UNIVERSIDADE NOVE DE JULHO - UNINOVE

Programa de Mestrado em Engenharia de Produção

INTELIGÊNCIA COMPUTACIONAL APLICADA NA ANÁLISE E RECUPERAÇÃO DE PORTFÓLIOS DE CRÉDITOS DO TIPO *NON-PERFORMING LOANS*

FLÁVIO CLESIO SILVA DE SOUZA

SÃO PAULO

2015

FLÁVIO CLESIO SILVA DE SOUZA

INTELIGÊNCIA COMPUTACIONAL APLICADA NA ANÁLISE E RECUPERAÇÃO DE PORTFÓLIOS DE CRÉDITOS DO TIPO *NON-PERFORMING LOANS*

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Nove de Julho, como parte dos requisitos exigidos para a obtenção do grau de Mestre em Engenharia de Produção.

Prof. Renato José Sassi, Dr.– Orientador, UNINOVE

SÃO PAULO

2015

INTELIGÊNCIA COMPUTACIONAL APLICADA NA ANÁLISE E RECUPERAÇÃO DE PORTFÓLIOS DE CRÉDITOS DO TIPO *NON-PERFORMING LOANS*

Por

FLÁVIO CLÉSIO SILVA DE SOUZA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Nove de Julho, como parte dos requisitos exigidos para a obtenção do grau de Mestre em Engenharia de Produção, pela Banca Examinadora, formada por:

_____

Presidente: Prof. Dr. Renato José Sassi, UNINOVE

_____

Membro: Prof. Dr. Cleber Gustavo Dias, UNINOVE

_____

Membro: Prof. Dr. Reinaldo Castro Souza, PUC-RIO

SÃO PAULO

2015

FICHA CATALOGRÁFICA

Souza, Flávio Clesio Silva de.
Inteligência computacional aplicada na análise e recuperação de portfólios de créditos do tipo non-performing loans. / Flávio Clesio Silva de Souza. 2015.
169 f.
Dissertação (mestrado) – Universidade Nove de Julho - UNINOVE, São Paulo, 2015.
Orientador (a): Prof. Renato José Sassi.
1. Non-performing loans. 2. Fundos de investimento em direitos creditórios. 3. Inteligência computacional.
I. Sassi, Renato José. II. Titulo

CDU 658.5

Dedico este trabalho ao Grande Arquiteto da vida e a Maria Francisca da Silva.

"Somente a consciência individual do agente dá testemunho dos atos sem testemunha, e não há ato mais desprovido de testemunha externa do que o ato de conhecer." (Olavo de Carvalho)

AGRADECIMENTOS

Primeiramente a Deus, pela ótima vida e por me mostrar que o conhecimento sempre será a única forma de chegar a Ele.

A minha mãe Maria Francisca da Silva por acreditar nas minhas aspirações e por me mostrar que o conhecimento é a melhor forma de transformação de sonhos em realidade.

Ao meu orientador Dr. Renato Sassi por acreditar no potencial do projeto desde a sua concepção inicial em 2012, e por ser um excelente parceiro intelectual.

Aos membros da banca, Dr. Cleber Gustavo Dias, e Dr. Reinaldo Castro Souza pelas importantes sugestões e críticas que enriqueceram o trabalho.

A Universidade Nove de Julho pela bolsa de estudos concedida, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro através da bolsa PROSUP, e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) por premiar o projeto com a bolsa PIBITI.

Aos docentes da Universidade Nove de Julho em especial: Dr. Fabio Henrique Pereira, Dr. Geraldo Cardoso de Oliveira Neto, Dr. Leonardo Junqueira, Dr. André Felipe Henriques Librantz, e Dr. Sidnei Alves de Araújo.

A todos os alunos do programa de mestrado, em especial: Ellen Martins Lopes da Silva, Elizangela de Jesus Gibelati, Fábio Cosme Rodrigues dos Santos, Stanley Jefferson de Araújo Lima, Valdemar Modolo Junior, Flávio Grassi, Marcelo Drudi Miranda, e Paulo Kaupa.

A Paola Ceccaci pelo companheirismo ao longo da última década e principalmente por acreditar e me incentivar a realizar todos os meus sonhos.

À RCB Investimentos pelo suporte acadêmico e profissional em ceder parte das bases de dados para os experimentos, em especial: Renato Proença Prudente de Toledo, Pedro Luiz Melchiades Gomes, Marco Aurélio de Camilo Mattos, Maurício de Figueiredo Longato, Tadeu Marques de Barros Deperon, Sidney Augusto Gamito, Jefferson Almeida, Daniel Augusti Graziano, e Daniel Fernando Cypas.

Ao Diego Farias, pela revisão na língua vernácula.

RESUMO

Uma das externalidades econômicas relativas ao aumento do crédito é o consequente aumento da taxa de inadimplência. Em face deste cenário econômico surgiu no Brasil a oferta por meio da venda dos direitos desses créditos por parte das mais diversas instituições financeiras e bancárias. Esses créditos inadimplidos são chamados de *Non-Performing Loans* (NPLs), ou créditos não-performados. Já a demanda em relação à aquisição dos NPLs é feita por estruturas econômicas denominadas de Fundos de Investimento em Direitos Creditórios (FIDC) que objetivam o retorno financeiro aos investidores por meio da recuperação desses créditos. Os fundos realizam diversas análises de viabilidade de negócios por meio de técnicas estatísticas, financeiras, e econômicas em busca dos fatores determinantes para a recuperação desses créditos. No entanto, outras técnicas como as de Inteligência Computacional podem ser aplicadas na análise desses créditos. Este trabalho tem como objetivo principal a aplicação de técnicas de Inteligência Computacional na análise e recuperação de portfólios de créditos do tipo *Non-Performing Loans*. Os trabalhos da literatura, até o momento, não abordam diretamente questões de como essas análises influenciam diretamente na capacidade de avaliação ou precificação desses ativos financeiros, como também na recuperação dos créditos partindo da perspectiva do crédito já inadimplido, e de quais são os determinantes que influenciam na recuperação desses créditos. Foram realizados três experimentos utilizando as seguintes técnicas de Inteligência Computacional: Redes Neurais Artificiais, Teoria dos *Rough Sets* e Árvores de Decisão, aplicadas de forma isolada ou de forma combinada. Os resultados obtidos com a aplicação das técnicas foram conclusivos ao destacar que os fatores relacionados às formas de localização dos devedores, a idade do crédito, e o valor da dívida são os principais determinantes na recuperação dos NPLs, e que, portanto devem ser levados em consideração no suporte à decisão seja para atividades de avaliação e precificação desses ativos, seja para elaboração de estratégias de recuperação desses ativos.

Palavras-chave: *Non-Performing Loans*, Fundos de Investimento em Direitos Creditórios, Inteligência Computacional, Tomada de Decisão, Créditos Não-Perfomados.

ABSTRACT

One of the economic externalities for the credit increase is the resulting increase in the default rate. In the face of this economic scenario emerged in Brazil financial and banking institutions are offer through the sale of such claims. These defaulted credits are called Non-Performing Loans (NPLs). The demand to the purchase of NPLs is made by economic structures called Fundos de Investimentos em Direitos Creditórios (FIDC), that are aimed to get financial return through the recovery of such credits. The funds perform several analysis of business viability through statistical techniques, financial, and economic conditions in search of the determinants that influences directly in the recovery of these credits. However, other techniques of the computational intelligence can be applied in the analysis of such credits. This main objective of this work is the application of Computational Intelligence techniques in the recovery of Non-Performing Loans. Studies in the literature, yet not directly address questions of how these analyzes directly influence the evaluation capacity or pricing of these financial assets, the recovery of credits based on the already defaulted credit perspective; and also what are the determinants that influence the recovery of these credits. Studies in the literature, yet not directly address questions of how these analyzes directly influence the evaluation capacity or pricing of these financial assets, the recovery of credits based on the already defaulted credit perspective; and also what are the determinants that influence the recovery of these credits. Three experiments were conducted using the following Computational Intelligence techniques: Artificial Neural Networks, Theory of *Rough Sets* and Decision Trees. The results obtained with the application of techniques were conclusive to point out that the factors related to the forms of location of debtors, aging, and the value of debt are the main determinants in the recovery of NPLs; and therefore must be taken into account in decision support is for evaluation activities and pricing of these assets, is to prepare recovery strategies of these assets.

Keywords: Non-Performing Loans, *Fundos de Investimento em Direitos Creditórios*, Computational Intelligence, Decision Making, Default Loan.

LISTA DE QUADROS

Quadro 1 - Características principais dos algoritmos de Árvore de Decisão. 52

Quadro 2 - Algoritmo de Treinamento do *Backpropagation*. 56

Quadro 3 - Caracterização da presente pesquisa, conforme os tópicos de coberta de revisão de literatura proposto por Cooper (1988). 58

Quadro 4 - Levantamento Amostral. 60

Quadro 5 - *Softwares* utilizados para a realização dos experimentos e criação de gráficos, quadros e tabelas. 63

Quadro 6 - Tabela de custos. Pontuações para recompensa e penalização do modelo. 66

Quadro 7 - Descrição dos parâmetros utilizados para todas as cinco Redes Neurais Artificiais do experimento. 67

Quadro 8 - Atributos e as respectivas descrições da base de dados utilizada para os experimentos com as RNAs. 70

Quadro 9 - Parâmetros utilizados para configuração da rede SOM para geração dos mapas. 74

Quadro 10 - Parâmetros utilizados para configuração dos *Rough Sets* para a extração de regras. 76

Quadro 11 - Tabela de custos do modelo. 81

Quadro 12 - Parâmetros utilizados para configuração dos *Rough Sets* para redução de atributos utilizando o Algoritmo de Johnson. 82

Quadro 13 - Parâmetros utilizados para configuração dos *Rough Sets* para redução de atributos utilizando os Algoritmos Genéticos. 82

Quadro 14 - Parâmetros utilizados para geração das três Árvores de Decisão. 84

Quadro 15 - Parâmetros utilizados para configuração das cinco RNAs. 88

Quadro 16 - Parâmetros das topologias das RNAs utilizadas nos experimentos. 89

Quadro 17 - Parâmetros utilizados para configuração da rede SOM. 94

Quadro 18 - Parâmetros utilizados para configuração dos *Rough Sets* para a extração de regras. 94

Quadro 19 - Regras para a classe de decisão dos débitos não recuperados. 102

Quadro 20 - Regras para a classe de decisão dos débitos recuperados. 102

Quadro 21 - Parâmetros utilizados para configuração dos *Rough Sets* utilizando o método do Algoritmo de Johnson. 111

Quadro 22 - Parâmetros utilizados para configuração dos *Rough Sets* utilizando o método de Algoritmos Genéticos. 111

Quadro 23 - Parâmetros utilizados para geração das Árvores de Decisão. 113

Quadro 24 - Resultados dos Redutos de acordo com os algoritmos utilizados. 113

Quadro 25 - Regras de decisão provenientes da Árvore de Decisão com todas as variáveis. 122

Quadro 26 - Regras de decisão provenientes da Árvore de Decisão com o Algoritmo de Johnson. 132

Quadro 27 - Regras de decisão provenientes da Árvore de Decisão com os Algoritmos Genéticos. 141

LISTA DE ILUSTRAÇÕES

Figura 1 - Saldo das Operações de Crédito no Brasil por instituições financeiras sob controle público e privado. Fonte: Sistema Gerenciador de Séries Temporais - https://www3.bcb.gov.br/sgspub/ 30

Figura 2 - Crédito classificado como risco H em instituições privadas. Fonte: Sistema Gerenciador de Séries Temporais - https://www3.bcb.gov.br/sgspub/ 32

Figura 3 - Crédito classificado como risco H em instituições públicas. Fonte: Sistema Gerenciador de Séries Temporais - https://www3.bcb.gov.br/sgspub/ 33

Figura 4 - Fluxo Operacional simplificado de um FIDC. Fonte: Elaborada pelo Autor. 42

Figura 5 - Modelo de Estratégia para Créditos Não-Performados. Fonte: Adaptado de International Finance Corporation (2012) 45

Figura 6 - Arquitetura da rede SOM. Fonte: Adaptada de Kohonen (1982). 54

Figura 7 - Arquitetura da MLP do Modelo 4 (M4) com camada de entrada, camadas escondidas e camada de saída. Fonte: Elaborada pelo Autor. 55

Figura 8 - Distribuição geográfica dos saldos dos NPLs relativa à base de dados utilizada para os experimentos com as RNAs. Fonte: Elaborada pelo Autor. 71

Figura 9 - Evolução temporal dos NPLs ao longo dos anos de acordo com a data de celebração dos contratos, relativa à base de dados utilizada para os experimentos com as RNAs. Fonte: Elaborada pelo Autor. 72

Figura 10 - Distribuição geográfica dos saldos dos NPLs relativa à base de dados utilizada para os experimentos com a rede SOM e *Rough Sets*. Fonte: Elaborada pelo Autor. 78

Figura 11 - Evolução temporal dos NPLs ao longo dos anos de acordo com a data de celebração dos contratos, relativa à base de dados utilizada para os experimentos com a rede SOM e *Rough Sets*. Fonte: Elaborada pelo Autor. 79

Figura 12 - Evolução temporal dos NPLs ao longo dos anos de acordo com a data de celebração dos contratos, relativa à base de dados utilizada para os experimentos com *Rough Sets* e Árvores de Decisão. Fonte: Elaborada pelo Autor. 86

Figura 13 - Distribuição geográfica dos saldos dos NPLs relativa à base de dados utilizada para os experimentos com *Rough Sets* e Árvores de Decisão. Fonte: Elaborada pelo Autor 87

Figura 14 - Distribuição do erro médio ao longo das épocas durante a etapa de treinamento. Fonte: Elaborada pelo Autor. 90

Figura 15 - Mapa Topológico com a determinação dos *Clusters*. Fonte: Elaborada pelo Autor. 95

Figura 16 - Mapa Topológico das Dívidas Pagas. Fonte: Elaborada pelo Autor. 97

Figura 17 - Mapa Topológico do saldo do crédito na data do atraso. Fonte: Elaborada pelo Autor.....98

Figura 18 - Mapa Topológico do saldo principal das dívidas sem a implicação dos juros. Fonte: Elaborada pelo Autor. 99

Figura 19 - Mapa Topológico do saldo na abertura do crédito (*i.e.* valor do contrato). Fonte: Elaborada pelo Autor. 100

Figura 20 - Mapa Topológico Saldo Principal. Fonte: Elaborada pelo Autor.....101

Figura 21 - Nó Raiz do Modelo 1. Fonte Elaborada pelo Autor. 119

Figura 22 - Modelo 1 - Nó 2. Fonte: Elaborada pelo Autor. 121

Figura 23 - Modelo 3 - Nó 2. Fonte: Elaborada pelo Autor. 124

Figura 24 - Modelo 3 - Nó 3. Fonte: Elaborada pelo Autor. 126

Figura 25 - Modelo 3 - Nó 4. Fonte: Elaborada pelo Autor. 127

Figura 26 - Modelo 3 - Nó 5. Fonte: Elaborada pelo Autor. 128

Figura 27 - Modelo 3 - Nó 6. Fonte: Elaborada pelo Autor. 129

Figura 28 - Modelo 3 - Nó 7. Fonte: Elaborada pelo Autor. 130

Figura 29 - Modelo 2 - Nó 0. Fonte: Elaborada pelo Autor. 133

Figura 30 - Modelo 2 - Nó 2. Fonte: Elaborada pelo Autor. 135

Figura 31 - Modelo 2 - Nó 3. Fonte: Elaborada pelo Autor. 136

Figura 32 - Modelo 2 - Nó 5. Fonte: Elaborada pelo Autor. 138

Figura 33 - Modelo 2 - Nó 6. Fonte: Elaborada pelo Autor. 140

LISTA DE TABELAS

Tabela 1 - Exemplo de um Sistema de Informação (S). 50

Tabela 2 - Sistema de Decisão com os elementos Ação2 e Ação3 indiscerníveis. 50

Tabela 3- Saldos e quantidade de contratos da base de dados para os experimentos com as RNAs distribuídos nas classes de dívidas pagas e não pagas. 69

Tabela 4 - Saldos e quantidade de contratos da base de dados para os experimentos com as RNAs distribuídos no atributo tipo de crédito e posteriormente nas classes de dívidas pagas e não pagas. 70

Tabela 5 - Saldos e quantidade de contratos da base de dados para os experimentos com a rede SOM e *Rough Sets* nas classes de dívidas pagas e não pagas. 77

Tabela 6 - Saldos e quantidade de contratos da base de dados para os experimentos com a rede SOM e *Rough Sets* distribuídos nas classes de dívidas pagas e não pagas e de acordo com o atributo categoria. 77

Tabela 7 - Saldos e quantidade de contratos da base de dados para os experimentos *Rough Sets* e Árvores de Decisão distribuídos no atributo tipo de crédito e posteriormente nas classes de dívidas pagas e não pagas. 86

Tabela 8 - Resultados dos experimentos de acordo com as métricas de avaliação de modelos. 90

Tabela 9 - Resultados dos experimentos de acordo com as métricas de avaliação de classificadores. 90

Tabela 10 - Resultado final dos custos de cada modelo através da abordagem sensível ao custo. 91

Tabela 11 - Distribuição dos créditos nos *Clusters*. 95

Tabela 12 - Distribuição dos créditos nos *Clusters* 96

Tabela 13 - Distribuição dos créditos nos *Clusters*. 96

Tabela 14 - Resultados dos Modelos de Árvores de Decisão. 115

Tabela 15 - Matriz de confusão com todos os três modelos de Árvores de Decisão. 116

Tabela 16 - Resultados dos modelos de Árvores de Decisão de acordo com métricas de avaliação. 117

Tabela 17 - Resultados dos modelos de acordo com métricas de avaliação de classificadores. 118

LISTA DE ABREVIATURAS E SIGLAS

AG	-	Algorítmos Genéticos
AID	-	*Automatic Interaction Detector*
BANCO CENTRAL DO BRASIL	-	Banco Central do Brasil
CC	-	Cartão de Crédito
CDC	-	Crédito Direto ao Consumidor
CF	-	Constituição Federal do Brasil de 1988
CHAID	-	*Chi-squared Automatic Interaction Detector*
CP	-	Crédito Pessoal
CVM	-	Comissão de Valores Mobiliários
FIDC	-	Fundos de Investimento em Direitos Creditórios
FMI	-	Fundo Monetário Internacional
GLM	-	*Generalized Linear Model*
IA	-	Inteligência Artificial
IC	-	Inteligência Computacional
IMF	-	*International Monetary Fund*
LMT	-	*Logistic Model Tree*
MAID	-	*Multivariate Automatic Interaction Detector*
MARS	-	*Multi Adaptative Regression Splines*
MLP	-	*Multilayer Perceptron*
MtM	-	*Mark-To-Market*
NPL	-	*Non-Performing Loans*
PDD	-	Provisão para Devedores Duvidosos
RN	-	Redes Neurais
RNA	-	Redes Neurais Artificiais
RS	-	*Rough Sets*
RST	-	*Rough Sets Theory*
SOM	-	*Self-Organizing Maps*
SVM	-	*Support Vector Machines*
THAID	-	*THeta Automatic Interaction Detector*
TL	-	*Total Loss*
WCL	-	*Weighted Capital Loss*

SUMÁRIO

1 INTRODUÇÃO 18
1.1 *NON-PERFORMING LOANS* (NPLs) 20
1.2 FUNDOS DE INVESTIMENTO EM DIREITOS CREDITÓRIOS E ANÁLISE DE PORTFÓLIOS 21
1.3 INTELIGÊNCIA COMPUTACIONAL 22
1.4 JUSTIFICATIVA E MOTIVAÇÃO 24
1.5 PROBLEMA DE PESQUISA 25
1.6 OBJETIVOS 25
1.6.1 Objetivo Geral 25
1.6.2 Objetivos Específicos 25
1.7 HIPÓTESES DE PESQUISA 26
1.8 DELIMITAÇÕES DO ESTUDO 26
1.9 ORGANIZAÇÃO DO TRABALHO 27

2 FUNDAMENTAÇÃO TEÓRICA 29
2.1 *NON-PERFORMING LOANS* (CRÉDITOS NÃO-PERFORMADOS) 29
2.1.1 Pré-Crise – Antes de 2007 34
2.1.2 Crise e Pós-Crise – Após 2007 36
2.2 FUNDOS DE INVESTIMENTO EM DIREITOS CREDITÓRIOS 42
2.3 INTELIGÊNCIA COMPUTACIONAL 46
2.3.1 Árvores de Decisão 46
2.3.2 Teoria dos *Rough Sets* 48
2.3.3 *Chi-squared Automatic Interaction Detection* - CHAID 51
2.3.4 Redes Neurais Artificiais 53

3 MATERIAIS E MÉTODOS 57
3.1 CARACTERIZAÇÃO METODOLÓGICA 57
3.2 ORGANIZAÇÃO DA REVISÃO DE LITERATURA 58
3.3 PLANEJAMENTO AMOSTRAL 59
3.4 METODOLOGIA EXPERIMENTAL 62
3.4.1 Bases de Dados, Ferramentas e Plataformas de Experimentos 62
3.4.2 Extração e Tratamento Inicial dos Dados 64
3.5 CONDUÇÃO DOS EXPERIMENTOS 64
3.5.1 Experimento 1: Experimento com Redes Neurais Artificiais 64
3.5.2 Experimento 2: Experimento com *Self-Organizing Maps* conjuntamente com a Teoria dos *Rough Sets* 72
3.5.3 Experimento 3: Experimento com a Teoria dos *Rough Sets* conjuntamente com as Árvores de Decisão 79

4 REALIZAÇÃO DOS EXPERIMENTOS E DISCUSSÃO DOS RESULTADOS 88
4.1 REALIZAÇÃO DOS EXPERIMENTOS 88

4.1.1 Experimento 1: Experimento com Redes Neurais Artificiais 88
4.1.2 Discussão dos Resultados 91
4.1.3 Experimento 2: Experimento com *Self-Organizing Maps* conjuntamente com a Teoria dos *Rough Sets* 93
4.1.4 Discussão dos Resultados 103
4.1.5 Experimento 3: Experimento com a Teoria dos *Rough Sets* conjuntamente com Árvores de Decisão 110
4.1.6 Discussão dos Resultados 143

5 CONCLUSÕES 144

PUBLICAÇÕES DO AUTOR 147

REFERÊNCIAS BIBLIOGRÁFICAS 148

APÊNDICE A – ATRIBUTOS E AS RESPECTIVAS DESCRIÇÕES DA BASE DE DADOS UTILIZADA PARA OS EXPERIMENTOS COM A REDE SOM E A EXTRAÇÃO DE REGRAS UTILIZANDO ROUGH SETS. 159

APÊNDICE B – ATRIBUTOS E AS RESPECTIVAS DESCRIÇÕES DA BASE DE DADOS UTILIZADA PARA OS EXPERIMENTOS COM ROUGH SETS E AS ÁRVORES DE DECISÃO. 163

APÊNDICE C – EXPANSÃO COMPLETA DO MODELO 1 - NÓ 1. 166

APÊNDICE D – EXPANSÃO COMPLETA DO MODELO 3 - NÓ 0. 167

APÊNDICE E – EXPANSÃO COMPLETA DO MODELO 3 - NÓ 8. 168

APÊNDICE F – EXPANSÃO COMPLETA DO MODELO 2 - NÓ 4. 169

1 INTRODUÇÃO

A possibilidade de criação dos Fundos de Investimento em Direitos Creditórios (FIDCs), ou Fundo de Recebíveis como um instrumento de captação de recursos foi uma decisão do Conselho Monetário Nacional e a sua regulamentação foi exposta na Resolução nº 2907, de 29 de Novembro de 2001 do Banco Central do Brasil (BANCO CENTRAL DO BRASIL, 2001).

Estes fundos são constituídos e regulamentados de acordo com a instrução CVM nº 356, de 17 de dezembro de 2001, e que através da Instrução CVM nº 471 tiveram as suas ofertas públicas de distribuição de valores simplificados. (COMISSÃO DE VALORES MOBILIÁRIOS, 2001; COMISSÃO DE VALORES MOBILIÁRIOS, 2008).

Os direitos creditórios de acordo com a definição do Banco Central configuram-se como títulos de representatividade creditícia realizadas nos segmentos financeiro, comercial, industrial, imobiliário, de hipotecas, de arrendamento mercantil e de prestação de serviços, entre outros (BANCO CENTRAL DO BRASIL, 2001).

Desta forma, esta modalidade de fundos de investimento tem como principal finalidade a aquisição de direitos creditórios, em que a remuneração dos investidores se dará pelo recebimento e/ou recuperação de tais direitos de crédito.

O crescimento econômico nacional dos últimos dez anos que possibilitou que mais famílias tivessem acesso à combinação consumo e crédito, juntamente com a evolução do crédito bancário em relação ao Produto Interno Bruto (PIB), trás um problema relativamente novo no aspecto microeconômico que é a inadimplência que atinge, segundo estimativas, a 57 milhões de brasileiros (BANCO CENTRAL DO BRASIL, 2009; SERASA EXPERIAN, 2014).

Com o aumento da inadimplência, as instituições financeiras e bancárias perdem liquidez monetária devido à interrupção do fluxo de pagamentos ocasionado pela ausência dos pagamentos desses créditos que se transformaram em dívidas.

Desta forma, as instituições financeiras como forma de compensar o efeito dessa queda em seus respectivos fluxos de pagamentos, realizam a distribuição dessas perdas seja na forma de maior controle intrabancário no momento da concessão do crédito com políticas restritivas, isto é, menos propensas a emprestar ou a vender a prazo; ou em um segundo fenômeno, que é o mais comum, essas instituições elevam a taxa de juros para as operações creditícias.

Com essa inadimplência, os agentes econômicos que têm a possibilidade de realizar a transferência desses direitos creditórios executam essa atividade para fins de antecipação de recebíveis para fins de geração de caixa, mitigação de risco de inadimplência, eliminar a

necessidade de realizar empréstimos bancários e o consequente pagamento de juros para investimentos, ou reestruturar as perdas contábeis nos casos em que proceder qualquer tipo de ação para recuperação desse ativo, mas isso já não é economicamente viável.

Uma das formas que essas instituições podem usar como forma de liquefação desses ativos, isto é, trazer do estado de perda contábil para o estado de liquidez monetária, é a venda desses direitos creditórios.

Em outras palavras, pode-se dizer que com as perdas contábeis já configuradas e com essas perdas já diluídas através dos juros e da restrição creditícia, o crédito fica mais caro para o sistema nacional de crédito como um todo. Esta operação de transferência de perdas contábeis desse tipo de título é conhecida internacionalmente como *Write-Off*.

Em face desse problema relativo à inadimplência a Inteligência Computacional pode auxiliar na tomada de decisão e da análise desses ativos em três fases:

- **Precificação (*Pricing*) e Pré-Aquisição:** Consiste em um estudo inicial dos créditos a serem adquiridos. A prática comum no mercado é a disponibilização de bases de dados contendo informações sobre os devedores, avalistas, dívidas, garantias, entre outras informações relevantes para que os investidores possam avaliar o valor do ativo que se pretende vender. Neste momento os investidores propõem o valor que acreditam ser justo dada a relação risco-retorno do ativo e o vendedor toma a decisão sobre a venda. É a etapa da precificação e venda da carteira de créditos inadimplidos.

- **Operacionalização e acompanhamento de desempenho:** Nesta fase, após a compra dos ativos os FIDCs realizam a estruturação de acordo com a forma em que esses ativos serão recuperados. Com as informações provenientes da base de dados os FIDCs realizam diligências e em seguida a cobrança desses ativos utilizando atividades como prestar informações aos órgãos de restrição de crédito sobre o devedor, realização de ações de cobrança por canais telefônicos ou através de envio de cartas, chegando até em alguns casos a cobrança via esfera judicial.

- ***Mark-To-Market* (MtM) ou Marcação à Mercado:** Nesta última fase ou após o processo de operacionalização do portfólio de créditos a serem cobrados é realizada a avaliação do desempenho da recuperação *vis-à-vis* a precificação do ativo na fase de *Pricing* para determinação do preço justo do ativo após um determinado período de

tempo. Essa atividade é denominada *Mark-To-Market* (MtM) ou Marcação à Mercado (ALLEN; CARLETTI, 2008 & HEATON; LUCAS, 2010). Esse desempenho leva em consideração aspectos de eficiência operacional na recuperação, questões ligadas à liquidação do portfólio, e verificação das estimativas de recuperação. Com essas informações são elaboradas estratégias de recuperação, e posterior confronto das estimativas com o que foi realizado de fato para ajustes na precificação de ativos de mesma natureza. O MtM também avalia e atualiza a variação do ativo ao longo tempo tornando o processo de marcar o valor ao mercado menos opaco, isto é, fundamentando e justificando valorizações ou desvalorizações do portfólio de créditos.

1.1 *NON-PERFORMING LOANS* (NPLs)

A definição de *Non-Performing Loans* (NPLs) ou Créditos Não-Performados é baseada na provisão de créditos inadimplentes acima do limite de 90 dias e abrangem pessoas jurídicas e também pessoas físicas (CORTAVARRIA, 2000).

Os motivos que causam a inadimplência, e por consequência os NPLs são explicados diante de quatro hipóteses: (i) eventos externos que podem causar o aumento dos créditos inadimplidos; (ii) empréstimos sem bons critérios de avaliação e ausência de monitoramento dos empréstimos atrasados; (iii) os bancos podem atingir baixos custos por meio de subutilização da subscrição de empréstimos e monitoramento no curto prazo e ao longo do tempo isso resulta em aumento dos créditos problemáticos; e (iv) tendência a aceitação de riscos devido ao fato de que os custos não serão incorridos sobre a parte que está aceitando correr o risco (BERGER; DEYOUNG, 1997).

A partir da aquisição dessas dívidas por parte dos Fundos de Investimento em Direitos Creditórios (FIDCs) devem ser definidas estratégias para a sua recuperação. Um modelo de estratégia de recuperação e estruturação de um portfólio de NPLs pode ser encontrado em Vogiazas e Nikolaidou (2011).

O problema dos NPLs é exposto de acordo com o Índice de Créditos Não-Performados que é um dos indicadores chave para atestar a qualidade, o grau de risco e a solvência dos bancos. Se uma instituição bancária tem um índice elevado é uma indicação que o portfólio de crédito está deteriorado, isto é, se o portfólio de crédito sofreu decomposição do seu potencial de pagamento e aferição de lucro (HERRERIAS, MORENO; 2011).

Outro estudo que aborda a constituição de créditos não performados está relativo a questões de governança bancária na qual os ativos relativos a empréstimos bancários ocorrem em perdas bancárias devido ao *insider lending*, que são os empréstimos de alto valor concedidos para diretores, e executivos das próprias instituições bancárias. (Bonin, Hasan e Wachtel; 2008).

Um trabalho sobre a cointegração de métodos de avaliação desses portfólios de créditos inadimplidos é exposto por Toledo (2013), no qual o panorama no aspecto macro econômico dentro do mercado brasileiro é tratado, e também é proposto um modelo de precificação relativo às dinâmicas do mercado brasileiro em NPL.

No estudo recente do European Central Bank (2013) foi constatado que as dinâmicas dos NPLs obedecem a determinantes empíricos como crescimento do Produto Interno Bruto (PIB), índices de preços, taxas de cambio, e a taxa de juros sobre os empréstimos.

O estudo em desenvolvimento pelo International Monetary Fund (2013) teve como principal objetivo o estudo dos NPLs em países de diferentes regiões da Europa, e revelou que os motivos do aumento dos NPLs devem-se a condições macroeconômicas e fatores específicos dos bancos como má administração, custos de desempenho, risco moral e excesso de empréstimos.

1.2 FUNDOS DE INVESTIMENTO EM DIREITOS CREDITÓRIOS E ANÁLISE DE PORTFÓLIOS

De acordo com a Comissão de Valores Mobiliários (CVM) a constituição de um Fundo de Investimento em Direitos Creditórios (FIDC) caracteriza-se pela política de investimento em aplicações sobre o patrimônio líquido na aquisição de créditos vencidos e pendentes de pagamento na ocasião de sua cessão ao fundo (CVM, 2006).

A estrutura para a composição desses fundos se dá na forma em que os mesmos FIDCs participam de leilões de instituições financeiras (em geral bancos, e demais instituições financeiras) que são denominados cedentes, os quais vendem essas dívidas vencidas; e os FIDCs denominados cessionários através de operações de cobrança tentam recuperar esses créditos.

Um dos métodos de análise de portfólio foi concebido por Herry Markowitz na década de 50 e tinha como objetivo realizar a seleção de ativos disponíveis listados na bolsa de valores

através de inferências matemáticas, em que o principal objetivo era a diversificação e posterior alocação de ativos a fim de minimizar o risco (MARKOWITZ, 1953).

Na década de 1960, Sharpe (1963) realizaria uma revisão de literatura sobre a teoria inicial de Markowitz de alocação de ativos e considerou que os critérios relativos à escolha desses ativos teriam uma etapa de análises relativas a) à capacidade de securitização desses ativos, b) determinação de um conjunto de carteiras de investimentos através de estimativas e c) seleção de aplicações em portfólios de acordo com as preferências do investidor (SHARPE, 1963).

Mesmo que as duas abordagens anteriormente citadas tenham o seu desenvolvimento original no mercado de ações, a análise de FIDCs pode obedecer a essa mesma lógica considerando as suas especificidades, e mais que isso essa análise deve ser capaz de responder perguntas como a) qual o potencial de recuperação dentro desses créditos vencidos? b) qual é a curva de recuperação necessária para liquidar esse portfólio e d) qual é o preço justo considerando as perguntas anteriores?

1.3 INTELIGÊNCIA COMPUTACIONAL

Como apresentado por Duch (2007) em um trabalho de revisão de literatura a Inteligência Computacional é definida como uma ramificação da ciência da computação que estuda problemas para o qual não há algoritmos computacionais efetivos.

No trabalho de Engelbrecht (2007) a Inteligência Computacional é definida como uma sub-ramificação da Inteligência Artificial que é responsável pelo estudo de mecanismos adaptativos para habilitar ou facilitar o comportamento inteligente em ambientes complexos e de mudanças constantes. Esses mecanismos incluem paradigmas computacionais que possuem a habilidade de aprender ou adaptar-se em novas situações, para generalizar, abstrair, descobrir e associar.

Em Bezdek (1994) um sistema só é computacionalmente inteligente quando lida apenas com dados numéricos, tem componentes de reconhecimento de padrões, não utiliza conhecimento dentro da definição de inteligência artificial; e adicionalmente quando exibe (i) adaptabilidade computacional, (ii) tolerância à falha computacional, (iii) tem rapidez na abordagem de algum tipo de reviravolta como os seres humanos, e (iv) detêm taxas de erros que se aproximam do desempenho humano.

A Inteligência Computacional também pode ser definida como o estudo de agentes inteligentes, em que o agente inteligente age de acordo com as circunstâncias para atingir um

determinado objetivo, é flexível para a mudança de ambientes e objetivos, aprende através da experiência e realiza escolhas apropriadas de acordo com as limitações da computação (MACKWORTH, GOEBEL, POOLE; 1998).

Em contraste com essas abordagens que colocam a Inteligência Computacional em um nível de abstração mais alto, no trabalho de Craenen e Eiben (2002) os autores realizam a distinção de Inteligência Artificial (IA) e Inteligência Computacional (IC).

Os autores afirmam que a IA lida com a representação simbólica do conhecimento, enquanto a IC lida com a representação numérica da informação; a IA preocupa-se em si mesma com a questão das funções cognitivas de alto-nível, enquanto a IC se preocupa com funções cognitivas de baixo nível.

Além do mais, conforme afirmam os autores, a IA analisa a estrutura de um determinado problema e tenta construir um sistema inteligente baseado nesta estrutura, isto é, operando de maneira *top-down*; enquanto espera-se a estrutura emerja de uma forma não ordenada da IC, isto é, uma abordagem *bottom-up.*

Na literatura não há ponto pacífico no que tange as abordagens e/ou técnicas que definam a Inteligência Computacional de forma clara.

No geral consideram-se abordagens bio-inspiradas (Algoritmos Genéticos (Evolucionária), Enxame de Partículas (Inteligência Coletiva), Redes Neurais Artificiais (Neurológica) e também abordagens que estão ligadas a problemas de incerteza/probabilísticos (Probabilidade Bayesiana), métodos de kernel (Máquinas de Vetor de Suporte), derivações dos métodos lineares (regressão, GLM), métodos adaptativos (*Multi Adaptative Regression Splines*), aprendizado por reforço, heurísticas e metaheurísticas (*Simulated Annealing*, Algoritmos de Busca) além das técnicas aplicadas em Mineração de Dados (*Self-Oganizing Maps*, Teoria dos *Rough Sets*, Regras de Associação) (Engelbrecht, 2007).

Embora haja divergências em questões relativas à abordagem da Inteligência Computacional, este trabalho por questões metodológicas adotou o conceito asseverado por Craenen e Eiben (2002), isto é, a IC será realizada através da construção de estruturas para a resolução do problema de forma não ordenada.

1.4 JUSTIFICATIVA E MOTIVAÇÃO

Com as sucessivas reduções na taxa de juros promovidas pelo Comitê de Política Monetária (BANCO CENTRAL DO BRASIL, 2012), decorrente da estratégia governamental de incentivo ao consumo; bem como os estímulos ao setor produtivo e comércio, ampliou-se a oferta de crédito por parte das instituições financeiras.

Tendo em vista o cenário macroeconômico que vinha beneficiando o Brasil nos últimos anos, tornou-se uma crescente a democratização do crédito, seja na modalidade para pessoas físicas nas formas de Crédito Direto ao Consumidor (CDC), CRÉDITO IMOBILIÁRIO, crédito para aquisição de veículos; para pessoas jurídicas nas modalidades crédito para formação de fluxo de caixa, empréstimos para expansão de negócios, renegociação de dívidas; ou na modalidade crédito direcionado que são créditos com juros subsidiados para fins de investimentos dentro de uma estratégia de expansão de crédito pelo governo.

Junto com a expansão do crédito, também cresceu a inadimplência desses setores; e com este fato o mercado de transferência de direitos creditórios tornou-se uma alternativa para essas instituições financeiras para recuperar parte dos recebíveis que devido à regulação do Banco Central devem ser movidos de forma compulsória para perdas após 180 dias de atraso (BANCO CENTRAL DO BRASIL, 1999; 2000).

No entanto, por se tratar de ativos não estruturados e que obedecem à regulação específica da Comissão de Valores imobiliários (CVM) este trabalho servirá como bússola para ações que possam potencialmente serem adotadas em sua totalidade ou parcialmente; sejam de forma direta ou em conjunto com outras ações. Isso faz que os resultados sejam generalizados para a tomada de decisões gerais sobre esses ativos e elaboração de estratégias, e não somente para o que a técnica por ventura pode recomendar.

Dessa forma o presente estudo é motivado pela oportunidade de colocar em intersecção a análise dos portfólios FIDC, em especial aqueles que realizam a recuperação de créditos do tipo NPL, e a Inteligência Computacional, pois existe uma lacuna encontrada na literatura sobre a utilização dessas técnicas para análise dos NPLs, em especial de abordagens que trabalhem diretamente com bases de dados com créditos já inadimplidos. Estas abordagens da literatura geralmente trabalham com dados consolidados de agregados econômicos ou de instituições específicas.

A importância dessa natureza de análise utilizando Inteligência Computacional possibilita que decisões econômicas e financeiras relativas à análise de NPLs sejam tomadas com um espectro de informações mais amplo, em que há diversas possibilidades como: de

análise de padrões em bases de dados de créditos não estruturados; identificação de anomalias, classificação de devedores através da criação de escores, regras de associações, agrupamento entre classes de devedores de acordo com a possibilidade de recuperação de crédito, *etc*.

1.5 PROBLEMA DE PESQUISA

Como a aplicação das técnicas de Inteligência Computacional podem auxiliar na análise e recuperação de créditos do tipo *Non-Performing Loans*?

1.6 OBJETIVOS

1.6.1 Objetivo Geral

O objetivo deste trabalho foi a aplicação de técnicas de Inteligência Computacional na análise e recuperação de portfólios de créditos do tipo *Non-Performing Loans*.

1.6.2 Objetivos Específicos

Os objetivos específicos são, portanto, voltados para:

- Realização dos três experimentos utilizando as seguintes técnicas de Inteligência Computacional: Redes Neurais Artificiais, Teoria dos *Rough Sets* e Árvores de Decisão, *Self-Organizing Maps*, aplicadas de forma isolada ou de forma combinada;
- Especificar os aspectos determinantes de recuperação, isto é, fatores específicos que auxiliem na reestruturação ou que caracterizem um NPLs com potencial de pagamento dos devedores;
- Construção de modelos usando técnicas de Inteligência Computacional em algumas combinações para atividades de classificação, análise de *cluster*, e geração de regras de decisão que deem subsídios para formulação de estratégias de recuperação; e
- Através desse problema dos NPLs realizar a comparação de modelos seja utilizando abordagem sensível ao custo dada a natureza do estudo que trata de

uma questão de fundo econômica, ou na comparação entre técnicas em termos de acurácia e custo computacional de acordo com a base de dados utilizada.

1.7 HIPÓTESES DE PESQUISA

O presente trabalho busca validar as seguintes hipóteses ao problema de pesquisa:

H1) A aplicação de Redes Neurais Artificiais (RNA) para a classificação e posterior criação de modelos preditivos de recuperação crédito do tipo NPL se submete não somente em sua acurácia, mas sim na estrutura de custos envolvida em uma operação de recuperação de crédito, isto é, os Erros do Tipo I (Falsos Positivos) e Erro do Tipo II (Falsos Negativos) influenciam diretamente no modelo de recuperação;

H2) A combinação das técnicas da Teoria dos Conjuntos Aproximados (*Rough Sets Theory*, RST) juntamente com as redes neurais do tipo Mapas Auto-Organizáveis (*Self-Organizing Maps*, SOM) pode auxiliar na construção de modelos de recuperação, com a primeira técnica aplicada na extração de regras; e a segunda na criação de modelos descritivos (através da análise de *cluster*); e

H3) O número de atributos tem influência direta no modelo de recuperação, contudo utilizando-se a abordagem de *Rough Sets* para pré-processamento e posterior classificação via Árvore de Decisão há um grau de especificidade maior que apresenta de forma mais precisa os determinantes na recuperação de crédito.

1.8 DELIMITAÇÕES DO ESTUDO

a) Foram utilizados nos experimentos bases de dados compostas de registros de créditos cedidos por bancos de médio e grande porte. A população constante na base (*i.e.* os clientes) não será identificada devido a questões de (i) ordem legal relativas ao sistema bancário e, (ii) ao direito da confidencialidade prevista na Constituição Brasileira de 1988 (CF/88);

b) A abrangência da pesquisa será limitada a um grupo amostral seleto, tal qual pode não permitir uma generalização para todos os ativos financeiros dessa natureza do mercado nacional, mas sim observar o comportamento específico de *tranches* (*i.e.* safras de dívidas ou fatias do portfólio de crédito) analisadas pelos cedentes; *tranches* estas que poderão conter algum tipo de viés amostral anterior à análise que está sendo procedida devido ao fato de sua

seleção não obedecer os critérios do pesquisador mas sim interesses exclusivos do(s) cedente(s). Contudo a extração dessas dívidas foi realizada de forma aleatória buscando apenas manter a proporção de créditos pagos e não pagos da base completa;

c) Na mesma seleção das dívidas, buscou-se escolher apenas ativos financeiros ligados a pessoas físicas; como, por exemplo, crédito direto ao consumidor (CDC) como cartão de crédito (CC), crédito pessoal (CP), ou empréstimos com débito em conta bancária. Isso significa que outros produtos creditícios que obedecem outras dinâmicas de recuperação não foram tratados neste trabalho;

d) O trabalho manterá o foco apenas no que concerne à aplicação das técnicas nas bases de dados cedidas, com a abrangência somente nos ativos creditícios sob análise. Isto significa que aspectos conjeturais podem ser levados em consideração, contudo somente de maneira secundária;

e) Este trabalho não contempla análises conjecturais a despeito do momento econômico seja relativo a fatores macroeconômicos ou microeconômicos no momento da elaboração dos resultados. Este trabalho está orientado a realizar análises descritivas e/ou preditivas baseado somente nos dados apresentados para avaliação das técnicas empregadas, que por ventura podem levar a *insights* (*i.e.* processo inicial de dedução ou indução intuitiva), entretanto, este não é o foco do trabalho.

1.9 ORGANIZAÇÃO DO TRABALHO

Este trabalho está estruturado em cinco capítulos. Além da introdução, o trabalho compõe-se das seguintes partes:

Capítulo 2 - Fundamentação Teórica. Neste capítulo é apresentado o referencial teórico através de revisão bibliográfica dos principais elementos do trabalho que são os créditos do tipo *Non-Performing Loans* no período de antes e após a crise financeira, a estrutura dos Fundos de Investimento em Direitos Creditórios, e aspectos teóricos relativos à Inteligência Computacional.

Capítulo 3 - Materiais e Métodos. Neste capítulo é realizada uma descrição detalhada da caracterização metodológica do trabalho, além de expor também a organização da revisão de literatura que contou com técnicas provenientes revisão sistemática da bibliometria. Neste capítulo também foram expostos o planejamento amostral, a metodologia experimental e o fluxo no qual os experimentos foram conduzidos.

Capítulo 4 - Realização dos Experimentos e Discussão dos Resultados. Neste capítulo estão organizados, apresentados e discutidos os resultados dos três experimentos realizados.

Capítulo 5 - Conclusões. Neste capítulo são apresentadas as conclusões do trabalho, além da sumarização das principais contribuições e perspectivas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é exposta a literatura que trata dos *Non-Performing Loans*, Fundos de Investimentos em Direitos Creditórios e Inteligência Computacional para fins de fundamentação dos pressupostos teóricos envolvidos nesta pesquisa.

O quadro teórico através da revisão de literatura deu-se em três aspectos:

- O primeiro aspecto são os *Non-Performing Loans* e a sua contextualização no aspecto bancário e a importância do seu entendimento em face de dois momentos da economia mundial que é o antes e o depois da crise de 2007/2008;

- O segundo aspecto a ser fundamentado são os Fundos de Investimento em Direitos Creditórios (FIDCs) e como esta estrutura financeira realiza a estruturação desses ativos para aferição de lucros; e

- O terceiro e último aspecto a ser abordado é a fundamentação teórica da Inteligência Computacional e as técnicas que foram utilizadas neste trabalho.

Busca-se com esse referencial teórico a sustentação do objeto de estudo que está sendo apresentado.

2.1 *NON-PERFORMING LOANS* (CRÉDITOS NÃO-PERFORMADOS)

Créditos do tipo *Non-Performing Loans* (NPL), ou créditos não-performados, são caracterizados como créditos que estão vencidos e sem pagamentos a mais de 90 dias (CORTAVARRIA, 2000).

A definição de NPL pode ser baseada na provisão de créditos inadimplentes acima do limite de 90 dias e abrange tanto pessoas jurídicas e pessoas físicas (INTERNATIONAL MONETARY FUND, 2013).

A classificação em termos de limites de dias para caracterização de crédito não-performados é apresentada pelo International Monetary Fund (2013), em que há uma variação em termos de dias de acordo com cada país e a sua regulação bancária. Contudo, de maneira geral considera-se o crédito inadimplido como NPL quando o mesmo ultrapassa 90 dias após a data do último vencimento.

A importância dos NPLs dentro do contexto brasileiro se dá na forma em que as operações a crédito tem crescido de forma expressiva, em especial ao longo dos últimos 11 anos como mostra a Figura 1.

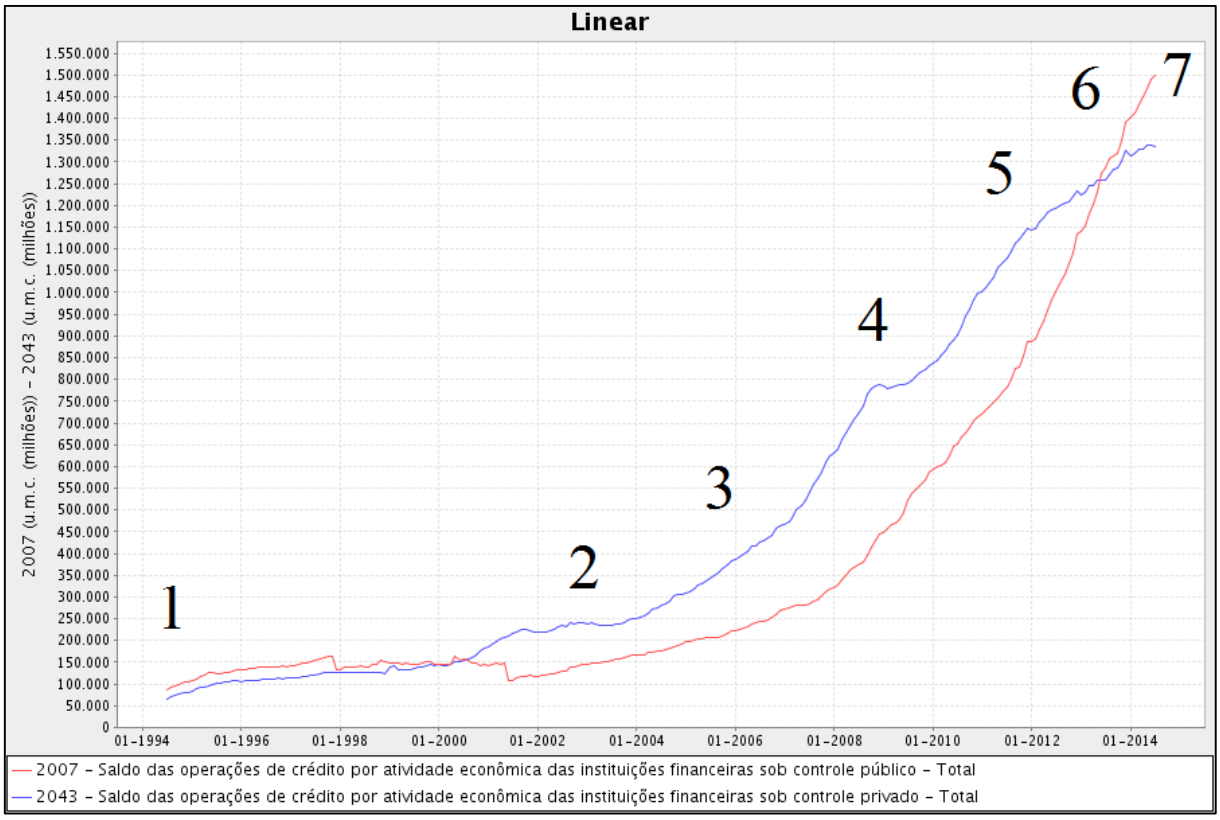


Figura 1 - Saldo das Operações de Crédito no Brasil por instituições financeiras sob controle público e privado. Fonte: Sistema Gerenciador de Séries Temporais - https://www3.bcb.gov.br/sgspub/

A Figura 1 representa todas as operações de crédito realizadas no Brasil em que a linha azul representa as operações de crédito realizadas por instituições privadas e a linha vermelha representa as operações realizadas por instituições sob controle público.

