



**UNIVERSIDADE ESTADUAL DO CEARÁ**  
**CENTRO DE ESTUDOS SOCIAIS APLICADOS**  
**PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO**

**SIDARTA SILVA GALAS**

**PREVISÃO DE RUPTURA DE CLIENTES DE PLANOS DE SAÚDE**

**FORTALEZA - CEARÁ**

**2016**

SIDARTA SILVA GALAS

PREVISÃO DE RUPTURA DE CLIENTES DE PLANOS DE SAÚDE

Dissertação apresentada ao Programa de Pós-Graduação em Administração, do Centro de Estudos Sociais Aplicados da Universidade Estadual do Ceará, como requisito parcial para obtenção do grau de Mestre em Administração.

Área de concentração: Gestão, Organizações e Ambientes.

Orientador: Prof. Dr. Jerffeson Teixeira de Souza

Coorientador: Prof. Dr. Marcio de Oliveira Mota

FORTALEZA - CEARÁ

2016

Dados Internacionais de Catalogação na Publicação

Universidade Estadual do Ceará

Sistema de Bibliotecas

Galas, Sidarta Silva.

Previsão de ruptura de clientes de planos de saúde [recurso eletrônico] / Sidarta Silva Galas. - 2016.

1 CD-ROM: il.; 4 ¾ pol.

CD-ROM contendo o arquivo no formato PDF do trabalho acadêmico com 116 folhas, acondicionado em caixa de DVD Slim (19 x 14 cm x 7 mm).

Dissertação (mestrado acadêmico) - Universidade Estadual do Ceará, Centro de Estudos Sociais Aplicados, Mestrado Acadêmico em Administração, Fortaleza, 2016.

Área de concentração: Gestão, Organizações e Ambientes.

Orientação: Prof. Dr. Jerffeson Teixeira de Souza.

Coorientação: Prof. Dr. Marcio de Oliveira Mota.

1. Mercado de planos de saúde. 2. Marketing de relacionamento. 3. Retenção de clientes. 4. Churn. 5. CRISP-DM. I. Título.

SIDARTA SILVA GALAS

PREVISÃO DE RUPTURA DE CLIENTES DE PLANOS DE SAÚDE

Dissertação apresentada ao Programa de Pós-Graduação em Administração, do Centro de Estudos Sociais Aplicados da Universidade Estadual do Ceará, como requisito parcial para obtenção do grau de Mestre em Administração.

Orientador: Prof. Dr. Jerffeson Teixeira de Souza  
Coorientador: Prof. Dr. Marcio de Oliveira Mota

Aprovada em: 10 de abril de 2016.

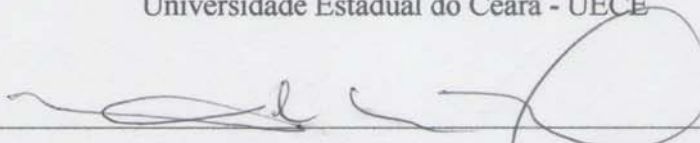
BANCA EXAMINADORA



---

Prof. Dr. Jerffeson Teixeira de Souza (Orientador)

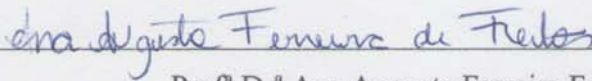
Universidade Estadual do Ceará - UECE



---

Prof. Dr. Marcio de Oliveira Mota (Coorientador)

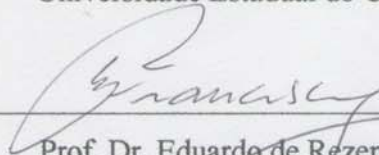
Universidade Estadual do Ceará - UECE



---

Profª Drª Ana Augusta Ferreira Freitas

Universidade Estadual do Ceará - UECE



---

Prof. Dr. Eduardo de Rezende Francisco

MPCC-ESPM e FGV-EAESP

Às três pessoas mais importantes da  
minha vida: Magna Coeli de Sousa e Silva  
Galas (mãe), Eduardo Santos Galas (pai)  
e Kandarpa Silva Galas (irmão).

## AGRADECIMENTOS

O mestrado é uma realização muito importante e desejada em minha vida. Realizar um trabalho como este implica empenho, concentração, rigor e o acompanhamento e estímulo das pessoas que estão mais próximas. Há muitas pessoas que tiveram importância durante esta difícil jornada.

Primeiro, agradeço a Deus, por ter me dado forças para superar os obstáculos desta difícil caminhada, e à Maria, que nunca parou de interceder por mim em todos os momentos.

Aos meus pais, Eduardo e Magna, pelo amor incondicional, pelo esforço, pela credibilidade, pela torcida, pelo incentivo, pelas orações que fizeram para que eu chegasse até aqui e por serem os melhores pais do mundo. Vocês foram meus féis incentivadores para crescer por meio dos estudos.

Ao meu irmão, Kandarpa, por me suportar, me incentivar, me ajudar e me apoiar em tudo que almejo.

Ao meu orientador, o Prof. Dr. Jerffeson Teixeira de Souza, pela sua experiência, sugestões, sua visão do todo, paciência nos momentos de indefinição, a sua clareza de pensamento, e, acima de tudo, pela sua enorme disposição em me ajudar.

Ao meu coorientador, o Prof. Dr. Marcio de Oliveira Mota, pelo estímulo, parceria, aprendizado e por todo o seu tempo dedicado a mim durante a minha estadia no mestrado. Agradeço, principalmente, pelas dicas preciosas e inteligentes nos momentos mais difíceis da construção deste projeto. Sempre se colocou à disposição e atendeu aos meus gritos de desespero. MUITÍSSIMO OBRIGADO!

À Prof<sup>a</sup>. Dr<sup>a</sup>. Ana Augusta Freitas, que me ajudou e colaborou para o desenvolvimento desta dissertação. Seus conhecimentos e contribuições foram de primeira importância na elevação e no aperfeiçoamento da construção dessa pesquisa.

Ao Prof. Dr. Eduardo de Rezende Francisco, por ter cedido seu tempo para conhecer este trabalho e pela valiosa contribuição na defesa.

À Livia A., pela sua alegria, atenção, vibração com as minhas conquistas e seu apoio em cada momento difícil que me ajudou a atravessar. Sem você, essa conquista não teria o mesmo sabor.

Aos amigos do PPGA/UECE, pela companhia, ajuda e atenção dispensadas a

mim. Vocês me ajudaram de várias formas na realização deste trabalho. Vocês fizeram esta etapa da minha vida inesquecível. Agradeço a todos que acreditaram em mim e contribuíram para que eu pudesse concretizar mais um objetivo em minha vida, o título de Mestre em Administração de Empresas.

## RESUMO

Embora exista uma vasta literatura referente à ruptura de cliente em empresas de telefonia, as empresas operadoras de planos de saúde têm dedicado pouca importância ao tema. O mercado brasileiro de planos de saúde movimentou, em 2015, R\$ 117,3 bilhões, até o terceiro trimestre, somadas as receitas das operadoras de planos de assistência médica e odontológicos, onde a maioria deste valor pertence a planos de assistência médica. Verificando a mudança do ano de 2014 para 2015, houve uma grande ruptura de beneficiários em planos de assistência médica, cerca de 766 mil, mostrando a importância de se identificar o risco de ruptura para a realização de ações de retenção de cliente. Diante da lacuna e da importância acima indicadas, esse estudo tem como objetivo classificar o risco de ruptura de clientes (beneficiários) de operadora de plano de saúde privado. O estudo compreendeu uma pesquisa exploratória-aplicada de natureza quantitativa. A partir do levantamento bibliográfico da pesquisa, foram apresentadas variáveis adequadas à composição de identificação à propensão de ruptura de clientes. O banco de dados trabalhado foi disponibilizado por uma operadora de planos de saúde, especificamente de assistência médica, com 21.074 beneficiários, com informações desde o ano 2000. A metodologia para analisar os dados se deu por meio do CRISP-DM, onde foi possível aplicar, na etapa de *data mining*, um comitê de classificadores (árvore de decisão, regressão logística e redes neurais) para prever o risco. Os resultados indicaram que a árvore de decisão foi ligeiramente melhor que redes neurais MLP. Em uma análise financeira dos resultados obtidos para esta operadora, verificou-se que os clientes previstos como possíveis *churn* eram prejudiciais ao negócio da empresa. Deste modo, o trabalho mostra que uma boa previsão de *churn* pode ajudar a empresa a definir se irá realizar ações de incentivo a ruptura de clientes não rentáveis e/ou se fará ações de retenção de clientes lucrativos.

**Palavras-chave:** Mercado de planos de saúde. *Marketing* de relacionamento. Retenção de clientes. *Churn*. CRISP-DM. *Data mining*. Comitê de classificadores.



## ABSTRACT

Although there is a vast literature related to churn in telephone companies, private health insurance companies have paid little attention to this issue. In 2015, the Brazilian's health insurance market drives R\$ 117.3 billion until the third quarter, counting all the revenues of health care and dental plans, where most of this amount belongs to health care plans. Checking the change from 2014 to 2015, there was a considerable churn of beneficiaries in health care plans, about 766 thousand people, pointing out the importance of identifying the risk of churn to do customer retention actions. Given the gap and the importance above, this study aims to classify the risk of customers churn (beneficiaries) in a private health plan firm. This study is an exploratory-applied quantitative research. The bibliographic research brought some appropriated variables and was presented to identify the churn propensity of customers. The database was provided by a health plans firm, specifically health care, in which had 21,074 beneficiaries with information since 2000. The methodology to analyze the data was by the CRISP-DM, where it was possible to apply, in data mining stage, a classifier committee (decision tree, logistic regression and neural networks) to predict the risk. The results indicated that the decision tree was slightly better than MLP neural networks. In a financial analysis of the produced results for this firm, we found that customers provided as possible churn were harmful to the business. Thus, this work shows that a good prediction of churn can help the company to set up the actions to stimulate the churn of unprofitable customers and/or make retention actions on profitable customer.

**Keywords:** Health Plans Market. Relationship Marketing. Customer Retention. Churn. CRISP-DM. Data Mining. Classifiers Committee.

## LISTA DE ILUSTRAÇÕES

Figura 1 - Exemplo de um classificador utilizando árvore de decisão.....	45
Figura 2 - Neurônio real e modelo neurônio artificial.....	51
Figura 3 - Simples rede neural artificial.....	52
Figura 4 - Representação das fases do modelo de referência do CRISP-DM.....	79
Figura 5 - Apresentação do WEKA.....	84
Figura 6 - Modelo do banco de dados.....	92
Figura 7 - Custos médios (despesa e receita) por beneficiário da análise de risco.....	101

## LISTA DE QUADROS

Quadro 1 - Visão geral da literatura de modelagem para predição de churn .....	31
Quadro 2 - Quantidade de publicações por área de estudo.....	39
Quadro 3 - Variáveis demográficas propostas para previsão de ruptura em um banco de dados de plano de saúde .....	66
Quadro 4 - Variáveis relacionadas a gastos propostas para previsão de ruptura em um banco de dados de plano de saúde .....	67
Quadro 5 - Variáveis relacionadas a reclamações propostas para previsão de ruptura em um banco de dados de plano de saúde .....	68
Quadro 6 - Variáveis de contagem propostas para previsão de ruptura em um banco de dados de plano de saúde .....	70
Quadro 7 - Variáveis relacionadas ao produto oferecido propostas para previsão de ruptura em um banco de dados de plano de saúde .....	72
Quadro 8 - Variáveis relacionadas ao serviço propostas para previsão de ruptura em um banco de dados de plano de saúde .....	73
Quadro 9 - Variáveis relacionadas a adicionais propostas para previsão de ruptura em um banco de dados de plano de saúde .....	74
Quadro 10 - Variáveis relacionadas a ofertas feitas propostas para previsão de ruptura em um banco de dados de plano de saúde .....	75
Quadro 11 - Descrição das variáveis da primeira base de dados (dados cadastrais).....	77
Quadro 12 - Descrição das variáveis da segunda base de dados (dados de uso e custo)	78
Quadro 13 - Descrição das variáveis da terceira base de dados (dados de mensalidades) .....	78
Quadro 14 - Variáveis excluídas do estudo .....	90
Quadro 15 - Descrição das variáveis da base final de dados (dados dos beneficiários do plano de saúde) .....	93

## LISTA DE TABELAS

Tabela 1 - Tabela de frequência Estado civil vs. Churn.....	88
Tabela 2 - Tabela de frequência Sexo vs. Churn.....	88
Tabela 3 - Tabela de frequência por região vs. Churn.....	89
Tabela 4 - Tabela de frequência de titulares e dependentes vs. Churn.....	89
Tabela 5 - Tabela de frequência do modo de pagamento vs. Churn .....	90
Tabela 6 - Peso das variáveis selecionadas para modelo final .....	96
Tabela 7 - Matriz de confusão da árvore de decisão (validação) .....	96
Tabela 8 - Matriz de confusão da regressão logística (validação).....	97
Tabela 9 - Matriz de confusão do RBF (validação) .....	97
Tabela 10 - Matriz de confusão do MLP (validação).....	98
Tabela 11 - Resultados dos modelos de previsão .....	99
Tabela 12 - Direções das variáveis .....	99

## SUMÁRIO

<b>1 INTRODUÇÃO</b> .....	<b>14</b>
<b>2 MARKETING</b> .....	<b>22</b>
<b>3 GERENCIAMENTO DE CHURN</b> .....	<b>27</b>
3.1 DEFINIÇÃO E CAUSAS DO <i>CHURN</i> .....	27
3.2 IMPLICAÇÕES SOBRE <i>CHURN</i> .....	29
3.3 LIMITAÇÕES DA UTILIZAÇÃO DE GERENCIAMENTO DE <i>CHURN</i> .....	39
3.4 DATA MINING .....	40
<b>3.4.1 Árvore de decisão</b> .....	<b>44</b>
<b>3.4.2 Regressão logística</b> .....	<b>47</b>
<b>3.4.3 Redes neurais</b> .....	<b>50</b>
3.4.3.1 <i>Multilayer Perceptron</i> (MLP) .....	53
3.4.3.2 <i>Radial Basis Function</i> (RBF) .....	54
<b>4 GERENCIAMENTO DE CHURN EM PLANOS DE SAÚDE</b> .....	<b>56</b>
4.1 O MERCADO DE PLANO DE SAÚDE .....	56
4.2 RETENÇÃO E IMPLICAÇÕES EM PLANOS DE SAÚDE .....	58
<b>5 PROPOSTA DE VARIÁVEIS PARA COMPOSIÇÃO DE BASE PARA PREVISÃO DE RUPTURA EM PLANO DE SAÚDE</b> .....	<b>64</b>
5.1 VARIÁVEIS DEMOGRÁFICAS .....	64
5.2 VARIÁVEIS DE GASTOS .....	65
5.3 VARIÁVEIS DE RECLAMAÇÕES .....	67
5.4 VARIÁVEIS DE CONTAGEM.....	68
5.5 VARIÁVEIS DO PRODUTO .....	71
5.6 VARIÁVEIS DO SERVIÇO .....	72
5.7 VARIÁVEIS DE ADICIONAIS .....	74
5.8 VARIÁVEIS DE OFERTAS .....	75
<b>6 PROCEDIMENTOS METODOLÓGICOS</b> .....	<b>76</b>
6.1 CARACTERIZAÇÃO DO ESTUDO .....	76
6.2 BANCO DE DADOS .....	76
6.3 A METODOLOGIA CRISP-DM.....	78
<b>6.3.1 Entendimento do Negócio (<i>Business Understanding</i>)</b> .....	<b>79</b>
<b>6.3.2 Entendimento dos Dados (<i>Data Understanding</i>)</b> .....	<b>80</b>
<b>6.3.3 Preparação dos dados (<i>Data Preparation</i>)</b> .....	<b>81</b>

<b>6.3.4 Modelagem (<i>Modelling</i>)</b> .....	<b>82</b>
<b>6.3.5 Avaliação (<i>Evaluation</i>)</b> .....	<b>82</b>
<b>6.3.6 Implementação (<i>Deployment</i>)</b> .....	<b>83</b>
<b>6.3.7 Ferramenta de <i>Data mining</i> utilizada</b> .....	<b>84</b>
<b>7 ESTUDO DE CASO</b> .....	<b>86</b>
7.1 ENTENDIMENTO DO NEGÓCIO ( <i>BUSINESS UNDERSTANDING</i> ).....	86
7.2 ENTENDIMENTO DOS DADOS ( <i>DATA UNDERSTANDING</i> ).....	87
7.3 PREPARAÇÃO DOS DADOS ( <i>DATA PREPARATION</i> ) .....	90
7.4 MODELAGEM ( <i>MODELLING</i> ).....	95
7.5 AVALIAÇÃO ( <i>EVALUATION</i> ).....	98
7.6 IMPLEMENTAÇÃO ( <i>DEPLOYMENT</i> ) .....	100
7.7 ANÁLISE FINANCEIRA DO NEGÓCIO .....	100
<b>CONSIDERAÇÕES FINAIS</b> .....	<b>103</b>
<b>REFERÊNCIAS</b> .....	<b>107</b>

## 1 INTRODUÇÃO

A ruptura de cliente foi estudada por autores (BOTELHO; TOSTES, 2010; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; KNOX; OEST, 2014; OWCZARCZUK, 2010; VAFEIADIS *et al.*, 2015) que mostraram que a ruptura de cliente dá prejuízo à empresa. Encontra-se na literatura várias técnicas estatísticas para previsão do possível clientes que vai romper com a empresa. Esse trabalho se propõe a abordar e apresentar uma análise sobre o tema ruptura de clientes, utilizando técnicas de classificação e previsão para identificar possíveis abandonos de clientes.

Com a crescente concorrência atual no mercado, empresas de inúmeros ramos buscam novas estratégias para tornar-se competitivas. Entretanto, o cliente é quem define se uma companhia se sobressairá frente às demais ou não. Assim, as organizações devem oferecer produtos/serviços que atendam às expectativas e às necessidades do cliente, uma vez que estes estão mais exigentes, reconhecem a atual concorrência e sabem a sua importância para a empresa.

Kakwani, Neri e Son (2010), além de afirmarem que o cliente tem se tornado mais exigente, relatam que o número de consumidores brasileiros e transações em grandes empresas tem aumentado, tendo como causas possíveis a concentração de poucas grandes companhias e o maior acesso ao consumo que a população brasileira tem tido na última década. Assim, buscando aproveitar esse momento de acessibilidade do indivíduo ao consumo, empresas buscam maximizar seu lucro diante das concorrentes. Para tal realização, as organizações podem vender para mais clientes, vender mais para clientes atuais, vender para clientes antigos, aumentar preços ou buscar a menor perda de consumidores possível. Contudo, os consumidores não aceitam aumentos constantes de preços, e vender é uma operação cara. Desta forma, a retenção de clientes, vulneráveis ou não à migração, se transforma em uma das mais importantes saídas para elevar o lucro da empresa, além de garantir resultados consistentes a longo prazo (GOMES, 2011).

Para alcançar a retenção, desenvolver projetos e planos que superem as expectativas dos clientes e compreender as suas necessidades é algo que assegura a escolha do cliente pela empresa, ao invés da concorrente. Essas ações são bastante importantes e de responsabilidade da área funcional de marketing, cuja finalidade é satisfazer, fidelizar e reter clientes (SHETH; SISODIA, 1995; SHETH; PARVATIYAR, 1996).

O marketing, durante as décadas de 1960 e 1970, era responsável pelo estudo da troca transacional (KOTLER, 1972; BAGOZZI, 1974) e da troca social (BARTELS, 1968). O estudo de Hunt (1983), influenciado por Howard e Sheth (1969), propôs uma expansão do foco do marketing de trocas (transacionais e sociais) para trocas relacionais, no qual há uma mudança de paradigma implícita no conceito. Em vista disso, a construção de relacionamento passa a ser o foco do marketing, e Hunt (1983) o define como sendo a ciência que procura explicar as trocas relacionais.

O marketing de relacionamento, primeiramente citado por Berry (1983), foi conceituado como todos os serviços organizacionais de uma empresa que buscam atrair, conquistar e manter o cliente. O autor declara que a captação de clientes é algo intermediário no processo do marketing, haja vista que o cerne do marketing é transformar clientes casuais em clientes fiéis, consolidando o mútuo relacionamento. Anos depois, Morgan e Hunt (1994, p. 22) apresentaram uma nova proposta de definição para o marketing de relacionamento, que “[...] se refere a todas as atividades de marketing direcionadas a estabelecer, desenvolver e manter trocas relacionais bem-sucedidas”. Seguindo a mesma ideia, Gordon (1998) define o *marketing* de relacionamento como um processo constante de identificação e criação de valores novos com clientes individuais e a partilha de seus benefícios durante toda a parceria existente.

Em vista disso, o *marketing* de relacionamento, juntamente à estratégia de retenção de clientes e fidelização, transforma-se em uma alternativa atraente para as organizações no combate ao abandono (*churn*) de clientes da empresa.

Observando mais atentamente o aspecto de investimentos para reter um cliente a uma determinada empresa, vertentes de vantagens e desvantagens surgem quanto à sua aplicação. Segundo Reinartz, Thomas e Kumar (2005), a mobilização na redução dos investimentos para reter um cliente seria negativa no tocante à lucratividade com clientes de longo prazo, ao invés de colocar baixos investimentos para obtenção de novos consumidores.

Sobre isso, comenta-se que em muitos setores o custo de aquisição de novos clientes pode ser cinco vezes superior ao seu custo de retenção (KURTZ; CLOW, 1998). Em outros casos, como os explicitados por Asbrand (1997) e Knowles (1997), a retenção pode custar quatro a oito vezes menos do que a aquisição de clientes. Logo, Reichheld (1996) e Reichheld, Markey Jr. e Hopton. (2000) reiteram que conquistar novos clientes é mais oneroso e custoso que mantê-los.

Consequentemente, uma retenção do consumidor objetivando evitar a ruptura



do relacionamento tornou-se uma questão importante para gestores (BUCKNIX; VAN DEN POEL, 2005). Tendo em mente o desafio, as empresas devem estar capacitadas com os métodos mais eficientes e eficazes para analisar o comportamento de seus clientes e prever a sua possível ruptura com a empresa no futuro.

Após observação e aplicação de pesquisas, Vavra (1994) reforça que por quanto mais tempo se mantém um cliente, mais lucrativo ele poderá ser para o negócio, e por quanto mais tempo o cliente comprar de uma mesma organização, tornar-se-á mais dependente de seus produtos ou serviços e estará menos suscetível a ofertas com preços mais baixos da concorrência. Os clientes podem abandonar uma organização ainda que altos investimentos em prospecção e retenção sejam realizados, o que requer diagnóstico e compreensão sempre de forma objetiva (BOTELHO; TOSTES, 2010).

Para isso, as atividades de relacionamento com o cliente estão sendo incorporadas ao negócio das empresas, pois são práticas de marketing cujo foco é assegurar seus clientes, adotando o grande desafio de reconhecê-los, mostrando-lhes o quanto a empresa os estima por terem lhe conferido a preferência (VAVRA; PRUDEN, 1995).

Além da melhor agilidade da empresa, a taxa de retenção de clientes é vista por Rust e Zahorik (1993) como o componente mais importante para que a empresa consolide sua participação no mercado, sendo direcionada pela satisfação do consumidor. Um mercado se torna seguro e consolidado por meio de suas boas e construídas relações.

Com o mercado saturado, as organizações percebem que precisam alinhar seus esforços na identificação de clientes que são predispostos ao abandono da empresa (HADDEN *et al.*, 2005). A taxa de abandono de clientes (*Customer Churn*) diz respeito à propensão dos clientes a parar de realizar negócios com uma determinada empresa por um período de tempo específico (QIAN; JIANG; TSUI, 2006). Assim, analisar o comportamento de seus clientes durante sua permanência na empresa pode contribuir para a redução dessa taxa, devido à identificação de fatos que expliquem a insatisfação do cliente (GALVÃO; GONZALEZ, 2011). Esses autores indicam que a taxa de abandono em níveis baixos pode refletir em uma melhoria do relacionamento entre empresa e cliente e, por conseguinte, aumentar a possibilidade de retenção.

Botelho e Tostes (2010) explicam que a perda de consumidor por atrito no relacionamento (e.g. autorização não permitida em uma venda por cartão de crédito) ou o abandono do cliente (e.g. troca de produto sem motivo explícito) pode acontecer a qualquer momento do ciclo de vida do cliente junto à empresa.

O gerenciamento do *churn* permite o uso de estratégias de relacionamento que previnam tal atrito ou abandono de consumidor, podendo, assim, manter os seus clientes mais lucrativos (BOTELHO; TOSTES, 2010; GALVÃO; GONZALEZ, 2011).

O progresso em sistemas de armazenamentos e programas de banco de dados possibilita que empresas reúnam grandes registros de transações de clientes e possam manipular tal massa de dado de maneira rápida, o que facilita a exploração dos dados para acompanhar as atividades de *churn* (QIAN; JIANG; TSUI, 2006). A massa de dados produzida por meio de atividades pelo CRM nas empresas e colhida de forma digital possibilita ações individualizadas aos consumidores, por meio da utilização de técnicas de mineração de dados (BOTELHO; TOSTES, 2010).

Um *Database Marketing* bem estruturado armazena o histórico de transações dos clientes de forma clara e objetiva, entre outras informações, de tal forma que o monitoramento seja contínuo e assegure a integridade das informações dos consumidores. Para Stone e Shaw (1987), *Database Marketing* é um banco de dados eletrônico de clientes em potencial e de todos os contatos comerciais ou de comunicação. A partir dele, o *marketing* pode expandir seu público-alvo, estimular a demanda desta audiência e estar próximo a ela. Segundo os autores, tal base de dados visa ao melhoramento de contatos futuros.

De acordo com Hughes (1998), *Database Marketing* é planejado com os dados que revelam o perfil dos clientes atuais e potenciais, guardando seu histórico de compra, com o propósito de agrupar informações individualizadas. Possuindo este o relacionamento personalizado, pode ser construído e gerenciado. Reforçando esta definição, Pedron (2011) declara que *Database Marketing* é um gerenciamento dinâmico de uma base de dados onde estas possuem informações relevantes e atualizadas de clientes atuais e de clientes potenciais.

Conforme Holtz (1994), o levantamento de informações para o *Database Marketing* vem diretamente dos clientes e, conseqüentemente, faz com que a empresa estreite o relacionamento com o mesmo. A produção desse banco de dados mostra o interesse que a empresa tem em seu cliente, buscando compreender o que ele deseja e pensa. Stone e Shaw (1987) declaram que a grande capacidade de armazenamento e processamento de dados dos atuais computadores são responsáveis pela possibilidade de existência do *Database Marketing*.

O processamento de dados é uma atividade que está associada ao *Database Marketing* por meio do *data mining*, ou mineração de dados, que é a utilização de

algoritmos para encontrar padrões implícitos dos dados, pretendendo resolver problemas como classificação de entidades, descrição, agrupamento e predição (LAROSE, 2005; MAIMON; ROKACH, 2010). Hongxia, Min e Jianxia (2009) ressaltam que esses algoritmos procuram, de forma automática, padrões previamente desconhecidos e possivelmente úteis para solucionar problemas.

Conforme Gomes (2011), a ferramenta de *data mining* é substancialmente importante nos modelos de negócio atuais, pois possibilita a transformação dos dados em informações objetivas e confiáveis para os gestores tomarem decisões. O autor indica que a classificação e a predição de indivíduos são os trabalhos mais usuais na área de apoio à tomada de decisão nas empresas.

O desenvolvimento de modelos que buscam prever o *churn* se torna mais difícil quando se trata de uma situação não contratual (e.g. supermercado), dado que os clientes podem mudar constantemente o seu comportamento de compra sem informar à empresa. Consequentemente, acompanhar e estimar o *churn* tem sido um transtorno de alta complexidade para as organizações (BUCKNIX; VAN DEN POEL, 2005). Já em um ambiente contratual (e.g. operadora de telefonia, plano de saúde), o cliente precisa informar qualquer mudança que deseja para a empresa, tornando um pouco mais fácil acompanhar e estimar o *churn*, mas não menos desafiadora (HUNG; YEN; WANG, 2006).

Operadoras de planos de saúde, por exemplo, buscam maneiras diferentes para estimar o *churn*, diante da grande concorrência do mercado e por ser um mercado regulamentado (MENDES, 2008). Este mercado, em 2015, possuía 1.340 operadoras com registro ativo no Brasil e 71.680.868 beneficiários (assistência médica com ou sem odontologia e exclusivamente odontológico). Dentre os planos de assistência médica, houve uma queda de 1,5% (cerca de 766 mil beneficiários) em relação ao ano anterior (BRASIL, 2016). Tal informação mostra o quanto é importante o estudo deste setor.

O avanço relevante de inovações tecnológicas na área médica e o aumento da demanda fez com que, segundo Alvez (2008), os custos com a prestação de serviços médicos aumentassem. O mercado de planos de saúde brasileiro teve uma receita de 115,26 bilhões de reais e uma despesa de 115,50 bilhões de reais até o terceiro trimestre de 2015, fechando o ano de 2015 com um prejuízo de 238 milhões (BRASIL, 2015). Esses dados corroboram a relevância de se estudar esse mercado à luz do gerenciamento de *churn*, ajudando a prever quando o cliente abandonará a operadora do plano de saúde.

Atualmente, as bases de dados disponíveis para as empresas possuem grandes

desafios para a realização de modelos generalizados de previsão de *churn* (QIAN; JIANG; TSUI, 2006), e isso não é diferente em operadoras de planos de saúde. A título de exemplo de problemas possíveis, podem-se citar perfis com informações históricas curtas, presença de tendência e a existência de correlações seriais ao longo do tempo. Ademais, perfis são repetidamente interrompidos por mudanças no negócio, tais como vários ajustes contábeis, inovações tecnológicas e substituição de produtos no varejo (QIAN; JIANG; TSUI, 2006).

Em um *Database Marketing*, milhares de clientes são modelados simultaneamente, a fim de que um modelo global e flexível capture as dinâmicas comerciais e não comerciais e categorize os consumidores de acordo com o seu comportamento. Desta maneira, campanhas de *marketing* podem ser mais eficazes, pois seria possível concentrar esforços em grupos específicos de usuários (QIAN; JIANG; TSUI, 2006). Na mesma direção, Last, Kandel e Bunke (2004) explicam que, por modelar todos os perfis simultaneamente, a classificação de clientes ajuda na identificação de grupos homogêneos e “toma emprestado à força” (*borrow strength*) de perfis do mesmo *cluster*, assegurando estimações precisas.

