Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Uma Abordagem para a Caracterização do Cancelamento Eletivo de Contratos em Planos de Saúde Privados

Jefferson Henrique Camelo Soares

Número de Ordem PPGCC: M001

Teresina-PI, 29 de Fevereiro de 2016

Jefferson Henrique Camelo Soares

Uma Abordagem para a Caracterização do Cancelamento Eletivo de Contratos em Planos de Saúde Privados

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Erick Baptista Passos

Teresina-PI

29 de Fevereiro de 2016

Jefferson Henrique Camelo Soares
Uma Abordagem para a Caracterização do Cancelamento Eletivo de Contratos em Planos de Saúde Privados/ Jefferson Henrique Camelo Soares. – Teresina-PI, 29 de Fevereiro de 2016-
60 p. : il. (algumas color.) ; 30 cm.

Orientador: Erick Baptista Passos

Dissertação (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, 29 de Fevereiro de 2016.

1. Palavra-chave1. 2. Palavra-chave2. I. Orientador. II. Universidade xxx. III. Faculdade de xxx. IV. Título

CDU 02:141:005.7

Jefferson Henrique Camelo Soares

Uma Abordagem para a Caracterização do Cancelamento Eletivo de Contratos em Planos de Saúde Privados

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 29 de Fevereiro de 2016:

Erick Baptista Passos
Orientador

Antonio Helson Mineiro Soares
Professor

Pedro de Alcântara dos Santos Neto
Professor

Ricardo Augusto Souza Fernandes
Professor

Teresina-PI
29 de Fevereiro de 2016

"I hold the world but as the world, Gratiano;
A stage where every man must play a part..."
(Shakespeare)

Resumo

Uma diversidade de fatores influencia na expectativa de vida de uma pessoa, e um fator fundamental é o cuidado com a própria saúde. Porém, o cuidado com a saúde não possui um baixo custo; empresas privadas, denominadas Operadoras de Plano de Saúde (OPS), são geralmente responsáveis pelo pagamento das contas de medicamentos, exames médicos, internações e outros custos hospitalares. O funcionamento financeiro estável da OPS está diretamente relacionado à permanência dos beneficiários na empresa e, portanto, inversamente relacionado à quantidade de cancelamentos eletivos desses contratos. O cancelamento eletivo tipifica-se quando o beneficiário decide, de forma deliberada, cancelar o contrato estabelecido com a OPS, fato que pode comprometer a receita desta. O objetivo principal deste trabalho consiste em desenvolver uma abordagem para caracterizar o cancelamento eletivo de contratos em planos de saúde privados. A abordagem proposta é constituída por três fases: Pré-Processamento, Mineração de Dados e Priorização de Contratos. A fase de Pré-Processamento visa garantir uma maior qualidade às informações de contratos extraídos da base de dados de uma OPS real. A fase de Mineração de Dados explora os dados pré-processados com o intuito de descobrir novos conhecimentos, ou seja, padrões, relacionamento entre atributos e tendências ainda não conhecidos pela gestão da OPS. Essa fase de Mineração de Dados é responsável por reconhecer contratos ativos com características de contratos já cancelados, por meio de modelos de classificação; identificar que tipos de ações e comportamentos levam os beneficiários da OPS a cancelarem seus contratos, por meio da análise de árvores de decisão; e estimar o tempo restante até o cancelamento do contrato, por meio de técnicas de regressão. Dessa forma, a gestão da OPS pode interceder de forma proativa no problema do cancelamento, ou seja, anteceder-se à possível saída de um beneficiário da empresa e promover ações que evitem tal evento. A fase de Priorização de Contratos objetiva evidenciar quais contratos apresentam um maior risco de serem cancelados, permitindo que a gestão da OPS possa avaliá-los de forma prioritária. Por fim, é realizado um conjunto de experimentos demonstrando passo-a-passo a execução prática da abordagem proposta, com a apresentação dos resultados e discussões.

Palavras-chaves: mineração de dados. classificação. regressão. saúde suplementar. plano de saúde.

Abstract

A range of different factors influences the life expectancy of a person, and a key factor for this is healthcare. However, this concern with healthcare is not cheap, there is usually a need for some private company, called Health Insurance Provider (HIP), which is generally responsible for payment of medical exams, hospitalizations, medications and other medical costs. The stable financial operation of HIP is directly related to keep beneficiaries in the company and therefore inversely related to elective cancellations of these contracts. Elective cancellation is typified when the beneficiary decides deliberately, cancel the contract with the HIP, which may compromise the company's revenue. The main objective of this work is to develop an approach to characterize the elective cancellation of contracts in private health insurances. The proposed approach consists of three phases: Pre-Processing, Data Mining and Prioritization. The Pre-Processing phase aims to ensure greater quality to contract information extracted from a real HIP database. The Data Mining phase explores the pre-processed data in order to discover new knowledge, which means patterns, relationships between attributes and trends not known by HIP managers. This Data Mining phase is responsible for recognizing through classification models active contracts that share features with contracts already canceled; identify what types of actions and behaviors lead beneficiaries of the HIP to cancel their contracts, through decision tree analysis; and estimate the time remaining until the cancellation of the contract, through regression techniques. Thus, the management of the HIP can intervene proactively in the cancellation problem by preceding to possible cancelattions and promote actions to prevent it. The Prioritization phase has to evidence contracts that have a higher risk of being canceled, allowing the management of HIP to analyse them on a priority basis. Finally, it conducted a set of experiments demonstrating step-by-step practical implementation of the proposed approach by presenting results and discussions.

Keywords: data mining. classification. regression. supplementary health. healthcare insurance provider.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM	Aprendizado de Máquina
ANS	Agência Nacional de Saúde
AUC	*Area Under ROC Curve*
CART	*Classification and Regression Tree*
DCBD	Descoberta de Conhecimento em Banco de Dados
FN	Falsos Negativos
FP	Falsos Positivos
KNN	*K-Nearest Neighbors*
LD	Limiar de Discriminação
MCR	Modelo Classificador Refinado
MLP	*Multilayer Perceptron*
MRR	Modelo Regressor Refinado
OPS	Operadora de Plano de Saúde
RMSE	*Root Mean Square Error*
ROC	*Receiver Operating Characteristic*
SUS	Sistema Único de Saúde
SVM	*Support Vector Machine*
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
WEKA	*Waikato Environment for Knowledge Analysis*

Sumário

Introdução

Contexto e Motivação

Muitos fatores influenciam na vida duradoura de uma pessoa, e um fator fundamental é o cuidado com a saúde física e mental. Mas preservar pela própria saúde tem custos elevados (MARINER, 2014), geralmente se faz necessária a participação de um plano de saúde responsável por pagar as contas hospitalares, seja ele um plano de saúde público ou privado (GOULãO, 2014). As empresas privadas, em países em desenvolvimento, geralmente têm maior senso de responsabilidade com o atendimento prestado, são mais eficientes e autossustentáveis que a iniciativa pública (BASU et al., 2012). Mesmo em países mais desenvolvidos, nos quais os investimentos em saúde pública são elevados, os planos privados possuem uma representação significativa no atendimento à saúde (SEKHRI; SAVEDOFF, 2005).

O Brasil é um exemplo de país em desenvolvimento com um amplo sistema de planos de saúde privados, também conhecido como sistema suplementar de saúde. Informações da Agência Nacional de Saúde Suplementar (ANS, 2015) mostram, desde o ano 2001, um aumento no número de beneficiários[1] associados a Operadoras de Plano de Saúde (OPS). A Figura 1 ilustra uma evolução da quantidade de beneficiários, no Brasil, entre março do ano 2000 e março de 2015.

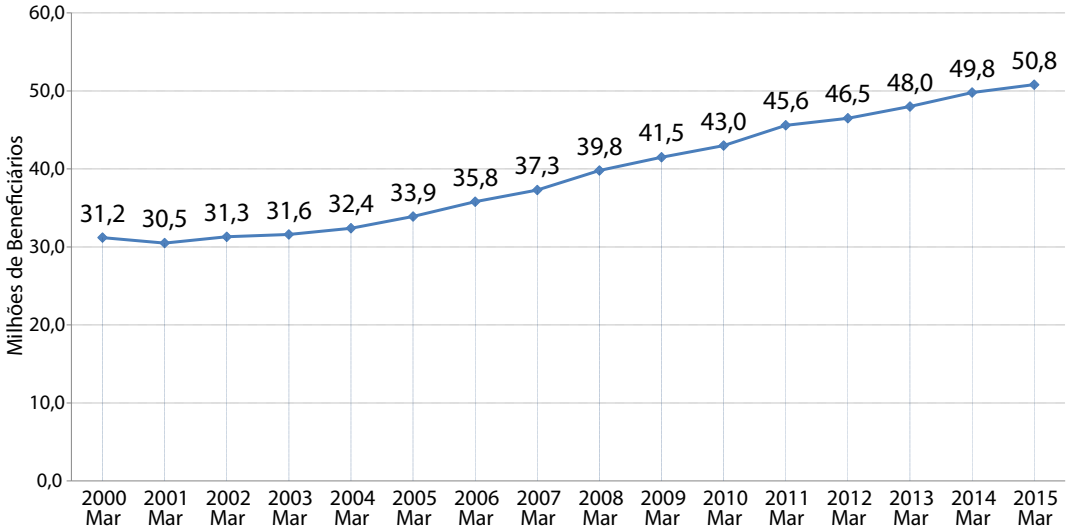

Figura 1 – Quantidade de beneficiários associados a planos privados entre março de 2000 e março de 2015 no Brasil (ANS, 2015).

[1] Pessoa segurada que usufrui dos benefícios oferecidos por um plano de saúde.

Nota-se que a quantidade registrada de beneficiários, até março de 2015, totalizou quase 51 milhões de pessoas, representando mais de 26% da população brasileira coberta pela iniciativa privada. Esse elevado número de beneficiários influencia diretamente na receita obtida pelas OPSs. Entretanto, essa elevada quantidade influencia também no aumento das despesas assistenciais. Despesa assistencial é toda despesa resultante da utilização das coberturas oferecidas pela OPS, representando, portanto, todo gasto que precisa ser despendido quando um beneficiário precisa de atenção médica, como: consultas, exames, internações e terapias. A Figura 2 mostra a evolução das receitas e das despesas assistenciais das OPSs, entre o primeiro trimestre de 2007 e o primeiro trimestre de 2015.

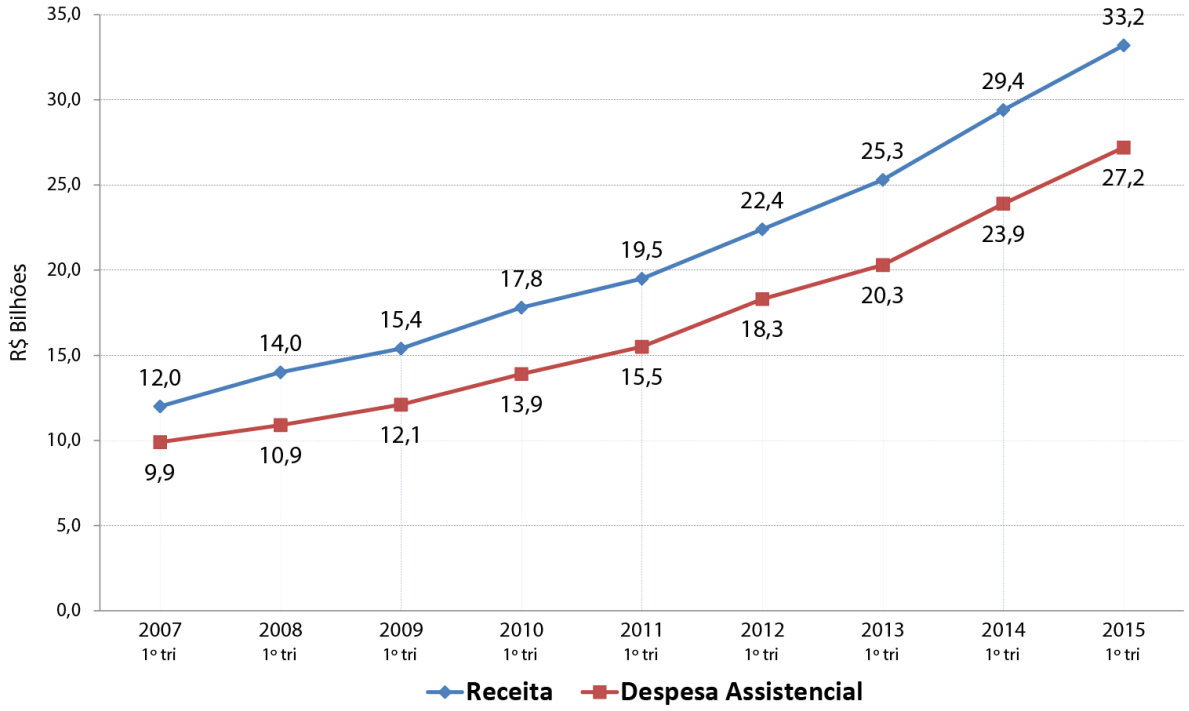

Figura 2 – Evolução das receitas e das despesas assistenciais dos planos de saúde privados entre o primeiro trimestre de 2007, e o primeiro trimestre de 2015. (ANS, 2015).

É possível notar pela Figura 2, que o crescimento das despesas assistenciais segue um ritmo similar ao crescimento das receitas, o que evidencia a importância para a gestão da OPS de se gerenciar os custos com saúde ao longo do ano, pois quanto mais próximo esses custos forem do rendimento da empresa, menor será o caixa disponível para investimentos, pagamento de salários e outros tipos de gastos administrativos. Nesse contexto de proximidade entre receita e despesa assistencial, qualquer método, técnica ou proposta, que reduza os gastos com saúde ou mantenha a receita estável, é relevante para o bom funcionamento financeiro do plano de saúde.

Definição do Problema

Devido a essa margem estreita de lucro, as OPSs, juntamente com pesquisadores, têm investido tempo e esforços na utilização de técnicas de mineração de dados. Boa parte das aplicações tem por finalidade reduzir as despesas assistenciais ou administrativas, como: prever erros na regulação de guias médicas (KUMAR; GHANI; MEI, 2010), detectar abusos em serviços requisitados pelos médicos (ORTEGA; RUZ; FIGUEROA, 2006) e reduzir custos com análises incorretas de guias (WOJTUSIAK et al., 2011; KUMAR; GHANI; MEI, 2010).

Outro ponto importante na estabilidade financeira é a receita da OPS. Como provedoras de seguro, as OPSs têm sua receita pautada no pagamento de uma taxa periódica, geralmente mensal, por parte do titular[2] do seguro. Essa forma de receita se baseia na metodologia de gerenciamento de risco conhecida como *risk pool* ou grupo de risco (CUTLER; ZECKHAUSER, 2000). Nessa metodologia, representada na Figura 3, um grupo de agentes compartilha o risco de que algo não desejável aconteça a algum agente específico. Dessa forma, ao invés de um eventual agente lesado arcar de forma individual com um débito alto, uma parte do valor contribuído pelo grupo é utilizado no pagamento dos gastos, mitigando de forma substancial o impacto financeiro para o agente envolvido (NORMAND, 2009).

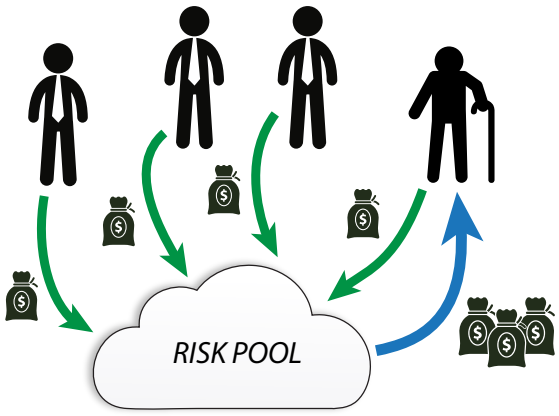

Figura 3 – Representação da metodologia de gerenciamento de risco *"Risk Pool"*.

Visto que o valor pago por cada titular é importante para o funcionamento estável da OPS, uma redução não prevista desses pagamentos pode comprometer o caixa financeiro da empresa. Essa redução não prevista acontece, geralmente, quando o titular resolve cancelar o contrato, encerrando a necessidade de futuros pagamentos. Baseando-se nisso, o problema específico abordado por este trabalho é o cancelamento eletivo dos contratos

[2] Pessoa responsável pelo pagamento do contrato estabelecido com a OPS. Essa pessoa não precisa ser necessariamente um beneficiário.

por parte do titular, ou seja, quando este decide, deliberadamente, encerrar sua parcela mensal de contribuição financeira, deixando de fazer parte do *risk pool* formado pela OPS.

Visão Geral da Proposta

A abordagem proposta objetiva caracterizar o cancelamento eletivo de contratos em planos de saúde privados. Essa caracterização compreende a capacidade de distinguir aspectos, padrões e propriedades que possam moldar, baseado em eventos passados, o perfil de um contrato cancelado; dessa forma, a gestão da OPS pode interceder previamente no problema do cancelamento, antecedendo-se ao possível encerramento de um contrato. A definição da abordagem é realizada por meio de etapas, que por sua vez pertencem às seguintes fases: Pré-Processamento, Mineração de Dados e Priorização de Contratos.

A fase de Pré-Processamento visa garantir uma maior qualidade às informações extraídas da base de dados de uma operadora de plano de saúde real, que abrange mais de 25.000 beneficiários, distribuídos principalmente nos estados brasileiros do Piauí e Maranhão. As etapas realizadas nessa fase são:

- Seleção de Contratos: é selecionado um sub-conjunto dos contratos presentes na base de dados da OPS. A partir dessa etapa, os contratos considerados não relevantes para a abordagem são desconsiderados;
- Limitação de Histórico: nessa etapa, os dados históricos dos contratos selecionados são limitados de acordo com um intervalo específico de datas;
- Limpeza dos Dados: os atributos que contêm informações pessoais como nome, endereço e dados bancários são removidos. Além disso, atributos com valores redundantes ou preenchidos de forma padrão também são retirados;
- Construção de Atributos: o conhecimento do especialista é utilizado para a criação e definição de atributos. Esses novos atributos representam informações que não podem ser diretamente mapeadas a atributos no banco de dados;
- Seleção de Atributos: são removidos os atributos que não apresentam relevância para a identificação de contratos ativos com características de contrato cancelado.