No ponto 1 do gráfico é o início da série que ocorreu com a normalização da base monetária ocorrida pelo início do Plano Real em 1994. Após alguns anos de estacionareidade da série e o descolamento em meados de 2001, no momento 2 com o início de políticas de crescimento econômico baseadas no incentivo ao crédito tanto instituições públicas e privadas começam a ter uma aceleração no ritmo de empréstimos e concessão de crédito. No momento 3, há uma forte aceleração na concessão de crédito e ambas as curvas apresentam o formato exponencial.

No momento 4, com a crise financeira mundial, o setor privado que concentrava um maior portfólio de crédito até o momento apresentou uma retração. Em contrapartida o setor público foi insensível a este evento e mesmo assim continuou com o mesmo ritmo de concessão de empréstimos. Em meados de 2011 no ponto 5, o setor privado iniciou uma desaceleração do ritmo de empréstimos, causando um arrefecimento da oferta de crédito. Em 2013 no momento 6 há uma mudança em que o setor público torna-se responsável pela maior parte da carteira de

crédito do sistema nacional. Em outras palavras o setor público é majoritariamente responsável pela maioria de empréstimos, financiamentos, *etc*.

Já no momento 7 mesmo com o setor privado em sua tendência de arrefecimento na concessão de empréstimos o setor público ainda prossegue em aceleração da atividade de concessão de empréstimos.

Observando o gráfico, pode-se dizer que 2008 a 2014 foi colocado na economia cerca de R$ 1 trilhão pelos bancos públicos, estes que por sua vez podem não atender demandas genuínas de mercado, ou mesmo podem estar submetidas a demandas políticas ou eleitorais.

Um problema que mostra a importância desses ativos é exposto de acordo com o Índice de NPLs que é um dos indicadores chave para atestar a qualidade, o grau de risco e a solvência dos bancos. Se uma instituição bancária tem um índice de NPLs elevado é uma indicação que o portfólio de crédito está deteriorado, isto é, com perda do seu potencial de aferição de lucros e com perspectivas de prejuízos (HERRERIAS; MORENO, 2011).

Na perspectiva da exposição bancária do cenário brasileiro, a Resolução nº2682 do Banco Central do Brasil estabeleceu os critérios de classificação para as operações de crédito e também nas regras de constituição de provisão para créditos de liquidação duvidosa ou Provisão de Devedores Duvidosos (PDD).

Todas as instituições financeiras reguladas obedecem a essa forma de classificação em que são levados em consideração a) aspectos do devedor e seus garantidores (*e.g.* situação econômica, grau de endividamento e capacidade de geração de resultados, *etc*.) e b) operação de crédito (*e.g.* natureza e finalidade da operação, características das garantias, e valor) (BANCO CENTRAL DO BRASIL, 1999).

A PDD é uma métrica que pode ser utilizada para avaliação do grau de solvência da instituição bancária; e também, como forma de avaliar se o portfólio de crédito está deteriorado.

A forma de constituição da PDD tem a denominação de risco de forma crescente com utilização das letras AA, A, B, C, D, E, F, G, H; cuja classificação obedece aos seguintes critérios em função do atraso do pagamento:

a) atraso entre 15 e 30 dias: risco nível B, no mínimo;
b) atraso entre 31 e 60 dias: risco nível C, no mínimo;
c) atraso entre 61 e 90 dias: risco nível D, no mínimo;
d) atraso entre 91 e 120 dias: risco nível E, no mínimo;
e) atraso entre 121 e 150 dias: risco nível F, no mínimo; e
f) atraso entre 151 e 180 dias: risco nível G, no mínimo.

Cabe ressaltar que o risco AA e A representam o mais alto grau de confiabilidade para pagamento, em que AA é o crédito em dia e A crédito com data de vencimento até 14 dias. Já o risco H representa o crédito vencido a mais de 180 dias e que pode ser movido para perdas bancárias.

O crédito risco H no Brasil tem crescido nos últimos anos, como mostram a Figura 2 e Figura 3.

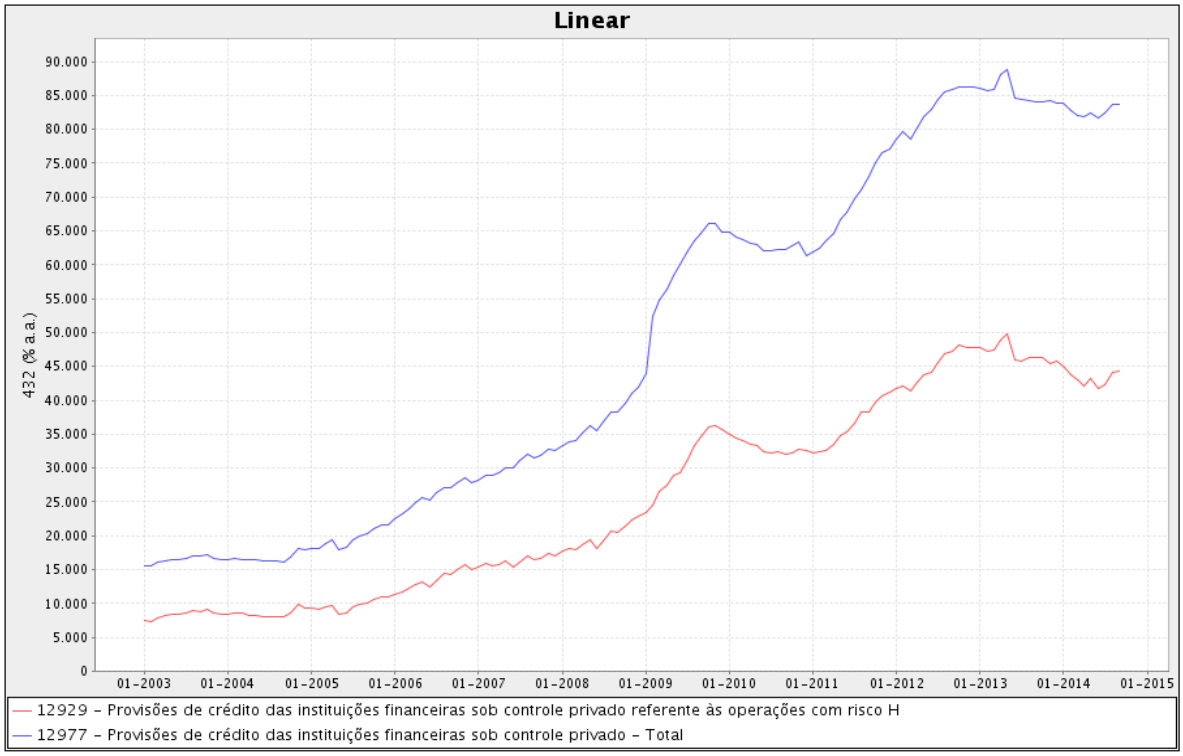


Figura 2 - Crédito classificado como risco H em instituições privadas. Fonte: Sistema Gerenciador de Séries Temporais - https://www3.bcb.gov.br/sgspub/

De acordo com a Figura 2 pode ser verificado que as instituições sob controle privado tiveram um grande aumento de crédito risco H após a crise financeira de 2007/08 e que esses patamares permaneceram altos mesmo com o final da crise.

Na Figura 3 é apresentada a carteira de crédito risco H das instituições públicas desde 2003 que foi o início da série.

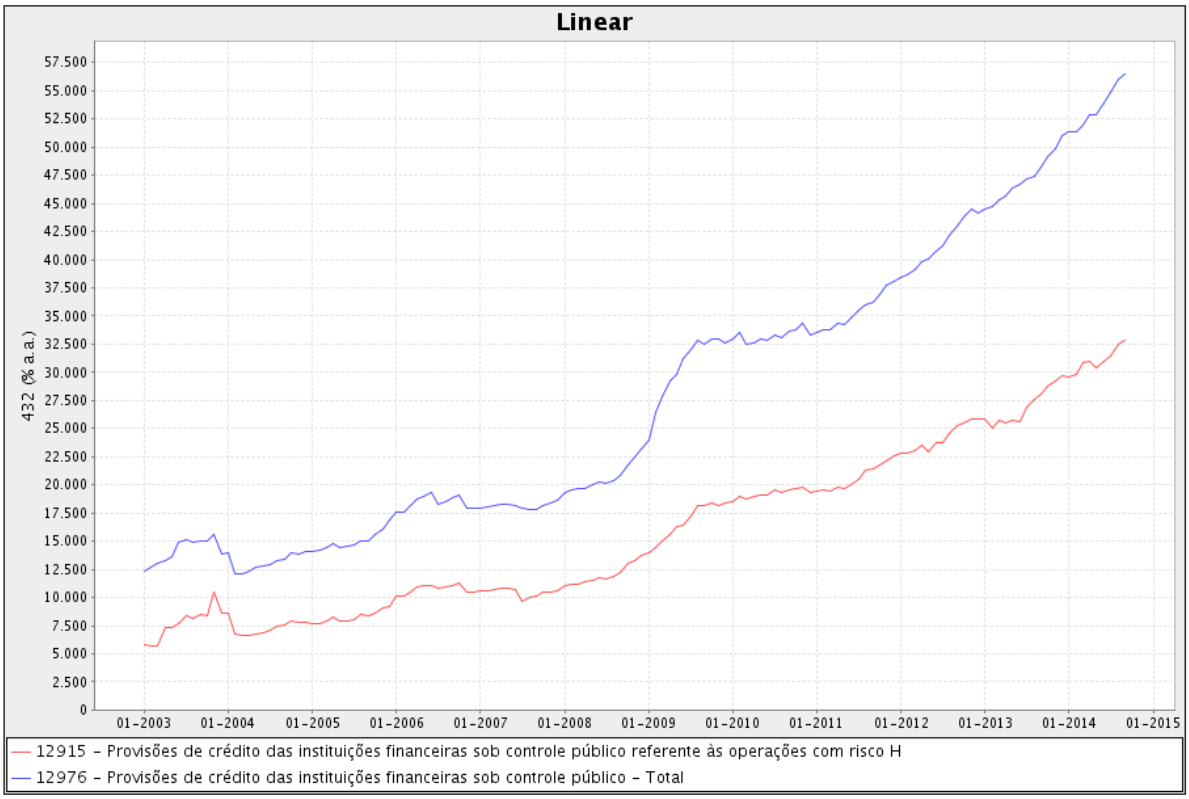


Figura 3 - Crédito classificado como risco H em instituições públicas. Fonte: Sistema Gerenciador de Séries Temporais - https://www3.bcb.gov.br/sgspub/

Na Figura 3 fica evidente que depois de 2008 houve um crescimento relativo ao portfólio de crédito em risco H, no entanto com uma aceleração um pouco menor do que no setor privado.

O artigo de Berger e DeYoung (1997) atribuiu o NPL a quatro hipóteses que são ligadas ao aspecto infrabancário, ou seja, aspectos estritamente internos dos bancos. Entretanto, o artigo não mostra uma perspectiva quantitativa na forma de minimizar esses créditos. Essas hipóteses levantadas pelos autores são (i) má sorte da instituição: eventos exógenos que podem causar o aumento dos créditos inadimplidos; (ii) má administração: empréstimos sem bons critérios de avaliação e ausência de monitoramento dos empréstimos atrasados; (iii) descuido: os bancos podem atingir baixos custos através de subutilização de subscrição de empréstimos e monitoramento no curto prazo e ao longo do tempo isso resulta em aumento dos créditos problemáticos; e (iv) risco moral: tendência à aceitação de riscos devido ao fato dos custos não serão incorridos sobre a parte que está aceitando correr o risco.

Na perspectiva do mercado brasileiro de NPLs, a dissertação de Toledo (2013) realizou uma abordagem inédita em expor aspectos operacionais e de custos na reestruturação dos NPLs, e também realiza análises de créditos de forma consolidada, isto é, trata da agregação desses créditos como um portfólio propriamente dito.

De acordo com o trabalho de Nkusu (2011) a literatura que trata de NPL pode ser dividida em três perspectivas; (i) trabalhos que têm como finalidade explicar o papel do desempenho macroeconômico sobre as instituições de crédito em relação aos NPLs; (ii) análise do relacionamento entre NPL e condições financeiras e macroeconômica mostrando o impacto positivo dos NPLs na probabilidade de crise e subsequentemente o papel chave dos NPLs na predição de crises bancárias, e (iii) trabalhos com a abordagem orientada a explicar ou realizar análises preditivas dos NPLs em nível macroeconômico, como, por exemplo, analisando os índices de NPLs agregados.

A análise e discussão sobre os NPLs neste trabalho foi realizada partindo de duas perspectivas: a primeira trata dos NPLs no momento pré-crise, isto é, o que estava discutido antes de 2007; e a literatura após a crise de 2007/2008. Essa divisão faz-se necessária devido ao fato de que os NPLs foram um dos determinantes da crise financeira mundial, em especial os créditos imobiliários e a sua cadeia de securitização.

2.1.1 Pré-Crise – Antes de 2007

Em um dos primeiros artigos em que aparecem os NPLs foi realizado um estudo sobre maturidade da estrutura de ativos em relação aos NPLs em que foi proposto o seguinte modelo de Meeker e Gray (1987) que é representado pela Equação 1:

$$\frac{WCL}{TL} = a + \beta_1 \frac{(PD89)}{TL} + \beta_2 \frac{(PD90)}{TL} + \beta_3 \frac{(N)}{TL} + \beta_4 \frac{(R)}{TL} + \varepsilon \quad (1)$$

No qual:

WCL (Weighted Capital Loss) é o peso da perda de capital, que em termos práticos é a diferença entre o valor de face do ativo e o valor presente;
TL (Total Loans) é o total de empréstimos;
a é uma constante;
β_n são os coeficientes regressores;
PD89 (Past Due) são empréstimos vencidos entre 30 e 90 dias;
PD90 (Past Due) são empréstimos com mais de 90 dias de atraso;
N (Non-Accrual Loans) são os empréstimos de provisão duvidosa, isto é com menos 90 dias de atraso;
R (*Renegotiated Loans*) são os créditos renegociados; e
ε é o termo de erro que não é explicado pelos coeficientes regressores.

O artigo de Berger e DeYoung (1997) examina a intersecção entre o problema dos empréstimos e a eficiência bancária na literatura. Os autores levantam quatro hipóteses das origens dos problemas que são: Má Sorte: problemas externos ao banco que por ventura venham aumentar o problema dos empréstimos no banco; Má Administração: más práticas operacionais como baixa eficiência na mensuração de custos e falta de qualificação para escoragem do crédito, superestimação da garantia exposta ao tomador do crédito, falta de monitoração dos empréstimos; Limitações Operacionais: quantidade de recursos alocados para monitoração direta dos empréstimos, em que os custos operacionais não são levados em conta no momento da concessão do empréstimo quando no longo prazo o portfólio de crédito pode ter problemas de performance; e Risco Moral: Excesso de tomada de risco por quem não é parte direta nos prejuízos, no caso, gerentes que ganham comissões sobre empréstimos concedidos não tem estímulos diretos para uma avaliação mais apurada devido ao fato de que eles não fazem parte do prejuízo.

O estudo de Ferguson (2007) mostrou os NPLs e a sua importância no cenário de securitização, em que segundo o autor, não aconteceram na Rússia conforme o esperado; e coloca também os NPLs como ativos que podem ser securitizados e comercializados em pacotes de diversos ativos consolidados chamados *Asset-Baked (i.e.* modalidade de FIDC), em que esses passam de um estado de baixa liquidez para serem convertidos em ativos comercializáveis a investidores através de margens de acordo com diferentes estruturas de riscos.

Em Fofack (2005) o autor realiza um estudo dos NPLs no contexto da África Saariana em meados dos anos 90. Os resultados indicam que uma economia estabilizada em termos macroeconômicos e o crescimento da economia são fatores que são considerados no aumento dos NPLs.

No trabalho de Rajan e Dhal (2003) foi realizada uma análise empírica sob o prisma de três aspectos majoritários que são crédito, preferência e exposição a riscos de acordo com o tamanho do banco e choques macroeconômicos usando modelos de regressão. Os autores chegaram aos resultados cujas variáveis ligadas ao crédito em si exercem influência direta nos índices de NPLs. Outro resultado relevante é que o tamanho das instituições é um fator que determina a sua disposição em expor-se ao risco e por fim, os autores verificaram que uma alta taxa de juros tem relação direta com o aumento dos NPLs.

Em Alton e Hazen (2001) os autores identificaram que no momento em que a atividade econômica tem o seu ritmo desacelerado, esta exerce influência negativa na capacidade dos

clientes de realização do pagamento de seus empréstimos; especialmente aqueles que possuem débitos em bancos de pequeno porte.

Ainda no estudo de Meeker e Gray (1987) foi realizada uma análise de regressão em relação aos dados de NPL no sentido de avaliar a informação que até então era inédita ao grande público, no qual os resultados mostraram que há uma ligação entre o volume de NPLs e a qualidade dos ativos do sistema bancário.

O que pode ser visto nessa parte da literatura pré-crise é que há uma predominância do estudo dos NPLs de acordo com a perspectiva intrabancária, em outras palavras, havia uma preocupação com os efeitos imediatos de más práticas bancárias das mais diversas naturezas.

A seguir, será discutida a literatura que trata dos NPLs no momento da crise de 2007/2008 e no momento pós-crise até os dias atuais.

2.1.2 Crise e Pós-Crise – Após 2007

Na literatura que trata dos NPLs nos momentos de crise e pós-crise temos os trabalhos que ressaltam os aspectos infrabancários (administração, risco moral, *etc*.) e aspectos macroeconômicos; porém, sem nenhuma abordagem direta que trata o aspecto microeconômico na reestruturação do crédito.

Em Makri, Tsagkanos e Bellas (2014) há o estudo sobre os fatores macroeconômicos que influenciam o volume de NPLs na Zona do Euro no período pré-crise no qual os autores encontraram relação em aspectos específicos do setor bancário para o aumento dos NPLs.

No trabalho de Zaib, Farid e Khan (2014) os autores realizaram um estudo em oito bancos paquistaneses no período entre 2003-2011, e o estudo indicou que a disposição das instituições financeiras em tomar riscos e o crescimento do PIB são aspectos relacionados aos NPLs. O trabalho indicou que as estratégias e fatores administrativos também influenciam no aumento ou redução dos NPLs.

Em IMF (2013) foram investigadas as principais causas na inadimplência de crédito no sistema bancário da Bulgária em que o ponto de partida da pesquisa foi baseado na hipótese que indicadores ligados a ciclos macroeconômicos e agregados financeiros do setor bancário tinham papel importante na constituição dos NPLs. Os resultados mostraram que os indicadores macroeconômicos como a taxa de desemprego, índice de construção civil, produção industrial juntamente com crescimento da atividade creditícia e o cenário de crise influenciam na qualidade dos ativos.

No estudo de Jordan e Tucker (2013) foi aplicada a técnica de regressão para a análise dos dados e verificou-se que aumento da atividade econômica está relacionado com a diminuição dos NPLs assim como o crescimento do emprego, condições de negócios e de pagamentos.

Em Škarica (2013) foram apresentados resultados de que o crescimento no índice de NPLs gera o consequente crescimento da inflação; o que indica que o aumento da taxa de juros devido à inflação exerce influência nos NPLs. Este trabalho indicou que políticas macroeconômicas de expansão do crédito para estimular o crescimento proveniente de autoridades monetárias como o banco central podem ter exercido influência no crescimento dos NPLs, no entanto, isso não fica exposto de forma direta no artigo.

A abordagem de Vazquez, Tabak e Souto (2012) trás o desenvolvimento de um modelo em *stress-test*, isto é, a simulação na qual o modelo incorpora uma crise econômica, que através das análises de empréstimos de pessoas físicas e pessoas jurídicas para analisar o risco de crédito e o possível aumento dos NPLs.

Em Herrerias e Moreno (2012) é apresentado um estudo que faz a mensuração do risco de crédito de acordo com o índice de NPLs no sistema bancário mexicano. A abordagem verificou cada instituição bancária e a sua contribuição ao resto do sistema através da difusão de risco; e chegou-se a margem de 60-75% de variação dependendo da instituição bancária.

O trabalho publicado pelo IFC (2012) mostra uma abordagem em que há alguns pontos de atenção na construção de uma estrutura específica para reestruturação de NPLs, no qual leva em conta a análise de ambiente externo e aspectos microeconômicos/operacionais para a reestruturação desses ativos.

Na pesquisa de Vogiazas e Nikolaidou (2011) os autores investigaram os determinantes de crédito no sistema bancário da Bulgária através da técnica de análise de series temporais com agregados macroeconômicos e aspectos específicos do sistema bancário. Os resultados indicam que variáveis ligadas a atividade industrial juntamente com a expansão do crédito influenciaram na qualidade dos ativos dos bancos búlgaros.

No trabalho de Espinoza e Prasad (2010) foi realizada uma análise em 80 bancos e foi constatado que o índice de NPLs piorou quando caíram respectivamente o crescimento econômico e a taxa de juros. No entanto, de acordo com os autores isso aconteceu devido a fatores específicos ligados a aspectos infrabancários.

Em Greenidge e Grosvenor (2010) os autores argumentam que a magnitude dos NPLs e o seu impacto na economia é um elemento chave na iniciação e na progressão de crises bancárias.

No trabalho de Kalluci e Kodra (2010) foi analisado que o índice de NPLs no sistema bancário na Albânia pode ser separado em duas subcategorias, empréstimos a empresas e empréstimos ao consumidor final. Essa caracterização mostrou as diferentes segmentações no portfólio de crédito do país em questão, mas nada que lançasse luz à reestruturação dos NPLs.

No trabalho de Louzis, Vouldis e Metaxas (2010) os autores realizam o estudo do setor bancário grego em que foi realizada uma análise sobre três tipos de NPL: crédito direto ao consumidor, crédito para empresas, e crédito imobiliário. Os autores partiram da hipótese que aspectos macroeconômicos e aspectos infrabancários influenciam no aumento dos NPLs.

O estudo econométrico de Khemraj e Pasha (2009) teve como objetivo saber as causas de NPLs no sistema bancário da Guiana. Os resultados indicam que o crescimento do PIB é inversamente relacionado aos NPLs, e que um cenário de recuperação econômica se traduz em redução dos NPLs. Um fator que foi levantado pelos autores é que os bancos que aplicam taxas de juros altas e que realizam empréstimos excessivamente são mais sujeitos a terem um crescimento dos NPLs em seu portfólio de crédito.

Khemraj, Tarron e Sukrishnalall (2009) usaram modelo baseado em regressão na relação de aspectos macroeconômicos e aspectos infrabancários na Guiana. O modelo apresentou o resultado de que a inflação está negativamente associada ao crescimento dos índices de NPLs.

Em Weissbach e Wilkau (2008) foram desenvolvidos dois modelos de previsão de NPLs, um baseado em uma função de perda que leva em consideração a média de cada empréstimo, e um modelo misto de créditos performados e não-performados. O principal resultado foi que o risco de NPL em um portfólio de crédito depende da volatilidade da atividade econômica e da granularidade do portfólio entre créditos performados e não-performados.

Na abordagem usada no trabalho de Bonin, Hasan e Wachtel (2008) foi constatado que o declínio macroeconômico, juntamente com maus empréstimos, e ausência de uma regulação bancária forte implicam em NPLs. Os autores atestam em seu estudo que o índice de NPLs chegou a 58% em 1998. Os autores também falam a respeito do aspecto da recapitalização desses débitos, contudo de forma vaga.

Explorando a literatura que faz a ligação entre aspectos infrabancários e aspectos macroeconômicos; embora essa abordagem possa nortear pesquisas em aspectos ligados a visões gerais sobre os NPL e suas tendências, os dados são limitados a economias específicas, e não há nenhuma proposta efetiva, pois há muitos fatores exógenos a serem considerados.

Por outro lado, a presente abordagem envolve aspectos puramente operacionais através de características dos dados de acordo com informações provenientes de um sistema de cobrança.

Em Herrerias e Moreno (2011) os autores analisam o risco de crédito de acordo com o índice de NPLs usando o sistema bancário mexicano; e de acordo com os estudos dos autores mostrou-se que o processo de difusão dos riscos de NPL no sistema bancário depende de um determinado período de tempo onde no curto prazo (de 1 até 6 meses) os riscos são da instituição, e no longo prazo 70% do risco de crédito é atribuível ao risco sistêmico.

Em Impavido, Klingen e Sun (2012) os autores estudam os NPLs em relação aos países da Europa nas regiões Central, Leste e Sudeste em que os autores identificam que os NPLs têm papel crucial na saúde econômica através dos mais diversos sistemas bancários, e elencam uma série de questões em aberto no entendimento desses ativos.

No trabalho de Saba, Kouser e Azeem (2012) para analisarem os principais dados macroeconômicos em face de fatores internos, os autores utilizam o método de regressão e chegaram à conclusão que os coeficientes macroeconômicos não foram tão altos, e advertiram que os bancos devem ter melhores controles e políticas de crédito baseadas em boas práticas.

O modelo de pesquisa apresentou a seguinte fórmula proveniente da Equação 2:

$$NPLR = a_0 + \beta_1 TL + \beta_2 IR + \beta_3 GDPPC + \varepsilon \qquad (2)$$

No qual:

NPLR (Non-Performing Loans Rate) é o estimador usado para a taxa de NPLs;
TL (Total Loans) para o total de empréstimos;
IR (Interest Rate) é a taxa de Juros;
GDPPC (Gross Domestic Product per Capita) corresponde ao PIB per capita;
β_n são os coeficientes regressores; e
ε é o termo de erro.

Outro estudo relativo à constituição de créditos não performados está ligado a questões de governança bancária cujos ativos relativos a empréstimos bancários ocorrem em perdas bancárias devido ao *Insider Lending*, isto é, empréstimos realizados sem nenhum tipo de controle ou restrição específica para membros da própria instituição financeira e seus apaniguados (BONIN, HASAN E WACHTEL, 2008).

Weißbach e Wilkau (2013) realizaram estudo sobre o capital econômico empregado em previsões de portfólios de NPLs, e encontraram que o risco desses tipos de portfólios juntamente com os portfólios de créditos adimplidos depende da volatilidade da atividade

econômica sobre a granularidade do portfólio (*i.e.* diversidade de produtos) e sobre o portfólio performado.

No estudo recente do European Central Bank (2013) foi constatado que as dinâmicas dos NPLs obedecem as determinantes empíricas como crescimento do PIB, índices de preços, taxas de cambio, e a taxa de juros sobre os empréstimos.

No estudo em desenvolvimento pelo International Monetary Fund (2013) que teve como principal objetivo o estudo dos NPLs em países de diferentes regiões da Europa, até o momento revelou-se que os motivos do aumento dos NPLs devem-se a condições macroeconômicas e fatores específicos dos bancos como má administração, custos de desempenho, risco moral e excesso de empréstimos.

No trabalho de Weißbach e Wilkau (2013) os autores desenvolvem dois modelos de previsão de perdas em NPL de portfólios através de estimação distribuição de probabilidade de acordo com algumas características do tipo de crédito.

No artigo de Klein (2013) foram investigados os NPLs no período de 1998 até 2011 na região dos países do centro, leste e sudeste da Europa. O autor concluiu que os NPLs podem ser atribuídos a condições macroeconômicas e fatores bancários específicos, porém, com esses últimos com baixo poder de explicação. O autor sugere que o desemprego, a depreciação da taxa de juros e alta inflação contribuíram para o aumento dos NPLs.

Ainda de acordo com o Klein (2013) a literatura explica os determinantes dos NPLs ao longo do tempo de duas formas (i) foco nos eventos externos que abordam condições macroeconômicas no qual esses eventos influenciam na capacidade de pagamento dos tomadores de empréstimo, e (ii) que os NPLs são determinados por aspectos infrabancários. No qual através de evidência empírica utilizando os dois conjuntos de fatores mostrou que esses fatores são suportados.

O artigo de Louzis, Vouldis e Metaxas (2010) foi o primeiro trabalho que ressalta os determinantes de NPLs de acordo com o tipo de empréstimo (cartão de crédito, empresas e crédito hipotecário). Os autores realizam o estudo em relação ao setor bancário grego, e os resultados indicaram que esses determinantes podem ser explicados por fundamentos macroeconômicos como PIB, desemprego e taxa de desemprego e qualidade da administração. Em relação aos tipos de empréstimos os autores atestam que o crédito hipotecário é o menos responsivo às condições macroeconômicas.

Estudos conjecturais que mesclam as duas abordagens (intrabancárias e macroeconômica) mostraram uma menor fragilidade em relação aos resultados apresentados e menos questionamentos de ordem metodológica.

De acordo com a revisão bibliográfica, há evidências de que o volume de artigos relevantes e demais estudos sobre os NPLs cresceu após a crise financeira. Isto leva a crer que os NPLs exerceram papel fundamental durante a crise de 2007/2008 que teve como fator fundamental a estrutura da cadeia de securitização de ativos ligados ao crédito hipotecário dos Estados Unidos da América.

Como visto em alguns artigos, inúmeras variáveis exógenas (*e.g.* aspectos macroeconômicos, crescimento do PIB, *etc.*) e variáveis endógenas (*e.g.* má administração, *insider lending*, excesso de empréstimos, *etc.*) exercem influência direta nos NPLs. Entretanto, pouco se falou da influência do que pode ser considerado o agente econômico importante em qualquer sistema financeiro nacional que são os Bancos Centrais.

Os bancos centrais, que em sua maioria, têm o poder de realizar a expansão monetária, e por consequência o crescimento do crédito através de expedientes como controle da taxa de juros e da inflação, e através de regulação do sistema bancário tornando estes mais propensos a expandir os seus portfólios de crédito. Esta questão foi levantada no trabalho de Paul (2011), contudo são necessários estudos mais aprofundados para validação dessa hipótese.

Outra observação sobre a revisão de literatura é que nenhum artigo endereçou a questão dos NPLs partindo da perspectiva da recuperação do crédito, ou mesmo métodos, técnicas ou metodologias orientadas a análise dos créditos já inadimplidos.

A seguir são examinadas as estruturas financeiras que realizam a absorção dos NPLs para fins de aferição de lucros que são os Fundos de Investimento em Direitos Creditórios (FIDCs); estruturas estas que realizam a compra e a posterior operação dos *Non-Performing Loans* para realizar a liquefação desses ativos financeiros.

2.2 FUNDOS DE INVESTIMENTO EM DIREITOS CREDITÓRIOS

De acordo com a Comissão de Valores Imobiliários a constituição de um Fundo de Investimento em Direitos Creditórios (FIDC) caracteriza-se pela política de investimento em aplicações sobre o patrimônio líquido na aquisição de créditos vencidos e pendentes de pagamento na ocasião de sua cessão ao fundo (CVM, 2006).

A estrutura para a composição desses fundos se dá na forma em que os mesmos participam de leilões de instituições financeiras (em geral bancos, e financeiras) denominados cedentes, estes últimos que vendem os direitos de recebimento dessas dívidas vencidas, e o fundo através da contratação de empresas de *call-center* realiza operações de cobrança para recuperação desses créditos, e o posterior pagamento dos investidores.

Para a operacionalização desses créditos, isto é, ao recondicionamento desses ativos financeiros, a legislação brasileira permite a criação de fundos chamados Fundos de Investimento em Direitos Creditórios (FIDCs). Esses fundos têm como principal finalidade a obtenção de lucro através da cobrança de recebíveis das mais diversas modalidades que vão desde crédito imobiliário até dívidas relativas a cartão de crédito.

A Figura 4 apresenta de forma simplificada todo o fluxo operacional que vai desde a originação dos NPLs até a remuneração final do investidor. Cabe ressaltar que os arranjos podem ser distintos, mas de maneira geral esse é o modelo mais usual.

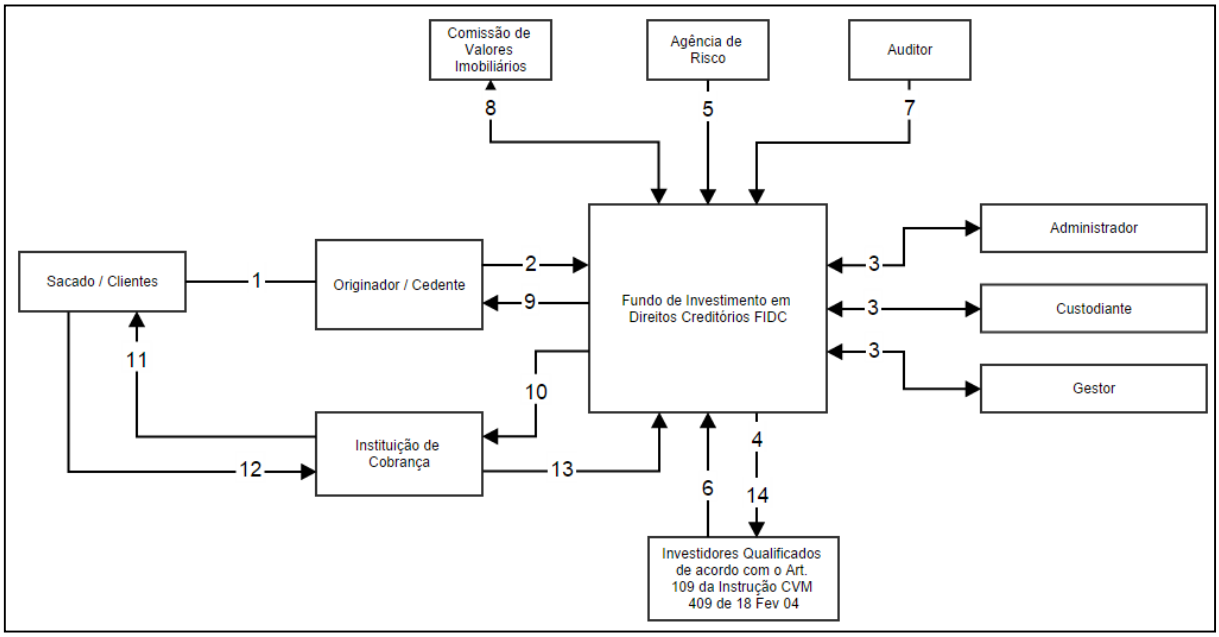


Figura 4 - Fluxo Operacional simplificado de um FIDC. Fonte: Elaborada pelo Autor.

Começando a sequência operacional de um FIDC temos o número 1 que é a relação de originação do NPL, ou seja, no momento em que o cliente fica inadimplente com a instituição originadora do crédito. Um fato a ser observado é que a relação "Originador" e "Sacado" será permanente por questão de rastreabilidade do crédito.

No momento 2 o Originador oferece os NPLs ao FIDC para análise e precificação do ativo através de leilão. A composição desses créditos após a escolha do Originador é chamada de t*ranche*, isto é, o originador 'fatiou' o seu portfólio e ofereceu um conjunto específico de dívidas. No momento 3 o Gestor do Fundo realiza a análise e precificação dos ativos constantes da base de dados da *tranche* enviado pelo Originador. Após a precificação, o Gestor atesta o preço justo do ativo NPL dado à relação risco e retorno é feita uma oferta ao Originador, e caso seja o melhor preço para o Originador, é realizada a captação de recursos junto aos investidores que é o momento 4.

Depois da verificação da viabilidade econômica para a aquisição dos direitos creditórios, é feita uma avaliação de uma agência de risco; esta que por sua vez tem a função de analisar e informar as potencialidades de ganho e avaliar os riscos desses ativos, em que são consideradas desde questões de risco de mercado até mesmo fatores inerentes ao ativo. Esta atividade está descrita no passo 5 e é chamada de *Rating*.

Após a aplicação do *Rating* no conglomerado de ativos, ou seja, esses ativos recebem uma nota que atestem a sua qualidade, no passo 6, os investidores realizam o aporte no fundo de investimentos, isto é, eles disponibilizam o dinheiro para realização da compra dos NPLs junto ao Originador.

No passo 7 é realizada uma auditoria dos créditos que serão disponibilizados, para verificar questões ligadas a consistência dos dados e detecção de possíveis casos de fraude para correção ou eliminação dos créditos para a cessão dos créditos.

Com a verificação realizada é solicitada uma autorização junto à Comissão de Valores Mobiliários (CVM) para aquisição e operação de reestruturação dos NPLs. Essa atividade está no passo 8.

Uma questão que não pode ser deixada de lado é que se caso a CVM não autorize o FIDC a operar, toda a estrutura operacional é desfeita e o FIDC é obrigado a entrar com uma nova requisição de permissão operacional junto à CVM.

Depois de concedida a autorização para reestruturação do NPLs, no passo 9 é realizado o pagamento ao Originador dos NPLs, que nesse momento se transforma em Cedente, pois o mesmo a contar desse momento cede o direito de cobrança do NPL ao FIDC.

Com a cessão, os FIDCs podem optar em internalizar as operações de cobrança; ou como descrito no passo 10 podem alocar/pulverizar esses NPLs em instituições de cobrança como Bancos, Empresas especializadas em cobrança, ou mesmo vender esses ativos para outros FIDCs. Para efeitos de simplificação trabalharemos neste exemplo com uma instituição de cobrança que tem a capacidade de operacionalização das atividades de recuperação através de um *Call-Center*.

No passo 11, a instituição de cobrança inicia as atividades de recuperação ligando para os clientes donos dos NPLs no qual envolve atividades de negociação para liquidação da dívida.

Caso o cliente resolva realizar um acordo, e posteriormente liquidar ou pagar parcialmente o débito, no passo 12 à instituição de cobrança fecha um acordo com o cliente e posteriormente repassa essa informação para o FIDC em 13, que remunera a instituição de cobrança através de comissão pelo acordo fechado.

Com a recuperação efetuada, o FIDC após um determinado tempo inicia a atividade de remuneração dos Investidores como está no passo 14. Essa remuneração pode ser feita através de resgate de cotas, ou reaplicação do dinheiro em novas aquisições.

Em outras palavras, estes fundos obtêm o lucro através da recuperação desses créditos, seja através de cobranças via *Call-Center* ou mesmo através de cobranças jurídicas. Uma proposta de estruturação e criação de estratégias de recuperação desses créditos é apresentada no trabalho de Vogiazas e Nikolaidou (2011) que mostra de forma generalista todo o processo de formulação dessas estratégias.

Com a redução na taxa de juros promovida pelo Comitê de Política Monetária (COPOM) decorrente da estratégia governamental de incentivo ao consumo, e também os estímulos ao setor produtivo e comércio; ampliou-se a oferta de crédito por parte das instituições financeiras (BANCO CENTRAL DO BRASIL, 2012).

Cabe ressaltar que as justificativas para este tipo de expansão podem ser de diversas naturezas como governamental, política, eleitoral, ou mesmo ajustes de cunho macroeconômico. O trabalho de Paul (2011) exemplifica essas motivações de forma detalhada e as consequências nefastas destas práticas de expansão creditícia.

Com essa expansão, cresceu também a inadimplência desses setores; e com este fato o mercado de transferência de direitos creditórios tornou-se uma alternativa para essas instituições financeiras para recuperar parte dos recebíveis que por força de regulação do Banco Central devem ser movidos de forma compulsória para perdas (BANCO CENTRAL DO BRASIL, 1999 & 2000).

A partir da aquisição dessas dívidas por parte de FIDCs devem ser definidas estratégias para a sua recuperação. Um modelo de estratégia foi proposto por *International Finance Corporation* (2012), conforme a Figura 5.

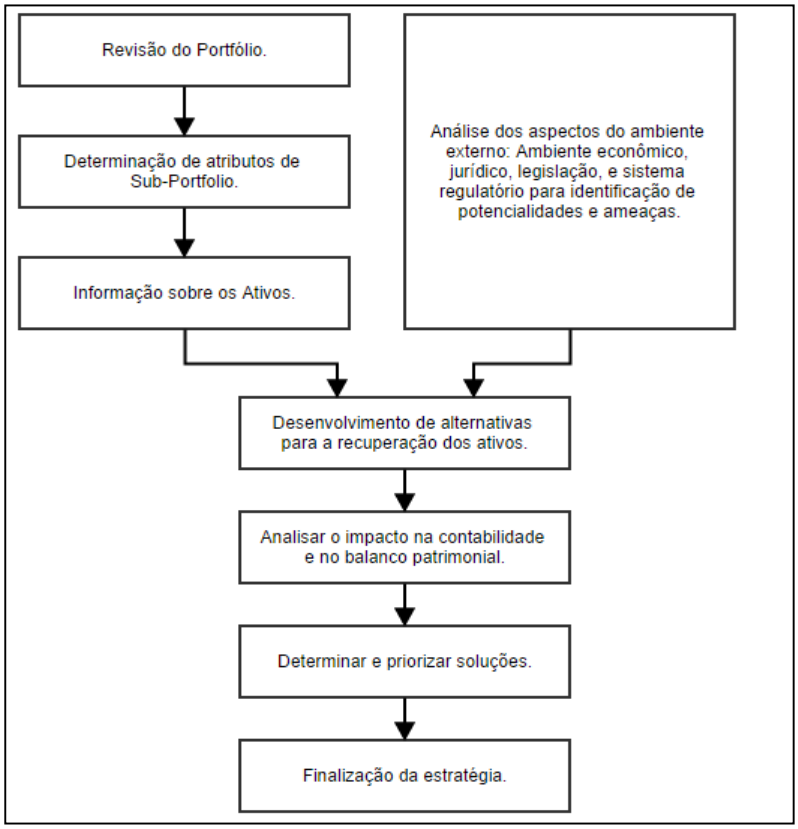


Figura 5 - Modelo de Estratégia para Créditos Não-Performados. Fonte: Adaptado de International Finance Corporation (2012)

A Figura 2 apresenta um método de construção de estratégia para reestruturação desses créditos em que são levados aspectos endógenos relativos aos créditos não performados, como a revisão do portfólio e determinação de atributos desses ativos; como também aspectos exógenos ligados ao ambiente econômico em que essa estratégia de reestruturação é realizada.

Com esse tipo de avaliação são construídas estratégias para a recuperação e reestruturação desses créditos. A partir da aquisição dessas dívidas por parte dos FIDCs devem ser definidas estratégias para a sua recuperação.

Após a apresentação e definição dos aspectos relativos aos NPLs e aos FIDCs, a próxima seção fará a introdução à Inteligência Computacional e algumas de suas técnicas.

2.3 INTELIGÊNCIA COMPUTACIONAL

Nesta seção são apresentados os principais conceitos das técnicas de inteligência computacional utilizadas neste trabalho como (i) Árvores de Decisão, (ii) Teoria dos *Rough Sets*, (iii) Redes Neurais Artificiais, (iv) *Self-Organizing Maps*, e (v) *Chi-squared Automatic Interaction Detection*.

Busca-se com esse referencial conceitual uma apresentação mais em nível de entendimento dos mecanismos básicos de cada uma dessas técnicas, do que uma abordagem profunda em cada uma dessas técnicas.

2.3.1 Árvores de Decisão

O trabalho de Cho, Hong e Ha (2010) aplicou árvores de decisão dentro de um contexto híbrido para previsão de falência, em que as árvores foram responsáveis pela seleção das variáveis para determinar o modelo.

Em Mandala, Nawangpalupi e Praktikto (2012) os autores aplicaram árvores de decisão para levantamento dos fatores para critérios para o fornecimento de empréstimos em um banco rural. Este trabalho usou NPLs para treinamento da árvore usando o algoritmo C5.0 e mostrou que a árvore reduziu cerca de 7% os NPL na classificação dos créditos na escoragem.

Os algoritmos de árvores de decisão têm obtido uma grande evolução ao longo dos anos, tendo em vista que até mesmo métodos de randomização e re-amostragem estão sendo incorporados aos classificadores, de forma em que tenha a eliminação de qualquer tipo de viés ou variância no modelo como assevera o trabalho de Dietterich (2000).

A utilização das árvores de decisão para a modelagem do problema bancário foi exposta no trabalho de Matuszyk, Mues, e Thomas (2010). No trabalho os autores ressaltaram a aplicação das árvores de decisão na modelagem de um modelo de provisionamento de perdas bancárias, modelagem esta que foi ocasionada devido a uma mudança de regulação bancária mundial do acordo de Basiléia. Os resultados do trabalho indicaram que mesmo com uma sobreposição de assuntos oriundos da disciplina de escoragem de crédito (*Credit Scoring*) questões operacionais como se o devedor já tinha ou não registros são de fundamental importância no modelo de provisionamento de perdas.

O trabalho de Bensic, Sarlija, Zekic-Susac (2006) foi explorada as árvores de decisão para a atividade de escoragem de crédito de forma e também foi realizada a comparação entre outros algoritmos para verificação de acurácia e contraste em termos de classificações erradas.

Com os cenários testados os autores chegaram à conclusão que as características do programa de crédito, e características ligadas a informações pessoais e de negócios dos tomadores de empréstimos pessoa jurídica são determinantes para a construção do modelo.

Em Lin, McClean (2001) as árvores de decisão foram utilizadas para a predição de falência corporativa seja de forma isolada ou combinada com outros métodos. Os autores chegaram ao resultado que os modelos combinados de técnicas são os melhores para a tarefa de classificação em comparação com os modelos isolados.

No artigo de Zurada (2010) ele realiza a comparação de três árvores de decisão utilizando uma base de concessão de empréstimos de uma instituição financeira alemã, e os autores verificaram que mesmo as árvores obtendo um nível de acurácia do que outras técnicas como as RNAs, as árvores destacam-se pelo fato de permitirem um grau maior de interpretação e com isso fundamentar de uma maneira muito mais transparente a negativa de um empréstimo.

Loterman e outros (2012) analisaram as perdas bancárias provenientes da inadimplência dos clientes usando árvores regressoras (*Classification and Regression Trees* - CART), entretanto de acordo com os experimentos constatou-se que as Máquinas de Vetor de Suporte (*Support Vector Machines*–SVM) e as Redes Neurais apresentaram melhores resultados.

Tsai, Lu e Yen (2012) usaram árvores de decisão para pré-processamento, isto é, para seleção de atributos em uma base de dados. Os resultados indicaram que a combinação de técnicas, com as árvores de decisão incluídas apresentaram uma acurácia de 75% na valoração de ativos intangíveis como fator representativo.

2.3.2 Teoria dos *Rough Sets*

A Teoria dos *Rough Sets* ou *Rough Sets* (RS) proposta por Zdzislaw Pawlak (1982) é uma teoria matemática aplicada que tem como conceitos de aproximações/classificações para descrição de dados que estejam incompletos ou que contenham um dado grau de incerteza.

Essa teoria encaixa-se com precisão nos problemas atuais em relação ao tratamento de dados, os quais se por um determinado prisma ferramentas de tratamento estatístico podem apresentar problemas de robustez, por outro lado em termos computacionais processar uma base de dados completa pode ter um alto custo computacional, sendo que pode haver diversos problemas como dados faltantes, covariância, *Overfitting* entre outros.

A teoria consiste em realizar aproximações de atributos de acordo com seus valores de atributo e considerando uma variável de decisão; em que os conjuntos de atributos mais relevantes para explicação da variável de decisão são considerados conjuntos aproximados (PAWLAK, 1991). Em outras palavras, o conjunto de características que melhor explica uma variável de decisão é um conjunto aproximado.

O conceito fundamental da Teoria dos *Rough Sets* são os Sistemas de Informação (SI) que são representados por um conjunto de pares ordenados (U,A) no qual U é um conjunto finito e não vazio de objetos, e A é um conjunto finito não vazio de um conjunto de atributos no qual $a : U \rightarrow V_a$ para todo $a \in A$ no qual V_a é o valor de a.

Outro conceito importante são as tabelas de decisão ou Sistemas de Decisão (SD) que são representados por $T = (U, A \cup \{d\})$ no qual $d \notin A$ é o atributo de decisão e os elementos de A são chamados de atributos de condição.

O conceito de indiscernibilidade baseia-se no princípio da relação binária $R \subseteq X \times X$ que obedece aos conceitos de equivalência (isto é tem as propriedades de reflexividade (xRx para qualquer objeto x), simetria (se xRy então yRx) e transitividade (se xRy e yRz então xRz)). A classe de equivalência $[x]_R$ de um elemento $x \in X$ consiste em todos os objetos $y \in X$ os quais são xRy.

Dessa forma sendo o SI= (U, A), então com qualquer $B \subseteq A$ há uma relação associada de equivalência na Equação 3:

$$IND_{IS}(B) = \{(x, x') \in U^2 \mid \forall a \in B, a(x) = a(x')\} \qquad (3)$$

No qual $IND_{IS}(B)$ é chamada de relação de indiscernibilidade. Se $(x, x') \in IND_{IS}(B)$ então os objetos x e x’ são indiscerníveis dado o conjunto de atributos de B.

Sendo T = (U, A); $B \subseteq A$ e $X \subseteq U$ pode ser realizada a aproximação de X usando apenas a informação contida em B construindo a aproximação inferior e superior de B denotadas respectivamente como $\underline{B}X$ e $\overline{B}X$ no qual são representadas pelas Equações 4 e 5.:

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \phi\}. \qquad (4)$$

$$\underline{B}X = \{x \mid [x]_B \subseteq X\}, \qquad (5)$$

A região de fronteira de X consiste de objetos que não podem ser decisivamente classificados em X em B, e é definida pela Equação 6:

$$BN_B(X) = \overline{B}X - \underline{B}X, \qquad (6)$$

Um conjunto é denominado irregular quando a sua região de fronteira não é vazia, ao contrário, pode-se dizer que o conjunto é regular; isto é, apresenta distinções claras de suas regiões de aproximação.

A forma mais comum para representação dos dados em RS é por meio de um sistema de informação (S) que contém um conjunto de elementos, sendo que cada elemento tem uma quantidade de atributos condicionais. Esses atributos são os mesmos para cada um dos elementos, mas os seus valores nominais podem diferir (Tabela 1).

Desta forma, um sistema de informação é um par ordenado S = (U, C), onde U é um conjunto finito e não-vazio de elementos chamado de universo, e C é um conjunto finito e não-vazio formado pelos atributos. Cada atributo a $\in$ C é uma função a: U $\rightarrow$ Va, onde Va é o conjunto dos valores permitidos para o atributo a (sua faixa de valores).

Na Tabela 1, onde é apresentado o Sistema de Informação S, podem-se observar os principais conceitos de RS, o espaço aproximado A = (U, R), o universo U formado pelos elementos Ação1; Ação2; Ação3; Ação4; Ação5; Ação6 e os Atributos (C) Fechamento, Nova Mínima, Preço Máximo e R a relação de equivalência sobre U.

Tabela 1 - Exemplo de um Sistema de Informação (S).