Metodologicamente, inúmeras técnicas para previsão de ruptura já foram aplicadas, dentre elas temos: a árvore de decisão (ABBASIMEHR; SETAK; TAROKH, 2014); *random forest* (BUREZ; VAN DEN POEL, 2008); regressão logística (COUSSEMENT; VAN DEN POEL, 2008a) e redes neurais (YU *et al.*, 2011).

Entretanto, diante da diversidade e da complexidade existentes, o presente estudo não pretende abranger todas as técnicas já utilizadas. Assim, a presente dissertação busca explorar e comparar as seguintes técnicas: árvore de decisão, regressão logística e redes neurais em análises de *churn*. Tais técnicas foram escolhidas por serem as mais utilizadas, conforme o levantamento da literatura para previsão de *churn*.

Deste modo, ressalta-se que a justificativa de se estudar o tema ruptura de clientes visa colaborar para o crescimento teórico e prático desta área que vem ganhando notoriedade na última década, especialmente na modelagem do risco de ruptura do cliente (GUSTAFSSON; JOHNSON; ROOS, 2005; KNOX; OEST, 2014; LAMBRECHT; SKIERA, 2006; LEMMENS; CROUX, 2006; LEWIS, 2004; NESLIN *et al.*, 2006; NITZAN; LIBAI, 2011; SCHWEIDEL; FADER; BRADLOW, 2008). Assim, a questão principal para a qual se busca resposta por meio deste estudo é: qual o risco de ruptura de clientes (beneficiários) de operadoras de planos de saúde privados?

A pergunta de pesquisa acima mencionada origina o objetivo principal do

estudo: classificar o risco de ruptura de clientes (beneficiários) de operadora de plano de saúde privado. Especificamente, pretende-se (1) examinar modelos de classificação de risco de ruptura de clientes; (2) apresentar variáveis adequadas à composição de identificação à propensão de ruptura de clientes; (3) identificar o principal modelo que se adequa às condições de previsão de classificação de ruptura de clientes.

Nesta introdução, foram expostas as lacunas teóricas e a delimitação da temática desta dissertação com os principais temas para possibilitar a pesquisa. Foram apresentados o problema de pesquisa, seus objetivos (geral e específicos) e a relevância do estudo.

O capítulo 2, apresentará, inicialmente, considerações gerais sobre o tema *marketing*, conceituando-o e mostrando suas perspectivas. Em seguida, uma revisão teórica dos principais conceitos sobre *marketing* de relacionamento, retenção de clientes e, especificamente, *churn* e gerenciamento de *churn*.

O capítulo 3 abordará o conceito, as causas, alguns estudos realizados e as limitações da utilização de gerenciamento de *churn* em uma empresa, apontando a necessidade do desenvolvimento de modelos capazes de provisionar o tempo de vida do cliente dentro da organização. Para tal desenvolvimento, serão apresentadas considerações sobre *Data mining* (regressão logística, árvore de decisão e redes neurais).

O gerenciamento de *churn* em planos de saúde é apresentado no capítulo 4, onde o mercado de planos de saúde é abordado. Implicações e estudos sobre retenção de clientes neste mercado são contemplados. Diretamente associado a este capítulo e buscando atingir um dos objetivos específicos da dissertação, são propostas, no Capítulo 5, variáveis para compor base para previsão de ruptura em planos de saúde.

Os procedimentos metodológicos são descritos, explicados e justificados no capítulo 6, no qual o tipo de pesquisa, a definição da população, a coleta de dados, o tratamento dos dados, os procedimentos de análise dos dados, construção do modelo proposto e sua verificação e validação serão explicados.

Com o título de “Estudo de Caso”, o capítulo 7 apresentará as análises relacionadas ao desenvolvimento e aplicação dos modelos propostos. Estatísticas descritivas da população estudada serão apresentadas. Por fim, buscaremos informar se os objetivos propostos neste estudo foram atingidos.

No último capítulo, serão tecidas as devidas conclusões, bem como as limitações da pesquisa, implicações gerenciais e recomendações para trabalhos futuros. Por fim, seguem as referências utilizadas.

## 2 MARKETING DE RELACIONAMENTO

No presente capítulo, será realizada uma revisão teórica dos principais conceitos sobre *marketing* de relacionamento, retenção de clientes e *churn*.

Com o surgimento das trocas relacionais citadas por Hunt (1983), Berry (1983) foi o primeiro a citar o termo *marketing* de relacionamento, conceituando-o como aqueles serviços de uma empresa, a nível organizacional, que buscam atrair, conquistar e manter o cliente. No processo de *marketing*, a conquista de clientes tem papel intermediário, dado que o foco principal é transformar o cliente casual em um cliente leal, através do estabelecimento de um mútuo relacionamento.

Para envolver todas as formas de trocas relacionais e com o foco no processo de *marketing* de relacionamento, Morgan e Hunt (1994, p. 22) determinaram que “[...] o *marketing* de relacionamento se refere a todas as atividades de *marketing* direcionadas a estabelecer, desenvolver e manter trocas relacionais bem-sucedidas”. Semelhantemente, Gordon (1998) apresenta o *marketing* de relacionamento como sendo um processo contínuo que identifica e cria valores em cada cliente, compartilhando os benefícios construídos em toda a permanência de tal relação.

O comportamento do consumidor, definido por Engel, Blackwell e Miniard (2000) como atividades comprometidas na obtenção, consumação e disposição de produtos e serviços, abrangendo os processos decisórios que as antecedem e sucedem, é algo fundamental para que uma relação entre empresa e cliente se estabeleça. A compreensão do comportamento do cliente se torna relevante para compreender as suas implicações na retenção do cliente. Assim, Kotler e Armstrong (1999) destacam que o comportamento do consumidor não é algo permanente, mas sofre mudanças ocasionadas por alguns fatores (e.g. sociais, culturais, pessoais e psicológicas).

Essas mudanças de comportamento, associadas às novas exigências dos consumidores e o avanço tecnológico, impactaram na administração do *marketing*, fazendo com que estratégias voltadas ao cliente passassem a ser utilizadas com o intuito de conseguir acompanhar esta evolução (DARÉ, 2007). A busca das organizações para oferecer os produtos que os clientes almejam se torna cada vez mais difícil, fazendo com que as empresas assumam tais estratégias para obter vantagem competitiva sustentável (HUNT; LAMBE; WITTMANN, 2002).

Entretanto, com o surgimento do *e-commerce*, manter os clientes deixou de ser uma tarefa fácil (JAHROMI *et al.*, 2010). Com o poder que a *internet*, com suas facilidades de acesso, proporcionou aos consumidores, pela maior liberdade para tomar

decisões, é gerado um agravamento da concorrência (PEPPARD, 2000). Diante disso, o *marketing* de relacionamento, combinado com estratégias de retenção e fidelização de clientes, apresenta-se como uma maneira interessante de combater a ruptura (*churn*) de clientes da empresa.

Logo, no tocante ao aspecto investimento, observa-se, em alguns casos, que para a reter clientes em uma empresa, menores investimentos se tornam menos prejudiciais à lucratividade (REINARTZ; THOMAS; KUMAR, 2005). Enquanto que para uns o custo de aquisição de novos clientes pode ser cinco vezes superior ao seu custo de retenção (KURTZ; CLOW, 1998), para outros a retenção pode custar quatro a oito vezes menos do que a aquisição de clientes (ASBRAND, 1997; KNOWLES, 1997).

Outros estudos mostram que, a cada cinco anos, as empresas americanas têm os seus clientes reduzidos pela metade; mantendo tal taxa, as empresas reduzem de 25% a 50% o seu desempenho financeiro (REICHHELD; TEAL, 1996). Deste modo, percebe-se que conquistar novos clientes acaba sendo mais caro do que manter os clientes já existentes (REICHHELD, 1996; REICHHELD; MARKEY Jr.; HOPTON, 2000).

Vavra (1994) reitera que quanto mais tempo se sustenta um cliente, mais lucrativo ele poderá ser, e quanto mais tempo o cliente comprar de uma mesma empresa, tornar-se-á mais dependente de seus produtos e/ou serviços, e estará menos vulnerável a ações da concorrência.

Segundo Botelho e Tostes (2010), clientes podem romper com uma organização mesmo que altos investimentos em retenção sejam feitos, o que requer identificação e entendimento sempre de forma objetiva. Corroborando, Reichheld e Sasser (1990) afirmam que quando uma empresa investe 5% em retenção de clientes, os lucros sobem entre 25% a 85%, dependendo do segmento da empresa, o que só reforça o fato de que a retenção (i.e. não abandono) pode está diretamente ligada ao grau de satisfação do cliente.

A retenção de clientes tem amplo valor econômico para as empresas (BUCKNIX; VAN DEN POEL, 2005; VAN DEN POEL; LARIVIÈRE, 2004; JACOB, 1994). Clientes antigos compram mais (PAULIN *et al.*, 1998; GANESH; ARNOLD; REYNOLDS, 2000) e, caso estejam satisfeitos, podem fazer um boca-a-boca positivo para a companhia (GANESH; ARNOLD; REYNOLDS, 2000). Uma retenção de clientes bem realizada diminui a necessidade de recrutar novos clientes, permitindo que as empresas se concentrem nas necessidades de seus clientes (VAN DEN POEL; LARIVIÈRE, 2004; DAWES; SWAILES, 1999).

Complementando, clientes antigos, além de serem menos propícios às ações de *marketing* da concorrência (VAN DEN POEL; LARIVIÈRE, 2004), tornam-se menos dispendiosos, devido ao melhor conhecimento do seu comportamento e à redução dos seus custos de manutenção (PAULIN *et al.*, 1998; GANESH; ARNOLD; REYNOLDS, 2000). Consequentemente, a perda de clientes não só acarreta custos de oportunidade ocasionados pela redução das vendas, mas também uma maior necessidade de conquistar clientes novos (VAN DEN POEL; LARIVIÈRE, 2004).

A retenção do consumidor para prevenir a sua ruptura se torna, pois, um ponto relevante para gestores de CRM (BUCKNIX; VAN DEN POEL, 2005). Para Grönroos (1995), visualizar o processo de relacionamento com o cliente como um ciclo de vida entre cliente-empresa é relevante, uma vez que está diretamente associado ao valor do cliente durante o tempo e com as ações necessárias para ampliá-lo e sustentá-lo.

Percebendo o processo de relacionamento como um ciclo de vida, agrupam-se em quatro estágios as experiências e interações vividas dos consumidores (GRÖNROOS, 1995). O primeiro estágio começa quando o consumidor ainda não conhece os serviços ofertados pela empresa (cliente potencial). O cliente avalia o produto ou serviço, contrastando com suas expectativas, e realiza sua primeira compra no segundo estágio (novo cliente). O estágio seguinte consiste no processo de consumo ou de uso, no qual se observa a competência da empresa em solucionar problemas e entregar o serviço ou produto (usuário). Finalmente, o processo termina quando o consumidor rompe o relacionamento com a empresa (ex-cliente).

Percebe-se que o objetivo do marketing de relacionamento nos dois estágios iniciais é estimular o interesse no consumidor e transformá-lo em venda. Quando o cliente se torna usuário, a empresa concentra os esforços para promover ações que proporcionem experiências positivas ao cliente, na esperança do cliente efetivar novas compras com a empresa (GRÖNROOS, 1995). Para tanto, as ações de relacionamento com o cliente estão sendo incorporadas ao negócio das empresas, pois são condutas de marketing, cujo cerne é garantir seus clientes, assumindo o desafio de reconhecê-los, mostrando-lhes o quanto a empresa os respeita por terem lhe dado a predileção (VAVRA; PRUDEN, 1995).

A necessidade de encontrar maneiras que permitam elevar a taxa de retenção de seus clientes é enfrentada, por muitas organizações, como a melhor estratégia de sobrevivência (WEI; CHIU, 2002). A retenção de clientes se transformou em algo essencial nas estratégias de marketing, centrando-se mais nos clientes e menos nos produtos (GANESH; ARNOLD; REYNOLDS, 2000; HADDEN *et al.*, 2006). A taxa de



retenção é percebida como elemento principal para que a empresa estabilize a perda de clientes, objetivando a satisfação do consumidor (RUST; ZAHORIK, 1993).

Com o propósito de melhorar a retenção de clientes, Galvão e Gonzalez (2011) sugerem que analisar o comportamento de seus clientes ao longo de sua existência na empresa pode ajudar na redução da taxa de ruptura, devido à identificação de fatos que expliquem a insatisfação do cliente.

Os autores afirmam que a taxa de ruptura, em níveis baixos, pode indicar uma melhoria do relacionamento entre cliente e empresa, tornando possível a retenção do mesmo. A taxa de ruptura (*churn rate*) de clientes diz respeito à tendência de os clientes pararem de efetuar negócios com uma determinada empresa por um determinado período de tempo (QIAN; JIANG; TSUI, 2006).

A maior atratividade de opções apresentadas por concorrentes, o baixo custo de troca de fornecedores, os elevados custos associados à atividade de fazer compras, o declínio da qualidade e o baixo valor percebido dos produtos ofertados, a maneira que a empresa reage às tentativas de ruptura de clientes e o baixo esforço da empresa em suprir as novas exigências e preferências dos consumidores são alguns fatores que colaboram para a ruptura de clientes com uma empresa (STROUSE, 1999; ASAARI; KARIA, 2000; MICHALSKI, 2004). Sabendo disso, as organizações precisam intensificar o seu empenho na identificação de clientes que, por inúmeros motivos, são predispostos a romper com a empresa (HADDEN *et al.*, 2005).

Day (2001) e Strouse (1999) declaram que não há uma razão determinante para que se ocorra a cisão, mas sim um conjunto de frustrações ou desapontamentos os quais levam à decisão de procurar bens ou serviços em outro lugar, e/ou deixar que um concorrente seja capaz de convencer o cliente de que pode lhe prestar melhores serviços ou bens.

A crescente facilidade que o cliente tem para trocar de empresa força as empresas a examinarem os dados e identificarem exatamente quem são os clientes que romperam e suas causas, para que se possa reduzir futuros abandonos. Day (2001) sustenta que alguns clientes são fáceis de serem identificados, haja vista que cancelaram seus contratos. Por outro lado, existem outros clientes que apenas diminuíram o volume de suas compras, passando a comprar da empresa uma fração menor dos serviços e/ou bens das suas necessidades.

Apesar de ser intuitivo admitir que os clientes que rompem com a empresa são aqueles mais sensíveis ao preço, Hoffman e Bateson (1997) afirmam que clientes que

mudam de empresa por causa do preço são, possivelmente, os mais infíéis, e precisam de descontos financeiros para continuarem sendo clientes. Pesquisas indicam que investir na qualidade dos serviços e/ou bens ofertados ao consumidor é a melhor solução para esse tipo de cliente (DESOUZA, 1992).

Botelho e Tostes (2010) esclarecem que a perda de consumidor por atrito no relacionamento (e.g. autorização não concedida em uma venda por cartão) ou o abandono do cliente (e.g. troca de produto/serviço sem motivo claro) pode ocorrer a qualquer momento do ciclo de vida deste relacionamento. Para isso, o gerenciamento do *churn* possibilita o uso de estratégias que impeçam tal atrito ou abandono, podendo manter seus clientes mais lucrativos (BOTELHO; TOSTES, 2010; GALVÃO; GONZALEZ, 2011).

### 3 GERENCIAMENTO DE CHURN

Neste capítulo, serão abordados conceitos, causas, alguns estudos realizados e as limitações da utilização de gerenciamento de *churn* em uma empresa, apontando a necessidade do desenvolvimento de modelos capazes de prever o tempo de vida do cliente dentro da organização. Para tal desenvolvimento, serão feitas considerações sobre data mining, focando em regressão logística, árvore de decisão e redes neurais.

#### 3.1 DEFINIÇÃO E CAUSAS DO *CHURN*

Originado da expressão inglês “*change and turn*”, o termo *churn* é comumente utilizado na literatura e na indústria para representar a descontinuação ou rompimento de um contrato (LAZAROV; CAPOTA, 2007), que pode ser definido como o movimento de clientes entre inúmeros fornecedores dentro de determinado serviço em comum (AU; MA, 2003; HADDEN *et al.*, 2005; HUNG; YEN; WAG, 2006; HONGXIA; MIN; JIANXIA, 2009). Entende-se tal termo como uma medida de deslealdade de uma base de clientes (ANDRADE, 2007; LEJEUNE, 2001), e sua operacionalização pode ser feita através da porcentagem de clientes que uma empresa perde num determinado intervalo de tempo (LEJEUNE, 2001).

Na literatura, identifica-se dois tipos de *churn*: involuntário e o voluntário. *Churn* involuntário acontece, por exemplo, quando um cliente para de pagar o serviço comprado e tem o fornecimento cancelado pela empresa (HADDEN *et al.*, 2005). Cister (2005) explica que as razões involuntárias são consequências de uma ação da própria empresa, que por algum motivo (e.g. fraude, falta de pagamentos, falta de utilização do serviço) viu-se obrigada a romper sua relação com o cliente.

O *churn* voluntário ocorre quando, de forma racional, o cliente resolve romper com uma companhia e não mais utilizar seus serviços (HADDEN *et al.*, 2005). Esta é dividida em dois tipos: acidental e o deliberado. No conceito de *churn* acidental, inúmeras situações fazem com que o cliente não seja capaz de manter o serviço (e.g. desemprego, mudança para uma região onde o serviço não é ofertado), entretanto este tipo de *churn* reflete uma pequena parcela da taxa de ruptura das organizações (HADDEN *et al.*, 2005; LAZAROV; CAPOTA, 2007).

*Churn* deliberado acontece em situações em que o cliente opta por mudar de fornecedor (HADDEN *et al.*, 2005; HADDEN *et al.*, 2006). Para este tipo, as causas mais comuns são encontradas em fatores associados às relações entre empresa e cliente e que

podem ser administrados pela própria empresa (por exemplo, clareza de faturação e serviço pós-venda). As organizações buscam, com mais vigor, combater esse tipo de *churn* (HADDEN *et al.*, 2005; HADDEN *et al.*, 2006; LAZAROV; CAPOTA, 2007).

Por um lado, em empresas de serviço, como empresas de telefonia, geralmente o cancelamento do relacionamento do cliente com a empresa precisa se dar de maneira formal. Em empresas americanas de cartão de crédito, o cliente pode facilmente deixar de efetuar pagamentos das faturas, mandar o cartão de volta à empresa ou deixar de usar o cartão de crédito. Por outro lado, para romper o relacionamento com empresas de varejo (e.g. supermercado), basta apenas o cliente deixar de comprar os produtos (BERRY; LINOFF, 1997).

As estratégias e os esforços de uma empresa para manter seus clientes mais lucrativos recebem o nome de gestão de *churn* (HUNG; YEN; WANG, 2006). Segundo Burez e Van den Poel (2007), tais estratégias são frequentemente apresentadas dentro de duas categorias: as não dirigidas (*untargeted*) e as dirigidas (*targeted*). Para os autores, as estratégias não dirigidas sustentam-se na publicidade em massa, com o objetivo de ampliar os níveis de lealdade da marca, enquanto as estratégias dirigidas evidenciam esforços em clientes que possuem a maior probabilidade de romper com a empresa.

As organizações criam desde programas de frequência até a participação em clubes para gratificar comportamentos de clientes que se mantêm consumidores, fortalecendo, assim, a marca da empresa (DAY, 2001; RUST *et al.*, 2001).

Rust *et al.* (2001) mostram cinco áreas de ação em que as companhias podem desenvolver programas para aumentar a probabilidade do cliente retornar a comprar da empresa, ou seja, criar uma necessidade no cliente de manter o contato com a empresa. Tais ações podem ser: (1) programas de criação de comunidade; (2) programas de criação de conhecimento; (3) programas de afinidade; (4) programas de lealdade e; (5) programas de reconhecimento e tratamento especiais.

Todavia, essas estratégias são dispendiosas, logo, falar de gestão de *churn* para vários autores (AU; MA, 2003; HUNG; YEN; WANG, 2006) significa ser capaz de prever o comportamento de um determinado consumidor, como trocar um serviço de uma empresa pelos dos seus concorrentes. Day (2001) entende que se uma empresa construiu um relacionamento firme e cativante com seus clientes, a lealdade foi atingida. Assim, a organização é capaz de reunir informações suficientes para desenvolver aptidões únicas, difíceis de serem equiparadas pela concorrência.

Por meio das estratégias dirigidas, a empresa enfatiza em clientes com maior

risco de abandoná-la, e oferece incentivos para a não efetivação da ruptura (BUREZ; VAN DEN POEL, 2007). As estratégias dirigidas podem ser divididas em dois tipos: reativas e as proativas.

As abordagens reativas acontecem quando a empresa aguarda o momento em que o cliente a aciona para efetuar o cancelamento do relacionamento. É neste momento que a empresa reage, apresentando-lhe um incentivo para manter tal relacionamento (e.g. futuros descontos). Quando se fala em abordagens proativas, a companhia se esforça para reconhecer, de forma antecipada, os clientes com maior probabilidade de romper com a mesma num futuro próximo (BUREZ; VAN DEN POEL, 2007). Tais abordagens não buscam caracterizar o *churn* ou observar a sua evolução ao longo do tempo, mas sim perceber e identificar os clientes em risco de romper com a empresa antes que isso aconteça de fato (MORIK; KOPCKE, 2004).

Burez e Van den Poel (2007) entendem que uma abordagem proativa pode ter grandes vantagens (e.g. incentivos com menores custos), uma vez que, normalmente, não se necessita gastar grandes valores se comparado com os valores das abordagens reativas. Entretanto, podem ser desperdiçadoras se o reconhecimento dos potenciais clientes de ruptura for impreciso, o que pode levar as empresas a gastarem recursos com os clientes errados, ou seja, os consumidores que não tinham a pretensão de romper o relacionamento com a empresa.

Assumindo que as abordagens proativas se respaldam na máxima que a empresa é capaz de identificar, antecipadamente, os clientes que possuem um maior risco de *churn* (MORIK; KOPCKE, 2004), estas precisam se equipar com ferramentas capazes de determinar, com o maior rigor possível, a probabilidade de um cliente específico romper com a empresa. Visto o desafio, as companhias devem estar preparadas com metodologias eficientes e eficazes para analisar o comportamento de seus clientes e prever a sua possível ruptura com a empresa num futuro não tão distante.

### 3.2 IMPLICAÇÕES SOBRE *CHURN*

A gestão de relacionamento com o cliente e a previsão da ruptura de clientes têm recebido uma atenção crescente na última década, tornando-se um desafio para empresas de vários ramos. O Quadro 1 apresenta uma visão global da literatura sobre o uso de técnicas de mineração de dados para modelos de previsão de *churn* pertencentes aos últimos anos, resumindo as técnicas de modelagem aplicadas, a área de estudo onde

os dados foram adquiridos, as revistas publicadas, o título do trabalho e seus autores.

Como pode ser visto a partir do referido quadro, várias técnicas de modelagem de dados têm sido testadas, com o propósito de encontrar a modelagem mais acurada: regressão logística, árvores de decisão, redes neurais, *support vector machine*, *random forests*, dentre outras. A compreensão de modelos de previsão de *churn* tem recebido pouca atenção na literatura (VERBEKE *et al.*, 2011), entretanto alguns autores já centraram seus esforços em analisar os gatilhos do rompimento com a empresa (BUCKNIX; VAN DEN POEL, 2005; KUMAR; RAVI, 2008), o que ilustra a necessidade de construir modelos que mostrem as causas da ruptura e sejam compreensíveis.

Com a ideia de perceber quais técnicas de *data mining* são mais utilizadas por acadêmicos e não acadêmicos, Neslin *et al.* (2006) realizaram uma avaliação comparativa entre 44 modelos preditivos de evasão para o contexto de telefonia móvel. O trabalho dos autores resumiu os resultados de um torneio de previsão de ruptura, o *Churn Modeling Tournament*, promovido pelo *Teradata Center for Customer Relationship Management* na *Duke University - The Fuqua School of Business*. Nesta competição, foi disponibilizado um banco de dados com 171 variáveis potencialmente preditivas a respeito de consumidores anônimos, foram definidas métricas de desempenho para definir os melhores modelos e uma análise detalhada do impacto econômico das campanhas de retenção também foi apresentada neste mesmo trabalho.

Ao final, a equipe vencedora do torneio, *Cardell, Golovnya e Steinberg*, apresentou a metodologia que utilizaram, aplicando uma técnica baseada em árvore de decisão, a fim de selecionar, de forma automática, os parâmetros relevantes para a modelagem preditiva (NESLIN *et al.*, 2006).

Seguindo a mesma direção de fazer um levantamento do estado da arte em *churn*, Hadden *et al.* (2005) revisaram algumas das tecnologias mais populares (e.g. árvore de decisão, redes neurais) identificadas na literatura, para o desenvolvimento de uma plataforma de gestão de *churn*. As vantagens e desvantagens das tecnologias identificadas são discutidas ao passo que os autores oferecem direções futuras de pesquisas na área. Outros autores (VERBEKE *et al.*, 2011; ABBASIMEHR; SETAK; SOROOR, 2013; ABBASIMEHR; SETAK; TAROKH, 2014) também realizam uma revisão da literatura, apontando inúmeros trabalhos já realizados sobre predição de *churn*.

**Quadro 1 - Visão geral da literatura de modelagem para predição de *churn***

N	Autores	Título	Journal	Ano	Área de estudo	Metodologia
1	WEI; CHIU, 2002;	<i>Turning telecommunications call details to churn prediction: a data mining approach</i>	<i>Expert Systems with Applications</i>	2002	Telecomunicação	Árvore de decisão
2	AU; CHAN; YAO, 2003;	<i>A novel evolutionary data mining algorithm with applications to churn prediction</i>	<i>Computers &amp; Operations Research</i>	2003	Telecomunicação	Árvore de decisão, Redes neurais (DMEL)
3	LEWIS, 2004;	<i>The Influence of Loyalty Programs and Short-Term Promotions on Customer Retention</i>	<i>Journal of Marketing Research</i>	2004	Venda Online (Itens de mercearia e drogaria)	Modelo de programação dinâmica
4	GUSTAFSSON; JOHNSON; ROOS, 2005;	<i>The Effects of Customer Satisfaction, Relationship Commitment Dimensions, and Triggers on Customer Retention</i>	<i>Journal of Marketing</i>	2005	Telecomunicação	Regressão logística
5	BUCKNIX; VAN DEN POEL, 2005;	<i>Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting</i>	<i>European Journal of Operational Research</i>	2005	Varejo de supermercado	Regressão logística, Redes neurais e Florestas aleatórias
6	HADDEN <i>et al.</i> , 2005;	<i>Computer assisted customer churn management: State-of-the-art and future trends</i>	<i>Computers &amp; Operations Research</i>	2005	Estudo teórico	Árvore de decisão, Redes neurais
7	HUNG; YEN; WANG, 2006;	<i>Applying data mining to telecom churn management</i>	<i>Expert Systems with Applications</i>	2006	Telecomunicação	Árvore de decisão, Redes neurais
8	LAMBRECHT; SKIERA, 2006;	<i>Paying Too Much and Being Happy About It: Existence, Causes, and Consequences of Tariff-Choice Biases</i>	<i>Journal of Marketing Research</i>	2006	Provedor de internet	Regressão logística
9	LEMMENS; CROUX, 2006;	<i>Bagging and boosting classification trees to predict churn</i>	<i>Journal of Marketing Research</i>	2006	Telecomunicação	Regressão logística, Bagging, Boosting,
10	NESLIN <i>et al.</i> , 2006;	<i>Defection detection: Measuring and understanding the predictive accuracy of customer churn models</i>	<i>Journal of Marketing Research</i>	2006	Telecomunicação	Regressão logística, Árvore de decisão, Redes neurais, Análise discriminante, Bayes

Fonte: Autoria própria.

**Quadro 1 - Visão geral da literatura de modelagem para predição de *churn* (Cont.)**

N	Autores	Título	Journal	Ano	Área de estudo	Metodologia
11	QIAN; JIANG; TSUI, 2006;	<i>Churn detection via customer profile modelling</i>	<i>International Journal of Production Research</i>	2006	Telecomunicação	Modelo misto funcional, Análise de <i>cluster</i>
12	BUREZ; VAN DEN POEL, 2007;	<i>CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services</i>	<i>Expert Systems with Applications</i>	2007	TV por assinatura	Regressão logística com cadeia de Markov, Florestas aleatórias
13	FIGUEIREDO; SILVERMAN, 2007;	<i>Churn, Baby, Churn: Strategic Dynamics Among Dominant and Fringe Firms in a Segmented Industry</i>	<i>Management Science</i>	2007	Indústria de impressora a laser	Análise regressão com dados em painel
14	BUREZ; VAN DEN POEL, 2008;	<i>Separating financial from commercial customer churn: a modeling step towards resolving the conflict between the sales and credit department</i>	<i>Expert Systems with Applications</i>	2008	TV por assinatura	Florestas aleatórias, Análise de sobrevivência
15	COUSSEMENT; VAN DEN POEL, 2008a;	<i>Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques</i>	<i>Expert Systems with Applications</i>	2008	Assinatura de Jornal	SVM, Regressão logística, Florestas aleatórias
16	COUSSEMENT; VAN DEN POEL, 2008b;	<i>Integrating the voice of customers through call center emails into a decision support system for churn prediction</i>	<i>Information and Management</i>	2008	Assinatura de Jornal	Regressão logística
17	SCHWEIDEL; FADER; BRADLOW, 2008;	<i>Understanding Service Retention Within and Across Cohorts Using Limited Information</i>	<i>Journal of Marketing</i>	2008	Telecomunicação	Análise de sobrevivência
18	KUMAR; RAVI, 2008;	<i>Predicting credit card customer churn in banks using data mining</i>	<i>International Journal of Data Analysis Techniques and Strategies</i>	2008	Cartão de Crédito	Regressão logística, Árvore de decisão, Redes neurais, SVM, Florestas aleatórias
19	MENDES, 2008;	Modelos de <i>churn</i> de clientes em planos de saúde	Dissertação - UFF	2008	Plano de saúde	Regressão logística, <i>Propensity Score Matching</i>
20	GLADY; BAESENS; CROUX, 2009;	<i>Modeling churn using customer lifetime value</i>	<i>European Journal of Operational Research</i>	2009	Companhia de Serviços Financeiros	Regressão logística, Árvore de decisão, Redes neurais

Fonte: Autoria própria.