A fase de Mineração de Dados explora os dados pré-processados, à procura de padrões, relacionamento entre atributos e tendências com o objetivo de evidenciar informações ainda desconhecidas pela gestão da OPS. Essa fase consiste das seguintes etapas:

- Reconhecimento de Contratos: visa reconhecer contratos ativos com características de contrato cancelado. Esse reconhecimento é dado por meio do uso de um modelo

classificador. O resultado dessa classificação indica se o contrato está mais relacionado ao rótulo “cancelado” ou “não cancelado”;

- Definição do Classificador: contempla a comparação de performance entre classificadores de diferentes paradigmas de aprendizado. O objetivo é determinar, dentre os algoritmos pré-selecionados, quais constroem os modelos mais adaptados à base de dados;

- Identificação de Características: tem como objetivo identificar que tipos de ações e comportamentos levam os beneficiários da empresa a cancelarem seus contratos. Essa identificação é realizada por meio da análise dos ramos gerados por classificadores baseados em árvore de decisão;

- Previsão de Cancelamento: realiza uma regressão para estimar o tempo entre a data de início de um contrato na OPS e o fim desse contrato, ou seja, a data de cancelamento. Esse tempo estimado é importante na fase de Priorização de Contratos, pois compõe uma das formas utilizadas para destacar os contratos com maior risco de cancelamento;

- Definição do Regressor: possui o mesmo papel de comparação entre algoritmos da etapa Definição do Classificador, porém, nessa etapa, é avaliada a performance para realizar a regressão que estima o tempo até o cancelamento de um contrato;

- Refinamento do Modelo: objetiva criar um modelo mais robusto, de classificação e regressão, a partir, respectivamente, dos dois melhores algoritmos encontrados na etapa Definição do Classificador e Definição do Regressor. Esse modelo mais robusto visa encorporar o melhor de cada algoritmo isolado, visto que diferentes paradigmas de aprendizado são utilizados.

Por fim, a fase de Priorização de Contratos objetiva priorizar os contratos ativos classificados como “cancelado”, de acordo com o risco de cancelamento associado a cada contrato. Para isso, é desenvolvida a etapa a seguir:

- Ordenação de Contratos: nessa etapa, os contratos provenientes da fase de Mineração de Dados são organizados/ordenados visando destacar aqueles com maior risco de cancelamento e que, de forma complementar, representam uma maior contribuição financeira para a OPS.

Objetivos

O objetivo principal deste trabalho consiste em desenvolver uma abordagem para caracterizar o cancelamento eletivo de contratos em planos de saúde privados. Ressalta-se

que o principal guia para essa caracterização é a descoberta de conhecimento implícito na base de dados da OPS, pois, *a priori*, apesar de desconhecida para a gestão, esse conhecimento deve apresentar, e representar, particularidades expressivas para o problema em foco.

Além do objetivo principal, pretende-se alcançar os seguintes objetivos específicos:

- Classificar contratos ativos em contratos com características de contrato cancelado. Essa classificação é realizada, principalmente, na etapa Reconhecimento de Contratos (fase de Mineração de Dados);
- Analisar regras geradas por classificadores baseados em árvore de decisão, para identificar que padrões caracterizam os contratos ativos rotulados com a classe "cancelado". Essa análise é executada na etapa Identificação de Características (fase de Mineração de Dados);
- Estimar, por de meio de regressão, o tempo entre a data de início de um contrato e a data de seu cancelamento. Essa regressão é realizada na etapa Previsão de Cancelamento (fase de Mineração de Dados);
- Realizar uma comparação entre diferentes classificadores e regressores para definir os paradigmas de aprendizado e os algoritmos que apresentam melhores resultados para o problema. Essa comparação é realizada tanto na etapa Definição do Classificador como na etapa Definição do Regressor (fase de Mineração de Dados);
- Priorizar os contratos ativos rotulados como "cancelado" de forma a evidenciar os mais importantes para a gestão dentro do contexto do cancelamento. Essa priorização é realizada na etapa Ordenação de Contratos (fase de Priorização de Contratos).

Justificativa

Escolheu-se a área de planos de saúde privados pela sua representação no contexto nacional, com uma abrangência superior a 25% da população brasileira. Além disso, constitui-se como um mercado crescente devido ao baixo e precário investimento em saúde pública no Brasil. É notável, na maioria das cidades brasileiras, o descaso com os cidadãos que precisam ser atendidos pelo SUS[3] (Sistema Único de Saúde) e têm seu acesso negligenciado a medicamentos, médicos, internações e até mesmo urgências. Soluções que permitam reduzir gastos, ou que mantenham a receita superior às despesas, podem viabilizar uma maior popularização de planos de saúde privados, levando ao restante da população artifícios que contribuam para um melhor cuidado com a saúde. Salienta-se que os planos de saúde públicos não fazem parte do escopo do trabalho, devido a alguns fatores

[3] Sistema público que oferece, de forma gratuita, acesso a serviços de saúde a todo cidadão brasileiro.

externos que poderiam dificultar o decorrer de um estudo, como: burocracia na aquisição dos dados, interesses políticos, mudanças na gestão administrativa, amparo financeiro das esferas públicas, dentre outros.

Apesar da existência de vários estudos na literatura sobre redução de despesas em planos de saúde privados, não foram encontrados trabalhos com o objetivo claro de manter o nível de receita estável, promovendo, por exemplo, abordagens para inibir uma redução na receita. Devido a essa lacuna na análise da receita, foi escolhido como problema-alvo o cancelamento eletivo do contrato estabelecido com o plano de saúde. Uma saída deliberada do plano pode comprometer o planejamento financeiro da empresa, pois, *a priori*, a gestão da OPS não possui mecanismos que identifiquem se um determinado titular irá cancelar seu contrato.

O aprofundamento em técnicas de mineração de dados foi motivado por se acreditar que informações relevantes, sobre o problema-alvo, estejam implicitamente armazenadas no banco de dados da OPS, informações essas que podem ajudar a caracterizar um contrato com potenciais riscos de cancelamento. Por isso, desenvolveu-se uma abordagem voltada para a descoberta de conhecimento em banco de dados.

Acredita-se também que, apenas a rotulação de um contrato em "cancelado" ou "não cancelado" não produziria informação suficiente para a gestão da OPS, pois essa rotulação é apenas um efeito do cancelamento e não a causa real. Diz-se apenas um efeito porque, mesmo sendo uma informação interessante, não indica à gestão que motivos levaram o titular a cancelar seu contrato. Portanto, de modo complementar a essa informação disponibilizada, identifica-se possíveis razões que motivem um titular a realizar o cancelamento de seu contrato ou, de forma igualmente importante, motivem-lo a manter seu contrato ativo na OPS.

Contribuições

Como contribuições relevantes do estudo realizado, destacam-se:

- Identificação de padrões, ações e comportamentos, que caracterizam um perfil de um contrato cancelado. De posse dessa informação a gestão da OPS pode, por exemplo, desenvolver políticas que atenuem algum dos comportamentos identificados no perfil. Dessa forma, age-se diretamente sobre os motivos que levam os beneficiários a cancelarem deliberadamente seus contratos;
- Rotulação de quais contratos ainda ativos possuem características de contrato cancelado. Por meio dessa rotulação, a gestão da OPS pode executar medidas proativas sobre os contratos rotulados, com o objetivo de evitar que os titulares em questão realmente concretizem o cancelamento;

- Estimação do tempo esperado até que o contrato seja efetivamente cancelado. A partir dessa informação, a OPS pode ter uma ideia de quando os titulares do contrato pretendem realizar o cancelamento, servindo assim de base para ações preventivas adequadas à dimensão do tempo estimado;
- Priorização dos contratos rotulados como "cancelado" pela abordagem. Por meio dessa priorização, amplia-se o horizonte de recursos da gestão, permitindo a seleção dos contratos de acordo com a sua relevância para empresa;
- Uma abordagem completa, desde a extração dos dados da base de dados até a entrega de diferentes artefatos para a gestão da OPS, contemplando quais contratos ativos podem ser cancelados, o tempo estimado até esse cancelamento e a priorização necessária para facilitar o manuseio de toda essa informação por parte da OPS.

Estrutura do Trabalho

O restante deste trabalho está estruturado da seguinte forma: o Capítulo 1 elenca trabalhos encontrados na literatura que desenvolveram estudos relacionados a planos de saúde e ao cancelamento de contratos; o Capítulo 2 retrata aspectos relacionados à Descoberta de Conhecimento em Banco de Dados, Aprendizado de Máquina e aos algoritmos utilizados neste estudo; o Capítulo 3 detalha a estrutura da abordagem proposta e de cada uma das suas fases e etapas; o Capítulo 4 expõe e discute os resultados encontrados com a execução prática da abordagem em uma OPS real e o Capítulo 5 retrata as conclusões, limitações e pontos de continuidade da pesquisa desenvolvida neste trabalho.

1 Trabalhos Relacionados

O problema que motiva o presente trabalho é caracterizar, por meio de mineração de dados, o cancelamento eletivo de contratos em planos de saúde, portanto, pode-se entender o estudo realizado como a união de dois temas distintos: cancelamento de contratos e planos de saúde. O primeiro tema serve como base para muitos trabalhos na literatura, que visam construir modelos para prever essa perda de clientes e proporcionar uma robustez estratégica à gestão da empresa (Gür Ali; ARITüRK, 2014). No outro tema tem-se, também, uma variedade de pesquisas com o objetivo de fundamentar a decisão dos gestores e manter as OPSs funcionando de forma estável (TOMAR; AGARWAL, 2013), até mesmo o próprio crescimento do mercado e da competitividade que gira em torno de planos de saúde (KOSE; GOKTURK; KILIC, 2015). Em ambos os temas a aplicação de mineração de dados ocorre de diferentes formas, por meio do uso de diferentes estratégias e algoritmos. O propósito deste capítulo é destacar algumas das pesquisas realizadas na literatura, evidenciando o objetivo de cada uma e comparando-as, a partir de questões pré-estabelecidadas, com o presente trabalho.

1.1 Cenários de Planos de Saúde

O uso de mineração de dados na área de planos de saúde tem sido uma importante fonte para descoberta de conhecimento nesse ramo de negócio. Além disso, por meio da aplicação de técnicas de aprendizado de máquina, um conjunto de comportamentos no cenário médico pode ser aprendido, detectado e até mesmo antecipado. Ortega, Ruz e Figueroa (2006) descreveram um sistema para detectar fraudes e abusos em guias médicas[1] de um plano de saúde privado chileno. Um comitê de *perceptron* multicamadas (do inglês *multilayer perceptron* ou MLP) foi utilizado para relacionar um evento fraudulento às quatro entidades envolvidas no processo entre uma OPS e um hospital: guias médicas, beneficiários, médicos e funcionários. A utilização desse comitê resultou em um sistema capaz de detectar, aproximadamente, 75% de casos fraudulentos e abusivos por mês. Kose, Gokturk e Kilic (2015) desenvolveram uma ferramenta de suporte a decisão baseada no conceito de aprendizado de máquina interativo que, diferente do modelo tradicional, envolve a participação de especialistas dentro da fase de treinamento. A ideia dessa ferramenta é identificar e priorizar casos suspeitos de fraudes em guias médicas, formatando tais casos em um painel de visualização que permita, de forma mais clara e direta, ajudar o especialista a avaliar se existe ou não realmente um abuso por parte dos envolvidos. Após essa

[1] Conjunto de informações sobre o atendimento realizado em um paciente, como: tipo de acomodação em caso de internação, medicamentos utilizados, tratamentos executados, materiais utilizados, etc.

avaliação, o especialista pode realimentar o sistema com possíveis novos padrões detectados, visando melhorar o processo de detecção de fraudes. Araújo, Macedo e SANTOS NETO (2015) apresentaram um processo para realizar o aprendizado automático da regulação médica/odontológica de uma operadora específica de plano de saúde. Foi executado um conjunto de experimentos para avaliar o resultado preditivo logrado por classificadores de diversos paradigmas de aprendizado. Além disso foram utilizadas técnicas para combinar os melhores classificadores com o objetivo de obter um processo decisivo mais robusto.

Os trabalhos apresentados sobre mineração de dados em planos de saúde utilizaram diferentes técnicas e algoritmos para solucionar problemas ligados a esse cenário. Entretanto, não há, em nenhum deles, um objetivo bem definido visando evidenciar quais foram as possíveis razões e motivos que levaram aos resultados da classificação. Lembra-se que no presente trabalho, aborda-se o problema do cancelamento eletivo de contratos em plano de saúde, e que um dos principais objetivos é evidenciar quais fatores possuem influência nesse cancelamento.

Os estudos de Wojtusiak et al. (2011) e Kumar, Ghani e Mei (2010), apesar de experimentarem um pequeno número de algoritmos, apresentaram um interesse em entender os comportamentos que influenciaram nos resultados obtidos. Wojtusiak et al. (2011) apresentaram um método que derivava regras interpretáveis para ajudar na preparação e investigação de guias médicas antes de submetê-las às empresas responsáveis pelo pagamento, o que reduziu custos com erros e imperfeições na análise das guias. Kumar, Ghani e Mei (2010) também propuseram uma abordagem para reduzir os gastos excessivos no processamento de guias médicas. Foi descrito um sistema com o objetivo de prever quais guias precisariam ser reprocessadas, gerando automaticamente um conjunto de motivos para explicar o porquê dessas guias necessitarem de uma segunda análise, facilitando o trabalho desenvolvido pelos auditores responsáveis por esse processamento. Delen et al. (2009) detalharam uma metodologia para classificar se uma determinada pessoa tem cobertura ou não de plano de saúde, levando em consideração para isso uma série de respostas obtidas por meio de um questionário sobre fatores de risco. Delen et al. (2009) compartilharam dos mesmos interesses base deste presente trabalho: aplicar diferentes técnicas de mineração de dados para se obter resultados promissores de classificação, e de forma complementar, descobrir, analisar e evidenciar possíveis motivos, razões ou ações que influenciam nesses resultados.

1.2 Cenários de Cancelamentos

Cancelamento de contratos, perda de clientes ou *churn* em inglês são sinônimos que representam um problema comum a diversos segmentos empresariais (XIAO-BING; JIE; ZAI-WU, 2012). Geralmente se configura um cenário no qual clientes de uma "Empresa A"

optam por cancelar seus contratos e associar-se a uma "Empresa *B*". Devido ao custo de aquisição de um novo cliente ser maior que a permanência dele, empresas têm apostado em formas de fidelizar a clientela, o que motivou/motiva uma série de pesquisas a proporem métodos para descobrir com antecedência futuros cancelamentos ou mesmo entender o porquê da saída de um cliente.

Huang et al. (2015) propuseram um sistema para prever a saída de clientes de uma das maiores companhias telefônicas da China. O conjunto de atributos utilizados consistiu do agrupamento entre as ferramentas de apoio a negócio e a operação, já existentes na operadora chinesa. Além da previsão de cancelamento, o sistema proposto classificava grupos de consumidores mais adequados às campanhas de marketing disponíveis. Um teste A/B[2] era realizado para iterativamente rotular os clientes mais suscetíveis a cada campanha, objetivando melhorar a classificação de futuros clientes. Huang, Kechadi e Buckley (2012) elaboraram e avaliaram um conjunto próprio de atributos para caracterizar um cliente no setor de telecomunicações, visando melhorar o poder preditivo no processo de classificação. Foram realizados experimentos comparativos entre os atributos normalmente utilizados no setor e os atributos propostos, evidenciando a melhoria acarretada por meio da nova modelagem. Moeyersoms e Martens (2015) descreveram uma variedade de métricas que podem ser utilizadas na inclusão de atributos categóricos de alta cardinalidade[3], utilizando como problema-alvo o cancelamento de consumidores em uma concessionária de energia elétrica belga. Por meio da comparação entre diversos classificadores o estudo mostrou uma melhoria no processo de predição ao se utilizar as métricas evidenciadas. Gür Ali e Aritürk (2014) propuseram uma nova metodologia para descrever as características dos consumidores em variados períodos de tempo, enquanto que a metodologia padrão é observar o consumidor apenas em um período específico. A metologia proposta foi aplicada ao cancelamento efetuado por clientes de um banco privado europeu, demonstrando uma melhoria significativa no poder preditivo da classificação quando comparada à metodologia padrão.

No parágrafo anterior são relacionados trabalhos que compartilham de uma característica em comum: analisar o estudo proposto por meio de variados algoritmos de classificação. O presente trabalho possui a mesma característica, porém há também neste trabalho a preocupação em realizar um ajuste dos parâmetros dos algoritmos, de forma a conseguir uma configuração mais adequada à base de dados utilizada. Outro ponto destacado neste estudo é o interesse em elencar as razões que podem ter influenciado no cancelamento dos usuários, algo não endereçado nos estudos relacionados anteriormente.

[2] Um tipo de teste no qual os elementos envolvidos são separados, randomicamente, entre grupo de controle (não recebe o tratamento) e grupo de tratamento. O objetivo da divisão é avaliar se o tratamento em questão influencia nos resultados.

[3] Atributos que possuem um grande número de valores, como CEPs, número de identidade, nomes de família, etc.

Dentre os trabalhos encontrados na literatura, dois deles possuem o mesmo domínio de aplicação deste estudo: cancelamentos em planos de saúde. Su et al. (2009) utilizaram regressão logística para elencar as características mais relevantes para a saída de um cliente do plano de saúde, assim como atribuir a cada um dos consumidores um risco de cancelamento, permitindo a priorização dos casos mais graves. É importante frisar que as características foram elencadas de forma independente, ou seja, motivos de cancelamento envolvendo mais de uma variável não foram evidenciados, diferente do corrente trabalho que analisa esse aspecto também sobre a ótica de interdependência entre os fatores. Em Su et al. (2009), o treinamento/validação do classificador todos os consumidores foram agrupados em dois grupos, por meio do algoritmo *k-Means*, e o grupo no qual a taxa de cancelamento era maior foi escolhido. Goonetilleke e Caldera (2013) realizaram experimentos com diferentes classificadores para rotular um determinado consumidor como alguém que iria continuar ou sair do plano de saúde. O problema do desbalanceamento entre classes foi abordado por meio de um aprendizado baseado em custos, no qual há uma diferença de custo para cada possibilidade de classificação. O teste do modelo foi realizado por meio de validação cruzada, com a separação dos clientes em 10 grupos aleatórios.

Para evitar uma comparação repetitiva de abordagens, métodos e técnicas utilizadas entre os trabalhos anteriores e o presente trabalho, realizou-se a sumarização das características de cada estudo na Tabela 1, sendo observados os seguintes pontos:

- **P1**: "Experimentos com diferentes classificadores?"
- **P2**: "Evidencia os motivos/razão que influenciaram na classificação?"
- **P3**: "Ajuste automático ou semi-automático das configurações dos algoritmos?"
- **P4**: "Aplicação de técnicas para balanceamento das classes?"
- **P5**: "Uso de modelo interativo para mineração de dados?"
- **P6**: "Utilização de técnicas para combinar classificadores?"
- **P7**: "Uso de algum método de teste, como o teste A/B, para avaliar se a modelagem proposta também obtém resultados semelhantes na prática?"
- **P8**: [Para os estudos aplicados ao cancelamento] "É realizada uma estimativa de tempo para a saída do cliente/consumidor?"

Percebe-se que, dentre os oito pontos destacados, o presente trabalho não aborda um modelo interativo para mineração de dados e não há um método para avaliação dos resultados na prática. O modelo interativo não foi adotado por requerer, previamente, uma ideia mais sólida de como funciona o contexto de aplicação da mineração de dados,

que no caso deste estudo é uma OPS. Acredita-se que, com um amadurecimento da abordagem proposta, deve ser possível intercalar gradativamente o especialista no processo de aprendizado, potencializando o conhecimento adquirido internamente pela OPS. Considera-se também, de importante relevância, a avaliação dos resultados na prática. Essa avaliação é um dos pontos a serem desenvolvidos como continuidade da pesquisa, de forma a complementar a aplicação da abordagem proposta, como ferramenta de apoio, dentro da OPS. Ressalta-se que, apesar de não abordar os dois pontos anteriormente discutidos, este estudo diferencia-se por: executar um ajuste automático das configurações dos algoritmos utilizados, o que pode evitar possíveis perdas de performance utilizando a configuração padrão; e, ainda mais importante, é realizada uma estimativa do tempo até o cancelamento de um contrato, tempo esse que pode potencializar o valor da informação a ser consumida pela gestão da OPS.