Elementos - Ações	Preço de Fechamento	Nova Mínima	Preço Máximo
Ação1 {e1]	Positivo	Não	Não Rompeu
Ação2 {e2}	Neutro	Não	Não Rompeu
Ação3 {e3}	Neutro	Não	Não Rompeu
Ação4 {e4}	Negativo	Sim	Não Rompeu
Ação5 [e5}	Neutro	Sim	Rompeu
Ação6 {e6}	Positivo	Não	Rompeu

Fonte: Kaupa e Sassi (2013).

O principal conceito envolvido em RS é a Relação de Indiscernibilidade a qual normalmente está associada a um conjunto de atributos (PAWLAK, 1982). Se tal relação existe entre dois elementos, isso significa que todos os valores nominais dos seus atributos são idênticos com respeito aos atributos considerados, portanto não podem ser discernidos (distinguidos) entre si.

Ao utilizar todos os atributos condicionais do sistema de informação S da Tabela 1 obtêm-se os seguintes conjuntos elementares: {Ação1}, {Ação2, Ação3}, {Ação4}, {Ação5} e {Ação6}. Ao sistema de informação que contém o atributo de decisão, no caso desse trabalho o atributo Tendência (Tabela 2), dá-se o nome de sistema de decisão.

Observando a Tabela 2, pode-se perceber que existem 2 (dois) elementos (casos) {Ação2} e {Ação3} iguais (destacados em negrito), no que diz respeito aos valores de atributos condicionais, mas que tem o atributo de decisão diferente.

Tabela 2 - Sistema de Decisão com os elementos Ação2 e Ação3 indiscerníveis.

Elementos - Ações	Preço de Fechamento	Nova Mínima	Preço Máximo	Tendência
Ação1 {e1}	Positivo	Não	Não Rompeu	Alta
Ação2 {e2}	**Neutro**	**Não**	**Não Rompeu**	**Baixa**
Ação3 {e3}	**Neutro**	**Não**	**Não Rompeu**	**Alta**
Ação4 {e4}	Negativo	Sim	Não Rompeu	Baixa
Ação5 {e5}	Neutro	Sim	Rompeu	Baixa
Ação6 {e6}	Positivo	Sim	Rompeu	Alta

Fonte: Kaupa e Sassi (2013).

Existindo a Relação de Indiscernibilidade entre os elementos {Ação2} e {Ação3} como mostrado na Tabela 2, significa que todos os valores nominais de seus atributos são idênticos com relação ao subconjunto de atributos B ($B \subseteq S$) considerado, ou seja, são indiscerníveis,

não podem ser diferenciados entre si. A Relação de Indiscernibilidade é o conceito utilizado por RS para gerar regras de decisão usadas em classificação (PAWLAK, 1996).

A Teoria dos *Rough Sets* ou *Rough Sets* (RS) é aplicada também em uma série de artigos provenientes da área de finanças (ZURADA e ZURADA, 2002; HE, LIU e XIA, 2010; MANDALA, NAWANGPALUPI, PRAKTIKTO, 2012; CLÉSIO e SASSI, 2014).

2.3.3 *Chi-squared Automatic Interaction Detection* - CHAID

De acordo com Mitchell (1997) aprendizado por árvore de decisão é um método de aproximação de funções objetivo com valores discretos em que a função é representada por uma árvore de decisão e que esse método pode ser representado por um conjunto de regras "Se...Então" que melhora a compreensão humana do modelo.

As árvores de decisão representam uma estrutura que através de conjunções e disjunções explicam a estrutura de consequências de um conjunto de dados através do aprendizado indutivo das árvores (MITCHELL, 1997).

Para realizar a construção das árvores de decisão também são utilizadas métricas como entropia, que mede a impureza das instâncias dado o conjunto de dados ou heterogeneidade dos registros; e, Ganho de Informação que é a medida que mostra a efetividade do atributo de acordo com os dados de teste dada uma variável dependente.

Um dos métodos de aprendizado indutivo via árvore de decisão é o algoritmo *Chi-Squared Automatic Interaction Detecto*r (CHAID) que foi proposto por Kass (1980) baseado na modificação do algoritmo *Automatic Interaction Detection* (AID) de Morgan e Sonquist (1963).

A proposta original do AID foi a criação de um método de árvore de decisão que usa o particionamento recursivo para previsão de variáveis de decisão quantitativas. Uma das variações do AID é o THAID que foi proposto por Morgan e Messenger (1973) que através do critério de Theta para resultados categóricos. A extensão do THAID é o MAID que é o acrônimo do *Multivariate Automatic Interaction Detector* proposto por Gillo (1972) e Gillo e Shelly (1974).

O algoritmo propunha fusão e divisões baseadas no teste estatístico do Chi-quadrado, pois Kass (1980) acreditava que o tempo computacional era uma preocupação dentro do que tange ao processamento das árvores de decisão, dessa forma ao invés de realizar todos os cálculos relativos à árvore, o algoritmo encontrava divisões sub-ótimas para geração dos nós e folhas.

Cabe ressaltar que no período em que compreendeu o desenvolvimento desses algoritmos, algumas distinções relativas aos seus mecanismos de funcionamento ficam bem evidentes. No trabalho de Ritschard (2010) o autor apresenta uma tabela com as características dos algoritmos de árvore de decisão, conforme o Quadro 1.

Quadro 1 - Características principais dos algoritmos de Árvore de Decisão.

Algoritmo	**Particionamento Local**	**Variável Dependente**		**Critério de Particionamento**		
		Quantitativa	**Categórica**	**Associação**	**Pureza**	**p-valor**
AID	Binário	X		X		
MAID	Binário	X		X		
THAID	Binário		X	X	X	
CHAID	enésimo	X	X	X		X

Fonte: Adaptado de Ritschard (2010).

De acordo com o Quadro 1 a técnica CHAID é o único algoritmo da família AID que consegue trabalhar tanto com variáveis discretas quanto contínuas. Isso permite uma maior flexibilidade no uso do algoritmo, além de eliminar a necessidade de ajustes nos dados, e assim reduzindo o tempo de pré-processamento da base de dados.

De acordo com Magidson e Vermunt (2005) uma limitação do CHAID é que os segmentos são definidos baseados no critério de seleção de uma única variável, e dado que situações em que possam ter critérios múltiplos, não é tão claro como o algoritmo se comportaria para segmentar de forma única um dado conjunto de dados.

Uma vantagem do CHAID é que como o seu algoritmo original previa o uso de variáveis numéricas (contínuas) ou categóricas (discretas) esse algoritmo tem a vantagem de trabalhar com dados não paramétricos, o que representa uma vantagem sobre métodos como, por exemplo, os métodos regressores. No entanto, como descrito originalmente no trabalho de Kass (1980) no momento em que há variáveis categóricas na base de dados, todas as outras variáveis são categorizadas pelo algoritmo.

Outra característica do CHAID é que enquanto os outros algoritmos de árvore de decisão como o CART, C4.5, C5.0, e ID3 usam as métricas de entropia e ganho de informação, o CHAID tem como principal métrica para definição de particionamento o teste estatístico de Chi-quadrado de acordo com as instâncias no conjunto de dados.

O algoritmo CHAID funciona através do particionamento dos registros de um conjunto de dados e para cada variável de decisão (dependente) são agrupadas as respectivas categorias (variáveis).

A partir do agrupamento é realizado o teste de significância estatística Chi-quadrado de Pearson de acordo com a tabulação cruzada de cada uma das variáveis de decisão. Finalmente

as variáveis com menores p-valor são escolhidas como divisoras (*i.e.* particionamento). Uma demonstração algorítmica mais detalhada pode ser verificada no trabalho de Melo e Murakami (2010).

2.3.4 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são modelos computacionais distribuídos compostos de unidades de processamento densamente conectadas e realizam atividades de otimização de uma função objetivo em relação a tarefas de aprendizado (FACELI *et. al.*, 2011)

As RNAs são usadas em muitos estudos devido ao seu poder de modelagem sobre dinâmicas de dados que possuem comportamento de natureza não linear, e também pelo seu poder de acurácia relativa à predição (HAYKIN, 1999).

Várias são as arquiteturas de RNAs, dentre elas pode-se destacar os *Self-Organizing Maps* (SOM) ou Mapa Auto-Organizável de Kohonen, e a *Multilayer Perceptron*.

- *Self-Organizing Map* (SOM) ou Mapa Auto-Organizável de Kohonen

Uma Rede Neural Artificial do tipo Mapa Auto-Organizável de Kohonen ou *Self-Organizing Map* (SOM) é uma arquitetura de rede neural artificial com aprendizado não supervisionado, baseada em um mapa de neurônios, cujos pesos são adaptados para verificar padrões semelhantes em relação a um conjunto de treinamento (KOHONEN, 2001).

Sua principal característica é o mapeamento ordenado dos padrões de entrada de elevada dimensão em reticulados de neurônios de saída com dimensão menor, comumente duas, o que facilita a visualização dos dados.

Para uma dada base de dados com N amostras com d atributos cada, em que *d* determina a dimensão dos padrões de entrada, ocorrerá um mapeamento desses padrões para um reticulado de neurônios de saída arranjados em 2D, como mostra a Figura 6.

A rede SOM é uma arquitetura de Rede Neural Artificial, estruturada em duas camadas, entrada e saída. Os neurônios da camada de saída são comumente dispostos em um mapa de duas dimensões, com uma dada relação de vizinhança.

A Figura 6 ilustra essa arquitetura, com **d** atributos na camada de entrada e um conjunto de unidades u (neurônios) arranjados na forma de um mapa em **2D** na camada de saída. Cada **u** é caracterizado por sua posição **x** e **y** no mapa, que é representado por **ux** e **uy**, respectivamente, que resulta em um vetor 2D igual a u=[ux uy]. Cada u tem associado um vetor protótipo mu = [m1u, m2u,..., mdu], sendo d a dimensão do protótipo, a mesma do padrão de entrada.

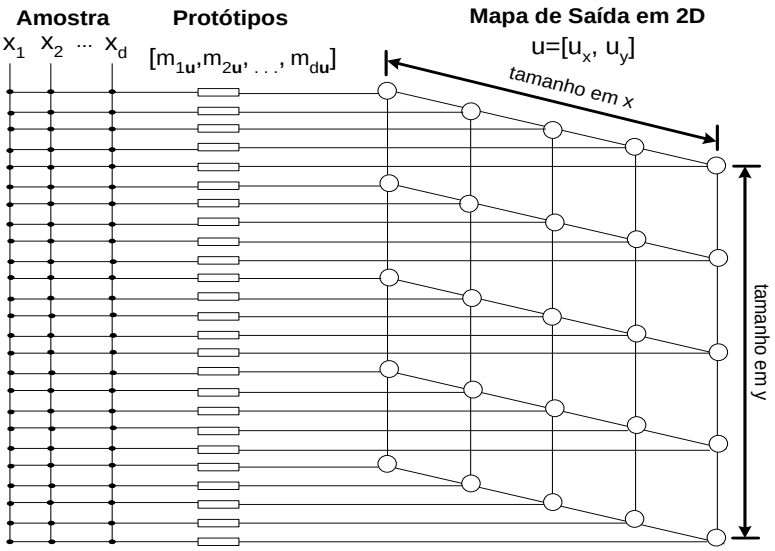


Figura 6 - Arquitetura da rede SOM. Fonte: Adaptada de Kohonen (1982).

O algoritmo de aprendizado da rede SOM é realizado em um processo iterativo, onde, no primeiro passo, t=0, inicializa o vetor protótipo (m) randomicamente. Porém, a inicialização do m, pode ser feita de outras maneiras (KOHONEN, 2001).

O algoritmo de treinamento da rede SOM é também chamado de competitivo. Em cada passo do processo (ou época), uma amostra x é randomicamente escolhida do conjunto de treinamento. A distância, geralmente euclidiana, entre x e todos os vetores protótipos m é calculada. A unidade com menor distância, chamada de BMU (*best-matching unit*) é o u com protótipo m mais próximo a x, conforme a Equação 7.

$$||x - mbmu|| = arg||x-mu|| \qquad (7)$$

-*Multilayer Perceptron* (MLP)

Uma rede artificial do tipo Multilayer Pereceptron (MLP) é uma arquitetura de rede neural artificial organizada em camadas como pode ser visualizado na Figura 7.

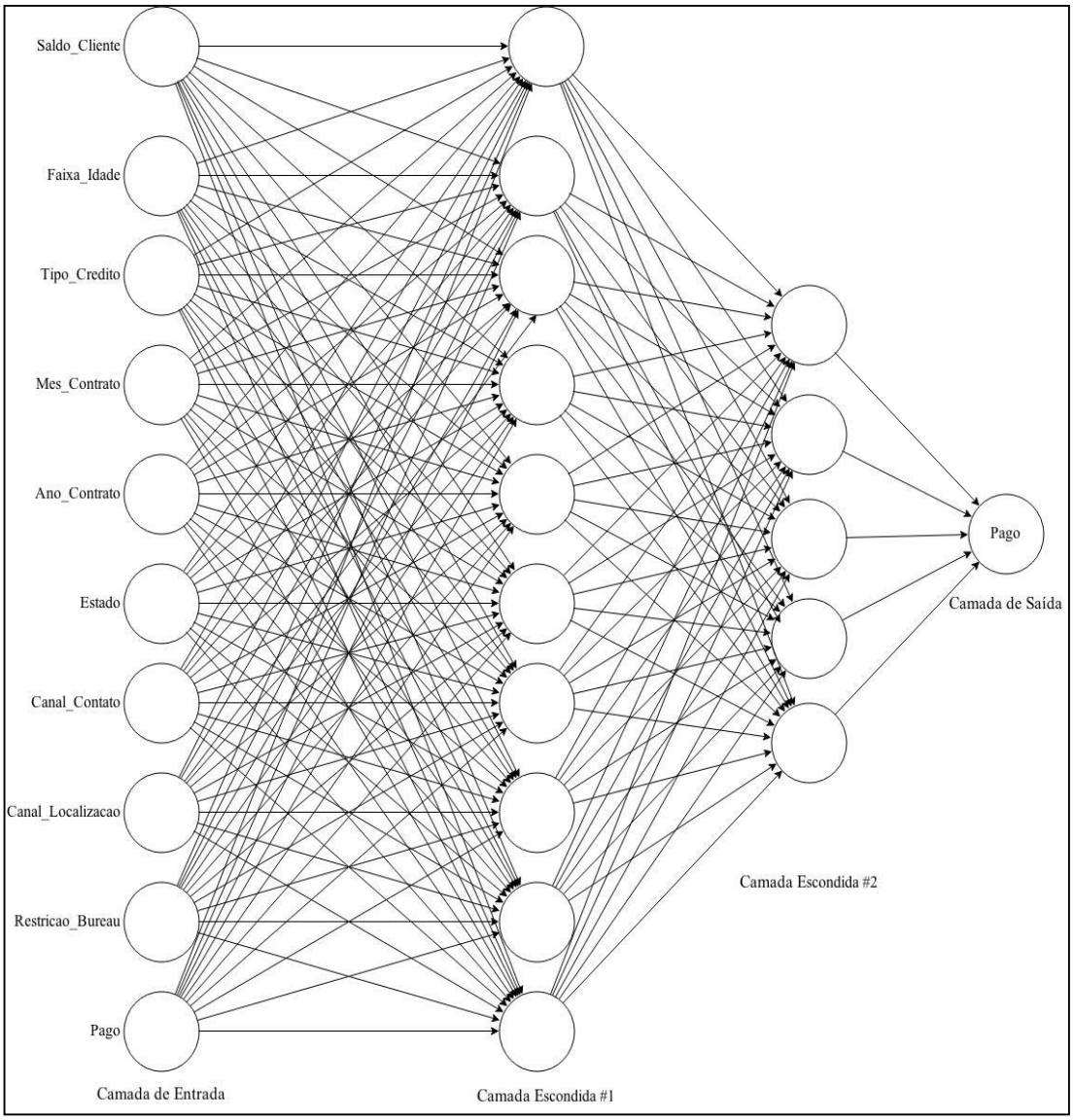


Figura 7 - Arquitetura da MLP do Modelo 4 (M4) com camada de entrada, camadas escondidas e camada de saída. Fonte: Elaborada pelo Autor.

A Figura 7 apresenta uma RNA com a arquitetura de dez neurônios na camada de entrada, dez neurônios na primeira camada escondida, cinco neurônios na segunda camada escondida e com um neurônio na camada de saída. Como as interações entre os neurônios se fazem apenas com os neurônios da camada imediatamente posterior, este modelo de aprendizado da RNA é conhecido com *feedforward*.

Outro processo de aprendizado da MLP consiste na apresentação do conjunto de dados de treinamento, e na medida em que haja erros de classificação eles são ajustados e voltam de forma iterativa influenciando nos pesos sinápticos, a fim de minimizar os erros nas próximas iterações.

Este método de aprendizado da MLP é denominado de retropropagação (ou retropropagação de erros - *backpropagation*) criado por Werbos (1974), que consiste em executar dois passos relativos ao método de treinamento, um passo para frente para formulação do resultado e a propagação reversa para com os erros obtidos na classificação ou regressão para ajuste dos pesos sinápticos.

O algoritmo do *backpropagation* é descrito no Quadro 2.

Quadro 2 - Algoritmo de Treinamento do *Backpropagation*.

#	**Atividade**
1	Inicialização dos Pesos de Forma Randômica
2	Repita
3	Repita Época
4	Escolha uma instância do conjunto de treinamento
5	Aplique a instância na rede
6	Avalie a saída da rede
7	Compare o valor da saída da rede com o valor desejado
8	Realize a ponderação dos pesos até a escolha de todas as instâncias do conjunto de treinamento
9	Até que o erro global < critério

Fonte: Adaptado de Drchal (2012)

No algoritmo descrito no Quadro 2, o passo fundamental que diferencia o processo de aprendizado *backpropagation* são os passos 7 e 8 em que o algoritmo não apenas aprende com as interações imediatamente posteriores, mas também após a ponderação dos pesos informa os erros para a camada imediatamente anterior, assim melhorando o aprendizado da rede como um todo.

A ideia geral é que a rede aprenda com a minimização dos erros, ou seja, quanto menor o erro, mais a rede se ajustará aos dados do modelo. Sendo assim, unindo as características de processamento distribuído e síncrono, juntamente com o aprendizado baseado na minimização do erro a abordagem com MLP torna-se uma boa alternativa para problemas de generalização de classificação de bases de dados.

No próximo capítulo os Materiais e Métodos utilizados nesta pesquisa são apresentados e discutidos.

3 MATERIAIS E MÉTODOS

3.1 CARACTERIZAÇÃO METODOLÓGICA

A abordagem metodológica definida para este trabalho foi de caráter experimental e empírica. Um experimento é um método de pesquisa científica que visa testar o impacto da variação de determinado aspecto sobre um fenômeno, controlando-se todos os demais aspectos que atuem sobre ele (CRESWELL, 2009).

Montgomery (2008) define experimento como um teste ou uma série de testes no qual são propostas mudanças nas variáveis de entrada de um processo ou sistema de modo em que possam ser observadas as influências na variável de saída.

A pesquisa empírica é a aquela utilizada com o objetivo de conseguir informações e/ou conhecimento acerca de: (a) Determinado problema para o qual se procura uma resposta; (b) uma hipótese que se queira comprovar; e (c) descobrir novos fenômenos ou uma relação entre eles (MARCONI; LAKATOS, 2010).

Este trabalho contou também com ampla pesquisa bibliográfica por meio da utilização da ferramenta de revisão sistemática proveniente da disciplina de blibliometria; atividade esta que permitiu realizar uma varredura de forma sistematizada na literatura. A pesquisa bibliográfica trata-se de uma leitura atenta e sistemática que se faz acompanhar de anotações e fichamentos que, eventualmente, poderão servir à fundamentação teórica do estudo (GIL, 2002).

As palavras chave escolhidas foram: *Non-Performing Loans*, Inteligência Computacional, NPL, NPA, *Non-Performing Assets*, Créditos Não-Performados, e *Computacional Intelligence*.

As bases de periódicos consultadas foram: IEEE Xplore, Periódicos Capes, Science Direct, Google Scholar, ProQuest, e EBSCO. Foram coletados 416 artigos, com a seleção de 104 artigos que foram triados através de leitura e posterior indicação de relevância de acordo com a aderência no trabalho.

Esta abordagem foi escolhida devido ao que assevera Coughlan e Coughlan (2002) em que os autores afirmam que pesquisas dessa natureza podem auxiliar nas deficiências associadas aos métodos e tópicos de pesquisa tradicionais, bem como é de alta relevância para profissionais das áreas investigadas (pesquisador-profissional) para aplicação em questões integrativas ou desestruturadas.

Neste trabalho foi tratada a situação dos créditos em situação de *Write-Off* em que são os FIDCs especializados em aquisição de créditos já inadimplidos e massificados, isto é, créditos que podem ter uma miríade de produtos financeiros envolvidos nas carteiras estudadas, o que corresponde a uma parte da inadimplência no sistema de crédito nacional.

A abordagem desse trabalho é realizada através de créditos inadimplidos reais que já são NPLs e terá como propósito a apresentação de padrões que geralmente não estão implícitos nos dados, ou minimizar aspectos relativos à assimetria de informação entre as duas partes envolvidas na negociação; isto é, o vendedor dos créditos (cedente) e o comprador (cessionário).

3.2 ORGANIZAÇÃO DA REVISÃO DE LITERATURA

A organização da revisão de literatura foi realizada de acordo com a estrutura de Cooper (1988) quando o autor elenca os principais tópicos para a cobertura correta de uma revisão de literatura para o posicionamento do trabalho que são: (i) Foco, (ii) Objetivo, (iii) Perspectiva, (iv) Cobertura, e (v) Organização. De acordo com os pressupostos do autor, esta pesquisa está posicionada dentro dos seguintes aspectos, conforme o Quadro 3:

Quadro 3 - Caracterização da presente pesquisa, conforme os tópicos de coberta de revisão de literatura proposto por Cooper (1988).

Característica	**Categoria**
Foco	Aplicação
Objetivo	Identificação de questões centrais
Perspectiva	Exposição de posição
Cobertura	Exaustiva com citações seletivas
Organização	Histórica

Fonte: Cooper (1988)

Com os tópicos listados no Quadro 3 foi possível estruturar a revisão de literatura não somente em sistematizar a coleta de artigos, mas permitiu posicionar a presente pesquisa na literatura e principalmente colocar o trabalho em comunicação com os trabalhos dos outros autores da literatura.

De acordo com os tópicos para a cobertura da revisão de literatura de Cooper (1988) em *foco*a pesquisa caracteriza-se por realizar a busca em bases teóricas voltadas na aplicação de Inteligência Computacional em bases de dados de NPLs com o objetivo de identificar determinantes para a sua recuperação.

Em *objetivo* o trabalho vai para uma linha para a identificação das questões centrais, em especial dos determinantes e na caracterização dos ativos NPL, isto é, no problema da recuperação dos NPLs o que está sendo discutido para a resolução ou minimização desse problema.

Em *perspectiva* a pesquisa vai para a linha de exposição de posição na qual como é a única pesquisa em que saí da perspectiva macroeconômica dos créditos e enfatiza os determinantes para recuperação de NPLs.

Já em *cobertura* a pesquisa foi feita de forma exaustiva com a seleção e posterior análise de todos os artigos nas bases de dados IEEE Xplore Digital Library, Periódicos Capes, Science Direct, Google Scholar, ProQuest, e EBSCO, e mais artigos através do Google com as palavras-chave *Non-Performing Loans, Computacional Intelligence,* Inteligência Computacional, Créditos Não-Performados; que iniciou com 416 artigos e terminou com 104 selecionados de acordo com a relevância.

E por último a *organização* escolhida foi a organização histórica tendo em vista que os *Non-Performing Loans* foram um dos determinantes da crise mundial de 2007/2008, e revisar as origens desse problema específico se faz necessário para o seu entendimento e contextualização na pesquisa.

3.3 PLANEJAMENTO AMOSTRAL

Devido à heterogeneidade da natureza desses créditos, isto é, estes créditos podem ter diversas fontes ou formas de acordo com o tipo do originador; este trabalho considera sensível a generalização de resultados em análises de NPLs. Isto significa que cada tipo de crédito inadimplido pode obedecer a diversas dinâmicas de recuperação devido às suas características como se o crédito é do tipo Crédito Direto ao Consumidor (CDC) se é crédito hipotecário, se é crédito direcionado para atividades específicas ligadas a uma estratégia governamental de subsídio de créditos, entre outros.

Outro fato a ser considerado é que devido ao fato que a seleção, *i.e.*a composição das *tranches*, submetem-se a questões que podem enviesar a amostra como:

(i) Os créditos estarem enviesados devido a questões ligadas a falhas no programa de escoragem de crédito se houver, da instituição que está cedendo o crédito;

(ii) Enviesamento da amostra devido ao fato de que a própria concepção de *tranche*, a qual pode obedecer a critérios estratégicos e subjetivos da instituição que cede o crédito; e

(iii) A composição dos créditos que foram pagos (variável de resposta desejada para análise dos classificadores) obedecer a critérios randômicos ou não listados nas variáveis independentes.

Isso de nenhuma forma invalida os resultados das análises, mas sim reforça o fato de que esse estudo obedece a características particulares, e que a sua reprodutibilidade em outros cenários de NPL pode sofrer alterações devido às características listadas acima.

Dessa forma, o levantamento amostral foi adaptado de acordo com o apresentado na obra de Bolfarine e Bussab (2005) e adaptado, conforme o Quadro 4.

Quadro 4 - Levantamento Amostral.

Fase da Elaboração Amostral	**Descrição**
a) Identificação de Objetivos e Populações	**Estabelecer a população-alvo:** A população alvo deste estudo consiste em créditos do tipo *Non-Performing Loans* de um Fundo de Investimento em Direitos Creditórios com valor de face (*i.e.* valor nominal) de 13 Bilhões de Reais.
	Especificação dos parâmetros populacionais de interesse: Os parâmetros populacionais seguem os seguintes critérios: (i) créditos vencidos a mais de 180 dias, (ii) dívidas provenientes de cartão de crédito ou produtos bancários como empréstimos pessoais, (iii) sejam de a) instituições bancárias, ou b) instituições que trabalham especificamente com operações de crédito (financeiras)
	Descrição da população amostrada: a população amostrada consiste de três portfólios de créditos de instituições com as seguintes características de acordo com os respectivos experimentos: No experimento 1 com as RNAs foi realizada a extração de uma base de dados de um banco múltiplo de abrangência multinacional e de grande porte, que será denominado como '*Banco 1'*. No experimento 2 foi selecionada uma base de dados de um banco múltiplo nacional de médio porte, com foco nas operações de varejo e arrendamento mercantil, que será denominado como '*Banco 2'*. No terceiro e último experimento foi utilizada uma base de dados de um banco múltiplo nacional de médio porte, com foco nas operações de crédito direto ao consumidor, financiamento de veículos, cartão de crédito e nas modalidades de empréstimo pessoal e consignado, que será denominado como '*Banco 3'*. Por motivos de privacidade e manutenção de contratos de confidencialidade os nomes dos portfólios foram omitidos.

Fase da Elaboração Amostral	Descrição
b) Coleta das Informações	**Escolher o tipo de investigação:** A amostragem foi escolhida devido ao fato de que mesmo tratando-se de um estudo relativo à aplicação de técnicas de Inteligência Computacional em um alto volume de dados; nota-se que pela utilização da técnica de inferência amostral para obtenção de tamanho de amostra são obtidos os benefícios operacionais de pesquisa como (i) menor chance de erros ligados à manipulação de dados, (ii) minimizar fragilidade estatística no que tange a confiança de que a base é estatisticamente significante para aumentar a robustez dos modelos sejam eles de classificação, associação, ou agrupamento, (iii) e melhoria no sentido ao uso dos recursos computacionais, *i.e.* minimizar a complexidade computacional do estudo (*e.g.* complexidade temporal e complexidade espacial).
	Estabelecer o modo de coleta: O método de coleta foi consulta ao banco de dados com exportação automática e direta a planilhas do *software* Microsoft Excel 2010. Foi escolhida essa forma devido ao fato de que a manipulação dos dados de sua fonte não teria nenhum tipo de interferência humana. Isto é, os dados extraídos seriam brutos e a manipulação ocorreria em um momento posterior de acordo com a técnica a ser empregada para o estudo (*e.g.* binarização para Redes Neurais, discretização para o uso de árvores de decisão ou regras de associação, etc.).
c) Planejamento e seleção da amostra	**Fixar tamanho da amostra:** Para o experimento 1 com as RNAs foram escolhidas todas dívidas do *Banco 1* que ao menos tiveram uma ação de cobrança. Isto significa que todas as dívidas da amostragem tiveram algum tipo de tentativa de recuperação em algum ponto do tempo. A distribuição dos créditos que foram pagos e não pagos obedeceu a mesma proporção da base de dados que está atualmente em cobrança referente ao portfólio completo. No experimento2 com *Rough Sets* e SOM, foram escolhidas todas as dívidas do *Banco 2* que não tivessem nenhum tipo de suspeita de fraude, e que tiveram algum tipo de contato com os respectivos devedores no último ano. A distribuição dos créditos pagos e não pagos está na mesma proporção do que a base de dados do portfólio como um todo. Já no experimento 3 com *Rough Sets* e Árvores de Decisão foram escolhidos de forma aleatória1% do total de créditos de todo o portfólio do *Banco 3*, respeitando-se as proporções de clientes pagos e não pagos da base real. Um aspecto a ser observado é que esses créditos podem ter um forte viés seja (i) a uma possível do sistema de escoragem que antecedeu a classificação, (ii) vieses dos problemas de NPL explicados no trabalho de Berger e Young (1997), e (iii) determinação de viés em quem realizou o corte da 'tranche' como é apresentado em Chacko (2006) e Motohashi *et. al.* (2006).
d) Processamento dos dados	Planejar e criar banco de dados e dicionário de variáveis: o dicionário das variáveis extraídas do banco de dados estão na seção que descreve cada uma das bases de dados.

Conforme a estruturação do planejamento amostral apresentado no Quadro 4 a caracterização amostral de cada uma das bases de dados foi realizada obedecendo a critérios distintos de seleção. Essa caracterização teve como principal objetivo minimizar qualquer viés amostral anterior à análise que pudesse influenciar nos resultados finais dos experimentos.

3.4 METODOLOGIA EXPERIMENTAL

Foram realizados três experimentos utilizando as seguintes técnicas de Inteligência Computacional: Redes Neurais Artificiais, Teoria dos *Rough Sets* e Árvores de Decisão, aplicadas de forma isolada ou de forma combinada.

A Metodologia Experimental foi dividida da seguinte forma:

a) Experimento 1: Experimento com Redes Neurais Artificiais;
b) Experimento 2: Experimento com *Self-Organizing Maps* conjuntamente com a Teoria dos *Rough Sets*;
c) Experimento 3: Experimento com a Teoria dos *Rough Sets* conjuntamente com as Árvores de Decisão

Buscou-se com a aplicação de técnicas de Inteligência Computacional a obtenção de regras, padrões, relações e tendências nos dados armazenados nas bases para subsidiar a análises e obtenção dos resultados dos experimentos.

3.4.1 Bases de Dados, Ferramentas e Plataformas de Experimentos

As bases de dados utilizadas neste trabalho para a realização dos experimentos são de créditos das mais distintas naturezas como descrito na seção 3.3. No entanto, correspondem a créditos que já tiveram ao menos uma ação de cobrança junto aos seus clientes. Foram incluídos também créditos da mesma *tranche* que por ventura foram liquidados para as atividades de classificação e análise de *cluster*; além de testar o poder de discriminação dos classificadores envolvidos nos experimentos. A caracterização das bases de dados utilizadas nos experimentos pode ser encontrada na subseção 3.5 (Condução dos Experimentos).

As máquinas utilizadas para a plataforma de *hardware* para a realização dos experimentos foram um *notebook* Samsung Modelo RV410 com Windows 8 *Professional* com 2,3 GHz de processamento, 2Gb de memória e 100Gb de disco; e um *notebook* Asus modelo K45VM processador Intel core i7com Windows 8 *Professional*, 2.30Ghz de processamento e 8Gb de memória RAM com 500Gb de disco.

Os *softwares* utilizados para os experimentos e demais atividades como construção de gráficos e armazenamento de dados estão no Quadro 5.

Quadro 5 - *Softwares* utilizados para a realização dos experimentos e criação de gráficos, quadros e tabelas.

Software	Função Principal	Utilização	URL
SQL Server 2008 R2	Sistema Gerenciador de Banco de Dados (SGBD)	Acesso às bases e extração de dados do SGBD	http://www.microsoft.com/pt-br/download/details.aspx?id=30438
Excel 2010	Editor de planilhas eletrônicas	Armazenamento dos dados extraídos e criação de gráficos	http://office.microsoft.com/pt-br/excel/
Tabelau Desktop 8.2	*Software* de visualização de dados	Criação de gráficos	http://www.tableausoftware.com/pt-br/products/desktop
Weka 3.7.7	*Software* de aprendizado de máquina	Experimentos com as técnicas de Redes Neurais Artificiais, LibSVM, Regressão Logística e *Logistic Model Trees*	http://www.cs.waikato.ac.nz/ml/weka/downloading.html
IBM SPSS 20	*Software* de aprendizado de máquina e estatística	Experimento com a técnica de Árvores de Decisão.	http://www-01.ibm.com/software/analytics/spss/products/statistics/
Viscovery Mine	*Software* de *Data Mining*	Experimento com a técnica de *Self-Organizing Maps*	http://www.viscovery.net/viscovery-suite
Rosetta	*Software* de aprendizado de máquina e *Rough Sets*	Experimento com a Teoria dos *Rough Sets*	http://www.lcb.uu.se/tools/rosetta/
RSES	*Software* de aprendizado de máquina e *Rough Sets*	Experimento com a Teoria dos *Rough Sets*	http://www.mimuw.edu.pl/~szczuka/rses/start.html

O Quadro 5 apresenta cinco ferramentas proprietárias (*e.g.* SQL Server 2008 R2, Excel 2010, Tableau Desktop 8.2, IBM SPSS 20, e Viscovery Mine) e três ferramentas de código livre (*e.g.* Weka 3.7.7, Rosetta e RSES) que foram utilizadas neste trabalho. Todas as ferramentas proprietárias do Quadro 5 foram utilizadas em seu período de avaliação, *i.e.* período em que todas as suas funcionalidades estavam liberadas para uso para fins de testes por parte dos potenciais compradores.

3.4.2 Extração e Tratamento Inicial dos Dados

Inicialmente os dados foram extraídos de um banco de dados relacional com o Sistema Gerenciador de Bancos de Dados SQL Server 2008 por meio de consultas usando a linguagem de programação Transact-SQL (T-SQL).

Após a extração dos dados passaram pelo processo de anonimização, isto é, foram retiradas todas as informações que pudessem caracterizar os clientes; ou quaisquer informações que pudessem caracterizar a instituição originadora do NPL, ou mesmo qualquer tipo de informação que potencialmente pudesse ferir questões relativas ao sigilo ou confidencialidade.

Após isso, os dados foram armazenados de forma tabular, sem nenhum tipo de modelagem de banco de dados específica, no formato para o Microsoft Excel ® (.xls ou .xlsx).

Posteriormente por meio de ferramentas de visualização como o Tableau® e o Microsoft Excel® foram construídos gráficos e elaboração de estatísticas descritivas para sumarização e/ou apresentação das principais características dos dados seja através de tabelas ou gráficos.

3.5 CONDUÇÃO DOS EXPERIMENTOS

3.5.1 Experimento 1: Experimento com Redes Neurais Artificiais

A escolha das Redes Neurais Artificiais (RNAs) para este problema foi devido à sua característica de aprendizado através dos dados para tarefas de classificação sobre problemas não-lineares.

O objetivo principal deste experimento foi realizar a aplicação das Redes Neurais Artificias (RNA) com cinco diferentes topologias para a tarefa de classificação de uma base de dados de créditos NPL, em que estão em uma mesma base de dados créditos NPL que já foram pagos e créditos ainda pendentes de pagamento.

Com as diferenciações das arquiteturas das RNAs buscou saber se dados os mesmos parâmetros para o conjunto de cinco RNAs com topologias distintas, qual arquitetura da rede obtêm o melhor desempenho.

Este experimento foi realizado de forma isolada de outras técnicas, *i.e.* não houve aplicação em conjunto de outras técnicas seja para tarefas de pré-processamento ou mesmo pós-processamento para auxílio na tarefa de classificação, devido ao fato de que as RNAs possuem inúmeras possibilidades de parametrizações, seja à nível da técnica no que concerne à parte algorítmica (*e.g. momentum*, taxa de aprendizado, número de épocas) como também à parametrizações relativas à arquitetura da rede (*e.g.* número de neurônios na camada de entrada, número de neurônios na camada escondida, *etc*.).

Dessa forma escolheu-se apenas a realização de modificações na arquitetura de cada uma das redes para verificar qual tipo de arquitetura tem a melhor adaptação para que em um segundo momento possa ser verificada a influência dos parâmetros algorítmicos da técnica.

Sendo assim a validação ou verificação empírica de qual das cinco arquiteturas têm o melhor desempenho em relação à classificação será feita através da abordagem sensível ao custo.

Essa abordagem baseia-se na estrutura de recompensa e penalização de classificadores, em que de acordo com diferentes qualificações de erros a rede sofrerá uma penalização com a atribuição de uma pontuação negativa, e conforme o classificador tenha diferentes acertos haverá em contrapartida atribuição de pontuações positivas.

A escolha dessa abordagem como medida de desempenho dos modelos constituídos pelas diferentes arquiteturas foi realizada devido à natureza do problema de classificação que é de cunho financeiro. Em outras palavras, o objetivo dessa abordagem é fornecer subsídios para escolher uma técnica que tenha os menores custos possíveis, o que é um problema tipicamente econômico.

Para determinar estes custos foi adotada a matriz de confusão tal como proposta por Kohavi e Provost (1998) que informa os valores previstos em relação aos valores atuais após a classificação.

Para avaliação dos modelos de classificação, foi criada uma estrutura de custos econômicos que consta no Quadro 6 determinando parâmetros de acordo com um custo relativo de sucesso ou erro de classificação em uma abordagem relativa à cobrança de uma dívida. Isto significa que os custos relativos a uma operação de recuperação desses ativos podem variar de acordo com os custos diretos da instituição como, por exemplo, campanhas de cartas para os devedores, comunicação telefônica, custo de operadores em um *Call-Center*, custo dos negociadores, telefonia, espaço físico, entre outros.

A tabela de custos foi elaborada de forma empírica, já que este tipo de operação de estruturação contém inúmeras variáveis e dinâmicas os quais podem aumentar ou diminuir os valores determinados na tabela de custos. A tabela dos custos que será usada para atribuição nos modelos está disposta no Quadro 6.

Quadro 6 - Tabela de custos. Pontuações para recompensa e penalização do modelo.

	Parâmetros de Custos	
Custo/Benefício do Modelo	**TP – Verdadeiros Positivos**	-1
	FP – Falsos Positivos	9
	FN – Falsos Negativos	4
	TN – Verdadeiros Negativos	-2

Essas recompensas e penalizações constantes no Quadro 6 foram estabelecidas de forma empírica de acordo com o conhecimento da estrutura de custos devidamente conhecida. Contudo esses pesos são subjetivos e podem ter configurações diferentes conforme a estrutura de custos de recuperação de um FIDC.

Toledo (2013) propõe um modelo de fluxo de recuperação líquida em que um dos parâmetros utilizados é um vetor de custo mensal que considera os atributos citados anteriormente, conforme a Equação 8:

$$C^*_{l.xn} = \rho * \mu^*_{l.xn} \qquad (8)$$

Na qual:

$C^*_{l.xn}$ = vetor de custo mensal; e

ρ = custo médio operacional por negociador por mês.

Com isso o modelo de fluxo de recuperação é definido, conforme a Equação 9:

$$\lambda^*_{l.xn} = \Psi^*_{l.xn} - c^*_{l.xn} \quad (9)$$

Na qual:

$\Psi^*_{l.xn}$ = Vetor de recuperação; e

$c^*_{l.xn}$ = Vetor de custos estimados.

Dessa forma, como a estrutura de custos do Quadro 6 foi aplicada de forma empírica foram usadas métricas de classificação com pesos artificiais para não somente validar o resultado das classificações, como também verificar o custo econômico de cada modelo devido a natureza financeira do estudo. Dessa forma, os pesos relativos às classificações são considerados:

TP (Verdadeiros Positivos) – Dívidas com alto potencial de recuperação classificadas corretamente, neste caso o custo econômico que bonifica o modelo é uma economia em termos de custo de **-1** unidade monetária;

FP (Falsos Positivos) – Dívidas com baixo potencial de recuperação classificadas incorretamente, neste caso o custo econômico que penaliza o modelo é **9** unidades monetárias;

FN (Falsos Negativos) – Dívidas com alto potencial de recuperação classificadas incorretamente, neste caso o custo econômico que penaliza o modelo é **4** unidades monetárias;

TN (Verdadeiros Negativos) - Dívidas com baixo potencial de recuperação classificadas corretamente, neste caso o custo econômico que bonifica o modelo é **-2** unidades monetárias.

A ferramenta escolhida para os experimentos foi o Weka versão 3.7.7 Developer Edition. O critério de escolha dessa ferramenta foi devido à facilidade de implementação e configuração das redes, e a familiaridade do autor com a ferramenta.

Os parâmetros de configuração de todas as RNAs utilizadas no estudo são apresentados no Quadro 7.

Quadro 7 - Descrição dos parâmetros utilizados para todas as cinco Redes Neurais Artificiais do experimento.

Parâmetro	**Tradução**	**Descrição**	**Valor do Parâmetro**
Decay	Decaimento	Decréscimo da taxa de aprendizado. Este parâmetro divide a taxa de aprendizado inicial pelo número de épocas para determinar qual deve ser a taxa de aprendizado correta. Este parâmetro ajuda a rede a parar de processar caso os resultados tenham uma divergência muito grande da variável de resposta.	False
Learning Rate	Taxa de Aprendizado	Taxa de aprendizado. Parâmetro de aprendizado que controla o tamanho dos pesos de atualização dos neurônios e as mudanças do *bias*durante o aprendizado da rede.	0,7

Parâmetro	Tradução	Descrição	Valor do Parâmetro
Training Time	Tempo de Treinamento	Número de épocas de treinamento da rede, *i.e.* a quantidade de vezes que todo o conjunto de treinamento irá passar na rede para o aprendizado.	5000
Seed	Semente	Número aleatório utilizado para dar os pesos iniciais na rede.	0
Validation Threshold	Limite de Validação	Limite de validação. Este limite determina a quantidade de vezes consecutivas que o conjunto de validação tem uma piora no aprendizado antes da parada da rede.	20
Momentum	Momento	Parâmetro que adiciona uma porção do peso anterior na atualização de pesos corrente. Este parâmetro serve para prevenir uma convergência muito rápida que por ventura venha a cair em um mínimo local. Isto é, este parâmetro serve para criar artificialmente um 'distúrbio' durante a convergência. Um *momentum* muito alto pode aumentar a velocidade de convergência, com o risco de se cair em um mínimo local, enquanto um valor baixo deixa a rede mais estável, no entanto deixando a velocidade de convergência extremamente mais baixa.	0,7

Para validação dos modelos foi escolhida a técnica de validação cruzada com o número de partições n=10. Essa técnica de validação de modelo tem como finalidade testar o modelo na fase de treinamento, e eliminar a necessidade de composições amostrais em arquivos diferentes. O particionamento é realizado de forma automática em que o conjunto de dados é dividido pelo número de partições n é determinado anteriormente ao experimento. Dessa forma ao escolher um número, por exemplo, n=5, o conjunto de dados é particionado em cinco partes,

em que quatro partes são utilizadas para treinamento do modelo e a parte restante serve para o teste do modelo.

Essa abordagem permite uma melhor generalização do modelo tendo em vista que como cada conjunto de dados é testado à medida que ocorre a fase de treinamento após o final de cada teste de todas as partições *n* tem-se a estimativa de como o modelo irá desempenhar na prática.

Esta abordagem de validação cruzada foi escolhida devido ao fato de que como as RNAs são métodos que buscam aproximações sucessivas até uma solução de acordo com aprendizado da rede para a tarefa de classificação; naturalmente esse método encontra sempre as melhores soluções possíveis o que coloca um elemento de incerteza que não pode ser quantificável. Sendo assim a validação cruzada através da estimativa após o cruzamento de todas as partições fará a exposição dessa incerteza (ou variância do modelo) de forma mais robusta do que se houvesse somente a fase de treino e testes realizadas de formas separadas.

Para maior aprofundamento da validação cruzada recomenda-se a leitura do trabalho de Hastie, Tibshirani, e Friedman (2009) em que são apresentadas as vantagens e desvantagens da validação cruzada, e também como a técnica auxilia na decomposição e redução da variância entre os modelos.

Caracterização da base de dados

A base de dados consiste em 19.845 registros relativos a crédito direto ao consumidor (CDC) e a créditos para pequenas e médias empresas já vencidos e não-performados. Como se tratam de dados reais, para a proteção da confidencialidade os nomes de alguns atributos foram modificados e informações relativas aos clientes foram eliminadas da base. A relação de atributos utilizados na amostra dos créditos está apresentada no Quadro 8.

A base de dados utilizada no experimento de Redes Neurais Artificias contém as seguintes informações relativas à quantidade de dívidas e o saldo correspondente, conforme a Tabela 3.

Tabela 3- Saldos e quantidade de contratos da base de dados para os experimentos com as RNAs distribuídos nas classes de dívidas pagas e não pagas.

Dívida Paga	**Saldo das Dívidas**	**Qtde Contratos**	**%Saldo das Dívidas**	**%Qtde Contratos**
Não	R$10.606.307,06	7.287	*39,29%*	*36,72%*
Sim	R$16.388.598,74	12.558	*60,71%*	*63,28%*
Total Geral	*R$26.994.905,80*	*19.845*	*100,00%*	*100,00%*

Os campos selecionados para compor a base de dados têm os seguintes nomes e descrições apresentados no Quadro 8.

Quadro 8 - Atributos e as respectivas descrições da base de dados utilizada para os experimentos com as RNAs.

Atributo	Tipo	Descrição
Saldo_Cliente	Monetário	Valor do contrato do cliente no momento da concessão do crédito.
Faixa_Idade	Numérico	Faixa de idade da dívida em meses.
Tipo_Credito	Nominal	Modalidade de crédito que foi concedido ao cliente. 1- Pessoa Física, 2-Pessoa Jurídica.
Mes_Contrato	Data	Número do mês em que o contrato foi celebrado.
Ano_Contrato	Data	Ano em que o contrato foi celebrado.
Estado	Binário	Estado no qual o cliente fez o contrato para a concessão de crédito.
Canal_Contato	Nominal	Canais disponíveis para o contato com o cliente. 0 - Nenhum, 1 - Telefone, 2 - Endereço, 3 - Telefone e Endereço, 4 - Telefone, endereço e SMS.
Canal_Localizacao	Nominal	Forma na qual foi obtido o contato com o cliente.
Restricao_Bureau	Binário	Se há indicação a órgãos de restrição de crédito como SPC ou Serasa.
Pago	Binário	Indica se a dívida foi quitada ou não.

Como apresentado no Quadro 8 as RNAs tiveram disponíveis para o aprendizado dados ligados a característica da dívida, informações geográficas, dados relativos a forma em que o cliente foi contatado, forma de contato e informações se a dívida foi paga ou não. Com esse conjunto de atributos esperou-se da rede uma boa capacidade de aprendizado dado também o volume de registros (*i.e.* contratos) contidos na Tabela 3.

Em relação à distribuição das dívidas dado o tipo de produto - cartão de crédito ou crédito para pequena e média empresa foram dispostos de acordo com a Tabela 4.

Tabela 4 - Saldos e quantidade de contratos da base de dados para os experimentos com as RNAs distribuídos no atributo tipo de crédito e posteriormente nas classes de dívidas pagas e não pagas.

Tipo de Crédito	Saldo das Dívidas	Qtde Contratos	%Saldo das Dívidas	%Qtde Contratos
CreditCard	R$26.381.342,60	19.741	*97,73%*	*99,48%*
Não	R$10.244.086,02	7.232	38,83%	36,63%
Sim	R$16.137.256,58	12.509	61,17%	63,37%
SME	R$ 613.563,20	104	*2,27%*	*0,52%*
Não	R$ 362.221,04	55	59,04%	52,88%
Sim	R$ 251.342,16	49	40,96%	47,12%
Total Geral	*R$26.994.905,80*	*19.845*	*100,00%*	*100,00%*

Como pode ser observado na Tabela 4 em uma pequena parte da base de dados contém créditos voltados para pessoa jurídica; porém, esses créditos representam menos de 3% do saldo total das dívidas analisadas.

A composição geográfica dos créditos, em relação à concentração de saldo pode ser verificada na Figura 8, em que uma maior concentração econômica está indicada nas áreas com menor intensidade de cores.

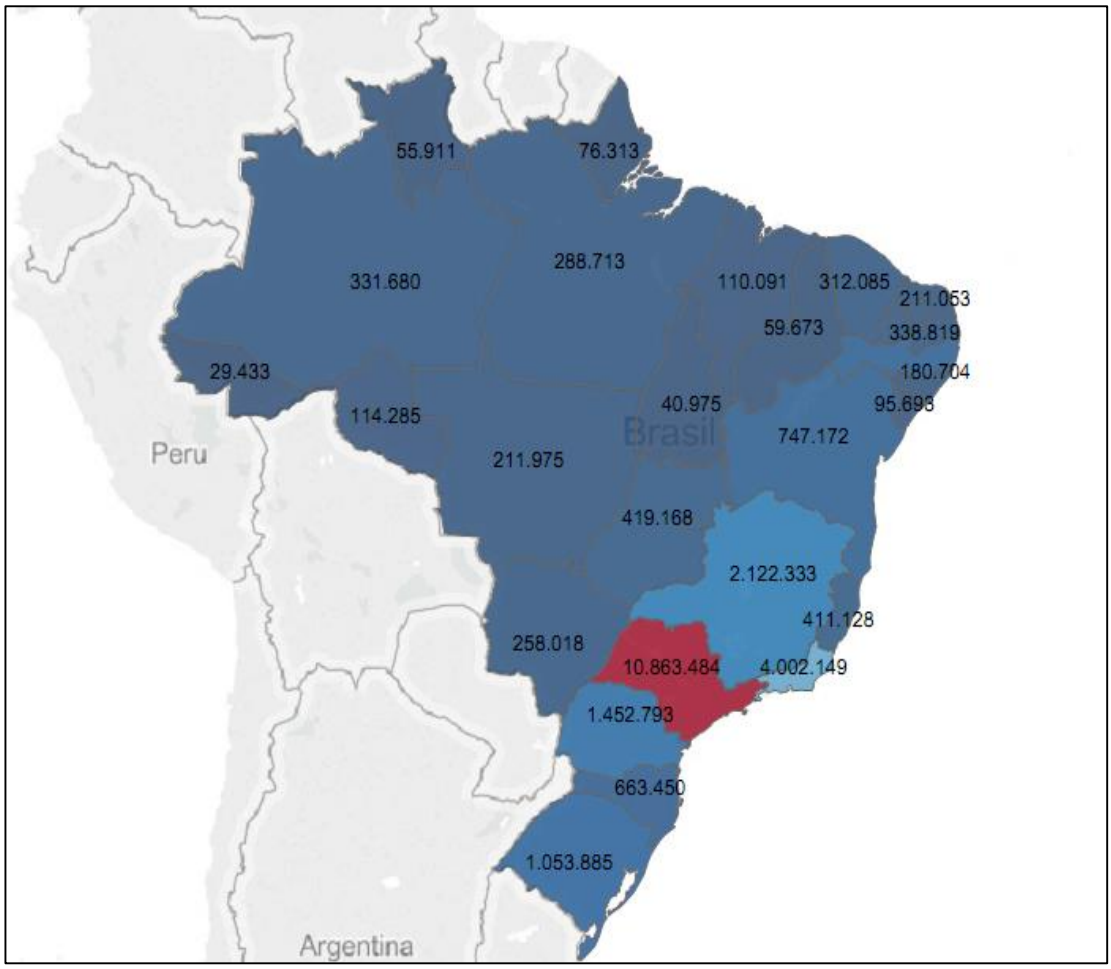


Figura 8 - Distribuição geográfica dos saldos dos NPLs relativa à base de dados utilizada para os experimentos com as RNAs. Fonte: Elaborada pelo Autor.