**Quadro 1 - Visão geral da literatura de modelagem para predição de *churn* (Cont.)**

N	Autores	Título	Journal	Ano	Área de estudo	Metodologia
21	PENDHARKAR, 2009;	<i>Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services</i>	<i>Expert Systems with Applications</i>	2009	Telecomunicação	Algoritmo genético baseado em redes neurais
22	TSAI; LU, 2009;	<i>Customer churn prediction by hybrid neural networks</i>	<i>Expert Systems with Applications</i>	2009	Telecomunicação	Redes neurais (Self-Organising Maps)
23	XIE <i>et al.</i> , 2009;	<i>Customer churn prediction using improved balanced random forests</i>	<i>Expert Systems with Applications</i>	2009	Banco	Redes neurais, Árvore de decisão, SVM, Florestas aleatórias, Florestas aleatórias melhor equilibradas
24	BOTELHO; TOSTES, 2010;	Modelagem de probabilidade de <i>churn</i>	RAE-Revista de Administração de Empresas	2010	Cartão de Crédito	Regressão logística
25	COUSSEMENT; BENOIT; VAN DEN POEL, 2010;	<i>Improved marketing decision making in a customer churn prediction context using generalised additive models</i>	<i>Expert Systems with Applications</i>	2010	Assinatura de Jornal	Regressão logística, Modelos aditivos generalizados (GAM)
26	HUANG; BUCKLEY; KECHADI, 2010;	<i>Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications</i>	<i>Expert Systems with Applications</i>	2010	Telecomunicação	Árvore de decisão
27	JAHROMI <i>et al.</i> , 2010;	<i>Modeling customer churn in a non-contractual setting: the case of telecommunications service providers</i>	<i>Journal of Strategic Marketing</i>	2010	Telecomunicação	Análise de <i>cluster</i> , Árvore de decisão
28	OWCZARCZUK, 2010;	<i>Churn models for prepaid customers in the cellular telecommunication industry using large data marts</i>	<i>Expert Systems with Applications</i>	2010	Telecomunicação	Regressão logística, Regressão linear, Análise discriminante de Fisher, Árvore de decisão
29	KARAOHOCA; KARAOHOCA, 2011;	<i>GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system</i>	<i>Expert Systems with Applications</i>	2011	Telecomunicação	Fuzzy C-means <i>clustering</i> , adaptive neuro-fuzzy inference system (ANFIS)
30	NIE <i>et al.</i> , 2011;	<i>Credit card churn forecasting by logistic regression and decision tree</i>	<i>Expert Systems with Applications</i>	2011	Cartão de Crédito	Regressão logística, Árvore de decisão

Fonte: Autoria própria.

**Quadro 1 - Visão geral da literatura de modelagem para predição de churn (Cont.)**

N	Autores	Título	Journal	Ano	Área de estudo	Metodologia
31	KISIOGLU; TOPCU, 2011;	<i>Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey</i>	<i>Expert Systems with Applications</i>	2011	Telecomunicação	Redes Bayesianas
32	NITZAN; LIBAI, 2011;	<i>Social Effects on Customer Retention</i>	<i>Journal of Marketing</i>	2011	Telecomunicação	Análise de sobrevivência
33	VERBEKE, 2011;	<i>Building comprehensible customer churn prediction models with advanced rule induction techniques</i>	<i>Expert Systems with Applications</i>	2011	Telecomunicação	Antminer+, (ALBA), RIPPER, SVM, Regressão logística, Árvore de decisão
34	YU <i>et al</i> , 2011;	<i>An extended support vector machine forecasting framework for customer churn in e-commerce</i>	<i>Expert Systems with Applications</i>	2011	Venda Online (E-commerce)	SVM, ANN, <i>extended support vector machine</i> (ESVM)
35	ABBASIMEHR; SETAK; SOROOR, 2013;	<i>A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques</i>	<i>International Journal of Production Research</i>	2013	Telecomunicação	Neuro-Fuzzy
36	PHADKE <i>et al</i> , 2013;	<i>Prediction of Subscriber Churn Using Social Network Analysis</i>	<i>Bell Labs Technical Journal</i>	2013	Telecomunicação	Modelo híbrido para análise de redes sociais
37	ABBASIMEHR; SETAK; TAROKH, 2014;	<i>A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction</i>	<i>The International Arab Journal of Information Technology</i>	2014	Telecomunicação	SVM, Árvore de decisão, Redes neurais, Regressão logística
38	KNOX; OEST, 2014;	<i>Customer Complaints and Recovery Effectiveness: A Customer Base Approach</i>	<i>Journal of Marketing</i>	2014	Venda Online (Varejo)	Regressão logística, Distribuição beta geométrica/binomial negativa, Distribuição de Weibull
39	VAFEIADIS <i>et al.</i> , 2015;	<i>A comparison of machine learning techniques for customer churn prediction</i>	<i>Simulation Modelling Practice and Theory</i>	2015	Telecomunicação	SVM, Árvore de decisão, Redes neurais, Regressão logística
40	MOEYERSOMS; MARTENS, 2015;	<i>Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector</i>	<i>Decision Support Systems</i>	2015	Sector Energético	SVM, Árvore de decisão, Regressão logística

Fonte: Autoria própria.

Observa-se no Quadro 1 que o setor de telecomunicações é a área que mais abre espaço para os pesquisadores desenvolverem novas técnicas e tecnologias para predição de ruptura de clientes. Qian, Jiang e Tsui (2006), por exemplo, propuseram a utilização de um modelo misto funcional para predizer o comportamento do consumidor em relação à sua ruptura como cliente. Tal estudo foi dividido em 5 etapas: (1) padronização de perfis; (2) triagem de perfis irrelevantes; (3) projeção de perfis em um espaço representado por um conjunto de funções base; (4) aplicação do algoritmo de *cluster* para os coeficientes resultantes no espaço; e (5) a identificação de perfis interessantes (clientes com mais risco de romper com a empresa). Os autores perceberam que, para o setor de telecomunicações, o modelo proposto foi eficaz para detectar atividades de ruptura.

Reconhecendo que a perda de clientes de baixo valor é, naturalmente, menos danosa do que a ruptura de clientes leais e de alto valor, Abbasimehr, Setak e Soroor (2013) apresentaram um estudo dividido em duas fases para a predição de clientes valiosos propensos ao rompimento. Com dados de uma grande empresa de telecomunicações, a primeira fase consistiu-se na identificação dos clientes de alto valor, por meio da análise de *cluster* (*K-means*), no qual os *clusters* foram ranqueados e o grupo mais bem classificado foi usado na segunda fase. Nesta fase de desenvolvimento do modelo, duas técnicas de *neuro-fuzzy* (ANFIS e LLNF) foram aplicadas em conjunto com algoritmo de aprendizagem LoLiMoT. Os pesquisadores desenvolveram um algoritmo para comparar os modelos propostos com os métodos de redes neurais mais utilizados (MLP e RBF). Os autores concluíram que as técnicas de *neuro-fuzzy* tiveram melhores desempenhos do que os modelos de redes neurais.

Schweidel, Fader e Bradlow (2008) propõem um quadro para examinar fatores que podem ser a base da retenção em um ambiente contratual em uma empresa de telecomunicações. Especificamente, eles utilizam um modelo de retenção que corresponde a: (1) dependência de duração; (2) efeitos de promoção; (3) heterogeneidade de assinante; (4) os efeitos de grupo cruzado; e (5) os efeitos do tempo (e.g. sazonalidade). De acordo com o levantamento feito, percebe-se que ambientes contratuais são os mais estudados pelos pesquisadores.

Diferentemente de Schweidel, Fader e Bradlow (2008), Jahromi *et al.* (2010) buscaram desenvolver um modelo preditivo para a ruptura de clientes no Irã em configurações não-contratuais, ou seja, em clientes que possuem celulares pré-pagos. Em um campo pouco explorado, os autores construíram um modelo de duas etapas, análise

de *cluster* e classificação, em uma base de 34.504 clientes. Os resultados confirmam o desempenho satisfatório dos modelos propostos, sendo assim mostra-se um modelo promissor para identificação de clientes propensos a romper com a empresa.

O desenvolvimento de modelos que buscam prever o *churn* se torna mais difícil quando se trata de uma situação não contratual, dado que os clientes podem mudar constantemente o seu comportamento de compra sem informar à empresa. Consequentemente, acompanhar e estimar o *churn* têm sido um transtorno de alta complexidade para as organizações (BUCKNIX; VAN DEN POEL, 2005).

Os clientes brasileiros não têm diversas opções para trocar de fornecedor com facilidade, porém, já começam a sentir a liberdade de escolha no setor de telecomunicações. Outros setores da economia aprenderam a administrar o *churn* há tempos, e administradoras de cartão de crédito e bancos são exemplos bastante conhecidos.

Botelho e Tostes (2010) propuseram modelar a probabilidade de clientes romperem com a empresa e descrever as possíveis variáveis que influenciam tal rompimento/permanência do cliente. A base de dados consistia em 100 mil (70mil para calibração e 30 mil para validação) clientes que possuíam cartão de crédito próprio de uma grande rede varejista. O modelo de regressão logística desenvolvido foi avaliado pelo teste de KS (*Kolmogorov-Smirnov*) e pela curva ROC (*Receiver Operating Characteristic*), que apresentaram boa adequação do modelo aos dados e possuíam 16 variáveis, sendo 14 informações individuais, em relação ao cadastro demográfico do cliente, e duas comportamentais, em relação ao uso do cartão de crédito pelo cliente.

Em um trabalho similar ao de Botelho e Tostes (2010), Nie *et al.* (2011) utilizam dados de uma operadora de cartão de crédito coletados por um banco chinês. Em seu trabalho, os autores examinaram 135 variáveis, divididas em quatro categorias: informações do cliente, informações do cartão, informações de risco e informações das atividades de transações. Aplicando técnicas de seleção de variáveis, foram construídos inúmeros modelos de regressão logística e árvores de decisão, cujos resultados do teste mostram que a regressão logística executa uma performance um pouco melhor do que a árvore de decisão.

Ainda utilizando banco de dados de um banco chinês, Xie *et al.* (2009) buscaram solucionar o problema de desequilíbrio na distribuição de dados. Assim, os autores propuseram um novo método de aprendizagem, chamado florestas aleatórias (em inglês, *random forest*) equilibradas melhoradas. Foi descoberto que este método produz

uma precisão melhor da previsão, em comparação com outros algoritmos, tais como redes neurais e árvores de decisão. Entretanto, tal método não é o foco desta dissertação, uma vez que foi definido trabalharmos apenas com as técnicas árvore de decisão, regressão logística e redes neurais. Conforme o Quadro 1, a literatura de modelagem para predição de *churn* indica essas técnicas de modelagem como as mais utilizadas nas pesquisas.

Usando uma base de dados de uma empresa varejista (supermercado), Bucknix e Van den Poel (2005) se concentraram no tratamento dos clientes mais leais em um ambiente não contratual. As técnicas de regressão logística, determinação de relevância automática, redes neurais e *random forest* foram utilizadas para o estudo. Não foram encontradas diferenças significativas em termos de desempenho entre elas.

Identificando que nenhum estudo mediu o impacto das reclamações formais e ações de recuperações em compras posteriores dos clientes, Knox e Oest (2014) desenvolvem um modelo para investigar a eficácia das ações de recuperação na prevenção de ruptura de clientes. Os dados do estudo foram fornecidos por uma empresa de varejo pela internet, com informações de reclamações e ações de recuperação de novos clientes em um período de dois anos e meio.

Os pesquisadores (KNOX; OEST, 2014) perceberam que reclamações estão associadas a um aumento significativo da probabilidade de o cliente parar de comprar, entretanto o tamanho desse aumento depende de experiências anteriores dos clientes. Seguindo esse contexto, compras anteriores diminuem o efeito, e o seu impacto é de longa duração, ao passo que queixas anteriores aumentam o efeito e seu impacto é de curta duração. Os autores usaram técnica de simulação para entender os resultados ao impacto financeiro.

O estudo realizado por Yu *et al.* (2011) comparou redes neurais, árvore de decisão, *support vector machine* (SVM) e *extended support vector machine* (ESVM) para a previsão de *churn* de clientes em *e-commerce*. O resultado mostrou que, entre as outras técnicas, a ESVM executou melhor.

Burez e Van den Poel (2007) analisaram a base de dados de uma empresa que fornece TV por assinatura na Europa. O artigo desenvolve diferentes modelos de previsão de *churn*: cadeias de Markov e um modelo de *random forest* são referenciados para um modelo logístico. Os resultados obtidos no experimento de campo mostraram que os lucros podem ser duplicados usando o modelo proposto de previsão de *churn*.

No ano seguinte, Burez e Van den Poel (2008), novamente com uma empresa de TV por assinatura, propõem a existência de dois grupos de clientes que abandonam a

empresa: os que não renovam seus contratos fixos no final do contrato e os que simplesmente param de pagar durante o contrato a que estão legalmente obrigados. Os autores chamam o primeiro tipo de *churn* comercial e o segundo de *churn* financeiro. Os resultados mostram que o conhecimento prévio de um mal é mais importante como um *input* para o *churn* financeiro do que para *churn* comercial. Ademais, pode-se prever, com mais precisão, o *churn* financeiro do que o comercial.

Por outro lado, ao tentar persuadir os clientes a ficar com a empresa, o impacto das ações de "fidelização" é muito maior com potenciais *churners* comerciais, em comparação com *churners* financeiros. Isso considerando que os *churners* comerciais são clientes com poder de compra mais equilibrado em relação ao dos *churners* financeiros, que são clientes de risco financeiro no primeiro trimestre (BUREZ; VAN DEN POEL, 2008).

Em um contexto onde pessoas são socialmente afetadas por outras, a rede social dos clientes é um importante objeto de estudo. Diante disso, Nitzan e Libai (2011) criaram um sistema social em larga escala, composto de redes sociais individuais dos clientes. Isso foi feito por meio de uma grande base de dados (entre um milhão de clientes) obtida de uma empresa de telecomunicação. O estudo indicou que a exposição a um vizinho que já rompeu está filiada com um aumento de 80% no risco de ruptura, após o controle de uma série de variáveis sociais, pessoais e relacionadas à compra. Clientes altamente conectados são mais afetados, enquanto que clientes fiéis são menos afetados por rupturas que ocorrem em suas redes sociais.

Em resumo, percebe-se que o número de áreas de aplicação desses estudos é grande, uma vez que já existem estudos realizados em empresas de telecomunicação, TV por assinatura, supermercados, vendas *online*, jornais, bancos e cartões de crédito. Nos anos de 2008 e 2011 houve bastante publicação sobre o tema, constando 6 artigos em cada ano. Observa-se que, em 2015, foram publicados dois artigos sobre o tema, o que mostra a atualidade do assunto trabalhado nesta dissertação.

O Quadro 2 apresenta um resumo da quantidade de artigos publicados em cada área. Há inúmeras técnicas estatísticas que podem ser utilizadas para tal análise (e.g. regressão logística, redes neurais, árvore de decisão, *neuro-fuzzy*, etc.).

**Quadro 2 - Quantidade de publicações por área de estudo**

Área de estudo	Frequência
Telecomunicação	21
Cartão de Crédito	3
Venda Online	3
Assinatura de Jornal	3
TV por assinatura	2
Banco	1
Companhia de Serviços Financeiros	1
Estudo teórico	1
Indústria de impressora a laser	1
Provedor de internet	1
Varejo de supermercado	1
Setor Energético	1
Plano de saúde	1

Fonte: Autoria própria.

Vale ressaltar que foi encontrada apenas uma dissertação que abordasse a ruptura de cliente na área de plano de saúde, mostrando, assim, que ainda é uma área bastante inexplorada na academia.

### 3.3 LIMITAÇÕES DA UTILIZAÇÃO DE GERENCIAMENTO DE *CHURN*

Algumas limitações para a utilização de gerenciamento de *churn* podem ser identificadas, assim como o custo de sua aplicação. Investimentos em tecnologia para armazenagem, processamento de dados e capacitação profissional (teórica e técnica) para exploração dos dados e análise estatística seriam necessários (DARÉ, 2007).

Contribuindo para aumentar o risco do gerenciamento de *churn*, Berry e Linoff (2000) atentam para as técnicas de modelagem preditiva utilizadas, pois suas aplicações seriam perigosas se fossem realizadas por pessoas sem conhecimento necessário ou sem a supervisão de pessoas especializadas.

A utilização do *churn* como mecanismo de controle necessita da existência de um *Database Marketing* e de um gerenciamento do relacionamento entre o cliente e a empresa, pressupondo investimentos por parte da organização, que deveria perguntar antes se estaria predisposta a realizar estes investimentos (DARÉ, 2007).

Conforme Daré (2007), para que o gerenciamento de *churn* exista, entende-se que esse relacionamento seja de longo prazo. Tal afirmação se justifica pelo fato de existirem produtos para uso apenas uma vez. Nesses casos, talvez não tenha lógica manter um relacionamento subsequente, haja vista que o gerenciamento de *churn* não se mostra aplicável (DARÉ, 2007).

Em setores altamente competitivos (como os setores de cartões de crédito e telecomunicação), a satisfação do cliente possui papel notório para o seu relacionamento e para sua manutenção (BOTELHO; TOSTES, 2010), tornando o gerenciamento de *churn* vital para a sobrevivência das empresas. Entretanto, em setores onde a competição não é acirrada, talvez o gerenciamento de *churn* não seja necessário (DARÉ, 2007).

O gerenciamento de *churn* se importa com o tipo de cliente que romperá com a empresa; assim, a gestão precisa identificar quem são os bons ou maus clientes e os clientes mais ou menos lucrativos, para poder decidir quais esforços serão tomados para mantê-los (ABBASIMEHR; SETAK; SOROOR, 2013; DARÉ, 2007).

### 3.4 DATA MINING

O crescimento tecnológico registrado nas últimas décadas mostra que é cada vez mais simples capturar e armazenar dados dos clientes, possibilitando, assim, que empresas reúnam grandes registros de transações de clientes (QIAN; JIANG; TSUI, 2006). Gomes (2011) diz que é muito difícil encontrar uma empresa de médio ou de grande porte que não use sistemas avançados de informação para a melhoria da gestão do seu negócio. Porém, tal aumento na simplicidade em obter e guardar grandes volumes de dados não foi acompanhado de perto pela capacidade de interpretar esses mesmos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Os dados produzidos e colhidos de forma digital viabilizam ações individualizadas aos consumidores por meio da utilização de técnicas de mineração de dados, juntamente ao CRM (BOTELHO; TOSTES, 2010). Contudo, os dados armazenados não são, em grande maioria dos casos, de fácil interpretação, ao se aplicar meios de análise convencionais (GOMES, 2011).

*Data mining*, ou mineração de dados, é a aplicação de um algoritmo específico para a extração de padrões implícitos dos dados (LAROSE, 2005; MAIMON; ROKACH, 2010). Mais especificamente, busca a detecção automática de padrões



previamente desconhecidos e potencialmente úteis que não se encontram explicitamente discriminados numa base de dados (HONGXIA; MIN; JIANXIA, 2009). Ferramenta imprescindível nos modelos de negócio atuais, o *data mining* possibilita a transformação dos dados em informações confiáveis, a fim de dar suporte ao processo de tomada de decisão (GOMES, 2011).

Percebe-se que o uso de *data mining*, nos últimos anos, se tornou uma alternativa acessível e importante, tanto na capacidade de processamento e de armazenamento de dados quanto na geração de vantagem estratégica para a empresa (GOMES, 2011). Oferece uma enorme variedade de algoritmos para extração de conhecimento das bases de dados que pretendem resolver problemas como a descrição ou classificação de entidades, agrupamento, descoberta de afinidades, estimação de propriedades, predição (LAROSE, 2005; MAIMON; ROKACH, 2010). A classificação de entidades e a predição são as tarefas mais comuns na área de apoio à tomada de decisões nas empresas (GOMES, 2011).

Pretendendo transformar os dados em conhecimento, surge o processo chamado de CRISP-DM (*Cross-Industry Standard Process for Data Mining*), criado em 1996 para padronizar conceitos e técnicas na busca de informações específicas oriundos de grandes bases de dados para tomada de decisões. A metodologia é formada por um conjunto de fases e processos para *data mining*, independentemente de ferramentas e da área de negócios. Não necessariamente executadas nessa ordem, as fases são: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação (SHEARER, 2000). O processo CRISP-DM será melhor detalhado por esta dissertação nos procedimentos metodológicos.

O recente crescimento dos campos de *data mining* e a descoberta de conhecimento não são imprevisíveis do modo que uma variedade de métodos esteja acessível para os pesquisadores e profissionais, lembrando que, para todos os casos, não há método superior aos outros (MAIMON; ROKACH, 2010). A disponibilidade de dados cresce exponencialmente, à medida que o nível de processamento humano é quase constante. Deste modo, surge uma oportunidade para o CRISP-DM se tornar cada vez mais necessário.

Em *marketing*, a principal aplicação de *data mining* é em sistemas de *database marketing*, que analisam dados de clientes para identificar diferentes grupos de consumidores, de forma a prever seus comportamentos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Berry (1994) estima que mais da metade de todos os

varejistas usam ou planejam usar um *database marketing*, e completa afirmando que aqueles que o utilizam alcançam bons resultados. Com o uso apropriado, a empresa *American Express* relata um aumento de 10 a 15% no uso do cartão de crédito por parte dos clientes (BERRY, 1994).

*Database marketing* é um banco de dados de clientes em potencial e de todos os contatos comerciais ou de comunicação, podendo o *marketing* utilizar-se dele para expandir seu público-alvo, estimular a demanda deste público e estar próximo a ele, visando ao melhoramento de contatos futuros (STONE; SHAW, 1987). Um *database marketing* bem estruturado e que armazene o histórico de transações dos clientes (entre outras informações), de tal forma que o monitoramento seja contínuo e assegure a integridade das informações dos consumidores, facilita a aquisição de diferencial competitivo para as empresas (STONE; SHAW, 1987).

Segundo Stone e Shaw (1987) e corroborado por Larose (2005) e Maimon e Rokach (2010), as empresas utilizam *database marketing* de diferentes maneiras, e por isso nem todas as características deste são visíveis. Para os autores, as características básicas de um *database marketing* são:

- a) cada cliente potencial ou atual é identificado em um registro no banco de dados. Mercados e segmentos de mercado não são identificados através de dados agregados, que não podem ser decompostos em clientes individuais, mas como aglomerações de clientes individuais;
- b) cada registro de cliente contém não apenas a identificação e acesso à informação, mas também uma série de informações importantes para o *marketing*. Isso inclui informações sobre as necessidades e características dos clientes (e.g. informações demográficas e psicográficas), que são usadas para identificar os prováveis compradores de produtos específicos e como eles devem ser abordados. Também inclui informações sobre transações passadas, como o histórico de concorrentes, e as comunicações sobre a campanha (e.g. caso o cliente tiver sido exposto a campanhas de *marketing*);
- c) a informação está disponível para a empresa durante cada transação com o cliente, para que se possa decidir como responder às necessidades do cliente;
- d) o banco de dados registra as respostas dos clientes aos incentivos da empresa (e.g. campanhas de vendas);
- e) a informação está também disponível para os formuladores de políticas de *marketing*, para que possam decidir quais segmentos de mercado são

apropriados para cada produto ou serviço e qual *mix* de *marketing* (preço, comunicações de *marketing*, canais de distribuição etc.) é apropriado para cada produto em cada mercado-alvo;

f) em grandes empresas que vendem muitos produtos para cada cliente, a base de dados é utilizada para assegurar que uma abordagem coordenada e consistente ao cliente seja desenvolvida;

g) o banco de dados eventualmente substitui a pesquisa de mercado. Campanhas de *marketing* são realizadas de maneira que a resposta dos clientes para a campanha forneça a informação que a empresa está procurando;

h) automação da gestão de *marketing* é desenvolvida para trabalhar com o grande volume de informações geradas pelo *database marketing*, identificando oportunidades e ameaças de forma automática e recomendando formas de capturar as oportunidades e neutralizar as ameaças.

A vasta capacidade de armazenamento e processamento de dados dos atuais computadores é responsável pela viabilidade de existência do *database marketing* (STONE; SHAW, 1987; LAROSE, 2005; MAIMON; ROKACH, 2010). O levantamento de dados para o *database marketing* vem diretamente dos clientes, fazendo com a empresa estreite o relacionamento com o consumidor e, além disso, mostra o interesse que a empresa tem no cliente, buscando compreender o que ele pensa e deseja (HOLTZ, 1994).

Qian, Jiang e Tsui (2006) afirmam que as bases de dados disponíveis para as empresas são desafiadoras para o desenvolvimento de modelos generalizados de previsão para *churn*. Os autores indicam alguns exemplos de possíveis problemas que os analistas podem encontrar, tais como: perfis com informações históricas curtas, presença de tendência e a existência de correlações seriais ao longo do tempo. Os pesquisadores completam dizendo que os perfis são repetidamente interrompidos por mudanças no negócio, como vários ajustes contábeis, inovações tecnológicas e substituição de produtos no varejo.

Para que campanhas de *marketing* sejam mais eficazes e atinjam grupos específicos de consumidores, milhares de clientes de um *database marketing* são modelados simultaneamente em um modelo global e flexível, que captura as dinâmicas comerciais e não comerciais e que categoriza os consumidores de acordo com o seu comportamento (QIAN; JIANG; TSUI, 2006). O problema do *churn*, foco do presente estudo, é tradicionalmente contextualizado como um problema de classificação, e é

resolvido, na maioria dos casos, com recursos a técnicas de *data mining* (ANDRADE, 2007; AU; MA, 2003; BUREZ; VAN DEN POEL, 2007; HUNG; YEN; WANG, 2006).

Em uma análise da literatura sobre *data mining*, inúmeras técnicas para previsão e classificação de ruptura já foram aplicadas. A presente dissertação busca comparar as seguintes técnicas:

- a) árvore de decisão (e.g. ABBASIMEHR; SETAK; TAROKH, 2014; GLADY; BAESENS; CROUX, 2009; HUANG; BUCKLEY; KECHADI, 2010; HUNG; YEN; WANG, 2006; NESLIN *et al.*, 2006; NIE *et al.*, 2011; VERBEKE, 2011; WEI; CHIU, 2002; XIE *et al.*, 2009);
- b) regressão logística (e.g. ABBASIMEHR; SETAK; TAROKH, 2014; BOTELHO; TOSTES, 2010; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; COUSSEMENT; VAN DEN POEL, 2008a; COUSSEMENT; VAN DEN POEL, 2008b; GLADY; BAESENS; CROUX, 2009; LEMMENS; CROUX, 2006; NESLIN *et al.*, 2006; NIE *et al.*, 2011; VERBEKE *et al.*, 2011);
- c) e redes neurais (e.g. ABBASIMEHR; SETAK; TAROKH, 2014; GLADY; BAESENS; CROUX, 2009; NESLIN *et al.*, 2006; XIE *et al.*, 2009; YU *et al.*, 2011).

### 3.4.1 Árvore de decisão

Árvore de decisão é uma das técnicas de *data mining* mais utilizadas, pois é de fácil implementação e compreensão de seu resultado final (QUINLAN, 1986). Podem, ainda, ser caracterizadas como robustas ferramentas de classificação e representações do conhecimento (QUINLAN, 1986; PETERMANN, 2006).

As árvores de decisão são constituídas por (1) nós, representando os atributos; (2) ramos, que recebem os valores possíveis provenientes dos nós; e (3) folhas, representando as diferentes classes da variável dependente de um conjunto de dados de treino (GARCIA, 2003). Diferentemente das árvores na natureza, que crescem da raiz para as folhas, as árvores de decisão são delineadas numa metodologia *topdown*, ou seja, a raiz da árvore está no topo, e os ramos vão crescendo até chegar a uma folha (QUINLAN, 1986). Neste sentido, percebe-se que os ramos mais superiores da análise representam as variáveis que mais oferecem ganho de informação e que convergem mais

rapidamente para uma folha e, por conseguinte, para uma conclusão (MAIMON; ROKACH, 2010; LAROSE, 2005; PETERMANN, 2006).

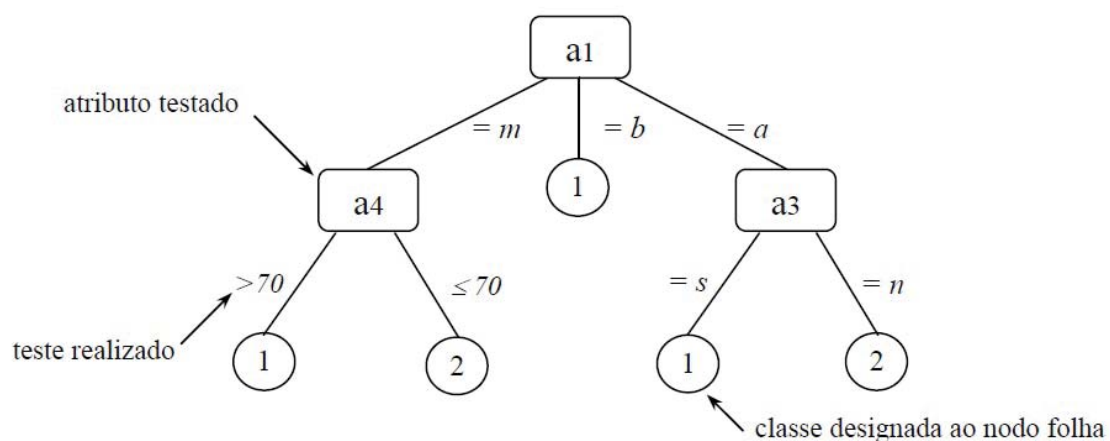
Sabe-se que cada folha representa uma classe e que cada nó intermediário representa um teste, o qual leva em consideração um ou mais atributos; cada possível resultado do teste realizado origina uma nova subárvore (QUINLAN, 1987). Assim, a construção de um modelo de árvores de decisão parte de um banco de dados de treino, sendo um método de aprendizagem supervisionada.

A aprendizagem supervisionada ocorre a partir de exemplos, ou seja, são definidas classes e exemplos para cada classe, e o sistema precisa formular a regra de classificação que pode ser utilizada para prever a classe. Na aprendizagem não supervisionada, a rede irá descobrir, sozinha, padrões, relações ou categorias nos dados. Ao final, são apresentadas descrições de classes, uma para cada classe descoberta na base de dados (MAIMON; ROKACH, 2010; GARCIA, 2003)

Ao dividir o banco de dados com base no resultado de um teste em uma de suas variáveis, o processo de crescimento acontece de forma iterativa. As divisões ocorrem de forma recursiva em cada subárvore formada da divisão anterior, e acabam quando todas as observações de um nó pertencerem à mesma classe ou quando a realização de mais divisões não originar aumento de valor nas previsões (MAIMON; ROKACH, 2010; GARCIA, 2003).

Uma árvore de decisão simples, construída a partir de uma base com três atributos, é apresentada como um exemplo no trabalho de Garcia (2003), conforme Figura 1 a seguir.

**Figura 1 - Exemplo de um classificador utilizando árvore de decisão**



Fonte: Garcia (2003).