Tabela 1 – Sumarização dos trabalhos relacionados.

Trabalho	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
Este trabalho	X	X	X	X		X		X
Ortega, Ruz e Figueroa (2006)						X		
Kose, Gokturk e Kilic (2015)					X	X		
Araújo, Macedo e SANTOS NETO (2015)	X			X		X		
Wojtusiak et al. (2011)	X	X						
Kumar, Ghani e Mei (2010)	X	X						
Delen et al. (2009)	X	X						
Huang et al. (2015)	X			X			X	
Huang, Kechadi e Buckley (2012)	X							
Moeyersoms e Martens (2015)	X							
Gür Ali e Aritürk (2014)	X			X				
Su et al. (2009)		X						
Goonetilleke e Caldera (2013)	X	X				X		

2 Descoberta de Conhecimento em Banco de Dados

A Descoberta de Conhecimento em Banco de Dados (DCBD) é o estudo de coletar, limpar, processar, analisar e obter informações úteis e relevantes a partir de um conjunto de dados (AGGARWAL, 2015). Devido a capacidade da DCBD de revelar padrões, até então desconhecidos em um determinado banco de dados, diversas pesquisas fazem uso de técnicas de DCBD em variados tipos de problema, como: prever em qual curso um universitário irá se matricular (OGNJANOVIC; GASEVIC; DAWSON, 2016); determinar automaticamente o assunto de uma notícia divulgada na internet (IGLESIAS et al., 2016); analisar sentimentos em conteúdos gerados por redes sociais (GASPAR et al., 2016); e operar, de forma autônoma, trens "inteligentes" (YIN; CHEN; LI, 2016). A DCBD pode ser dividida em três momentos/estágios (AGGARWAL, 2015):

- **Coleta de Dados:** representa a parte basilar de todo o processo de descoberta, pois é nesse momento que os dados contidos na base de dados são selecionados e minimamente organizados para identificar o artefato de estudo;

- **Extração dos Atributos e Limpeza de Dados:** nesse estágio, os dados passam por um processo maior de organização, transformando-se em um formato adequado para viabilizar o processamento e a aplicação de técnicas analíticas. É importante salientar a limpeza dos dados realizada, pois o que é recebido da coleta de dados ainda pode representar atributos incompletos, mal-concebidos e até mesmo irrelevantes para um problema específico;

- **Mineração de Dados:** o estágio final do processo de DCBD tem como objetivo analisar os dados pré-processados; essa análise dá-se pela aplicação de diferentes estratégias e técnicas visando produzir informações relevantes.

A área Aprendizado de Máquina está intimamente relacionada à DCBD, muitas vezes são até confundidas como um mesmo tema de estudo (LESKOVEC, 2014). Porém, isso não é exatamente verdade, mas muitas das técnicas utilizadas na Mineração de Dados da DCBD são oriundas do Aprendizado de Máquina (LESKOVEC, 2014); principalmente no que diz respeito ao Aprendizado Supervisionado (AGGARWAL, 2015). Por esse motivo, a próxima seção aborda mais claramente o Aprendizado de Máquina.

2.1 Aprendizado de Máquina

O Aprendizado de Máquina (AM), compreende o estudo de técnicas que possibilitem a um programa de computador aprender alguma tarefa por meio de experiências (MITCHELL, 1997). Mitchell (1997) define precisamente o AM da seguinte forma: "Um programa de computador é dito que aprendeu da experiência E, com respeito a algum tipo de tarefa T e métrica de performance P, quando sua performance, na tarefa T, medida por P, melhora a partir da experiência E".

Existem diversas formas de dividir a pesquisa e aplicação de AM no que diz respeito ao tipo de aprendizado dos algoritmos envolvidos. Uma definição comumente adotada na literatura é a divisão em (MOHRI, 2012):

- **Aprendizado Supervisionado:** está disponível para o aprendizado tanto exemplos como contra-exemplos, seja o rótulo correto para classificação ou o valor numérico para regressão; isso permite que o algoritmo possa avaliar e validar quão próximo o conhecimento que foi aprendido está do conhecimento correto;

- **Aprendizado Semi-Supervisionado:** o algoritmo não possui clareza para determinar se o aprendizado realizado está 100% certo, pois há a definição apenas de uma pequena quantidade de contra-exemplos. Nesse caso costuma-se bonificar ou penalizar o algoritmo caso a inferência realizada esteja mais para um conceito correto ou errado, respectivamente;

- **Aprendizado Não-Supervisionado:** não há uma definição de contra-exemplos para os dados disponíveis, o que implica na incapacidade, *a priori*, de separar os exemplos a partir de um rótulo bem definido.

Como neste estudo há a presença tanto de exemplos (contratos ativos) como de contra-exemplos (contratos cancelados), foca-se no Aprendizado Supervisionado. O Aprendizado Supervisionado visa solucionar um determinado problema ao aprender a mapear uma determinada entrada X para uma saída Y (ALPAYDIN, 2010). Chama-se esse mapeamento de regressão quando a saída Y é numérica, e chama-se de classificação quando a saída representa um rótulo. Para o processo de aprendizado geralmente se particiona os exemplos ou instâncias disponíveis em três conjuntos:

- **Conjunto de treinamento:** conjunto principal de instâncias utilizado pelo algoritmo para gerar um modelo que represente o aprendizado. A partir desse modelo o algoritmo utilizado pode então executar inferências sobre novos dados e realizar uma regressão ou classificação;

- **Conjunto de validação:** conjunto de instâncias utilizado para validar o modelo aprendido e verificar sua capacidade de generalização; ou seja, quão bem o modelo consegue gerar uma saída correta de acordo com as saídas esperadas pelo conjunto de validação. O melhor modelo encontrado tem o conjunto de validação incorporado ao seu conjunto de treinamento, gerando-se um novo modelo;

- **Conjunto de teste:** conjunto com instâncias até então desconhecidas para o modelo, com o intuito de avaliar o poder de generalização obtido por meio do aprendizado.

Uma analogia presente em Alpaydin (2010) ajuda no entendimento das diferenças entre os conjuntos supracitados. A analogia é: para um aluno se formando em um determinado curso, os problemas resolvidos em classe representam o conjunto de treinamento, os problemas abordados em uma prova representam o conjunto de validação, e os potencias problemas que o aluno enfrentará na sua carreira representam o conjunto de teste.

No que diz respeito aos métodos ou paradigmas de aprendizado, este trabalho foca em cinco tipos: baseado em exemplos, bayesiano, baseado em árvore de decisão, conexionistas e estatístico. Cada um desses paradigmas é apresentado nas próximas subseções, juntamente com os algoritmos utilizados de cada um.

2.1.1 Paradigma Baseado em Exemplos

Nesse paradigma o conjunto de treinamento não recebe nenhum tipo complexo de processamento, o treinamento do modelo consiste apenas do armazenamento das instâncias de treinamento em uma estrutura de dados de rápido acesso ou em uma combinação entre essas instâncias (ALPAYDIN, 2010). A computação em cima do modelo de aprendizado é realizada ao se receber uma instância de validação ou teste. Os algoritmos utilizados desse paradigma são: *K-Nearest Neighbors* (KNN) e K*.

O KNN é conhecido pela sua estrutura simples e ao mesmo tempo por apresentar performance competitiva com relação a outros algoritmos de Aprendizado de Máquina (PARK; LEE, 2013). O funcionamento básico do KNN é atribuir a um elemento do conjunto de teste uma determinada classe baseada nos vizinhos mais próximos. O K do nome do algoritmo representa a quantidade dos vizinhos mais próximos que serão utilizados para classificar o elemento de teste. Essa métrica de proximidade na versão padrão do KNN é a distância Euclidiana (TANEJA et al., 2014).

O K* difere dos outros algoritmos baseados em exemplos por utilizar como métrica de distância a entropia (CLEARY; TRIGG, 1995). Por meio de uma série de definições especializadas para atributos do tipo real, e atributos sem um valor definido, o K* consegue combinar atributos de diferentes instâncias, determinar a vizinhança mais próxima da

instância de teste e classificar essa instância de acordo com os vizinhos.

2.1.2 Paradigma Bayesiano

O paradigma bayeasiano, baseado no Teorema de Bayes (SWINBURNE, 2002), é centrado no uso de distribuição probabilística para descrever todos os aspectos relevantes à classificação de um elemento; a partir disso, define-se a probabilidade de um determinado evento baseado no acontecimento ou não de uma série de condições (BERNARDO, 2000).

O classificador *Naive Bayes* é um dos métodos mais práticos e simples do aprendizado bayesiano, ele se baseia no pressuposto de que todos atributos ou eventos que influenciam uma determinada classe são independentes entre si, o que diminui drasticamente a complexidade para realizar a classificação (MITCHELL, 1997).

Diferente do *Naive Bayes*, o classificador *BayesNet* permite a definição da probabilidade condicional entre diferentes atributos, o que aumenta a complexidade ao se construir um modelo de classificação, porém garante uma maior maleabilidade para definir como um conjunto de condições e suas potenciais inter-relações influenciam em um evento (MITCHELL, 1997).

2.1.3 Paradigma Baseado em Árvore de Decisão

Por meio de árvores de decisão, os classificadores desse paradigma tanto representam a informação, como utilizam essa estrutura para decidir por intermédio de regras bem definidas qual a classe de uma determinada instância (QUINLAN, 1987). Os algoritmos utilizados nesse paradigma são: C4.5, *RandomForest*, CART (do inglês *Classification and Regression Tree*) e M5.

No algoritmo C4.5 (QUINLAN, 1993) a árvore de decisão é dividida a partir do atributo que representar a maior razão de ganho (STREIB, 2009) dentre os outros atributos. Esse processo é continuamente repetido até que as instâncias estejam separadas por classe nas folhas da árvore de decisão, ou seja, nas folhas não se encontram instâncias de classes diferentes.

O *RandomForest* (BREIMAN, 2001) é uma combinação entre diferentes árvores de decisão isoladas, nas quais a seleção dos atributos, que compõem o aprendizado, é feita de forma aleatória. A moda do valor final dessas árvores isoladas é utilizada em problemas de classificação, e a média dos valores é utilizada quando o problema é de regressão. Uma das grandes vantagens do algoritmo *RandomForest*, em relação a utilizar apenas uma dessas árvores de decisão isolada, é uma maior robustez a ruídos e atributos que dificultam o aprendizado (BREIMAN, 2001).

O classificador CART (BREIMAN et al., 1984) permite a utilização tanto de variáveis contínuas como variáveis categóricas para identificar uma determinada classe. Para os atributos contínuos é construída uma sub-árvore para regressão e para as categóricas

é construída uma sub-árvore para classificação.

O algoritmo M5 (QUINLAN, 1992) serve exclusivamente para a tarefa de regressão. Esse algoritmo gera sub-árvores de regressão nas quais as folhas possuem um modelo linear para caracterizar os dados, dessa forma cada folha realiza uma estimativa do valor esperado para um determinado conjunto de entradas.

2.1.4 Paradigma Conexionista

Nesse paradigma, o processo de aprendizado é constituído de unidades processadoras simples que armazenam o conhecimento adquirido por meio de pesos entre suas conexões com outras unidades, realizando de forma global, uma computação dinâmica que leva em consideração cada unidade do sistema (DINSMORE, 2014). O nome desse paradigma teve origem com o trabalho de (RUMELHART; MCCLELLAND; GROUP, 1986), no qual esse paralelismo entre unidades foi chamado de "revolução conexionista" (do inglês *connectionist revolution*). O algoritmo utilizado desse paradigma é o MLP (do inglês *Multilayer Perceptron*).

O MLP é um algoritmo que pode possuir múltiplas camadas das mencionadas unidades processadoras, essas unidades são conhecidas como neurônios artificiais, constituindo dessa forma uma rede neural artificial. Um conjunto de neurônios artificiais é agrupado em cada camada, e cada camada é responsável pelo processamento do sinal desde a entrada (camada de entrada), "descoberta" de atributos que caracterizam os dados (camada escondida) à saída do algoritmo (camada de saída) (HAYKIN, 2009). O treinamento de uma MLP é popularmente executado por outro algoritmo conhecido como "*back-propagation*" (HAYKIN, 2009).

2.1.5 Paradigma Estatístico

No paradigma estatístico, o objetivo é formular um modelo, juntamente com seus parâmetros, que possa, de forma aproximada, representar a estrutura fundamental que um conjunto de dados obedece (HASTIE, 2009). A partir de um processo de aprendizado iterativo, os parâmetros são ajustados a fim de garantir uma maior qualidade ao modelo que é formado. Esse paradigma pode ser entendido como uma representação generalizada e mais abrangente dos outros paradigmas de aprendizado detalhados. Os algoritmos utilizados são: SVM (do inglês *Support Vector Machine*) e Regressão Linear.

O SVM realiza um mapeamento no conjunto de treinamento de tal forma que se encontra o hiperplano de separação ótimo entre instâncias de classes diferentes. Utiliza-se uma abordagem de cunho geométrico e matematicamente bem fundamentada como base para o processo de definição do hiperplano e consequente da maior distância que separe as classes (BENNETT; CAMPBELL, 2000).

A Regressão Linear (CHATTERJEE, 2006) é uma das formas mais básicas de se

estimar uma variável final (dependente), a partir de uma ou mais variáveis independentes. É dita básica por supor que a relação entre esses variáveis dá-se por uma expressão linear, cujos parâmetros são ajustados a partir das observações/dados do problema. A Regressão Linear pode ser simples quando há apenas uma variável independente, ou múltipla quando há mais de uma variável independente envolvida.

2.1.6 Avaliação dos Classificadores

Para avaliar e comparar os resultados de diferentes classificadores, é necessária a computação de métricas que demonstrem a performance de cada algoritmo. Uma diversidade de métricas foi desenvolvida na literatura, abordando diferentes conceitos de performance. A maioria dessas métricas é calculada por meio de uma matriz de confusão, que serve para facilitar a visualização dos acertos e erros de um classificador (PATRO; PATRA, 2015). A estrutura da matriz de confusão (WITTEN; FRANK; HALL, 2011) aplicada a uma classificação binária[1], classe positiva A e classe negativa B, é ilustrada na Figura 4. As linhas da matriz de confusão representam a classe correta dos elementos a serem classificados, e as colunas representam qual a classe foi escolhida pelo classificador. A explicação dos termos VP, FP, VN, FN é feita a seguir:

- **VP - Verdadeiros Positivos:** representa a quantidade de elementos da classe A que corretamente foram classificados como pertencentes à classe A;
- **FP - Falsos Positivos:** representa a quantidade de elementos da classe B que incorretamente foram classificados como pertencentes à classe A;
- **VN - Verdadeiros Negativos:** representa a quantidade de elementos da classe B que corretamente foram classificados como pertencentes à classe B;
- **FN - Falsos Negativos:** representa a quantidade de elementos da classe A que incorretamente foram classificados como pertencentes à classe B.

	CLASSIFICADO COMO **CLASSE *A***	CLASSIFICADO COMO **CLASSE *B***
CLASSE *A*	VP	FN
CLASSE *B*	FP	VN

Figura 4 – Matriz de confusão para uma classificação binária.

Neste trabalho são utilizadas quatro medidas de performance, baseadas na matriz de confusão, para avaliar os classificadores, sendo elas: *recall*, precisão, curva ROC (do

[1] Só existem duas classes possíveis, comumente uma é chamada de positiva e a outra de negativa.

inglês *Receiver Operating Characteristic*) (FAWCETT, 2006) e AUC (do inglês *Area Under ROC Curve*).

O *recall*, Equação 2.1, representa a percentagem de elementos que foram corretamente rotulados dentre todos aqueles presentes em uma base de dados, sendo importante para demonstrar a cobertura da classificação nessa base. A precisão, Equação 2.2, representa a taxa de corretude da classificação apenas dentre os elementos rotulados, sendo relevante para avaliar a taxa preditiva do modelo classificador. Ambas as métricas medem, da sua maneira, o acerto de um determinado algoritmo. Por exemplo, imagina-se um cenário para classificar se um determinado contrato está "cancelado", no qual a base de dados possui 60 contratos ativos e 40 cancelados. A Figura 2.4 ilustra a matriz de confusão caso o classificador rotula-se 80 contratos como "cancelado", dos quais 40 são realmente contratos cancelados mas os outros 40 são contratos ativos. O *recall* nesse caso seria de 100% (Equação 2.3), pois o classificador rotulou corretamente todos os contratos cancelados existentes. Porém, a precisão seria de apenas 50% (Equação 2.4), visto que metade dos contratos rotulados como "cancelado" eram na verdade contratos ativos.

$$recall = \frac{VP}{VP + FN} \tag{2.1}$$

$$precisão = \frac{VP}{VP + FP} \tag{2.2}$$

	CLASSIFICADO COMO **CANCELADO**	CLASSIFICADO COMO **NÃO CANCELADO**
CANCELADO	VP = 40	FN = 0
NÃO CANCELADO	FP = 40	VN = 20

Figura 5 – Matriz de confusão para um cenário de exemplo.

$$recall\ (Figura\ 5) = \frac{VP}{VP + FN} = \frac{40}{40 + 0} = \frac{40}{40} = 100\% \tag{2.3}$$

$$precisão\ (Figura\ 5) = \frac{VP}{VP + FP} = \frac{40}{40 + 40} = \frac{40}{80} = 50\% \tag{2.4}$$

A curva ROC é um dos tipos de medidas de performance mais utilizado para avaliar classificadores (ZHANG et al., 2015). O objetivo dessa medida é retratar a relação entre o *recall* e a taxa de falsos positivos ao se variar o limiar de discriminação (LD) da classificação. O LD representa o valor ou ponto fundamental que diferencia, por exemplo, uma instância entre classe A e classe B. Um classificador pode expressar o rótulo de uma

instância a ser classificada por meio de uma probabilidade, variando de 0 a 1; caso essa probabilidade seja maior que 0,5 (LD) então a instância pertence a classe A; caso contrário, pertence a classe B. Para cada valor de LD há um valor para o *recall* e outro para a taxa de falso positivos, dessa forma é possível construir a curva ROC propriamente dita, como pode ser visto na Figura 6.

Ao se comparar dois ou mais classificadores nem sempre é possível que haja uma única curva ROC que se sobressaia, o que motivou a comunidade científica a adicionalmente utilizar a métrica AUC (ZHANG et al., 2015). Essa métrica representa o valor da área sob a curva ROC, servindo como uma forma de resumir, em apenas um valor numérico, a performance de um classificador (XU et al., 2013). A métrica AUC possibilita a comparação mais objetiva entre diferentes algoritmos, quanto maior essa métrica, melhor pode ser considerado um classificador no domínio da curva ROC.