A Figura 8 mostra que grande parte do saldo constante na base de dados está concentrada na região sudeste e sul do Brasil, o que indica que possivelmente na composição da *tranche* não houve uma extração randomizada dos créditos, tendo em vista que a soma dos saldos não se aproxima de uma distribuição semelhante à composição populacional do Brasil.

A Figura 9 tem como principal função apresentar a evolução temporal dos créditos constantes na base de dados, tanto em termos de saldo quanto também em termos da data de criação desses débitos.

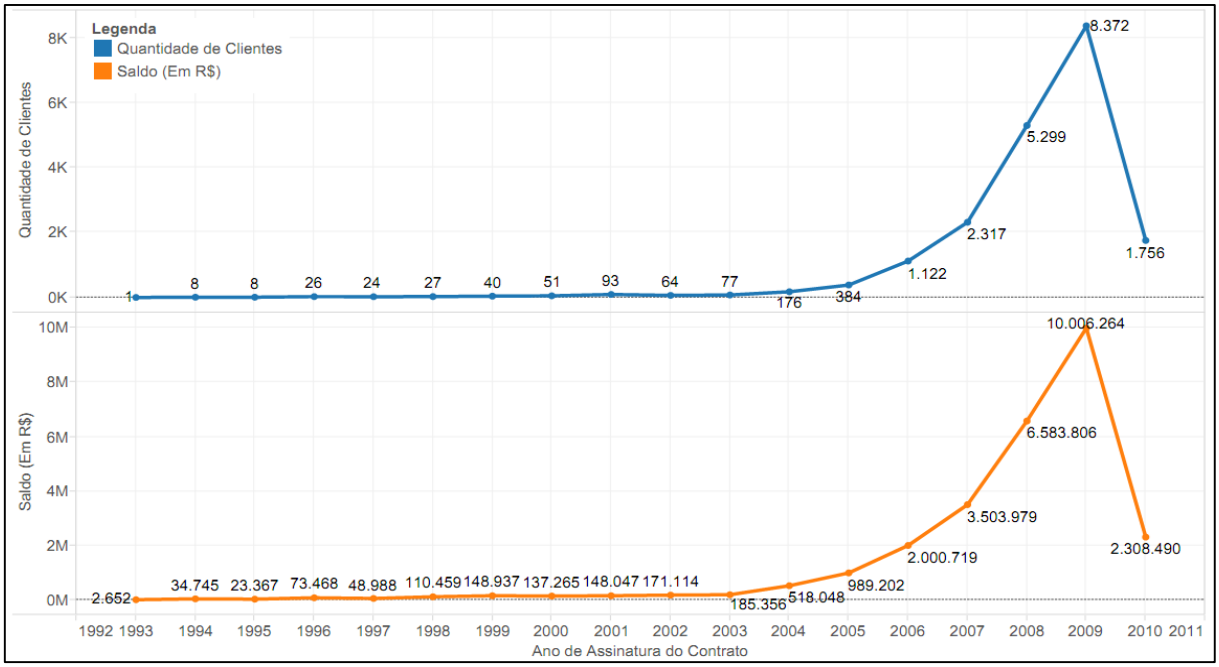


Figura 9 - Evolução temporal dos NPLs ao longo dos anos de acordo com a data de celebração dos contratos, relativa à base de dados utilizada para os experimentos com as RNAs. Fonte: Elaborada pelo Autor.

Na Figura 9 a curva laranja que representa o saldo das dividas no momento da celebração dos contratos mostra um comportamento exponencial aproximadamente na última parte do ano de 2007 com o pico no inicio do ano de 2008 e posterior queda na forma de exponencial negativa no qual após a metade do ano houve um arrefecimento menos abrupto que se manteve até 2009.

Uma característica importante relativa aos créditos NPL é a idade média da dívida, a qual uma dívida com uma idade maior tem um menor potencial de recuperação. Isso obriga os gestores de FIDCs há considerarem o tempo relativo a essas dívidas em seus modelos de precificação bem como administração do risco inerente à aquisição de uma dívida com baixa possibilidade de recuperação (International Finance Corporation, 2012).

3.5.2 Experimento 2: Experimento com *Self-Organizing Maps* conjuntamente com a Teoria dos *Rough Sets*

Para este experimento foram escolhidas duas técnicas que são os *Self-Organizing Maps* (SOM) e *Rough Sets*, em que a primeira técnica realizará a tarefa de geração de *clusters* e a segunda técnica será responsável para a extração de regras da base de dados.

O principal objetivo deste experimento é a extração de conhecimento de uma base de dados utilizando duas técnicas distintas, em que uma realizará a tarefa de apresentação de uma estrutura implícita dos dados através da segmentação ou geração de *clusters*, enquanto a outra técnica tem como finalidade descobrir relações e coocorrências entre as variáveis em uma base de dados e realiza atribuições de causa e efeito (*e.g.* Se... Então).

O resultado esperado com a rede SOM foi que a mesma realizasse a formação de agrupamentos dentro do conjunto de dados apresentados, de forma que estes conjuntos de dados com maior similaridade representassem o maior grau de coesão possível; e que com os seus mapas topológicos pudessem ser extraídas informações descritivas que revelassem a natureza dos ativos; para posterior formação de estratégias de recuperação ou descoberta de informações previamente desconhecidas.

Já resultado esperado com *Rough Sets* foi que a mesma realizasse a geração de regras a partir da base de dados, em que essas regras tivessem significância estatística sobre a base de dados, apresentassem informações relevantes que dessem subsídios para formulação ou subsidiar estratégias de recuperação, e que por fim fossem acionáveis, isto é, dessem suporte a ações que podem ser realizadas com o conhecimento extraído das regras.

Com a aplicação destas técnicas busca-se saber primeiramente qual é a estrutura implícita no conjunto de dados através da criação de *clusters* sem que haja uma hipótese previamente definida para subsidiar análises iniciais sobre estes dados, e após isso verificar padrões existentes através de relação entre as variáveis dos elementos, isto é a presença de regras implícitas nos dados.

Apesar de terem sido utilizadas duas técnicas para este experimento, a aplicação de cada uma dessas técnicas foi realizada de forma isolada, em que primeiramente foi aplicada a técnica SOM e posteriormente a Teoria dos *Rough Sets*.

Essa escolha ocorreu de forma arbitrária em que buscou-se extrair o máximo de conhecimento previamente desconhecido com as duas técnicas utilizando uma abordagem mais exploratória, em detrimento de uma abordagem baseada em hipóteses propriamente ditas.

Com isso a aplicação da técnica de SOM tem como meta responder as perguntas ‘Qual a estrutura dos *clusters* deste conjunto de dados?’, ‘Quais são as principais características de cada um dos *clusters*?’, enquanto a técnica de *Rough Sets* tem como finalidade apresentar as relações entre cada uma das transações do conjunto de dados e responder as perguntas ‘Quais são as principais regras deste conjunto de dados?’, ‘Quais regras possuem conhecimento relevante?’.

Com esta abordagem das técnicas espera-se que sejam descobertas conjecturas através do resultado da aplicação das técnicas que subsidiem a elaboração de estratégias de recuperação efetivas através da exposição dos padrões, seja através da composição dos *clusters* ou por meio da geração de regras de associação.

Os parâmetros utilizados para a geração das redes SOM estão apresentados no Quadro 9.

Quadro 9 - Parâmetros utilizados para configuração da rede SOM para geração dos mapas.

Parâmetro	**Tradução**	**Descrição**	**Valor do Parâmetro**
Number of nodes	Número de Nós	Este parâmetro determina a granularidade do mapa de forma quanto maior o número de nós, mais detalhado é o mapa.	1000
Map Size	Tamanho do Mapa	Este parâmetro estabelece o formato e o tamanho do mapa. As opções incluem desde mapas de formato quadrado até geração automática de acordo com o número de instâncias.	Automatic Map
Trainning Schedule	Rotina de Treinamento	Parametrização interna do processo de treinamento, em que um processo rápido leva a um resultado com uma acurácia menor, a opção *Accurate* leva a um resultado com um grau de acurácia maior; porém, com tempo de treinamento maior.	Accurate
Tension	Tensão	Especifica a tensão resultante do mapa. Quanto menor for atensão, mais o mapa se adapta ao conjunto de dados (*i.e.*quanto mais diferenças nos dados, mais os atributos convergem aos valores da média).	0,75

Parâmetro	Tradução	Descrição	Valor do Parâmetro
Data Records Used	Dados Utilizados	Quantidade de registros utilizados para o treinamento.	500
Clusted Method	Método de *Cluster*	O método de Ward que é a opção padrão começa a realizar a *cluster*ização em que cada nó individual forma um *cluster*. Dois *clusters* são fundidos em cada passo do algoritmo: aqueles que detêm uma distância mínima de acordo com uma medida de distância, neste caso a distância de Ward. Essa medida leva em consideração as distâncias e o posicionamento dos dois *clusters*. Mais detalhes a respeito dessa técnica podem ser encontrados em Ward (1963).	SOM-Ward *Clusters*

Uma importante observação sobre o Quadro 9 é que apesar do Viscovery Mine ser uma ferramenta de mercado e ter como vantagem a facilidade para o uso, algumas parametrizações como número de neurônios, opções de formato de vizinhança, formato do arranjo, e número de épocas não estão disponíveis no programa.

O trabalho de Stefanovic e Kurasova (2011) trata de algumas dessas limitações do Viscovery Mine, contudo os autores ressaltam que a ferramenta está em uma posição intermediária para o uso em geração dos SOM o que de forma alguma não invalida os experimentos.

A análise de dados foi realizada a partir dos mapas topográficos gerados com a determinação dos *clusters* de acordo com o valor da variável no parâmetro *Clusted Method.* Essa análise envolveu a verificação de concentração de quantidade de clientes e o total de saldo em cada *cluster*.

A análise de dados também foi feita nos mapas topográficos de acordo com os atributos *Dívidas Pagas, Saldo do Crédito na Data do Atraso, Saldo Principal das Dívidas sem a Implicação de Juros, Saldo da dívida na data de celebração do contrato,* e *Saldo Principal*.

Para a condução desses experimentos com a técnica de *Rough Sets* para a geração de regras, foi utilizado o software Rough Set Exploration System (RSES). Os parâmetros

escolhidos para este experimento para a aplicação dos *Rough Sets* para a extração de regras estão listados no Quadro 10.

Quadro 10 - Parâmetros utilizados para configuração dos *Rough Sets* para a extração de regras.

Parâmetro	Tradução	Descrição	Valor do Parâmetro
Reduct / Rule Choice	Reduto / Geração de Regras	Este parâmetro oferece a opção entre a geração de redutos (*i.e.* redução de atributos) ou geração de regras de associação.	Rules
Method	Método	Este parâmetro oferece a opção da escolha do algoritmo de cálculo. A opção *Genetic Algorithm* constrói todas as regras baseado no conceito dos algoritmos genéticos como mutação e *cross-over*.	Genetic Algorithm
Genetic Algorithm Settings	Parâmetros dos Algoritmos Genéticos	Parâmetro para cálculo das regras usando algoritmos genéticos. A opção *High Speed* realiza os cálculos de forma mais rápida, mas com menos acurácia.	High Speed

Nota-se pelo Quadro 10 que a ferramenta dispõe de configurações simples para a parametrização de extração de regras, em que o parâmetro *Method* é o que efetivamente realiza as operações de cálculo para esta extração.

Para maior aprofundamento nos parâmetros de cálculo do parâmetro *Method* este trabalho recomenda os trabalhos de Bazan e outros (2000) e Grzymała-Busse (1997); em que o primeiro trabalho apresenta informações sobre os algoritmos de busca exaustiva e algoritmos genéticos, e o segundo trabalho fala a respeito da técnica da regra de indução LERS.

A seleção dos parâmetros foi determinada por principalmente ao tempo de processamento e resultados mais satisfatórios do ponto de vista de geração de regras mais significativas.

A análise de dados foi realizada nas regras de decisão através do suporte das regras, *i.e.* a cobertura da regra ao maior número de instâncias possível. Após isso foi realizada uma seleção de regras que não tivessem um alto número de atributos covariantes.

Caracterização da base de dados

Para os experimentos usando *Rough Sets* e SOM, foi realizada a base de dados com os atributos contidos no arquivo de extração encontram-se no Apêndice A. Propositalmente, como

apresentado no Apêndice A foi escolhido um alto número de atributos, muitos desses atributos covariantes, para testar o poder discriminatório dos *Rough Sets*, em especial para a extração de regras com um alto número de variáveis.

A base de dados possuí as seguintes informações de saldos, e quantidade de contratos conforme a Tabela 5.

Tabela 5 - Saldos e quantidade de contratos da base de dados para os experimentos com a rede SOM e *Rough Sets* nas classes de dívidas pagas e não pagas.

Dívida Paga	**Saldo das Dívidas**	**Qtde Contratos**	**%Saldo das Dívidas**	**%Qtde Contratos**
Não	R$ 478.786,24	412	89,67%	82,40%
Sim	R$ 55.175,38	88	10,33%	17,60%
Total Geral	*R$533.961,62*	*500*	*100,00%*	*100,00%*

A distribuição em termos de pessoa física e jurídica e se o crédito foi quitado ou não está na Tabela 6:

Tabela 6 - Saldos e quantidade de contratos da base de dados para os experimentos com a rede SOM e *Rough Sets* distribuídos nas classes de dívidas pagas e não pagas e de acordo com o atributo categoria.

Dívida Paga	**Categoria**	**Saldo das Dívidas**	**Qtde Contratos**	**%Saldo das Dívidas**	**%Qtde Contratos**
Não	PF	R$ 474.521,08	408	99,11%	99,03%
	PJ	R$ 4.265,16	4	0,89%	0,97%
	Total	*R$478.786,24*	*412*	*89,67%*	*82,40%*
Sim	PF	R$ 52.446,90	86	95,05%	97,73%
	PJ	R$ 2.728,48	2	4,95%	2,27%
	Total	*R$ 55.175,38*	*88*	*10,33%*	*17,60%*
Total Geral		*R$533.961,62*	*500*	*100,00%*	*100,00%*

A distribuição geográfica dos saldos relativos à base de dados é apresentada na Figura 10.

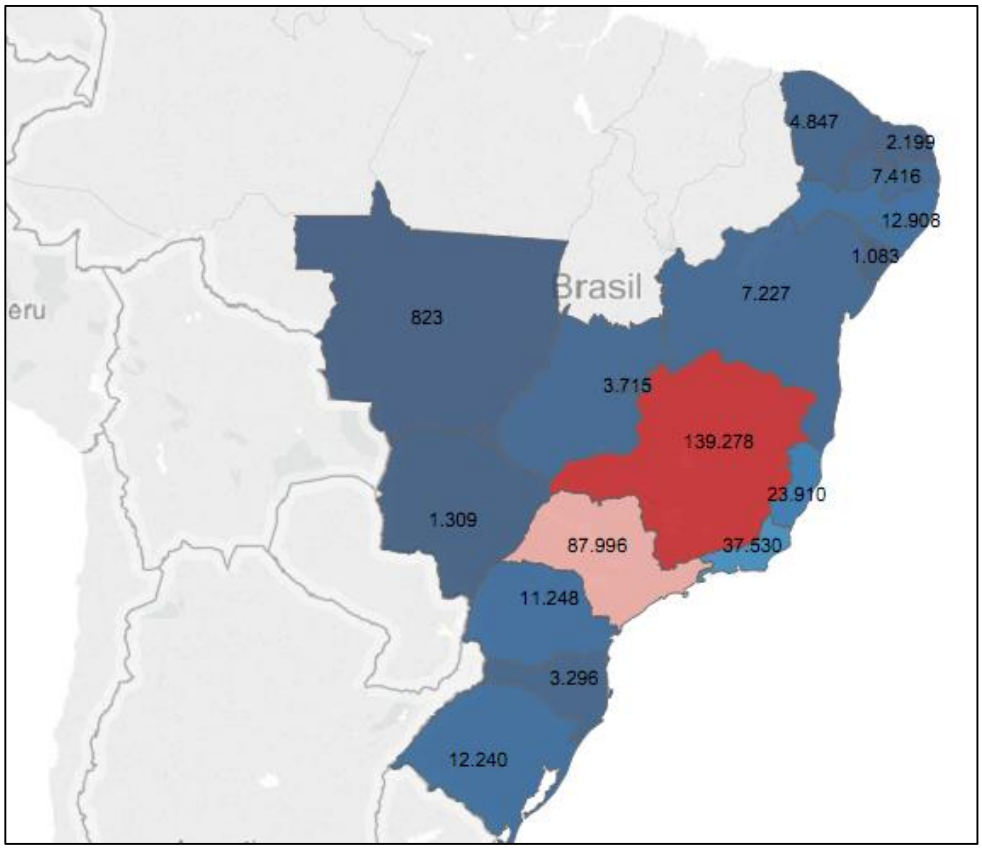


Figura 10 - Distribuição geográfica dos saldos dos NPLs relativa à base de dados utilizada para os experimentos com a rede SOM e *Rough Sets*. Fonte: Elaborada pelo Autor.

Na Figura 10 fica evidente que há dois vieses amostrais evidentes na seleção das dívidas em que (i) a região norte não contém nenhum estado com dívidas na base de dados, e (ii) há uma forte concentração de saldos no estado de Minas Gerais, o que indica que pode ter havido uma escolha de créditos desta região específica em detrimento das demais regiões.

A Figura 11 é apresentada a evolução temporal em termos de saldo e quantidade de contratos celebrados ao longo do tempo de vida desse conjunto de dívidas, em que a linha azul representa o número de dívidas e a linha laranja representa o saldo dessas respectivas dívidas.

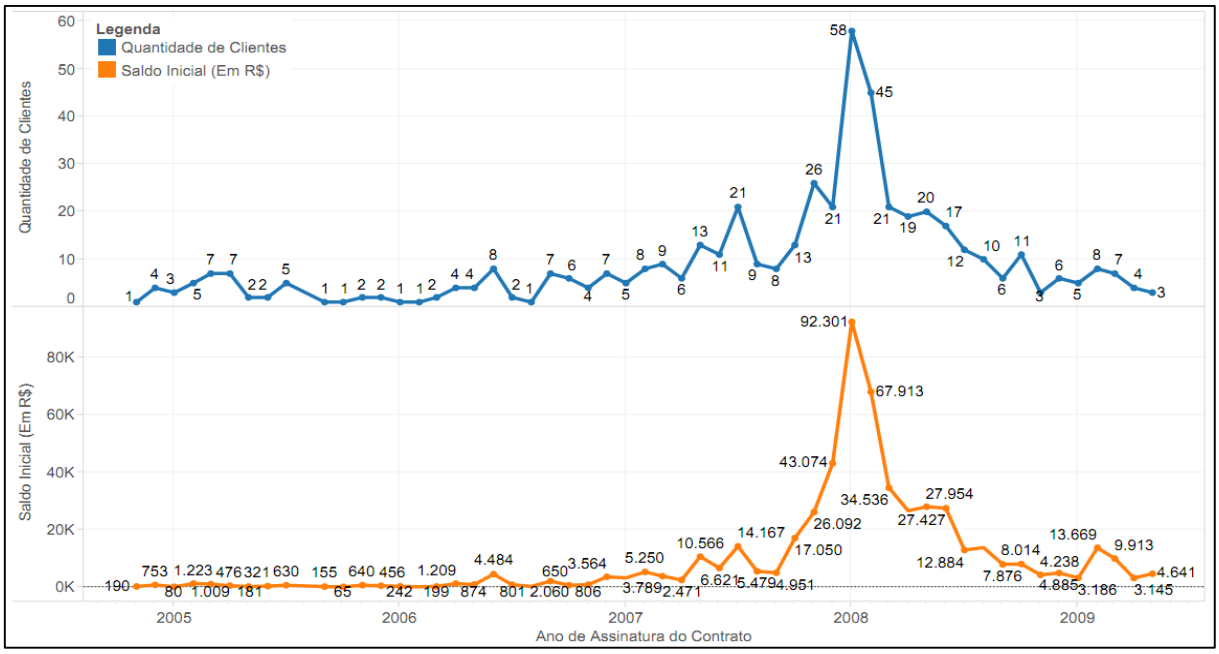


Figura 11 - Evolução temporal dos NPLs ao longo dos anos de acordo com a data de celebração dos contratos, relativa à base de dados utilizada para os experimentos com a rede SOM e *Rough Sets*.
Fonte: Elaborada pelo Autor.

Na Figura 11 a exemplo do que ocorreu na Figura 9 houve um crescimento exponencial do saldo total das dívidas na parte final do ano de 2007 atingindo o seu pico no início de 2008 com um decaimento vertiginoso até os primeiros meses do ano e após isso um arrefecimento até em meados de 2009 quando há um pequeno pico.

Uma observação que pode ser feita sobre a Figura 9 e a Figura 11 é que mesmo sendo bases de dados de diferentes cedentes, e com amostragens distintas é que o ano de 2007 e início de 2008 compõem a maioria das dívidas contidas nestes dois portfólios. Essa congruência de tendências temporais relativas à data de celebração dos contratos mostra indícios que as duas instituições realizaram as suas operações de *Write-Off* de forma seletiva e com mais intensidade nestes períodos.

3.5.3 Experimento 3: Experimento com a Teoria dos *Rough Sets* conjuntamente com as Árvores de Decisão

A escolha dos *Rough Sets* para este problema foi testar a capacidade de escolha de variáveis quando submetida a uma base de dados com diversos atributos, muitos dos quais são covariantes; e após essa escolha de variáveis essa base de dados foi submetida a um classificador de Árvore de Decisão.

O objetivo principal deste experimento é realizar a aplicação de *Rough Sets* em conjunto com a técnica de árvore de decisão, em que a primeira técnica fará a redução de atributos de uma base de dados, e a segunda técnica realiza a tarefa de classificação através da geração de expressões disjuntas.

Para efeitos de contraste será feito também um experimento em que a base de dados com todas as variáveis será submetida ao classificador de árvores de decisão. Esse contraste consiste em analisar o tempo de execução das árvores para medir o desempenho computacional do algoritmo dependendo do volume de atributos, e também realizar o contraste em termos de acurácia do modelo, isto é, a capacidade de classificações corretas geradas pelo algoritmo.

Para testar de forma exaustiva os *Rough Sets* serão utilizados os dois algoritmos, a geração de redutos via algoritmo de Johnson e a geração de redutos utilizando características dos algoritmos genéticos.

A escolha por uma abordagem conjunta neste experimento foi feita para verificar se um grande número exerce influência nos modelos de classificação, seja em questões computacionais, *i.e.* como o algoritmo de Árvore de Decisão é influenciável por essas variáveis, e se os modelos realizam melhores classificações com um número alto de classificações corretas a despeito do seu tempo de processamento.

Dessa forma escolheu-se apenas a realização de modificações no que diz respeito à escolha do método na aplicação da técnica de *Rough Sets* para geração de redutos. Foram escolhidos como métodos de geração de redutos o Algoritmo de Johnson, e os Algoritmos Genéticos ambos com os seus valores de configuração padrão.

A validação empírica de qual dos três modelos têm o melhor desempenho em relação à classificação será feita através de três formas: (a) abordagem sensível ao custo, (b) tempo total de processamento do modelo e (c) métricas de avaliação de modelos como taxa de falsos positivos, taxa de falsos negativos, taxa de verdadeiros positivos (Sensibilidade – Erro Tipo II), verdadeiros negativos (Especificidade – Erro Tipo I), acurácia, Coeficiente de Correlação de Matthews (1975); esta última métrica que foi utilizada devido ao desbalanceamento em termos de proporções das classes de predição.

A abordagem sensível ao custo baseia-se na estrutura de recompensa e penalização de classificadores, em que de acordo com diferentes qualificações de erros a rede sofrerá uma penalização com a atribuição de uma pontuação negativa, e conforme o classificador tenha diferentes acertos haverá em contrapartida atribuição de pontuações positivas.

O tempo de processamento vai levar em consideração o tempo total de processamento dos modelos e a sua acurácia final para estabelecer relação entre a redução de variáveis e o método de geração de redutos de acordo com a acurácia final dos modelos.

Por fim a avaliação via as métricas de avaliação de modelos fornecerão subsídios sobre a qualidade dos modelos, isto é, a sua qualidade em termos de gerar classificações corretas ou mesmo minimizar erros mais custosos, qualificações essas que podem subsidiar a escolha de um classificador ou com mais acertos na variável dependente positiva, *i.e.* maximizar acertos, ou um classificador que cometa menos erros na variável negativa, *i.e.* minimizar custos dos erros de predição.

Para determinar estes custos foi adotada a matriz de confusão tal como proposta por Kohavi e Provost (1998) que informa os valores previstos em relação aos valores atuais após a classificação.

A tabela de custos utilizada será a mesma que deu subsídios para os experimentos com as RNAs e é apresentada no Quadro 6. Essa tabela foi elaborada de forma empírica, já que este tipo de operação de estruturação contém inúmeras variáveis e dinâmicas os quais podem aumentar ou diminuir os valores determinados na tabela de custos.

Quadro 11 - Tabela de custos do modelo.

	Parâmetros de Custos	
Custo/Benefício do Modelo	**TP – Verdadeiros Positivos**	-1
	FP – Falsos Positivos	9
	FN – Falsos Negativos	4
	TN – Verdadeiros Negativos	-2

Dessa forma, os pesos relativos das classificações serão exatamente aos do capítulo de RNAs em que:

TP (Verdadeiros Positivos) – Dívidas com alto potencial de recuperação classificadas corretamente, neste caso o custo econômico que bonifica o modelo é uma economia em termos de custo de **-1** unidades monetárias;

FP (Falsos Positivos) – Dívidas com baixo potencial de recuperação classificadas incorretamente, neste caso o custo econômico que penaliza o modelo é **9** unidades monetárias;

FN (Falsos Negativos) – Dívidas com alto potencial de recuperação classificadas incorretamente, neste caso o custo econômico que penaliza o modelo é **4** unidades monetárias;

TN (Verdadeiros Negativos) - Dívidas com baixo potencial de recuperação classificadas corretamente, neste caso o custo econômico que bonifica o modelo é **-2** unidades monetárias.

As ferramentas escolhidas para os experimentos foram para a geração dos redutos utilizando a técnica de *Rough Sets* o *software* Rosetta em um ambiente Windows XP; e para geração das árvores de decisão foi utilizado o software IBM SPSS versão 20. Os critérios de escolha dessas ferramentas foram devido à facilidade de geração de redutos, qualidade dos gráficos das árvores de decisão que consequentemente facilita a leitura das mesmas, e a familiaridade do autor com as ferramentas.

Os parâmetros do experimento com *Rough Sets* utilizando o Método de Johnson estão contidos no Quadro 12.

Quadro 12 - Parâmetros utilizados para configuração dos *Rough Sets* para redução de atributos utilizando o Algoritmo de Johnson.

Parâmetro	**Tradução**	**Valor do Parâmetro**
Discernibility	Discernibilidade	Full
Modulo Decision	Módulo de Decisão	True
Discernibility predicate	Predicado de indiscernibilidade	False
Memory usage	Uso de memória	False
Approximate solutions	Soluções de aproximação	Compute approximate solutions
Hitting fraction	Fração de quebra para aproximação	0,95

Os parâmetros do experimento com *Rough Sets* utilizando o método de Algoritmos Genéticos estão no Quadro 13.

Quadro 13 - Parâmetros utilizados para configuração dos *Rough Sets* para redução de atributos utilizando os Algoritmos Genéticos.

Parâmetro	**Tradução**	**Valor do Parâmetro**
Discernibility	Discernibilidade	Full
Modulo Decision	Módulo de Decisão	True
Discernibility predicate	Predicado de Discernibilidade	False
Memory usage	Uso de memória	False
Approximate solutions	Soluções de aproximação	Compute approximate solutions

Parâmetro	**Tradução**	**Valor do Parâmetro**
Algorithm Variation	Variação do algoritmo	Modified
Options	Opções	Boltzmann scaling of fitness function
Temperature Range	Intervalo de temperatura	[6.45 - 1.45]
Delta	Delta	0.02
Sample parents with replacement	Amostra com elementos pai com substituição	True
Use elitism	Uso do elitismo	True
Crossover probability	Probabilidade de crossover	0.7
Mutation probabiliy	Probabilidade de mutação	0.07
Inversion probability	Probabilidade de Inversão	0.05
Nr. crossover points	Número de pontos de crossover	1
Nr. Mutations on an individual	Número de mutações sobre um indivíduo	1
Nr. Transpositions for inversion	Número de transposições para inversão	1
Nr. Generations to wait for fitness to improve	Número de gerações de espera para adaptações de melhora	70
Stop if average population fitness does not improve	Critério de parada da adaptação se a população média melhorar	True
Stop if keep list does not change	Critério de parade se a lista de indivíduos não mudar	True
Population size	Tamanho da população	70
Size of keep list	Tamanho da lista de indivíduos	256
Weighting between subset cost and hitting fraction	Peso ente um subconjunto de custo e a função de quebra para aproximação	0.4
Incorporate attribute cost information	Incorporação da informação sobre o atributo de custo	False
Random number generator seed	Semente do número aleatório	54321

Parâmetro	Tradução	Valor do Parâmetro
Compute approximate solutions	Computar soluções de aproximação	False

A parametrização das árvores de decisão para os três experimentos é apresentada no Quadro 14:

Quadro 14 - Parâmetros utilizados para geração das três Árvores de Decisão.

Parâmetro	Tradução	Descrição	Valor do Parâmetro
Tree Display	Apresentação da Árvore	Forma de visualização do nó raiz da árvore de decisão.	Top Down
Nodes	Nós	Informa se apresenta as estatísticas nos nós folha.	Statistics
Branch Statistics	Estatísticas dos Nós	Informa se apresenta as estatísticas nos nós, isto é, se apresentam os valores de decisão para cada uma das ramificações.	Yes
Growing Method	Método de Crescimento da Árvore	Método de crescimento da árvore de decisão.	CHAID
Minimum Cases in Parent Node	Número mínimo de casos no nó antecessor	Número de casos mínimos em um nó antecessor, isto é, nó pai.	100
Minimum Cases in Child Node	Número mínimo de casos no nó sucessor	Número de casos mínimos em um nó sucessor, isto é, nó filho.	50
Validation Type	Método de validação	Tipo de validação dos resultados pelo algoritmo.	None
Allow splitting of merged categories	Permitir redivisão de categorias	Permite a redivisão de categorias que tiveram algum tipo de processo de junção anteriormente.	True
CHAID Alpha Split	Critério de divisão	Critério de divisão do algoritmo de acordo com o p-valor da categoria.	0.05

Parâmetro	Tradução	Descrição	Valor do Parâmetro
Alpha Merge	Parâmetro de junção	Parâmetro que determina o critério de junção de categorias caso o p-valor seja menor do que o estabelecido.	0.05
Split Merged	Divisão de junções	Se a opção *Allow splitting of merged categories*estiver como *True*se o p-valor dessa categoria for menor ou igual ao parâmetro o algoritmo irá dividir a respectiva categoria.	No
Chi-Square	Qui-quadrado	Tipo de estatística para cálculo do Chi-quadrado	Pearson
Dependent Variable	Variável dependente	Variável dependente	Dummy_pago
Maximum Tree Depth	Profundidade máxima da árvore	Profundidade máxima da árvore de decisão.	3

As parametrizações do Quadro 12, Quadro 13 e Quadro 14 são as configurações padrão das ferramentas de estudo. O objetivo principal com essa escolha foi observar o poder do método de aplicação em conjunto dessas técnicas em detrimento de uma parametrização mais detalhada. Refinamentos de parametrização podem ser realizados, contudo isto ficou de fora do escopo deste trabalho.

O método CHAID foi escolhido para as experimentações devido ao fato de que esse tipo de árvore de decisão tem facilidade para lidar com dados categóricos, e também pelo fato de que o seu algoritmo conta com a incorporação de uma heurística em termos de adaptação dos testes de Chi-Quadrado para formação das classes de decisão e divisão dos nós; o que contribuí em termos de tempo de processamento do algoritmo.

As análises que serão realizadas serão de natureza de contraste entre a geração das árvores de decisão com todas as variáveis, em comparação com as árvores que por ventura tenham um conjunto de atributos escolhidos através da geração de redutos provenientes da técnica de *Rough Sets* com os métodos de Algoritmo de Johnson e Algoritmos genéticos.

Caracterização da base de dados

Para as experimentações realizadas foi realizada a extração de uma base de dados com os seguintes atributos descritos no Apêndice B. Intencionalmente conforme é mostrado no Apêndice B foi escolhido um alto número de atributos para testar o poder de redução de atributos dos *Rough Sets* para contraste dos experimentos.

Em relação aos saldos e quantidade de dívidas essas informações são apresentadas na Tabela 7.

Tabela 7 - Saldos e quantidade de contratos da base de dados para os experimentos *Rough Sets* e Árvores de Decisão distribuídos no atributo tipo de crédito e posteriormente nas classes de dívidas pagas e não pagas.

Dívida Paga	**Saldo das Dívidas**	**Qtde Contratos**	**%Saldo das Dívidas**	**%Qtde Contratos**
Não	R$3.131.771,56	4.898	70,09%	73,63%
Sim	R$1.336.465,45	1.754	29,91%	26,37%
Total Geral	*R$4.468.237,01*	*6.652*	*100,00%*	*100,00%*

Na Figura 12 é apresentada a evolução temporal dos saldos e da quantidade de contratos celebrados desse conjunto de dívidas, em que a linha azul representa o saldo das dívidas e a linha laranja a quantidade de contratos celebrados.

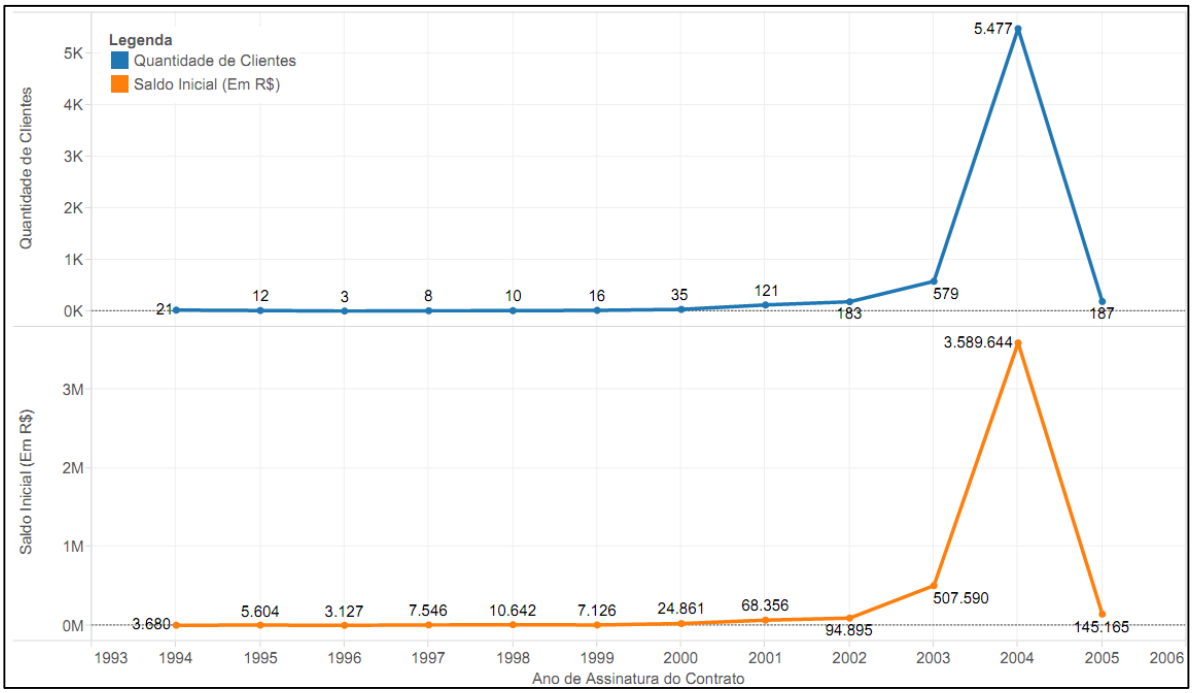


Figura 12 - Evolução temporal dos NPLs ao longo dos anos de acordo com a data de celebração dos contratos, relativa à base de dados utilizada para os experimentos com *Rough Sets* e Árvores de Decisão. Fonte: Elaborada pelo Autor.

Diferentemente da Figura 9 e Figura 11 quando foi analisada a distribuição temporal dos NPLs de acordo com a data de celebração dos contratos toda a série é anterior ao ano de 2005; ano este que foi o primeiro das séries da Figura 9 e Figura 11. Isso pode indicar que essa instituição atrasou a cessão desses créditos por diversos motivos como, por exemplo, houve uma estratégia clara para a liquefação desses ativos dentro desse tempo, falta de prioridade para o tratamento dos NPLs, questões de cunho operacional, etc.

Na Figura 13 pode-se verificar a concentração demográfica dos créditos em relação ao valor financeiro, no qual em vermelho verifica-se uma maior concentração e em azul uma menor concentração.

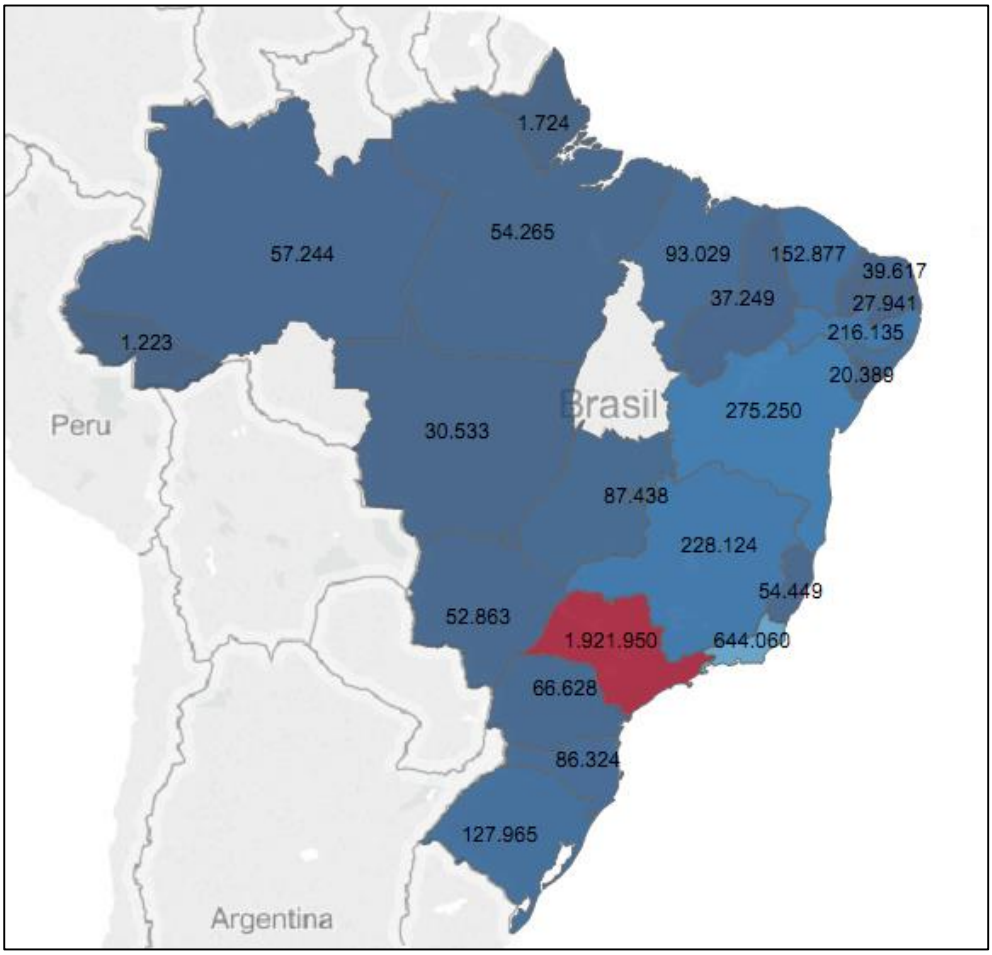


Figura 13 - Distribuição geográfica dos saldos dos NPLs relativa à base de dados utilizada para os experimentos com *Rough Sets* e Árvores de Decisão. Fonte: Elaborada pelo Autor.

De acordo com a distribuição geográfica em relação ao saldo dos créditos, nota-se que pode ter havido um viés amostral que eliminou os estados de Rondônia, Roraima e Tocantins.

No próximo capítulo será apresentada a forma na qual os experimentos foram realizados.

4 REALIZAÇÃO DOS EXPERIMENTOS E DISCUSSÃO DOS RESULTADOS

Neste capitulo apresenta-se a realização dos experimentos e a discussão dos resultados. Como foi descrito no capitulo 3 foram realizados três experimentos:

- Experimento 1: Experimento com Redes Neurais Artificiais;
- Experimento 2: Experimento com *Self-Organizing Maps* conjuntamente com a Teoria dos *Rough Sets*; e
- Experimento 3: Experimento com a Teoria dos *Rough Sets* conjuntamente com as Árvores de Decisão

4.1 REALIZAÇÃO DOS EXPERIMENTOS

4.1.1 Experimento 1: Experimento com Redes Neurais Artificiais

Nesta seção serão apresentados os parâmetros e a descrição da forma de como os experimentos com as RNAs foram conduzidos.

A topologia da MLP foi a seguinte: neurônios na camada de entrada, número de camadas ocultas, número de neurônios em cada camada oculta, número de neurônios na camada de saída e número máximo de iterações (épocas).

Os valores das parametrizações estão dispostos no Quadro 15.

Quadro 15 - Parâmetros utilizados para configuração das cinco RNAs.

Parâmetro	**Tradução**	**Valor do Parâmetro**
Decay	Decaimento	False
Learning Rate	Taxa de Aprendizado	0.7
Training Time	Tempo de Treinamento	5000
Seed	Semente	0
Validation Threshold	Limite de Validação	20
Momentum	Momento	0.7

Para este experimento foram construídos cinco modelos nos quais cada modelo conta com diferenças de arquitetura como apresentado no Quadro 16.

Quadro 16 - Parâmetros das topologias das RNAs utilizadas nos experimentos.

Arquitetura das Redes					
Modelo	**(#) Variáveis de Entrada**	**Amostragem**	**# Camada Escondida 1**	**# Camada Escondida 2**	**(#) Variáveis de Saída**
M1	10	Validação Cruzada *n*=10	6	- - - -	1
M2	10	Validação Cruzada *n*=10	10	- - - -	1
M3	10	Validação Cruzada *n*=10	2	- - - -	1
M4	10	Validação Cruzada *n*=10	10	5	1
M5	10	Validação Cruzada *n*=10	12	5	1

O objetivo das mudanças arquiteturais como apresentadas no Quadro 16 foi de verificar a capacidade de aprendizado da rede, dados os mesmos parâmetros, e verificar o aprendizado de cada rede de acordo com características de mudanças na arquitetura.

Cabe ressaltar que essas topologias foram escolhidas de maneira arbitrária no momento dos experimentos. Sendo as RNAs uma técnica não-determinísca, não há ponto pacífico relativo à melhor forma de se escolher uma arquitetura ou até mesmo critérios de parada em relação a convergência da rede. No entanto nos trabalhos de Koehn (1994) e de Priddy e Keller (2005) são apresentadas algumas heurísticas práticas que podem dar suporte à escolha de um método em específico.

Na Figura 14 foi apresentado o processo de aprendizado da rede durante o período de treinamento dos cinco modelos considerando erro médio por época.

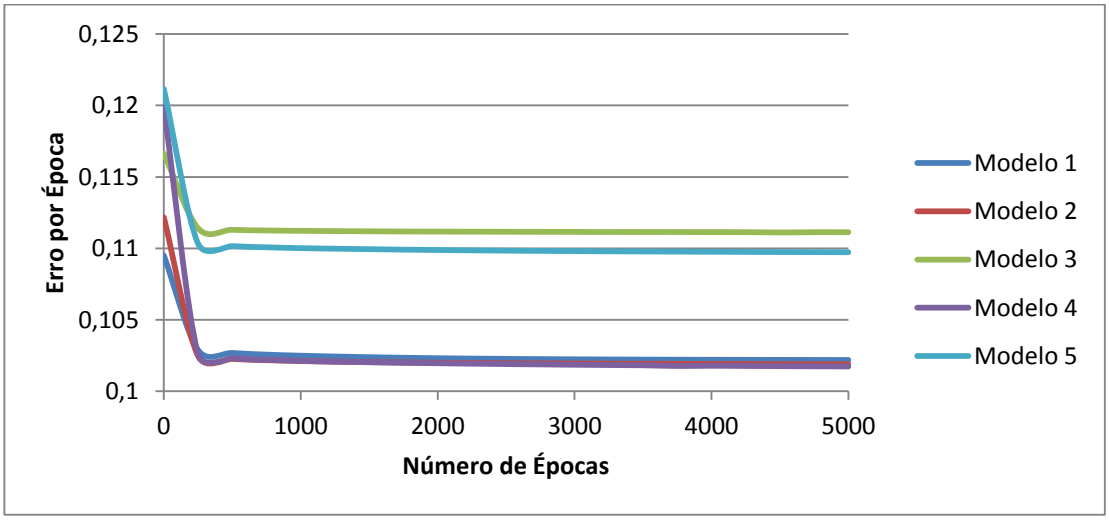


Figura 14 - Distribuição do erro médio ao longo das épocas durante a etapa de treinamento. Fonte: Elaborada pelo Autor.

A Figura 14 mostra que os modelos M3 e M5 tiveram os maiores índices de erros médios por época durante o treinamento da rede. Somente com essas informações não são possíveis maiores análises sobre o comportamento da rede, dado que esses dois modelos possuem diferenças de topologia em que o M3 tem apenas dois neurônios em uma camada escondida, e o M5 possuí duas camadas escondidas a primeira com 12 neurônios e a segunda com 5 neurônios.

Já a Tabela 8 apresenta o resultado da validação dos modelos da MLP na classificação dos créditos.

Tabela 8 - Resultados dos experimentos de acordo com as métricas de avaliação de modelos.

Tabela de Desempenho					
Modelo	**%Erro C-Positiva**	**%Erro C-Negativa**	**%Erro**	**%Acurácia**	**%Precisão**
M1	21.84%	9.28%	14.24%	85.76%	84.60%
M2	21.31%	9.81%	14.26%	85.74%	83.52%
M3	22.51%	8.78%	14.32%	85.68%	85.62%
M4	22.35%	**8.63%**	14.17%	85.83%	**85.88%**
M5	**20.61%**	9.72%	**13.90%**	**86.10%**	83.57%

A Tabela 9 apresenta os resultados dos experimentos de cada uma das RNAs de acordo com algumas métricas de avaliação de modelos.

Tabela 9 - Resultados dos experimentos de acordo com as métricas de avaliação de classificadores.

Tabela de Desempenho					
Modelo	**%Sensibilidade**	**%Especificidade**	**Acertos**	**Erros**	**Kappa**
M1	78.16%	90.72%	4,595	763	0.6980
M2	78.69%	90.19%	4,594	764	0.6963
M3	77.49%	91.22%	4,591	767	0.6978
M4	77.65%	**91.37%**	4,599	759	0.7010
M5	**79.39%**	90.28%	**4,613**	**745**	**0.7033**

Com os parâmetros estabelecidos no Quadro 11, os resultados para cada modelo em termos de custos econômicos estão escritos na Tabela 10.

Tabela 10 - Resultado final dos custos de cada modelo através da abordagem sensível ao custo.

Modelo	TP	FP	FN	TN	Custo do Modelo	Posição Geral do Modelo
M1	1,653	301	462	2,942	5,844	3º
M2	1,632	322	442	2,962	**18,054**	5º
M3	1,673	281	486	2,918	9,048	4º
M4	**1,678**	**276**	483	2,921	**-435**	1º
M5	1,633	321	**424**	**2,980**	2,980	2º

4.1.2 Discussão dos Resultados

No experimento do modelo 1 verificou-se que a taxa de acertos foi a 2ª maior (4,595). Contudo, o índice Kappa apresentou o pior desempenho e o seu custo econômico que foi o 3º maior (5,844). Com estes resultados o modelo 1 apresentou um resultado intermediário entre todos os modelos envolvidos nos experimentos.

No experimento do modelo 2 teve como principal característica o maior custo econômico em unidades monetárias (18,054), dentre todos os modelos. Isso pode ser explicado devido à alta penalização da MLP em termos de falsos positivos (322) e a alta porcentagem de erros na classe negativa (9,81%). Mesmo com a 2ª maior acurácia (85.74 %) a sua precisão apresentou o pior desempenho.

Isso possibilita ver que erros na classe negativa são fatores determinantes no mau desempenho desse modelo levando a indução que, antes de acertar a melhor dívida para ser cobrada é preciso minimizar erros decorrentes de má classificação, que são mais custosos.

No experimento do modelo 3 foi apresentado o maior erro relativo à classe positiva (22.51%) e o segundo menor erro na classe negativa (8.78%). Mesmo com o maior índice de erros (14,32%) o custo econômico ficou em 2º lugar em pior desempenho (9,048). Este modelo apresentou também a menor sensibilidade, isto é, a menor capacidade de identificar resultados positivos.

Com muitos erros na classificação negativa, este modelo teve um desempenho ruim em relação aos experimentos realizados devido à baixa generalização na classificação de dívidas com potencial de recuperação.