Neste exemplo são testados atributos quantitativos e categóricos. Os nós são representados pelos atributos  $a1$ ,  $a3$  e  $a4$ , e os testes são feitos de acordo com o tipo de variável presente no nó. Se for variável categórica, são representados por uma igualdade, e. g. “=  $b$ ”, onde  $b$  é um valor dos atributos testados. Se for uma variável quantitativa, são representados por um intervalo de valor, e. g. “> 70”, sendo este intervalo alcançado por meio de cálculo (GARCIA, 2003). As árvores podem avançar a um elevado nível de complexidade, principalmente quando existe um grande número de atributos e classes (GOMES, 2011).

Após sua implementação, a classificação/predição de um novo indivíduo é realizada navegando pela árvore, desde a raiz da árvore até uma folha, sendo este novo indivíduo classificado conforme a classe representada por essa folha (QUINLAN, 1986; PETERMANN, 2006; GOMES, 2011). Por exemplo, assume-se que se  $a1$  possui valor igual a  $m$  e se o atributo  $a4$  tem valor maior que 70, o indivíduo seria classificado na classe 1. Diante disso, pode-se assumir a seguinte regra:

$$SE\ a1 = m\ E\ a4 > 70\ ENTÃO\ classe\ 1.$$

Conforme Lewis (2000), existem algumas vantagens das árvores de decisão em relação aos outros métodos de classificação. Em primeiro lugar, existe uma grande simplicidade na compreensão e na análise dos seus resultados. Em segundo lugar, um resultado obtido num modelo é fácil de ser comprovado por operações lógicas, diferentemente de outras técnicas (e.g. redes neurais). Por último, opera tanto sobre dados numéricos como sobre dados categóricos, e não se faz um intenso tratamento prévio para assegurar a qualidade dos dados (e.g. normalização dos dados, tratamento de valores extremos), ao contrário de outras técnicas de *data mining*.

Por outro lado, também há desvantagens de se usar árvores de decisão. Maimon e Rokach (2010) apontam dois pontos fracos: (1) a maioria dos algoritmos exigem que o atributo de destino (variável resposta) tenha apenas valores discretos; e (2) trazem bons resultados se existirem alguns atributos altamente relevantes, mas resultados ruins se muitas interações complexas forem necessárias. Quinlan (1993) também afirma que a técnica possui excesso de sensibilidade em relação ao conjunto de treino, a atributos

irrelevantes e a ruídos.

Maximizar a acurácia dos resultados é o objetivo de todos os algoritmos para a construção de uma árvore de decisão. Entretanto, estes diferem na forma de realizar o crescimento da árvore, isto é, na técnica e na métrica aplicada para determinar qual variável utilizar e qual ponto de divisão escolher (MAIMON; ROKACH, 2010; GARCIA, 2003). Existem, na literatura, diversos algoritmos para induzir árvores de decisão. Entre os mais populares estão os algoritmos ID3 (QUINLAN, 1986) e C4.5 (QUINLAN, 1993).

O ID3 foi um dos primeiros algoritmos de árvore de decisão. Baseia-se em sistemas de inferência e em conceitos de sistemas de aprendizagem (QUINLAN, 1986). Inicialmente, era utilizado para tarefas de aprendizagem, assim como um jogo de xadrez, no qual a estratégia é parte crucial. Depois disso, o algoritmo passou a ser aplicado em atividades industriais e acadêmicas (QUINLAN, 1986; GARCIA, 2003).

Desenvolvido para solucionar problemas com atributos categóricos, o ID3 necessita que os valores sejam tratados previamente, uma vez que os valores dos atributos não podem ter ruídos (QUINLAN, 1986). Tal algoritmo assume o ganho de informação para a escolha do atributo que será posto em cada nó; cada nó permite a divisão da base de dados de treino num número de subconjuntos igual ao seu número de elementos (QUINLAN, 1986; GARCIA, 2003).

O algoritmo C4.5, aprimoramento do algoritmo ID3, é proposto pelo mesmo autor, e passa a utilizar atributos categóricos e quantitativos, bem como valores ausentes, adotando o sistema de poda, melhorando o desempenho computacional (QUINLAN, 1993; GARCIA, 2003; PETERMANN, 2006). As variáveis categóricas podem ser divididas de duas formas: (1) um ramo diferente para cada valor; e (2) ramos diferentes para agrupamento de valores.

Para variáveis quantitativas, é utilizado o método de pesquisa exaustiva do ponto de divisão (QUINLAN, 1993). Percebe-se que, frequentemente, existem mais ramos nos primeiros nós da árvore em relação aos restantes, contribuindo para uma convergência mais acelerada da uma folha da árvore (GARCIA, 2003).

### **3.4.2 Regressão logística**

A regressão logística se encaixa na classe de métodos estatísticos

multivariados de dependência, uma vez que é capaz de relacionar um conjunto de variáveis independentes com uma variável dependente categórica (HAIR *et al.*, 2009). Essa técnica surgiu por volta de 1960 e ganhou destaque com o famoso estudo *Framingham Heart Study*, cujo principal objetivo era identificar fatores que ocorreram para desencadear doenças cardiovasculares (CORRAR *et al.*, 2007). Tal estudo foi responsável por identificar vários fatores de risco, tais como hipertensão arterial, diabetes, obesidade e vida sedentária (CORRAR *et al.*, 2007).

A variável dependente da regressão logística assume valor zero (fracasso) ou um (sucesso). Pode-se citar, por exemplo, um estudo que deseja prever se um aluno será ou não aprovado num exame, ou se um gestor terá êxito ou não numa negociação internacional (CORRAR *et al.*, 2007). Assim, percebe-se que mesmo que essa técnica tenha surgido e se desenvolvido na medicina, a sua aplicação não se restringiu apenas a essa área.

Hair *et al.* (2009) sustentam que um pesquisador não pode utilizar a regressão múltipla para prever valores de uma variável dicotômica, pois busca prever valores de uma variável codificada (e. g. 0 ou 1), obtendo valores estimados como sendo a probabilidade de obter um valor previsto de 1, por exemplo. Já na regressão múltipla, o ajuste de linha reta aos dados faz com que, frequentemente, valores inferiores a 0 ou superiores a 1 sejam previstos.

No caso dessa dissertação, a regressão logística pode prestar relevantes contribuições, uma vez que se deseja explicar por que um cliente escolhe romper com a empresa. Deste modo, a variável resposta do modelo de previsão de *churn* assumiria 1, se o cliente não romper com a empresa, e 0 no caso contrário.

Supõe-se  $x_1, \dots, x_k$  como as variáveis explicativas ou independentes, e  $P(evento)$  a proporção de clientes que “não romperam” com a empresa em função do perfil desse cliente, caracterizado por  $x$ . Como dito anteriormente, o modelo é capaz de definir uma relação entre a probabilidade de um cliente não romper com a empresa e um vetor de características e comportamentos do cliente  $(X_1, \dots, X_k)$ , sendo definido pela função *logit*, dada pela seguinte expressão:

$$\log \left\{ \frac{P(evento)}{1 - P(evento)} \right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$



$$P(\text{evento}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

Esta função é interpretada como a probabilidade de o cliente não romper com a empresa de acordo com as suas características (WEISBERG, 2005; LAROSE, 2006; CORRAR *et al.*, 2007; HAIR *et al.*, 2009; BOTELHO; TOSTES, 2010; MALHOTRA, 2012).

Entende-se que  $\beta_0$  é a “intercepção”, que representa o valor da expressão quando todas as variáveis independentes são zero e  $\beta_1, \dots, \beta_k$  são os coeficientes de regressão para cada  $x_1, \dots, x_k$ . Vale salientar que a análise pode ser feita assumindo 1 se o cliente romper com a empresa 0 o caso contrário, pois a interpretação é a mesma (LAROSE, 2006; CORRAR *et al.*, 2007; HAIR *et al.*, 2009; BOTELHO; TOSTES, 2010; MALHOTRA, 2012). O modelo logístico é mais popularmente utilizado porque não impõe às variáveis explicativas condições como a homogeneidade de variância, a normalidade na distribuição dos erros e a igualdade das matrizes de variância/covariância dentre os grupos (CORRAR *et al.*, 2007; HAIR *et al.*, 2009).

As regressões logísticas são divididas em dois subtipos: binomiais e multinomiais. Segundo Weisberg (2005), as regressões logísticas binomiais são aplicadas quando se tem variável resposta dicotômica (assume apenas duas categorias), ao passo que as regressões logísticas multinomiais são usadas quando a variável dependente é nominal, ou seja, pode assumir várias categorias e seguem uma ordem de sentido.

Os modelos de regressão procuram minimizar o número de variáveis para que o modelo final seja mais facilmente generalizado e numericamente estável (CHEN; DEY, 2003; HAIR *et al.*, 2009). Diante disso, inúmeros métodos para seleção das variáveis estão disponíveis na literatura, e, geralmente, eles usam o teste estatístico *t-test* para fazer a seleção da variável mais significativa. Contudo, outros testes estatísticos, tais como o *R-square* ( $R^2$ ), *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC) podem ser utilizados (CHEN; DEY, 2003; GOMES, 2011). Os métodos mais conhecidos são: adição *forward*, eliminação *backward*, e estimação *stepwise* (CHEN; DEY, 2003; CORRAR *et al.*, 2007; HAIR *et al.*, 2009).

O método de adição *forward* parte de um modelo bastante simples (sem variáveis selecionadas) e vai acrescentando variáveis independentes ao modelo, sem alternativa de eliminar as que já foram introduzidas. Caso nenhuma das variáveis possíveis implique uma melhoria mínima à soma dos quadrados dos resíduos (menor

valor), previamente definida, o processo termina e o modelo final é construído (CORRAR *et al.*, 2007; HAIR *et al.*, 2009).

O método de eliminação *backward*, ao contrário da adição *forward*, começa com o modelo mais complexo possível, ou seja, com todas as variáveis independentes. A cada passo, as variáveis que não contribuem significativamente com o poder preditivo do modelo vão sendo excluídas e não poderão voltar ao modelo. Quando não for possível retirar mais atributos sem que a precisão do modelo seja fortemente prejudicada, o processo acaba e o modelo final é construído (CORRAR *et al.*, 2007; HAIR *et al.*, 2009).

A estimação *stepwise*, utilizada neste trabalho, é um método que parte de um modelo com apenas uma variável selecionada (a variável com o maior coeficiente de correlação com a variável dependente). Em cada passo (*step*), todos os termos que ainda não foram incluídos no modelo são testados em relação à sua contribuição incremental (correlação parcial) para a precisão do modelo, e a que contribuir mais é adicionada. Como diferencial, se a contribuição das variáveis antigas (já estavam no modelo) ainda são significativas, cada nova variável inserida no modelo é examinada pelo teste F, dada a adição da nova variável. Caso não seja, a estimação *stepwise* permite que as variáveis que já estão no modelo sejam eliminadas (CORRAR *et al.*, 2007; HAIR *et al.*, 2009).

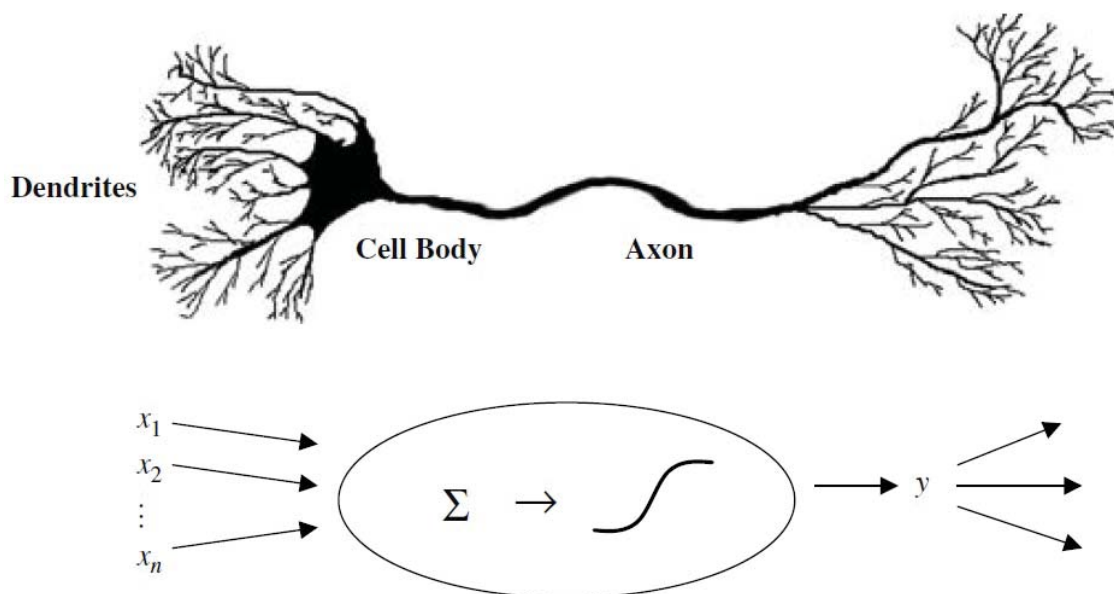
### 3.4.3 Redes neurais

Inspiradas no funcionamento e na estrutura das redes neurais biológicas e sendo modelos computacionais para processar informações, as redes neurais (RN) são vantajosas para a identificação de relações entre um agrupado de atributos ou padrões nos dados, ou seja, aprende através da experiência (MAIMON; ROKACH, 2010), característica responsável pela diferenciação das RN dos tradicionais programas computacionais, que somente seguem uma ordem de passos fixos (LEMOS, 2003). Para Lemos (2003), as RN artificiais buscam solucionar problemas por meio de um sistema de circuitos que simula o cérebro humano, inclusive aprendendo com os erros.

Um neurônio real usa dendritos (em inglês, *dendrites*) para receber as entradas de outros neurônios e combina a informação de entrada, gerando uma resposta que é enviada para outros neurônios usando o axônio (em inglês, *axon*), conforme Figura 2. A figura também apresenta um modelo de neurônio artificial presente em RN. As entradas ( $x_i$ , *inputs*) são coletadas a partir do conjunto de dados e combinadas através de uma função de combinação (e.g. somatório). Em seguida, são introduzidas numa função

de ativação, comumente não linear, para produzir uma resposta de saída ( $y$ ) (LAROSE, 2005). Os resultados (*output*) podem ser enviados para outro neurônio, servindo como *input* ou uma previsão do modelo (PETERMANN, 2006).

**Figura 2 - Neurônio real e modelo neurônio artificial**



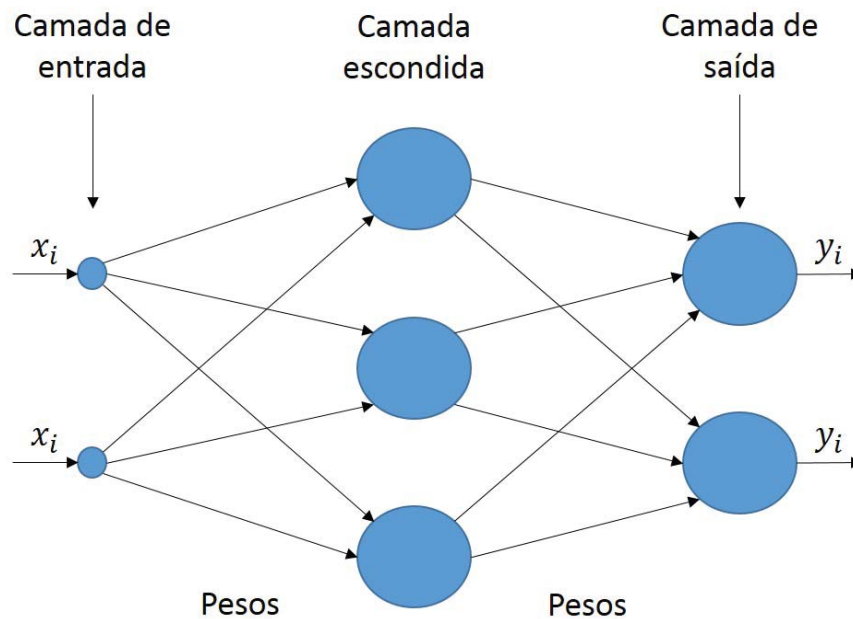
Fonte: Larose (2005).

Ressalta-se que a função de ativação é de suma importância para o comportamento de uma rede neural, haja vista que é a responsável por definir a saída do neurônio artificial e o caminho que a informação é conduzida (LEMOS, 2003; PETERMANN, 2006; MAIMON; ROKACH, 2010).

Observando a Figura 3, percebe-se que estes elementos estão organizados em camadas (*layers*) e conectados entre si por ligações com pesos associados (simulando as sinapses), sendo o peso responsável por determinar a influência de um elemento no outro. É por meio da modificação dessas conexões entre os *layers* que a RN aprende (LEMOS, 2003; LAROSE, 2005; MAIMON; ROKACH, 2010).

Segundo alguns autores (PETERMANN, 2006; MAIMON; ROKACH, 2010; GOMES, 2011), essa tecnologia é a que proporciona o maior poder de mineração de dados, mas, infelizmente, também é a mais difícil de se entender. Esses autores afirmam que as RN possuem um processo “caixa preta”, uma vez que os padrões achados e os modelos testados não são apresentados para o usuário. Por causa disso, muitos analistas de empresas não podem utilizá-las, pois os resultados finais não podem ser explicados.

**Figura 3 - Simples rede neural artificial**



Fonte: Adaptado de Larose (2005).

Petermann (2006) afirma que, geralmente, os pesquisadores trabalham somente com três camadas: camada de entrada, camada intermediária, ou escondida, e camada de saída. A primeira é onde os padrões são apresentados à rede, ou seja, a entrada dos dados. A segunda é o local onde a maior parte do processamento é realizado, através das conexões ponderadas. E, por fim, a última é onde o resultado final é finalizado e apresentado.

Analisando um pouco da história do campo das redes neurais, percebe-se que seu desenvolvimento se dá em períodos de entusiasmo e grandes progressos e em períodos de desconfiança e limitado progresso (MAIMON; ROKACH, 2010). Tratado como a origem desse campo, o estudo de McCulloch e Pitts (1943) constitui-se como base da visão moderna de redes neurais, fazendo a primeira tentativa de aplicar um modelo matemático para descrever o funcionamento do neurônio. A característica principal do seu modelo é que uma soma ponderada dos sinais de entrada é comparada com um limite para determinar a saída do neurônio. Os autores revelam que uma simples RN pode computar qualquer função (e.g. aritmética ou lógica).

Rosenblatt (1958) desenvolveu uma classe de redes neurais chamada de *perceptrons*, que são modelos de um neurônio biológico. O *perceptron*, com sua regra de

aprendizado associada, tinha gerado um grande interesse na pesquisa de rede neural. Todavia, Minsky e Papert (1969) publicaram o livro *Perceptrons*, no qual apontaram suas limitações e as grandes classes de problemas não linearmente separáveis existentes.

Mesmo os autores propondo uma RN de multicamadas com elementos escondidos, eles não foram capazes de treinar tal RN. Assim, afirmaram que talvez o problema de treinamento fosse insolúvel. Tal afirmação trouxe grande pessimismo para os pesquisadores da área. Com isso, durante a década de 1970 não houve praticamente nenhuma investigação (MAIMON; ROKACH, 2010).

Hopfield (1982) utilizou ferramentas estatísticas para provar que as redes neurais podem ser treinadas como uma memória associativa. Assim, as atividades de pesquisa dedicadas às redes neurais voltaram a acontecer. Conforme Maimon e Rokach (2010), dezenas de milhares de artigos foram publicados em revistas importantes e várias aplicações de sucesso têm sido relatadas nos últimos 15 anos.

Seguindo a ideia de que houve grande desenvolvimento em redes neurais, pode-se encontrar vários tipos de algoritmos para redes neurais. Assim como noticiam inúmeros autores (PETERMANN, 2006; MAIMON; ROKACH, 2010; GOMES, 2011), as redes neurais mais utilizadas para previsão são *multilayer perceptron* (MLP) e *radial basis function* (RBF), cujas implementações foram realizadas nesta dissertação.

#### 3.4.3.1 *Multilayer Perceptron* (MLP)

As redes neurais *Multilayer Perceptrons*, também conhecidas como *Multilayer Feedforward*, são os modelos de RN mais estudados e utilizados na prática (MAIMON; ROKACH, 2010). São organizadas em camadas, normalmente em três grupos (entrada, escondidas e saída), fazendo com que a informação siga em apenas um sentido: *entrada* → *escondidas* → *saída* (GOMES, 2011), conforme apresentado na Figura 3. Maimon e Rokach (2010) afirmam que embora seja possível ter mais de uma camada escondida dentro de uma MLP, a maioria das aplicações utiliza somente uma. Ademais, observa-se que neste tipo de RN a informação nunca volta para camadas anteriores.

Para casos com um grande número de relações complexas, o MLP necessita de um elevado tempo de treino e de teste, a fim de que sua precisão seja considerada satisfatória (MAIMON; ROKACH, 2010). Tal tempo se justifica porque os elementos

distribuídos pelas diversas camadas encontram-se completamente interligados, logo, os elementos da camada de entrada estão conectados aos da primeira camada escondida, e assim por diante, até a última camada escondida. Em seguida, todos os elementos da última camada escondida estão ligados aos elementos da camada de saída (MAIMON; ROKACH, 2010; GOMES, 2011).

Durante um processo de aprendizagem (o treino), cada registro é analisado individualmente, e à medida que a informação passa pelas camadas, é gerada uma previsão desse registro, baseada na variável resposta. Os valores dos pesos associados também são ajustados nesse momento, mudando o seu valor a cada previsão errada (MAIMON; ROKACH, 2010; GOMES, 2011). O treino se repete várias vezes, fazendo com que a capacidade de previsão do modelo melhore a cada vez que treine. Este processo acaba quando alguma determinação de paragem tenha sido alcançada (GOMES, 2011).

Quando se inicia a construção de um modelo MLP, os pesos associados são atribuídos de forma aleatória, tomando valores com mais sentido depois de um treino inicial (MAIMON; ROKACH, 2010; GOMES, 2011). Em seguida, após um treino inicial, o resultado previsto pela RN é confrontado com os valores reais do caso de treino, e o resultado desta comparação é inserido a uma nova RN, com o intuito de atualizar os pesos das ligações que estiveram envolvidas na previsão (GOMES, 2011).

#### 3.4.3.2 *Radial Basis Function* (RBF)

A Rede de Função de Base Radial (*Radial Basis Function* – RBF) possui esse nome devido à utilização, pelos neurônios da camada escondida, de funções de base radial. É uma RN inspirada pelas respostas que podem ser sintonizadas localmente por alguns neurônios biológicos. Estas células respondem a características selecionadas de algumas regiões finitas do espaço dos sinais de entrada (PETERMANN, 2006; BRAGA; CARVALHO; LUDERMIR, 2011).

A RBF possui uma estrutura similar a MLP, porém utiliza uma função de transferência linear para as unidades de saída e uma função de transferência não linear para as unidades ocultas. Além de um processo de aprendizado mais rápido (em relação ao MLP), viabiliza a utilização da RBF como um bom classificador de padrões (PETERMANN, 2006; BRAGA; CARVALHO; LUDERMIR, 2011). Uma rede RBF pode ser exposta sendo uma composição global de aproximações para funções, utilizando combinações de funções de base centralizadas em vetores de pesos (BRAGA;

CARVALHO; LUDERMIR, 2011).

A arquitetura típica pode ser visualizada na Figura 3. O que se altera é apenas que cada elemento da camada escondida utiliza funções de base radial. Neste modelo, cada neurônio define uma hiperelipsóide no espaço dos padrões de entrada, construindo estimadores locais. Esta camada transforma um conjunto de padrões de entrada não linearmente separáveis em um conjunto de saídas linearmente separáveis (BRAGA; CARVALHO; LUDERMIR, 2011).

As funções de base radiais geram uma saída significativa apenas quando o padrão de entrada está incorporado a uma região situada no espaço de entrada. Assim, cada função radial precisa de um centro e um parâmetro escalar (PETERMANN, 2006; BRAGA; CARVALHO; LUDERMIR, 2011). Apesar do exemplo apresentado na Figura 3 possuir apenas dois neurônios na camada de saída, esse número pode ser aumentado ou diminuído para um, dependendo do problema tratado (MAIMON; ROKACH, 2010; GOMES, 2011; BRAGA; CARVALHO; LUDERMIR, 2011).

Segundo Braga, Carvalho e Ludermir (2011), as funções radiais representam um conjunto especial de funções que aumentam ou diminuem em relação à distância de um ponto central. Existem diferentes funções para redes RBF, contudo. Segundo os autores, as mais comuns são: Gaussiana, Multiquadrática, *Thin-plate-spline* e Multiquadrática Inversa. Neste trabalho, utiliza-se a função Gaussiana. Para cada função é preciso entrar com a distância utilizada (e.g. euclidiana) e a largura da função radial como parâmetros (PETERMANN, 2006; BRAGA; CARVALHO; LUDERMIR, 2011).

Existem diferentes caminhos de se seguir quando se trata de estratégia de aprendizado em uma rede RBF, uma vez que há várias maneiras de o pesquisador escolher a maneira de definir os centros das funções de base radiais. Dentre as mais utilizadas, observa-se os centros fixos selecionados ao acaso, a seleção auto-organizada de centros e a seleção supervisionada de centros (BRAGA; CARVALHO; LUDERMIR, 2011).

Assim como ocorre com a rede MLP, também é preciso definir o número de elementos da camada escondida da rede RBF. Essa definição deve acontecer de forma criteriosa, haja vista que cada neurônio agrupa um conjunto de padrões que é aproveitado pelos neurônios da camada de saída (BRAGA; CARVALHO; LUDERMIR, 2011; GOMES, 2011; LAROSE, 2005; LEMOS, 2003; MAIMON; ROKACH, 2010; PETERMANN, 2006).

## 4 GERENCIAMENTO DE CHURN EM PLANOS DE SAÚDE

### 4.1 O MERCADO DE PLANO DE SAÚDE

No Brasil, os planos privados de assistência à saúde operaram desde a década de 1960 até 1998 sem regulação econômica. Assim, sua história divide-se em dois momentos distintos: antes e depois da Lei nº 9.656. Publicada no dia 3 de junho de 1998 e alterada em 24 de agosto de 2001 pela Medida Provisória nº 2.177-44, essa Lei criou novas categorias de planos de saúde. A diferenciação dos planos passou a ser determinada por características como a cobertura assistencial e a abrangência geográfica. Ademais, inúmeros procedimentos que não eram cobertos pelos planos passaram a ser obrigatórios (e.g. consultas sem limite e transplantes de órgãos).

Após a homologação da Lei, as operadoras de planos de saúde privados viram-se na obrigação de realizar uma reestruturação de seus planos. Também se deu início a um trabalho de persuasão dos seus assinantes para mudança de planos, uma vez que não era vantajosa a permanência dos clientes em planos no formato antigo (antes da lei), todavia os beneficiários não eram obrigados a migrar de plano, especialmente pelo acréscimo no valor da mensalidade (MENDES, 2008).

Criada pela Lei nº 9.961, de 28 de janeiro de 2000, a Agência Nacional de Saúde Suplementar (ANS), atual agência reguladora do setor de planos de saúde, originou-se em um setor específico do Ministério da Saúde e ficou responsável por garantir o cumprimento da Lei nº 9.656. Antes dessa lei, no âmbito dos planos de saúde, as seguradoras eram reguladas pela Superintendência de Seguros Privados (SUSEP). Atualmente, o setor brasileiro de planos e seguros de saúde é um dos maiores sistemas privados de saúde do mundo (BRASIL, 2016).

Os contratos de planos e seguros de saúde pós-regulamentação passaram a ser registrados, de forma obrigatória, na ANS, já em harmonia com as normas estabelecidas. A ANS categoriza os planos de saúde regulamentados segundo quatro dimensões que retratam as peculiaridades do produto comercializado. São elas:

- a) segmentação assistencial: a segmentação do plano advém da combinação da cobertura assistencial (ambulatorial, hospitalar, obstetrícia e odontológica) do plano de saúde. É de oferta obrigatória o plano de referência que contém o modelo mínimo de cobertura a ser ofertado pelas prestadoras;
- b) época de contratação do plano: planos novos (após a vigência da Lei no



9.656/98) e planos antigos (antes da vigência da Lei no 9.656/98);

c) tipo de contratação do plano: individual ou familiar, coletivo empresarial e coletivo por adesão;

d) abrangência geográfica: a abrangência do plano contratado pode ser municipal, conjunto de municípios, estadual, conjunto de estados ou nacional.

Graças a Resolução Normativa nº 186 de 2009 (alterada pela nº 252 de 2011), existe a oportunidade de portabilidade de carência, ou seja, a possibilidade de mudança de plano de saúde levando consigo os períodos de carência já realizados (BRASIL, 2016).

Segundo a legislação vigente, para ter direito à mudança de plano com a manutenção da carência, é necessário verificar os pré-requisitos presentes (e.g. o plano deve ter mais de dois anos, pagar em dia a mensalidade, ter plano individual/familiar ou coletivo por adesão etc).

Em 2015, segundo Brasil (2016), 1.340 operadoras apresentavam registro ativo no Brasil. As operadoras se diferenciam na forma de acesso, nos benefícios ofertados e no sistema de pagamento. As modalidades são: autogestão, cooperativa médica, filantropia, medicina de grupo, seguradora especializada em saúde, cooperativa odontológica, odontologia de grupo, administradora e administradora de benefícios.

Dados disponibilizados por Brasil (2016) mostram que o setor de saúde suplementar brasileiro atendeu, em 2015, 71.680.868 beneficiários, dentre os quais 21.950.463 são de planos exclusivamente odontológicos e 49.730.405 de planos de assistência médica. Comparando os anos de 2014 e 2015, observa-se que houve queda de 1,5% na quantidade de beneficiários em planos de assistência médica, cerca de 766 mil pessoas.

Delimitando os dados para o Nordeste, tem-se um total de 11.059.422 beneficiários. A região representa cerca de 15,43% de todos os beneficiários atendidos em 2015, dentre os quais 9,45% (6.777.298) são de assistência médica e 5,98% (4.282.124) exclusivamente odontológicos. Afunilando ainda mais a pesquisa, tem-se que o Ceará possui 2.042.903 beneficiários, dos quais 61,57% (1.257.825) são de assistência médica e 38,43% (785.078) exclusivamente odontológicos. O estado representa 18,47% dos beneficiários do Nordeste e 2,85% dos beneficiários do Brasil (BRASIL, 2016).

Conforme apresentado por Brasil (2016), o setor movimentou, em 2015, R\$ 117,3 bilhões até o terceiro trimestre, somadas as receitas de contraprestações e outras receitas operacionais das operadoras de planos assistência médica e odontológicos,

enquanto que despesas totais foram de R\$ 117,2 bilhões, gerando um resultado operacional de R\$ 80 milhões no período. Verificou-se que os planos de assistência médica representam, aproximadamente, 98,5% de todas a receita e despesas do setor, mostrando a importância de se estudar tal área. Deste modo, esse estudo se propõe a investigar somente os assinantes que possuem plano de assistência médica estabelecidos a partir do ano de 2000, por causa da regulamentação.