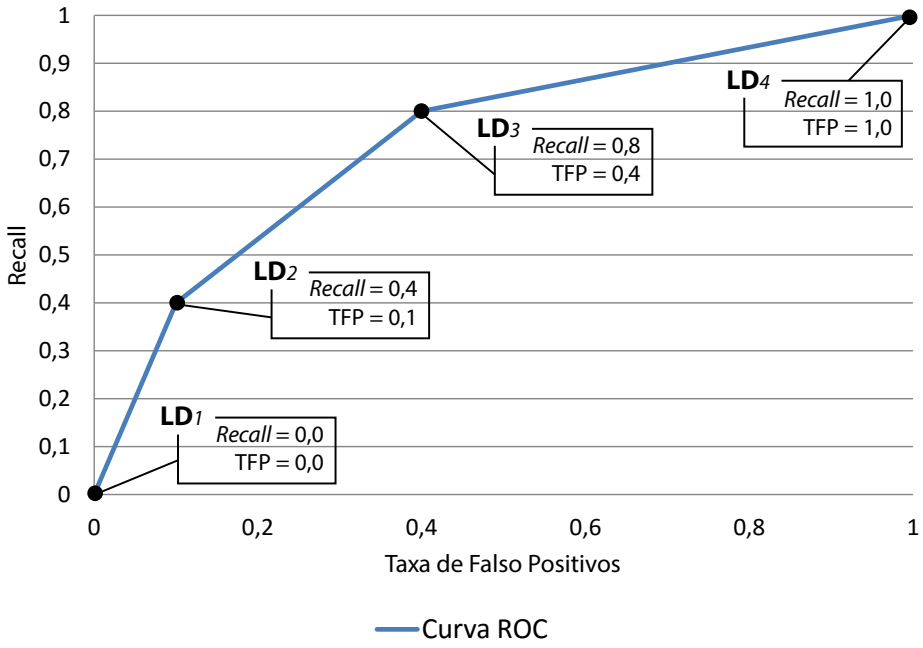

Figura 6 – Exemplo de uma curva ROC, no qual destacam-se os pontos LD utilizados para construir essa curva.

2.1.7 Avaliação dos Regressores

O objetivo da regressão é desenvolver um modelo que estime um valor de saída para um conjunto de entradas, a partir de uma série de observações (MONTGOMERY, 2012). Como esse modelo na maioria das vezes não é perfeito, ou seja, não estima exatamente o valor esperado, é necessário medir quão distante o modelo está do cenário real. A métrica utilizada neste trabalho, para realizar essa medição de performance do modelo regressor, é o RMSE (do inglês *Root Mean Square Error*).

O RMSE também é conhecido como o desvio padrão do erro preditivo, como pode ser notado na Equação 2.5, na qual N representa o número de elementos. O erro preditivo é a diferença entre o valor estimado e o valor esperado. Por ser equivalente ao desvio padrão desse erro, o RMSE serve para representar, em média, quanto de erro (para mais ou para menos) um regressor apresenta ao ser comparado ao modelo original.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(valorEstimado_i - valorEsperado_i)^2}{N}} \tag{2.5}$$

3 Abordagem Proposta

Para um agrupamento contextual das diferentes etapas participantes da abordagem proposta, definiu-se de três fases: Pré-Processamento, Mineração de Dados e Priorização de Contratos. As fases são detalhadas respectivamente na Seção 3.2, Seção 3.3 e Seção 3.4. A estrutura geral da abordagem é representada na Figura 7.

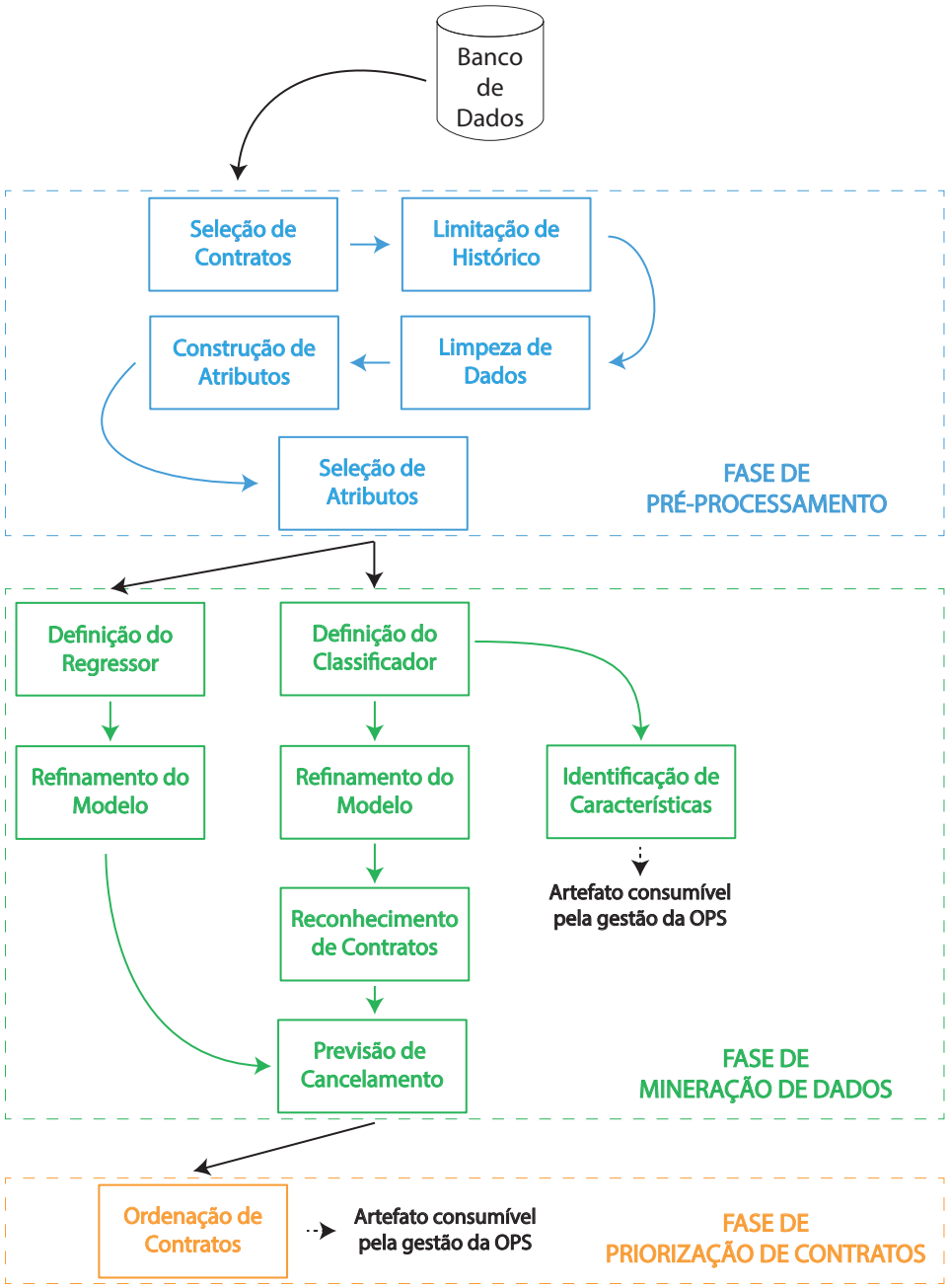

Figura 7 – Estrutura da abordagem proposta.

3.1 Base de Dados

A base de dados utilizada neste trabalho foi fornecida por uma OPS sediada no estado do Piauí, Brasil. A base possui informações de mais de 25.000 beneficiários, com contratos estabelecidos entre Março de 2005 e Novembro de 2015, presentes em 5 tabelas, que juntas totalizam 230 colunas/atributos. As definições de conceitos importantes presentes nos dados cedidos e comuns ao contexto de planos de saúde, são as seguintes:

- Beneficiário: segurado que possui um vínculo com a empresa, podendo usufruir das coberturas acordadas no contrato;
- Titular: pessoa responsável pelo pagamento das mensalidades ao plano. Essa pessoa pode ser um beneficiário ou apenas o responsável financeiro (sem direito aos benefícios do plano de saúde);
- Dependente: beneficiário do plano que não possui responsabilidades financeiras com a empresa;
- Cobertura: tipo de atendimento que o beneficiário tem direito caso necessite de alguma assistência;
- Produto do Contrato: conjunto de coberturas associadas a um contrato;
- Contrato Individual: contrato firmado diretamente com uma pessoa física;
- Contrato Corporativo: contrato firmado com uma empresa que deseja beneficiar um grupo de funcionários com um plano de saúde;
- Data de Adesão: data oficial do início do contrato de um beneficiário.

3.2 Pré-Processamento

A fase de Pré-Processamento tem como objetivo principal remover as impurezas, ruídos e características potencialmente presentes na base de dados original, e que podem afetar a qualidade dos dados envolvidos. A não ou má execução dessa fase pode limitar a eficácia da fase de Mineração de Dados e consequentemente do resultado esperado para a abordagem proposta (LTIFI; KOLSKI; AYED, 2015). As etapas que constituem o Pré-Processamento são: Seleção de Contratos, Limitação de Histórico, Limpeza de Dados, Construção de Atributos e Seleção de Atributos.

3.2.1 Seleção de Contratos

Na etapa Seleção de Contratos, há uma primeira redução na quantidade de contratos processados pela abordagem; essa redução é obtida pela seleção apenas dos contratos dos

titulares, ou seja, não são selecionados da base de dados os contratos que representam dependentes. Optou-se por selecionar somente os contratos de titulares porque é de responsabilidade do titular realizar o pagamento das mensalidades ao plano de saúde e decidir sobre o cancelamento do contrato. Vale ressaltar que as informações relacionadas aos dependentes não são totalmente descartadas, alguns dados considerados relevantes para a abordagem são adicionados no contrato do titular associado ao dependente. Na etapa Construção de Atributos há mais detalhes sobre quais dados dos dependentes são utilizados.

Como o objetivo principal deste trabalho está relacionado ao cancelamento eletivo de contrato, apenas os contratos que foram firmados diretamente entre a OPS e uma pessoa física são selecionados. Isso significa que contratos corporativos são desconsiderados, pois quando há um cancelamento nesse caso a motivação é, geralmente, originada pela empresa que contratou a OPS, e não pelo funcionário que perderá sua cobertura de saúde.

Da base de dados original foram selecionados somente contratos com data de adesão superior ou igual a 01 de Janeiro de 2013. Essa data não foi escolhida empiricamente, pois de acordo com especialistas da OPS a maior parte dos produtos relacionados a contratos individuais foi adicionada em meados de 2012. Portanto, como é feita a seleção apenas de contratos individuais, resolveu-se selecionar os contratos a partir do início de 2013.

Uma restrição adicional, nessa etapa de Seleção de Contratos, foi adicionada devido a uma política interna da OPS. Essa restrição desconsidera da base de dados os contratos relacionados a funcionários da empresa, pois, para estes, o plano de saúde é um benefício gratuito, livre de mensalidades.

3.2.2 Limitação de Histórico

O objetivo dessa etapa é limitar o histórico dos contratos selecionados na etapa Seleção de Contratos, pois existe uma diferença contextual entre contratos novos e contratos antigos dentro da OPS, como por exemplo, entre beneficiários cuja data de adesão é o início de 2013 e o final de 2015. Beneficiários recém-chegados ao plano de saúde (final de 2015) podem não ter tido tempo suficiente para avaliar, de forma positiva ou negativa, os serviços oferecidos pela empresa e, consequentemente, iniciar um processo de cancelamento; entretanto, beneficiários com mais histórico no plano de saúde tem mais chance de já terem passado por experiências boas o suficiente para se tornarem clientes fidedignos ou experiências desagradáveis que podem influenciar no cancelamento do contrato.

Outro fator importante que influenciou na decisão de limitação de histórico está relacionado a atributos numéricos acumulativos, como por exemplo a “quantidade de ligações telefônicas entre o cliente e a OPS”. Quanto maior a permanência de um cliente no plano de saúde, maior será o valor esperado para esse tipo de atributo; portanto, comparar

contratos com períodos de permanência bastante diferentes, poderia comprometer a aprendizagem de um modelo de classificação capaz de separar contratos ativos de contratos cancelados.

Apesar de um histórico limitado evitar uma disparidade entre beneficiários recentes e beneficiários antigos, existe o efeito colateral negativo de se ignorar parte da história conhecida de um contrato. A Figura 8 mostra um exemplo desse efeito colateral, no qual o período máximo de histórico é de 6 meses para dois contratos fictícios: Contrato A e Contrato B. O Contrato A tem data de adesão em 01/07/2013 e data de cancelamento em 01/08/2014. O Contrato B tem data de adesão em 01/09/2013 e data de cancelamento em 01/02/2014. É possível perceber que apesar de ambos os contratos estarem com estado de "cancelado" antes de entrarem na etapa, apenas o Contrato B permaneceu nesse estado após ter seu histórico limitado. O Contrato A passou a ser considerado "não cancelado", pois houve uma diferença de 13 meses entre a adesão e o cancelamento, diferença superior ao limite de 6 meses, implicando que para as próximas etapas e fases da abordagem, esse contrato não será mais considerado "cancelado".

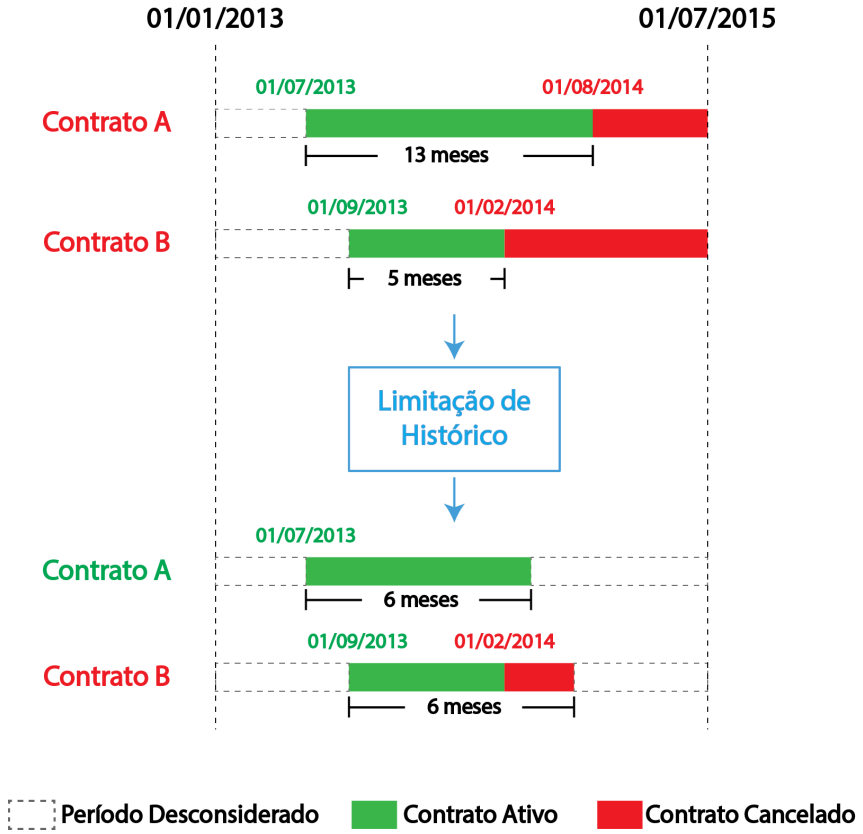

Figura 8 – Exemplificação do processo executado pela etapa Limitação de Histórico.

Um passo essencial para a limitação do histórico é definir o período máximo entre a data de adesão e a data limite final. Como base para essa escolha apresenta-se a Figura 9, por meio da qual é possível verificar o gráfico da percentagem acumulada de cancelamentos entre 1 mês e 18 meses após a data de adesão, para contratos cancelados entre 2013 e 2015.

É importante frisar que esse gráfico representa apenas o espaço amostral dos contratos já cancelados, portanto não se pode concluir que, por exemplo, 40% de todos os contratos da OPS são cancelados em até 6 meses.

Analisando a Figura 9, optou-se, na abordagem proposta, em fixar o período máximo em 12 meses, representando assim 80% dos contratos cancelados. Dessa forma, para a especificação dos contratos e construção dos modelos preditivos, são utilizados até 1 ano de dados do histórico. Um período maior que 12 meses não foi escolhido porque quanto maior é esse período, maior é o tempo de histórico ideal (no mínimo igual ao período escolhido) que um contrato deve ter.

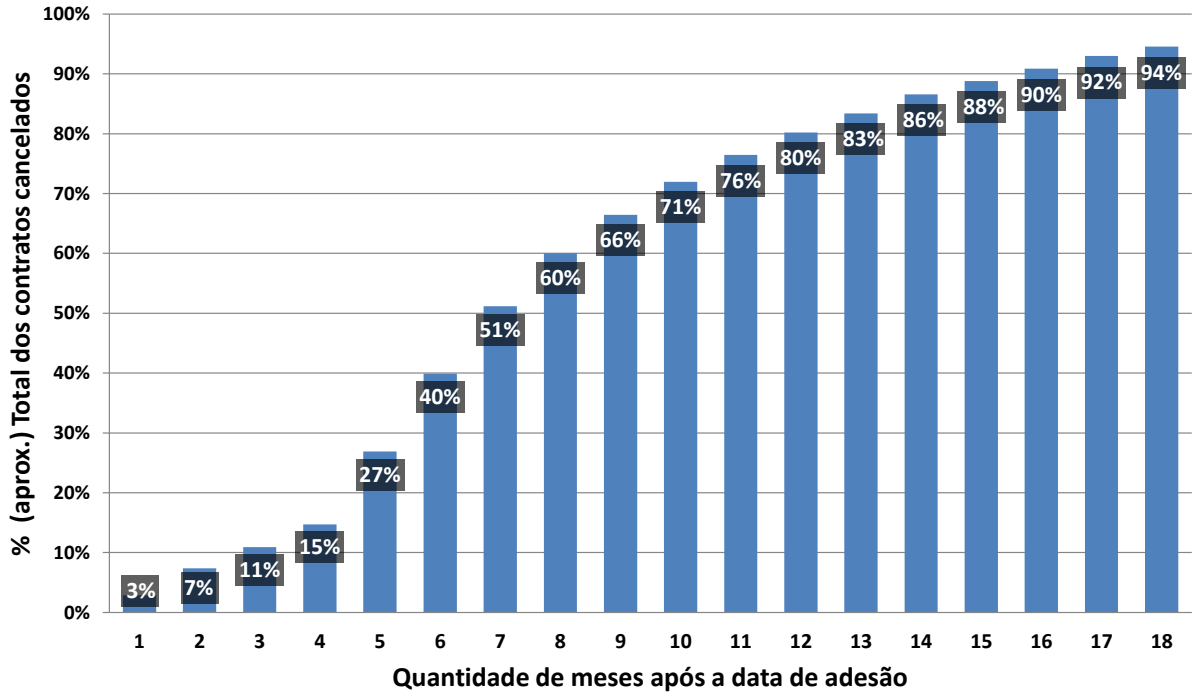

Figura 9 – Percentagem acumulada do total de contratos cancelados, de acordo com o número de meses entre a data de adesão e a data de cancelamento.

3.2.3 Limpeza de Dados

Nessa etapa são removidos todos os atributos que podem de alguma forma prejudicar a fase de Mineração de Dados, representando dados de má qualidade. Para cada uma das 5 tabelas presentes na base de dados foram definidos 6 grupos de atributos a serem removidos: poluídos, preenchidos com valor padrão, duplicado ou redundante, irrelevante, ético/legal e correlato.