O experimento do modelo 4 apresentou o melhor custo (-435) mesmo com o segundo maior erro na classe positiva (22,35%). Outro fato é que esse modelo teve o menor erro na classe negativa (8,63%) e a maior precisão (85.88 %).

O melhor resultado obtido nos experimentos por esse modelo explica-se pela baixa penalização do classificador devido ao baixo índice de erros, e a precisão nos acertos da classe negativa. Isso possibilita ver que além de minimizar custos relativos a erros de falsos positivos, os acertos relativos aos créditos ruins são determinantes em termos de custos; o que leva a crer que este é o melhor modelo por induzir uma situação de minimização de custos.

Finalmente no experimento do modelo 5, foi apresentado o melhor índice de Kappa (0.7033), que indica que a classificação foi satisfatória, bem como o menor erro na classe positiva (20,61%) e o menor erro no geral (13.90%). Este modelo também teve a maior sensibilidade (79.39%), o que indica a sua boa capacidade de identificar casos da classe positiva.

Pode-se concluir com os resultados desse modelo que mesmo com as melhores classificações por época (índice de Kappa), e a melhor capacidade de identificação de positivos; os erros que implicam em penalizações em termos de custos econômicos são fator fundamental no desempenho desse modelo.

Verificou-se também nos resultados que os modelos que tiveram duas camadas escondidas (M4, e M5) obtiveram um melhor desempenho em relação os modelos que continham apenas uma camada escondida

O modelo 5 teve um percentual menor de erros na classe negativa, enquanto o modelo 4 teve um percentual menor de erros na classe positiva. Isso indica que o modelo 5 pode ser utilizado para subsidiar uma estratégia que tenha como foco a identificação de potenciais pagadores dos NPLs; enquanto o modelo 4 se enquadra em uma estratégia de recuperação em que busca-se minimizar a identificação errada de clientes que irão realizar o pagamento.

A métrica de acurácia que mede a proporção dos resultados genuinamente verdadeiros indica que o modelo M5 obteve o melhor desempenho na identificação correta se os clientes pagaram ou não a dívida durante o aprendizado. Já a métrica de precisão que é a proporção entre verdadeiros positivos e o total de resultados positivos (negativos ou não) indica que o modelo M4 obteve um melhor desempenho neste aspecto.
Estas duas métricas estão intrinsicamente ligadas e os resultados indicam que ao passo que a acurácia indica melhores classificações corretas, a métrica de precisão mostra uma maior capacidade de acertos nas classificações de uma forma mais consistente.

Na Tabela 9 são apresentados os percentuais de Sensibilidade e Especificidade dos modelos. O modelo M5 apresentou o melhor percentual em Sensibilidade, o que significa que esse modelo foi o melhor na identificação dos elementos verdadeiramente positivos. Já o

modelo M4 apresentou o melhor percentual em Especificidade, que indica que este modelo é o melhor na identificação das classificações verdadeiramente negativas.

Com os resultados de Sensibilidade e Especificidade, podemos afirmar que o modelo M5 pode ser utilizado em estratégias de recuperação voltadas a identificação imediata de potenciais pagadores dos créditos, ou mesmo quando o ônus econômico-financeiro de não se identificar um cliente em potencial é alto. O modelo M4 pode ser utilizado em estratégias de recuperação em que a identificação errônea de um cliente que por ventura não pague o crédito incorre em custos altos até mesmo causando prejuízo, ou mesmo em situações em que a estratégia tenha que obter o máximo de informações corretas antes de qualquer tipo de ação prévia.

Cabe ressaltar que esse equilíbrio entre sensibilidade e especificidade deve ser encontrado de acordo com a estratégia, esta que por ventura pode ter formas de ações diferentes de acordo com as qualidades e deficiências do modelo adotado.

Dessa forma, os resultados obtidos confirmam que o modelo 4 apresentou o melhor desempenho por meio do balanceamento do número de neurônios em duas camadas escondidas (10 neurônios na primeira camada e 5 neurônios na segunda camada), o que o coloca como o melhor modelo dentre os cinco na classificação de créditos NPL.

Conclui-se, então, que a MLP pode ser considerada uma importante ferramenta na análise e classificação em bases de dados heterogêneas, como a de créditos não-performados, sendo capaz de criar um modelo de generalização robusto para análises de NPLs.

4.1.3 Experimento 2: Experimento com *Self-Organizing Maps* conjuntamente com a Teoria dos *Rough Sets*

Nesta seção serão apresentados os parâmetros e a descrição da forma de em que os experimentos com os *Self-Organizing Maps* e a Teoria dos *Rough Sets* foram conduzidos.

Os parâmetros utilizados para a geração das redes SOM foram os que estão no Quadro 17.

Quadro 17 - Parâmetros utilizados para configuração da rede SOM.

Parâmetro	Tradução	Valor do Parâmetro
Number of nodes	Número de Nós	1000
Map Size	Tamanho do Mapa	Automatic Map
Trainning Schedule	Rotina de Treinamento	Accurate
Tension	Tensão	0,75
Data Records Used	Dados Utilizados	500
Clusted Method	Método de *Cluster*	SOM-Ward *Clusters*

Os parâmetros escolhidos para este experimento foram à aplicação da Teoria dos *Rough Sets* para a extração de regras estão listadas no Quadro 18.

Quadro 18 - Parâmetros utilizados para configuração dos *Rough Sets* para a extração de regras.

Parâmetro	Tradução	Valor do Parâmetro
Reduct / Rule Choice	Reduto / Geração de Regras	Rules
Method	Método	Genetic Algorithm
Genetic Algorithm Settings	Parâmetros dos Algoritmos Genéticos	High Speed

- Experimento com a rede SOM

O resultado esperado neste experimento foi que a rede SOM realizasse a formação de agrupamentos dentro do conjunto de dados apresentados, de forma que estes conjuntos de dados com maior similaridade representassem o maior grau de coesão possível; e que com os seus mapas topológicos pudessem ser extraídas informações descritivas que revelassem a natureza dos ativos; para posterior formação de estratégias de recuperação ou descoberta de informações previamente desconhecidas.

Utilizando a técnica de Ward para a formação dos *clusters*, foram obtidas as seguintes segmentações, apresentadas no mapa topológico da Figura 15.

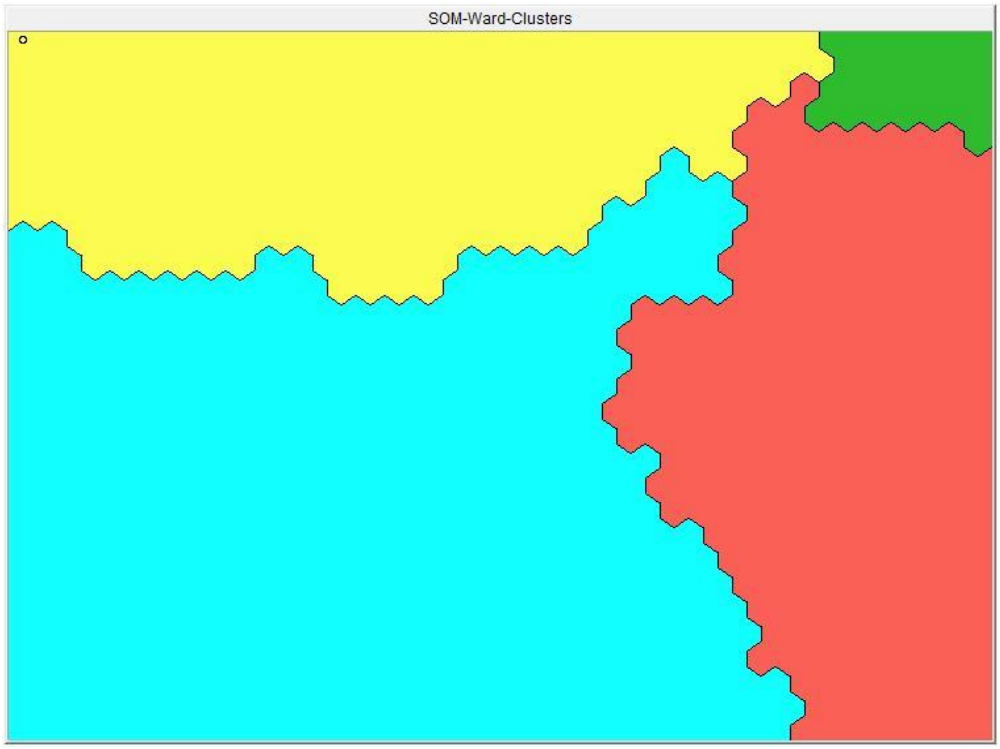


Figura 15 - Mapa Topológico com a determinação dos *Clusters*. Fonte: Elaborada pelo Autor.

De acordo com o mapa gerado, foram geradas as seguintes estatísticas relativas às características desses *clusters*, conforme foi apresentado na Tabela 11.

Tabela 11 - Distribuição dos créditos nos *Clusters*.

Cluster	**Qtde Registros**	**Saldo Inicial Médio**	**Divida Paga?**
S 1 - Azul Claro	*59,60%*	R$ 571	0
S 2 - Vermelho	*20,80%*	R$ 2.320	0
S 3 - Amarelo	*17,60%*	R$ 627	1
S 4 - Verde	*2,00%*	R$ 6.732	0

De acordo com a Tabela 11 o *cluster* S1 possuí grande parte dos créditos em termos de volume de dívidas e com o menor Saldo Inicial Médio. Uma observação que pode ser vista é que o *cluster* S4 apesar de ter 2% do número total de dívidas, possuí a maior concentração em termos de Saldo Inicial Médio de dívidas.

Analisando o Saldo Corrente médio das dívidas, isto é o saldo das dívidas no momento da extração da base de dados para a presente pesquisa constante na Tabela 12 observa-se que o *cluster* S4 detém uma concentração de saldo desproporcional em relação aos demais *clusters*.

Tabela 12 - Distribuição dos créditos nos *Clusters*

Cluster	**Saldo Corrente Médio**	**Divida Paga?**
S 1 - Azul Claro	R$ 1.204	0
S 2 - Vermelho	R$ 4.923	0
S 3 - Amarelo	R$ -	1
S 4 - Verde	R$ 14.358	0

Para uma possível estratégia de recuperação de crédito, a informação na qual os *clusters* distribuídos pelo saldo corrente médio menor podem indicar uma potencial estratégia de recuperação no sentido de iniciar a cobrança de acordo com os *clusters* com menor valor ou de maior valor.

Na Tabela 13 são apresentadas as médias em relação ao Saldo de Abertura (*i.e.* valor do contrato no momento da assinatura do contrato) em relação ao Saldo Principal.

Tabela 13 - Distribuição dos créditos nos *Clusters*.

Cluster	**Saldo de Abertura Médio**	**Saldo Principal Médio**	**Divida Paga?**
S 1 - Azul Claro	R$ 846	R$ 566	0
S 2 - Vermelho	R$ 2.243	R$ 2.314	0
S 3 - Amarelo	R$ 1.348	R$ -	1
S 4 - Verde	R$ 6.380	R$ 6.732	0

Na Tabela 13 observou-se que com exceção do *cluster* S1 os demais *clusters* possuem semelhanças em relação aos valores de Saldo de Abertura Médio e Saldo Principal Médio; o que indica que os créditos tiveram reajustes relativos aos juros implícitos na transação devido ao atraso no pagamento.

Para a análise topográfica dos mapas foi utilizado o campo *Pago*, no qual o valor 1 representa os créditos pagos e o valor 0 representa os créditos não pagos. O mapeamento das fronteiras dos *clusters* mantém-se consistente como pode ser visto na Figura 16.

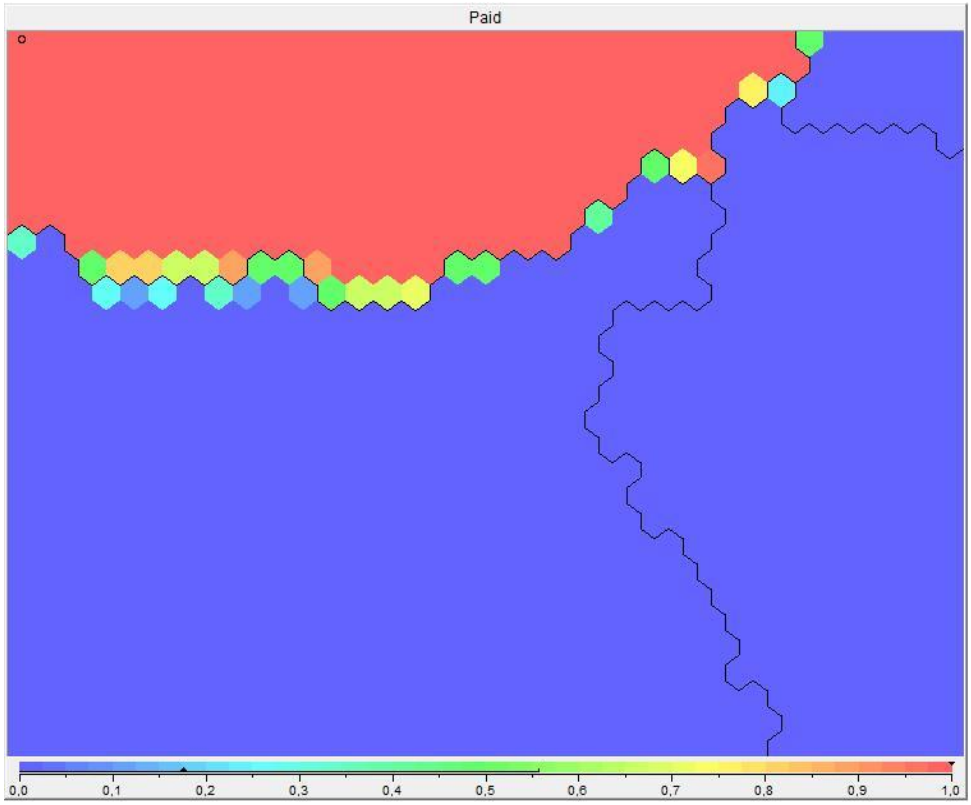


Figura 16 - Mapa Topológico das Dívidas Pagas. Fonte: Elaborada pelo Autor.

Na Figura 16, podemos verificar que de acordo com as determinações realizadas anteriormente, as dívidas pagas obedecem corretamente à determinação de sua segmentação.

Um fato a ser observado através dessa análise visual é que algumas regiões de fronteiras entre os *clusters* há a atribuição de cores diferentes do azul (que representa os créditos não pagos) e do vermelho (que representa os créditos pagos). Essa indefinição na região de fronteira são casos em que há a confusão de características em que o registro pode ‘pular’ para qualquer um dos lados da fronteira. Este trabalho define esses registros como *jumpers*.

Quando é utilizada a dimensão de saldo da dívida no momento da abertura em relação à dimensão topológica dos grupos há a seguinte distribuição mostrada na Figura 17. Essa análise apresenta evidências que há uma concentração de saldo na região do *cluster* de cor verde.

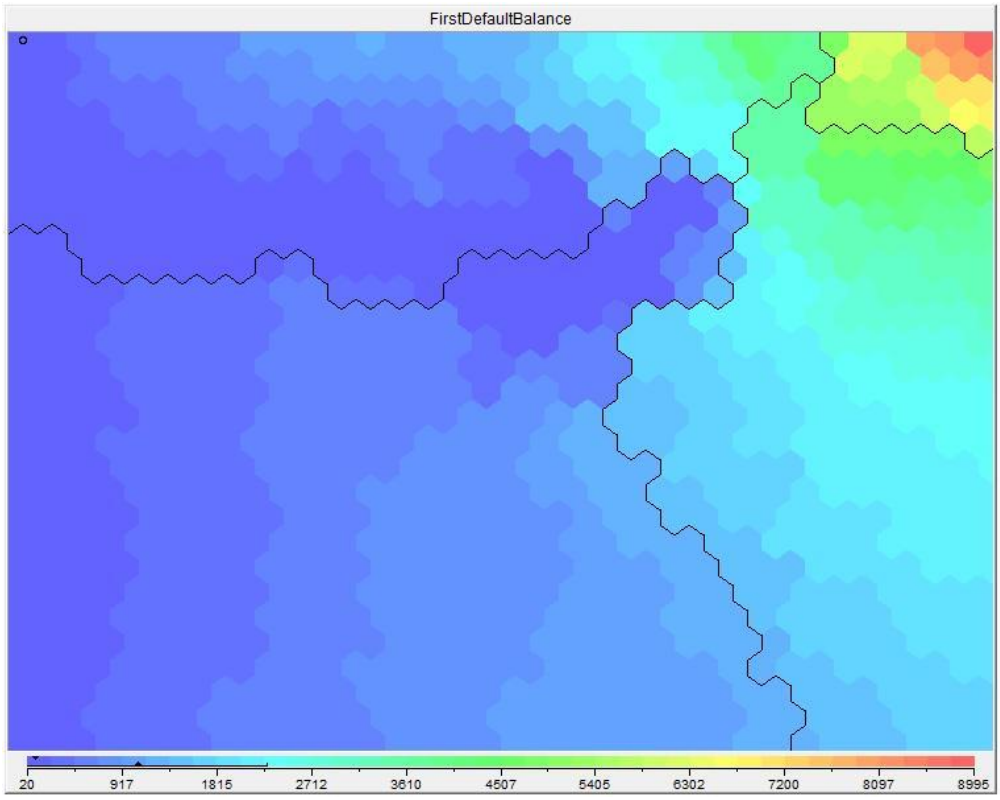


Figura 17 - Mapa Topológico do saldo do crédito na data do atraso. Fonte: Elaborada pelo Autor.

Com o mapa da Figura 17 foi visto que as dívidas relativas ao saldo total na data de atraso estão majoritariamente concentradas em valores abaixo de R$ 3.000, e que também os clientes com um saldo relativamente maior concentram-se em uma região específica do mapa.

Outra variável utilizada para a análise foi o saldo da dívida no momento da aquisição do crédito sem a implicação dos juros decorrentes da operação financeira. Mais uma vez, como podemos ver na Figura 18 ainda há uma clara divisão de acordo com as fronteiras dos *clusters* em relação aos saldos.

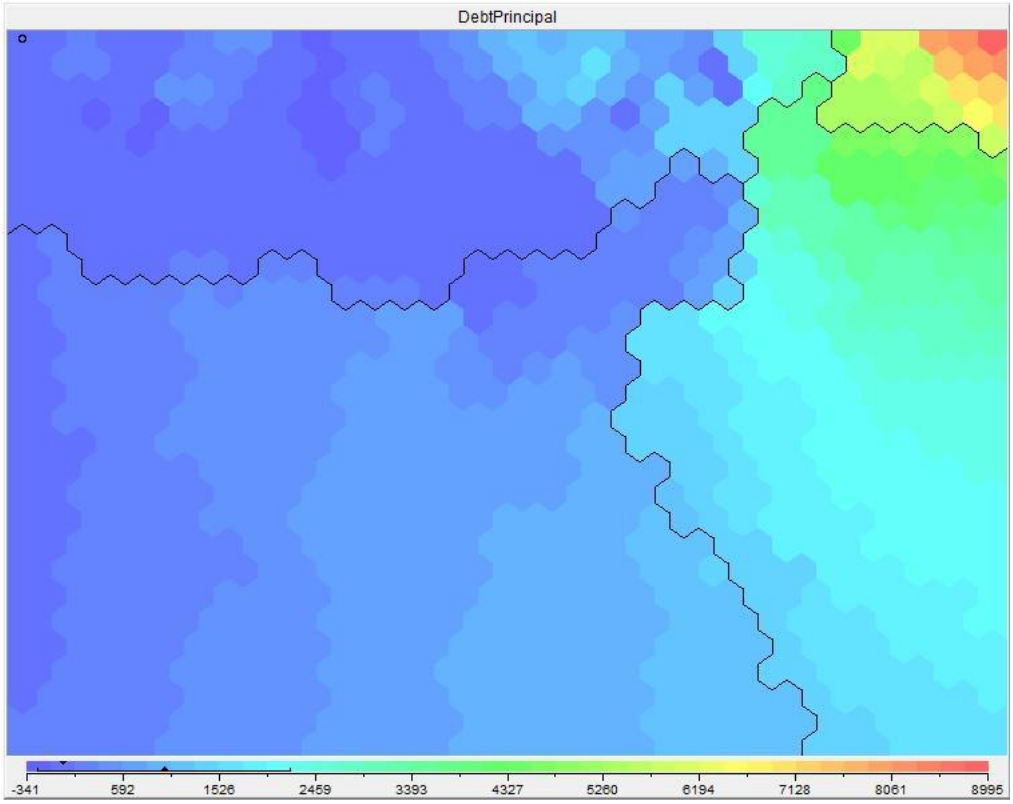


Figura 18 - Mapa Topológico do saldo principal das dívidas sem a implicação dos juros. Fonte: Elaborada pelo Autor.

O mapa da Figura 18 mostra que a concentração em termos de saldo principal (*i.e.* valor do contrato com a incidência de juros) obedece a mesma lógica relativa ao saldo na data do primeiro atraso.

Em outras palavras isto significa que pode haver evidências de que mesmo com a incidência de acréscimos monetários no valor da dívida (*e.g.* atribuição de juros pelo atraso, multas, etc) os créditos detêm o mesmo comportamento em relação à concentração em regiões específicas do mapa. Isso pode indicar que a incidência de mais juros não muda a dinâmica dos pagamentos das dívidas.

Quando realizamos a experimentação em relação ao valor do contrato, vemos que há uma pequena mudança em relação à intensidade do mapa em relação à área dos *clusters*, como mostra a Figura 19.

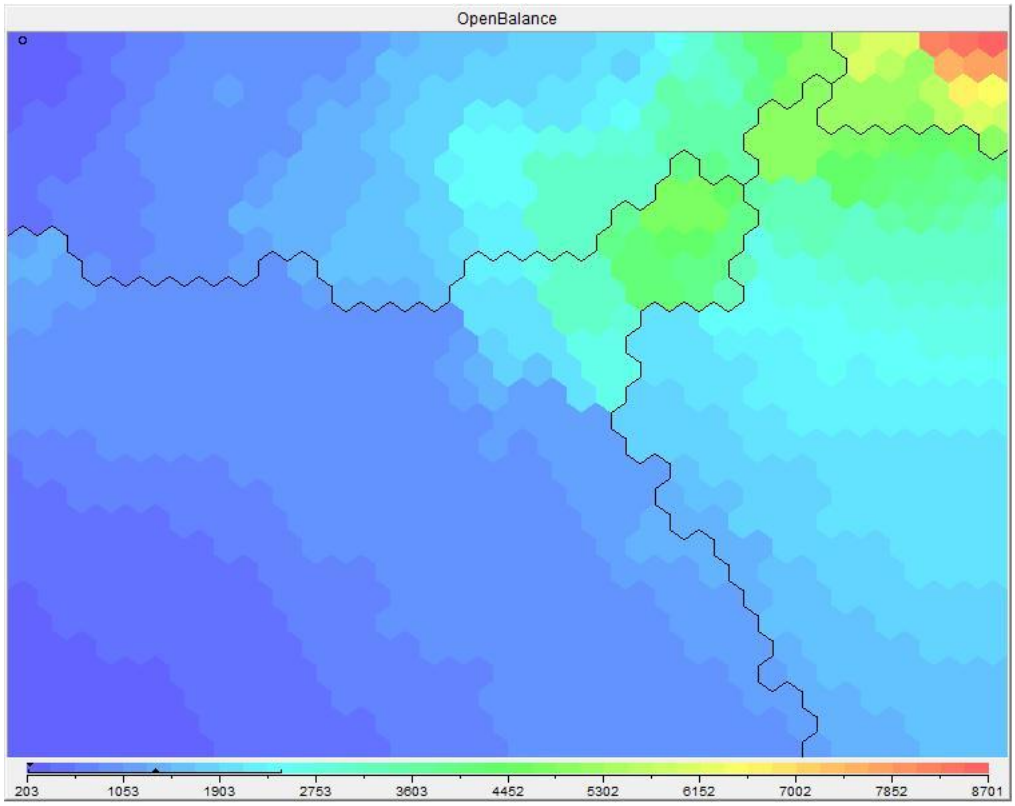


Figura 19 - Mapa Topológico do saldo na abertura do crédito (*i.e.* valor do contrato). Fonte: Elaborada pelo Autor.

Na Figura 19 uma observação que pode ser realizada é que neste portfólio específico é que os créditos que têm o valor entre R$ 3.603 e R$ 5.302 estão em todos os *clusters*, aparentemente em uma distribuição similar. Isso pode apontar que dívidas com o valor de contrato entre esses valores são mais difíceis de caracterizar em uma estratégia inicial de segmentação, somente por saldo e que mais análises devem ser procedidas para entender as dinâmicas de recuperação desses créditos.

No mesmo experimento, quando utilizamos o saldo principal da dívida, sem os emolumentos monetários (*e.g.* tarifas bancárias, impostos, juros decorrentes da transação, *etc*.) vemos ainda uma clara distinção dos *clusters* como na Figura 20.

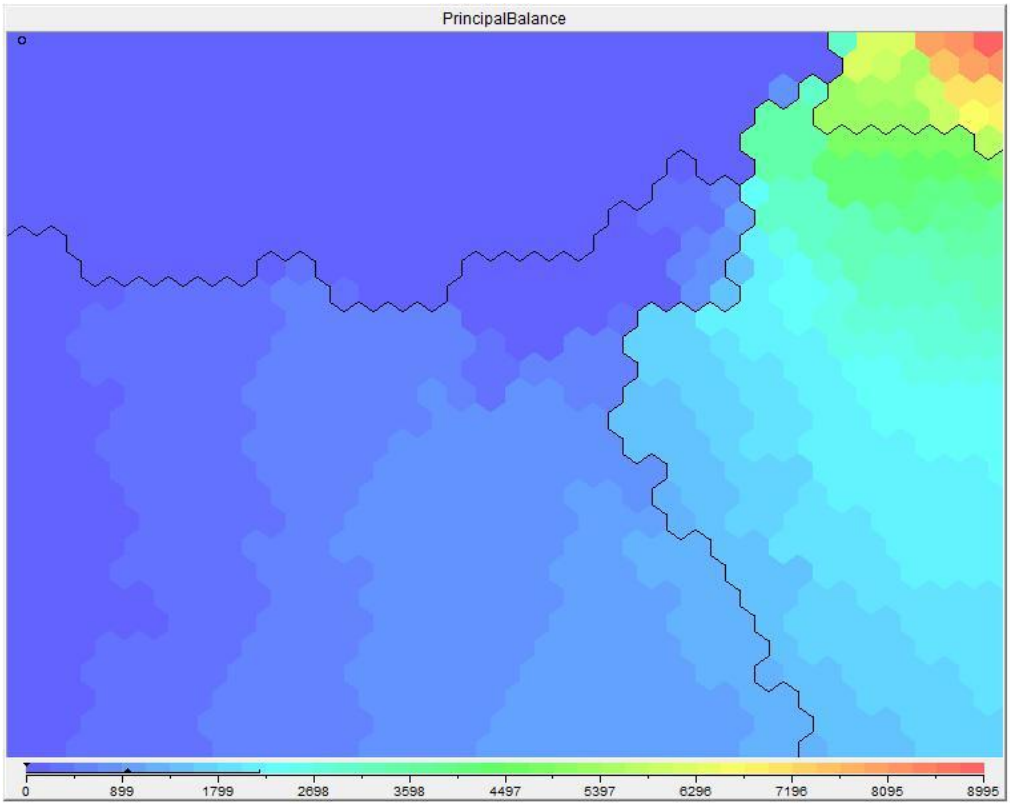


Figura 20 - Mapa Topológico Saldo Principal. Fonte: Elaborada pelo Autor.

Esses experimentos evidenciam que os mapas topológicos possuem um grande poder de discretização em relação à construção de segmentações e aproveitam as facilidades da questão visual para facilitação da compreensão da estrutura de uma base de dados.

- Experimento com a Teoria dos *Rough Sets*

Com a aplicação desses parâmetros foram geradas 4.538 regras com o suporte (i.e. frequência de transações que contenham mais de dois atributos em conjunto) mínimo de 1 atributo na regra e com o suporte máximo de 412 atributos na regra; com o tamanho de atributos na regra de em média de 6,6 dimensões.

Na distribuição das regras através dos atributos de decisão foram: Classe de Decisão Pago: 790 casos; Classe de Decisão Não-Pago: 3.748 casos.

Por meio da análise empírica das regras geradas foi escolhido o seguinte conjunto de regras constantes no Quadro 19 e Quadro 20 como as mais relevantes.

Quadro 19 - Regras para a classe de decisão dos débitos não recuperados.

Regras para a Classe de Decisão "Débito Não-Pago"	Suporte
SE (ID Combinação de Canal de Contato=4) **E** (ID Telefone Primário =FALTANTE) **ENTÃO** (Débito Pago={[116]})	*30,83%*
SE (Numero de Combinações de Contato=3) **E** (Telefone=N) **ENTÃO** (Débito Pago={[110]})	*29,37%*
SE (Concatenação de Canal de Contato="Call Center-Campanha de Carta") **E** (Telefone=N) **ENTÃO** (Débito Pago={[110]})	*28,16%*
SE (Categoria da Dívida=Pessoa Física) **E** (ID Combinação da Localização do Devedor=1) **E** (Idade da Dívida=48 a 54 Meses) **ENTÃO** (Débito Pago={[109]})	*26,70%*
SE (ID Endereço Primário=FALTANTE) **E** (ID Produto=716) **E** (CallCenter=Y) **ENTÃO** (Débito Pago={[108]})	*26,46%*
SE (ID Combinações de Canais de Localização=1) **E** (Numero de Combinações de Contato=3) **ENTÃO** (Débito Pago={[106]})	*26,46%*
SE (Concatenação de Canais de Contato="Call Center-Letter") **E** (Numero de Combinações de Contato=0) **ENTÃO** (Débito Pago={[106]})	*26,46%*
SE (Concatenação de Canais de Contato="Call Center-Letter") **E** (ID Combinações de Canais de Contato=1) **ENTÃO** (Débito Pago={[106]})	*25,73%*
SE (Numero de Combinações de Contato=3) **E** (Localizado=N) **ENTÃO** (Débito Pago={[106]})	*25,73%*
SE (Concatenação de Bureau de Restrições="[Bureau de Crédito 1]") **ENTÃO** (Débito Pago={[89]})	*25,73%*
SE (Restrito em Bureau de Crédito=Y) **ENTÃO** (Débito Pago={[89]})	*25,24%*
SE (Descrição da Idade da Dívida="60 a 66 Meses") **E** (Contato Prévio=Y) **E** (Localizado=N) **ENTÃO** (Débito Pago={[51]})	*12,62%*

Quadro 20 - Regras para a classe de decisão dos débitos recuperados.

Regras para a Classe de Decisão "Débito Pago"	Suporte
SE (Valor Inicial da Dívida=Até 200 Reais) **E** (Concatenação do Canal de Localização=Telefone-Endereço) **ENTÃO** (Débito Pago={1[32]})	*36,36%*
SE (ID Produto=707) **E** (Mensagem de Voz=N) **E** (Número de Combinações de Contato=3) **ENTÃO** (Débito Pago={1[28]})	*31,82%*
SE (Valor Inicial da Dívida=Até 200 Reais) **E** (Campanha de Cartas=Y) **b** (Concatenação de Canais de Localização=Telefone-Endereço) **ENTÃO** (Débito Pago={1[22]})	*26,14%*

De acordo com o número de registros alcançados pelas regras, bem como a complexidade das regras, os *Rough Sets* mostram uma grande capacidade em relação à extração de regras acionáveis, isto é, regras que dão algum tipo de sugestão de ações a serem tomadas.

4.1.4 Discussão dos Resultados

- Experimento com a rede SOM

As análises dos resultados experimentais com as redes SOM, de acordo com a Figura 12 foram gerados quatro *clusters* dentro da população de créditos disponíveis para a análise.

Uma das características das análises de *clusters*, e em especial da análise topológica de dados é a definição de *personas* que são perfis para ilustrar determinada característica de um grupo de instâncias em uma base de dados, ou mesmo a caracterização dos grupos de acordo com características implícitas nos dados.

De acordo com os dados provenientes da Figura 16, Figura 17 e Figura 18, os segmentos, isto é as suas composições topográficas no mapa, foram caracterizados da seguinte forma:

***Cluster 1* - Região Azul:** Nesta região há uma grande concentração de débitos (59,60%), no qual o Saldo Inicial das dívidas é em média de cerca de R$ 571,00, sendo que estes créditos em sua maioria ainda estão pendentes de recuperação. Nas métricas relativas a Saldo de Abertura do Crédito (R$ 846,00), Saldo Principal (R$ 566,00), e o Saldo Corrente da dívida (R$ 1.204,00) este *cluster* apresenta o menor valor entre todos os *clusters*. Dessa forma este *cluster* será denominado como "*Segmento dos Pequenos Devedores*".

***Cluster 2* - Região Vermelha:** Nesta região, temos a concentração de cerca de 1/5 dos créditos (20.80%) totais do portfólio estudado e o mesmo é composto majoritariamente de créditos ainda inadimplidos. Com o Saldo Inicial médio de R$ 2.320,00, este segmento apresenta um valor médio quase quatro vezes maior que o *Cluster* 1 do segmento dos pequenos devedores. Este fato pode ser constatado em relação às métricas de Saldo na Aquisição (R$ 2.320,00), Saldo Corrente (R$ 4.923,00) e Saldo na Abertura do Crédito (R$ 2.243,00). Como esses créditos apresentam um valor médio relativamente alto em relação ao primeiro segmento, denominaremos esse grupo como "*Segmento dos créditos de valor Médio*".

***Cluster 3* - Região Amarela:** Nesta região que concentra 17,60% do total de créditos, é exclusivamente o único *cluster* que majoritariamente é composto de dívidas pagas. Isto quer dizer que dentro de sua região de fronteira, todas as informações correlatas serão relativas a débitos já pagos pelos devedores. A análise das características deste segmento em especifico serve não somente para saber quais créditos foram pagos, mas também observar as dinâmicas por trás dos pagamentos das dívidas e determinar quais estratégias devem ser adotadas para recuperação dos créditos. Com um Saldo Inicial Médio e Saldo Médio de Aquisição de R$ 627,00, e também com um saldo de abertura sendo o segundo menor de todos os segmentos

(R$ 1.348,00) há evidências que os créditos recuperados obedecem a dinâmicas de recuperar débitos de pequenos valores de forma majoritária. Isso pode ser levado em consideração no momento de elaboração de uma estratégia com o foco no aumento do volume de recuperações, em detrimento de uma recuperação mais focada em valor monetário. Este *cluster* será conhecido como o "*Segmento dos Créditos Recuperados*".

***Cluster 4* - Região Verde:** Por último, neste *cluster* composto de 2% do total de créditos, vemos que de acordo com suas métricas como Saldo Inicial Médio e Saldo Principal (R$ 6.732,00), Saldo Corrente (R$ 14.358,00) e Saldo de Abertura (R$ 6.380,00) devido aos seus altos valores indicam que essas dívidas possuem um alto valor monetário. Dessa forma determinaremos esse *cluster* como o "*Segmento de Alto Valor Monetário*".

Esses pontos são caracterizados por clientes que já fizeram o pagamento e estão na região topológica de créditos não pagos, ou clientes que estão na região de fronteira de créditos pagos, mas que efetivamente não realizaram o pagamento dos NPLs. Esses pontos de dados no mapa que estabelecem uma situação de incerteza serão chamados neste trabalho de *jumpers*, isto é, pontos em regiões de fronteira sem nenhum tipo de caracterização perfeita.

Dentro das fronteiras topológicas um gráfico no qual o valor da dívida apresenta uma concentração no *cluster* 4 (Segmentação de Alto Valor Monetário); pode ser visto que dentro da mesma intensidade de valores no "Segmento dos créditos de valor Médio" há dívidas tanto que tem um valor perto dos débitos de alto valor, como perto do "Segmento dos Pequenos Devedores"; o que pode indicar que estrategicamente esse *cluster* seja mais difícil de trabalhar em termos operacionais de cobrança. No *cluster* dos débitos pagos, podemos verificar que a grande maioria esses créditos quitados são de baixo valor monetário.

Um fato importante é que o "Segmento dos Pequenos Devedores" possuí uma linha de fronteira bem definida com uma uniformidade de seus valores em relação aos *clusters* vizinhos; o que pode indicar que essa definição é fruto de algum tipo de promoção em relação a um tipo de produto creditício, ou mesmo no momento que esses créditos foram prestados aos clientes foi elaborada uma estratégia de expansão de créditos de baixo valor.

Quando é feita a análise do valor da dívida em relação à da data de assinatura do contrato, pode ser visto que de uma maneira mais clara que o *cluster* relativo ao "Segmento de Alto Valor Monetário" apresenta uma distinção bem clara em relação a os outros *clusters*.

Um fato que chama atenção é que os segmentos de "Segmento dos Pequenos Devedores", "Segmento dos créditos de valor Médio", e "Segmento dos Créditos Recuperados apresentam uma faixa de intersecção significativa (área em azul claro/verde piscina) o que pode

indicar que os contratos não apresentaram dinâmicas semelhantes no momento de sua aquisição.

Essa informação ainda precisa ser mais bem observada devido ao fato de que devido essa sobreposição de valores, não há nenhum tipo de indicação de como é o comportamento dos *clusters* de acordo com os valores de abertura.

No momento em que se analisa o saldo da dívida com os juros e emolumentos (saldo principal) como métrica que faz a intensidade topológica, mais uma vez o *cluster* relativo ao "Segmento de Alto Valor Monetário" apresenta uma distinção interessante em estratégias especificas que podem ser elaboradas para este grupo especifico dado o seu alto grau de coesão que pode facilitar a generalização de abordagens de recuperação.

Mais uma vez o *cluster* relativo ao "Segmento dos créditos de valor Médio" apresenta uma particularidade na qual possui créditos tanto de valor alto, médio e baixo; no qual as dinâmicas dentro deste *cluster* precisam de um melhor entendimento para que possam ser elaboradas deduções a respeito do comportamento desses créditos dentro da composição da carteira.

- Experimento com *Rough Sets*

Das análises dos resultados experimentais com *Rough Sets* no que concerne as regras para a classe de decisão de créditos “Não-Pagos” apresentou o suporte médio (*i.e.* a conjunção de itens em relação à base de dados) de 25,79%. Este resultado indica que mesmo em uma base de dados com um volume relativamente pequeno de instâncias (500 registros) o algoritmo conseguiu extrair regras úteis para as análises.

Entre as 14 regras extraídas, 10 apresentaram algum atributo relativo a questões de localização como telefone, endereço, canal de localização. Dessa forma, há evidências que grande parte dos débitos não pagos apresentaram forte relação com o fato de questões relativas à consistência dos endereços de localização dos devedores.

As regras geradas para a classe de decisão de créditos não pagos apresentaram os seguintes resultados de análise, para as conseguintes conclusões no nível de determinação de estratégias:

Regra 1 – **SE** (ID Combinação de Canal de Contato=4) **E** (ID Telefone Primário =FALTANTE) **ENTÃO** (Débito Não-Pago={[116]})

Esta regra apresenta maior suporte dentro da classe créditos “Não-Pago” com 30.83%; o que indica que é uma regra com alcance significativo de instâncias para tomada de ações. Essa regra mostra que quando não há a informação de telefone primário do devedor (*i.e.* quando o devedor não possuí nem mesmo telefone fixo para qualquer tipo de contato telefônico) e a combinação de canal de contato (*i.e.* isso é se o cliente tem endereço-telefone, só telefone, só endereço, ou nenhum) é igual a 4 há uma forte relação com a não recuperação desses créditos. Uma forma de transformar essa regra em acionável é a execução de medidas como enriquecimento dos dados cadastrais de endereços.

Regra 2 - SE (Numero de Combinações de Contato=3) **E** (Telefone=N) **ENTÃO** (Débito Não-Pago={[110]})

Essa regra apresenta o suporte de 29.37%. Essa regra indica que quando o número de combinações de contato, isto é, a forma de se conseguir contato com o devedor é igual a 3, e o devedor não tem telefone implica necessariamente que a dívida não está paga. Neste caso um fator importante a ser destacado é a ausência de telefone na regra uma forma de deixar a regra acionável é providenciando algum tipo de estratégia para enriquecimento dos telefones dos clientes seja via compra de dados cadastrais, ou mesmo validação prévia dos telefones antes da compra do ativo NPL.

Regra 3 - SE (Concatenação de Canal de Contato="Call Center-Campanha de Carta") **E** (Telefone=N) **ENTÃO** (Débito Não-Pago={[110]})

Essa regra apresentou um suporte em relação ao número de instâncias na base de dados de 28,16%. Essa regra diz que quando o canal de contato com o devedor foi realizado ou por *call-center* e por campanha de carta, mas que também o telefone não estava registrado na base de dados, o devedor não pagou a dívida. O que explica essa regra é que quando o contato foi realizado pelo *Call-Center*, o telefone pode ter sido invalidado (o que implicaria a sua exclusão dos registros do devedor) e posteriormente, sem as informações desses clientes foi realizada uma campanha de envio de cartas de pré-acordo para estes clientes para que os mesmos pudessem retornar ao telefone do *Call-Center* para procederem com as negociações.

Regra 4 - SE (Categoria da Dívida=Pessoa Física) **E** (ID Combinação da Localização do Devedor=1) **E** (Idade da Dívida=48 a 54 Meses) **ENTÃO** (Débito Não-Pago={[109]})

Esta regra apresentou 26,70% de suporte de transações. A descrição dessa regra diz que se o tomador do crédito foi pessoa física e a ID da sua localização é igual a 1 e a idade da dívida

é de 48 a 54 meses (4 até 4 anos e meio) o débito continua inadimplido. Essa regra permite induzir que a partir de um determinado momento no tempo em que as dívidas ficam mais antigas o índice de recuperação é maior. Isso deve ser confirmado realizando inferências em relação à idade média das dívidas contra o que está especificado na regra para que a afirmação seja consistente.

Regra 5 - SE (ID Endereço Primário=FALTANTE) **E** (ID Produto=716) **E** (CallCenter=Y) **ENTÃO** (Débito Não-Pago={[108]})

Essa regra possuí 26.46% de suporte em relação às transações do conjunto de dados. Essa regra indica mais uma vez que o fator endereço exerce influência na questão das dívidas não quitadas pelos devedores. Outra informação na regra é que em conjunto com a informação de endereços, o produto 716 pela primeira vez aparece como um fator de manutenção da inadimplência. Outra informação dessa regra é que mesmo com o contato via *Call-Center*, ainda sim o crédito permanece inadimplido. Isso leva a induzir que é necessária uma análise em relação às estratégias de recuperação do produto 716, e mais que isso: analisar a questão de endereços e se mesmo com o contato via *Call-Center* os telefones estão gerando contatos efetivos com os clientes ou não.

Regra 6 - SE (ID Combinações de Canais de Localização=1) **E** (Numero de Combinações de Contato=3) **ENTÃO** (Débito Não-Pago={[106]})

Essa regra é responsável por 26,46% do suporte da base de dados. Essa regra mostra que as combinações de canais de localização quando tem o valor 1 e o número de combinações de canais de contato é igual a 3 parte dos créditos não é quitado. Isso mostra que o fator de localização dos clientes, isto é, contato por telefone e endereço e a forma na qual a localização é realizada são importantes para a recuperação dos créditos.

Regra 7 - SE (Concatenação de Canais de Contato="Call Center-Letter")E (Numero de Combinações de Contato=0) **ENTÃO** (Débito Não-Pago={[106]})

Essa regra apresenta 26.46% de suporte em relação a todas as transações constantes na base de dados. Essa regra dentro da sua constituição em um primeiro momento pode parecer contra intuitiva já que os canais de contato conseguidos foram o *Call-Center* (via ligação) e envio de cartas para os devedores foi realizado, mas mesmo assim os débitos não foram pagos. Um fator importante é que mesmo com esses canais de contato, há evidências que o número de combinações de contato, isto é, se há mais de uma forma de contatar o devedor (*e.g.* telefone,

e-*mail*, carta, *etc*.) em que o contato é estabelecido exercem influência na capacidade de recuperar o ativo inadimplido.

Regra 8 - SE (Concatenação de Canais de Contato="Call Center-Letter") **E** (ID Combinações de Canais de Contato=1) **ENTÃO** (Débito Não-Pago={[106]})

Essa regra possuí 26.46% de suporte em relação a todo conjunto de dados estudado. Ao exemplo da regra 7 há uma pequena contra intuição em relação ao que a regra diz que é que mesmo com o contato via *Call-Center* e envio de cartas o ID das combinações dos canais de contato com o devedor exercem influência em relação à recuperação dos créditos. Neste caso, ao exemplo da regra 7, é necessário analisar quais combinações de contato são mais efetivas, bem como verificar se as combinações de fato melhoram a capacidade de recuperação.

Regra 9 - SE (Numero de Combinações de Contato=3) **E** (Localizado=N) **ENTÃO** (Débito Não-Pago={[106]})

Essa regra é responsável por 26,46% do suporte da base de dados. Novamente a questão da localização entra em análise já que a regra diz que quando o número de combinações de contato é igual 3 e o cliente não está localizado a dívida não é paga. Isso leva a análise da decisão se é necessário o enriquecimento dos dados cadastrais para exercer a atividade de cobrança.

Regra 10 - SE (Concatenação de Bureau de Restrições="[Bureau de Crédito 1]") **ENTÃO** (Débito Pago={[89]}) e Regra **SE** (Restrito em Bureau de Crédito=Y) **ENTÃO** (Débito Não-Pago={[89]})

Essa regra possuí respectivamente 25.73%e 25.24% de suporte em relação a todo conjunto de dados estudado. Um novo fator que ainda não tinha aparecido nas análises anteriores foi a indicação de restrição de crédito do devedor. Essa regra mostra que quando o devedor é restrito no Bureau de Crédito 1, implica que ele ainda não exerceu o pagamento da dívida. Essa regra tem um efeito apenas descritivo já que se o cliente pagou todos os seus débitos o mesmo não deve constar nos órgãos de restrição de crédito.

Regra 11 - SE (Descrição da Idade da Dívida="60 a 66 Meses") **E** (Contato Prévio=Y) **E** (Localizado=N) **ENTÃO** (Débito Não-Pago={[51]})

Essa regra é responsável por 12,62% do suporte da base de dados. Apesar de ter um suporte relativamente baixo às outras regras, mostra que quando a idade da dívida é de 60 a 66

meses (5 a 5 anos e meio) e foi realizado algum tipo de contato com o devedor e o mesmo não foi localizado mais o débito não foi pago. Devido aspectos relativos à legislação quando o débito possuí mais de 5 anos desde o último contato com o devedor, o mesmo saí de todos os registros dos órgãos de restrição. Dessa forma, uma estratégia para inibir custos de ir atrás de clientes que são resistentes à ideia de pagamento é inibir o contato neste intervalo de clientes e manter o foco em dívidas mais recentes, isto é, com uma idade média menor.

Das 3 regras extraídas todas apresentam algum atributo relativo a questões de localização e/ou se o cliente estava localizado para realização do contato para a negociação. Dessa forma, há fortes evidências que grande parte dos débitos só foram quitados devido ao fato dos clientes possuírem endereços ou telefones consistentes.

As regras geradas para a classe de decisão de créditos pagos apresentaram os seguintes resultados de análise, para as conseguintes conclusões em nível de determinação de estratégias:

Regra 12- SE (Valor Inicial da Dívida=Até 200 Reais) **E** (Concatenação do Canal de Localização=Telefone-Endereço) **ENTÃO** (Débito Pago={1[32]})

Essa regra é responsável por 36,36% do suporte da base de dados. Essa regra diz que se o valor inicial da dívida é de até 200 reais e o canal de localização do cliente é realizado via telefone e endereço o crédito foi recuperado. O conhecimento que pode ser extraído dessa regra é que dívidas que tenham um valor baixo, e com o contato via telefone e endereço com os devedores sugerem uma recuperação maior. Dessa forma no momento de aquisição de novas dívidas pode ser adotada a estratégia de preferência para dívidas com um valor médio menor em detrimento a dívidas com valores maiores.

Regra 13 - SE (ID Produto=707) **E** (Mensagem de Voz=N) **E** (Número de Combinações de Contato=3) **ENTÃO** (Débito Pago={1[28]})

Essa regra apresenta 31.82% de suporte em relação a todas as transações constantes na base de dados. Essa regra mostra que quando o produto é o 707 e que também não houve o envio de mensagens de voz para o cliente e o número de combinações de contato com o devedor é igual a 3 o crédito foi recuperado. Essa regra apresenta a título de informações que o produto 707 mostra um bom índice de recuperação e que o envio de mensagens de voz para os clientes não apresenta uma aderência interessante à título de recuperação dos créditos. Mais uma vez vemos também que as combinações de contato com os devedores são de fundamental importância para a recuperação.

Regra 14 - SE (Valor Inicial da Dívida=Até 200 Reais) **E** (Campanha de Cartas=Y) **E** (Concatenação de Canais de Localização=Telefone-Endereço) **ENTÃO** (Débito Pago={1[22]})

Essa regra possuí 26.14% de suporte em relação às transações do conjunto de dados. Nesta regra é apresentada a informação de que se o valor inicial da dívida é de até 200 reais e houve envio de cartas para o devedor e se a forma de localização os canais de comunicação com o cliente foi telefone juntamente com o endereço o crédito foi recuperado com sucesso. Novamente pode ser visto que há evidências que o fator de recuperação da dívida tem algum tipo de ligação com o valor inicial da mesma, isto é, quando mais baixo o valor inicial maior o potencial de recuperação. Outra informação importante é que a campanha de cartas juntamente com o contato via os canais de localização de telefone e endereço foram importantes para a recuperação de crédito. Isso pode indicar, por exemplo, que a carta foi enviada para um registro de endereço consistente (geralmente com algum tipo de pré-acordo com descontos no valor da dívida) e que o devedor fez contato com o *Call-Center* via telefone e posteriormente a isso houve uma negociação e o débito foi pago com sucesso.

A aplicação da rede SOM gerou *clusters* que revelaram a concentração de clientes em relação ao seu valor monetário dentro de uma carteira de crédito. Os *clusters* gerados apresentaram fronteiras bem definidas e boa coesão interna dos registros, o que permitiu a caracterização dos grupos de acordo com informações implícitas nos dados e a análise visual desses dados.

O experimento com RS evidenciou que a análise proveniente dos *clusters* possibilitou a extração de regras de associação que identificaram com clareza os tipos de créditos não pagos presentes em cada *cluster* apesar da alta dimensionalidade da base.

Pode-se afirmar, então, com base nos resultados positivos que a aplicação da rede SOM em conjunto com *Rough Sets* pode ser um importante instrumento seja para criação de estratégias ou desenvolvimento de modelos de recuperação desses créditos, ou para criação de modelos de recuperação baseado em escoragem (*Recovery Scoring*).