Alves (2008) atribui o aumento dos custos com prestação de serviços médicos à força da expressiva inovação tecnológica na área médica e ao aumento da demanda por serviços de saúde. O autor afirma que lidar com as demandas crescentes da população e os recursos limitados da sociedade exige que as organizações de saúde, públicas e privadas, busquem melhorar a qualidade e a eficiência de forma constante. Além disso, o elevado número de empresas e as exigências do mercado consumidor também contribuem para que as organizações busquem alcançar vantagens competitivas em relação aos seus concorrentes, objetivando melhores resultados (PORTER, 1989).

Para Oliver *et al.* (1994), o sucesso das operadoras de planos de saúde está ligado diretamente à sua competência para tomada de decisões, mas não exclusivamente a isso. Um Sistema de Informações Gerenciais é capaz de reunir a percepção da qualidade do serviço por parte do cliente. Além disso, desempenho financeiro e resultados clínicos seriam necessários para definir caminhos a seguir.

Piva *et al.* (2007) afirmam que analisar a relação de um beneficiário com a empresa vai além de suas transações individuais. Cada particularidade do relacionamento do consumidor deve ser levada em conta quando se quer avaliar seu relacionamento com a empresa (e.g. questões emocionais e percepção da marca), haja vista que cada ponto de interação pode ter importância diferente para todos os clientes.

## 4.2 RETENÇÃO E IMPLICAÇÕES EM PLANOS DE SAÚDE

É de conhecimento geral no meio empresarial que a satisfação de clientes torna-os mais leais aos produtos e marcas (e.g. RUST; ZAHORIK, 1993; RUST; ZEITHAML; LEMON, 2004; REICHHELD, 2006), fato este que leva diversas organizações a questionamentos sobre a adesão de programas para aumento da qualidade e implantação de planos de fidelização de clientes. Tal dilema torna-se complexo, uma vez que os empresários possuem, frequentemente, poucas bases para tomada de decisão,

além de sua própria prática e intuição (RUST; ZEITHAML; LEMON, 2004). O interesse na relação entre satisfação e retenção fomentou uma sequência de artigos e publicações específicas em torno do tema, quando evidenciado o ponto de necessidade de mudanças estratégicas pautadas no produto, tendo como base o cliente e o melhor atendimento de suas necessidades.

Em meio a ambientes gradativamente mais dinâmicos e competitivos, as instituições empresariais são obrigadas a levar vigorosamente em consideração um terceiro elemento, o qual está intimamente ligado à relação recém-observada: a rentabilidade. Tal ponto se justifica, pois não há sentido em investimentos comerciais se não houver clientes, e principalmente, a ideia de estes serem rentáveis. Afinal, trata-se aqui de satisfação para reter clientes, gerando relações de longo prazo, com elevação na rentabilidade (PIVA *et al.*, 2007).

Integrando a teoria à prática em planos de saúde, a satisfação do beneficiário está ligada à classificação de uma ou mais variedades de opiniões de outros beneficiários, que podem incluir a avaliação da qualidade dos serviços prestados, as intenções comportamentais futuras, o entendimento do beneficiário sobre os próprios resultados clínicos observados e a visão de satisfação geral com a organização de saúde (OLIVER *et al.*, 1994). Desta maneira, a satisfação passa a ser um relevante indicador da qualidade dos serviços em saúde, e está associada ao resultado financeiro em hospitais.

Agregando o desempenho financeiro com a qualidade percebida dos serviços prestados e os resultados clínicos, obtém-se o conhecimento primordial para se tornar um provedor de serviços de saúde de baixo custo e com alta qualidade. No entanto, não é possível redução de custos se não existir contribuição para o beneficiário e sua saúde (PORTER, 2006).

No caso de prestadoras de serviço de saúde privada, a rentabilidade é mensurada através do Índice de Sinistralidade do Seguro, preferencialmente o menor possível; portanto, quanto menor for a utilização dos serviços da prestadora, mais o contrato será rentável (PIVA *et al.*, 2007).

Para as operadoras de planos de saúde, é interessante perceber que não há vinculação direta entre o aumento da utilização dos serviços e a lealdade. Outro fator a ressaltar para as prestadoras é que a percepção dessa qualidade ainda é advinda da utilização dos serviços, e não de sua melhoria de condições de saúde. Neste intuito, a busca pela manutenção do cliente vem por meio do aumento da sua percepção de valor sobre o atendimento prestado, ligado à melhoria de seus indicadores de saúde, os quais,

consequentemente, alcançariam a objetivo da seguradora de diminuir no cliente a imprescindibilidade de uso das funções vinculadas a exames, consultas e procedimentos cirúrgicos (PIVA *et al.*, 2007).

Para gerar melhorias nos indicadores de saúde, Piva *et al.* (2007) afirmam que são necessários investimentos na qualidade das atividades prestadas, atuando nas relações a longo prazo, ou seja, na retenção. A melhoria no tocante à saúde não se dá em um curto período de tempo, dependendo diretamente das atividades da medicina preventiva, medicina curativa e das ações de promoção e preservação de saúde. Tudo isso dentro do aspecto subjetivo do processo saúde-doença, no que diz respeito à questão bem-estar (social, emocional e físico).

Propositando promover a real rentabilidade e criar valores para o cliente, a estratégia será desenvolver propostas a partir de dois pontos iniciais: focar nos clientes certos, subdividindo de forma adequada a base de clientes, e ser capaz de proporcionar uma experiência única que encante cada segmento criado. Tais planos podem ser meticulosamente realizados do início ao fim, e a empresa deve, por sua vez, prosperar na capacidade de repetir estas atitudes por várias vezes, sempre reciclando e recriando a experiência do cliente (PIVA *et al.*, 2007).

Inúmeras pesquisas já buscaram entender o setor de planos privados de saúde. Pinto e Soranz (2004) utilizaram o cadastro de beneficiários da ANS combinado com a Pesquisa Nacional por Amostra de Domicílios (PNAD/IBGE) para descrever o perfil da cobertura dos serviços oferecidos por planos de saúde privados. Tal estudo mostra que as capitais brasileiras são grandes centros de concentração de beneficiários para as operadoras de planos de saúde, e quem mais utiliza esse tipo de serviço são crianças menores de cinco anos, mulheres na idade fértil e idosos. Ademais, importa ressaltar que 70% da população coberta por planos de saúde está concentrada na região Sudeste.

Os autores (PINTO; SORANZ, 2004) concluem a pesquisa afirmando que os planos de saúde cobrem apenas uma parcela pequena da população, limitando-se a pessoas com maior grau de escolaridade, maior renda familiar, moradores das capitais e regiões metropolitanas e de cor branca. Campos (2006) apresentou um estudo sobre as relações entre operadoras de planos de saúde e prestadores de serviços hospitalares em um ambiente contratual, tentando identificar fatores que possam influenciar essa relação.

Em sua pesquisa, Campos (2006) entrevistou diretores e executivos de três operadoras de planos de saúde e três hospitais privados, e constatou que as operadoras de planos de saúde credenciam a sua rede de serviços hospitalares fundamentadas no

tamanho de rede, definições técnicas para a hierarquização desta e o volume de clientes. Por outro lado, os hospitais habilitam as operadoras para as suas organizações de acordo com a divulgação de imagem e os preços competitivos.

A qualidade da assistência dos prestadores de serviços é avaliada pelas operadoras, porém elas não possuem programas formais para tal. Além disso, monitoram o cumprimento das cláusulas contratuais celebradas entre as partes (CAMPOS, 2006). Ainda no sentido de investigar mais a fundo empresas prestadoras de serviços em saúde, Piva *et al.* (2007) fizeram uma análise com dados extraídos de uma pesquisa de opinião aplicada a clientes de uma empresa do setor sobre a existência de evidências que possam indicar as relações entre: satisfação, retenção de clientes e a rentabilidade da empresa prestadora.

O estudo de Piva *et al.* (2007) utilizou as técnicas de análise fatorial e de regressão linear como metodologia de análise de dados, e os seus indicadores foram tratados com o objetivo de formular um modelo teórico adaptado ao comportamento do consumidor de serviços de saúde. Os resultados obtidos apontaram que a satisfação e retenção impactam na rentabilidade.

Outro estudo que busca entender melhor as organizações de prestação de serviços médicos é o estudo de Pastrana (2012), que investigou como ocorreu a estruturação de duas unidades de saúde suplementar, comparando-as quanto à existência de modelos de benefícios motivacionais, que seriam o diferencial competitivo e de eficiência produtiva.

A autora (PESTRANA, 2012) verifica que as disposições das gestões são divergentes em suas complexidades. Por um lado, a empresa “A” possui diversos benefícios motivacionais interligados a propostas direcionadas para cargos em que se observam maiores riscos de ruptura da conexão com o profissional competente. Por outro lado, a empresa “B” depara-se com um espaço de transição entre o discurso de competência dos empregados e a mera remuneração em equivalência pelo trabalho prestado.

Abordando de forma mais quantitativa o mercado de planos de saúde, Leal e Matos (2009) analisaram a evolução dos custos de assistência médica dos planos de saúde brasileiros. Isso foi feito tomando como base a distinção marcante entre as categorias de contratação: individual ou coletivo. Segundo Brasil (2016), contrato individual é aquele oferecido para a livre adesão do consumidor pessoa física, enquanto que o contrato coletivo é aquele firmado por uma pessoa jurídica, cujas pessoas na condição de

empregado, associado ou sindicalizado podem se tornar beneficiários a partir de duas modalidades: contrato coletivo por adesão e plano coletivo empresarial (com e sem patrocínio).

Quando a adesão dessas pessoas ao plano é espontânea, o contrato é coletivo por adesão, ao passo que quando a adesão é automática, consequente de vínculo, e abrange todos ou a maioria das pessoas vinculadas, o contrato é coletivo empresarial. Se o contrato for com patrocinador, significa que a pessoa jurídica assinante do contrato com a operadora é responsável pelo pagamento parcial ou integral do plano. Por outro lado, se o contrato for sem patrocinador, a pessoa jurídica assinante do contrato com a operadora não é responsável pelo pagamento do plano, ou seja, o beneficiário deve fazê-lo diretamente com a operadora do plano. (BRASIL, 2016).

Leal e Matos (2009) fizeram uso de informações divulgadas pela ANS, estimando índices de variação baseados na metodologia de índices de valor e de dois de seus componentes: o preço e a quantidade. O estudo encontra que o índice de preços (custo médio por evento) representa o aumento dos custos unitários, e pode ser determinado principalmente pela inflação dos insumos e pela incorporação tecnológica. O índice de quantidade (frequência de utilização) representa o incremento de utilização pelos beneficiários, que pode ter relação com algumas categorias de fatores: (1) sociais; (2) demográficos e (3) perfil epidemiológico.

Com o intuito de unir o assunto retenção de cliente, *churn* e planos privados de assistência à saúde, Mendes (2008) apresentou um estudo que teve o objetivo de desenvolver um modelo estatístico que relacionasse variáveis transacionais, demográficas e dados sobre o histórico do cliente com a probabilidade de cancelamento dos beneficiários de um plano de saúde. Esse estudo teve um banco de dados de 130.552 registros e utilizou apenas a regressão logística para a realização da previsão e, em seguida, com o objetivo de definir perfis estratégicos de clientes, utilizou-se PSM (*Propensity Score Matching*).

Mendes (2008) explica que o PSM tem como principal objetivo descobrir o efeito médio do tratamento, ou seja, a técnica mostra o que aconteceria se o grupo de tratamento não recebesse tratamento e se o grupo controle tivesse recebido o tratamento.

O autor (MENDES, 2008) também utilizou a estatística de Wald para identificar a importância de cada variável no modelo proposto, fazendo com que se identificasse quais variáveis mais impactavam no modelo. Tais variáveis, da mais importante à menos importante, são: tempo de exame, segmento de utilização, atraso,

tempo de consulta, ano de inclusão, valor pago à empresa no penúltimo mês, faixa etária, área de rendimento, valor pago à empresa no último mês, rede de produto, opcional, gênero, estado civil.

Utilizando as variáveis citadas, Mendes (2008) aplicou uma regressão logística aos dados, e obteve 87,7% de taxa de acerto geral do modelo. Mais detalhadamente, o grupo que rompeu com a empresa teve uma taxa de acerto de 84,4%, enquanto o grupo que não rompeu com a empresa teve taxa de acerto de 91,1%.

Interpretando os resultados obtidos pela regressão logística no trabalho de Mendes (2008), o perfil do beneficiário com maior risco de cancelar o contrato com a operadora do plano de saúde é aquele com mais tempo entre exames e consultas; paga mensalidades mais baratas; possui uma rede de produtos inferior; paga em atraso; tem menos opcionais; idade baixa; e que utiliza pouco o plano.

## 5 PROPOSTA DE VARIÁVEIS PARA COMPOSIÇÃO DE BASE PARA PREVISÃO DE RUPTURA EM PLANO DE SAÚDE

Após uma vasta pesquisa, verifica-se que inúmeras variáveis foram utilizadas para tentar prever o *churn* em diferentes setores (e.g. telecomunicação e venda online). Dos artigos apresentados no Quadro 1, variáveis foram elencadas e agrupadas em categorias que mais às representavam. Deste modo, foram criadas 8 categorias. São elas: Demográficas, Gasto, Reclamações, Contagem, Produto, Serviço, Adicionais e Oferta. Cada categoria possui inúmeras variáveis; cada atributo pode ser sugerido de forma idêntica ao que foi utilizado na pesquisa original ou pode ser adaptado para que faça sentido em um plano de saúde.

Diante disso, este capítulo visa atingir o objetivo específico proposto por esta dissertação, que é apresentar variáveis adequadas à composição de identificação à propensão de ruptura de clientes em planos de saúde. Sabe-se que raramente uma empresa terá todas as variáveis aqui citadas, porém julga-se importante mostrar variáveis já testadas em outros estudos e setores sobre *churn* que possam ser utilizadas para o setor de plano de saúde, com algumas modificações ou não.

### 5.1 VARIÁVEIS DEMOGRÁFICAS

O Quadro 3 apresenta as variáveis que fazem referência às informações demográficas dos clientes. Idade, estado civil, sexo, escolaridade, renda familiar e localização do cliente são variáveis comuns em um cadastro de cliente. Explicando melhor a variável localização, pode-se entender como qualquer codificação do local onde o cliente reside (e.g. rua, bairro, cidade, CEP).

Contabiliza-se que idade, estado civil, sexo e localização foram as variáveis mais utilizadas nos artigos pesquisados. Todavia, existem outras variáveis utilizadas nos artigos que também são importantes. Acredita-se que saber o tipo de residência que o cliente mora e informações sobre o emprego dele pode melhorar a classificação do perfil do cliente, bem como o número de membros do agregado familiar.

Mendes (2008) afirma que quanto maior for a idade, maior é a chance de haver clientes ativos, ou seja, que não rompem com o plano de saúde. Para o autor, isso se explica, principalmente, pelas necessidades inerentes à idade. Clientes idosos recorrem



mais ao plano de saúde do que clientes mais jovens (MENDES, 2008).

Verifica-se que, nesta categoria, todas as variáveis propostas são iguais às variáveis base, ou seja, iguais aos atributos encontrados nos artigos.

## 5.2 VARIÁVEIS DE GASTOS

Observando a categoria gastos apresentada no Quadro 4, tem-se a sugestão de quatro variáveis, que são descritas a seguir.

Percebe-se a concentração de autores em duas variáveis. A primeira, “despesa mensal do serviço/produto?”, é citada em sete artigos (AU; CHAN; YAO, 2003; LEWIS, 2004; BUCKNIX; VAN DEN POEL, 2005; BOTELHO; TOSTES, 2010; ABBASIMEHR; SETAK; SOROOR, 2013; LEMMENS; CROUX, 2006; YU *et al.*, 2011), e entende-se que tal variável pode ser utilizada em um banco de dados de planos de saúde sem qualquer modificação.

Por outro lado, a terceira, “valor gasto em chamadas (acumulado total, de dia, de tarde, de noite, internacional)”, é citada em oito artigos (LEWIS, 2004; BUCKNIX; VAN DEN POEL, 2005; PHADKE *et al.*, 2013; LEMMENS; CROUX, 2006; COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; VAN DEN POEL, 2008a; ABBASIMEHR; SETAK; TAROKH, 2014; VERBEKE *et al.*, 2011), e precisa ser adaptada para sua aplicação em planos de saúde, pois é referente à operadora de telefonia. Desta maneira, sugeriu-se a variável “valor total gasto (acumulado total, de dia, de tarde, de noite, internacional)”, referente a valores gastos em serviços do plano de saúde por períodos ou fora do país.

Ressalta-se que, para ser possível obter as variáveis acima citadas, é preciso ter cada valor gasto registrado com identificador de tempo (dia, tarde, noite e período) e de local. Sem a devida identificação, tais variáveis se tornam impossíveis de serem calculadas.

Uma variável utilizada no trabalho de Abbasimehr, Setak e Soroor (2013) é a *Customer Lifetime Value* (CLV). Kotler (1974) definiu CLV como o valor presente dos fluxos de caixa descontados que a empresa espera obter do cliente no tempo. Portanto, tal variável ajudaria a identificar os clientes que têm expectativa de maior e menor lucro, podendo ser uma variável importante para previsão de ruptura em planos de saúde.

**Quadro 3 - Variáveis demográficas propostas para previsão de ruptura em um banco de dados de plano de saúde**

Categoria	Variável proposta para Plano de Saúde	Variável base	Autores
Demográficas	Igual à variável base	Idade do cliente	KUMAR; RAVI, 2008; XIE <i>et al.</i> , 2009; BOTELHO; TOSTES, 2010; NIE <i>et al.</i> , 2011; LEMMENS; CROUX, 2006; YU <i>et al.</i> , 2011; FIGUEIREDO; SILVERMAN, 2007; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; COUSSEMENT; VAN DEN POEL, 2008a; MENDES, 2008;
	Igual à variável base	Número de membros do agregado familiar	BUCKNIX; VAN DEN POEL, 2005; XIE <i>et al.</i> , 2009; LEMMENS; CROUX, 2006;
	Igual à variável base	Estado civil	KUMAR; RAVI, 2008; XIE <i>et al.</i> , 2009; BOTELHO; TOSTES, 2010; NIE <i>et al.</i> , 2011; LEMMENS; CROUX, 2006;
	Igual à variável base	Gênero do cliente	KUMAR; RAVI, 2008; BOTELHO; TOSTES, 2010; NIE <i>et al.</i> , 2011; YU <i>et al.</i> , 2011; COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; VAN DEN POEL, 2008a;
	Igual à variável base	Escolaridade do cliente	KUMAR; RAVI, 2008; XIE <i>et al.</i> , 2009; BOTELHO; TOSTES, 2010; NIE <i>et al.</i> , 2011;
	Igual à variável base	Renda familiar do cliente	KUMAR; RAVI, 2008; XIE <i>et al.</i> , 2009; LEMMENS; CROUX, 2006;
	Igual à variável base	Localização do cliente	AU; CHAN; YAO, 2003; BUCKNIX; VAN DEN POEL, 2005; BOTELHO; TOSTES, 2010; ABBASIMEHR; SETAK; TAROKH, 2014; LEMMENS; CROUX, 2006; VERBEKE <i>et al.</i> , 2011; YU <i>et al.</i> , 2011;
	Igual à variável base	Tipo de residência (própria ou alugada)	BOTELHO; TOSTES, 2010; LEMMENS; CROUX, 2006;
	Igual à variável base	Empregado? (sim ou não)	XIE <i>et al.</i> , 2009; NIE <i>et al.</i> , 2011; YU <i>et al.</i> , 2011;
	Igual à variável base	Tempo de emprego do cliente? (em anos)	BOTELHO; TOSTES, 2010;

Fonte: Autoria própria.

A variável “forma de pagamento” refere-se à maneira que o cliente efetua o pagamento da fatura do plano de saúde. Observando o uso desta variável por Wei e Chiu (2002) e por Au, Chan e Yao (2003), percebe-se que a forma de pagamento pode discriminar um beneficiário que possivelmente cancelará o seu contrato com o plano de saúde.

**Quadro 4 - Variáveis relacionadas a gastos propostas para previsão de ruptura em um banco de dados de plano de saúde**

<b>Categoria</b>	<b>Variável proposta para Plano de Saúde</b>	<b>Variável base</b>	<b>Autores</b>
Gastos	Igual à variável base	Despesa mensal do serviço/produto? (Mensalidade atual)	AU; CHAN; YAO, 2003; LEWIS, 2004; BUCKNIX; VAN DEN POEL, 2005; BOTELHO; TOSTES, 2010; ABBASIMEHR; SETAK; SOROOR, 2013; LEMMENS; CROUX, 2006; YU <i>et al.</i> , 2011;
	Igual à variável base	<i>Customer Lifetime Value</i> (CLV)	ABBASIMEHR; SETAK; SOROOR, 2013;
	Valor total gasto (acumulado total, de dia, de tarde, de noite, internacional)	Valor gasto em chamadas (acumulado total, de dia, de tarde, de noite, internacional)	LEWIS, 2004; BUCKNIX; VAN DEN POEL, 2005; PHADKE <i>et al.</i> , 2013; LEMMENS; CROUX, 2006; COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; VAN DEN POEL, 2008a; ABBASIMEHR; SETAK; TAROKH, 2014; VERBEKE <i>et al.</i> , 2011;
	Forma de pagamento	Tipo de pagamento	WEI; CHIU, 2002; AU; CHAN; YAO, 2003;

Fonte: Autoria própria.

### 5.3 VARIÁVEIS DE RECLAMAÇÕES

Segundo a literatura já levantada (e.g. LEMMENS; CROUX, 2006; COUSSEMENT; VAN DEN POEL, 2008a; COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; VERBEKE *et al.*, 2011; ABBASIMEHR; SETAK; TAROKH, 2014), informações sobre reclamações do cliente são de grande importância para previsão de ruptura do cliente com a empresa. Dessa forma, o Quadro 5 apresenta quatro variáveis que foram utilizadas em artigos e podem ser aplicadas em um banco de dados de plano de saúde.

Coussement e Van den Poel (2008a, 2008b) são os autores que mais procuram utilizar variáveis que façam referência às reclamações dos clientes. O número de reclamações e o tempo decorrido desde a última reclamação, segundo os autores, estão

diretamente ligados ao cancelamento de contrato pelo cliente. Desta forma, tais variáveis poderiam ser utilizadas sem qualquer alteração.

Associado ao número de reclamações, o número de chamadas realizadas ao Serviço de Atendimento ao Consumidor (SAC) também é uma variável importante que discrimina os clientes que possivelmente irão cancelar seus contratos (ABBASIMEHR; SETAK; TAROKH, 2014; LEMMENS; CROUX, 2006; VERBEKE *et al.*, 2011). Esta variável serve, além de compor uma base para previsão de ruptura em planos de saúde, para medir a qualidade do serviço oferecido pela empresa.

Coussement e Van den Poel (2008a), por meio da variável “custo médio de uma reclamação”, apresentaram uma maneira de associar um valor monetário à reclamação. Deste modo, torna-se público à empresa o quanto custa a reclamação de cada cliente. Isto posto, entende-se que essa variável deva compor a base proposta.

**Quadro 5 - Variáveis relacionadas a reclamações propostas para previsão de ruptura em um banco de dados de plano de saúde**

<b>Categoria</b>	<b>Variável proposta para Plano de Saúde</b>	<b>Variável base</b>	<b>Autores</b>
Reclamações	Igual à variável base	<b>Número de chamadas realizadas pelo cliente ao SAC</b>	ABBASIMEHR; SETAK; TAROKH, 2014; LEMMENS; CROUX, 2006; VERBEKE <i>et al.</i> , 2011;
	Igual à variável base	<b>Custo médio de uma reclamação (em termos monetários)</b>	COUSSEMENT; VAN DEN POEL, 2008a;
	Igual à variável base	<b>O tempo decorrido desde a última reclamação (dias)</b>	COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; VAN DEN POEL, 2008a; COUSSEMENT; BENOIT; VAN DEN POEL, 2010;
	Igual à variável base	<b>Número de reclamações</b>	COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; COUSSEMENT; VAN DEN POEL, 2008a;

Fonte: Autoria própria.

#### 5.4 VARIÁVEIS DE CONTAGEM

Esta categoria contempla atributos que apresentam a frequência em relação a alguma atividade, totalizando dez variáveis. Observa-se que sete atributos precisaram ser adaptados de outras áreas, por exemplo, o “número de cartões de crédito adicionais” deve se transformar em “número de adicionais que possui”, tornando-se útil para a técnica de

*data mining* em um banco de dados de uma operadora de plano de saúde.

O número de adicionais indica a quantidade de serviços extras que o cliente contratou no plano original (e.g. serviços odontológicos). Ressalta-se que esta variável apenas conta a quantidade de adicionais, mas não informa quais são. Para isso, a seção 4.8 apresenta variáveis para contemplar as informações sobre adicionais do beneficiário.

Com o intuito de identificar a influência de um cliente em outro, alguns autores (WEI; CHIU, 2002; ABBASIMEHR; SETAK; SOROOR, 2013; PHADKE *et al.*, 2013) utilizaram o total de chamadas realizadas para números diferentes para medir essa influência.

Buscando adaptar isto para uma base de plano de saúde, entende-se que um cliente é influente quando ele consegue atrair outros clientes para o seu plano de saúde. Então, em planos de saúde, uma maneira que a influência de um beneficiário pode ser medida é por meio do número de indicações de clientes.

A frequência de uso do plano de saúde é uma variável que representa a intensidade que um beneficiário usufrui do plano. É de se esperar que clientes com baixo número de uso tenham uma maior probabilidade de romperem com a operadora. Para se ter uma melhor descrição dos beneficiários do plano de saúde, sugere-se ter a frequência de uso estratificada por turno (dia, tarde e noite), por local (internacional ou não) e por tempo, possibilitando a criação do atributo “frequência de atendimentos rápidos (menos de 10 minutos)”. Ressalta-se que tais identificadores também foram sugeridos na seção 5.2.

Seguindo a mesma ideia de observar a intensidade de uso, Mendes (2008) utiliza duas variáveis para contabilizar o uso para cirurgias e internações. Espera-se que beneficiários que passaram por esse tipo de procedimento e tiveram uma boa experiência possuam menor probabilidade de romper com a empresa.

Para abranger os dependentes, as variáveis “número de dependentes que possui” e “número de dependentes cancelados” foram sugeridas a partir de variáveis utilizadas por vários autores (KUMAR; RAVI, 2008; BOTELHO; TOSTES, 2010; NIE *et al.*, 2011). Botelho e Tostes (2010) afirmam que a quantidade de cartões de créditos de um cliente influencia na ruptura do mesmo com a empresa. Dito isto, acredita-se que a quantidade de dependentes de um beneficiário também tenha influência na sua ruptura.

**Quadro 6 - Variáveis de contagem propostas para previsão de ruptura em um banco de dados de plano de saúde**

<b>Categoria</b>	<b>Variável proposta para Plano de Saúde</b>	<b>Variável base</b>	<b>Autores</b>
Contagem	Frequência de uso do plano de saúde (total, de dia, de tarde, de noite, uso internacional)	Frequência de uso (total de chamadas realizadas, de dia, de tarde, de noite, ligações internacionais)	WEI; CHIU, 2002; AU; CHAN; YAO, 2003; ABBASIMEHR; SETAK; SOROOR, 2013; ABBASIMEHR; SETAK; TAROKH, 2014; PHADKE <i>et al.</i> , 2013; LEMMENS; CROUX, 2006; VERBEKE <i>et al.</i> , 2011;
	Frequência de atendimentos rápidos (menos de 10min)	Frequência de ligações com duração menor que um minuto	PHADKE <i>et al.</i> , 2013; LEMMENS; CROUX, 2006;
	Influência (número de indicações de clientes)	Esfera de influência (total de chamadas realizadas para números diferentes)	WEI; CHIU, 2002; ABBASIMEHR; SETAK; SOROOR, 2013; PHADKE <i>et al.</i> , 2013;
	Número de extratos/faturas não pagas	Número de extratos/faturas que não foram pagas até o vencimento	BOTELHO; TOSTES, 2010;
	Número de dependentes que possui	Número de cartões de crédito o cliente possui no mês (t)	KUMAR; RAVI, 2008; BOTELHO; TOSTES, 2010; NIE <i>et al.</i> , 2011;
	Número de adicionais que possui	Número de cartões de crédito adicionais	BOTELHO; TOSTES, 2010;
	Número de dependentes cancelados	Número de cartões cancelados	NIE <i>et al.</i> , 2011;
	Igual à variável base	Número de rupturas anteriores	COUSSEMENT; BENOIT; VAN DEN POEL, 2010; COUSSEMENT; VAN DEN POEL, 2008a;
	Igual à variável base	Número de cirurgias	MENDES, 2008;
	Igual à variável base	Número de internações	MENDES, 2008;

Fonte: Autoria própria.

Coussement e Van den Poel (2008a, 2010) afirmam que a reincidência de ruptura de um cliente é uma variável importante para a previsão de *churn*. Assim sendo, o atributo “número de rupturas anteriores” entra, sem alteração, nas variáveis para compor a base de plano de saúde.

Botelho e Tostes (2010) identificam os bons e maus pagadores observando se o extrato ou fatura foi pago ou não. Os autores afirmam que este atributo influencia bastante a previsão de ruptura no setor de cartão de crédito. Diante do exposto, sugere-se esta variável para compor a base.

## 5.5 VARIÁVEIS DO PRODUTO

Quando se fala em produto de uma operadora de planos de saúde, está se falando nos planos oferecidos pelas mesmas (BRASIL, 2016). Deste modo, duas variáveis são apresentadas no Quadro 7 foram adaptadas de variáveis utilizadas na pesquisa original.

A primeira variável, “Qual o tipo de contratação do plano do cliente?”, é uma adaptação da variável base que informa se o plano contratado é um plano individual ou familiar, coletivo empresarial, coletivo por adesão, ou simplesmente coletivo não identificado. Essa variável faz parte do cadastro do beneficiário no momento da contratação do plano de saúde.

Já a segunda variável, “Qual a abrangência do plano de saúde contratado?”, deixa bem claro que faz referência à abrangência geográfica do plano (nacional, estadual, municipal). Acredita-se que essa variável tenha grande relação com a variável “Renda familiar do cliente”, da categoria demografia, e relação direta com a variável “Despesa mensal do serviço/produto”, da categoria gastos, pois quanto maior for a renda familiar do cliente, maior será a abrangência e mais caro será o valor gasto na mensalidade.

