

```
require(e1071)  
require(cvFolds)
```

```
# Abrindo conjunto de dados
```

```
dados=read.csv("C:\\...\\GermanCreditDataset.csv", sep=';', header=T)
```

```
#Entrada dos dados
```

```
dados=read.csv("C:\\...\\GermanCreditDataset.csv", sep=';', header=T)
```

```
# Discretizando os dados (tornado-os categorias)
```

```
require(infotheo)
```

```
dados=discretize(dados)
```

```
#Calculando porção de treinamento
```

```
Require(cvFolds)
```

```
porc=cvFolds(200, K=10, R=1, type=c("random"))
```

```
m.fin<-porc
```

```
#Classificando via Naïve Baeyes
```

```
classifier<-naiveBayes(m.fin[,1:20], m.fin[,21])
```

```
table(predict(classifier, m.fin[, -21]), m.fin[, 21], dnn=list('predicted','actual'))
```

```
classifier$apriori
```

```
classifier$tables
```

APÊNDICE B – Classificação via kdB

```
# Calculando Redes Bayesianas kdB-1
# Número de dependências
K=1

#Entrada dos dados
dados=read.csv("C:\\...\\GermanCreditDataset.csv", sep=';', header=T)

# Discretizando os dados (tornado-os categorias)
require(infotheo)
dados=discretize(dados)

#Calculando porção de treinamento
Require(cvFolds)
porc=cvFolds(200, K=10, R=1, type=c("random"))

# Priori para modelo bayesiano multinomial + dirichlet - não informativo
priori=0.000001

# Classificação

# Algoritmo kdB Louzada e Ara (2012a)

# Resultado das métricas de desempenho
result
```

APÊNDICE C – Classificação via SVC

```
# Abrindo pacotes
require(e1071)
require(infotheo)
require(ROCR)

#Entrada dos dados
dados=read.csv("C:\\...\\GermanCreditDataset.csv", sep=';', header=T)

# Discretizando os dados
dados=discretize(dados)

# Transformando em uma dataframe
dat=data.frame(dados)

# Suppor Vector Classifier

# Aplicando o svc aos dados
x <- subset(dat, select = -Y)
y <- dat$Y
dat=data.frame(x=x, y=as.factor (y))

# Calculando o treinamento SVC
svmfit=svm(y~., data=dat, kernel = "linear", cost = 10, scale=FALSE)
plot(svmfit, dat)
summary(svmfit)
tune.out=tune(svm, y~., data=dat, kernel = "linear", ranges = list(cost=c(0.001 , 0.01, 0.1,
1,5,10,100)))
summary(tune.out)

# Salvando o melhor modelo

bestmod = tune.out$best.model
summary(bestmod)

# Calculando a amostra

randomSample = function(df,n) {
  return (df[sample(nrow(df), n),])
}
train<-randomSample(dat, 200)
x <- subset(train, select = -y)
y <- train$y
testdat = data.frame(x=, y=as.factor(Y))

# Calculando a tabela matriz de confusão para Support Vector Classifier
train<-randomSample(dat, 200)
ypred = predict(bestmod, train)
table(predict=ypred, truth=train$y)
```


APÊNDICE D – Classificação via SVM

```
# Abrindo pacotes
require(e1071)
require(infotheo)
require(ROCR)

#Entrada dos dados
dados=read.csv("C:\\...\\GermanCreditDataset.csv", sep=';', header=T)

# Discretizando os dados
dados=discretize(dados)

# Transformando em uma dataframe
dat=data.frame(dados)

# SVM - Support Vector Machine

svmfit=svm(y~., data = dat, kernel= "radial", gamma=1, cost=1)
summary(svmfit, dat)

tune.out=tune(svm, y~., data=dat, kernel = "radial", ranges = list(cost=c(0.001 , 0.01, 0.1,
1,5,10,100)))
bestmod = tune.out$best.model
summary(tune.out, dat)

# Calculando a amostra
randomSample = function(df,n) {
  return (df[sample(nrow(df), n),])
}

# Calculando a tabela matriz de confusão para Support Vector Machine
train<-randomSample(dat, 200)
ypred = predict(bestmod, train)
table(predict=ypred, truth=train$y)
```