4.1.5 Experimento 3: Experimento com a Teoria dos *Rough Sets* conjuntamente com Árvores de Decisão

Nesta seção são apresentados os parâmetros e a descrição da forma de em que os experimentos com a Teoria dos *Rough Sets* e as Árvores de Decisão foram conduzidos.

- Redução de Atributos com a Teoria dos *Rough Sets*

Os parâmetros do experimento com *Rough Sets* utilizando o Algoritmo de Johnson estão no Quadro 21.

Quadro 21 - Parâmetros utilizados para configuração dos *Rough Sets* utilizando o método do Algoritmo de Johnson.

Parâmetro	**Tradução**	**Valor do Parâmetro**
Discernibility	Discernibilidade	Full
Modulo Decision	Módulo de Decisão	True
Discernibility predicate	Predicado de indiscernibilidade	False
Memory usage	Uso de memória	False
Approximate solutions	Soluções de aproximação	Compute approximate solutions
Hitting fraction	Fração de quebra para aproximação	0.95

Os parâmetros do experimento com *Rough Sets* utilizando o método de Algoritmos Genéticos estão no Quadro 22.

Quadro 22 - Parâmetros utilizados para configuração dos *Rough Sets* utilizando o método de Algoritmos Genéticos.

Parâmetro	**Tradução**	**Valor do Parâmetro**
Discernibility	Discernibilidade	Full
Modulo Decision	Módulo de Decisão	True
Discernibility predicate	Predicado de Discernibilidade	False
Memory usage	Uso de memória	False
Approximate solutions	Soluções de aproximação	Compute approximate solutions
Algorithm Variation	Variação do algoritmo	Modified
Options	Opções	Boltzmann scaling of fitness function
Temperature Range	Intervalo de temperatura	[6.45 - 1.45]
Delta	Delta	0.02
Sample parents with replacement	Amostra com elementos pai com substituição	True
Use elitism	Uso do elitismo	True
Crossover probability	Probabilidade de crossover	0.7
Mutation probabiliy	Probabilidade de mutação	0.07
Inversion probability	Probabilidade de Inversão	0.05

Parâmetro	**Tradução**	**Valor do Parâmetro**
Nr. crossover points	Número de pontos de crossover	1
Nr. Mutations on an individual	Número de mutações sobre um indivíduo	1
Nr. Transpositions for inversion	Número de transposições para inversão	1
Nr. Generations to wait for fitness to improve	Número de gerações de espera para adaptações de melhora	70
Stop if average population fitness does not improve	Critério de parada da adaptação se a população média melhorar	True
Stop if keep list does not change	Critério de parade se a lista de indivíduos não mudar	True
Population size	Tamanho da população	70
Size of keep list	Tamanho da lista de indivíduos	256
Weighting between subset cost and hitting fraction	Peso ente um subconjunto de custo e a função de quebra para aproximação	0.4
Incorporate attribute cost information	Incorporação da informação sobre o atributo de custo	False
Random number generator seed	Semente do número aleatório	54321
Compute approximate solutions	Computar soluções de aproximação	False

A parametrização para geração das árvores de decisão para os três experimentos foi a que está no Quadro 23.

Quadro 23 - Parâmetros utilizados para geração das Árvores de Decisão.

Parâmetro	Tradução	Valor do Parâmetro
Tree Display	Apresentação da Árvore	Top Down
Nodes	Nós	Statistics
Branch Statistics	Estatísticas dos Nós	Yes
Growing Method	Método de Crescimento da Árvore	CHAID
Minimum Cases in Parent Node	Número mínimo de casos no nó antecessor	100
Minimum Cases in Child Node	Número mínimo de casos no nó sucessor	50
Validation Type	Método de validação	None
Allow splitting of merged categories	Permitir redivisão de categorias	True
CHAID Alpha Split	Critério de divisão	0.05
Alpha Merge	Parâmetro de junção	0.05
Split Merged	Divisão de junções	No
Chi-Square	Qui-quadrado	Pearson
Dependent Variable	Variável dependente	Dummy_pago
Maximum Tree Depth	Profundidade máxima da árvore	3

Com os parâmetros especificados na Metodologia Experimental, foram obtidos os seguintes resultados exibidos no Quadro 24.

Quadro 24 - Resultados dos Redutos de acordo com os algoritmos utilizados.

Algoritmo	**Qtde Redutos**	**Principais Redutos**
Algoritmos Genéticos	43	**#1:**{**DIV_SaldoInicial**, DIV_DataContrato, **DIV_CodigoCombinacaoCanalLocalizacao**, DIV_UF, COMBCANALCONT_NumeroCombinacao}, **#2:**{**DIV_SaldoInicial**, DIV_DataContrato, **DIV_CodigoCombinacaoCanalContato**, DIV_UF, COMBCANALLOCAL_NumeroCombinacao}, **#3:**{**DIV_SaldoInicial**, DIV_CodigoDataPrimeiroAtraso, **DIV_CodigoCombinacaoCanalLocalizacao**, DIV_UF, COMBCANALCONT_NumeroCombinacao},

Algoritmo	Qtde Redutos	Principais Redutos
		#4:{**DIV_SaldoInicial,** DIV_CodigoDataPrimeiroAtraso, DIV_UF, COMBCANALCONT_NumeroCombinacao, COMBCANALLOCAL_NumeroCombinacao}, **#5:**{**DIV_SaldoInicial**, DIV_DataContrato, DIV_UF, DIV_CodigoCanalLocalizacao, COMBCANALCONT_NumeroCombinacao}
Algoritmo de Johnson	1	{DIV_SaldoInicial, DIV_DataPrimeiroAtraso, DIV_CodigoCombinacaoCanalContato, DIV_CodigoCombinacaoCanalLocalizacao}

Em negrito no Quadro 24 os atributos que aparecem tanto nos resultados provenientes do algoritmo genético, quanto no algoritmo de Johnson. Esse resultado indica que esses atributos são os mais importantes para a análise.

Foi escolhido para representar o experimento com algoritmos genéticos o reduto #1 composto dos atributos {DIV_SaldoInicial, DIV_DataContrato, DIV_CodigoCombinacaoCanalLocalizacao, DIV_UF, COMBCANALCONT_NumeroCombinacao}. Essa escolha foi devido ao Rosetta indicar esse reduto em primeiro lugar em ordem de importância no resultado final.

- Geração das Árvores de Decisão

Após a geração dos redutos, foram geradas 3 árvores de decisão as quais levam em consideração os seguintes aspectos (i) acurácia, (ii) tempo de processamento, e (iii) quantidade de nós terminais.

A primeira árvore de decisão foi gerada com todos os atributos constantes na base de dados; e esta árvore será denominada como M1. A segunda árvore conteve somente os atributos do reduto gerado pelos *Rough Sets* utilizando os Algoritmos Genéticos; esta árvore denominada como M2. E a terceira e última árvore que foi gerada conteve os atributos contidos no reduto gerado pelos *Rough Sets* através do método proveniente do Algoritmo de Johnson; esta árvore foi chamada de M3.

A acurácia determina a taxa de acerto da árvore em relação ao conjunto de dados treinado e determina com precisão qual é o grau de generalização na qual a árvore pode ser submetida.

O tempo de processamento indica o tempo total em que a árvore levou para ser treinada e testada dado o número de registros e atributos disponíveis. A escolha do tempo de processamento como medida de desempenho foi realizada para obter informações de contraste com os modelos com atributos reduzidos dado que a base de dados contém uma alta dimensionalidade, isto é, possuí muitos atributos, muitos deles covariantes.

E por último, a quantidade de nós terminais leva em consideração a complexidade da árvore para que a mesma seja compreendida por um analista humano ou mesmo o quão boa é a sua capacidade de generalização.

Todas as árvores de decisão geradas pelo SPSS tiveram a profundidade máxima de três níveis, com o número mínimo de 100 casos no nó precedente e de 50 casos no nó consequente.

Na Tabela 14 abaixo foi verificado desempenho em termos de custos relativos às métricas de avaliação de modelos, e também em termos de custo temporal computacional.

Tabela 14 - Resultados dos Modelos de Árvores de Decisão.

Modelo	**Acurácia**	**Δ – Acurácia**	**Nós**	**Δ – Nós**	**Nós Terminais**	**Δ –Nós Terminais**	**Tempo de Execução**	**Δ - Execução**
M1	*93,90%*		24		16		00:03,020	
M2	*88,39%*	*-5,50%*	42	18	31	15	00:01,370	-54,64%
M3	*91,16%*	*-2,74%*	58	34	37	21	00:00,390	-71,53%

A árvore de decisão gerada a partir de todas as variáveis teve a maior acurácia dos 3 experimentos escolhidos (93,90%), e também obteve a menor quantidade de nós (24), a menor quantidade de nós terminais (16). Contudo, esse desenho experimental com todos os atributos teve o maior tempo de processamento em relação aos demais experimentos (3s020).

A árvore de decisão gerada a partir das variáveis escolhidas via Algoritmo Genético obteve a menor acurácia de todos os experimentos (88,39%) cerca de -5,50% em relação ao modelo com todas as variáveis. Este modelo teve a segunda maior quantidade de nós (42) e a segunda maior quantidade de nós terminais (31). O modelo também obteve o segundo lugar em tempo de processamento (01s370).

A árvore de decisão gerada por meio das variáveis escolhidas via Algoritmo de Johnson obteve a segunda maior acurácia (91, 16%) com a perda de -2,74% em relação ao modelo com

todas as variáveis. Este modelo obteve as maiores quantidades em quantidade de nós (58) e no número de nós terminais (37).

No entanto, em termos de processamento este modelo obteve o menor tempo em relação a todos os outros (0s390) com a redução de -71,53% do tempo de processamento em relação ao modelo com todas as variáveis que está servindo de referência.

- Resultados das árvores decisão

As três árvores de decisão geradas nos experimentos apresentaram as seguintes matrizes de confusão apresentadas na Tabela 15 e mostram como ficaram as distribuições das classificações:

Tabela 15 - Matriz de confusão com todos os três modelos de Árvores de Decisão.

Modelo	**Total Registros**	**Verdadeiros Positivos (VP)**	**Verdadeiros Negativos (VN)**	**Falsos Positivos (FP) Erro Tipo I**	**Falsos Negativos (FN) Erro Tipo II**
M1	6652	**1508**	**4738**	**160**	**246**
M2		1205	4675	223	549
M3		1412	4652	246	342

Na Tabela 15 em que são apresentadas as matrizes de confusão relativas aos três modelos, o modelo M1 que contém todas as variáveis obteve o melhor desempenho em todas as variáveis da matriz. Isso pode ser devido ao fato de que quanto maior o numero de atributos contidos na base de dados, melhor será o desempenho do classificador.

Se isso por um lado há vantagens em termos de obtenção dos melhores resultados, essa abordagem pode levar ao fenômeno do *Overfitting*, *i.e.*os dados obtêm um ajuste muito preciso no classificador e isso acontece por meio da incorporação do erro como característica e não um problema. Em outras palavras o classificador memoriza ao invés de aprender com a amostra.

Ainda na Tabela 15 observando-se os resultados entre os modelos compostos com os conjuntos de atributos selecionados pelos redutos (M2 e M3), observa-se que o modelo M3 que é o modelo proveniente do conjunto de atributos reduzidos utilizando o método do Algoritmo de Johnson, obteve o melhor resultado sobre o modelo M2 (Algoritmos Genéticos) em todas as variáve is da matriz de confusão, exceto o número de Verdadeiros Negativos.

Conforme apresenta a Tabela 15 mesmo com uma melhor classificação na variável de Verdadeiros Negativos em que a estrutura de penalização é maior, o modelo M3 obteve um melhor desempenho devido ao fato de que teve um menor numero de Erros do Tipo I (Falsos Positivos), e também uma maior diferença no número Erros do Tipo II (Falsos Negativos) que acabou realizando essa compensação de desempenho por parte do modelo M3.

Pode-se dizer então que, enquanto o modelo M2 poderia ser utilizado em estratégias de recuperação orientadas a minimizar os clientes que verdadeiramente não irão pagar (*e.g.* identificação prévia de clientes ruins), seja para um tratamento especifico dessas dívidas em outro momento, ou questão de economia operacional (*e.g.* geração de menores filas em um *call-center*); o modelo M3 pode ser usado em uma estratégia de recuperação para a maximização de clientes potencialmente bons pagadores, e ao mesmo tempo para a minimização de clientes que por ventura estejam erroneamente classificados como ruins.

A Tabela 16 apresenta os resultados dos três modelos conforme, as métricas de avaliação de desempenho dos modelos como Precisão, Sensibilidade, e Especificidade.

Tabela 16 - Resultados dos modelos de Árvores de Decisão de acordo com métricas de avaliação.

Modelo	**Precisão**	**Sensibilidade (Taxa de Verdadeiros Positivos)**	**Especificidade (Taxa de Verdadeiros Negativos)**
M1	**90.41%**	**85.97%**	**96.73%**
M2	84.38%	68.70%	95.45%
M3	85.16%	80.50%	94.98%

Na Tabela 16 o modelo com todas as variáveis apresenta novamente os melhores resultados em todas as métricas, o que reforça a hipótese de que quanto mais atributos dispostos no classificador, melhor será o desempenho como visto também na Tabela 15.

Na comparação entre os modelos M2 e M3, o modelo M3 apresentou melhor desempenho nas métricas de Precisão e Sensibilidade; o que pode indicar que este modelo enquadra-se em uma estratégia de recuperação voltada à captação e identificação de potenciais

bons clientes; enquanto o modelo M2 enquadra-se em uma estratégia de recuperação voltada na identificação imediata de clientes ruins.

Na Tabela 17 um fato a ser considerado é que em termos de acurácia dos modelos a variação percentual entre o modelo com todas as variáveis, ou M1, contra o modelo com os atributos escolhidos através do Algoritmo de Johnson é de menos de 3%; o que reforça a hipótese de estar havendo algum tipo de *overfitting* por parte do modelo M1 por conta do alto número de variáveis.

Tabela 17 - Resultados dos modelos de acordo com métricas de avaliação de classificadores.

Modelo	**Acurácia**	**Coeficiente de Correlação de Matthews**
M1	**93.90%**	**0.8408**
M2	88.39%	0.6884
M3	91.16%	0.7688

Ainda na Tabela 17 através do Coeficiente de Correlação de Matthews pode ser visto que as predições estão em total acordo, isto é, há uma correlaçãoentre as variáveis presentes nas bases de dados com os resultados de classificação, o que mostra que mesmo com o desbalanceamento das classes de predição não foram gerados resultados provenientes do acaso.

Dessa forma de acordo com os resultados da Tabela 15, Tabela 16 e Tabela 17, viu-se que em termos de desempenho no geral, como era de se esperar, o modelo com todas as variáveis obteve o melhor resultado com 93.9% de classificações corretas, seguido do modelo gerado através da eliminação de atributos via *Rough Sets* através do método de Algoritmo de Johnson (91.2%) e por último os Algoritmos Genéticos (88.4%).

São apresentados a seguir os resultados das Árvores de Decisão com todas as variáveis, com as variáveis provenientes do reduto gerado pelo Algoritmo de Johnson e com as variáveis provenientes do reduto gerado pelos Algoritmos Genéticos.

Uma vez em que foram discutidos os aspectos relativos aos modelos e a sua capacidade de classificação através de uma comparação entre os modelos M1, M2 e M3, o próximo passo foi verificar cada uma das árvores de decisão geradas.

- Resultados das Árvores de Decisão com todas as variáveis

A Árvore de Decisão gerada pelo modelo M1 que tem todas as variáveis da base de dados está representada através de seu nó (*node*) raíz (nó 0) na Figura 21.

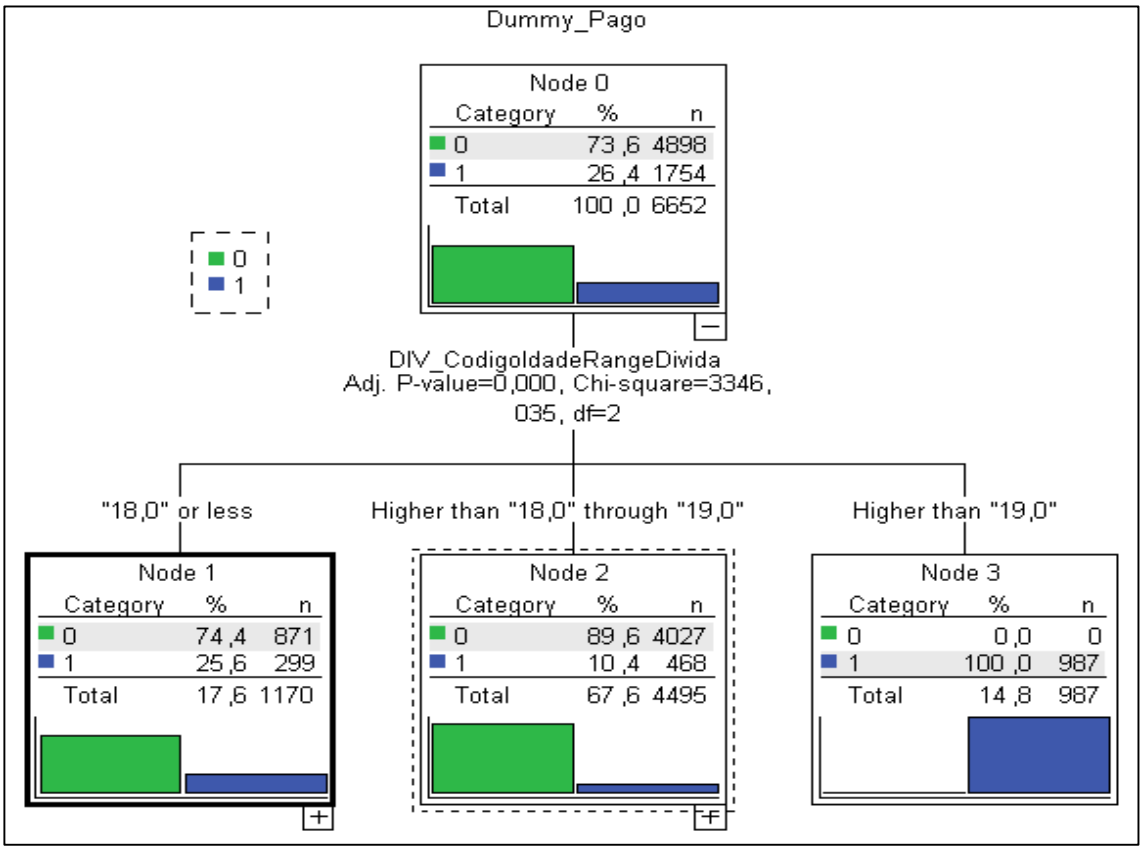


Figura 21 - Nó Raiz do Modelo 1. Fonte Elaborada pelo Autor.

A Figura 21 apresenta a primeira expansão do nó raiz que mostra que o algoritmo do CHAID escolheu como primeiro atributo de particionamento o atributo DIV_CodigoIdadeDivida. Pode-se afirmar que pelo valor do Chi-quadrado (3.346) o algoritmo CHAID considera que o atributo DIV_CodigoIdadeDivida possuí uma maior proporção de registros ligados diretamente ao evento de pagamento ou não da dívida.

O teste de decisão no atributo DIV_CodigoIdadeDivida gerou três nós folha a. “18” ou menos este código indica que a idade da dívida é menor do que 108 meses, b. Maior do que “18” até “19” indica que a dívida tem entre 108 meses até 114 meses, e c. Maior do que “19” que indica que as dívidas têm mais de 114 meses.

De acordo com os nós Folha apresentados na Figura 18 considerando os aspectos que influenciam na recuperação do crédito (*i.e.* variável Dummy_Pago = 1), pode-se extrair uma regra do tipo SE...ENTÃO, conforme descrito no capítulo II, item 2.3.3 com a seguinte regra:

SECódigo do Range relativo à Idade da dívida > 19 (Mais de 114 meses)
ENTÃO Crédito Pago (n=987)

Na expansão parcial da árvore como está disposta no Apêndice C é realizada uma expansão do nó Folha 1 (DIV_CodigoIdadeRangeDivida <= 18) e mais algumas regras relativas aos aspectos que auxiliaram na recuperação dos créditos podem ser extraídas como:

SE DIV_CodigoIdadeRangeDivida<= 18
E Código de Combinação de Canal de Contato for = 12
E a Combinação do Canal de Localização tiver Endereço
ENTÃO Crédito Pago (n=33)

SE DIV_CodigoIdadeRangeDivida<= 18
E Código de Combinação de Canal de Contato for = 1 **OU** Código de Combinação de Canal de Contato for = 14
ENTÃO Crédito Pago (n=71)

SE DIV_CodigoIdadeRangeDivida<= 18
E Código de Combinação de Canal de Contato for = 8 **OU** Código de Combinação de Canal de Contato for = 16 **OU** Código de Combinação de Canal de Contato for = 2
E Código de combinação de Canal de Localização = 4
ENTÃO Crédito Pago (n=72)

SE DIV_CodigoIdadeRangeDivida<= 18
E Código de Combinação de Canal de Contato for = 3 **OU** Código de Combinação de Canal de Contato for = 4 **OU** Código de Combinação de Canal de Contato for = 6 **OU** Código de Combinação de Canal de Contato for = 15
E Combinação de Canal de Localização Endereço = S
ENTÃO Crédito Pago (n=82)

Nesta expansão de nó folha do atributo COMBCANALCONT_MensagemVoz na Figura 22, podemos observar que em termos quantitativos há um fator que soa contra intuitivo no que diz respeito ao fato de que 71,9% dos créditos recuperados não teve esse canal de

comunicação (mensagem). Isso leva a crer que talvez esse meio de comunicação deva ser utilizado de maneira complementar em uma estratégia de recuperação dos créditos.

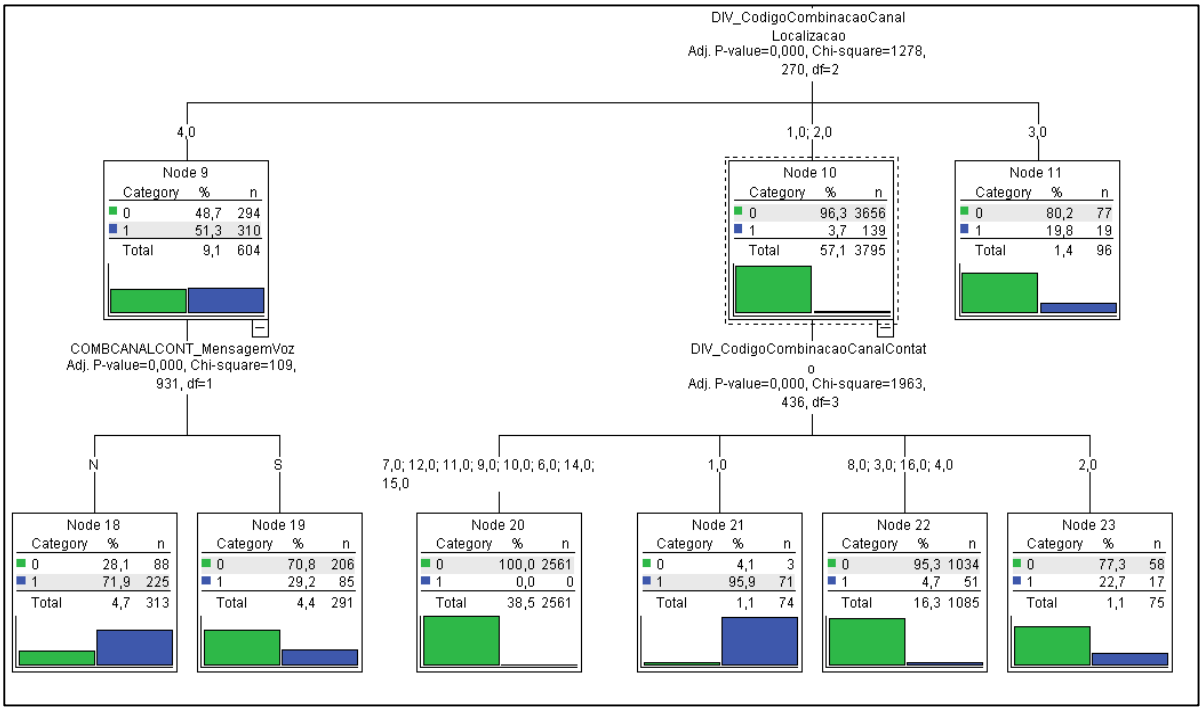


Figura 22 - Modelo 1 - Nó 2. Fonte: Elaborada pelo Autor.

Na Figura 22 é apresentada a expansão do Nó folha 2 com a apresentação das seguintes regras que elencam aspectos que influenciaram na recuperação desses créditos:

SE DIV_CodigoIdadeRangeDivida<= 18
E Código da Combinação do Canal de Localização = 4
E Combinação de Canal de Contato por Mensagem de voz = N
ENTÃO Crédito Pago (n=225)

SE DIV_CodigoIdadeRangeDivida<= 18
E Código da Combinação do Canal de Localização = 4
E Combinação de Canal de Contato por Mensagem de voz = S
ENTÃO Crédito Pago (n=85)

SE DIV_CodigoIdadeRangeDivida<= 18

E Código da Combinação do Canal de Localização = 1 **OU** Código da Combinação do Canal de Localização = 2

E Código de Combinação de Canal de Contato = 1

ENTÃO Crédito Pago (n=71)

A cobertura de cada uma das regras geradas pela Árvore de Decisão que conta com todas as variáveis em relação ao total de registros é apresentada no Quadro 25.

Quadro 25 - Regras de decisão provenientes da Árvore de Decisão com todas as variáveis.

Nr	**Descrição da Regra**	**Qte NPL Pagos**	**%Total NPL Pagos**
1	SE Código do Range relativo à Idade da dívida > 19 (Mais de 114 meses) ENTÃO Crédito Pago (n=987)	987	56,27%
2	SE DIV_CodigoIdadeRangeDivida<= 18 E Código de Combinação de Canal de Contato for = 12 E a Combinação do Canal de Localização tiver Endereço ENTÃO Crédito Pago (n=33)	33	1,88%
3	SE DIV_CodigoIdadeRangeDivida<= 18 E Código de Combinação de Canal de Contato for = 1 OU Código de Combinação de Canal de Contato for = 14 ENTÃO Crédito Pago (n=71)	71	4,05%
4	SE DIV_CodigoIdadeRangeDivida<= 18 E Código de Combinação de Canal de Contato for = 8 OU Código de Combinação de Canal de Contato for = 16 OU Código de Combinação de Canal de Contato for = 2 E Código de combinação de Canal de Localização = 4 ENTÃO Crédito Pago (n=72)	72	4,10%
5	SE DIV_CodigoIdadeRangeDivida<= 18 E Código de Combinação de Canal de Contato for = 3 OU Código de Combinação de Canal de Contato for = 4 OU Código de Combinação de Canal de Contato for = 6 OU Código de Combinação de Canal de Contato for = 15 E Combinação de Canal de Localização Endereço = S ENTÃO Crédito Pago (n=82)	82	4,68%

Nr	Descrição da Regra	Qte NPL Pagos	%Total NPL Pagos
6	SE DIV_CodigoIdadeRangeDivida<= 18 E Código da Combinação do Canal de Localização = 4 E Combinação de Canal de Contato por Mensagem de voz = N ENTÃO Crédito Pago (n=225)	225	12,83%
7	SE DIV_CodigoIdadeRangeDivida<= 18 E Código da Combinação do Canal de Localização = 4 E Combinação de Canal de Contato por Mensagem de voz = S ENTÃO Crédito Pago (n=85)	85	4,85%
8	SE DIV_CodigoIdadeRangeDivida<= 18 E Código da Combinação do Canal de Localização = 1 OU Código da Combinação do Canal de Localização = 2 E Código de Combinação de Canal de Contato = 1 ENTÃO Crédito Pago (n=71)	71	4,05%
		%Total Pago	92,70%

- Resultados das árvores com as variáveis provenientes do reduto pelo Algoritmo de Johnson

A Árvore de Decisão gerada pelo modelo M3 que tem as variáveis selecionadas de acordo com o Algoritmo de Johnson da base de dados está representada através de seu nó raíz (nó 0) no Apêndice D. Apêndice D que mostra as ramificações das árvores com as variáveis provenientes do reduto pelo Algoritmo de Johnson que apresenta logo de início a seguinte regra:

SE Data do Primeiro Atraso <= 19 de julho de 2003

ENTÃO Crédito Pago (n=685)

O Nó 0 é o nó raiz que contém a variável binária DummyPago e tem a distribuição de dívidas pagas em 26,4% e não quitadas em 73,6%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

Ao realizar as expansões da árvore como apresentado na Figura 23, o Nó 2 representa dívidas que têm a data de primeiro atraso compreendida entre 19 de julho de 2004 a 07 de janeiro de 2005 e tem a distribuição de dívidas pagas em 53,7% e não quitadas em 46,3%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

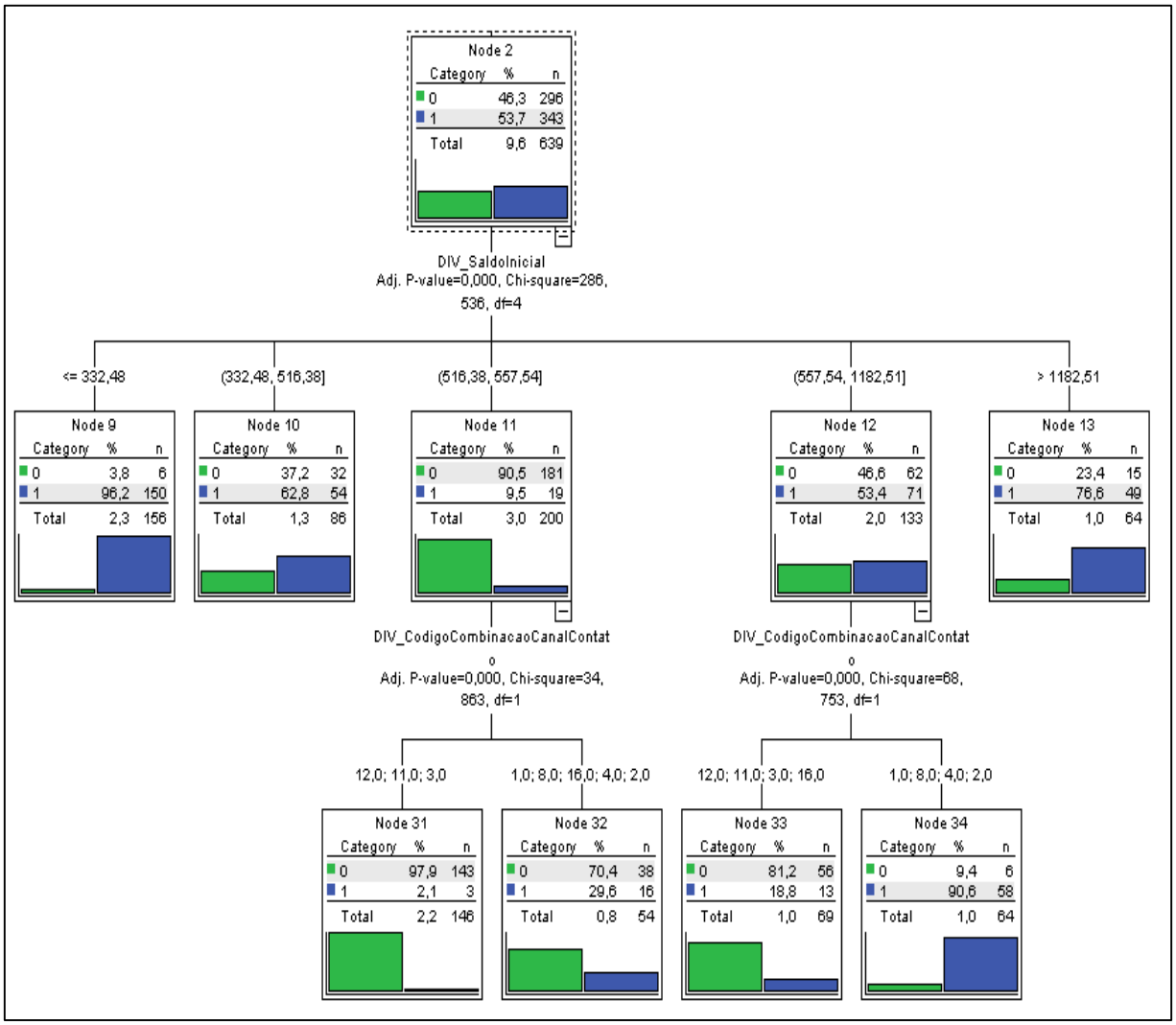


Figura 23 - Modelo 3 - Nó 2. Fonte: Elaborada pelo Autor.

A Figura 23 apresenta a expansão do nó 2 em que a divisão foi realizada com as dívidas têm a data de primeiro atraso compreendida entre 19 de julho de 2004 a 07 de janeiro de 2005, e foram obtidas as seguintes regras:

SE Data do Primeiro Atraso >= 19 de julho de 2004

EData do Primeiro Atraso <= 07 de janeiro de 2005

E Saldo Inicial <= R$ 332,48

ENTÃO Crédito Pago (n=150)

SE Data do Primeiro Atraso >= 19 de julho de 2004

EData do Primeiro Atraso <= 07 de janeiro de 2005

E Saldo Inicial >= R$ 557,54

E Saldo Inicial <= R$ 1182,51

E Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2

ENTÃO Crédito Pago (n=58)

As Figura 24, Figura 25 e Figura 26 apresentam ramificações da árvore de decisão que vão a aspectos muito específicos para o baixo retorno em termos de créditos pagos. Nestes casos, ao ser elaborada uma estratégia para estes créditos, os mesmos podem ser considerados para outra etapa de análise para verificação das dinâmicas desses créditos de forma específica.

Na Figura 24 o Nó 3 representa dívidas que têm a data de primeiro atraso compreendida entre 07 de janeiro de 2005 a 16 de janeiro de 2005 e tem a distribuição de dívidas pagas em 5,2% e não quitadas em 94,8%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

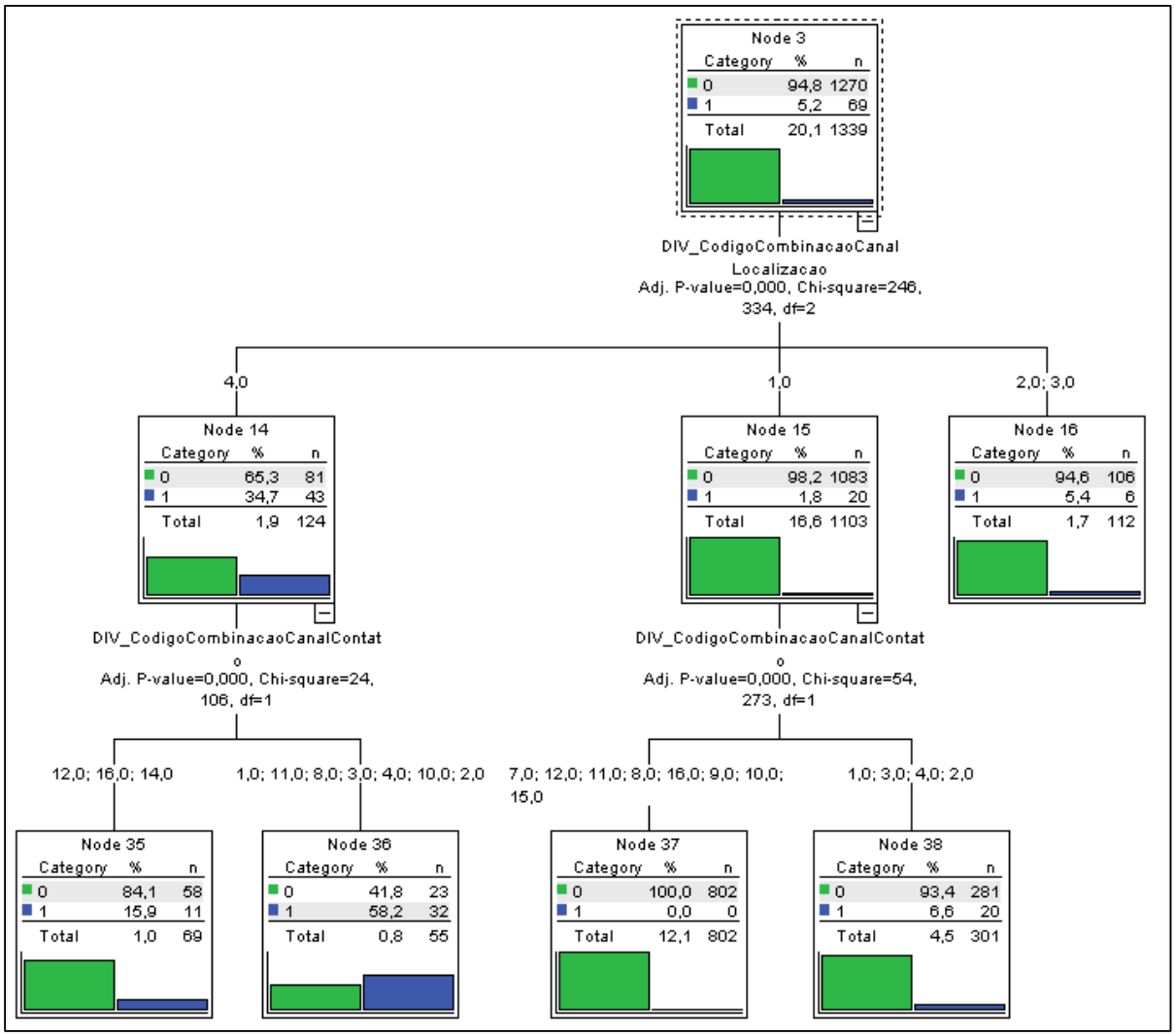


Figura 24 - Modelo 3 - Nó 3. Fonte: Elaborada pelo Autor.

A Figura 24 mostra que o Nó 4 representa dívidas que têm a data de primeiro atraso compreendida entre 16 de janeiro de 2005 a 25 de janeiro de 2005 e tem a distribuição de dívidas pagas em 7,4% e não quitadas em 92,6%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

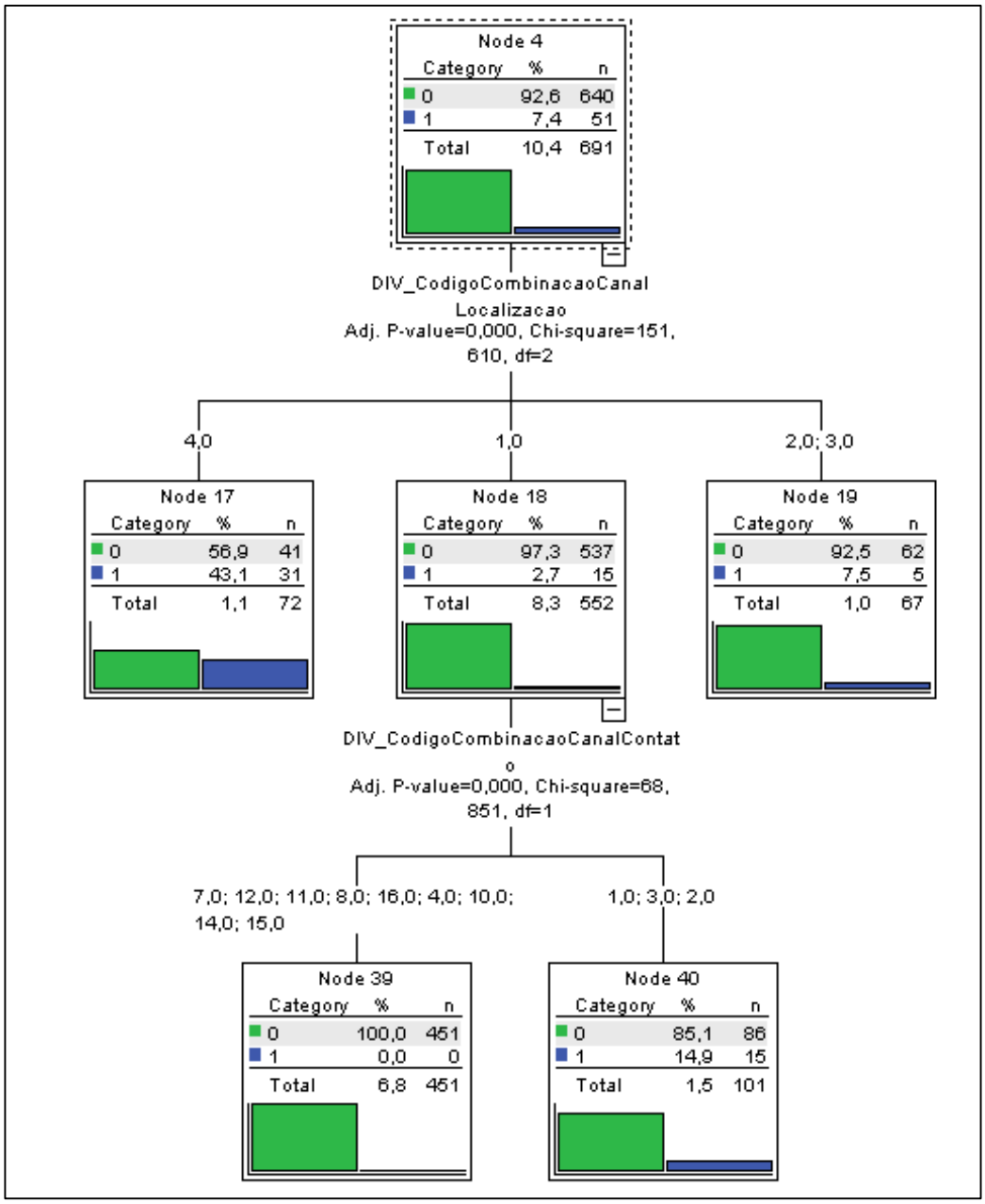


Figura 25 - Modelo 3 - Nó 4. Fonte: Elaborada pelo Autor.

De acordo com a Figura 25, o Nó 5 representa dívidas que têm a data de primeiro atraso compreendida entre 25 de janeiro de 2005 a 10 de fevereiro de 2005 e tem a distribuição de dívidas pagas em 12,2% e não quitadas em 87,8%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

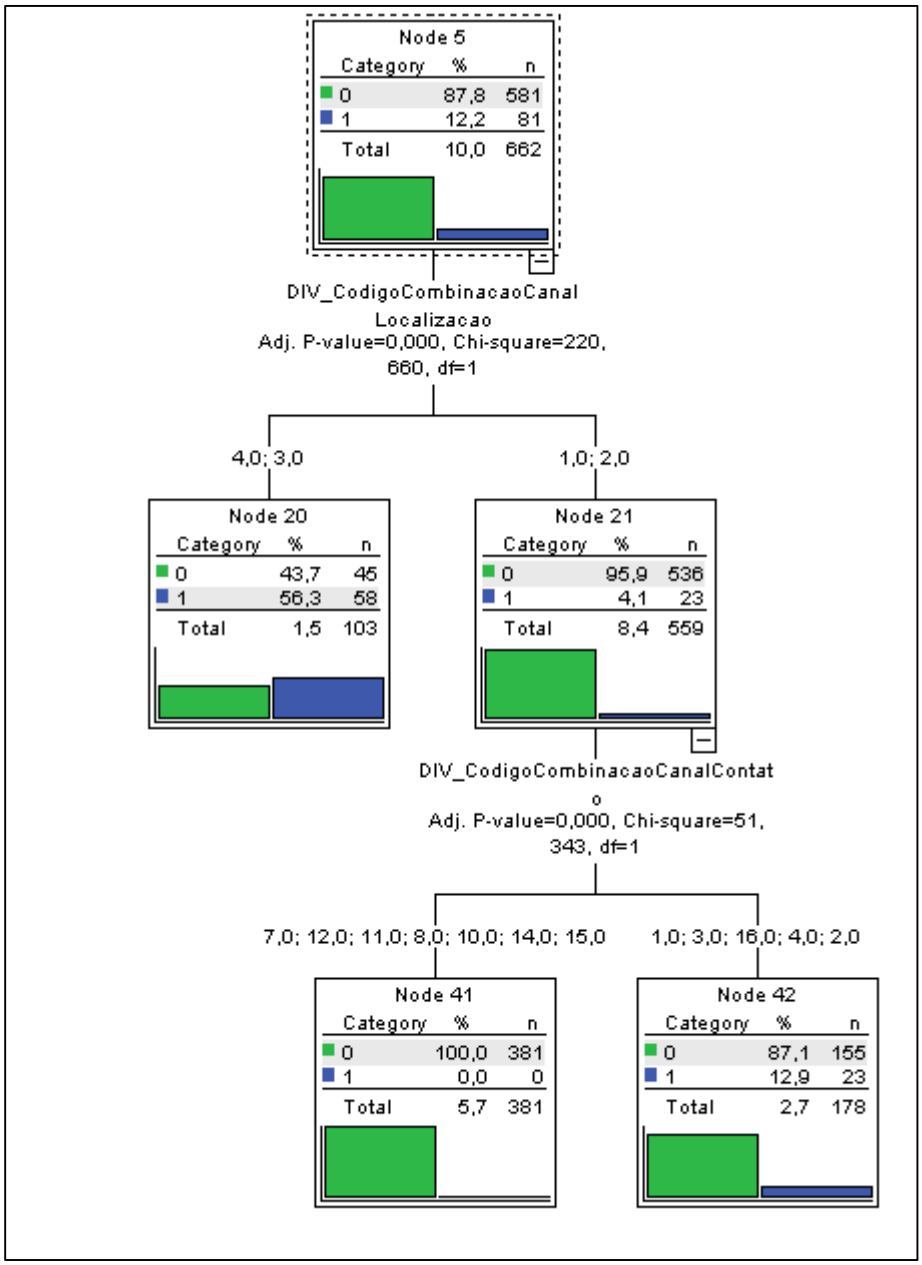


Figura 26 - Modelo 3 - Nó 5. Fonte: Elaborada pelo Autor.

Conforme a Figura 26, o Nó 6 representa dívidas que têm a data de primeiro atraso compreendida entre 10 de fevereiro de 2005 a 09 de abril de 2005 e tem a distribuição de dívidas pagas em 16,8% e não quitadas em 83,2%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

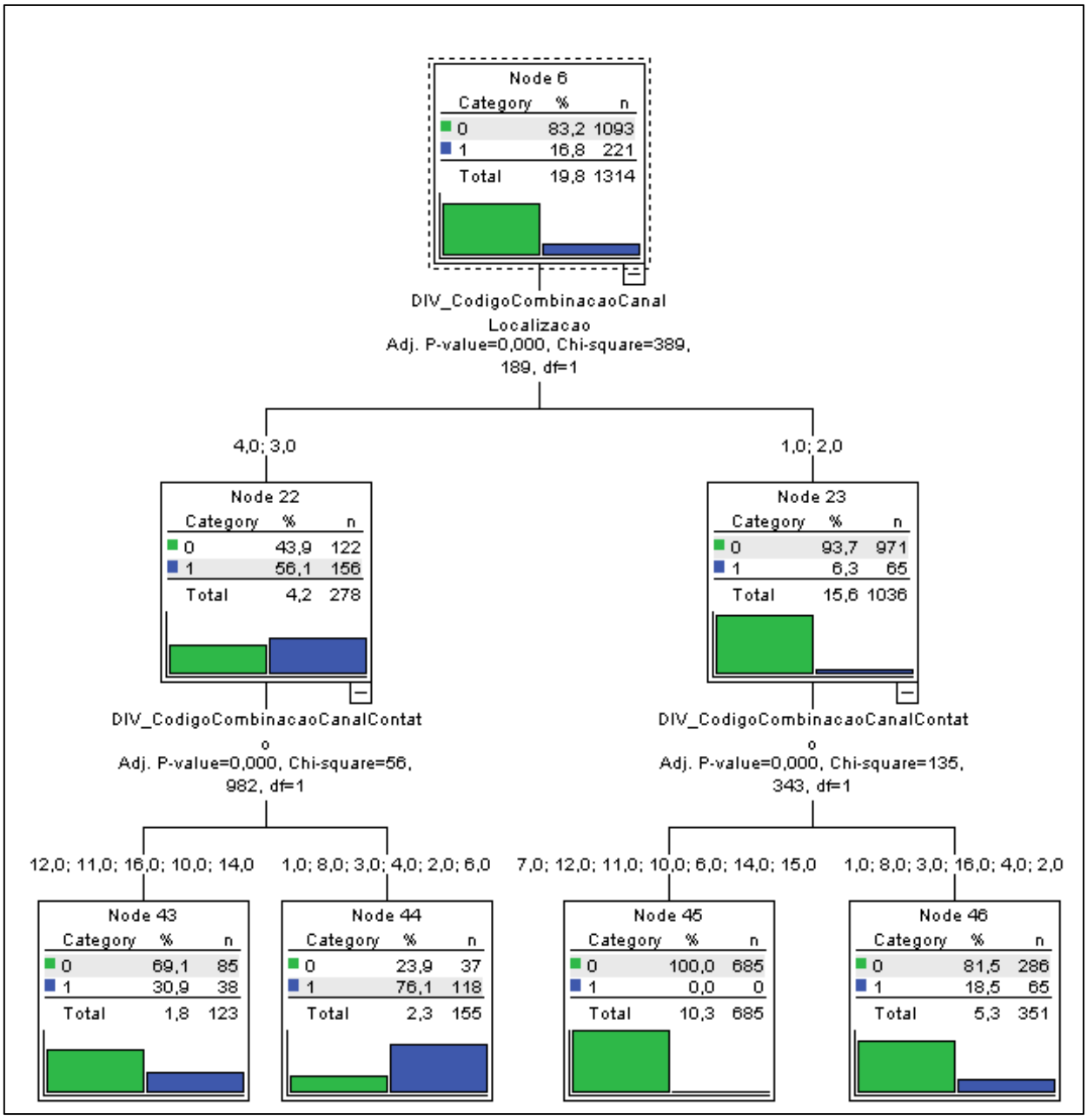


Figura 27 - Modelo 3 - Nó 6. Fonte: Elaborada pelo Autor.