**Quadro 7 - Variáveis relacionadas ao produto oferecido propostas para previsão de ruptura em um banco de dados de plano de saúde**

<b>Categoria</b>	<b>Variável proposta para Plano de Saúde</b>	<b>Variável base</b>	<b>Autores</b>
Produto	Qual o tipo de contratação do plano do cliente?	Qual o produto o assinante possui; Tipo de plano do serviço	COUSSEMENT; VAN DEN POEL, 2008a; AU; CHAN; YAO, 2003; NIE <i>et al.</i> , 2011;
	Qual a abrangência do plano de saúde contratado?	Onde o serviço/produto é entregue?	COUSSEMENT; VAN DEN POEL, 2008a; COUSSEMENT; VAN DEN POEL, 2008b;

Fonte: Autoria própria.

## 5.6 VARIÁVEIS DO SERVIÇO

Nesta categoria, as variáveis propostas são tentativas de quantificar o serviço prestado ao beneficiário. Acredita-se que mapeando o tempo decorrido em todos os processos, podendo ser em dias e minutos, de cada beneficiário, essas variáveis de tempo contribuem para prever e caracterizar um beneficiário que irá cancelar o seu contrato de plano de saúde.

Vários artigos afirmam que o tempo de relacionamento do cliente com a empresa é uma variável importante para previsão de ruptura, podendo ser verificado no Quadro 8. Juntamente com o tempo de relacionamento, o ano de início e o mês de vencimento (COUSSEMENT; VAN DEN POEL, 2008a) de contrato são considerados atributos que ajudam a identificar possíveis clientes que romperão seus contratos com a empresa.

Constata-se, também, que variáveis associadas ao tempo de prestação do serviço são relevantes para o estudo. Neste momento, a estratificação sugerida nas seções 5.2 e 5.4 contribuem para as variáveis aqui sugeridas. O tempo médio de atendimento em minutos pode ser dividido em todas as estratificações já propostas.

Realizando o agrupamento em dias, temos o tempo desde o último uso do plano de saúde, o tempo decorrido até o uso do plano de saúde pela primeira vez, o maior intervalo entre consultas, o tempo desde a última consulta e o último exame. Essas variáveis surgiram a partir de variáveis balizadas em artigos de outros setores, como o de cartão de crédito e o de telecomunicações. Lembrando que todas as variáveis citadas são apresentadas no Quadro 8.



**Quadro 8 - Variáveis relacionadas ao serviço propostas para previsão de ruptura em um banco de dados de plano de saúde**

Categoria	Variável proposta para Plano de Saúde	Variável base	Autores
Serviço	Igual à variável base	Mês de vencimento do contrato	COUSSEMENT; VAN DEN POEL, 2008a;
	Tempo de Relacionamento (em dias) Ano de início de contrato	Duração do serviço; Início e Fim de contrato	WEI; CHIU, 2002; LEWIS, 2004; BUCKNIX; VAN DEN POEL, 2005; XIE et al., 2009; ABBASIMEHR; SETAK; SOROOR, 2013; ABBASIMEHR; SETAK; TAROKH, 2014; PHADKE et al., 2013; LEMMENS; CROUX, 2006; VERBEKE, 2011; YU et al., 2011; COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; COUSSEMENT; VAN DEN POEL, 2008a; MENDES, 2008;
	Tempo médio, em minutos, de atendimento (acumulado total, de dia, de tarde, de noite, internacional)	Minutos de uso (total de minutos em chamadas) por mês; Tempo médio de compras na loja	WEI; CHIU, 2002; AU; CHAN; YAO, 2003; PHADKE et al., 2013; LEMMENS; CROUX, 2006; ABBASIMEHR; SETAK; TAROKH, 2014; VERBEKE et al., 2011; BUCKNIX; VAN DEN POEL, 2005; MENDES, 2008;
	Tempo desde o último uso do plano de saúde (em dias)	Tempo em semanas/dias desde a compra anterior/uso do cartão	LEWIS, 2004; BUCKNIX; VAN DEN POEL, 2005; NIE et al., 2011; COUSSEMENT; VAN DEN POEL, 2008b;
	Tempo decorrido até o uso do plano de saúde pela primeira vez (em dias)	Tempo decorrido até o uso do cartão pela primeira vez	NIE et al., 2011;
	O maior intervalo (em dias) entre consultas	O maior intervalo entre as transações	NIE et al., 2011;
	Tempo, em dias, desde a última consulta	Minutos de uso (total de minutos em chamadas) por mês	WEI; CHIU, 2002; AU; CHAN; YAO, 2003; PHADKE et al., 2013; LEMMENS; CROUX, 2006; MENDES, 2008;
	Tempo, em dias, desde o último exame		

Fonte: Autoria própria.

## 5.7 VARIÁVEIS ADICIONAIS

Observando as variáveis presentes nos artigos apresentados no Quadro 1, foi possível verificar em dois artigos (VERBEKE *et al.*, 2011; ABBASIMEHR; SETAK; TAROKH, 2014) variáveis que englobam os adicionais contratados pelo cliente. Seguindo este caminho, decidiu-se criar a categoria adicionais para representar variáveis que caracterizem os serviços extras contratados pelo beneficiário.

Dentre esses dois artigos citados, sendo do setor de telecomunicações, dois atributos deram origem a cinco variáveis propostas, das quais contribuem para traçar o perfil do cliente e, conseqüentemente, melhorar a qualidade das previsões de ruptura. Essas variáveis são dicotômicas, recebendo 0 (zero) para a não existência do adicional e 1 (um) para a existência do adicional.

A primeira variável base faz referência a viagens de modo geral. Desta feita, sugerem-se as variáveis que contemplam os adicionais de plano internacional, viagens, emergência e emergência aérea. Já a segunda variável base faz referência a um adicional específico; deste modo, sugere-se uma variável para abranger um adicional específico também, o de odontologia. Um resumo das variáveis citadas nesta seção é apresentado no Quadro 9.

**Quadro 9 - Variáveis relacionadas a adicionais propostas para previsão de ruptura em um banco de dados de plano de saúde**

Categoria	Variável proposta para Plano de Saúde	Variável base	Autores
Adicionais	Possui adicional de plano internacional? (Sim ou não)	Possui ou não plano internacional de ligação?	ABBASIMEHR; SETAK; TAROKH, 2014; VERBEKE <i>et al.</i> , 2011;
	Possui adicional de emergência? (Sim ou não)		
	Possui adicional de emergência aérea? (Sim ou não)		
	Possui adicional de viagens? (Sim ou não)		
	Possui adicional de odontologia? (Sim ou não)	Possui ou não plano de mensagem de voz?	

Fonte: Autoria própria.

## 5.8 VARIÁVEIS DE OFERTAS

Acredita-se que aceitar uma oferta ou propaganda feita pela empresa contribua para a identificação de possíveis clientes que irão romper com a empresa (LEMMENS; CROUX, 2006; COUSSEMENT; VAN DEN POEL, 2008a; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; COUSSEMENT; VAN DEN POEL, 2008b; YU *et al.*, 2011). Por consequência, o Quadro 11 apresenta proposta de variáveis associadas a ofertas feitas pela empresa ao cliente que poderiam ser agregadas a bancos de dados de planos de saúde.

A primeira variável proposta é uma variável dicotômica, que indica se o beneficiário aceitou algum tipo de propaganda ou oferta de desconto oferecida pelo plano de saúde. Em seguida, a segunda e a terceira variáveis propostas utilizam como base o trabalho do Yu *et al.* (2011), que busca identificar aspectos de descontos fornecidos ao cliente. Assim, buscam apresentar se o beneficiário obteve desconto na assinatura do contrato e qual foi o valor de desconto oferecido.

**Quadro 10 - Variáveis relacionadas a ofertas feitas propostas para previsão de ruptura em um banco de dados de plano de saúde**

<b>Categoria</b>	<b>Variável proposta para Plano de Saúde</b>	<b>Variável base</b>	<b>Autores</b>
Ofertas	Aceitou oferta/propaganda para assinar o contrato do plano de saúde? (Sim ou não)	Aceitou oferta/propaganda para comprar (pode ser por e-mail ou não)	LEMMENS; CROUX, 2006; COUSSEMENT; VAN DEN POEL, 2008b; COUSSEMENT; BENOIT; VAN DEN POEL, 2010; COUSSEMENT; VAN DEN POEL, 2008a;
	Houve desconto na assinatura do contrato? (Sim ou não)	Desconto (houve desconto? - Sim ou não; quanto?)	YU <i>et al.</i> , 2011;
	Qual o valor de desconto oferecido ao cliente?		

Fonte: Autoria própria.

Ao final desse capítulo, o objetivo específico de apresentar variáveis adequadas à composição de identificação à propensão de ruptura de clientes.

## 6 PROCEDIMENTOS METODOLÓGICOS

### 6.1 CARACTERIZAÇÃO DO ESTUDO

Com a finalidade de atender ao objetivo proposto, que é classificar o risco de ruptura de clientes (beneficiários) de operadora de plano de saúde privado, um estudo exploratório (GIL, 2010) foi feito para conhecer as características do objeto de estudo e ter uma maior familiaridade com tema escolhido, no caso, a ruptura de clientes. Quanto à natureza da pesquisa, o estudo se classifica como pesquisa aplicada, por procurar gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos (GIL, 2010).

### 6.2 BANCO DE DADOS

Para o trabalho que foi desenvolvido nesta dissertação, uma operadora de planos de saúde que solicitou não ser identificada, especificamente de assistência médica e que abrange todo o Brasil, disponibilizou todo seu banco de dados. Os dados foram selecionados com a empresa, e três bases com informações desde o ano de 2000, dada a regulamentação, foram repassadas ao pesquisador para a realização deste estudo. O acesso e disponibilidade à base se deu por meio de parceria acadêmica entre o Programa de Pós-Graduação em Administração da Universidade Estadual do Ceará (PPGA/UECE) e a operadora de planos de saúde.

A primeira base apresenta 23 variáveis com informações sobre os clientes, onde estas são descritas e apresentadas no Quadro 11. No total, a base possui 21.074 beneficiários. A segunda base é referente a despesas médicas de cada beneficiário, possuindo quatro variáveis que classificam quando, quanto, quem e em que esse valor foi gasto. O Quadro 12 descreve as variáveis presentes nessa base. Esta base contém 298.375 *data points*. A terceira e maior base, que possui 430.233 registros, mostra a mensalidade que o beneficiário paga para cada competência. O Quadro 13 apresenta as quatro variáveis presentes nessa base de dados.

Vale ressaltar que os dados são oriundos de uma operadora de plano de saúde real; assim, nem todas as variáveis mencionadas no capítulo 5 puderam ser contempladas, dada a disponibilidade de dados da empresa operadora.

**Quadro 11 - Descrição das variáveis da primeira base de dados (dados cadastrais)**

Variáveis	Descrição
DATA_PRIMEIRO_PROCEDIMENTO	Data da realização do primeiro procedimento do beneficiário
DATA_ULTIMO_PROCEDIMENTO	Data da realização do último procedimento do beneficiário
SEGMENTO_DE_UTILIZACAO	Tipo de contratação do plano do beneficiário (plano individual ou familiar, coletivo empresarial, coletivo por adesão, coletivo não identificado)
VENCIMENTO_ULT_PAGAMENTO	Data do vencimento da última fatura
PAGAMENTO_ULT_PAGAMENTO	Data de pagamento da última fatura
DATAULTIMACONSULTA	Data da última consulta realizada pelo beneficiário
DATAINCLUSAO	Data de início do contrato
NASCIMENTO	Data de nascimento do beneficiário
SEXO	Sexo do beneficiário
ESTADOCIVIL	Estado civil do beneficiário
CEP	CEP do beneficiário
CIDADE	Cidade onde o beneficiário reside
ESTADO	Unidade federativa que o beneficiário reside
EH_TITULAR	Código identificador do beneficiário titular
EH_OPCIONALODONTOLOGIA	Se o cliente possui ou não opcional de odontologia
OPCIONAL_DE_EMERGENCIA	Se o cliente possui ou não opcional de emergência
OPCIONAL_DE_EMERGENCIA_AEREA	Se o cliente possui ou não opcional de emergência aérea
CARTEIRA	Indica a forma de pagamento da fatura
COBERTURA	Abrangência geográfica do plano
DATADEFALECIMENTO	Data de falecimento do cliente
AGRUPADORDAFAMILIA	Código para associar dependentes
IDENTIFICADORBENEFICIARIO	Código identificador do beneficiário
DATADEEXCLUSAO	Apresenta a data que o cliente rompeu com a empresa. Recebe vazio se o cliente não tiver cancelado o contrato.

Fonte: Autoria própria.

**Quadro 12 - Descrição das variáveis da segunda base de dados (dados de uso e custo)**

Variáveis	Descrição
CUSTO_TOTAL_BENEF	Valor gasto para aquele tipo de despesa
COMPETENCIA	Referência de qual ano e mês aquele valor foi gasto
IDENTIFICADORBENEFICIARIO	Código identificador do beneficiário
TIPO DE DESPESA	Tipo de despesa, podendo ser Consulta, Serviço Auxiliar Diagnóstico e Terapia (SP/SADT), guias de PCMSO, Internação, Internação (Honorários) ou odontológica.

Fonte: Autoria própria.

**Quadro 13 - Descrição das variáveis da terceira base de dados (dados de mensalidades)**

Variáveis	Descrição
TOTAL	Valor da mensalidade gasto para aquela competência
COMPETENCIA	Referência de qual ano e mês aquele valor foi gasto
IDENTIFICADOR_BENEFICIARIO	Código identificador do beneficiário

Fonte: Autoria própria.

### 6.3 A METODOLOGIA CRISP-DM

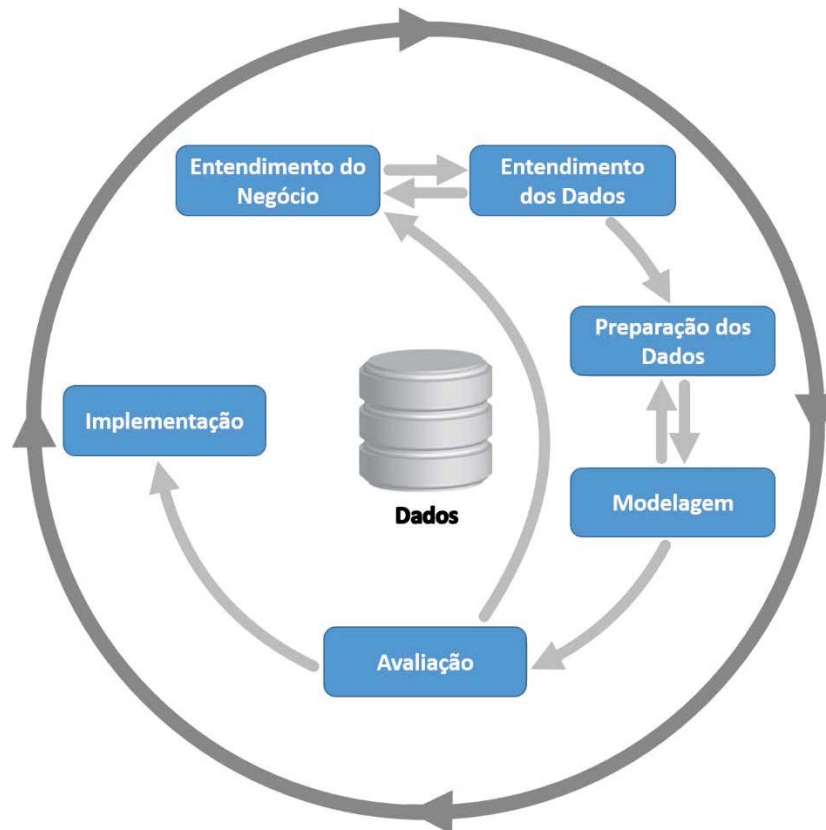
Apresentada na Figura 4, a metodologia CRISP-DM é composta por seis fases: (1) entendimento do negócio; (2) entendimento dos dados; (3) preparação dos dados; (4) modelagem; (5) avaliação; e (6) implementação (SHEARER, 2000). Embora esses passos devam ser executados na ordem apresentada, o processo é interativo — uma vez que o analista pode intervir as atividades realizadas — e iterativo, por ser uma sequência finita de operações em que o resultado de cada uma depende da saída das operações que a antecedem (CHAPMAN *et al.*, 2000).

Esta metodologia é descrita em termos de um modelo hierárquico de processos que se fundamenta num conjunto de tarefas por níveis (do geral para o específico): fases, tarefas genéricas, tarefas especializadas e instâncias de processos (CHAPMAN *et al.*, 2000). Deve-se destacar que entre as etapas de Entendimento do Negócio e Entendimento dos dados não há hierarquia, bem como entre Preparação dos Dados e Modelagem.

Os principais objetivos são: transformar as necessidades de negócios em

tarefas de *data mining*, promover mudanças nas técnicas e nos dados, fazer uso de métricas para avaliação dos resultados e elaborar a documentação do projeto (SHEARER, 2000).

**Figura 4 - Representação das fases do modelo de referência do CRISP-DM**



Fonte: CHAPMAN *et al.*, 2000.

Shearer (2000) afirma que o sucesso de tal metodologia decorre do fato de ter sido desenvolvida à luz da prática, fazendo com que as melhores rotinas utilizadas em um projeto de análise de dados sejam unidas e aplicadas sobre todo o processo de mineração de dados. A seguir, é feita uma descrição de cada uma das fases e das tarefas que as compõem da metodologia CRISP-DM, conforme apresentado por Chapman *et al.* (2000).

### 6.3.1 Entendimento do Negócio (*Business Understanding*)

O primeiro passo é compreender a perspectiva do problema, identificar os objetivos e restrições segundo a perspectiva organizacional. Descobrir os fatores

importantes que influenciam os resultados. O conhecimento adquirido nesse estudo será utilizado para a definição do problema de mineração de dados e na concepção do plano preliminar. O entendimento do negócio realiza-se pelas seguintes tarefas:

- a) **determinação dos objetivos do negócio:** compreender todos os aspectos que influenciam o negócio e conhecer os objetivos fundamentais do cliente;
- b) **avaliação da situação atual:** fazer um levantamento dos recursos disponíveis, dos requisitos, pressupostos e restrições do projeto. Identificar todos os riscos, ameaças ou eventos que possam comprometer a realização do projeto;
- c) **definição dos objetivos de *data mining*:** descrever os objetivos de *data mining* e os critérios de seu sucesso (classificação, previsão, segmentação);
- d) **definição do plano para o projeto:** elaborar um plano para o projeto que tenha a duração, os recursos e as fases. Apresenta-se, também, as ferramentas e as técnicas que serão utilizadas no projeto.

Os frutos deste estudo do negócio aparecem em forma de um plano do projeto que inclui toda a informação acerca do negócio e os resultados de cada tarefa.

### 6.3.2 Entendimento dos Dados (*Data Understanding*)

Nesta etapa, se faz necessário obter o conteúdo inicial dos dados, os quais seguirão com uma devida análise, de forma que se apontem os problemas de qualidade. Logo, será importante levar em conta algumas tarefas para que se possam aplicar as técnicas de DM aos dados, tais como:

- a) **obtenção inicial dos dados:** compreende na obtenção dos dados e de seu entendimento conciso e claro. Desta sequência adquire-se uma listagem dos dados com subdivisões de localização, métodos de aquisição, problemas diagnosticados e soluções constatadas;
- b) **descrição dos dados:** tendo os pontos com seus devidos enfoques, a atitude neste momento será de descrição e reconhecimento de seus formatos, perceber a quantidade de registros nas tabelas, identificando-os, para que se facilite o foco no(s) problema(s);
- c) **exploração dos dados:** por meio desta etapa, será criada, no projeto como um todo, uma listagem inicial de hipóteses e seus devidos impactos. Neste



intuito, ferramentas como histogramas, gráficos etc. indicariam as características dos dados de forma mais prática. Para tanto, as características são listadas de duas formas: quantitativa (seus devidos valores numéricos) e qualitativas (dados ordinais e nomes específicos). Dependendo do objetivo da mineração dos dados, cada classificação seguirá para uma análise específica;

d) **verificação da qualidade dos dados**: nesta subfase, faz-se um relatório que deverá incluir os problemas de qualidade e possíveis modos de solução.

### 6.3.3 Preparação dos dados (*Data Preparation*)

O conjunto final de dados é construído ao final dessa fase; assim, entende-se que todas as atividades relacionadas ao desenvolvimento do banco de dados a ser trabalhado ocorrem aqui. Esta fase contém cinco subfases: (1) seleção de dados, (2) limpeza de dados, (3) derivação de dados, (4) integração de dados e (5) formatação de dados. Cada uma delas é descrita a seguir.

- a) **seleção de dados**: consiste na escolha dos dados a utilizar na análise. Restrições técnicas e de qualidade podem influenciar no volume e tipo de dados a serem trabalhados. É apresentada, ao final da tarefa, uma listagem dos dados incluídos e excluídos e suas razões para tal;
- b) **limpeza de dados**: existem inúmeras técnicas para otimizar a qualidade dos dados. Usualmente, a normalização dos dados e tratamento dos dados omissos são aplicados nessa tarefa;
- c) **derivação de dados**: momento em que novos atributos são gerados (e.g. calcula-se a idade do indivíduo a partir do atributo data de nascimento);
- d) **integração de dados**: criação de novos registros ou valores por meio da combinação de múltiplas tabelas ou registros (e.g. junção e/ou agregação de tabelas ou registros); e
- e) **formatação de dados**: consiste em modificações sintáticas nos dados (e.g. reordenação dos atributos e registros), de modo que não mudem o seu significado, mas que os tornem utilizáveis pela ferramenta de modelagem escolhida.

### 6.3.4 Modelagem (*Modelling*)

Nesta fase, ocorre a seleção de várias técnicas de modelagem, e os seus parâmetros são ajustados para melhorar os resultados. Existem inúmeras técnicas disponíveis. Resta saber qual delas se aplica melhor aos dados e aos objetivos apresentados, podendo ser necessário voltar à fase anterior para algumas adaptações. As seguintes tarefas são abordadas nessa fase:

- a) **seleção de técnicas de modelagem**: escolha da técnica mais adequada ao tipo de problema, às ferramentas e aos objetivos da mineração de dados. Esta tarefa inclui decidir quais os modelos e parâmetros podem ser apropriados para o caso em estudo (e.g. árvore de decisão, regressão logística e redes neurais), tendo em vista que cada algoritmo tem parâmetros e táticas de aprendizado diferentes;
- b) **definição de uma concepção de teste**: importa, antes de construir o modelo, definir um procedimento ou um mecanismo para testar o desempenho do próprio modelo, por exemplo, curva de ROC (*Receiver Operating Curve*) e percentual de acerto (e.g. relação de Verdadeiros Positivos com Falsos Positivos);
- c) **construção do modelo**: consiste na aplicação da ferramenta de modelagem ao conjunto de dados preparados anteriormente, possibilitando a criação de um ou mais modelos. Ao final, os modelos resultantes devem ser interpretados e o seu desempenho explicado; e
- d) **revisão do modelo**: interpretar os modelos criados de acordo com o domínio do conhecimento e com o mecanismo de teste pré-definido. A avaliação deve levar em consideração o impacto dos resultados no contexto do negócio.

Vale lembrar que a maioria das técnicas de mineração de dados é baseada no aprendizado indutivo, no qual um modelo é construído de forma explícita ou implícita, por generalizar a partir de um número suficiente de exemplos de treinamento (LAROSE, 2005, 2006; MAIMON; ROKACH, 2010).

### 6.3.5 Avaliação (*Evaluation*)

Consiste na validação do(s) modelo(s), revisão dos passos já executados e verificação se os objetivos foram alcançados. As seguintes tarefas fazem parte desta fase:

- a) **avaliação dos resultados:** determinar se o(s) modelo(s) atingi(ram) os objetivos do negócio. Visualizar os dados e determinar se as descobertas e similaridades dos padrões encontrados fazem sentido para o negócio;
- b) **revisão do processo:** analisar todas as fases do processo, buscando destacar eventuais atividades que possam ter sido esquecidas ou que precisem ser repetidas; e
- c) **determinação dos próximos passos:** considera-se o projeto concluído se todas as fases anteriores tiverem sido satisfatórias e os resultados atingidos suprirem os objetivos. Caso não se possa afirmar que o projeto esteja concluído, faz-se necessária uma nova iteração das fases anteriores, utilizando novos parâmetros. Também é possível que os resultados obtidos sugiram um novo projeto.

### 6.3.6 Implementação (*Deployment*)

Nesta fase, o conhecimento descoberto incorpora-se a outro sistema. Na verdade, o sucesso deste passo determina a eficácia global do projeto, e um dos desafios desta etapa é não ter mais as condições de laboratório em que todo o processo e construção do modelo foi realizado (LAROSE, 2005; MAIMON; ROKACH, 2010). A implementação pode ser simples (e.g. gerar um relatório) ou pode ser complexa (e.g. integrar os resultados nos sistemas da organização). As tarefas envolvidas nesta fase são:

- a) **planejamento da avaliação dos resultados:** definição da estratégia para a implementação dos resultados, incluindo os passos e a forma de execução.
- b) **planejamento do controle e manutenção:** consiste na definição da estratégia de controle e manutenção. Com o retorno desse monitoramento e manutenção, é possível verificar se os modelos são usados corretamente e trazem os resultados esperados.
- c) **produção um relatório final:** é a subfase de conclusão do projeto. Elaborase um relatório final resumindo os pontos mais importantes no projeto, experiência adquirida, explicação dos resultados mais importantes produzidos.
- d) **revisão do projeto:** avaliação dos acertos e dos erros. Resume-se as

experiências mais importantes do projeto, com o intuito de dar suporte a projetos futuros, em situações similares (e.g. apontar aproximações erradas).

### 6.3.7 Ferramenta de *Data mining* utilizada

Como ferramenta para o desenvolvimento dos modelos de data mining desta dissertação, foi escolhido o WEKA (Waikato Environment for Knowledge Analysis). O WEKA é um *software* opensource de análise estatística e de data mining desenvolvido pela Universidade de Waikato, na Nova Zelândia. É uma ferramenta que permite, respeitando o processo de CRISP-DM, realizar o processo de implementação de um modelo preditivo, desde o carregamento até a geração dos resultados finais. Tal ferramenta será utilizada para explorar e comparar as técnicas de árvore de decisão (C4.5), regressão logística e redes neurais (MLP e RBF), buscando atender os objetivos específicos (vide Figura 5).

**Figura 5 - Apresentação do WEKA**



Fonte: HALL *et al.*, 2009.

Para confecção da base, agrupamento e estatísticas descritivas, utilizou-se o Excel, uma vez que tal ferramenta funciona bem para o tamanho da base de dados trabalhada, mesmo com 741.864 mil *data points* (somando as três bases). Caso a base de dados fosse muito maior, outra ferramenta precisaria ser utilizada.

O tipo de arquivo de dados aceito no WEKA é o *.arff*. Para isso, utilizou-se a ferramenta estatística R para a conversão do formato dos dados, de *.xlsx* para *.arff*. O R é

um ambiente de *software* livre para análise de dados estatísticos, gráficos e manipulação de dados. Além disso, esta ferramenta foi utilizada para a junção das três bases de dados.

## 7 ESTUDO DE CASO

Neste capítulo, são apresentadas as análises relacionadas ao desenvolvimento e aplicação dos modelos propostos. Estatísticas descritivas da amostra estudada serão reveladas e, por fim, os resultados obtidos serão dissertados na fase de implementação do processo de CRISP-DM desenvolvido neste estudo de caso. Neste capítulo, serão delineados como os objetivos propostos foram atingidos.

### 7.1 ENTENDIMENTO DO NEGÓCIO (*BUSINESS UNDERSTANDING*)

Com base no levantamento bibliográfico feito nesta dissertação, entende-se que a ruptura de clientes é um problema frequentemente enfrentado pelas empresas de diversos setores, em especial por operadoras de planos de saúde. Uma abordagem para diminuir o *churn* poderia ser o uso da mineração de dados para determinar as causas desse evento (ruptura do contrato) para que medidas preventivas possam ser tomadas.

Identifica-se que o problema de alta taxa de *churn* é geralmente entendido como um problema de classificação, e utilizam-se técnicas de *data mining* para sua resolução. Assim, um processo de CRISP-DM pode ser aplicado, e técnicas de árvore de decisão (C4.5), regressão logística e redes neurais (MLP e RBF) serão utilizadas na etapa de modelagem, uma vez que foi feito o levantamento das técnicas mais utilizadas para predição de *churn*.

Para a avaliação e comparação dos modelos, a curva ROC (*Receiver Operating Characteristic*) e o percentual de acerto serão considerados. A curva ROC coloca em um gráfico a sensibilidade e a especificidade para todos os possíveis pontos de corte entre 0 (zero) e 1 (um) e apresenta o valor da área sob a curva.

Neste estudo, utiliza-se dados de um plano de saúde com cadastro de clientes a partir de 2000, pois beneficiários com data de inclusão anterior a 1999 seriam amparados por uma legislação diferente do que os planos regulamentados após esse ano. Especificamente, a base de clientes oferece dados dos beneficiários entre os anos de 2000 e 2015.

Este capítulo busca alcançar o objetivo principal de classificar o risco de ruptura de clientes (beneficiários) de operadora de plano de saúde privado, bem como os objetivos específicos de examinar modelos de classificação de risco de ruptura de clientes e identificar o principal modelo que se adequa às condições de previsão de classificação de ruptura de clientes.

## 7.2 ENTENDIMENTO DOS DADOS (*DATA UNDERSTANDING*)

Buscou-se, nesta fase, analisar a base de dados obtida de um plano de saúde com a finalidade de apontar problemas de qualidade nas variáveis. Desta forma, cada variável foi avaliada em relação aos seus valores faltantes (*missing values*), *outliers* e consistência com outras variáveis.

A variável `DATA_ULTIMO_PROCEDIMENTO` possui datas de procedimentos que aconteceram depois da exclusão do beneficiário da base. Logo, há inconsistência nesta variável. Nesta mesma direção, existem valores da variável `DATAULTIMACONSULTA` que apontam para datas de consultas que ocorreram depois que o cliente cancelou seu contrato com o plano de saúde. Esta variável também possui problema de qualidade.

Foi também identificado que existem beneficiários que pagaram seus vencimentos muito tempo depois. Por exemplo, quando a operadora entra na justiça para cobrar as faturas atrasadas, mesmo que o vínculo do beneficiário tenha sido cancelado.