3.2.3.1 Atributos Poluídos

Atributos poluídos representam colunas que são preenchidas sem seguir um padrão estabelecido. Na Tabela 2 segue o exemplo de um atributo “idade” preenchido de forma a

poluir o conteúdo, assumindo que os valores esperados para esse atributo deveriam ser apenas números.

Tabela 2 – Exemplo de atributo poluído.

ID	Idade
1	“10”
2	“velho”
3	“menor de idade”
4	“completou 20”
5	“solteiro”

3.2.3.2 Atributos Preenchidos com Valor Padrão

Atributos preenchidos com valor padrão podem representar, por exemplo, atributos não utilizados pelo sistema durante algum processo. Dessa forma, todas as instâncias da tabela que contêm essa coluna possuem o valor preenchido de acordo com o tipo de dados do atributo. Na Tabela 3 segue o exemplo de um atributo “preço”, do tipo numérico, preenchido com valor padrão.

Tabela 3 – Exemplo de atributo preenchido com valor padrão.

ID	Preço
1	0.0
2	0.0
3	0.0
4	0.0
5	0.0

3.2.3.3 Atributos Duplicados ou Redundantes

Atributos duplicados ou redundantes representam colunas que contêm informações presentes em outras colunas. Na Tabela 4 segue o exemplo de um atributo “valor final”, e de um atributo “resultado”, que possuem os mesmos valores em todas as instâncias.

3.2.3.4 Atributos Irrelevantes

Atributos irrelevantes representam colunas que contêm informações não consideradas úteis para a resolução do problema. Essa atribuição de relevância está relacionada a

Tabela 4 – Exemplo de atributos duplicado ou redundante.

ID	Valor Final	Resultado
1	13.5	13.5
2	1.7	1.7
3	5.4	5.4
4	120.0	120.0
5	86.9	86.9

Tabela 5 – Exemplo de atributo irrelevante.

ID	PessoaId
1	10
2	12
3	16
4	19
5	23

uma tarefa direta do especialista. Na Tabela 5 segue o exemplo de um atributo "pessoaId" contendo informação considerada irrelevante.

3.2.3.5 Atributos Legais/Éticos

Atributos legais/éticos representam colunas que contêm informações de cunho sigiloso ou privado. Esse tipo de informação geralmente está relacionado a dados pessoais, como nome, endereço, dados bancários, etc. Na Tabela 6 segue o exemplo de um atributo "nome" contendo informação de natureza legal.

Tabela 6 – Exemplo de atributo com informação legal/ética.

ID	nome
1	"João Emilio"
2	"Carlos Tereza"
3	"Tereza Cristina"
4	"Paulo Cardoso"
5	"Maria de Fátima"

3.2.3.6 Atributos Correlatos

Atributos correlatos são aqueles que carregam dados não exatamente duplicados ou redundantes mas que estatisticamente representam a mesma informação. Para identificar pares de atributos correlatos é utilizado o coeficiente de Pearson (BENESTY et al., 2009)

para atributos numéricos e o coeficiente *phi* de Cramer (AGRESTI, 2013) para atributos categóricos. A Tabela 7 mostra um exemplo de correlação entre dois atributos numéricos, é possível notar que o atributo "Valor2" é sempre duas vezes maior que o atributo "Valor1", portanto um dos dois poderia ser removido da base de dados.

Tabela 7 – Exemplo de atributo com dados correlatos.

Valor1	Valor2
1	2
2	4
3	6
4	8
5	10

3.2.4 Construção de Atributos

Essa etapa é importante para adicionar informação aos contratos, pois nem tudo aquilo que pode ser considerado útil está diretamente mapeado a uma tabela ou coluna do banco de dados. Um exemplo pode ser visto na Tabela 8, na qual a criação do atributo "idade" pode ser mais representativa para a fase de Mineração de Dados do que utilizar a coluna "nascimento" com valores do tipo data.

Tabela 8 – Exemplo da construção de um atributo "idade" baseado no atributo "nascimento".

ID	Nascimento	Idade
1	22/10/1991	24
2	25/11/1980	35
3	10/09/2000	15
4	07/09/1999	16
5	29/12/1993	22

É nessa etapa que informações dos dependentes, ignoradas na etapa de Seleção de Contratos, são adicionadas ao contrato dos respectivos titulares. Julgou-se importante adicionar dados dos dependentes porque, apesar de não serem responsáveis pelos pagamentos, os dependentes representam uma extensão do contrato do titular, seja influenciando no valor total da mensalidade a ser paga, seja na própria utilização da cobertura do plano de saúde. A quantidade de dependentes de um contrato titular é um exemplo básico de atributo que pode ser construído para expandir a informação contida no contrato, pois esse dado não está presente de forma explícita na base de dados original cedida pela OPS.

3.2.5 Seleção de Atributos

Antes de passar por essa etapa, cada instância que representa um contrato ainda possui atributos que não se encaixam em nenhum grupo da etapa de Limpeza de Dados e que foram adicionados na etapa de Construção de Atributos, acarretando em um alto número de dimensões para o contrato. Como um valor elevado de dimensões provoca um crescimento exponencial do espaço que representa uma determinada instância (CHEN; MONTGOMERY; BOLUFÉ-RöHLER, 2015), o papel da etapa Seleção de Atributos é reduzir a dimensionalidade do problema. Para isso é aplicada uma estratégia de seleção de atributos do tipo filtro, na qual a métrica utilizada como divisa é a razão de ganho. Apenas os atributos que tiverem razão de ganho superior a zero serão mantidos na instância.

3.3 Mineração de Dados

A fase de Mineração de Dados explora a massa de dados oriunda do Pré-Processamento a fim de descobrir padrões, relacionamentos e tendências que promovam a descoberta de conhecimento (SIM; KWON; LEE, 2016). É nessa fase que as razões para saída do beneficiário e o tempo estimado até essa possível saída são inferidos. As etapas que constituem a fase de Mineração de Dados são: Definição do Classificador, Definição do Regressor, Identificação de Características, Refinamento do Modelo, Reconhecimento de Contratos e Previsão de Cancelamento.

3.3.1 Definição do Classificador

Essa etapa visa determinar os classificadores mais adequados à abordagem proposta e para o domínio de dados em questão. O objetivo da classificação é determinar, para contratos ativos, aqueles que possuem características de contratos cancelados. São realizados experimentos com algoritmos de diversos paradigmas de aprendizado, cujo resultado é avaliado a partir da métrica AUC. Essa métrica é levada em consideração por ser comumente adotada na literatura para comparar a performance entre classificadores, além de avaliar o comportamento de um modelo classificador sob a ótica das métricas *recall* e taxa de falso positivo (ZHANG et al., 2015; ROUHI; JAFARI, 2016). Os algoritmos utilizados para os experimentos são os seguintes:

- Paradigma Baseado em Árvore de Decisão: C4.5, *RandomForest* e CART;
- Paradigma Bayesiano: *Naive Bayes* e *BayesNet*;
- Paradigma Baseado em Exemplos: KNN e K*;
- Paradigma Conexionista: MLP;
- Paradigma Estatístico: SVM.

Os experimentos são realizados por meio da utilização de toda a base de dados disponível, oriunda da fase de Pré-Processamento, e a métrica final (AUC) é calculada a partir da média obtida na validação cruzada (divisão em 10 grupos).

3.3.2 Definição do Regressor

Essa etapa possui um objetivo semelhante da etapa Definição do Classificador, porém, ao invés de se avaliar os classificadores são avaliados os regressores. O papel da regressão nesse caso é estimar o tempo restante de um contrato dentro do plano de saúde, caso esse contrato possua características de um contrato cancelado. Algoritmos também de variados paradigmas são avaliados, sob a ótica da métrica RMSE, popularmente adotada como critério de comparação entre regressores (ACAR, 2015). Os algoritmos experimentados nessa etapa são os seguintes:

- Paradigma Baseado em Árvore de Decisão: M5;
- Paradigma Baseado em Exemplos: KNN e K*;
- Paradigma Conexionista: MLP;
- Paradigma Estatístico: SVM e Regressão Linear.

Nessa etapa os experimentos são realizados apenas com os contratos cancelados existentes na base de dados, pois a partir desse tipo de contrato pode-se extrair o tempo entre a data de adesão e a data de cancelamento. A métrica final (RMSE) também é calculada a partir da média obtida na validação cruzada (divisão em 10 grupos). Ressalta-se que tanto nessa etapa como na etapa Definição do Classificador, os algoritmos são executados obedecendo a configuração padrão da ferramenta de apoio utilizada - a ferramenta WEKA (WEKA, 2016).

3.3.3 Identificação de Características

Essa etapa, paralela à etapa de Reconhecimento de Contratos, visa identificar quais ações, comportamentos ou padrões, levam o titular a cancelar seu contrato. Essa identificação é guiada pela análise dos ramos gerados por um classificador baseado em árvore de decisão; o classificador escolhido para essa tarefa é o melhor algoritmo baseado em árvore de decisão da etapa Definição do Classificador.

Além de estarem entre os tipos de classificadores mais poderosos e populares (SOK et al., 2016), as árvores de decisão podem ter seu modelo de aprendizado facilmente interpretado (DIAZ; THEODOULIDIS; DUPOUY, 2016). Visto isso, ao se utilizar árvores de decisão para classificar os contratos, pode-se investigar, por meio do modelo gerado, quais caminhos/ramos são mais relevantes para descrever a classe “cancelado”. Nessa etapa

de Identificação de Características, os ramos que possuem as maiores taxas de acerto são identificados, a fim de investigar o que cada atributo, contido em um desses ramos da árvore de decisão, representa no contexto do cancelamento. A taxa de acerto, Equação 3.1, equivale a precisão de um ramo, ou seja, é calculada como a razão entre o número de contratos corretamente classificados (VP) e o número total de contratos rotulados por um determinado ramo (VP + FP).

$$taxa\ de\ acerto = precisão\ (do\ ramo) = \frac{VP}{VP + FP} \tag{3.1}$$

3.3.4 Refinamento do Modelo

O Refinamento do Modelo consiste em ajustar as configurações/parâmetros de um algoritmo visando otimizar uma determinada métrica ou um conjunto de métricas. Por exemplo, em um *perceptron* multicamadas, a quantidade de camadas e neurônios em cada camada são parâmetros do algoritmo, em que configurações diferentes podem levar a um melhor ou pior poder de predição. Como o ajuste dos parâmetros não segue uma formulação exata, faz-se necessária a realização de experimentos que visem detectar as melhores configurações disponíveis. Para essa tarefa é utilizado nesse trabalho o algoritmo de otimização baseado em colônia de formigas - do inglês *Ant Colony Optimization* (DORIGO; BIRATTARI, 2011), com o objetivo de maximizar a área sob a curva ROC.

A Figura 10 mostra a estruturação em forma de grafo para se otimizar o algoritmo KNN, tendo como exemplo dois atributos: número de vizinhos e método de distância. A escolha do valor de cada atributo é feita de forma independente em relação aos outros atributos, pois pressupõe-se que não há dependência entre eles. Após a execução do Refinamento do Modelo, o melhor valor encontrado para cada um dos atributos é utilizado para compor a configuração final do algoritmo, que no caso do exemplo da Figura 10 seria o número de vizinhos igual a 3 e a distância euclidiana como o método de distância.

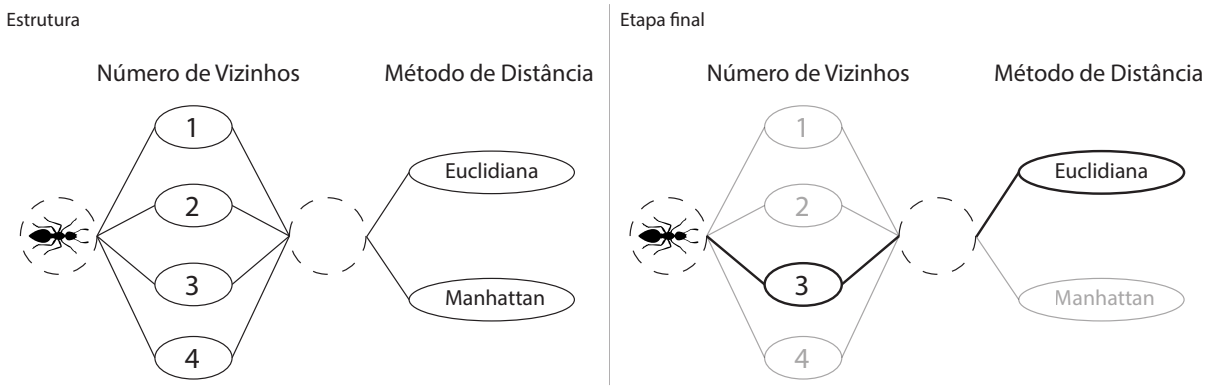

Figura 10 – Estrutura em forma de grafo para refinamento do algoritmo KNN.

É possível notar pela estrutura da abordagem proposta (vide Figura 7) que a etapa Refinamento do Modelo está conectada tanto à etapa Definição do Classificador como à

Definição do Regressor. Essa conexão visa identificar os dois melhores algoritmos (de forma separada entre classificação e regressão), refiná-los e então combiná-los por meio da técnica *stacking* (ZHU, 2010), obtendo-se ao final um classificador resultante da combinação entre os dois melhores classificadores e um regressor resultante da combinação entre os dois melhores regressores. Ressalta-se que a escolha de apenas dos dois melhores algoritmos, e não de um número maior, é realizada de forma empírica. A técnica *stacking* é utilizada devido a sua capacidade de lidar com algoritmos diferentes e de ser a base para importantes estudos sobre métodos de combinação (NGUYEN et al., 2016; ABUROMMAN; Ibne Reaz, 2015). Para o algoritmo de saída do *stacking*, responsável por combinar o modelo preditivo dos algoritmos de entrada, é escolhido o MLP. O MLP foi adotado de forma empírica, pois apresenta resultados expressivos na literatura na solução de problemas não-lineares, sem a necessidade de um conhecimento prévio sobre a relação entre a entrada e a saída (KORDOS; RUSIECKI, 2015; GALESHCHUK, 2016). A Figura 11 exemplifica mais claramente como é o processo geral da etapa Refinamento do Modelo.

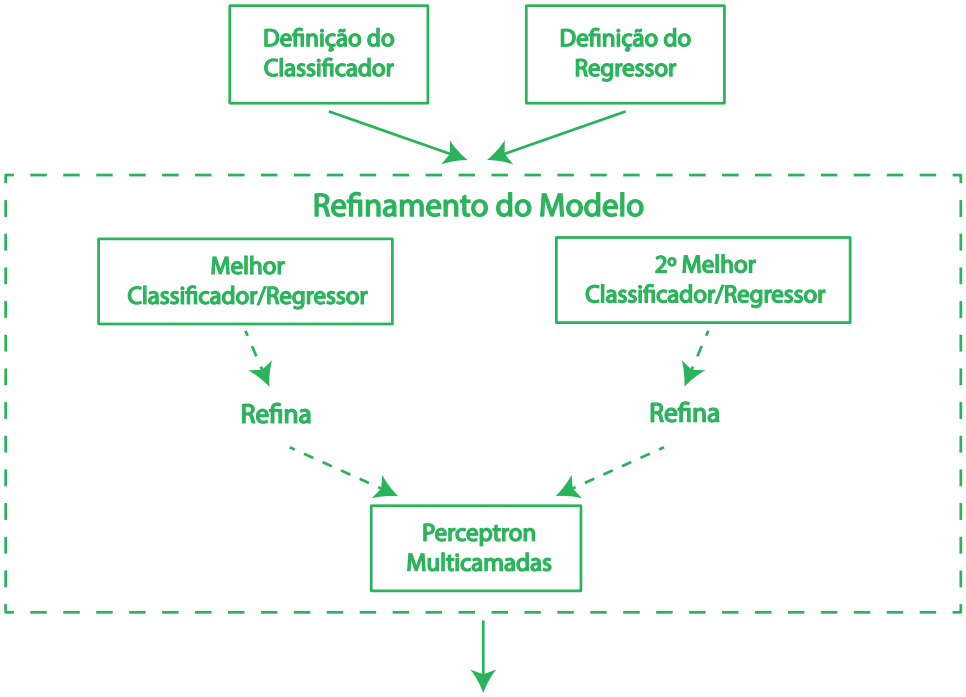

Figura 11 – Processo de funcionamento da etapa Refinamento do Modelo.

Uma parte importante do refinamento da classificação consiste em balancear o conjunto de treinamento. Como a frequência de contratos ativos é maior que a de contratos cancelados, a quantidade de elementos entre essas duas classes possui diferenças significativas. Visto que o desbalanceamento entre as classes pode influenciar em uma performance menor dos classificadores (GALAR et al., 2012), um dos objetivos do refinamento é definir uma estratégia para balancear a base de dados a ser utilizada na etapa Reconhecimento de Contratos. Portanto, especificamente para modelos de classificação, além de encontrar o melhor conjunto de parâmetros para um determinado algoritmo, a etapa Refinamento

do Modelo determina também a melhor técnica de balanceamento, dentre as seguintes técnicas: *undersampling* aleatório (GARCÍA; HERRERA, 2009), *oversampling* aleatório (GARCÍA; HERRERA, 2009), SMOTE (CHAWLA et al., 2002) e medóides (do inglês *k-medoids undersampling*) (DUBEY et al., 2014).

3.3.5 Reconhecimento de Contratos

Essa etapa visa reconhecer contratos ativos que possuem características de contratos anteriormente cancelados. Para isso é efetuada uma classificação com o objetivo de rotular um contrato ainda ativo com a classe "cancelado" caso haja valores de atributos comuns a um contrato cancelado, ou a classe "não cancelado" caso não haja indícios de características de cancelamento. Ressalta-se que o classificador utilizado nessa etapa é resultante da combinação, por meio de *stacking*, dos dois melhores classificadores da etapa Refinamento do Modelo.

A etapa Reconhecimento de Contratos desempenha um papel importante dentro da abordagem proposta, pois é partir desse ponto que os contratos são separados entre aqueles que não representam risco de cancelamento e aqueles que representam. Após essa separação, é possível focar apenas no grupo de contratos rotulados como "cancelado" e então aprimorar a informação que é entregue a gestão da OPS.

3.3.6 Previsão de Cancelamento

Essa etapa tem como objetivo estimar o tempo até o cancelamento de um cliente, ou seja, para cada contrato rotulado como "cancelado" realiza-se uma regressão para se obter um prazo até que o titular decida cancelar seu vínculo com o plano de saúde. Esse prazo é estimado em meses e experimentos são realizados visando demonstrar a performance da regressão. Enfatiza-se que o regressor utilizado nessa etapa é resultante da combinação, por meio de *stacking*, dos dois melhores regressores da etapa Refinamento do Modelo.

A etapa Previsão de Cancelamento possui um papel complementar à etapa Reconhecimento de Contratos, pois cada contrato além do rótulo de "cancelado" passa a ter uma estimativa do tempo até a efetivação do cancelamento; esse tempo estimado serve como informação adicional para a fase de Priorização de Contratos, pois ajuda na identificação de quais contratos são mais importantes para a gestão da OPS por estarem mais próximos de serem cancelados.