Como mostram as ramificações da Árvore de Decisão contida na Figura 27 pode ser extraída a seguinte regra:

SE Data do Primeiro Atraso >= 10 de fevereiro de 2005

EData do Primeiro Atraso <= 09 de abril de 2005

E Código de Combinação de Canal de Localização = 4 **OU** Código de Combinação de Canal de Localização = 3

E Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6 **OU** Código de Combinação de Canal de Contato = 3

ENTÃO Crédito Pago (n=118)

Como está apresentada na Figura 28, o Nó 7 representa dívidas que têm a data de primeiro atraso compreendida entre 09 de abril de 2005 a 10 de junho de 2005 e tem a distribuição de dívidas pagas em 21,8% e não quitadas em 78,2%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

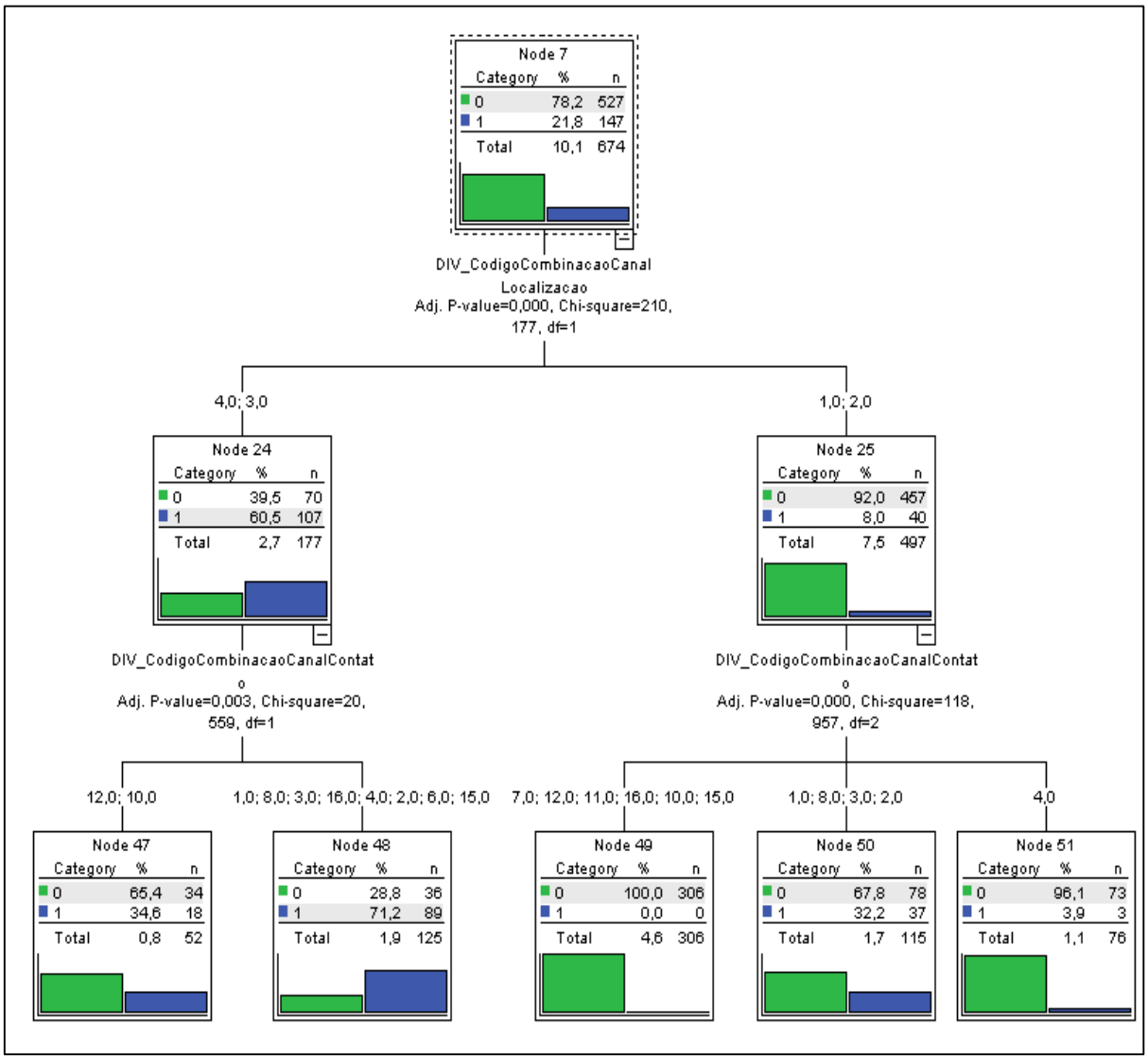


Figura 28 - Modelo 3 - Nó 7. Fonte: Elaborada pelo Autor.

Como mostram as ramificações da árvore contida Figura 28 que expande o nó 7 pode ser extraída também a seguinte regra:

SE Data do Primeiro Atraso >= 09 de abril de 2005

EData do Primeiro Atraso <= 10 de junho 2005

E Código de Combinação de Canal de Localização = 4 **OU** Código de Combinação de Canal de Localização = 3

E Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6 **OU** Código de Combinação de Canal de Contato = 3 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 15

ENTÃO Crédito Pago (n=89)

Com a expansão do nó 8 do Modelo 3 que encontra-se no Apêndice E pode ser extraída a seguinte regra de negócio:

SE Data do Primeiro Atraso > 10 de junho de 2005

E Código de Combinação de Canal de contato = 1 **OU** Código de Combinação de Canal de contato = 4

ENTÃO Crédito Pago (n=52)

Em conformidade com a Figura 28, o Nó 8 representa dívidas que têm a primeiro atraso maior do que 10 de junho de 2005 e tem a distribuição de dívidas pagas em 26,5% e não quitadas em 73,5%. O atributo utilizado para a divisão dos nós folha foi DIV_DataPrimeiroAtraso.

A cobertura de cada uma das regras geradas pela Árvore de Decisão que conta com atributos selecionados utilizando o Algoritmo de Johnson em relação ao total de registros é apresentada no Quadro 26.

Quadro 26 - Regras de decisão provenientes da Árvore de Decisão com o Algoritmo de Johnson.

Nr	Descrição da Regra	Qte NPL Pagos	%Total NPL Pagos
1	**SE** Data do Primeiro Atraso <= 19 Jul 2003 **ENTÃO** Crédito Pago (n=685)	685	39,05%
2	**SE** Data do Primeiro Atraso >= 19 Jul 04 **E** Data do Primeiro Atraso <= 07 Jan 05 **E** Saldo Inicial <= R$ 332,48 **ENTÃO** Crédito Pago (n=150)	150	8,55%
3	**SE** Data do Primeiro Atraso >= 19 Jul 04 **E** Data do Primeiro Atraso <= 07 Jan 05 **E** Saldo Inicial >= R$ 557,54 **E** Saldo Inicial <= R$ 1182,51 **E** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **ENTÃO** Crédito Pago (n=58)	58	3,31%
4	**SE** Data do Primeiro Atraso >= 10 Fev 05 **E** Data do Primeiro Atraso <= 09 Abr 05 **E** Código de Combinação de Canal de Localização = 4 **OU** Código de Combinação de Canal de Localização = 3 **E** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6 **OU** Código de Combinação de Canal de Contato = 3 **ENTÃO** Crédito Pago (n=118)	118	6,73%
5	**SE** Data do Primeiro Atraso >= 09 Abr 05 **E** Data do Primeiro Atraso <= 10 Jun 05 **E** Código de Combinação de Canal de Localização = 4 **OU** Código de Combinação de Canal de Localização = 3 **E** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6 **OU** Código de Combinação de Canal de Contato = 3 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 15 **ENTÃO** Crédito Pago (n=89)	89	5,07%

Nr	Descrição da Regra	Qte NPL Pagos	%Total NPL Pagos
6	**SE** Data do Primeiro Atraso > 10 Jun 05 **E** Código de Combinação de Canal de contato = 1 **OU** Código de Combinação de Canal de contato = 4 **ENTÃO** Crédito Pago (n=52)	52	2,96%
		%Total Pago	65,68%

- Resultados das árvores com as variáveis provenientes do reduto pelo Algoritmo Genético

A Árvore de Decisão gerada pelo modelo M2 que tem as variáveis selecionadas de acordo com o Algoritmo de Genético da base de dados está representada através de seu nó raíz (nó 0) na Figura 29.

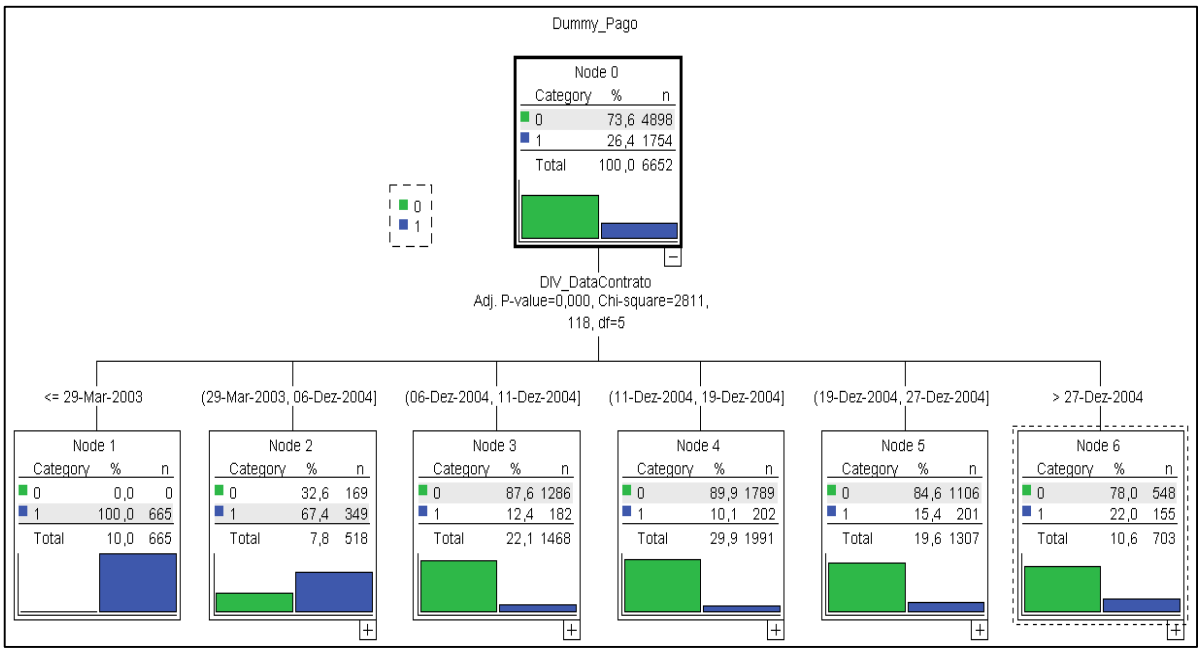


Figura 29 - Modelo 2 - Nó 0. Fonte: Elaborada pelo Autor.

Nesta primeira expansão apresentada na Figura 29 verifica-se um número de nós folha proveniente do nó raiz maior (6) do que a primeira expansão sem a eliminação dos atributos (3). Isso leva crer que *Rough Sets* realizando a eliminação de atributos, tornou a árvore mais específica em termos de discriminação de fatores, e, portanto a uma maior especificidade dos nós folha.

Nesta primeira análise pode ser visto que as dívidas recuperadas concentram-se majoritariamente nos nós 1 (665 casos) e 2 (349 casos) que correspondem ao período de tempo de dívidas com a data de contrato menor do que 6 de dezembro de 2004.

Uma hipótese para esse fato é de que de acordo com a legislação brasileira as dívidas têm um prazo de cobrança máximo de 10 anos, e optou-se uma estratégia de recuperação mais agressiva nesses créditos que poderiam ser perdidos.

Uma situação interessante ficou evidente nos nós 3, 4 e 5 que representam um espaço de tempo de apenas 21 dias no mês de dezembro de 2004 e que tem uma representatividade numérica. Houve evidências que no momento da venda dessa *tranche* a instituição bancária levou em consideração créditos contraídos muito possivelmente na época de final de ano, no qual no brasil temos historicamente um período de alto consumo seja por conta do natal ou a entrada de um ano novo.

Com a expansão do nó 2 da árvore têm-se as seguintes regras:

SE Data do Primeiro Atraso >= 29 de março de 2003

EData do Primeiro Atraso <= 6 de dezembro de 2004

E Código de Combinação de Canal de Contato = 7 **OU** Código de Combinação de Canal de Contato = 3

ENTÃO Crédito Pago (n=48)

SE Data do Primeiro Atraso >= 29 de março de 2003

EData do Primeiro Atraso <= 6 de dezembro de 2004

E Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 6

ENTÃO Crédito Pago (n=118)

SE Data do Primeiro Atraso >= 29 de março de 2003

EData do Primeiro Atraso <= 6 de dezembro de 2004

E Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2

E Saldo inicial <= R$ 646,05

ENTÃO Crédito Pago (n=85)

SE Data do Primeiro Atraso >= 29 de março de 2003

EData do Primeiro Atraso <= 6 Dez 04

ECódigo de Combinação de Canal de Contato = 8 **OU**Código de Combinação de Canal de Contato = 16 **OU**Código de Combinação de Canal de Contato = 4 **OU**Código de Combinação de Canal de Contato = 2

ESaldo inicial > R$ 646,05

ENTÃOCrédito Pago (n=75)

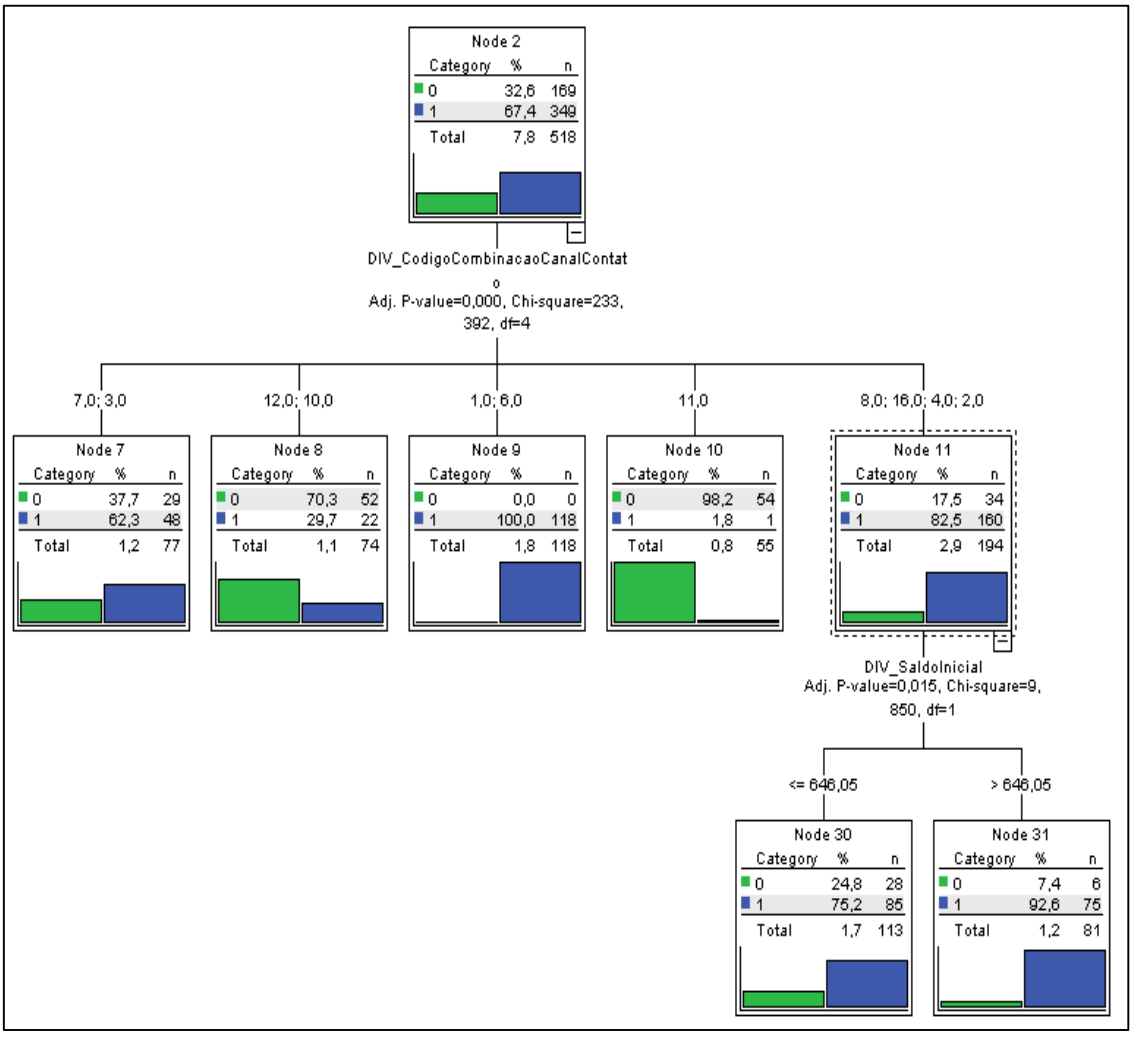


Figura 30 - Modelo 2 - Nó 2. Fonte: Elaborada pelo Autor.

Expandido o nó 2 conforme a Figura 30 que compreende de data de contrato no período de 29 de março de 2003 a 6 de dezembro de 2004, pode-se observar que o atributo utilizado para a divisão foi DIV_CodigoCombinacaoCanalContato, o qual apresenta 67,4% de créditos recuperados, o que indica que dentro desse período de tempo esse atributo pode ter influenciado positivamente na recuperação desses créditos. Dentro desse nó, pode-se ver que os códigos que

mais tiveram ocorrências de recuperação de créditos foram com os valores 1, 6, 8, 16, 4 e 2 e que em contrapartida os que tiveram valor 11 houve 98% de créditos não recuperados, o que pode sugestionar uma estratégia de recuperação orientada a dívidas que tenham este status.

Neste nó 3 do modelo 2 que representa dívidas com a data de contrato de 6 de dezembro de 2004 até 11 de dezembro 2004 pode ser visto que o atributo escolhido para realizar a divisão do nó foi DIV_CodigoCombinacaoCanalContato. Por tratar-se de um período de tempo muito curto (5 dias) esse nó mostra que com a utilização dos *Rough Sets* houve uma especificação maior dos nós folha. Grande parte das dívidas compreendidas neste nó é de débitos não recuperados. Em termos quantitativos observa-se que somente os códigos de combinação 1, 8, 2, 16, 4, e 15 representaram nós que obtiveram recuperação efetiva (Nó 14 com 52,3% (68 dívidas) e Nó 16 com 21,9% (67 dívidas).

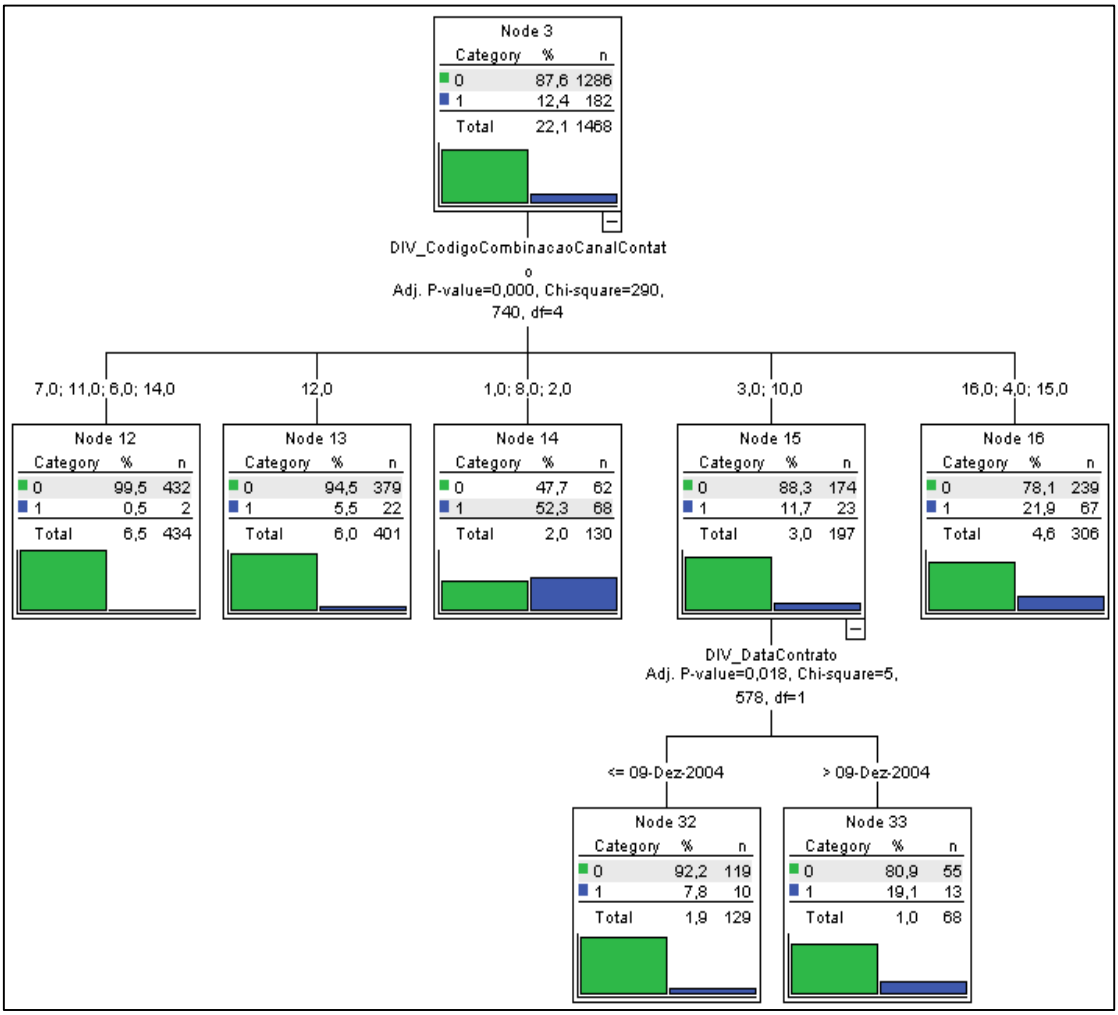


Figura 31 - Modelo 2 - Nó 3. Fonte: Elaborada pelo Autor.

A expansão do nó folha 4, que consta no Apêndice F, apresenta as seguintes regras de decisão:

SE Data do Primeiro Atraso >= 11 de dezembro de 2004

E Data do Primeiro Atraso <= 19 de dezembro de 2004

E Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 14

ENTÃO Crédito Pago (n=79)

SE Data do Primeiro Atraso >= 11 de dezembro de 2004

E Data do Primeiro Atraso <= 19 de dezembro de 2004

E Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 14

ENTÃO Crédito Pago (n=79)

De acordo com o Nó 4 dívidas que têm a data de contrato compreendida entre 11 de dezembro de 2004 a 19 de dezembro de 2004 e tem a distribuição de dívidas pagas em 10,1% e não quitadas em 89,9%. O atributo utilizado para a divisão dos nós folha foi DIV_CodigoCombinacaoCanalContato. A principal característica observada neste nó fica por conta dos códigos 1, 8, 2, 14, 16, 4, 6 que apresentam números de recuperação significativos (Nó 19 com 49,7% de recuperação e Nó 21 com 21,9% de recuperação).

Com a expansão do nó 5, constante na Figura 32 pode ser extraída a seguinte regra:

SE Data do Primeiro Atraso >= 11 de dezembro de 2004

EData do Primeiro Atraso <= 19 de dezembro de 2004

E Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 2

ENTÃO Crédito Pago (n=84)

SE Data do Primeiro Atraso >= 19 de dezembro de 2004

EData do Primeiro Atraso <= 27 de dezembro de 2004

E Código de Combinação de Canal de Contato = 3 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4

ENTÃO Crédito Pago (n=93)

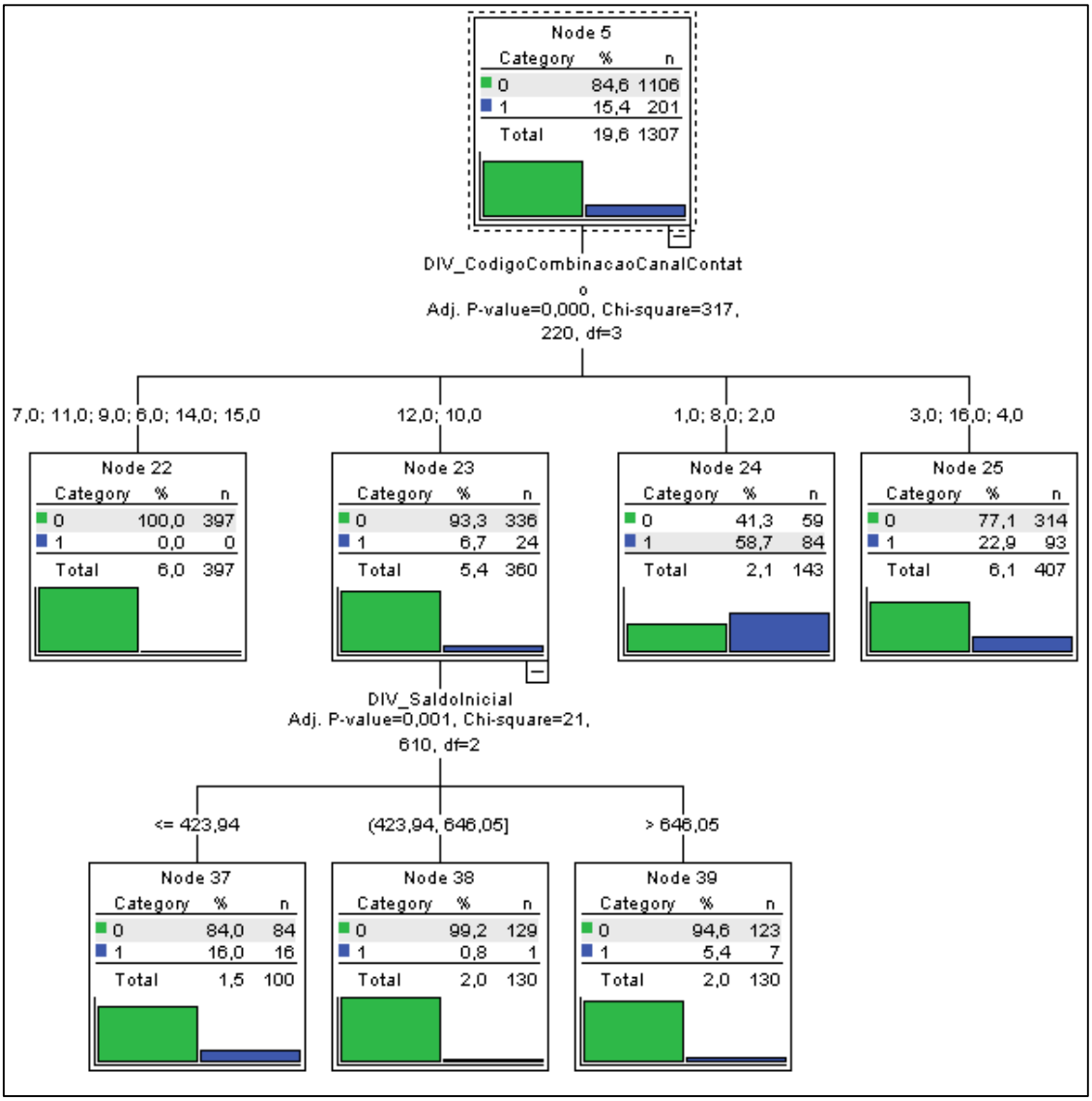


Figura 32 - Modelo 2 - Nó 5. Fonte: Elaborada pelo Autor.

A Figura 32 mostra que o Nó 5 representa dívidas que têm a data de contrato compreendida entre 19 de dezembro de 2004 a 27 de dezembro de 2004 e tem a distribuição de dívidas pagas em 15,4% e não quitadas em 84,5%. O atributo utilizado para a divisão dos nós folha foi DIV_CodigoCombinacaoCanalContato. Uma característica desse nó é que o nó folha

22 não apresentou nenhum crédito recuperado, o que pode demandar uma análise mais aprofundada para saber o porquê da não recuperação dos créditos nessa condição. Os nós folha 24 e 25 apresentaram desempenho satisfatório com respectivamente 58,7% e 22,9% de créditos recuperados em seus segmentos.

A Figura 33 mostra que o Nó 6 representa dívidas que têm a data de contrato compreendida maior que 27 de dezembro de 2004 e tem a distribuição de dívidas pagas em 22,1% e não quitadas em 78,0%. O atributo utilizado para a divisão dos nós folha foi DIV_CodigoCombinacaoCanalContato. O nó folha com o melhor desempenho em relação à recuperação foi o Nó 26 com 121 créditos recuperados (49,0%). Uma ramificação interessante desse nó foi que a métrica de DIV_SaldoInicial apareceu com o valor de divisão em R$ 516,38, o que leva a crer que esse valor pode ser utilizado como um componente estratégico nas políticas e/ou estratégias de recuperação.

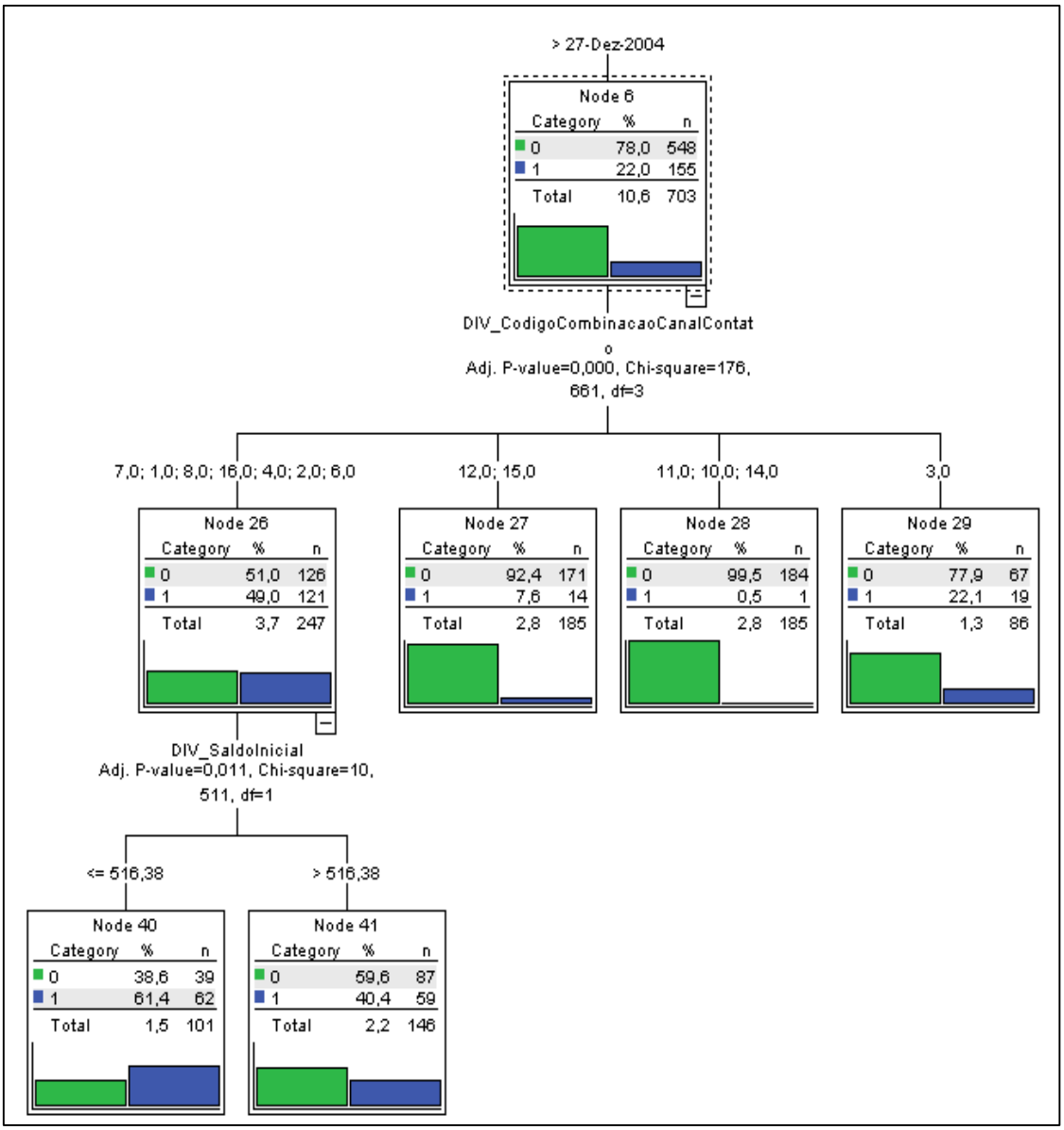


Figura 33 - Modelo 2 - Nó 6. Fonte: Elaborada pelo Autor.

Conforme a Figura 33, que apresenta a expansão do nó 6 podem ser extraídas as seguintes regras:

SE Data do Primeiro Atraso > 27 de dezembro de 2004

E Código de Combinação de Canal de Contato = 7 **OU** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6

E Saldo inicial <= R$ 516,38

ENTÃO Crédito Pago (n=62)

SE Data do Primeiro Atraso > 27 de dezembro de 2004

E Código de Combinação de Canal de Contato = 7 **OU** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6

E Saldo inicial > R$ 516,38

ENTÃO Crédito Pago (n=59)

A cobertura de cada uma das regras geradas pela Árvore de Decisão que conta com atributos selecionados utilizando os Algoritmos de Genéticos em relação ao total de registros é apresentada no Quadro 27.

Quadro 27 - Regras de decisão provenientes da Árvore de Decisão com os Algoritmos Genéticos.

Nr	Descrição da Regra	Qte NPL Pagos	%Total NPL Pagos
1	**SE** Data do Contrato <=29 Mar 2003 **ENTÃO** Crédito Pago (n=665)	665	37,91%
2	**SE** Data do Primeiro Atraso >= 29 Mar 03 **E** Data do Primeiro Atraso <= 6 Dez 04 **E** Código de Combinação de Canal de Contato = 7 **OU** Código de Combinação de Canal de Contato = 3 **ENTÃO** Crédito Pago (n=48)	48	2,74%
3	**SE** Data do Primeiro Atraso >= 29 Mar 03 **E** Data do Primeiro Atraso <= 6 Dez 04 **E** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 6 **ENTÃO** Crédito Pago (n=118)	118	6,73%
4	**SE** Data do Primeiro Atraso >= 29 Mar 03 **E** Data do Primeiro Atraso <= 6 Dez 04 **E** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **E** Saldo inicial <= R$ 646,05 **ENTÃO** Crédito Pago (n=85)	85	4,85%

Nr	Descrição da Regra	Qte NPL Pagos	%Total NPL Pagos
5	**SE** Data do Primeiro Atraso >= 29 Mar 03 **E** Data do Primeiro Atraso <= 6 Dez 04 **E** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **E** Saldo inicial > R$ 646,05 **ENTÃO** Crédito Pago (n=75)	75	4,28%
6	**SE** Data do Primeiro Atraso >= 11 Dez 04 **E** Data do Primeiro Atraso <= 19 Dez 04 **E** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 14 **ENTÃO** Crédito Pago (n=79)	79	4,50%
7	**SE** Data do Primeiro Atraso >= 11 Dez 04 **E** Data do Primeiro Atraso <= 19 Dez 04 **E** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 2 **ENTÃO** Crédito Pago (n=84)	84	4,79%
8	**SE** Data do Primeiro Atraso >= 19 Dez 04 **E** Data do Primeiro Atraso <= 27 Dez 04 **E** Código de Combinação de Canal de Contato = 3 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **ENTÃO** Crédito Pago (n=93)	93	5,30%
9	**SE** Data do Primeiro Atraso > 27 Dez 04 **E** Código de Combinação de Canal de Contato = 7 **OU** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6 **E** Saldo inicial <= R$ 516,38 **ENTÃO** Crédito Pago (n=62)	62	3,53%

Nr	Descrição da Regra	Qte NPL Pagos	%Total NPL Pagos
10	**SE** Data do Primeiro Atraso > 27 Dez 04 **E** Código de Combinação de Canal de Contato = 7 **OU** Código de Combinação de Canal de Contato = 1 **OU** Código de Combinação de Canal de Contato = 8 **OU** Código de Combinação de Canal de Contato = 16 **OU** Código de Combinação de Canal de Contato = 4 **OU** Código de Combinação de Canal de Contato = 2 **OU** Código de Combinação de Canal de Contato = 6 **E** Saldo inicial > R$ 516,38 **ENTÃO** Crédito Pago (n=59)	59	3,36%
		%Total Pago	77,99%

4.1.6 Discussão dos Resultados

Para realizar este tipo de análise foi investigada a utilização combinada de duas técnicas de inteligência computacional que são a Teoria dos *Rough Sets* e as árvores de decisão provenientes do algoritmo CHAID. Este estudo utilizou-se desses mecanismos e acredita-se que essa estrutura possa ser generalizada para créditos dessa natureza, desde que respeitadas às faixas de atraso das dívidas e os tipos de produtos.

Os experimentos realizados mostram que os *Rough Sets* exercem um papel importante na redução de atributos para efeitos de pré-processamento. Essa redução de atributos exerce uma alta redução do custo computacional em detrimento de uma baixa perda de acurácia do modelo de classificação em comparação com um modelo com todas as variáveis.

O Algoritmo de Johnson apresentou bons resultados em termos de processamento e acurácia. Nos experimentos as árvores de decisão CHAID apresentaram bons resultados em termos de acurácia e baixa complexidade em termos de geração de árvores com poucos níveis.

Com isso, cada folha da árvore possui um bom grau de interpretação e auxilia no endereçamento de recomendações para formulação de estratégias, avaliação dos ativos NPL de acordo com algumas características, e também evidenciamento de atributos chave para indicar as características dos créditos recuperados.

5 CONCLUSÕES

Neste trabalho foram aplicadas técnicas de Inteligência Computacional na análise e recuperação de portfólios de créditos do tipo *Non-Performing Loans*.

Foram realizados três experimentos descritos e resumidos a seguir:

- Experimento 1: Experimento com Redes Neurais Artificiais

- Experimento 2: Experimento com *Self-Organizing Maps* conjuntamente com a Teoria dos *Rough Sets*; e

- Experimento 3: Experimento com a Teoria dos *Rough Sets* conjuntamente com Árvores de Decisão.

No experimento 1 as RNAs aplicadas na tarefa de classificação dos NPLs apresentaram capacidade de generalização, em especial o modelo M4 que apresentou um resultado satisfatório, e que mostrou que quando utilizada uma medida de validação baseada na abordagem sensível ao custo, uma arquitetura com duas camadas escondidas reduz significativamente o número falsos positivos, e torna o modelo mais efetivo.

No experimento 2 com os *Self-Organizing Maps* aplicados na análise de *clusters* juntamente com a teoria dos *Rough Sets* aplicada na extração de regras apresentaram satisfatórios.

Os Self-Organizing Maps mostraram que um dos aspectos determinantes para o pagamento das dívidas está ligado diretamente ao Saldo da Dívida. Quanto menor a dívida maior a possibilidade de recuperação. A disposição dos *clusters* no mapa sugere que à segmentação de clientes de acordo com o seu saldo podem servir como subsídio para elaboração de estratégias de recuperação mais adequadas para cada um dos perfis.

Ainda no experimento 2, *Rough Sets* na tarefa de extração de regras mostrou que ter informações relativas à forma de localização (*e.g.* se o cliente tem telefone, endereço) e contato com o cliente (*e.g.* se o cliente recebeu uma notificação via telefonema, ou por carta) são fundamentais na recuperação desses créditos. Além disso, verificou-se nesse experimento que quanto menor for a idade da dívida menor é a capacidade de recuperação da mesma; e também que campanhas de carta sem a confirmação prévia do endereço do cliente torna essa atividade um desperdício de recursos.

No experimento 3 com a Teoria dos *Rough Sets* juntamente com as árvores de decisão o modelo proveniente da redução de atributos via algoritmo de Johnson teve o melhor desempenho em termos de tempo de processamento em decorrência da acurácia do modelo.

Este modelo apresentou evidências que aspectos ligados a localização dos clientes é uma determinante importante na recuperação, e que o canal de comunicação “SMS” mostrou baixa influência na recuperação dos NPLs.

Os experimentos mostraram que aplicações das técnicas de IC combinadas apresentam uma alta capacidade de gerar valiosos *insights* em relação à estruturação e suporte à decisão na elaboração de políticas de recuperação desses créditos dentro de uma abordagem mais micro, isto é, ao invés de analisar por agregados econômicos ou suposições administrativas em relação a fatores infrabancários o trabalho mostra os resultados a contar de um fundo de investimentos que está recuperando essas dívidas.

Além do mais, a abordagem experimental neste trabalho confirma que os créditos do tipo NPL obedecem a uma dinâmica diferente do que propõe a literatura no que diz respeito à estruturação e formas de recuperação. Além disso, o trabalho propõe que mesmo com abordagens únicas de implementação como no caso das Redes Neurais Artificias, ou com combinação de mais técnicas como nos demais experimentos; as técnicas de Inteligência Computacional mostram-se efetivas na apresentação dos elementos determinantes para a recuperação dos *Non-Performing Loans*.

A aplicação das técnicas de Inteligência Computacional auxiliou na análise e recuperação de créditos do tipo *Non-Performing Loans.* Desta forma, pode-se então considerar que o objetivo deste trabalho foi atingido.

Desta forma, o trabalho contribuiu no tema, uma vez que utiliza dados reais anonimizados de um fundo de investimentos ao invés de painéis de dados, ou mesmo agregados econômicos. Em outras palavras, o trabalho avança na literatura devido ao fato de que realiza a aplicação das técnicas de Inteligência Computacional em bases de dados reais de créditos *Non-Performing Loans* em operação para apresentar as determinantes específicas que influenciam na recuperação desses créditos.

Como continuidade dos estudos, pretende-se a realização de experimentos para a extração de conhecimento sobre aspectos que reforcem ou que tragam novos detalhes sobre determinantes que influenciam no processo de recuperação dos NPLs; e através das métricas de avaliação de modelos, tempo de processamento, e abordagem sensível ao custo determinar qual tipo de classificador é mais adequado para a construção de um modelo preditivo de créditos NPL.

Para continuidade da pesquisa está em desenvolvimento mais dois experimentos. No primeiro experimento será utilizada uma técnica regressora chamada *Multi-Adaptative Regression Splines* para levantamento de novos determinantes e verificação de poder preditivo;

e no segundo e último experimento será aplicada a técnica de *Self-Organizing Maps* para estabelecimentos de *clusters*, e posteriormente com a indicação desses *clusters* construir e comparar o desempenho de modelos preditivos com as técnicas de *Support Vector Machines*, Regressão Logística, Árvores Logísticas, e Lógica Fuzzy.

PUBLICAÇÕES DO AUTOR

Artigos completos publicados em periódicos

CLESIO, F. ; SASSI, R. J. . **Classificação de portfólio de créditos não-performados utilizando redes neurais artificiais Multilayer Perceptron.** GEPROS. Gestão da Produção, Operações e Sistemas (Online), v. 9, p. 27-40, 2014.

Trabalhos completos publicados em anais de congressos

CLESIO, F. ; SASSI, R. J. . **Aplicação de Redes Neurais Artificiais na Classificação de Créditos Não-Performados.** In: EMEPRO - Encontro Mineiro de Engenharia de Produção, 2013, Juiz de Fora. Anais do IX EMEPRO, 2013.

CLESIO, F. ; SASSI, R. J. . **Aplicação de Redes Neurais Artificiais do Tipo Multilayer Perceptron na Criação de Modelos para Classificação de um Portfólio de Crédito do Tipo Non-Performing Loan.** In: XXXIII Encontro Nacional de Engenharia de Produção - ENEGEP, 2013, Salvador. Anais ENEGEP, 2013.

CLESIO, F. ; SASSI, R. J. . **Classificação de Portfólio de Créditos Não-Performados Utilizando Redes Neurais Artificiais Multilayer Perceptron.** In: XX Simpósio de Engenharia de Produção - SIMPEP, 2013, Bauru. Anais SIMPEP, 2013.

Resumos expandidos publicados em anais de congressos

CLESIO, F. ; SASSI, R. J. . **Classificação de Créditos Não-Performados utilizando Redes Neurais Artificiais.** In: 16 SICT - Simpósio de Iniciação Científica e Tecnológica da FATEC-SP, 2014, São Paulo. Anais do 16 SICT - Simpósio de Iniciação Científica e Tecnológica da FATEC-SP, 2014.

Resumos publicados em anais de congressos

CLESIO, F. ; SASSI, R. J. . **Redes Neurais Artificiais aplicadas em uma base de dados de créditos não-performados**. In: Anais do 11º Encontro de Iniciação Científica da UNINOVE, 2014, São Paulo. Anais do 11º Encontro de Iniciação Científica da UNINOVE, 2014.

CLESIO, F. ; SASSI, R. J. . **Técnicas de Data Mining aplicadas à análise de portfólios de Créditos Não-Performados**. In: XXI Simpósio Internacional de Iniciação Científica - 21º SIICUSP, 2013, São Carlos. Anais SIICUSP, 2013.

CLESIO, F. ; SASSI, R. J. . **Técnicas de Data Mining Aplicadas à Análise de Portfólios de Créditos Não-Performados**. In: 10º Encontro de Iniciação Científica da UNINOVE, 2013, São Paulo. Anais do 10º Encontro de Iniciação Científica da UNINOVE, 2013.

REFERÊNCIAS BIBLIOGRÁFICAS

AHMAD, Nor Bahiah Hj; SHAMSUDDIN, Siti Mariyam; ABRAHAM, Ajith.**Granular Mining of Student's Learning Behavior in Learning Management System Using Rough Set Technique.**Em: Computational Intelligence for Technology Enhanced Learning. Springer Berlin Heidelberg, 2010. p. 99-124.

ALLEN, Franklin; CARLETTI, Elena.**Mark-to-market accounting and liquidity pricing.** Journal of accounting and economics, v. 45, n. 2, p. 358-378, 2008.

ALTON, R. G.; HAZEN, J. H. **As Economy Flounders, Do We See A Rise in Problem Loans???**, Federal Reserve Bank of St. Louis. 2001. Disponível em: <http://www.stlouisfed.org/publications/cb/articles/?id=1478> Acesso em:4 jun. 2014.

BANCO CENTRAL DO BRASIL. **Altera, no COSIF, procedimentos para registro das operações de credito e constituição de provisão para fazer face aos créditos de liquidação duvidosa.** 2000. Carta Circular 2.899, de 1/3/2000 Disponível em: <http://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=C_Circ&ano=2000&numero=2899 >Acesso em:24 mar.2013.

BANCO CENTRAL DO BRASIL. **Ata da 168ª Reunião do Comitê de Política Monetária (COPOM)**. 2012. Disponível em: <http://www.bcb.gov.br/?copom168>Acesso em: 4 nov. 12.

BANCO CENTRAL DO BRASIL. **Circular 3.098, de 20/3/2002 - Dispõe sobre a remessa adicional de informações no âmbito do sistema Central de Risco de Crédito**. 2002. Disponível em: <http://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=circ&ano=2002&numero=3098> Acesso em: 4 nov.2012.

BANCO CENTRAL DO BRASIL. **Comunicado Nº 21.928, de 25 de Janeiro de 2012 - Divulga autorização para funcionamento da Central de Cessão de Crédito - C3.** 2012. Disponível em: <https://www3.bcb.gov.br/normativo/detalharNormativo.do?method=detalharNormativo&N=112005200 > Acesso em: 4 nov.2012.

BANCO CENTRAL DO BRASIL. **Dispõe sobre critérios de classificação das operações de crédito e regras para constituição de provisão para créditos de liquidação duvidosa.** 1999. Resolução 2.682, de 21/12/1999 - http://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=Res&ano=1999&numero=2682 > Acesso em: 24 mar.2013.

BANCO CENTRAL DO BRASIL. **Inadimplência no Setor Bancário: uma avaliação e suas medidas.** 2009. Disponível em: < http://www.bcb.gov.br/pec/wps/port/wps192.pdf> Acesso em: 2 mar.2012.

BANCO CENTRAL DO BRASIL. **Resolução 2.907, de 29/11/2001 - Autoriza a constituição e o funcionamento de fundos de investimento em direitos creditórios e de fundos de aplicação em quotas de fundos de investimento em direitos creditórios.** 2001. Disponível em:

<http://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=res&ano=2001&numero=2907> Acesso em: 4 nov.2012.

BANCO CENTRAL DO BRASIL. **Resolução 3.334, de 22/12/2005 - Estabelece normas a serem observadas pelas instituições financeiras e demais instituições autorizadas relativas a fundos de investimento. 2005.** Disponível em: <http://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=res&ano=2005&numero=3334> Acesso em: 4 nov.2012.

BANCO CENTRAL DO BRASIL. **Resolução 2.682, de 21/12/1999 -Dispõe sobre critérios de classificação das operações de crédito e regras para constituição de provisão para créditos de liquidação duvidosa.** 2001. Disponível em: <http://www.bcb.gov.br/pre/normativos/res/1999/pdf/res_2682_v2_L.pdf> Acesso em: 4nov.2012.

BANCO CENTRAL DO BRASIL. **Resolução 2.907, de 29/11/2001 - Autoriza a constituição e o funcionamento de fundos de investimento em direitos creditórios e de fundos de aplicação em quotas de fundos de investimento em direitos creditórios.** 2001. Disponível em: <http://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=res&ano=2001&numero=2907> Acesso em: 4 nov.2012.

BAZAN J.; NGUYEN, H. S.; NGUYEN, S. H.; SYNAK, P.; WRÓBLEWSKI, J. **Rough set algorithms in classification problem.** Em: L. Polkowski, S. Tsumoto, and T. Lin, editors, Rough Set Methods and Applications, Physica-Verlag, Heidelberg New York, pp. 49–88, 2000

BECK, Roland; JAKUBIK, Petr; PILOIU, Anamaria. Non-performing loans: what matters in addition to the economic cycle? In: EUROPEAN CENTRAL BANK, Working Paper Series. n. 1515. 2013.

BERGER, Allen N.; DEYOUNG, Robert. **Problem loans and cost efficiency in commercial banks.** Journal of Banking & Finance, v. 21, n. 6, p. 849-870, 1997.

BENSIC, Mirta; SARLIJA, Natasa; ZEKIC-SUSAC, Marijana. **Modelling small-business credit scoring by using logistic regression, neural networks and decision trees.** Intelligent Systems in Accounting, Finance and Management, v. 13, n. 3, p. 133-150, 2005.

BEZDEK, James C. **What is computational intelligence.** Computational Intelligence: Imitating Life, p. 1-12, 1994.

BISHOP, C. **Pattern Recognition and Machine Learning.**Springer, 2006.

BM&F BOVESPA. **O que são FIDICs? 2013.** Disponível em: <http://www.bmfbovespa.com.br/pt-br/renda-fixa/o-que-sao-fidcs.aspx?idioma=pt-br >> Acesso em: 24 mar. 2013.

BUSSAB, Wilton de Oliveira; BOLFARINE, Heleno. **Elementos de amostragem.** São Paulo: Edgard Blücher, ABE, 2005.

BOLLINGER, Audrey S.; SMITH, Robert D. **Managing organizational knowledge as a strategic asset.** Journal of knowledge management, v. 5, n. 1, p. 8-18, 2001.

BONIN, John; HASAN, Iftekhar; WACHTEL, Paul.**Banking in transition countries.** 2008.Disponível em: < http://www.suomenpankki.fi/bofit/tutkimus/tutkimusjulkaisut/dp/Documents/dp1208.pdf >Acesso em: 4 jun. 2014.