Assim, foi decidido que as variáveis que representam a data de vencimento da última fatura (`VENCIMENTO_ULT_PAGAMENTO`) e a data de pagamento da última fatura (`PAGAMENTO_ULT_PAGAMENTO`) não seriam utilizadas para a construção do modelo, pois o modelo poderia produzir previsões ruins.

Verificou-se que algumas variáveis não poderiam ser utilizadas no estudo, uma vez que não há variação em seus valores. As variáveis são:

- a) **SEGMENTO\_DE\_UTILIZACAO**: todos os beneficiários possuem plano coletivo empresarial;
- b) **OPCIONAL\_DE\_EMERGENCIA**: nenhum dos beneficiários possui este opcional;
- c) **OPCIONAL\_DE\_EMERGENCIA\_AREA**: nenhum dos beneficiários possui este opcional; e
- d) **COBERTURA**: todos os beneficiários possuem cobertura nacional;

Em relação aos valores faltantes, apenas dois beneficiários não possuíam resposta na variável `SEXO`, três no atributo `DATA_PRIMEIRO_PROCEDIMENTO` e quatro na variável `NASCIMENTO`. Nenhum problema com *outliers* foi encontrado. Verificou-se que o conjunto de dados possui 21.074 beneficiários, onde 76,25% romperam com a empresa (variável que se deseja prever) e que a maioria dos clientes deste plano de saúde é solteiro (49,37%), seguido dos casados (28,70%), conforme Tabela

1. Deve-se destacar o grande número de pessoas solteiras que rompem com o plano de saúde, podendo ser uma oportunidade de melhorar o produto oferecido.

Descrevendo as variáveis, percebe-se que existem mais mulheres (51,07%) do que homens, entretanto constata-se que a diferença é de apenas, aproximadamente, 2%, o que leva a acreditar que a base de dados está bem distribuída quanto ao sexo dos beneficiários (Tabela 2).

**Tabela 1 - Tabela de frequência Estado civil vs. Churn**

Estado Civil	Churn?		Total	%
	0	1		
Solteiro	2500	7905	10405	49,37%
Casado	1933	4115	6048	28,70%
Outros	438	3700	4138	19,64%
Divorciado	85	201	286	1,36%
Viúvo	28	93	121	0,57%
Separado judicialmente	22	54	76	0,36%
<b>Total</b>	<b>5006</b>	<b>16068</b>	<b>21074</b>	<b>100%</b>
<b>%</b>	<b>23,75%</b>	<b>76,25%</b>	<b>100%</b>	

Fonte: Autoria própria.

**Tabela 2 - Tabela de frequência Sexo vs. Churn**

Sexo	Churn?		Total	%
	0	1		
<b>F</b>	2552	8211	10763	51,07%
<b>M</b>	2454	7857	10311	48,93%
<b>Total</b>	<b>5006</b>	<b>16068</b>	<b>21074</b>	<b>100%</b>
<b>%</b>	<b>23,75%</b>	<b>76,25%</b>	<b>100%</b>	

Fonte: Autoria própria.

Verificou-se que o plano de saúde abrange todas as regiões do Brasil, estando presente em 22 estados, faltando apenas os estados do Amapá, Mato Grosso, Rondônia e Roraima. A Tabela 3 mostra que os beneficiários deste plano de saúde estão concentrados nas regiões do Nordeste e Centro-Oeste, tendo, somente nessas duas regiões, mais de 85% de todos os beneficiários do plano de saúde. Desta maneira, observa-se que existem números elevados de beneficiários que romperam com o plano de saúde.



**Tabela 3 - Tabela de frequência por região vs. Churn**

Região	Churn?		Total	%
	0	1		
Nordeste	2764	9622	12386	58,77%
Centro-Oeste	1613	4395	6008	28,51%
Sudeste	623	2018	2641	12,53%
Norte	3	20	23	0,11%
Sul	3	13	16	0,08%
<b>Total</b>	<b>5006</b>	<b>16068</b>	<b>21074</b>	<b>100%</b>
<b>%</b>	<b>23,75%</b>	<b>76,25%</b>	<b>100%</b>	

Fonte: Autoria própria.

A base é composta por beneficiários titulares (26,07%) e dependentes (73,93%) que cada um possui registro diferente, assim, ambos podem romper com o plano de saúde sem interferir no outro. A grande maioria destes beneficiários paga o plano de saúde direto na folha de pagamento (63,11%), seguido pelo pagamento por boleto bancário (24,72%). Tais informações podem ser conferidas nas Tabelas 4 e 5.

**Tabela 4 - Tabela de frequência de titulares e dependentes vs. Churn**

Tipo de cadastro	Churn?		Total	%
	0	1		
Dependente	3312	12268	15580	73,93%
Titular	1694	3800	5494	26,07%
<b>Total</b>	<b>5006</b>	<b>16068</b>	<b>21074</b>	<b>100%</b>
<b>%</b>	<b>23,75%</b>	<b>76,25%</b>	<b>100%</b>	

Fonte: Autoria própria.

Diante da Tabela 5, observa-se que há uma maior tendência para romper com o plano de saúde beneficiários que o pagam por boleto bancário, haja vista que é o tipo de pagamento que possui a maior taxa de *churn* ( $9.330 / 13.300 = 91,05\%$ ). Em segundo lugar se encontra a modalidade de pagamento débito em conta, com uma taxa de 80,10% ( $447 / 558$ ).

**Tabela 5 - Tabela de frequência do modo de pagamento vs. Churn**

Tipo de pagamento	Churn?		Total	%
	0	1		
Folha	3970	9330	13300	<b>63,11%</b>
Boleto Bancário	466	4743	5209	<b>24,72%</b>
Folha Fundação	381	1321	1702	<b>8,08%</b>
Débito em Conta	111	447	558	<b>2,65%</b>
Carteira	78	213	291	<b>1,38%</b>
TPMF 95	-	14	14	<b>0,07%</b>
<b>Total</b>	<b>5006</b>	<b>16068</b>	<b>21074</b>	<b>100%</b>
<b>%</b>	<b>23,75%</b>	<b>76,25%</b>	<b>100%</b>	

Fonte: Autoria própria.

### 7.3 PREPARAÇÃO DOS DADOS (*DATA PREPARATION*)

Conforme avaliado na seção 7.2, algumas variáveis precisariam ser desconsideradas e alguns registros deveriam ser tratados. Nesta etapa, a base de dados para confecção do modelo será desenvolvida. O Quadro 14 apresenta as nove variáveis excluídas da base e seu motivo geral da exclusão. Ressalta-se que a variável *DATADEFALECIMENTO* foi retirada após a exclusão dos 71 registros que tiveram *churn* por morte, haja vista que este tipo de ruptura não é relevante para a análise.

**Quadro 14 - Variáveis excluídas do estudo**

Variáveis	Motivo
DATA_ULTIMO_PROCEDIMENTO	Inconsistência
SEGMENTO_DE_UTILIZACAO	Sem variação
VENCIMENTO_ULT_PAGAMENTO	Muitos pagamentos fora do prazo (mais de seis meses depois)
PAGAMENTO_ULT_PAGAMENTO	
DATAULTIMACONSULTA	Inconsistência
OPCIONAL_DE_EMERGENCIA	Sem variação
OPCIONAL_DE_EMERGENCIA_AEREA	Sem variação
COBERTURA	Sem variação
DATADEFALECIMENTO	Não relevante

Fonte: Autoria própria.

Quanto à limpeza dos dados, os valores faltantes foram substituídos pela média, pois é a opção oferecida pela ferramenta computacional utilizada. Já em relação à

derivação dos dados, novas variáveis foram criadas, buscando alcançar as variáveis sugeridas no capítulo 5, uma vez que a operadora do plano de saúde não possuía todas as variáveis propostas.

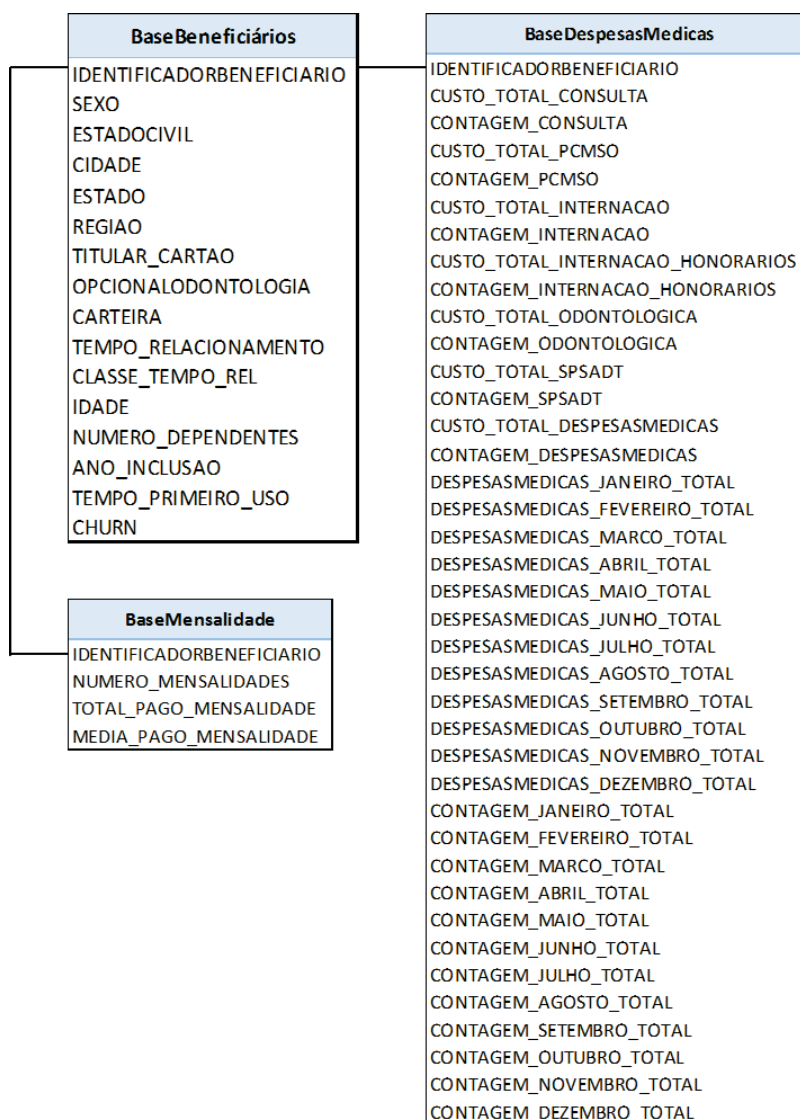
A idade de cada beneficiário foi calculada por meio da data de nascimento, onde o ano referência foi 2015. A variável dependente *churn* foi calculada por meio da data de exclusão do registro. Logo, se houvesse data na variável *DATADEEXCLUSAO*, o atributo receberia 1, caso contrário, 0.

O tempo de relacionamento, em dias, foi calculado pela diferença entre a data de exclusão e a data de inclusão. Se não houvesse data de exclusão, significa que o cliente não rompeu com o plano e a data referência para este procedimento foi 31 de dezembro de 2015. A partir desta variável, foi criada outra variável que agrupou, em quatro classes, o tempo de relacionamento de acordo com a classificação dos quartis.

O tempo decorrido até o primeiro uso, em dias, foi calculado a partir da diferença entre a data do primeiro procedimento e a data de inclusão. Outra variável calculada foi o número de dependentes que cada titular possui. Tal variável foi calculada por meio do agrupamento dos registros pela variável *AGRUPADORDAFAMILIA*.

O modelo do banco de dados apresentado na Figura 6 mostra como foi feita a junção das três bases de dados e quais variáveis foram criadas a partir de cada base, onde a ligação de cada base é feita por meio da chave de identificador do beneficiário. A base final foi composta por 56 variáveis independentes e uma dependente (*churn*). O Quadro 15 apresenta as variáveis presentes nesta base final e o tipo de cada variável (e.g. ordinais, nominais, discretas e contínuas).

**Figura 6 - Modelo do banco de dados**



Fonte: Autoria própria.

**Quadro 15 - Descrição das variáveis da base final de dados (dados dos beneficiários do plano de saúde)**

Variáveis	Descrição	Variáveis	Descrição	Variáveis	Descrição
IDENTIFICADORBENEFICIARIO	ID	CONTAGEM_INTERNACAO	Discreto	DESPESASMEDICAS_DEZEMBRO_TOTAL	Contínua
SEXO	Nominal	CUSTO_TOTAL_INTERNACAO_HONORARIOS	Contínua	CONTAGEM_JANEIRO_TOTAL	Discreto
ESTADOCIVIL	Nominal	CONTAGEM_INTERNACAO_HONORARIOS	Discreto	CONTAGEM_FEVEREIRO_TOTAL	Discreto
CIDADE	Nominal	CUSTO_TOTAL_ODONTOLOGICA	Contínua	CONTAGEM_MARCO_TOTAL	Discreto
ESTADO	Nominal	CONTAGEM_ODONTOLOGICA	Discreto	CONTAGEM_ABRIL_TOTAL	Discreto
REGIAO	Nominal	CUSTO_TOTAL_SPSADT	Contínua	CONTAGEM_MAIO_TOTAL	Discreto
TITULAR_CARTAO	Nominal	CONTAGEM_SPSADT	Discreto	CONTAGEM_JUNHO_TOTAL	Discreto
OPCIONALODONTOLOGIA	Nominal	CUSTO_TOTAL_DESPESASMEDICAS	Contínua	CONTAGEM_JULHO_TOTAL	Discreto
CARTEIRA	Nominal	CONTAGEM_DESPESASMEDICAS	Discreto	CONTAGEM_AGOSTO_TOTAL	Discreto
TEMPO_RELACIONAMENTO	Discreto	DESPESASMEDICAS_JANEIRO_TOTAL	Contínua	CONTAGEM_SETEMBRO_TOTAL	Discreto
CLASSE_TEMPO_REL	Nominal	DESPESASMEDICAS_FEVEREIRO_TOTAL	Contínua	CONTAGEM_OUTUBRO_TOTAL	Discreto
IDADE	Contínua	DESPESASMEDICAS_MARCO_TOTAL	Contínua	CONTAGEM_NOVEMBRO_TOTAL	Discreto
NUMERO_DEPENDENTES	Discreto	DESPESASMEDICAS_ABRIL_TOTAL	Contínua	CONTAGEM_DEZEMBRO_TOTAL	Discreto
ANO_INCLUSAO	Nominal	DESPESASMEDICAS_MAIO_TOTAL	Contínua	NUMERO_MENSALIDADES	Discreto
TEMPO_PRIMEIRO_USO	Discreto	DESPESASMEDICAS_JUNHO_TOTAL	Contínua	TOTAL_PAGO_MENSALIDADE	Contínua
CUSTO_TOTAL_CONSULTA	Contínua	DESPESASMEDICAS_JULHO_TOTAL	Contínua	MEDIA_PAGO_MENSALIDADE	Contínua
CONTAGEM_CONSULTA	Discreto	DESPESASMEDICAS_AGOSTO_TOTAL	Contínua	CHURN	Target
CUSTO_TOTAL_PCMSO	Contínua	DESPESASMEDICAS_SETEMBRO_TOTAL	Contínua		
CONTAGEM_PCMSO	Discreto	DESPESASMEDICAS_OUTUBRO_TOTAL	Contínua		
CUSTO_TOTAL_INTERNACAO	Contínua	DESPESASMEDICAS_NOVEMBRO_TOTAL	Contínua		

Fonte: Autoria própria.

Comparando as variáveis sugeridas no capítulo 5 com as variáveis encontradas no banco de dados real, 14 variáveis foram utilizadas, são elas: idade, estado civil, sexo, localização, forma de pagamento, abrangência do plano, número de dependentes, tempo de relacionamento (em dias e em classes), ano de início de contrato, tempo decorrido (em dias) até primeiro uso do plano, adicional de odontologia, informações sobre despesas mensais, valores totais gastos e frequência de uso.

Ressalte-se que inúmeras variáveis foram criadas para abranger, de forma ampla, as variáveis sugeridas para despesas, valores gastos e frequência de uso do plano de saúde.

#### 7.4 MODELAGEM (*MODELLING*)

Conforme estabelecido na primeira etapa do CRISP-DM, as técnicas selecionadas para essa dissertação são árvore de decisão (C4.5), regressão logística e redes neurais (MLP e RBF). Também, conforme a etapa, a avaliação e comparação dos modelos construídos serão feitas pela curva ROC (*Receiver Operating Characteristic*) e pelo percentual de acerto.

Para apresentação do percentual de acerto, utiliza-se a matriz de confusão. Estas tabelas apresentam as seguintes informações: verdadeiro negativo (observado não *churn* e previsto não *churn*); falso negativo (observado *churn* e previsto não *churn*); falso positivo (observado não *churn* e previsto *churn*); e verdadeiro positivo (observado *churn* e previsto *churn*).

Como um modelo de previsão precisa prever bem com poucas variáveis (HAIR *et al.*, 2009), decidiu-se, utilizando o WEKA, fazer uma seleção de variáveis por meio da estatística de Qui-Quadrado, que considera a importância de cada variável no modelo proposto. Desta maneira, oito variáveis foram selecionadas e apresentadas na Tabela 6. O valor percentual foi calculado a partir dos valores absolutos da estatística em estudo de cada variável, tornando possível comparar a importância das variáveis.

As três variáveis mais importantes, segundo seu critério de seleção, são, respectivamente: ano de inclusão na base, valor médio pago em mensalidades e a classe do tempo de relacionamento. Essas variáveis possuem a maior informação relacionada à probabilidade de o beneficiário cancelar o plano de saúde. Juntas, essas três carregam quase 60% de toda a variabilidade dos dados.

Utilizando esses oito atributos, os modelos foram construídos e avaliados segundo os critérios já mencionados.

**Tabela 6 - Peso das variáveis selecionadas para modelo final**

Variáveis	Estatística Qui-Quadrado	% Peso
ANO_INCLUSAO	15202,2504	22,02%
MEDIA_PAGO_MENSALIDADE	13585,5611	19,68%
CLASSE_TEMPO_REL	12552,3357	18,18%
TOTAL_PAGO_MENSALIDADE	11508,4016	16,67%
TEMPO_PRIMEIRO_USO	6681,1936	9,68%
CONTAGEM_SPSADT	3451,476	5,00%
CONTAGEM_DESPESASMEDICAS	3275,5741	4,74%
CUSTO_TOTAL_CONSULTA	2789,4491	4,04%

Fonte: Autoria própria.

Como é habitual em *data mining*, a base de dados foi dividida em dois conjuntos de dados: treino (60%, n=12.644) e validação (40%, n=8.430). Optou-se, nesta fase, por não considerar o terceiro conjunto, o de teste, tendo em vista dispor de mais observações para treino dos modelos.

O modelo de árvore de decisão (C4.5) foi desenvolvido com base no critério de decisão de redução da entropia, ou seja, a escolha do atributo de partição é feita considerando o ganho de informação. A árvore consegue prever corretamente em cerca de 99,25% dos casos no conjunto de validação, e obteve curva ROC de 0,993. Assim, observa-se que a exatidão do modelo é boa (acima de 90%) e se comporta bem com dados desconhecidos, ou quando dados futuros forem aplicados a ele. A Tabela 7 apresenta a matriz de confusão deste modelo.

**Tabela 7 - Matriz de confusão da árvore de decisão (validação)**

Tabela de classificação		Predito		Percentual
		Não Churn	Churn	
Observado	Não Churn	1964	22	98,89%
	Churn	41	6403	99,36%
Percentual Médio				99,25%

Fonte: Autoria própria.

Dadas as variáveis seleccionadas, o modelo de regressão logística desenvolvido obteve percentual de acerto de 98,49%, e curva ROC de 0,997, tendo, assim, apresentado boas previsões. Comparando este modelo com a árvore de decisão, tem-se que este é inferior no percentual de acerto e ligeiramente superior na curva ROC. A Tabela 8 apresenta a matriz de confusão deste modelo.

**Tabela 8 - Matriz de confusão da regressão logística (validação)**

Tabela de classificação		Predito		Percentual
		Não <i>Churn</i>	<i>Churn</i>	
Observado	Não <i>Churn</i>	1976	10	99,50%
	<i>Churn</i>	117	6327	98,18%
<b>Percentual Médio</b>				<b>98,49%</b>

Fonte: Autoria própria.

Os modelos de rede neurais produziram bons resultados, entretanto o modelo RBF, dentre os outros apresentados, apresentou o pior resultado, tanto no percentual de acerto (92,67%) quanto na curva ROC (0,971). Já o modelo MLP mostrou o segundo melhor resultado dentre os modelos desenvolvidos, tendo 99,12% de acerto e curva ROC de 0,999. As Tabelas 9 e 10 mostram a matriz de confusão destes modelos RBF e MLP, respectivamente.

**Tabela 9 - Matriz de confusão do RBF (validação)**

Tabela de classificação		Predito		Percentual
		Não <i>Churn</i>	<i>Churn</i>	
Observado	Não <i>Churn</i>	1896	90	95,47%
	<i>Churn</i>	528	5916	91,81%
<b>Percentual Médio</b>				<b>92,67%</b>

Fonte: Autoria própria.



**Tabela 10 - Matriz de confusão do MLP (validação)**

Tabela de classificação		Predito		Percentual
		Não <i>Churn</i>	<i>Churn</i>	
Observado	Não <i>Churn</i>	1984	2	99,90%
	<i>Churn</i>	72	6372	98,88%
<b>Percentual Médio</b>				<b>99,12%</b>

Fonte: Autoria própria.

Ao final desta seção, os resultados de um comitê de classificação foram apresentados, atingindo o objetivo específico de examinar modelos de classificação de risco de ruptura de clientes.

### 7.5 AVALIAÇÃO (*EVALUATION*)

Depois do desenvolvimento dos quatro modelos, observa-se que eles são robustos, sendo os seus comportamentos semelhantes uns aos outros. Sabe-se que um bom modelo de previsão não é só o que apresenta uma grande porcentagem de previsões corretas, mas sim um modelo que consiga prever bem tanto a classe *churn* como a não *churn*.

Analisando a Tabela 11, pode-se concluir que a principal técnica que se adequou melhor às condições de previsão de clientes em planos de saúde foi árvore de decisão, seguida pela técnica de redes neurais MLP, que por sua vez prevê *churn* = 0 melhor que a árvore de decisão. A decisão de escolha se deu pelos critérios definidos na seção 7.1 – Entendimento do Negócio, assumindo que a curva ROC dos dois modelos está muito próxima e que não existe diferença expressiva de um valor para o outro (diferença de 0,006).

**Tabela 11 - Resultados dos modelos de previsão**

Modelos	Previsão correta	Previsão correta (Churn = 0)	Previsão correta (Churn = 1)	ROC	Classificação
C4.5	99,25%	98,89%	99,36%	0,993	1
Regressão logística	98,49%	99,50%	98,18%	0,997	3
RBF	92,67%	95,47%	91,81%	0,971	4
MLP	99,12%	99,90%	98,88%	0,999	2

Fonte: Autoria própria.

Desta maneira, entende-se que, de posse das variáveis sugeridas no capítulo 5, é possível desenvolver modelos capazes de apresentar bons resultados de predição quanto à ruptura de clientes de um plano de saúde. Assim, algumas variáveis propostas foram avaliadas a partir de dados reais com técnicas de previsão.

Observando o resultado obtido pela regressão logística, foi possível identificar a direção de cada variável em relação à variável dependente, conforme apresentado na Tabela 12. Referindo-se à variável “ano de inclusão na base”, tem-se que quanto mais recente for o contrato, maior será a probabilidade de abandonar o plano de saúde. Quando se verifica o tempo decorrido, em dias, até o primeiro uso do plano de saúde, entende-se que a probabilidade de abandonar diminui à medida que os dias passam.

**Tabela 12 - Direções das variáveis**

Variáveis	Direção
ANO_INCLUSAO	↑
MEDIA_PAGO_MENSALIDADE	↑
CLASSE_TEMPO_REL	↓
TOTAL_PAGO_MENSALIDADE	↑
TEMPO_PRIMEIRO_USO	↓
CONTAGEM_SPSADT	↓
CONTAGEM_DESPESASMEDICAS	↓
CUSTO_TOTAL_CONSULTA	↑

Fonte: Autoria própria.

A cada acréscimo de exames feitos, a probabilidade de cancelar o contrato com o plano de saúde diminui. Na mesma direção, o aumento de procedimentos realizados, não só exames, para o beneficiário, também diminui a probabilidade de *churn*. Quando o custo total das consultas aumenta, a probabilidade de romper com o plano de saúde também aumenta, o que se julga uma reação natural do consumidor para um eventual aumento de preço, por exemplo. Tal comportamento se assemelha para cada acréscimo no valor médio pago por mensalidade e no total pago em mensalidades.

Interpretando a direção do atributo que representa as classes do tempo de relacionamento, tem-se que quanto mais tempo o beneficiário passa no plano de saúde, sua probabilidade de *churn* diminui.

Finalizando esta etapa, constatou-se que as fases anteriores foram consideradas satisfatórias e os resultados atingidos suprem os objetivos definidos na seção 7.1. Assim, entende-se que o projeto está concluído, conforme manual do CRISP-DM, descrito na seção 6.3. Nesta seção, o objetivo específico de identificar o principal modelo que se adequa às condições de previsão de classificação de ruptura de clientes foi atingido.

## 7.6 IMPLEMENTAÇÃO (*DEPLOYMENT*)

Diante dos resultados apresentados, a implementação desta pesquisa se dá na previsão de *churn* para clientes reais, não presentes na base de treino. Assim, a empresa de plano de saúde estudada terá informações para combater o *churn* deliberado (definido na seção 2.1), dado que os clientes com maior probabilidade de romperem com a empresa num futuro próximo estariam sendo indicados.

Uma estratégia dirigida com abordagem proativa da empresa (definida na seção 3.1) poderia fazer com que esses clientes não cancelassem seus contratos. Assim, a empresa do plano de saúde pode oferecer incentivos para a não efetivação dessa ruptura.

## 7.7 ANÁLISE FINANCEIRA DO NEGÓCIO

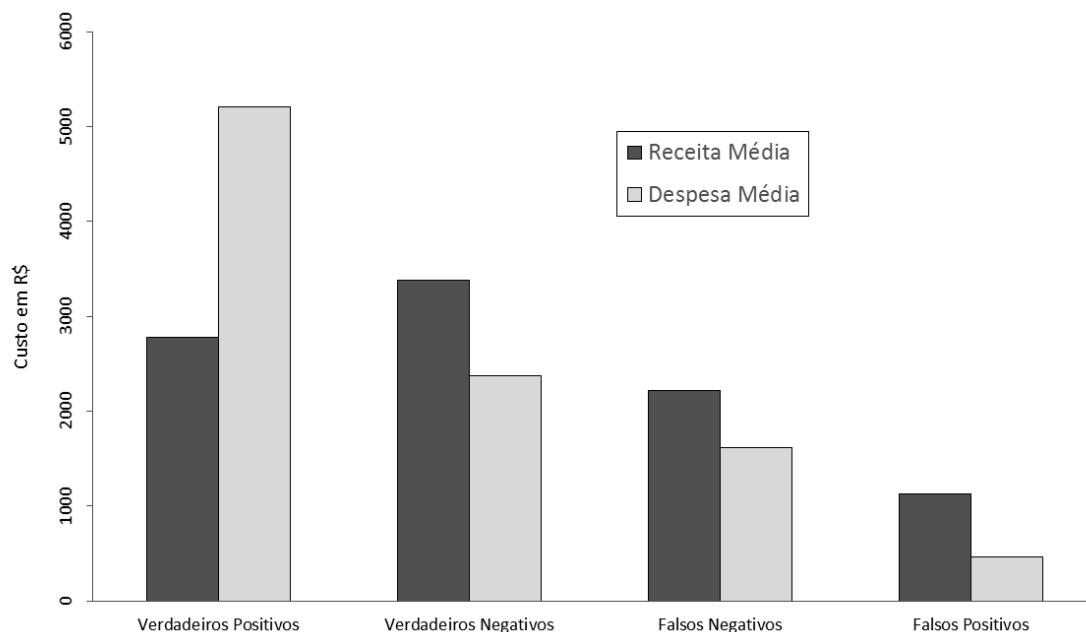
Sob o ponto de vista do negócio, pode-se verificar as receitas e despesas de beneficiários que romperam com a operadora. Se o beneficiário era prejudicial ao negócio da operadora, a ruptura dele é algo que aumenta a rentabilidade da empresa; neste mesmo sentido, a saída de um beneficiário superavitário à operadora de plano de saúde (boa

carteira de cliente) reduz a rentabilidade da empresa.

O sucesso de uma operadora de plano de saúde se dá ao conhecimento da carteira de clientes para agir (MENDES, 2008). Sob o ponto de vista de implicações gerenciais, a operadora pode não agir quando um “péssimo” cliente rompe ou pode “oferecer benefícios” quando o “bom” cliente rompe. As ações preventivas antes do rompimento podem fazer a diferença na retenção desses “bons” e “maus” beneficiários, consequentemente, impactar na rentabilidade da empresa.

A Figura 7 apresenta o custo médio (despesa e receita) por beneficiário da análise de risco apresentada na seção 7.4, conforme resultado do principal modelo (árvore de decisão). Os beneficiários que romperam com a empresa e o modelo previram corretamente (verdadeiro positivo), em média, que estes eram mais prejudiciais ao negócio da operadora do que rentáveis. Estima-se um prejuízo total de R\$ 15.601.060,26 para os 6.403 beneficiários corretamente classificados como *churn*. Assim, indicar previamente e de maneira correta tais beneficiários faria com que a operadora evitasse um investimento desnecessário em ações de retenção para "péssimos" beneficiários.

**Figura 7 - Custos médios (despesa e receita) por beneficiário da análise de risco**



Fonte: Autoria própria.

Analisando os beneficiários que continuam na empresa e o modelo previu corretamente esta manutenção (verdadeiro negativo), pode-se observar que são

beneficiários, em média, mais rentáveis ao negócio da operadora. Calcula-se que o lucro para cada beneficiário, em média, é de R\$ 1.004,08. Assim, estima-se um ganho total de R\$ 1.972.006,73 para os 1964 beneficiários corretamente classificados.

Como a quantidade de beneficiários classificados erroneamente é pequena (total de 63 em 8430) e estes são rentáveis para a operadora, em termos de valor monetário, não faz sentido a operadora realizar qualquer intervenção para estimular a retenção. Por outro lado, percebe-se que a operadora é muito bem administrada em termos de avaliação da carteira de cliente, tendo em vista que os clientes mais custosos rompem com a operadora. Tal entendimento se dá pelo fato de contratos empresariais poderem ser cancelados pela própria operadora de plano de saúde, o que não é verdade em planos individuais, onde só quem pode cancelar seu contrato é o próprio beneficiário.