3.4 Priorização de Contratos

Por fim, a fase de Priorização de Contratos tem como objetivo priorizar os contratos rotulados como "cancelado" de acordo com um conjunto específicos de atributos, descritos na única etapa dessa fase: Ordenação de Contratos.

3.4.1 Ordenação de Contratos

Essa etapa realiza um papel importante na entrega da informação à gestão, pois se houvesse apenas um conjunto com todos os contratos rotulados como "cancelado", o gestor ficaria encarregado de manualmente priorizar quais contratos devem ser analisados inicialmente. A etapa Ordenação de Contratos é responsável por automaticamente organizar os contratos do mais crítico ao menos crítico, visando melhorar e facilitar as ações da gestão; um contrato é mais crítico quando o risco de cancelamento é maior e ele é mais relevante para o caixa financeiro da OPS.

A ordenação dos contratos obedece a seguinte sequência de atributos:

1. **Grau de certeza da classificação:** probabilidade associada ao rótulo "cancelado", oriunda da etapa Reconhecimento de Contratos. Os contratos são ordenados do maior para o menor grau de certeza;
2. **Tempo estimado até o cancelamento:** tempo estimado na etapa Previsão de Cancelamento. Os contratos são ordenados do menor para o maior tempo estimado;
3. **Valor da mensalidade:** valor mensalmente pago pelo titular ao plano de saúde. Os contrato são ordenados do maior para o menor valor da mensalidade.

Ao seguir essa ordenação o gestor pode acompanhar primeiramente os contratos cuja probabilidade de serem realmente cancelados é maior, e caso essa probabilidade seja igual entre dois contratos passa-se a considerar o tempo estimado até o cancelamento. Contratos que tiverem um tempo estimado menor devem ser priorizados, pois é esperado que eles sejam os primeiros a cancelarem seus contratos. E em última situação, caso dois contratos tenham o mesmo grau de certeza e o mesmo tempo estimado, ordena-se pelo maior valor da mensalidade, porque é mais vantajoso para o plano investir esforços e dinheiro para manter um contrato que contribui mais para o caixa da OPS do que para um contrato que contribui menos. Ressalta-se que, essa ordem nos atributos é apenas pré-estabelecida na abordagem, ou seja, o gestor da OPS pode, de forma simplificada, analisar os contratos a partir de outros critérios.

4 Resultados e Discussões

Nesse capítulo são apresentados os resultados e as discussões referentes a aplicação da abordagem proposta na base de dados cedida pela OPS. Esses resultados e discussões são divididos em seções homônimas às fases da abordagem, respectivamente: Fase de Pré-Processamento (Seção 4.1), Fase de Mineração de Dados (Seção 4.1) e Fase de Priorização de Contratos (Seção 4.3). Como suporte para a execução dos algoritmos de AM é utilizada a ferramenta WEKA (WEKA, 2016) (do inglês *Waikato Environment for Knowledge Analysis*). Essa ferramenta é largamente aceita e utilizada na academia e na indústria, como um instrumento de referência no processo para descoberta de conhecimento (WITTEN; FRANK; HALL, 2011). Outro dois fatores contribuíram fortemente para essa escolha: facilidade de realizar alterações na execução e parametrização dos algoritmos; flexibilidade oferecida para se executar os algoritmos e manipular os resultados de forma separada da interface gráfica original.

Todo o código necessário para importação da base de dados, execução dos algoritmos, compilação dos resultados e geração de dados para análise é desenvolvido na linguagem de programação JAVA (JAVA, 2016), devido à compatibilidade com a ferramenta WEKA. Adicionalmente é utilizado o Eclipse (ECLIPSE, 2016) como plataforma de desenvolvimento e o repositório Gitlab (GITLAB, 2016) para o controle de versão dos elementos produzidos.

4.1 Fase de Pré-Processamento

Antes de iniciar a execução da fase de Pré-Processamento, a base de dados apresenta um total de 125.715 contratos; todos esses contratos são entradas para a etapa Seleção de Contratos. A Figura 12 apresenta a redução no número de contratos da base de dados, a partir da aplicação de cada uma das seguintes restrições: apenas contratos individuais, sem contratos de funcionários, somente contratos de titulares e contratos com data de adesão superior a 01 de Janeiro de 2013. Ao final, é possível notar que a quantidade de contratos é reduzida a 23.911 elementos, uma redução de quase 81% da quantidade original.

Após a etapa Limitação de Histórico, são removidos dos contratos os atributos que podem prejudicar a classificação. Cada contrato pode ser visto como um longo conjunto de colunas, totalizando 230 atributos. A Figura 13 mostra a redução na quantidade de atributos após a remoção dos seis grupos estabelecidos pela etapa Limpeza de Dados.

Os contratos, após a limpeza de dados, tiveram 219 atributos removidos, restando para as etapas posteriores um total de 11 atributos. Nota-se que, a maior parte dos atributos retirados pertence ao grupo de atributos irrelevantes, devido principalmente às

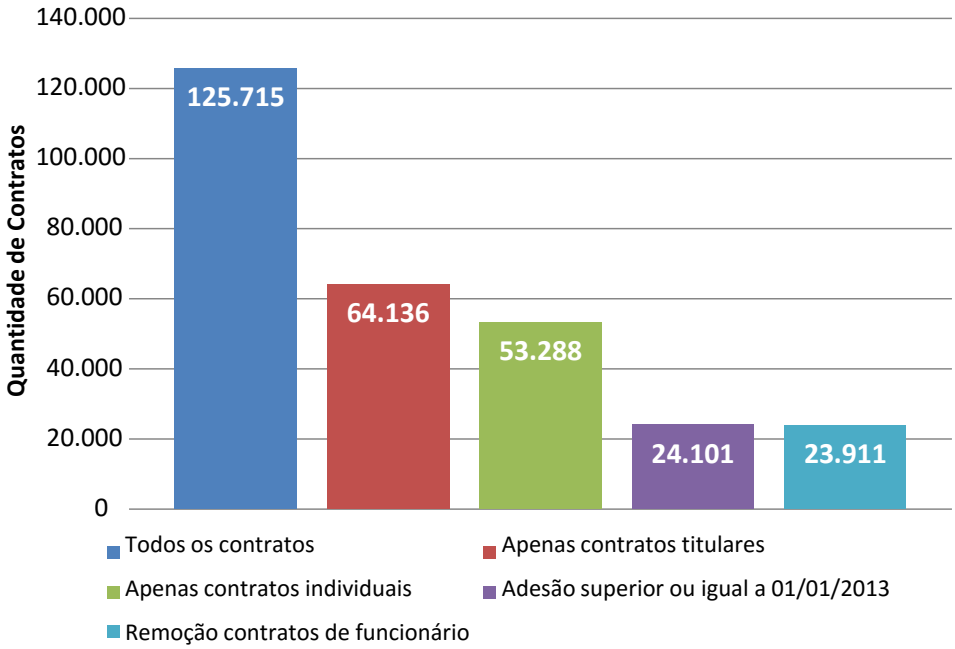

Figura 12 – Redução na quantidade de contratos efetuada pela etapa Seleção de Contratos.

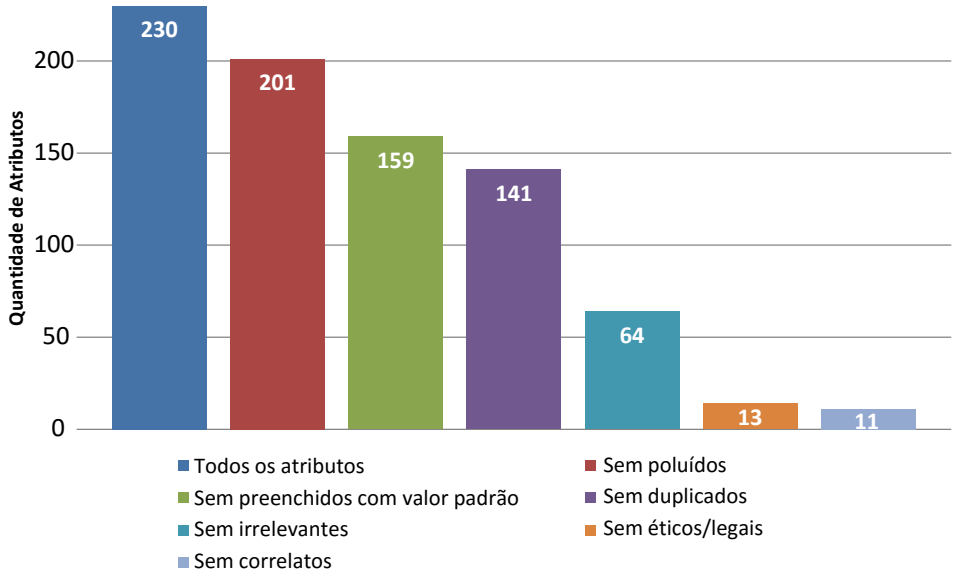

Figura 13 – Redução na quantidade de atributos efetuada pela etapa Limpeza de Dados.

colunas que representam chaves estrangeiras e valores sem representação para o contexto do cancelamento de contratos. O segundo grupo que obteve mais redução de elementos foi o de atributos preenchidos com valor padrão; deve-se isso ao longo tempo de funcionamento do sistema, desde 2005, o que acarretou em uma série de mudanças estruturais no banco de dados, tornando obsoletas diversas tabelas e colunas. Os atributos restantes são os seguintes:

1. **"beneficiario":** [nominal] representa se o titular do contrato é um beneficiário do plano de saúde. Pode assumir os valores: "SIM" e "NAO";

2. **“diabase”:** [numérico] representa qual o dia do mês foi escolhido pelo titular para realizar o pagamento da mensalidade;

3. **“estadocivil”:** [nominal] representa qual o estado civil do titular do contrato. Pode assumir os valores: “VIUVO”, “SOLTEIRO”, “CASADO”, “SEPARADO” e “OUTRO”;

4. **“faixapagamento”:** [nominal] representa em qual faixa de pagamento se encaixa o contrato. Pode assumir os valores: “0”, “1”, “2” e “3”;

5. **“idempresaterceirizada”:** [nominal] representa qual setor da OPS foi responsável pela efetivação do contrato. Pode assumir os valores: “1”, “2”, “3”, “4”, “5”, “6”, “7”, “8” e “9”;

6. **“idproduto”:** [nominal] representa qual o conjunto de coberturas está associado ao contrato. Pode assumir os valores: “1”, “25” e “26”;

7. **“seguradoodonto”:** [nominal] representa se o titular é um beneficiário da parte odontológica do plano de saúde. Pode assumir os valores: “SIM” e “NAO”;

8. **“sexo”:** [nominal] representa o sexo do titular. Pode assumir os valores: “Mulher” e “Homem”;

9. **“tipopagamento”:** [nominal] representa qual o tipo de pagamento padrão do contrato. Pode assumir os valores: “CARTAO”, “BOLETO”, “CONTA_CORRENTE” e “FOLHA_DE_PAGAMENTO”;

10. **“valor”:** [numérico] representa o valor referente ao contrato do titular do contrato, não estando inclusos os valores de possíveis dependentes;

11. **“valortotalcontrato”:** [numérico] representa o valor total do contrato, incluindo o valor do titular e dos dependentes.

Após a etapa Limpeza de Dados, 7 novos atributos foram adicionados ao contrato na etapa Construção de Atributos. Os atributos adicionados são os seguintes:

1. **“idade_media”:** [numérico] representa a idade média dos beneficiários envolvidos, sendo eles os dependentes de um titular e o próprio titular (caso este seja um beneficiário);

2. **“qtd_atendimentos”:** [numérico] representa a quantidade de atendimentos realizados pelo titular ou pelos dependentes;

3. **“qtd_dependentes”:** [numérico] representa a quantidade de dependentes do contrato;

4. **“qtd_ocorrencias”:** [numérico] representa a quantidade de contatos telefônicos realizados entre o titular, ou seus dependentes, com atendentes da OPS. Esses contatos podem significar dúvidas, sugestões ou reclamações provenientes dos beneficiários envolvidos, e também podem representar contatos iniciados pelas atendentes da OPS a fim de divulgar informações ou realizar cobranças aos titulares;

5. **“tipo_diabase”:** [nominal] representa qual período do mês foi escolhido pelo titular para realizar o pagamento da mensalidade. Pode assumir os valores: “INICIO_MES”, “MEIO_MES” e “FIM_MES”;

6. **“ultimo_atendimento_dias”:** [numérico] representa a quantidade de dias corridos entre a data do último atendimento realizado e a data limite do contrato (12 meses após a data de adesão);

7. **“valor_atendimentos”:** [numérico] representa o valor total dos atendimentos já realizados pelos dependentes e pelo titular (quando este é um beneficiário).

Após as etapas Limpeza de Dados e Construção de Atributos totalizam-se 18 atributos, dos quais apenas o atributo “diabase” foi removido na etapa Seleção de Atributos, por possuir valor de razão de ganho igual a zero.

4.2 Fase de Mineração de Dados

A etapa Definição do Classificador visa comparar a performance dos classificadores selecionados para rotular um contrato como “cancelado” ou “não cancelado”. A Tabela 9 evidencia o resultado obtido por cada classificador, ordenados da melhor para a pior performance. É possível notar que os dois melhores algoritmos representam dois paradigmas distintos, sendo eles, respectivamente, o bayesiano e o baseado em árvores de decisão. O algoritmo menos performático foi o KNN que, mesmo utilizando diferentes valores para K (1, 3 e 5) não superou os outros classificadores. Os algoritmos *BayesNet* e CART são utilizados para gerar um modelo de classificação mais robusto, abordado na etapa Refinamento do Modelo.

Na etapa Definição do Regressor, o objetivo é elencar os melhores regressores para estimar o tempo de permanência no plano de saúde dos contratos rotulados como “cancelado”. A Tabela 10 apresenta os resultados obtidos nesses experimentos, demonstrando que os dois algoritmos com menores valores para RMSE foram o M5 e Regressão Linear. Ambos os algoritmos apresentaram um RMSE próximo a 2,3, ou seja, em média, o erro na estimação do tempo de permanência é de cerca de 2 meses para mais ou para menos.

A próxima etapa a ter seus resultados demonstrados é a Identificação de Características. O algoritmo CART foi escolhido para esse momento da abordagem por apresentar,

Tabela 9 – Métrica AUC obtida para cada algoritmo utilizado na etapa Definição do Classificador.

Algoritmo	**AUC**
BayesNet	0,926
CART	0,882
RandomForest	0,861
C4.5	0,853
MLP	0,749
K*	0,748
NaiveBayes	0,662
SVM	0,653
KNN ($N = 5$)	0,624
KNN ($N = 3$)	0,603
KNN ($N = 1$)	0,561

Tabela 10 – Métrica RMSE obtida para os algoritmos da etapa Definição do Regressor.

Algoritmo	***RMSE***
M5	2,312
Regressão Linear	2,366
SVM	2,423
K*	2,706
KNN ($N = 5$)	2,752
MLP	2,860
KNN ($N = 3$)	2,933
KNN ($N = 1$)	3,638

na etapa Definição do Classificador, o melhor resultado dentre os classificadores baseados em árvore de decisão. Para destacar as regras mais relevantes construídas a partir da base de dados, optou-se por aquelas nas quais o acerto relativo ao rótulo ("cancelado" ou "não cancelado") é no mínimo de aproximadamente 5%; ou seja, as regras que definem um contrato cancelado devem abranger uma quantidade de acertos próxima a 303 contratos (5% dos 6063 contratos cancelados) e as regras que definem um contrato ativo devem abranger uma quantidade próxima a 386 contratos (5% dos 7725 contratos ativos presentes na base de dados).

As regras mais relevantes encontradas são:

- "Qtd. Ocorrências <= 2" → "Valor Total >= R$ 60" → "Tipo de Pagamento != Boleto" ↣ **"Não Cancelado" (91,7%: 1170 acertos e 106 erros)**;

  Essa regra demonstra que quando há poucas ocorrências dentro de um contrato, mesmo para mensalidades mais altas, uma parte representativa dos titulares não cancela. Esse baixo número de ocorrências pode indicar que os clientes não possuem graves reclamações ou dúvidas para com o plano de saúde, a ponto de não realizarem

um número expressivo de ligações telefônicas (ocorrências). Outro fato é que esse baixo número também indica que não há muitas interações partindo do plano de saúde para o cliente, como avisos de atraso no pagamento das mensalidades.

- "Qtd. Ocorrências > 2" $\rightarrow$ "Qtd. Atendimentos > 0" $\rightarrow$ "Último Atendimento $<$ 170 dias" $\rightarrowtail$ **"Não Cancelado" (89,6%: 1028 acertos e 119 erros)**;

  Nessa regra, para contratos nos quais o número de ocorrências pode ser alto, os titulares também tendem a não cancelar; porém, diferente da primeira regra, os beneficiários já utilizaram o plano de saúde, pois a quantidade de atendimentos é maior que zero. Esse uso do plano de saúde é importante para firmar as vantagens de ser possuir um seguro, pois mesmo havendo eventuais dúvidas e reclamações, o titular e dependentes já usufruíram de algum atendimento subsidiado pela OPS. É importante notar também que a regra especifica um valor máximo para o tempo transcorrido desde o último atendimento realizado (170 dias), pois quanto maior esse tempo, maior é a chance de os beneficiários do plano de saúde terem esquecido/relevado o fato de já terem efetuado algum atendimento.

- "Qtd. Ocorrências > 2" $\rightarrow$ "Qtd. Atendimentos $= 0$" $\rightarrow$ "Valor total entre R$ 70 e R$ 75" $\rightarrow$ "Idade Média < 16" $\rightarrowtail$ **"Não Cancelado" (86,4%: 388 acertos e 61 erros)**;

  A terceira regra parece ir de encontro às duas primeiras, pois uma significativa parte dos beneficiários não cancelam seus contratos, mesmo quando o número de ocorrências pode ser alto e o de atendimentos é igual a zero. Entretanto, a grande diferença da terceira regra é o baixo valor da idade média. Como a idade média dos envolvidos no contrato é menor que 16 anos, devem existir crianças ou pré-adolescentes como dependentes, motivo esse que pode tornar o cancelamento do contrato uma ação complicada para os pais ou responsáveis.

- "Qtd. Ocorrências entre 2 e 20" $\rightarrow$ "Qtd. Atendimentos $= 0$" $\rightarrow$ "Valor Total $<$ R$ 60" $\rightarrow$ "Produto $= 1$" $\rightarrowtail$ **"Cancelado" (92%: 589 acertos e 51 erros)**;

  Essa regra complementa os comentários da regra anterior, pois os beneficiários tendem a cancelar o contrato quando não há a restrição mencionada da baixa média de idade. No caso da quarta regra, ainda há o agravante do tipo de produto (Produto "1"), que dá direito ao beneficiário de ficar em apartamento isolado em situações de internação. Mesmo a mensalidade não sendo relativamente alta (menor que R$ 60), os clientes podem estar cancelando com o objetivo de firmar um novo contrato, dentro da própria OPS, que possua um produto ainda mais barato, sem direito a um apartamento.