BREIMAN, Leo et al. **Classification and regression trees.** CRC press, 1984.

CELEUX, Gilles; GOVAERT, Gérard.**A classification EM algorithm for clustering and two stochastic versions.** Computational statistics & Data analysis, v. 14, n. 3, p. 315-332, 1992.

CHACKO, George et al. **Credit derivatives.** Philadelphia: Wharton School, 2006.

CHAPLIN, Geoff. **Credit Derivatives: Trading, Investing, and Risk Management.**John Wiley & Sons, 2010.

CHEN, S. C.; HUANG, M. Y. **Constructing credit auditing and control & management model with data mining technique.** Expert Systems with Applications, v. 38, n. 5, p. 5359-5365, 2011.

CHO, Sungbin; HONG, Hyojung; HA, Byoung-Chun.**A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction.** Expert Systems with Applications, v. 37, n. 4, p. 3482-3488, 2010.

COMISSÃO DE VALORES MOBILIÁRIOS – CVM. **INSTRUÇÃO CVM Nº 444, DE 8 DE DEZEMBRO DE 2006. Dispõe sobre o funcionamento de Fundos de Investimento em Direitos Creditórios Não-Padronizados. 2006.** Dispobível em < http://www.abdir.com.br/legislacao/legislacao_abdir_12_12_1.doc > Acesso em: 3mar.2013.

COMISSÃO DE VALORES MOBILIÁRIOS – CVM**. INSTRUÇÃO CVM Nº 471, DE 8 DE AGOSTO DE 2008. Dispõe sobre o procedimento simplificado para registro de ofertas públicas de distribuição de valores mobiliários.** 2008. Dispobível em <http://www.cvm.gov.br/port/infos/inst471.doc > Acesso em: 3mar.2013.

COMISSÃO DE VALORES MOBILIÁRIOS – CVM. **INSTRUÇÃO CVM No 356, DE 17 DE DEZEMBRO DE 2001. Regulamenta a constituição e o funcionamento de fundos de investimento em direitos creditórios e de fundos de investimento em cotas de fundos de investimento em direitos creditórios.** 2001. Dispobível em < http://www.cvm.gov.br/asp/cvmwww/Atos/Atos/inst/inst356consolid.doc > Acesso em: 3mar.2013.

COMISSÃO DE VALORES MOBILIÁRIOS – CVM. **INSTRUÇÃO CVM Nº 489, DE 14 DE JANEIRO DE 2011. Dispõe Dispõe sobre a elaboração e divulgação das Demonstrações Financeiras dos Fundos de Investimento em Direitos Creditórios – FIDC e dos Fundos de Investimento em Cotas de Fundos de Investimento em Direitos Creditórios – FIC-FIDC, regidos pela Instrução CVM nº 356, de 17 de dezembro de 2001, dos Fundos de Investimento em Direitos Creditórios no âmbito do Programa de Incentivo**

à Implementação de Projetos de Interesse Social – FIDC-PIPS, regidos pela Instrução CVM nº 399, de 21 de novembro de 2003 e dos Fundos de Investimento em Direitos Creditórios Não Padronizados – FIDC-NP, regidos pela Instrução CVM nº 444, de 8 de dezembro de 2006. 2011. Disponível em: < www.cvm.gov.br/asp/cvmwww/Atos/Atos/inst/inst356consolid.doc > Acesso em: 3mar.2013.

COMISSÃO DE VALORES MOBILIÁRIOS– CVM. **INSTRUÇÃO CVM Nº 356, DE 17 DE DEZEMBRO DE 2001. Regulamenta a constituição e o funcionamento de fundos de investimento em direitos creditórios e de fundos de investimento em cotas de fundos de investimento em direitos creditórios.** 2001. Disponível em: <http://www.cvm.gov.br/asp/cvmwww/atos/exiato.asp?File=%5Cinst%5Cinst356.htm> Acesso em: 4 nov.2012.

COOPER, Harris M. **Organizing knowledge syntheses: A taxonomy of literature reviews.** Knowledge in Society, v. 1, n. 1, p. 104-126, 1988.

CORTAVARRIA, Luis et al. **Loan review, provisioning, and macroeconomic linkages.** 2000. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=880313 >Acesso em: 04 jun. 2014.

CRAENEN, B.; EIBEN, A. **Computational intelligence.**Encyclopedia of Life Support Sciences. EOLSS, EOLSS Co. Ltd, 2002.

CRESWELL, J.W. **Research design – qualitative, quantitative and mixed methods approaches.** 3. ed. Thousand Oaks, CA: Sage, 2009.

DIETTERICH, Thomas G. **An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization.** Machine learning, v. 40, n. 2, p. 139-157, 2000.

DRCHAL, Jan. Artificial **Neural Networks: MLP, Backpropagation.**Computational Intelligence Group. 2012. Disponível em::< http://cw.felk.cvut.cz/lib/exe/fetch.php/courses/a4m33bia/a4m33bia-03backprop-2012.pdf > Acesso em: 13 abr. 2013.

DUCH, Włodzisław. **What is Computational Intelligence and where is it going?.**Em: Challenges for computational intelligence. Springer Berlin Heidelberg, 2007. p. 1-13.

ENGELBRECHT, Andries P. **Computational intelligence: an introduction. J**ohn Wiley & Sons, 2007.

ESPINOZA, Raphael A.; PRASAD, Ananthakrishnan. **Nonperforming loans in the GCC banking system and their macroeconomic effects.** International Monetary Fund, 2010. Disponível em: < http://core.kmi.open.ac.uk/download/pdf/6544833.pdf>Acesso em: 04 jun. 2014.

EUROPEAN CENTRAL BANK. **Non-Performing Loans: What matters in addition to the economic cycle?** Working Paper Series. Nr 1515, 2013.

FACELI, K. et al**. Inteligência Artificial: Uma abordagem de aprendizado de Maquina.**Rio de Janeiro: Ltc, 2011.

FARHAN, Muhammad et al. **Economic Determinants of Non-Performing Loans: Perception of Pakistani Bankers.** European Journal of Business and Management, v. 4, n. 19, p. 87-99, 2012.

FÁVERO, L. P. et al. **Análise de Dados: Modelagem Multivariada para Tomada de Decisões**. Rio de Janeiro: Editora Campus, 2009.

FAYYAD, Usama M. et al. **Advances in knowledge discovery and data mining.**1996. Disponível em: < http://shawndra.pbworks.com/f/The%20KDD%20process%20for%20extracting%20useful%20knowledge%20from%20volumes%20of%20data.pdf > Acesso em: 24 mar.2013.

FERGUSON, Thomas P. **Observations on the Securitization of Non-Performing Loans in Russia.** Bucerius Law Journal, Bucerius Law School, Hamburg, Germany, 2008. Disponível em: <http://ssrn.com/abstract=1017288>Acesso em: 04 jun. 2014.

FOFACK, Hippolyte. **Nonperforming loans in Sub-Saharan Africa: causal analysis and macroeconomic implications.** World Bank Policy Research Working Paper, n. 3769, 2005.Disponível em: <https://openknowledge.worldbank.org/bitstream/handle/10986/8498/wps3769.pdf?sequence=1>Acesso em: 04 jun. 2014.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. **Knowledge discovery in databases: An overview.** AI magazine, v. 13, n. 3, p. 57, 1992.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. **Técnica de mineração de dados: uma revisão da literatura.** Acta Paulista de Enfermagem, v. 22, n. 5, p. 686-690, 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-21002009000500014&lng=en&nrm=iso> Acesso em: 2 mar.2013.

GIL, A.C. **Como elaborar projetos de pesquisa.** 4ª. ed. São Paulo: Atlas S/A.

GILLO, M. W. **Maid, a honeywell 600 program for an automatized survey analysis.** Behavioral Science, v. 17, n. 2, p. 251-&, 1972.

GILLO, Martin W.; SHELLY, Maynard W. **Predictive modeling of multivariable and multivariate data.** Journal of the American Statistical Association, v. 69, n. 347, p. 646-653, 1974.Disponível em <http://www.jstor.org/discover/10.2307/2285995?uid=3737664&uid=2&uid=4&sid=21104150695847 >Acesso em: 22 set.14.

GREENIDGE, Kevin; GROSVENOR, Tiffany.**Forecasting non-performing loans in Barbados.** Journal of Banking, Finance and Economics in Emerging Economies, June, 2010.Disponível em: <http://www.centralbank.org.bb/WEBCBB.nsf/vwPublications/53961B3EE2EEA1EA042577F2005E7CE6/$FILE/Forecasting%20Non-Performing%20Loans%20in%20Barbados.pdf>Acesso em: 04 jun. 2014.

GRZYMAŁA-BUSSE, J.**A New Version of the Rule Induction System LERS**. Fundamenta Informaticae, Vol. 31(1), pp. 27–39, 1997

HARRELL, Frank E**. Regression Modeling Strategies.**Springer-Verlag, 2001.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome.**The Elements of Statistical Learning: Data Mining, Inference, and Prediction.**2ª Edição.Springer Series in Statistics.Springer. 2009.

HAYKIN, S. **Neural Networks: A comprenhensive Foundation.** New York: Wiley & Sons, 1994.

HE, M.; LIU, N.; XIA, E. **Discrimination for non-performing loans recovery: a method of support vector machines based on wavelet transform.** Anais Third International Symposium on Information Science and Engineering, p. 88 - 92, 2010

HEATON, John C.; LUCAS, Deborah; MCDONALD, Robert L.**Is mark-to-market accounting destabilizing?** Analysis and implications for policy.Journal of Monetary Economics, v. 57, n. 1, p. 64-75, 2010.

HERRERIAS, Renata; MORENO, Jorge O. **Spillovers and Long Run Diffusion of Non-Performing Loans Risk.**Em: Midwest Finance Association 2012 Annual Meetings Paper. 2012.Disponível em::<http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1927982>Acesso em: 04 jun. 2014.

HVIDSTEN, Torgeir R. **A tutorial-based guide to the ROSETTA system: A Rough Set Toolkit for Analysis of Data.** 2006.Disponível em: < http://www.trhvidsten.com/docs/ROSETTATutorial.pdf>Acesso em: 3 jun.2014.

IMPAVIDO, Gregorio; KLINGEN, Christoph A.; SUN, Yan.**Non Performing Loans and the Macroeconomy.** 2012. Disponível em: < http://ssrn.com/abstract=2156325 >Acesso em: 3 jun. 2014.

INTERNATIONAL FINANCE CORPORATION. **NPL and Risk Management in ECA.**2012. Disponível em: <https://www.wbginvestmentclimate.org/advisory-services/regulatory-simplification/debt-resolution-and-business-exit/upload/IFC-contribution-to-foster-financial-stability-in-ECA.pdf>Acesso em: 3 jun.2014.

INTERNATIONAL MONETARY FUND.**Non-Performing Loans in CESEE: Determinants and Impact on Macroeconomic Performance.** Nir KLEIN. IMF Working Paper WP/13/72. 2013.

JOHNSON, David S. **Approximation algorithms for combinatorial problems. In:Proceedings of the fifth annual ACM symposium on Theory of computing.**ACM, 1973. p. 38-49. Disponível em: < http://www.cse.nd.edu/~izaguirr/papers/simon/ChenBang%20-%20proteinNetwork/literatures/SetCovering/Johnson.pdf >Acesso em: 3 jun.2014.

JORDAN, Alwyn; TUCKER, Carisma.**Assessing the Impact of Nonperforming Loans on Economic Growth in The Bahamas.** Monetaria, v. 1, n. 2, p. 371-400, 2013. Disponível em: <>Acesso em: 04 jun. 2014. http://www.cemla.org/PDF/monetaria/PUB-MON-I-02-04.pdf

KALLUCI, Irini; KODRA, Oriela.**Macroeconomic determinants of credit risk: The case of Albania Economic Policies in SEE: Design, performance and challenges.**Em: IMF Working Paper WP/11/161, 2010. Disponível em: <http://www.bankofalbania.org/previewdoc.php?crd=6200&ln=2&uni=20110706163638378 1 >Acesso em: 04 jun. 2014.

KASS, Gordon V. **An exploratory technique for investigating large quantities of categorical data.** Applied statistics, p. 119-127, 1980.

KAUPA, P. H.; SASSI, R. J. **Rough Sets: Técnica da Inteligência Computacional Aplicado no Mercado Financeiro.** Anais XX Simpósio de Engenharia de Produção, 2013.

KHEMRAJ, Tarron; PASHA, Sukrishnalall. **The determinants of non-performing loans: an econometric case study of Guyana.**2009. Disponível em: <http://mpra.ub.uni-muenchen.de/53128/1/MPRA_paper_53128.pdf>Acesso em: 04 jun. 2014.

KLEIN, Nir. **Non-Performing Loans in CESEE: Determinants and Impact on Macroeconomic Performance.**Em: INTERNATIONAL MONETARY FUND, IMF Working Paper WP/13/72, 2013.Disponível em: <http://www.imf.org/external/pubs/ft/wp/2013/wp1372.pdf>Acesso em: 04 jun. 2014.

KOEHN, Philipp. **Combining genetic algorithms and neural networks: The encoding problem.** 1994.

KOHAVI, R. ;PROVOST, F. **Machine Learning.** 2011.

KOHONEN, Teuvo. **Self-organized formation of topologically correct feature maps. Biological cybernetics**, v. 43, n. 1, p. 59-69, 1982.

LANDWEHR, Niels; HALL, Mark; FRANK, Eibe.**Logistic Model Trees.**Anais Proceedings of the 14th European Conference on Machine Learning (ECML-2003), Cavtat, Croária, 2003.

LANDWEHR, Niels; HALL, Mark; FRANK, Eibe.**Logistic model trees.**Em:Machine Learning: ECML 2003. Springer Berlin Heidelberg, 2003. p. 241-252.

LANDWEHR, Niels; HALL, Mark; FRANK, Eibe.**Logistic model trees.**Machine Learning, v. 59, n. 1-2, p. 161-205, 2005.

LIN, Feng Yu; MCCLEAN, Sally. **A data mining approach to the prediction of corporate failure.** Knowledge-based systems, v. 14, n. 3, p. 189-195, 2001.

LOTERMAN, Gert et al. **Benchmarking regression algorithms for loss given default modeling.** International Journal of Forecasting, v. 28, n. 1, p. 161-170, 2012.

LOUZIS, Dimitrios P.; VOULDIS, Angelos T.; METAXAS, Vasilios L. **Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of**

mortgage, business and consumer loan portfolios. Journal of Banking & Finance, v. 36, n. 4, p. 1012-1027, 2012.

LOUZIS, Dimitrios P.; VOULDIS, Angelos T.; METAXAS, Vasilios L. **Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios.** Journal of Banking & Finance, v. 36, n.4,pp. 1012-1027, 2012. Disponível em:<http://www.bankofgreece.gr/BogEkdoseis/Paper2010118.pdf >Acesso em: 04 jun. 2014.

MACKWORTH, Alan K.; GOEBEL, Randy G.; POOLE, David I. **Computational intelligence: a logical approach.** 1998.

MAGIDSON, Jay; VERMUNT, Jeroen K.**An extension of the CHAID tree-based segmentation algorithm to multiple dependent variables.** In:Classification—the Ubiquitous Challenge. Springer Berlin Heidelberg, 2005. p. 176-183.

MAKRI, Vasiliki; TSAGKANOS, Athanasios; BELLAS, Athanasios.**Determinants of non-performing loans: The case of Eurozone.**Panoeconomicus, v. 61, n. 2, p. 193-206, 2014.Disponível em: <http://www.doiserbia.nb.rs/img/doi/1452-595X/2014/1452-595X1402193M.pdf
>Acesso em: 04 jun. 2014.

MANDALA, I.; NAWANGPALUPI, Catharina Badra; PRAKTIKTO, Fransiscus Rian. **Assessing Credit Risk: An Application of Data Mining in a Rural Bank**.Procedia Economics and Finance, v. 4, p. 406-412, 2012.

MARCONI, M.A.; LAKATOS, E.M. **Fundamentos de metodologia científica.**7. ed. São Paulo, Atlas, 2010.

MARKOWITZ, Harry. **Portfolio selection.** The journal of finance, v. 7, n. 1, p. 77-91, 1952.Disponível em: <http://www.jstor.org/discover/10.2307/2975974?uid=3737664&uid=2&uid=4&sid=21101754273091 > Acesso em: 2 mar.2013.

MATUSZYK, Ania; MUES, Christophe; THOMAS, Lyn C**. Modelling LGD for unsecured personal loans: Decision tree approach.** Journal of the Operational Research Society, v. 61, n. 3, p. 393-398, 2010.

MEEKER, Larry G.; GRAY, Laura.**A note on non-performing loans as an indicator of asset quality.** Journal of banking & finance, v. 11, n. 1, p. 161-168, 1987. Disponível em: <http://www.sciencedirect.com/science/article/pii/0378426687900288
>Acesso em: 04 jun. 2014.

MITCHELL, Tom M. **Machine learning.** 1997. Burr Ridge, IL: McGraw Hill, v. 45, 1997.

MONTGOMERY, Douglas C. **Design and analysis of experiments.**John Wiley & Sons, 2008.

MORGAN, James N.; MESSENGER, Robert C. THAID: **A sequential analysis program for the analysis of nominal scale dependent variables.** 1973.

MORGAN, James N.; SONQUIST, John A. **Problems in the analysis of survey data, and a proposal. Journal of the American statistical association,** v. 58, n. 302, p. 415-434, 1963.

MOTOHASHI, Hideto et al. **Credit Derivatives: A Primer on Credit Risk, Modeling, and Instruments.** Pearson Education, 2006.

NIELS, Landwehr; HALL, Mark; FRANK, Eibe.**Logistic model trees.** 2005.

NKUSU, Mwanza. **Nonperforming loans and macrofinancial vulnerabilities in advanced economies.** IMF Working Papers, p. 1-27, 2011.

ØHRN, Aleksander. **Discernibility and Rough Sets in medicine: tools and applications.**2000. Disponível em: <https://wiki.eecs.yorku.ca/course_archive/2011-12/F/4403/_media/ohrn_thesis.pdf > Acesso em: 3 jun.2014.

ØHRN, Aleksander. **Rosetta technical reference manual.** Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, p. 1-66, 2000. Disponível em <https://wiki.eecs.yorku.ca/course_archive/2010-11/W/4403/_media/manual.pdf> Acesso em: 3 mar.2014.

PAUL, Ron. **O Fim do FED.**Tradução de Bruno Garschagen e Mônica Magalhães. É Realizações, Sao Paulo, 2011.

PAWLAK, Zdzislaw. **Rough Sets, rough relations and rough functions**.Fundamenta informaticae, v. 27, n. 2, p. 103-108, 1996.

PAWLAK, Zdzisław. **Rough Sets.** International Journal of Computer & Information Sciences, v. 11, n. 5, p. 341-356, 1982.

PAWLAK, Zdzisław**. Rough Sets: present state and the future**. Foundations of Computing and Decision Sciences, v. 18, n. 3-4, p. 157-166, 1993.

PAWLAK, Zdzisław. **Rough Sets: Theoretical aspects of reasoning about data.**Springer, 1991.

POLKOWSKI, Lech. **Rough Sets: Mathematical foundations.** Springer Science & Business, 2013.

PRIDDY, Kevin L.; KELLER, Paul E. **Artificial neural networks: an introduction**. SPIE Press, 2005.

RAJAN, Rajiv; DHAL, Sarat C. **Non-performing loans and terms of credit of public sector banks in India: An empirical assessment.** Occasional Papers, v. 24, n. 3, p. 81-121, 2003. Disponível em: <http://rbidocs.rbi.org.in/rdocs/Publications/PDFs/60613.pdf>Acesso em: 04 jun. 2014.

RITSCHARD, Gilbert. **CHAID and earlier supervised tree methods.** 2010. Disponível em <http://www.unige.ch/ses/metri/cahiers/2010_02.pdf>Acesso em: 22 set. 2014.

SABA, Irum; KOUSER, Rehana; AZEEM, Muhammad. **Determinants of Non Performing Loans: Case of US Banking Sector.** The Romanian Economic Journal, Year XV, v. 44, 2012.Disponível em <http://www.rejournal.eu/sites/rejournal.versatech.ro/files/issues/2012-06-02/556/15-determinantsofnon-performingloanscaseofusbankingsector.pdf >Acesso em: 22 set. 2014.

SASSI, Renato José. **An Hybrid Architecture For Clusters Analysis: Rough Sets Theory And Self-Organizing Map Artificial Neural Network.** Pesquisa Operacional. 2012. Disponível em: < http://www.scielo.br/pdf/pope/2012nahead/aop0512.pdf > Acesso em:24 mar.2013.

SERASA EXPERIAN. **Bate recorde o número de inadimplentes, revela levantamento inédito da Serasa Experian.** Disponível em: < http://noticias.serasaexperian.com.br/bate-recorde-o-numero-de-inadimplentes-revela-levantamento-inedito-da-serasa-experian/> Acesso em: 4nov.2012.

SHARPE, William F. **A simplified model for portfolio analysis.** Management science, v. 9, n. 2, p. 277-293, 1963.Disponível em: < http://www.jstor.org/stable/2627407> Acesso em: 2 mar.2013.

SCHÖNBUCHER, Philipp J. **Credit derivatives pricing models: models, pricing and implementation.** John Wiley & Sons, 2003.

ŠKARICA, Bruna.**Determinants of non-performing loans in Central and Eastern European countries.** Financial Theory and Practice, v. 38, n. 1, p. 37-59, 2014. Disponível em: <http://hrcak.srce.hr/119738>Acesso em: 04 jun. 2014.

STEFANOVIC, Pavel; KURASOVA, Olga.**Visual analysis of self-organizing maps.** Nonlinear Analysis, v. 16, n. 4, p. 488-504, 2011.

STROIEKEA, Renato Eduardo; FOGLIATTOB, Flavio Sanson; ANZANELLOC, Michel Jose.**Análise de conglomerados em curvas de aprendizado para formação de agrupamentos homogêneos de trabalhadores.** Revista Produção. São Paulo. Aguardando publicação, 2011. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-65132012005000084&lng=pt&nrm=iso > Acesso em: 24 mar.2013.

SUMNER, Marc; FRANK, Eibe; HALL, Mark.**Speeding up logistic model tree induction.**Em: Knowledge Discovery in Databases: PKDD 2005. Springer Berlin Heidelberg, 2005. p. 675-683.

TAVAKOLI, Janet M. **Credit derivatives and synthetic structures: A guide to instruments and applications.** John Wiley & Sons, 2001.

TOLEDO, Renato Proença Prudente de. **Mercado brasileiro de non-performing loans (NPL): uma abordagem teórica e prática na precificação de ativos.** São Paulo: FGV, 2013. 54 p. Dissertação de Mestrado -Escola de Economia de São Paulo, Fundação Getulio Vargas, São Paulo. 2013.

TSAI, Chih-Fong; LU, Yu-Hsin; YEN, David C. **Determinants of intangible assets value: The data mining approach.** Knowledge-Based Systems, v. 31, p. 67-77, 2012.

UCHÔA, J. Q. **Representação e Indução de Conhecimento usando Teoria de Conjuntos Aproximados.** Dissertação de Mestrado. Universidade Federal de São Carlos, 1998.

VAZQUEZ, Francisco; TABAK, Benjamin M.; SOUTO, Marcos. **A macro stress test model of credit risk for the Brazilian banking sector.** Journal of Financial Stability, v. 8, n. 2, p. 69-83, 2012. Disponível em: < http://www.bcb.gov.br/pec/wps/ingl/wps226.pdf>Acesso em: 04 jun. 2014.

VOGIAZAS, S.; NIKOLAIDOU, Eftychia.**Credit risk determinants in the Bulgarian banking system and the Greek twin crises.**Em: MIBES International conference. 2011. Disponível em: <http://mibes.teilar.gr/proceedings/2011/oral/14.pdf >Acesso em: 04 jun. 2014.

WARD JR, Joe H. **Hierarchical grouping to optimize an objective function**.Journal of the American statistical association, v. 58, n. 301, p. 236-244, 1963.

WEIßBACH, Rafael; UND WILKAU, Carsten von Lieres.**Capital for Non-Performing Loans.** .2008.Disponível em: < http://www.kaahlsfiles.com/thesis/thesis%20papers/2%20Medium/SSRN-id1098998.pdf >Acesso em: 04 jun. 2014.

WERBOS, P. J. **Beyond regression: new tools for prediction and analysis in the behavioral science.** Cambridge: Harvard, 1974. Tese (Doutorado). Harvard University, Cambridge, 1974.

WILLIAMS, Graham. **Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery.** Use R! Springer.1ª Edição. 2011.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. **Data Mining: Practical Machine Learning Tools and Techniques.** 3ª Edição.The Morgan Kaufmann Series in Data Management Systems. 2011. Morgan Kaufmann.

ZAIB, Amir; FARID, Faiza; KHAN, Muhammad Kamran.**Macroeconomic and Bank – specific Determinants of Non – performing Loans in the Banking Sector in Pakistan.** Business and Management, v. 6, n. 2, 2014. Disponível em: <http://ijibm.elitehall.com/IJIBM_Vol6No2_May2014.pdf>Acesso em: 04 jun. 2014.

ZURADA, Jozef; ZURADA, Martin.**How Secure Are "Good Loans": Validating Loan-Granting Decisions And Predicting Default Rates On Consumer Loans.**Review of Business Information Systems (RBIS), v. 6, n. 3, p. 65-84, 2011.

ZURADA, Jozef. **Could Decision Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decisions?**. In: System Sciences (HICSS), 2010 43rd Hawaii International Conference on. IEEE, 2010. p. 1-9.

APÊNDICE A – Atributos e as respectivas descrições da base de dados utilizada para os experimentos com a Rede SOM e a extração de regras utilizando Rough Sets.

Campo	**Descrição**	**Tipo**
Chave_Divida	Chave Divida na base de dados. Identificador da dívida.	Inteiro
Saldo_Inicial	Saldo Inicial da dívida no momento da celebração do contrato	Numérico (16,2)
CustomerID	Código do cliente.	Inteiro
DebtStatusCode1	Código Status Divida informa qual o status da dívida	Inteiro
CurrentBalance	Saldo Atual do cliente, até o momento do 'corte da base'	Numérico (16,2)
FirstDefaultDate	Data Primeiro Atraso	Data
DataVersion	Campo timestamp que indica o exato momento em que os dados foram extraídos	TimeStamp
CurrentAssignmentID	Chave que aponta a designação da unidade de negócio que trabalhou a dívida	Inteiro
ProductID1	Código que identifica o produto	Inteiro
CampaignID	Código que identifica a campanha	Inteiro
PortfolioID	Código do Portfólio na qual a dívida faz parte	Inteiro
Category	Categoria da dívida	Inteiro
OpenDate	Data da celebração do contrato	Data
FirstDefaultBalance	Saldo da dívida na data do Atraso	Data
OpenBalance	Saldo na Abertura do contrato	Numérico (16,2)
DebtNumber	Número do Contrato	Inteiro
DebtPrincipal	Saldo Principal	Numérico (16,2)
RecallDate	Se a dívida em algum momento foi devolvida, informa a data	Data
SettlementDate	Data Quitação da dívida	Data
DebtLastModificationDate	Data da Ultima Modificação de status da dívida	Data
AgeRangeID1	Código que apresenta a idade da dívida	Inteiro
InitialBalanceRangeID	Código que apresenta o saldo inicial da dívida	Inteiro
CustomerStatusCode1	Código do Status Cliente	Inteiro
CustomerLastWKTrackingID	Código do Ultimo Acionamento Cliente no sistema de cobrança	Inteiro
CustomerCreationDate	Data de Criação do Cliente no sistema de cobrança	Data
CustomerLastModificationDate	Data Ultima Modificação em algum atributo do Cliente	Data

Campo	Descrição	Tipo
CustomerDataVersion	Campo timestamp que indica o exato momento em que o cliente foi criado	TimeStamp
BinderID	Código que faz a ligação do cliente com a dívida	Inteiro
BindingStatusCode1	Código que informa o Status da ligação da dívida com o cliente	Inteiro
BindingLastWKTrackingID	Ultimo Acionamento da dívida	Inteiro
BindingCreationDate	Data de Criação do relacionamento dívida e cliente	Data
BindingLastModificationDate	Data Ultima Modificação da ligação cliente e dívida	Data
BindingDataVersion	Campo timestamp que indica o exato momento em que houve o relacionamento cliente e dívida	TimeStamp
CustodyStatusCode	Código Status da Custodia da dívida	Inteiro
ContactChannelCombinationID1	ID do Canal Combinação do Contato	Inteiro
LocalizationChannelCombinationID1	ID do Canal de Comunicação	Inteiro
RestrictionProviderCombinationID1	Código de Restrição	Inteiro
PreviousDataVersion	Marcação do tipo timestamp anterior relativo à dívida	TimeStamp
PrincipalBalance	Saldo Principal	Numérico (16,2)
CustomerName	Nome do Cliente	Texto
CustomerTypeCode	Código do Tipo de Cliente	Inteiro
PrimaryIdentityID	Código de Identidade de Cliente	Inteiro
PrimaryAddressID	Código de Endereço Primário	Inteiro
PrimaryPhoneID	Código do TelefonePrimário	Inteiro
PrimaryEmailID	Código do E-mail Primário	Inteiro
BestAddressID	Código do Melhor Endereço	Inteiro
BestPhoneID	Código do Melhor Telefone	Inteiro
BestEMailID	Código do Melhor E-mail	Inteiro
HomePage	Pagina Web do cliente	Texto
ContactCreationDate	Data da Criação do Contato	Data
ContactLastModificationDate	Data da Ultima Modificação Cliente1	Data
ContactDataVersion	Campo timestamp que informa quando o cliente foi criado	TimeStamp
State	Estado do cliente	Texto
FirstDefaultDateID	Código da Data do Primeiro Atraso	Data
SubPortfolioID	Código do Sub-Portfolio	Inteiro
DebtTotalBalance	Saldo Total Divida	Numérico (16,2)
LocalizationRoleCombinationID1	Código do Papel de Combinação de Localização	Inteiro
PortfolioKey	Identificador da Chave do Portfolio	Inteiro
CurrentAgentAssignmentID	Código que informa pra qual negociador a dívida esta alocada	Inteiro
DebtStatusCode	Código do Status da Divida	Inteiro
DebtStatusDesc	Descrição do Status da Divida	Texto

Campo	Descrição	Tipo
ProductID	Código do Produto	Inteiro
ProductCode	Código do Produto2	Inteiro
ProductName	Nome do Produto	Texto
ProductTypeID	Código do Tipo de Produto	Inteiro
ProductTypeAlias	Alias do Tipo de Produto	Texto
ProductTypeName	Nome do Tipo de Produto	Inteiro
AgeRangeID	Range relativo à Idade da Divida	Texto
AgeRangeStart	RangeInicial da dívida	Texto
AgeRangeEnd	RangeFinal da dívida	Texto
AgeRangeDesc	Descrição do Range da Divida	Texto
CustomerStatusCode	Código do Status do Cliente	Inteiro
CustomerStatusDesc	Descrição do Código do Status do Cliente	Texto
BindingStatusCode	Status do relacionamento cliente e dívida	Inteiro
BindingStatusDesc	Descrição do Status do relacionamento cliente e dívida	Texto
ContactChannelCombinationID	Código da Combinação do Canal de Contato	Inteiro
Penetrated	Penetração do canal de Contato	Inteiro
ContactChannelConcatenation	Concatenação do Canal de Contato	Inteiro
CombinationNumber1	Número da Combinacao1	Inteiro
CallCenter	Informa se o cliente foi contatado via call-center	Inteiro
Letter	Informa se o cliente foi contatado via Carta	Inteiro
SMS	Informa se o cliente foi contatado via Mensagem Texto	Inteiro
Voicer	Informa se o cliente foi contatado via Mensagem Voz	Inteiro
LocalizationChannelCombinationID	ID do Código de Concatenação do Canal de Localização	Inteiro
Localized1	Localizado	Inteiro
LocalizationChannelConcatenation	Concatenação do Canal de Localização	Inteiro
CombinationNumber2	Número da Combinacao2	Inteiro
Phone	Informa se o cliente tem Telefone	Inteiro
Address	Informa se o cliente tem Endereço	Inteiro
Email	Informa se o cliente tem E-mail	Inteiro
RestrictionProviderCombinationID	Código da Concatenação da Restrição de crédito	Inteiro
Restricted	Informa se o cliente está Restrito	Inteiro
RestrictionProviderConcatenation	Concatenação da Restrição	Inteiro
CombinationNumber3	Combinação do canal de localização do cliente	Inteiro
ACSP - SCPC	Informa se o cliente está com restrição no Bureau de cobrança 1	Inteiro

Campo	Descrição	Tipo
ACSP - UseCheque	Informa se o cliente está com restrição no Bureau de cobrança 2	Inteiro
ACSP - SCPC Empresas	Informa se o cliente está com restrição no Bureau de cobrança 3	Inteiro
SCPC/SERASA (Legado)	Informa se o cliente está com restrição no Bureau de cobrança 4	Inteiro
Restrição 021 (Legado)	Informa se o cliente está com restrição no Bureau de cobrança 5	Inteiro
SERASA - PEFIN	Informa se o cliente está com restrição no Bureau de cobrança 6	Inteiro
LocalizationRoleCombinationID	Código de Localização Concatenado	Inteiro
Localized	Informa se o cliente foi localizado	Inteiro
LocalizationRoleConcatenation	Localização Concatenada do Papel Divida	Inteiro
CombinationNumber	Numero Combinação de acordo com o Papelna Dívida	Inteiro
Paid	Informa se o crédito foi pago	Texto

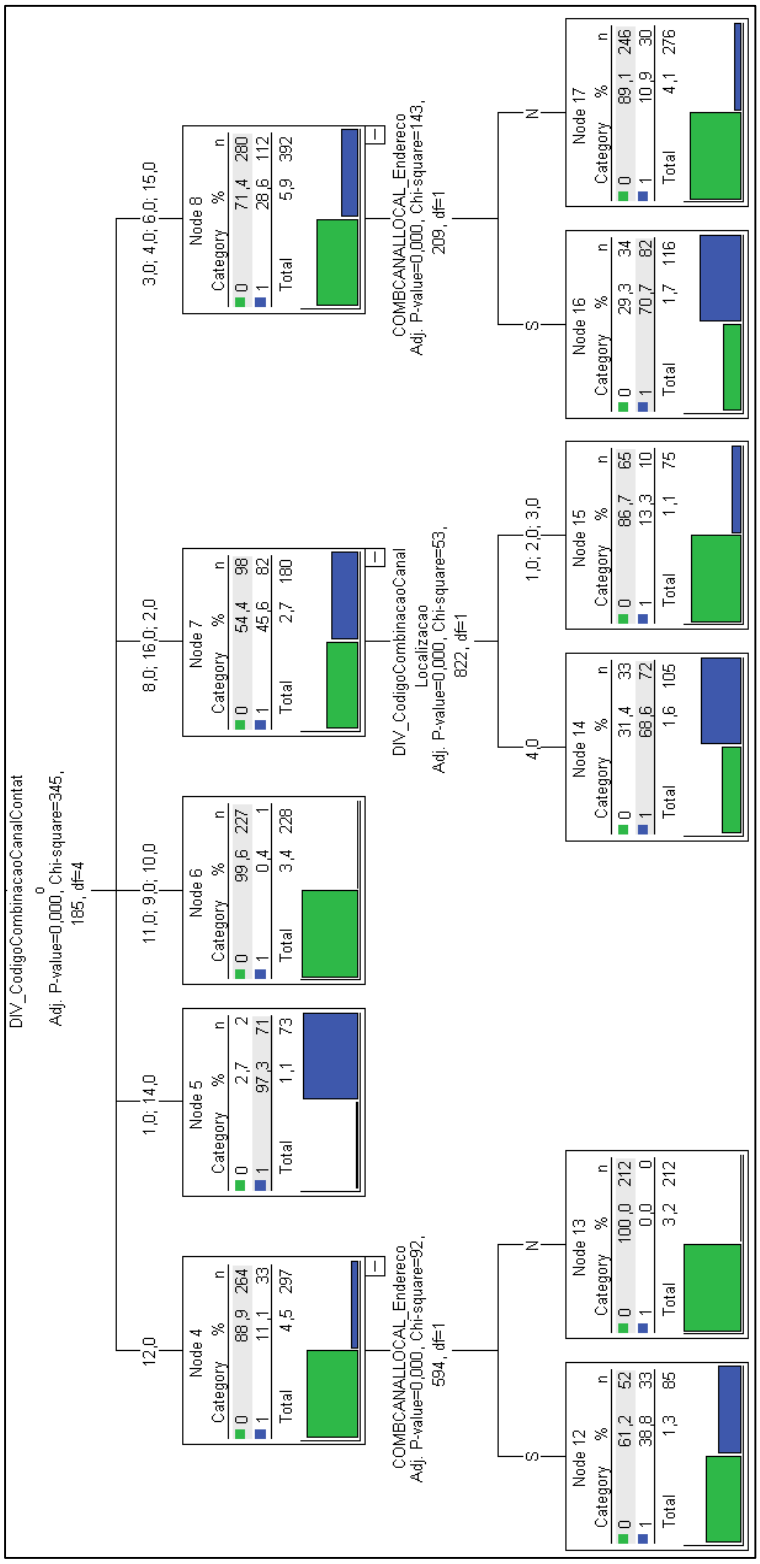
APÊNDICE B – Atributos e as respectivas descrições da base de dados utilizada para os experimentos com Rough Sets e as Árvores de Decisão.

Meta-Atributos	**Nr**	**Campo**	**Descrição**	**Tipo**
	2	DIV_SaldoInicial	Saldo inicial do cliente no momento em que a dívida foi passada para perda bancária.	Decimal (16,2)
	3	DIV_DataPrimeiroAtraso	Data de primeiro atraso da dívida.	Data
	4	DIV_CodigoDataPrimeiroAtraso	Código de data de primeiro atraso da dívida.	Inteiro
	5	DIV_CodigoProduto	Código do produto adquirido.	Inteiro
	6	DIV_DataContrato	Data o qual foi assinado o contrato relativo ao crédito cedido.	Data
	8	DIV_CodigoIdadeRangeDivida	Código que indica o range da dívida.	Inteiro
	9	DIV_CodigoRangeSaldoInicial	Código que indica o range relativo ao saldo da divida no momento em que foi passada para perdas.	Inteiro
	10	DIV_CodigoStatusCliente	Código: indica se cliente está ativo ou não.	Inteiro
	11	DIV_CodigoCombinacaoCanalContato	Código que indica o tipo de combinação no qual o cliente foi contatado.	Inteiro
	12	DIV_CodigoCombinacaoCanalLocalizacao	Código que indica as formas nas quais o cliente foi localizado.	Inteiro
	14	DIV_UF	Unidade da federação no qual a dívida foi originada.	Nominal
	15	DIV_CodigoCanalLocalizacao	Código que indica as formas que foram utilizadas para localizar o proprietário da divida.	Inteiro
	16	DIV_CodigoCanalLocalizacaoPapel	Código que indica se foi localizado o proprietário original da dívida ou o avalista.	Inteiro
	17	IDADERANGE_DescricaoRangeIdadeDivida	Descrição do período de vencimento da dívida a contar data de primeiro atraso.	Nominal
	18	SALDORANGE_DescricaoRangeSaldo	Descrição do range relativo ao saldo inicial da dívida no momento em que a mesma foi para perda.	Nominal
Combinação de	19	COMBCANALCONT_ContatoPenetrado	Flag que indica se o contato foi ao menos uma vez contatado.	Nominal

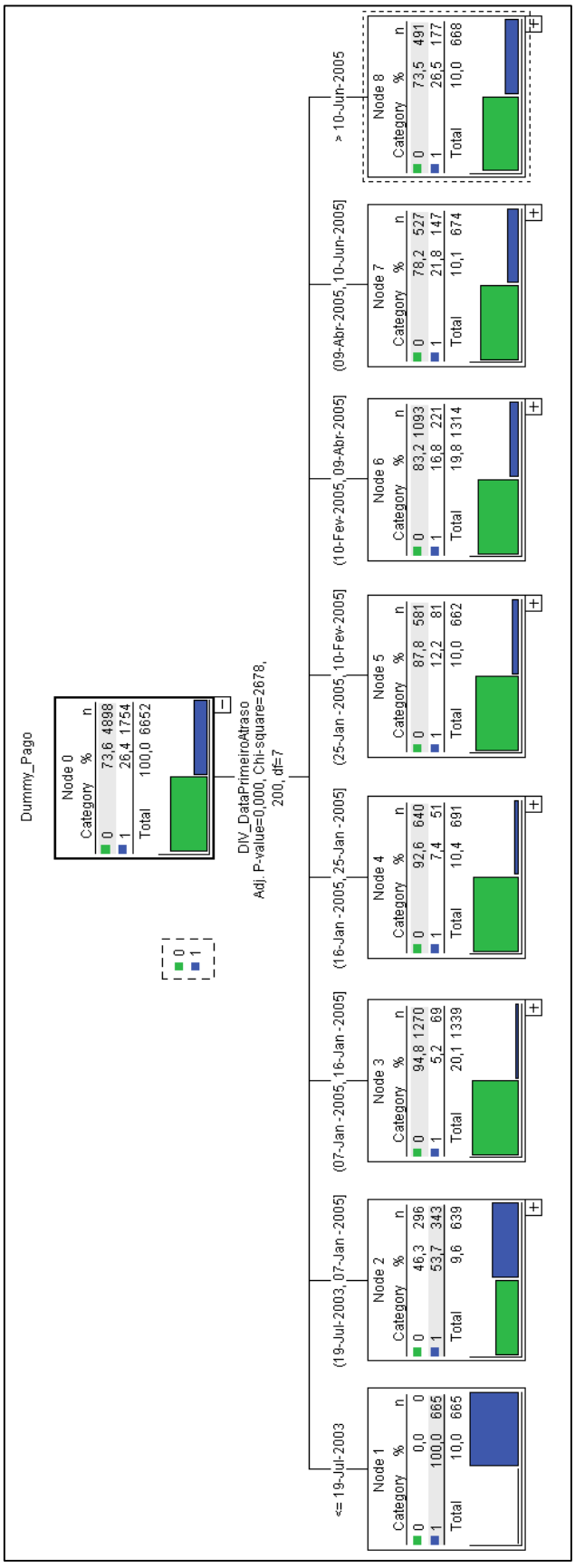
Meta-Atributos	Nr	Campo	Descrição	Tipo
Canal de Contato	20	COMBCANALCONT_ConcatenacaoCanalContato	Descrição dos canais utilizados para contactar cliente.	Nominal
	21	COMBCANALCONT_NumeroCombinacao	Número da combinação do canal de contato.	Inteiro
	22	COMBCANALCONT_CallCenter	Flag que indica se o cliente foi contatado via Call Center.	Nominal
	23	COMBCANALCONT_Carta	Flag que indica se o cliente foi contatado via Carta.	Nominal
	24	COMBCANALCONT_SMS	Flag que indica se o cliente foi contatado via SMS.	Nominal
	25	COMBCANALCONT_MensagemVoz	Flag que indica se o cliente foi contatado via Mensagem de Voz .	Nominal
Combinação de Forma de Restrição	26	COMBBUREAUREST_RestricaoCredito	Flag que indica se o CPF do cliente foi enviado para órgãos de restrição de crédito (Bureau de Crédito).	Nominal
	27	COMBBUREAUREST_ConcatenacaoBureauRestricao	Concatenação dos órgãos de restrição ao crédito os quais o CPF do cliente foi enviado.	Nominal
	28	COMBBUREAUREST_NumeroCombinacao	Código que indica a combinação das formas de restrição as quais o cliente foi submetido.	Inteiro
	29	COMBBUREAUREST_BureauCredito_1	Flag que indica se o cliente teve o CPF enviado para o Bureau de Restrição de Crédito número 1.	Nominal
	30	COMBBUREAUREST_BureauCredito_2	Flag que indica se o cliente teve o CPF enviado para o Bureau de Restrição de Crédito número 2.	Nominal
	31	COMBBUREAUREST_BureauCredito_3	Flag que indica se o cliente teve o CPF enviado para o Bureau de Restrição de Crédito número 3.	Nominal
	32	COMBBUREAUREST_BureauCredito_4	Flag que indica se o cliente teve o CPF enviado para o Bureau de Restrição de Crédito número 4.	Nominal
	33	COMBBUREAUREST_BureauCredito_5	Flag que indica se o cliente teve o CPF enviado para o Bureau de Restrição de Crédito número 5.	Nominal
	34	COMBBUREAUREST_BureauCredito_6	Flag que indica se o cliente teve o CPF enviado para o Bureau de Restrição de Crédito número 6.	Nominal
Combinação de Canali de	35	COMBCANALLOCAL_Localizado	Flag que indica se o cliente teve alguma combinação na sua localização.	Nominal

Meta-Atributos	**Nr**	**Campo**	**Descrição**	**Tipo**
Localização	36	COMBCANALLOCAL_ConcatenacaoCanalLocalizacao	Descrição dos canais nos quais o cliente foi localizado.	Nominal
	37	COMBCANALLOCAL_NumeroCombinacao	Código que indica a combinação dos canais de localização do cliente.	Inteiro
	38	COMBCANALLOCAL_Telefone	Flag que indica se o cliente foi contatado via telefone.	Nominal
	39	COMBCANALLOCAL_Endereco	Flag que indica se o cliente foi contatado via endereço.	Nominal
	40	COMBCANALLOCAL_Email	Flag que indica se o cliente foi contatado via e-mail.	Nominal
Produto	41	PROD_CodigoProduto	Código que indica o produto adquirido pelo cliente no momento da aquisição do crédito.	Inteiro
	42	PROD_NomeProduto	Nome do produto adquirido.	Nominal
	43	PROD_CodigoTipoProduto	Código que indica o tipo de produto adquirido.	Inteiro
	44	PROD_CodinomeProduto	Modalidade de crédito na qual foi submetido o cliente.	Inteiro
Combinação de Papel de Localização	45	COMBLOCALPAPEL_NumeroCombinacao	Código que indica se foi localizado o devedor principal ou o avalista (se houver).	Inteiro
	46	COMBLOCALPAPEL_CodigoCombinacaoPapelLocalizado	Código que indica combinação de localização do cliente ou do avalista.	Inteiro
	47	COMBLOCALPAPEL_ConcatenacaoPapelLocalizado	Descrição que indica se foi localizado o devedor principal ou avalista.	Nominal
	48	COMBLOCALPAPEL_Localizado	Flag que indica se o avalista ou o devedor principal foi localizado.	Nominal
Variáveis de Decisão	49	Dummy_Pago	Variável binária que indica se a dívida foi paga ou não.	Inteiro

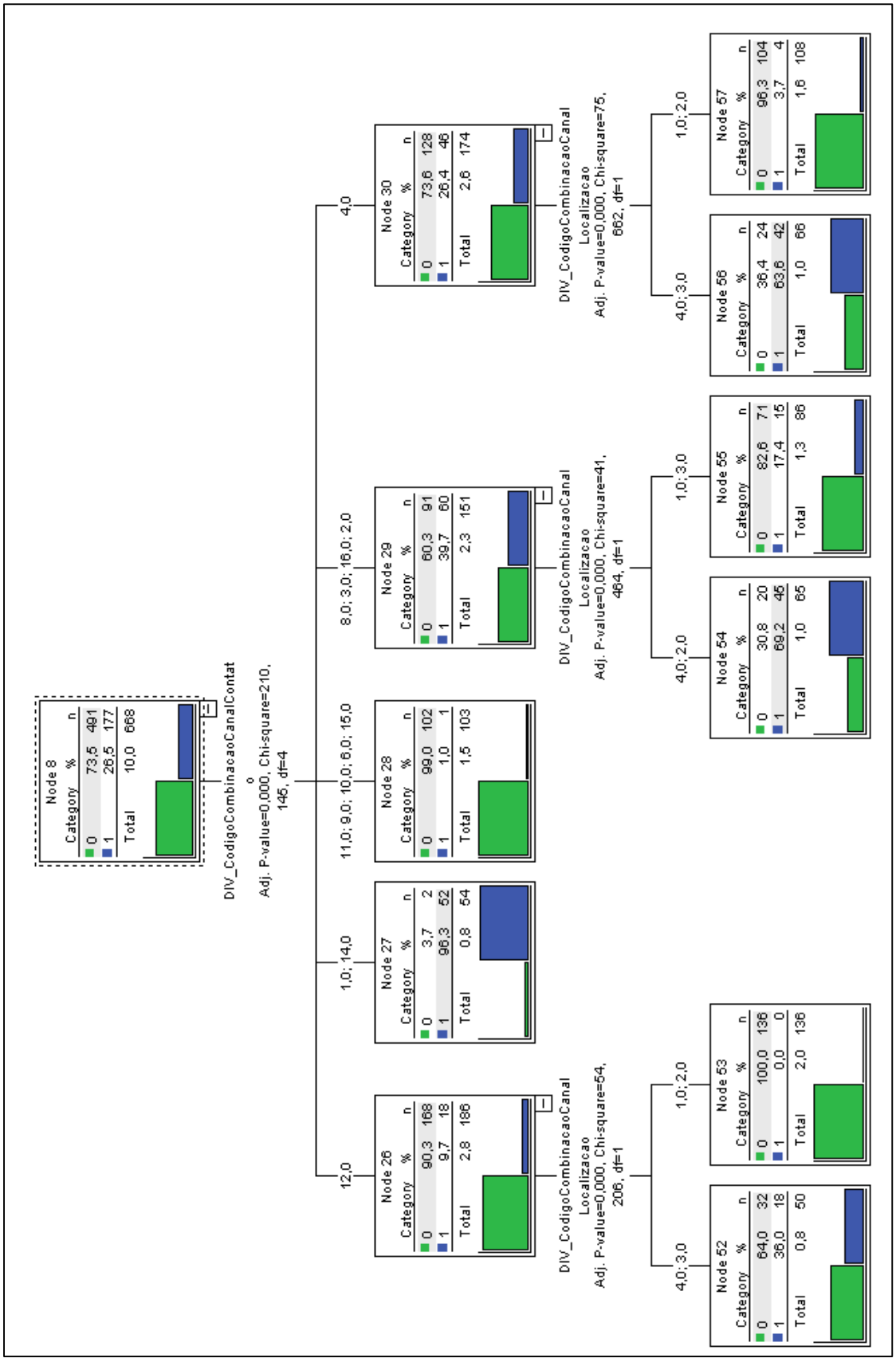
APÊNDICE C – Expansão completa do Modelo 1 - Nó 1.



APÊNDICE D – Expansão completa do Modelo 3 - Nó 0.



APÊNDICE E – Expansão completa do Modelo 3 - Nó 8.



APÊNDICE F – Expansão completa do Modelo 2 - Nó 4.