Acredita-se que medidas para acelerar ou facilitar a saída de clientes que estão dando prejuízos sejam ações que se adequam ao cenário atual da empresa, pois a operadora deve sempre buscar possuir uma boa carteira de clientes (receita maior que despesa).

## CONSIDERAÇÕES FINAIS

Percebendo que o tema retenção de clientes tem sido de bastante interesse das empresas, é de elevada relevância o desenvolvimento de técnicas capazes de detectar e antecipar a possível ruptura dos clientes em planos de saúde. O tema de ruptura de cliente tem ganhado visibilidade nos últimos anos, tanto no meio acadêmico como no meio empresarial. O *churn* tem sido estudado com mais vigor na área de telecomunicações, sendo possível achar um número grande de referências acadêmicas sobre o assunto.

Para que sejam possíveis as detecções e ações preventivas das empresas, uma base de dados com variáveis capazes de render boas previsões e técnicas adequadas para o contexto se faz necessário. Deste modo, o principal objetivo desta dissertação de mestrado foi classificar o risco de ruptura de clientes (beneficiários) de operadora de plano de saúde privado. Após sugestões de variáveis (vide capítulo 5), foi possível utilizar um comitê de classificadores em uma base de dados real com algumas variáveis sugeridas, cedida por um plano de saúde de abrangência nacional, com 21.074 registros (beneficiários), perfazendo o total de mais de 741 mil de *data points*.

Em um segundo momento do estudo, para efetuar a análise dos dados, adotou-se a metodologia de CRISP-DM, para direcionar e organizar as etapas de construção da base e desenvolvimento de modelos preditivos. As variáveis mais significativas, segundo a estatística de Qui-Quadrado, foram selecionadas para a construção dos modelos. São elas: ano de inclusão na base; valor médio pago em mensalidades; classe do tempo de relacionamento; total pago em mensalidades; tempo decorrido até o primeiro uso do plano; quantidade e despesas de serviço profissional ou serviço auxiliar de diagnóstico e terapia realizados; e o total de custo de consulta.

Em seguida, quatro modelos de previsão foram propostos (árvore de decisão, regressão logística e dois de redes neurais, RBF e MLP), avaliados e comparados uns com os outros. É possível afirmar que o modelo eleito (árvore de decisão) pode ser eficiente para determinar o *churn* em plano de saúde a partir das variáveis utilizadas. Ademais, a direção de influência de cada variável foi indicada.

O modelo de árvore de decisão mostrou uma taxa de acerto geral de 99,25%, taxas de acerto dos grupos individuais altas e curva ROC de 0,993, indicando uma consistência na previsão de qualquer um dos dois grupos. Observa-se que o modelo tem uma matriz de confusão mais “limpa”, ou seja, possui valores baixos de falso negativo (observado *churn* e previsto não *churn*) e falso positivo (observado não *churn* e previsto

*churn*), tornando-se um ótimo recurso para melhorar resultados de uma campanha de retenção de clientes.

Ao investir os recursos da operadora do plano de saúde nos clientes identificados e classificados pelo modelo eleito, a empresa seria capaz de agir de forma proativa para reter ou facilitar a saída do beneficiário do plano de saúde, visto que os resultados obtidos indicam que o modelo é bastante eficiente em discriminar beneficiários que provavelmente irão cancelar o contrato dos que não vão cancelar.

Verificou-se que os beneficiários classificados como *churn* e romperam com a operadora (verdadeiros positivos) trouxeram um prejuízo estimado para a empresa em cerca de R\$ 15 milhões, enquanto que os beneficiários classificados como não *churn* e não romperam com a operadora (verdadeiros negativos) trouxeram um lucro estimado em cerca de R\$ 2 milhões.

Para a operadora de plano de saúde em estudo, não faz sentido a operadora se esforçar para estimular a retenção, pois os beneficiários identificados como candidatos ao abandono não são rentáveis para o negócio. Desta forma, a operadora poderia facilitar ou incentivar a ruptura desses beneficiários, com o propósito de ter uma carteira apenas com clientes rentáveis.

Apesar dos resultados obtidos nesta dissertação, deve-se salientar as suas limitações. Uma das principais preocupações iniciais do estudo de caso foi, sem dúvida, a limpeza de dados. A maioria das variáveis obtidas na base de dados real que representava datas estava com problemas de consistência.

Outra limitação é que apenas três técnicas de previsão foram utilizadas. Assim, isso limita o resultado, pois outras técnicas poderiam apresentar melhores resultados. Assim, propõe-se o desenvolvimento de um comitê de classificadores na seção 7.3.

Esta dissertação se limitou a estudar apenas uma operadora de plano de saúde. Deste modo, obtendo dados de somente uma pequena parcela de todo o mercado de plano de saúde, não se pode generalizar os resultados para o setor. Além disso, esta operadora em questão só possuía beneficiários com plano coletivo empresarial, com cobertura nacional e sem opcional de emergência e emergência aérea.

Ressalta-se que, por causa da regulamentação imposta pela ANS, as operadoras de planos de saúde possuem apenas variáveis exigidas pelo órgão regulador, não alterando as variáveis. Apenas em casos excepcionais é possível conseguir de uma operadora variáveis não exigidas pela ANS, como é o caso do trabalho de Mendes (2008)

e, por exemplo, as variáveis tempo de exame e tempo de consulta. Diante disto, uma das principais limitações desse estudo refere-se ao fato de não usar outras variáveis que a literatura trata pelo fato de indisponibilidade da operadora em questão.

Para maior aprofundamento, é sugerido que sejam feitas pesquisas visando entender o comportamento dos clientes e prever sua ruptura, por meio do máximo de variáveis sugeridas nesta dissertação. Assim, seria possível registrar o relacionamento do serviço oferecido pelo plano de saúde e seus beneficiários. Ademais, investigar ex-beneficiários do plano de saúde e entender os principais motivos de seu abandono seria objeto de uma pesquisa futura.

Outro ponto para se pesquisar seria utilizar outras formas de se encontrar o escore de *churn*, podendo-se utilizar um comitê de classificadores com as seguintes técnicas: algoritmo genético, análise de *cluster*, análise discriminante, SVM, lógica *fuzzy*, outros tipos de redes neurais (e.g. SOM e *neuro-fuzzy*), entre outras já citadas nesta dissertação.

Conforme apresentado no final do capítulo 4, a técnica de PSM é ferramenta a mais na avaliação das técnicas do comitê classificador para regressão logística. Desta forma, será utilizada para encontrar e equiparar elementos com características semelhantes, com resultados contrários. Em outras palavras, encontrar beneficiários que se tornaram *churn* e beneficiários que não se tornaram. Executando esta equiparação, é possível definir dois grupos similares, a fim de comparar seus perfis e mostrar o efeito médio do tratamento.

Apesar dos resultados obtidos, sugerem-se melhorias no que diz respeito à tentativa de diminuição da classificação de Falsos Positivos, ou seja, beneficiários que não são *churners* e que foram classificados como tal. Desta forma, esta melhoria seria capaz de, ocasionalmente, evitar gastos “desnecessários”.

Pesquisas futuras poderiam não só indicar quem vai romper ou não com a empresa, mas também apontar qual seria a ação com maior probabilidade de efetividade. Por exemplo, indicar que um beneficiário vai cancelar seu contrato com o plano de saúde e informar que a ação proativa com maior probabilidade de sucesso deste cliente é oferecer um opcional de emergência. Em resumo, a análise de qual o melhor canal para comunicar com o cliente para ações de *marketing*.

Sugere-se fazer um estudo futuro que trabalhe com janelas temporais de *churners*, ou seja, estimar, para um intervalo de meses, clientes que estão mais propensos ao abandono e, se julgar necessário, realizar ações preventivas sobre eles. Em seguida,



deverá ser feita uma avaliação do resultado da ação realizada e verificar se a técnica ou comitê classificador produziu bons resultados de previsão.

Para finalizar, os pontos de melhorias e trabalhos futuros ressaltados anteriormente são meramente um exemplo do que poderá ser feito, mas que destacam a oportunidade e o espaço para melhorias que este trabalho representa.

## REFERÊNCIAS

- ABBASIMEHR, Hossein; SETAK, Mostafa; SOROOR, Javad. A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques. **International Journal of Production Research**, v. 51, n.4, p. 1279-1294, 2013.
- ABBASIMEHR, Hossein; SETAK, Mostafa; TAROKH, Mohammad. A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction. **The International Arab Journal of Information Technology**, v. 11, n. 6, p. 599-606, 2014.
- ALVES, S. L. Eficiência das Operadoras de Planos de Saúde. **Revista Brasileira de Risco e Seguro**, v. 4, n. 8, p. 87-112, 2008.
- ANDRADE, D. **Uma análise de cancelamentos em telefonia utilizando mineração de dados**. 67 fls. Dissertação (mestrado em Engenharia Civil) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2007.
- BRASIL. Agência Nacional de Saúde Suplementar. **Perfil do setor: Dados e Indicadores do Setor, Dados Gerais, Normas mais acessadas**. Disponível em: <<http://www.ans.gov.br/perfil-do-setor/dados-e-indicadores-do-setor>>. Acesso em: 10 mar. 2016.
- ASAARI, M. H. A. H.; KARIA, N. Churn Management towards Customer Satisfaction: A Case of Cellular Operators in Malaysia. In: **The International Conference on eCommerce: Emerging Trends in Electronic Commerce (ETEC2000)**, November, Kuala Lumpur, 21-23, 2000.
- ASBRAND, D. Is your automated customer service killing you? **Datamation**, v. 43, n. 5, p. 62-67, 1997.
- AU, T.; MA, G. Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques. **Journal of Comparative International Management**, v. 6, n. 1, p. 10-23, 2003.
- AU, W; CHAN, CC; YAO, X.A. Novel evolutionary data mining algorithm with applications to churn prediction. **IEEE Transactions on Evolutionary Computation**, v. 7, n. 6 p. 532-545, 2003.
- BAGOZZI, Richard P. Marketing as an organized behavioral system of exchanges. **Journal of Marketing**, v. 38, n. 4, p. 77-81, 1974.
- BARTELS, Robert D. W. The General Theory of Marketing. **Journal of Marketing**, v.32, n.1, p. 29-33, 1968.
- BERRY, J. Database Marketing. **Business Week**, v. 5, Setembro, p. 56-62, 1994.

- BERRY, L. Emerging perspectives on services marketing. Chicago: **American Marketing Association**, v. VI, p. 146, 1983.
- BERRY, M. J. A.; LINOFF, G. **Data mining techniques**: for marketing, sales, and customer support. NY: John Wiley & Sons: 1997.
- \_\_\_\_\_. **Mastering Data Mining**: The art and science of customer relationship management. New York: Wiley Computer Publishing, 2000.
- BOTELHO, Delane; TOSTES, Frederico Damian. Modelagem de probabilidade de Churn. **Revista de Administração de Empresas**, v. 50, n. 4, p. 396-410, 2010.
- BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais**: Teoria e Aplicações. 2a ed., Rio de Janeiro: LTC, 2011.
- BUCKNIX, W.; VAN DEN POEL, D. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. **European Journal of Operational Research**, v. 164, n.1, p.252–268, 2005.
- BUREZ, Jonathan; VAN DEN POEL. D. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. **Expert Systems with Applications**, v. 32, n. 2, p.277-288, 2007.
- \_\_\_\_\_. Separating financial from commercial customer churn: a modeling step towards resolving the conflict between the sales and credit department. **Expert Systems with Applications**, v. 35, n. 1-2, p. 497–514, 2008.
- CAMPOS, Carla da Costa. Um estudo das relações entre operadoras de planos de assistência à saúde e prestadoras de serviço. **O Mundo da Saúde São Paulo**, v. 30, n. 2, p. 228-238, 2006.
- CHAPMAN, Pete *et al.* **CRISP-DM 1.0** – Step-by-Step data mining guide. CRISPDM Consortium, 2000.
- CHEN, M. DEY, D. Variable selection for multivariate logistic regression models. **Journal of Statistical Planning and Inference**, v. 111, n.1-2 p. 37-55, 2003.
- CISTER, A. M. **Mineração de dados para a análise de atrito em telefonia móvel**. 2005. 158. fls. Tese (Doutorado em Engenharia). Rio de Janeiro, 2005. Programa de Pós-Graduação em Engenharia Civil, COPPE-Civil. Universidade Federal do Rio de Janeiro.
- CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada para os cursos de administração, ciências contábeis e economia**. São Paulo: Atlas, 2007.
- COUSSEMENT, Kristof; BENOIT, Dries F.; VAN DEN POEL, Dirk. Improved marketing decision making in a customer churn prediction context using generalised additive models. **Expert Systems with Applications**, v. 37, n. 3, p. 2132-2143, 2010.

COUSSEMENT, Kristof; VAN DEN POEL, Dirk. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. **Expert Systems with Applications**, v. 34, n. 1, p. 313-327, 2008a.

\_\_\_\_\_. Integrating the voice of customers through call center emails into a decision support system for churn prediction. **Information and Management**, v. 45, n. 3, p. 164-174, 2008b.

DARÉ, P. R. C. **Retenção de clientes à luz do gerenciamento de Churn: um estudo no setor de telecomunicações**. 2007. 166 fls. Dissertação (Mestrado em Administração) – Universidade de São Paulo, 2007.

DAWES, J.; SWAILES, S. Retention sans frontieres: issues for financial service retailers. **International Journal of Bank Marketing**, v. 17, n. 1, p. 36–43, 1999.

DAY, G.S. **A empresa orientada para o Mercado: compreender, atrair e manter clientes valiosos**. Porto Alegre: Bookman, 2001.

DESOUZA, G. Designing a Customer Retention Plan. **The Journal of Business Strategy**, v. 13, n. 2, p. 24-28, 1992.

ENGEL, J. F.; BLACKWELL, R. D.; MINIARD, P. W. **Comportamento do consumidor**. Rio de Janeiro: Livros técnicos e Científicos Editora S. A., 2000.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **American Association for Artificial Intelligence**, v.17, n. 3, p. 37-54, 1996.

FIGUEIREDO, John M. de; SILVERMAN, Brian S. Churn, Baby, Churn: Strategic Dynamics Among Dominant and Fringe Firms in a Segmented Industry. **Management Science**, v. 53, n. 4, p. 632–650, 2007.

GALVÃO, Marcela Squires; GONZALEZ, Mario Orestes Aguirre. Análise da utilização do churn em uma empresa de telecomunicações. In: Encontro Nacional de Engenharia De Produção, 31, 2011, Minas Gerais. **Anais...** Belo Horizonte: ENEGEP, 2011.

GANESH, J.; ARNOLD, M.; REYNOLDS, K. Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers. **Journal of Marketing**, v. 64, n. 3, p. 65-87, 2000.

GARCIA, S. O. **Uso de árvores de decisão na descoberta de conhecimento na área da saúde**. 2003. 87 fls. Dissertação (Mestrado em Computação), Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

GIL, A. C. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010.

GLADY, Nicolas; BAESENS, Bart; CROUX, Christophe. Modeling churn using customer lifetime value. **European Journal of Operational Research**, v. 197, n. 1, p. 402-411, 2009.

GOMES, B. M. V. **Previsão de churn em companhias de seguros**. Braga: Universidade do Minho, 2011. 113p. Dissertação (Mestrado em Engenharia Informática) - Universidade do Minho, Portugal, 2011.

GORDON, I. **Marketing de relacionamento: estratégias, técnicas e tecnologia para conquistar clientes e mantê-los para sempre**. São Paulo: Futura, 1998.

GRÖNROOS, C. **Marketing: Gerenciamento e Serviços: a competição por serviços na hora da verdade**. Rio de Janeiro: Campus, 1995.

GUSTAFSSON, Anders; JOHNSON, Michael D.; ROOS, Inger. The Effects of Customer Satisfaction, Relationship Commitment Dimensions, and Triggers on Customer Retention. **Journal of Marketing**, v. 69, n. 4, p. 210-218, 2005.

HADDEN, J. *et al.* Churn Prediction: Does Technology Matter? **International Journal of Intelligent Systems and Technologies**, v. 1, n. 2, p. 104-110, 2006.

\_\_\_\_\_. Computer assisted customer churn management: State-of-the-art and future trends. **Computers & Operations Research**, v. 34, n. 10, p. 2902–2917, 2005.

HAIR, F.J. *et al.* **Análise multivariada de dados**. 6a ed. Porto Alegre: Bookman, 2009.

HALL, M. *et al.* The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v. 11, issue 1, 2009.

HOFFMAN, K. D.; BATESON, J. E.G. **Essentials of Services Marketing**. Texas: The Dryden Press, 1997.

HOLTZ, Herman. **Databased marketing**. São Paulo: Makron, 1994.

HONGXIA, M.; MIN, Q.; JIANXIA, W. Analysis of the Business Customer Churn Based on Decision Tree Method. In: THE CONFERENCE ON ELECTRONICS MEASUREMENT & INSTRUMENTS, 9, 2009, Beijing. Anais... Beijing: China, 2009.

HOPFIELD, J.J. Neural networks and physical systems with emergent collective computational abilities. **Proceedings of National Academy of Sciences**, v. 79, n. 8, p. 2554-2558, 1982.

HOWARD, J.R.; SHETH, J.N. **The Theory of Buyer Behavior**. New York: John Wiley and Sons, 1969.

HUANG, Bingquan; BUCKLEY, B.; KECHADI, T.-M. Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. **Expert Systems with Applications**, v. 37, n. 5, p. 3638-3646, 2010.

- HUGHES, Arthur M. **Data Base Marketing Estratégico**. São Paulo: Makron Books, 1998.
- HUNG, S.; YEN, D.; WAG, H. Applying data mining to telecom churn management. **Expert Systems with Applications**, v. 31, n. 3, p. 515-524, 2006.
- HUNT, Shelby D.; LAMBE, C. J.; WITTMANN, C. M. A theory and model of business alliance success. **Journal of Relationship Marketing**, v. 1, n. 1, p. 17-35, 2002.
- HUNT, Shelby D. General theories and the fundamental explananda of Marketing. **Journal of Marketing**, Texas, v. 47, n. 4, p. 9- 17, 1983.
- JACOB, R. Why some customers are more equal than others. **Fortune**, v. 130, n. 6, p. 215-224, 1994.
- JAHROMI, Ali Tamaddoni *et al.* Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. **Journal of Strategic Marketing**, v. 18, n. 7, p. 587-598, 2010.
- KAKWANI, N; NERI, M. C; SON, H. Linkages between pro-poor growth, social programs and labor market: the recent Brazilian experience. **World Development**, v. 38, n. 6, p. 881-884, 2010.
- KARAOCA, Adem; KARAOCA, Dilek. GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system. **Expert Systems with Applications**, v. 38, n. 3, p. 1814-1822, 2011.
- KISIOGLU, Pinar; TOPCU, Y. Ilker. Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. **Expert Systems with Applications**, v. 38, n. 6, p. 7151-7157, 2011.
- KNOWLES, A. Get the complete picture. **Datamation**, v. 43, n. 10, p. 74-79, 1997.
- KNOX, George; OEST, Rutger van. Customer Complaints and Recovery Effectiveness: A Customer Base Approach. **Journal of Marketing**, v. 78, n. 5, p. 42-57, 2014.
- KOTLER, Philip. A Generic Concept of Marketing. **Journal of Marketing**, v. 36, n. 2, p. 46-54, 1972.
- \_\_\_\_\_. Marketing during periods of shortage. **Journal of Marketing**, v. 38, n. 3, p. 20-29, 1974.
- KOTLER, P.; ARMSTRONG, G. **Princípios de Marketing**. 7ª ed. Rio de Janeiro: LTC, 1999.
- KUMAR, D.; RAVI, V. Predicting credit card customer churn in banks using data mining. **International Journal of Data Analysis Techniques and Strategies**, v. 1, n.1, p. 4-28, 2008.

KURTZ, D. L.; CLOW, K. E. **Services marketing**. New York: John Wiley & Sons, 1998.

LAMBRECHT, Anja; SKIERA, Bernd. Paying Too Much and Being Happy About It: Existence, Causes, and Consequences of Tariff-Choice Biases. **Journal of Marketing Research**, v.43, n.2, p. 212-223, 2006.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. Hoboken, NJ: Wiley, 2005.

\_\_\_\_\_. **Data Mining Methods and Models**. Hoboken, NJ: Wiley, 2006.

LAST, W.; KANDEL, A.; BUNKE, H. Data Mining in Time Series Databases. Series in Machine Perception and Artificial Intelligence, **World Scientific Pub Co Inc**, v. 57, p. 192, Singapore, 2004.

LAZAROV, V.; CAPOTA, M. **Churn prediction, Business Analytics Course**. Munique: TUM Computer Science, 2007.

LEJEUNE, M. Measuring the impact of data mining on churn management. **Internet Research**, v. 11, n. 6, p. 375-387, 2001.

LEAL, R. M., MATOS, J. B. B de. Planos de saúde: uma análise dos custos assistenciais e seus componentes. **Rev. adm. Empresa - RAE**. vol. 49, n. 4, pp. 447-458, 2009.

LEMMENS, Aurélie; CROUX, Christophe. Bagging and boosting classification trees to predict churn. **Journal of Marketing Research**, v. 43, n. 2, p. 276-286, 2006.

LEMOS, Eliane P. **Análise de Crédito Bancário com o uso de Data Mining: Redes Neurais e Árvores de Decisão**. 2003. 147 f. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação, Curitiba, 2003.

LEWIS, Michael. The Influence of Loyalty Programs and Short-Term Promotions on Customer Retention. **Journal of Marketing Research**, v.41, n. 3, p. 281-292, 2004.

LEWIS, R. An introduction to classification and regression tree (CART) analysis. **Annual Meeting of the Society for Academy Emergency Medicine**, San Francisco, California, 2000.

MAIMON, Oded; ROKACH, Lior. **Data Mining and Knowledge Discovery Handbook**. Ed. 2, New York: Springer, 2010.

MALHOTRA, N. **Pesquisa de marketing: uma orientação aplicada**. Porto Alegre: Bookman, 2012.

McCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 5, n.4, p.115-133, 1943.

- MENDES, V. de P. **Modelos de churn de clientes em planos de saúde**. Niterói: Universidade Federal Fluminense, 2008. 113p. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal Fluminense, Niterói, Brasil, 2008.
- MICHALSKI, S. Types of customer Relationship ending processes. **Journal of Marketing Management**, v. 20, n. 9-10, p. 977-999, 2004.
- MINSKY, M. L.; PAPERT, S. A. **Perceptrons**. Oxford: MIT press, 1969.
- MOEYERSOMS, J.; MARTENS, D. Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. **Decision Support Systems**, 72, p. 72–81, 2015.
- MORGAN, Robert M.; HUNT, Shelby D. The commitment-trust theory of relationship marketing. **Journal of Marketing**, v. 58, n. 3, p. 20-38, Jul., 1994.
- MORIK, K.; KOPCKE, H. Analysing Customer Churn in Insurance Data – A Case Study. In: BOULICAUT, Jean-Francois (Ed.). **Proceedings of the Eight European Conference on Principles and Practice of Knowledge Discovery in Databases**. Vol. 3202, Pisa, Itália: Springer, 2004. p. 325-336.
- NESLIN, S. A. *et al.* Defection detection: Measuring and understanding the predictive accuracy of customer churn models. **Journal of Marketing Research**, v. 43, n. 2, p. 204-211, 2006.
- NIE, Guangli *et al.* Credit card churn forecasting by logistic regression and decision tree. **Expert Systems with Applications**, v. 38, n. 12, p. 15273-15285, 2011.
- NITZAN, Irit; LIBAI, Barak. Social Effects on Customer Retention. **Journal of Marketing**, v. 75, n. 6, p. 24-38, 2011.
- OLIVER, R.W. *et al.* Leveraging the value of customer satisfaction information. **Journal of Health Care Marketing**, v. 14, n. 3, p. 16-20, 1994.
- OWCZARCZUK, Marcin. Churn models for prepaid customers in the cellular telecommunication industry using large data marts. **Expert Systems with Applications**, v. 37, n. 6, p. 4710-4712, 2010.
- PASTANA, Greice Kelly Bussola. Benefícios motivacionais em empresas de saúde suplementar: estudo de caso de duas empresas do oeste paulista. **Omnia Saúde**, v. 9, n. 1, p. 23-37, 2012.
- PAULIN, M. *et al.* Relational norms and client retention: External effectiveness of commercial banking in Canada and Mexico. **International Journal of Bank Marketing**, v. 16, n. 1, p. 24-31, 1998.
- PEDRON, Cristiane Drebes. **Variáveis Determinantes no Processo de Implantação de CRM: Estudo de Casos Múltiplos em Empresas Gaúchas**. São Leopoldo: Unisinos, 2001.



PENDHARKAR, Parag C. Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. **Expert Systems with Applications**, v. 36, n. 3, p. 6714-6720, 2009.

PEPPARD, J. Customer relationship management (CRM) in financial services. **European Management Journal**, v.18, n. 3. p. 312–327, 2000.

PETERMANN, R. **Modelo de mineração de dados para classificação de clientes em telecomunicações**. 2006. 164 fls. Dissertação (Mestrado em Engenharia Elétrica). Universidade Católica do Rio Grande do Sul, Porto Alegre, 2006.

PHADKE, Chitra *et al.* Prediction of Subscriber Churn Using Social Network Analysis. **Bell Labs Technical Journal**, v. 17, n. 4, p. 63-76, 2013.

PINTO, Luiz Felipe; SORANZ, Daniel Ricardo. Planos privados de assistência à saúde: cobertura populacional no Brasil. **Ciênc. Saúde Coletiva**, vol. 9, n. 1, pp. 85-98, 2004.

PIVA, L. C. *et al.* Relação entre satisfação, retenção e rentabilidade de clientes no setor de planos de saúde. **Revista de Ciências da Administração**, v. 9, n. 19, p. 54-80, 2007.

PORTER, M. E. **Vantagem competitiva**: criando e sustentando um desempenho superior. Rio de Janeiro: Campus, 1989.

QIAN, Zhiguang; JIANG, Wei; TSUI, Kwok-Leung. Churn detection via customer profile modelling. **International Journal of Production Research**, v. 44, n. 14, p. 2913–2933, 2006.

QUINLAN, J. **C4.5**: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1993.

\_\_\_\_\_. Induction of decision Trees. **Machine Learning**, v. 1, n. 1, p. 81-106, 1986.

\_\_\_\_\_. Simplifying Decision Trees. **International Journal of Man-Machine Studies**, v. 27, n. 3, p. 221-234, 1987.

REICHHELD, F.F. **The loyalty effect**: the hidden force behind growth, profits, and lasting value. Boston, MA: Harvard Business School Press, 1996. 323 p.

\_\_\_\_\_. **A pergunta definitiva**: você nos recomendaria a um amigo? São Paulo: Bain & Company, 2006.

REICHHELD, F.F; MARKEY JR., R.G.; HOPTON, C. The loyalty effect: the relationship between loyalty and profits. **European Business Journal**, v.12, n.3, p.134-139, 2000.

REICHHELD, F.F; SASSER, E. Zero Defections: Quality Comes to Services. **Harvard Business Review**. v. 68, n. 5, p. 105-111, 1990.

REICHHELD, F.F.; TEAL, Thomas. **The Loyalty Effect**. Boston: Harvard Business School Press, 1996.

REINARTZ, W.; THOMAS, J.; KUMAR, V. Balancing acquisition and retention resources to maximize profitability. **Journal of Marketing**, v. 69, n. 1, p. 63-79, 2005.

ROSENBLATT; F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386-408, 1958.

RUST, Roland T.; ZAHORIK, Anthony J. Customer Satisfaction, Customer Retention and Market Share. **Journal of Retailing**, v. 69, n. 2, p.193–215, 1993.

RUST, R. T.; ZEITHAML, V.; LEMON, K. N. **O Valor do Cliente: O Modelo que está Reformulando a Estratégia Corporativa**. Porto Alegre: Bookman, 2001.

RUST, R.T.; ZEITHML, V.A.; LEMON, K.N. Customer-Centered Brand Management. **Harvard Business Review**, v., 82, n. 9, p. 110-118, 2004.

SCHWEIDEL, David A.; FADER, Peter S.; BRADLOW, Eric T. Understanding Service Retention Within and Across Cohorts Using Limited Information. **Journal of Marketing**, v. 72, n.1, p. 82-94, 2008.

SHEARER, C. The CRISP-DM Model: The New Blueprint for DataMining. **Journal of Data Warehousing**, v. 5, n. 4, p. 13-22, 2000.

SHETH, Jagdish N.; PARVATIYAR, Atul. The evolution of relationship Marketing. **International Business Review**, v. 4, n. 4, p. 397-418, 1995.

SHETH. Jagdish; SISODIA, Rajendra. Feeling the Heat-Part 2. **Marketing Management**, v.4, n.3, p. 19-33, 1995.

STONE, Merlin; SHAW, Robert. Database marketing for competitive advantage. **Long Range Planning**, v. 20, n. 2, p. 12-20, 1987.

STROUSE, K. G. **Marketing Telecommunications Services: new approaches for changing environment**. Norwood, MA: Artech House, Inc., 1999.

TSAI, Chih-Fong; LU, Yu-Hsin. Customer churn prediction by hybrid neural networks. **Expert Systems with Applications**, v. 36, n. 10, p. 12547-12553, 2009.

VAFEIADIS, T. *et al.* A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, p.1–9, 2015.

VAN DEN POEL, D.; LARIVIÈRE, B. Customer attrition analysis for Financial services using proportional hazard models. **European Journal of Operational Research**, v. 157, n. 1, p.196-217, 2004.

VAVRA, Terry. G.; PRUDEN, D. R. Using after marketing to maintain a customer base. **Discount Merchandiser**, v. 35, n. 5, p. 86-88, 1995.

VAVRA, Terry G. **Marketing de relacionamento: after marketing**. São Paulo: Atlas, 1994.

VERBEKE, Wouter *et al.* Building comprehensible customer churn prediction models with advanced rule induction techniques. **Expert Systems with Applications**, v. 38, n. 3, p. 2354-2364, 2011.

WEI, C.; CHIU, I. Turning telecommunications call details to churn prediction: a data mining approach. **Expert Systems with Applications**, v. 23, n. 2, p. 103-112, 2002.

WEISBERG, S. **Applied Linear Regression**. New Jersey: Wiley-IEEE, 3<sup>a</sup> ed., 336 p., 2005.

XIE, Yaya; LI, Xiu; NGAI, E.W.T.; YING, Weiyun. Customer churn prediction using improved balanced random forests. **Expert Systems with Applications**, v. 36, n. 3, p. 5445-5449, 2009.

YU, Xiaobing *et al.* An extended support vector machine forecasting framework for customer churn in e-commerce. **Expert Systems with Applications**, v. 38, n.3, p. 1425-1430, 2011.