- "Qtd. Ocorrências > 2" $\rightarrow$ "Qtd. Atendimentos $= 0$" $\rightarrow$ "Valor Total entre R$ 65 e R$ 70" $\rightarrowtail$ **"Cancelado" (88%: 397 acertos e 54 erros)**;

Essa regra demonstra que, o já comentado conjunto formado entre um número alto de ocorrências e nenhum atendimento realizado, é relevante para influenciar uma parte representativa dos titulares a cancelarem seus contratos. Um fator adicional a esse conjunto, no caso da quinta regra, é uma faixa de contratos com valores compreendidos entre R$ 65 e R$ 70.

- "Qtd. Ocorrências <= 2" → "Qtd. Atendimento = 0" → "Valor Total entre R$ 20 e R$ 60" → "Tipo de Pagamento = Boleto" ↣ **"Cancelado" (90.5%: 285 acertos e 30 erros)**;

  A sexta regra evidencia que, a não realização de atendimentos parece ser um fator pertinente na decisão de cancelamento do contrato. Com 90% de acerto, essa regra caracteriza contratos cancelados com poucas ocorrências, sem atendimento e com o pagamento por boleto. Nesse tipo de pagamento o débito não é imediato, ou seja, o cliente pode gerar o boleto na própria OPS mas não concluir a transação.

A extração das regras é importante para a gestão porque pode servir como um guia no desenvolvimento de políticas e campanhas dentro da OPS. Percebe-se, pelas regras mais relevantes encontradas, que a não utilização do plano de saúde, por meio de atendimentos subsidiados, parece influenciar na decisão de cancelar o contrato. Outro fator que poderia ganhar destaque é a criação de campanhas mais atrativas para dependentes menores de idade, visto que tal fato parece manter os contratos ativos. De uma forma geral, as regras oferecem uma fundamentação pautada no próprio histórico da OPS, contemplando de maneira mais precisa as características dos beneficiários presentes.

Seguindo o fluxo da abordagem, chega-se a etapa Refinamento do Modelo. A Figura 14 mostra o gráfico da curva ROC comparando a performance entre os dois melhores classificadores da etapa Definição do Classificador (*BayesNet* e CART) e o Modelo Classificador Refinado (MCR) formado a partir desses algoritmos. Pode-se notar que, o modelo refinado possui uma performance melhor que os classificadores isolados, principalmente em relação ao algoritmo CART. A Tabela 11 apresenta o valor da área sob a curva ROC dos três classificadores, resgatando os valores já presentes na Tabela 9. Ressalta-se que a técnica medóides, para tratar o desbalanceamento das classes, obteve os melhores em conjunto ao MCR.

Tabela 11 – Comparação de performance (métrica AUC) dos classificadores *BayesNet*, CART e do MCR formado a partir desses algoritmos.

Algoritmo	**AUC**
MCR	0,950
BayesNet	0,926
CART	0,882

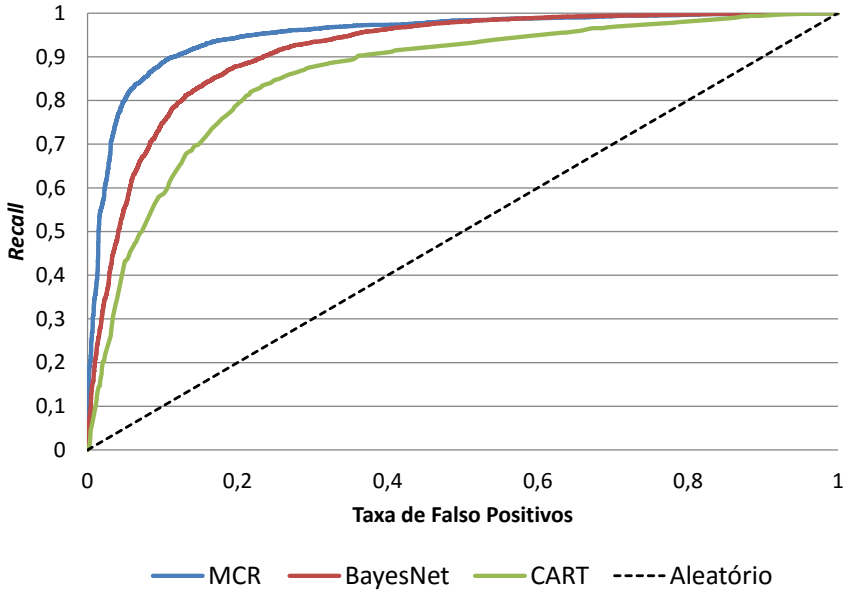

Figura 14 – Curva ROC dos classificadores *BayesNet*, CART e do MCR formado a partir desses algoritmos.

A Figura 15 apresenta o gráfico do ganho acumulado para o MCR aplicado a toda a base de dados. Esse gráfico é interessante pois mostra o total de contratos corretamente rotulados como "cancelado" a partir dos contratos com maior probabilidade de pertencerem a esse rótulo; ou seja, para a gestão da OPS é possível ter uma noção de que percentagem de contratos classificados é necessária para alcançar um determinado percentual de titulares que realmente irão efetuar o cancelamento. Por exemplo, caso a gestão concentrasse esforços apenas na metade dos melhores contratos rotulados como "cancelado", e esse contratos não cancelassem, isso já representaria 90% de todos os iminentes cancelamentos.

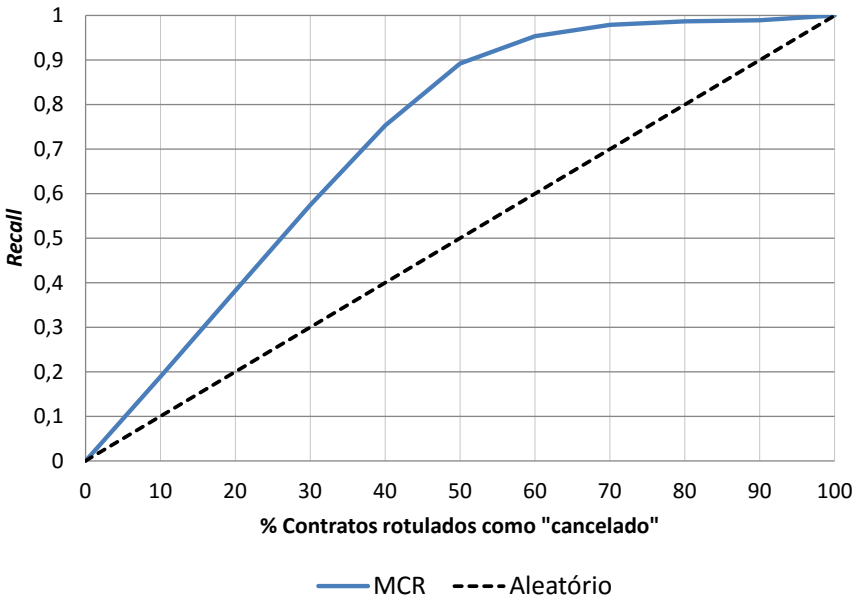

Figura 15 – Gráfico do ganho acumulado para o MCR aplicado a toda a base de dados.

Com base no MCR, é possível dar procedimento a etapa Reconhecimento de Contratos. Como mencionado na abordagem proposta, o objetivo dessa etapa é identificar contratos ativos que possuam características de contratos já cancelados, podendo, a partir disso, atribuir o rótulo "cancelado" ou caso contrário o rótulo "não cancelado". Para a validação da classificação, decidiu-se também utilizar um cenário mais próximo do real, ao invés de apenas uma validação cruzada em toda a base de dados disponível. Como cenário mais realístico, considerou-se o treinamento do classificador com um período T de contratos e a validação com o período seguinte $T+1$; dessa forma, é possível ter uma melhor impressão de como será o funcionamento da abordagem proposta ao ser adotada como ferramenta dentro da OPS.

Os experimentos a seguir visam identificar também o tamanho de período T mais adequado, dentre os seguintes: 3 meses, 4 meses, 6 meses e 12 meses. Toda a base de dados, contendo contratos de 2013 a 2015, é dividida de acordo com cada tamanho de período, originando do período T_0 ao período T_N. A partir disso é calculada a métrica AUC de T_0 (treinamento) com T_1 (validação), T_1 com T_2, T_2 com T_3 até T_{N-1} com T_N. A média das métricas representa o valor final obtido para o período em questão. A Tabela 12 demonstra um exemplo dessa bateria de treinamento/validação caso o período T seja igual a 6 meses.

Tabela 12 – Exemplo da bateria de treinamento/validação para um período T igual a 6 meses.

Treinamento	**Validação**
T_0: 2013.1 - Primeiro Semestre	T_1: 2013.2 - Segundo Semestre
T_1: 2013.2 - Segundo Semestre	T_2: 2014.1 - Primeiro Semestre
T_2: 2014.1 - Primeiro Semestre	T_3: 2014.2 - Segundo Semestre
T_3: 2014.2 - Segundo Semestre	T_4: 2015.1 - Primeiro Semestre
T_4: 2015.1 - Primeiro Semestre	T_5: 2015.2 - Segundo Semestre

O resultado dos experimentos com todos os tamanhos de período destacados é apresentado na Tabela 13. Além da média obtida para métrica AUC, são exibidas as médias das métricas *recall* e precisão. É possível notar que período trimestral é superior a todos os outros períodos na métrica AUC; porém, o período quadrimestral apresenta um resultado mais significativo na métrica *recall*, representando uma opção de período mais interessante nesse quesito. Para execução da abordagem, em um cenário real da OPS, julga-se mais apropriado o uso do período de tempo igual a 4 meses, dessa forma treina-se o MCR com um quadrimestre de contratos e então pode-se rotular, como "cancelado" ou "não cancelado" os contratos do próximo quadrimestre.

A partir de cada contrato rotulado como "cancelado", é possível estimar um tempo (medido em meses) até que o titular responsável efetue o cancelamento; essa tarefa é de responsabilidade da etapa Previsão de Cancelamento. Antes da execução dessa etapa também deve-se refinar o modelo regressor a fim de obter uma aproximação mais correta

Tabela 13 – Resultado da métrica AUC para os experimentos com os tamanhos de período: 3 meses, 4 meses, 6 meses e 12 meses.

Tamanho do período	**Média AUC**	**Média *recall***	**Média precisão**
3 meses	0,773	85,9%	56,4%
4 meses	0,746	88,6%	56,1%
6 meses	0,673	76,7%	53,6%
12 meses	0,643	36,2%	46,7%

do tempo estimado. A Tabela 14 apresenta o valor da métrica RMSE para os dois melhores regressores encontrados na etapa Definição do Regressor, com a adição do Modelo Regressor Refinado (MRR). É possível notar que, apesar de um menor valor para o RMSE, o MRR ainda mantém um erro médio de cerca de 2,3 meses (para mais ou para menos) para o tempo correto até a saída do beneficiário; o que não representa um ganho tão significativo em relação ao algoritmo M5 (o melhor dentre os algoritmos isolados).

Tabela 14 – Comparação de performance (métrica RMSE) dos regressores M5, Regressão Linear e do MRR formado a partir desses algoritmos.

Algoritmo	**RMSE**
MRR	2,302
M5	2,312
Regressão Linear	2,366

Para uma melhor visualização dos resultados da etapa Previsão de Cancelamento, obtidos por meio da aplicação do MRR, a Figura 16 apresenta a taxa de acerto a medida que se incorpora ao tempo estimado uma margem de erro; essa taxa de acerto mede qual o percentual de contratos tiveram seu tempo estimado exatamente igual ao valor esperado. Percebe-se que, utilizando o valor estimado, sem margem de erro, obtém-se apenas um acerto de cerca de 50%, ou seja, somente metade dos contratos tiveram seu tempo até o cancelamento estimado exatamente como o tempo esperado. Adicionando uma margem de erro de dois meses, para mais ou para menos, já se atinge quase 90% dos contratos, mais de 30% de melhoria. Vale ressaltar que, quanto maior a margem de erro, pior é a precisão do tempo estimado para a gestão; pois, por exemplo, se um contrato tem um cancelamento estimado em 6 meses e a margem de erro utilizada também fosse de 6 meses, o titular poderia, hipoteticamente, cancelar a qualquer momento dentro de um período de 1 ano.

Visando também demonstrar a performance do MRR em um cenário mais real, assim como foi feito na etapa Reconhecimento de Contratos, experimentos foram realizados seguindo a mesma estratégia de treinar o modelo com um período T e validar com um período $T+1$. A Tabela 15 apresenta os resultados da métrica RMSE para os tamanhos de período: 3 meses, 4 meses, 6 meses e 12 meses. No caso da etapa Previsão de Cancelamento, o tamanho de período com melhor resultado foi o de 3 meses, no qual o RMSE é de

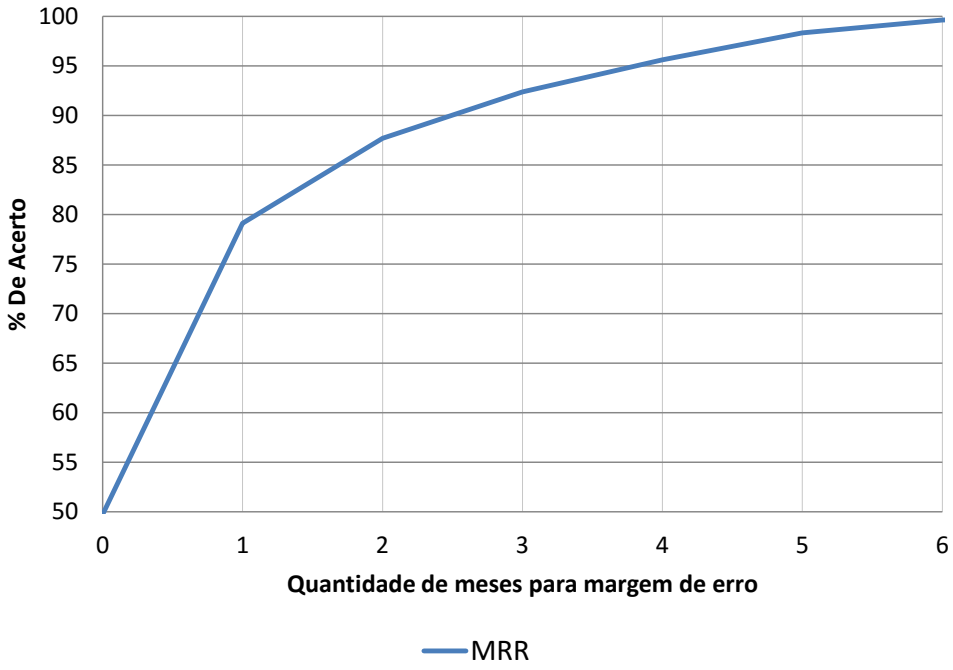

Figura 16 – Gráfico da taxa de acerto do MRR ao se variar a margem de erro em meses.

2,24 meses, um valor até menor que a aplicação do MRR em toda a base de dados. A partir desse resultado, percebe-se que a utilização do MRR, de forma trimestral, parece se manter estável no formato de experimento realizado, no qual treina-se com contratos de um período T e estima-se o tempo de cancelamento para os contratos do período seguinte $(T + 1)$.

Tabela 15 – Média da métrica RMSE para os experimentos com os tamanhos de período: 3 meses, 4 meses, 6 meses e 12 meses.

Tamanho do período	**RMSE**
3 meses	2,243
4 meses	2,274
6 meses	2,353
12 meses	2,369

É interessante notar que, tanto para a etapa Reconhecimento de Contratos como para a etapa Previsão de Cancelamento, os tamanhos de período maiores (6 meses e 12 meses) apresentaram os piores resultados para as métricas avaliadas; um possível motivo para isso é a formação de um conjunto de treinamento bastante diversificado, envolvendo contratos potencialmente influenciados por eventos sazonais distintos. Um tamanho de período menor não garante que os contratos envolvidos sejam semelhantes, mas tende a evitar uma variação muito grande por selecionar uma quantidade menor de contratos.

4.3 Fase de Priorização de Contratos

Na fase de Priorização de Contratos, são utilizados somente os contratos rotulados como "cancelado" a partir da etapa Reconhecimento de Contratos. A Tabela 16 apresenta um exemplo do resultado produzido pela etapa Ordenação de Contratos. Assim como já descrito na abordagem proposta, essa etapa ordena os contratos seguindo três fatores selecionados: grau de certeza da classificação (do maior para o menor valor), tempo estimado até o cancelamento (do menor para o maior valor) e valor da mensalidade (do maior para o menor valor). A ideia dessa ordenação é entregar para a gestão da OPS um conjunto de contratos priorizados, evidenciando aqueles que, são mais prováveis de cancelarem, com o menor tempo até a efetivação do cancelamento e de maior contribuição financeira para o caixa da OPS.

Tabela 16 – Exemplo dos 10 contratos mais prioritários seguindo as diretrizes da etapa Ordenação de Contratos.

#	**Grau de certeza**	**Tempo estimado**	**Valor da Mensalidade**
1	100%	4,8 meses	R$ 50,72
2	100%	5,4 meses	R$ 91,96
3	95%	4,2 meses	R$ 60,44
4	90%	3,6 meses	R$ 82,43
5	90%	5 meses	R$ 105,57
6	90%	5 meses	R$ 80,44
7	88%	6,1 meses	R$ 55,10
8	85%	4,7 meses	R$ 68,96
9	85%	7 meses	R$ 99,59
10	80%	3,6 meses	R$ 103,42

5 Conclusão

Neste trabalho foi proposta uma abordagem para caracterizar o cancelamento eletivo de contratos em planos de saúde privados, visando distinguir aspectos, padrões e propriedades que possam moldar, baseado em eventos passados, o perfil de um contrato cancelado. A definição dessa abordagem foi realizada por meio de etapas, que por sua vez pertencem às seguintes fases: Pré-Processamento, Mineração de Dados e Priorização de Contratos.

Na fase de Pré-Processamento, foi reduzida a quantidade de contratos da base de dados cedida em quase 81%, o que permitiu a análise de um conjunto de contratos mais relacionado à abordagem. Além disso, houve uma redução ainda maior na dimensionalidade do problema, pois menos de 5% dos 230 atributos disponíveis foram selecionados. Adicionalmente, 7 novos atributos foram construídos, pois representam informações, julgadas relevantes, que não fazem parte diretamente da base de dados. Dois desses atributos, quantidade de ocorrências e quantidade de atendimentos, tiveram papel fundamental para a descoberta de regras na etapa Identificação de Características.

Na fase de Mineração de Dados realizou-se uma comparação entre algoritmos (classificadores e regressores) de diferentes paradigmas de aprendizado, o que permitiu, de forma experimental, definir os algoritmos mais adequados para a base de dados estudada. Os dois melhores classificadores, *BayesNet* e CART, foram refinados e formaram um modelo mais robusto de classificação, utilizado para rotular como "cancelado" os contratos com características semelhantes a um contrato já cancelado. Por meio de um gráfico de ganho acumulado, pode-se perceber que, apenas com metade dos contratos que o classificador expressa com maior certeza o rótulo "cancelado", já foi possível prever 90% de todos os cancelamentos. Nos experimentos visando simular um cenário real, o MCR obteve uma média de 56% de precisão e 88% de *recall* para o período de avaliação a cada 4 meses.

Além da classificação realizada na fase de Mineração de Dados, foi realizado também uma regressão para estimar o tempo até a efetivação do cancelamento. O MRR foi formado a partir dos dois melhores regressores: algoritmo M5 e Regressão Linear. Pelos experimentos realizados, pode-se perceber que o MRR obteve em média 2,3 meses para a métrica RMSE, ou seja, em média o erro entre a estimativa correta e a esperada era um pouco maior que 2 meses. Demonstrou-se também a taxa de acerto da regressão ao se incorporar uma margem de erro ao tempo estimado. Com 2 meses de margem de erro, para mais ou para menos, já foi possível acertar, de forma exata, quase 90% do tempo entre a data de adesão e o cancelamento dos contratos. Complementarmente, identificou-se o período de 3 meses

como o mais performático para a aplicação do MRR dentro da OPS, seguindo o conceito de treinar o modelo com um trimestre T e aplicá-lo ao próximo trimestre $T + 1$, para estimar o tempo até o cancelamento de novos contratos.

A análise das árvores de decisão, realizada na etapa Identificação de Características, mostrou aspectos interessantes do perfil de um contrato cancelado. Com taxas de acerto acima de 85% e representando pelo menos 5% das amostras, notou-se características que têm influência direta na rotulação de um contrato como "cancelado" ou "não cancelado"; essa informação pode ser utilizada pelo gestor para criar políticas e ações visando evitar a saída de beneficiários do plano de saúde, ou manter aqueles que já parecem estar satisfeitos.

A fase de Priorização de Contratos serviu para evidenciar os contratos que devem receber uma maior prioridade por parte da gestão da OPS. A partir da ordenação dos contratos rotulados como "cancelado", é possível destacar os contratos com maior chance de serem realmente cancelados e que representam uma maior contribuição financeira para o plano de saúde. Sem a priorização, todos os contratos teriam a mesma importância ao saírem como artefatos da abordagem, algo que, na prática, pode não ser interessante para a gestão da OPS.

Conclui-se com este trabalho que, a abordagem proposta promoveu resultados promissores para: classificação e atribuição do rótulo "cancelado" para contratos ativos com características de um contrato cancelado; regressão para obter o tempo estimado entre a data de adesão e o cancelamento de um contrato; e a identificação de regras relevantes, tanto para caracterizar contratos cancelados como não cancelados, que podem auxiliar o gestor na tomada de decisão. Ratifica-se, então, o uso de técnicas de aprendizado de máquina e mineração de dados como estratégias aliadas no entendimento de problemas ligados à saúde suplementar.

5.1 Limitações

Durante o desenvolvimento, execução e análise deste trabalho, vários fatores foram considerados possíveis limitações para a abordagem proposta. As limitações mais importantes são as seguintes:

- A limitação de histórico dos contratos influencia diretamente na classificação dos contratos. Períodos maiores que o utilizado (12 meses), apesar de possuírem a desvantagem de necessitarem de um histórico maior do beneficiário, abrangem um maior número de cancelamentos. Outro fator crítico dessa limitação é que, mesmo quando se sabe que um contrato está cancelado, há a possibilidade de considerá-lo um contrato ativo pela restrição de histórico imposta;
- Os contratos são priorizados de forma simples, por meio de uma ordenação de acordo

com alguns fatores pré-estabelecidos. Uma metodologia que possa relacionar esses atributos de uma maneira mais robusta, como a atribuição de pesos a cada um deles, pode expandir a quantidade de opções disponibilizadas a gestão da OPS;

- A base de dados utilizada pertence a um plano de saúde privado específico, para outras empresas os algoritmos e técnicas aplicadas podem não oferecer um resultado satisfatório;
- Por fim, a abordagem proposta está limitada a variáveis internas da base de dados do plano de saúde. Fatores como desemprego, inflação, gastos com locomoção, escola para os filhos, dívidas e diversos outros elementos externos podem influenciar o titular do plano a realizar o cancelamento do contrato. Formas de correlacionar indicadores socioeconômicos e atributos internos da base de dados podem garantir uma maior robustez ao modelo de classificação.

5.2 Continuidade da Pesquisa

Baseando-se nas limitações encontradas e em ideias para complementação deste estudo, os seguintes itens são considerados pontos primordiais na continuidade da pesquisa:

- Realizar experimentos complementares para avaliar a aplicação da abordagem em um cenário mais próximo do real. Por exemplo, utilizar sempre uma janela deslizante de contratos passados para realizar o treinamento tanto do MCR como do MRR; ou, antes de classificar um determinado contrato, resgatar da base de dados os elementos mais semelhantes (clusterização) para treinar o modelo;
- Aprimorar a etapa Ordenação de Contratos, de forma a permitir um modelo mais robusto de priorização, que possua, por exemplo, a capacidade de ponderar os atributos utilizados de forma diferenciada. A finalidade de um modelo mais robusto é disponibilizar um maior conjunto de opções para que a gestão da OPS possa priorizar, de maneira cada vez mais precisa, quais contratos são mais importantes;
- Investigar uma potencial correlação entre o cancelamento dos contratos e indicadores socioeconômicos brasileiros. O objetivo é utilizar informações externas à base de dados visando melhorar o desempenho geral da fase de Mineração de Dados;
- Como principal ponto de continuidade da pesquisa, pretende-se executar a abordagem proposta dentro da própria OPS, como uma ferramenta constante de apoio estratégico; isso permite que desenhos de teste experimental, como o teste A/B (WOHLIN, 2012), possam ser aplicados para avaliar os resultados obtidos. Além disso, o fato de a abordagem ficar ativa em um contexto real pode gerar insumos, sugestões e mudanças que venham a enriquecer ainda mais a pesquisa desenvolvida.

Referências

ABUROMMAN, A. A.; Ibne Reaz, M. B. A novel SVM-kNN-PSO ensemble method for intrusion detection system. *Applied Soft Computing*, Elsevier B.V., v. 38, p. 360–372, 2015. Citado na página 36.

ACAR, E. Effect of error metrics on optimum weight factor selection for ensemble of metamodels. *Expert Systems with Applications*, Elsevier Ltd, v. 42, n. 5, p. 2703–2709, 2015. Citado na página 34.

AGGARWAL, C. C. *Data Mining*. [S.l.]: Springer Science + Business Media, 2015. Citado na página 15.

AGRESTI, A. *Categorical data analysis*. Hoboken, N.J: Wiley-Interscience, 2013. Citado na página 32.

ALPAYDIN, E. *Introduction to Machine Learning*. Cambridge, Mass: MIT Press, 2010. Citado 2 vezes nas páginas 16 e 17.

ANS. 2015. Disponível em: <http://www.ans.gov.br/perfil-do-setor/dados-e-indicadores-do-setor>. Acesso em: 27/07/2015. Citado 3 vezes nas páginas 11, 1 e 2.

ARAúJO, F. H. D. de; MACEDO, S. A.; SANTOS NETO, P. d. A. An Approach Influenced to Pre-processing for Learning Medical Claim Process. *Journal of Health Informatics*, v. 7, n. 1, p. 8–15, 2015. Citado 2 vezes nas páginas 10 e 13.

BASU, S. et al. Comparative performance of private and public healthcare systems in low- and middle-income countries: A systematic review. *PLoS Medicine*, Public Library of Science (PLoS), v. 9, n. 6, 2012. Citado na página 1.

BENESTY, J. et al. Pearson correlation coefficient. In: *Noise Reduction in Speech Processing*. [S.l.]: Springer Science + Business Media, 2009. p. 1–4. Citado na página 31.

BENNETT, K. P.; CAMPBELL, C. Support vector machines. *SIGKDD Explorations Newsletter*, Association for Computing Machinery (ACM), v. 2, n. 2, p. 1–13, 2000. Citado na página 19.

BERNARDO, J. M. *Bayesian theory*. Chichester New York: Wiley, 2000. Citado na página 18.

BREIMAN, L. *Machine Learning*, Springer Science + Business Media, v. 45, n. 1, p. 5–32, 2001. Citado na página 18.

BREIMAN, L. et al. *Classification and Regression Trees*. [S.l.]: Taylor & Francis, 1984. (The Wadsworth and Brooks-Cole statistics-probability series). Citado na página 18.

CHATTERJEE, S. *Regression analysis by example*. Hoboken, N.J: Wiley-Interscience, 2006. ISBN 0471746967. Citado na página 19.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 37.

CHEN, S.; MONTGOMERY, J.; BOLUFÉ-RöHLER, A. Measuring the curse of dimensionality and its effects on particle swarm optimization and differential evolution. *Applied Intelligence*, Springer Science + Business Media, v. 42, n. 3, p. 514–526, 2015. Citado na página 33.

CLEARY, J. G.; TRIGG, L. E. K*: An instance-based learner using an entropic distance measure. In: *International Conference on Machine Learning*. [S.l.: s.n.], 1995. p. 108–114. Citado na página 17.

CUTLER, D. M.; ZECKHAUSER, R. J. Chapter 11 the anatomy of health insurance. In: *Handbook of Health Economics*. [S.l.]: Elsevier BV, 2000. p. 563–643. Citado na página 3.

DELEN, D. et al. Analysis of healthcare coverage: A data mining approach. *Expert Systems with Applications*, Elsevier Ltd, v. 36, n. 2 PART 1, p. 995–1003, 2009. Citado 2 vezes nas páginas 10 e 13.

DIAZ, D.; THEODOULIDIS, B.; DUPOUY, C. Modelling and forecasting interest rates during stages of the economic cycle: A knowledge-discovery approach. *Expert Systems with Applications*, Elsevier BV, v. 44, p. 245–264, 2016. Citado na página 34.

DINSMORE, J. *The symbolic and connectionist paradigms: closing the gap*. [S.l.]: Psychology Press, 2014. Citado na página 19.

DORIGO, M.; BIRATTARI, M. Ant colony optimization. In: *Encyclopedia of Machine Learning*. [S.l.]: Springer Science + Business Media, 2011. p. 36–39. Citado na página 35.

DUBEY, R. et al. Analysis of sampling techniques for imbalanced data: An n=648 ADNI study. *NeuroImage*, Elsevier BV, v. 87, p. 220–241, 2014. Citado na página 37.

ECLIPSE. 2016. Disponível em: <https://eclipse.org/>. Acesso em: 25/01/2016. Citado na página 39.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, Elsevier BV, v. 27, n. 8, p. 861–874, 2006. Citado na página 21.

GALAR, M. et al. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, v. 42, n. 4, p. 463–484, 2012. Citado na página 36.

GALESHCHUK, S. Neural networks performance in exchange rate prediction. *Neurocomputing*, Elsevier BV, v. 172, p. 446–452, jan 2016. Citado na página 36.

GARCÍA, S.; HERRERA, F. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, MIT Press - Journals, v. 17, n. 3, p. 275–306, 2009. Citado na página 37.

GASPAR, R. et al. Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, Elsevier BV, v. 56, p. 179–191, 2016. Citado na página 15.

GITLAB. 2016. Disponível em: <https://gitlab.com/>. Acesso em: 25/01/2016. Citado na página 39.

GOONETILLEKE, T. L. O.; CALDERA, H. a. Mining Life Insurance Data for Customer Attrition Analysis. *Journal of Industrial and Intelligent Information*, v. 1, n. 1, p. 52–58, 2013. Citado 2 vezes nas páginas 12 e 13.

GOULãO, C. Voluntary public health insurance. *Public Choice*, v. 162, n. 1-2, p. 135–157, 2014. Citado na página 1.

Gür Ali, O.; ARITüRK, U. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications*, v. 41, p. 7889–7903, 2014. Citado 3 vezes nas páginas 9, 11 e 13.

HASTIE, T. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer, 2009. Citado na página 19.

HAYKIN, S. *Neural networks and learning machines*. New York: Prentice Hall/Pearson, 2009. Citado na página 19.

HUANG, B.; KECHADI, M. T.; BUCKLEY, B. Customer churn prediction in telecommunications. *Expert Systems with Applications*, Elsevier Ltd, v. 39, n. 1, p. 1414–1425, 2012. Citado 2 vezes nas páginas 11 e 13.

HUANG, Y. et al. Telco Churn Prediction with Big Data. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, p. 607–618, 2015. Citado 2 vezes nas páginas 11 e 13.

IGLESIAS, J. A. et al. Web news mining in an evolving framework. *Information Fusion*, Elsevier BV, v. 28, p. 90–98, 2016. Citado na página 15.

JAVA. 2016. Disponível em: <https://www.java.com/>. Acesso em: 25/01/2016. Citado na página 39.

KORDOS, M.; RUSIECKI, A. Reducing noise impact on MLP training. *Soft Comput*, Springer Science + Business Media, v. 20, n. 1, p. 49–65, may 2015. Citado na página 36.

KOSE, I.; GOKTURK, M.; KILIC, K. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing*, Elsevier B.V., v. 36, p. 283–299, 2015. Citado 2 vezes nas páginas 9 e 13.

KUMAR, M.; GHANI, R.; MEI, Z.-S. Data mining to predict and prevent errors in health insurance claims processing. In: *SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2010. p. 65–74. Citado 3 vezes nas páginas 3, 10 e 13.

LESKOVEC, J. *Mining of Massive Datasets*. Cambridge: Cambridge University Press, 2014. Citado na página 15.

LTIFI, H.; KOLSKI, C.; AYED, M. B. Combination of cognitive and HCI modeling for the design of KDD-based DSS used in dynamic situations. *Decision Support Systems*, v. 78, p. 51–64, 2015. Citado na página 26.

MARINER, W. K. Health Insurance Is Dead; Long Live Health Insurance. *American Journal of Law & Medicine*, v. 40, p. 195–214, 2014. Citado na página 1.

MITCHELL, T. *Machine Learning*. New York: McGraw-Hill, 1997. Citado 2 vezes nas páginas 16 e 18.

MOEYERSOMS, J.; MARTENS, D. Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, Elsevier B.V., v. 72, p. 72–81, 2015. Citado 2 vezes nas páginas 11 e 13.

MOHRI, M. *Foundations of machine learning*. Cambridge, MA: MIT Press, 2012. Citado na página 16.

MONTGOMERY, D. *Introduction to linear regression analysis*. Hoboken, NJ: Wiley, 2012. Citado na página 22.

NGUYEN, T. T. et al. A novel combining classifier method based on Variational Inference. *Pattern Recognition*, Elsevier, v. 49, p. 198–212, 2016. Citado na página 36.

NORMAND, C. *Social health insurance a guidebook for planning*. Bad Homburg v.d.H: VAS, 2009. Citado na página 3.

OGNJANOVIC, I.; GASEVIC, D.; DAWSON, S. Using institutional data to predict student course selections in higher education. *The Internet and Higher Education*, Elsevier BV, v. 29, p. 49–62, 2016. Citado na página 15.

ORTEGA, P. A.; RUZ, G. A.; FIGUEROA, C. J. A Medical Claim Fraud / Abuse Detection System based on Data Mining: A Case Study in Chile. *Proceedings of International Conference of Data Mining*, 2006. Citado 3 vezes nas páginas 3, 9 e 13.

PARK, S.-Y.; LEE, J.-J. An efficient differential evolution using speeded-up k-nearest neighbor estimator. *Soft Computing*, Springer Science + Business Media, v. 18, n. 1, p. 35–49, 2013. Citado na página 17.

PATRO, M.; PATRA, M. R. A novel approach to compute confusion matrix for classification of n-class attributes with feature selection. *TMLAI*, Scholar Publishing, 2015. Citado na página 20.

QUINLAN, J. Simplifying decision trees. *International Journal of Man-Machine Studies*, Elsevier BV, v. 27, n. 3, p. 221–234, 1987. Citado na página 18.

QUINLAN, J. R. Learning with continuous classes. In: . [S.l.]: World Scientific, 1992. p. 343–348. Citado na página 19.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. Citado na página 18.

ROUHI, R.; JAFARI, M. Classification of benign and malignant breast tumors based on hybrid level set segmentation. *Expert Systems with Applications*, Elsevier Ltd, v. 46, p. 45–59, 2016. Citado na página 33.

RUMELHART, D. E.; MCCLELLAND, J. L.; GROUP, C. P. R. (Ed.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986. Citado na página 19.

SEKHRI, N.; SAVEDOFF, W. Private health insurance: implications for developing countries. *Bulletin of the World Health Organization*, n. 03, p. 127–134, 2005. Citado na página 1.

SIM, J.; KWON, O.; LEE, K. C. Adaptive Pairing of Classifier and Imputation Methods Based on the Characteristics of Missing Values in Data Sets. Elsevier Ltd, v. 46, p. 485–493, 2016. Citado na página 33.

SOK, H. K. et al. Multivariate alternating decision trees. *Pattern Recognition*, Elsevier BV, v. 50, p. 195–209, 2016. Citado na página 34.

STREIB, F. *Information theory and statistical learning*. New York London: Springer, 2009. Citado na página 18.

SU, J. et al. Customer Retention Predictive Modeling in HealthCare Insurance Industry. In: *SESUG Southeast SAS Users Group*. [S.l.: s.n.], 2009. p. 1–8. Citado 2 vezes nas páginas 12 e 13.

SWINBURNE, R. *Bayes's theorem*. Oxford England New York: Published for the British Academy by Oxford University Press, 2002. Citado na página 18.

TANEJA, S. et al. An enhanced k-nearest neighbor algorithm using information gain and clustering. In: *2014 Fourth International Conference on Advanced Computing & Communication Technologies*. [S.l.]: Institute of Electrical & Electronics Engineers (IEEE), 2014. Citado na página 17.

TOMAR, D.; AGARWAL, S. A survey on data mining approaches for healthcare. *Internation Journal of Bio-Science and Bio-Technology*, Science and Engineering Research Support Society, v. 5, n. 5, p. 241–266, 2013. Citado na página 9.

WEKA. 2016. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 10/01/2016. Citado 2 vezes nas páginas 34 e 39.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)*. [S.l.]: Morgan Kaufmann, 2011. Citado 2 vezes nas páginas 20 e 39.

WOHLIN, C. *Experimentation in software engineering*. Berlin New York: Springer, 2012. Citado na página 53.

WOJTUSIAK, J. et al. Rule-based prediction of medical claims' payments: A method and initial application to medicaid data. *Proceedings of International Conference on Machine Learning and Applications*, v. 2, p. 162–167, 2011. Citado 3 vezes nas páginas 3, 10 e 13.

XIAO-BING, Y.; JIE, C.; ZAI-WU, G. Review on customer churn issue. *Computer Integrated Manufacturing Systems*, v. 18, 2012. Citado na página 10.

XU, W. et al. Estimating the area under a receiver operating characteristic (ROC) curve: Parametric and nonparametric ways. *Signal Processing*, Elsevier BV, v. 93, n. 11, p. 3111–3123, 2013. Citado na página 22.

YIN, J.; CHEN, D.; LI, Y. Smart train operation algorithms based on expert knowledge and ensemble CART for the electric locomotive. *Knowledge-Based Systems*, Elsevier BV, v. 92, p. 78–91, 2016. Citado na página 15.

ZHANG, X. et al. The use of ROC and AUC in the validation of objective image fusion evaluation metrics. *Signal Processing*, Elsevier, v. 115, p. 38–48, 2015. Citado 3 vezes nas páginas 21, 22 e 33.

ZHU, D. A hybrid approach for efficient ensembles. *Decision Support Systems*, Elsevier BV, v. 48, n. 3, p. 480–487, 2010. Citado na página 36.