Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 14, Iss. 2, pp. 16-35, 2016

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

Jefferson Henrique Camelo, Ricardo de Andrade Lira Rabêlo
Universidade Federal do Piauí
jeffersonpi@hotmail.com, ricardoalr@ufpi.edu.br

Erick Baptista Passos
Instituto Federal do Piauí
erickpassos@gmail.com

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, provedoras de planos de saúde, geralmente, são responsáveis pelo pagamento das contas de operações, internações, medicamentos e outros custos hospitalares. O bom funcionamento dessas empresas está diretamente relacionado à permanência dos segurados no plano de saúde e, portanto, inversamente relacionado à quantidade de cancelamentos desses seguros. O objetivo principal deste trabalho consiste em desenvolver uma abordagem projetada para caracterizar o cancelamento eletivo de contratos em planos de saúde privados. A abordagem proposta é constituída de várias etapas, que são organizadas em três fases: Pré-Processamento, Mineração de Dados e Priorização. A fase de Pré-Processamento visa garantir uma maior qualidade às informações extraídas da base de dados de um plano de saúde privado real. A fase de Mineração de Dados explora os dados pré-processados, à procura de padrões, relacionamento entre atributos e tendências, com o intuito de descobrir novos conhecimentos. Essa fase de Mineração de Dados é responsável por reconhecer contratos ativos com características de contratos já cancelados, e identificar que tipos de ações e comportamentos levam os clientes da empresa a cancelarem seus vínculos. Dessa forma, a gestão do plano de saúde pode interceder de forma proativa no problema, e não apenas de forma reativa. Uma comparação entre diferentes algoritmos de classificação, utilizados para o reconhecimento de contratos com potencial de cancelamento, é realizada com o objetivo de definir os paradigmas de aprendizado mais adequados à etapa de reconhecimento de contratos. Uma complementação à fase de Mineração de Dados é a experimentação de diferentes técnicas de balanceamento de classes e seu impacto nas métricas de precisão e *recall* dos resultados obtidos. A fase de Priorização de Contratos objetiva priorizar os contratos, com características de contrato cancelado, de acordo com um grau de cancelamento. Esse grau de cancelamento é estimado para cada contrato por meio de uma sistema de inferência *fuzzy*. Por fim, é realizado um conjunto de experimentos demonstrando passo-a-passo a execução prática da abordagem proposta, com a apresentação de resultados e discussões.

Palavras-chave – Mineração de dados, sistema de inferência *fuzzy*, saúde suplementar, plano de saúde.

Abstract – A range of different factors influences the life expectancy of a person, and a key factor is healthcare. However, this concern with healthcare is not cheap, there is usually a need for some private company, a provider of health insurance, which is responsible for payment of transaction accounts, hospitalizations, medications and other medical costs. The proper functioning of these companies is directly related to the stay of insured people in the health plan and therefore inversely related to the amount of such insurance cancellations. The main objective of this work is to develop an approach designed to characterize the elective cancellation of contracts in private health plans. The proposed approach consists of several steps, which are organized in three phases: Pre-Processing, Data Mining and Prioritization. The Pre-Processing phase aims to ensure greater quality to information obtained from the database of a real private health insurance. The Data Mining phase explores the pre-processed data, looking for patterns, relationships between attributes and trends, in order to discover new knowledge. This Data Mining phase is responsible for recognizing active contracts with features of contracts already canceled, and identify what types of actions and behaviors lead the company's customers to cancel their ties. Thus, the health plan management can intervene proactively in the problem, not just reactively. A comparison of different classification algorithms, used for recognition of contracts with a potential cancellation is performed aiming to define the most appropriate learning paradigms to contracts recognition step. A complement to Data Mining phase is experimenting with different class balancing techniques and their impact on precision metrics and recall of results. Prioritization phase objectively prioritize contracts with canceled contract characteristics, according to a degree of cancellation. This degree of cancellation is estimated for each contract by an fuzzy inference system. Finally, a set of experiments is conducted demonstrating step-by-step practical implementation of the proposed approach with the presentation of results and discussions.

Keywords – Data mining. fuzzy inference system. supplementary health. health plan. healthcare.

1. INTRODUÇÃO

Uma diversidade de fatores influencia na expectativa de vida de uma pessoa, e um fator fundamental é o cuidado com a própria saúde. Mas cuidar da saúde não é algo barato, 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 [21, 28]. Planos de saúde privados, em países em desenvolvimento, geralmente têm maior senso de responsabilidade com o atendimento prestado, são mais eficientes e autossustentáveis que a iniciativa pública [5]. Mesmo em países melhor desenvolvidos, naqueles onde uma boa cobertura de saúde foi alcançada, os planos privados possuem uma representação significativa no atendimento à saúde [36].

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 [1] mostram um constante aumento no número de beneficiários associados a alguma Operadora de Plano de Saúde (OPS). 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 clientes influencia diretamente na receita obtida pelos planos de saúde. Entretanto, a alta quantidade de clientes influencia também no aumento das despesas assistenciais. Despesa assistencial é toda despesa resultante da utilização, por parte do segurado, das coberturas oferecidas pelo plano de saúde, representando, portanto, todo gasto que precisa ser despendido quando um cliente da empresa precisa de atenção médica, como: consultas, exames, internações e terapias.

Como o crescimento das despesas segue o mesmo ritmo do crescimento das receitas, é importante para a gestão lograr êxito no gerenciamento e redução dos 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 com elevadas despesas assistenciais, qualquer método, técnica ou proposta que reduza os gastos com saúde ou mantenha a receita estável, é relevante para o funcionamento duradouro do plano de saúde.

O uso de mineração de dados na área de planos de saúde tem sido uma importante fonte para descoberta de conhecimento em bancos de dados. Além disso, por meio da aplicação de técnicas de aprendizado de máquina, um conjunto de comportamentos no cenário médico podem ser aprendidos, detectados e até mesmo antecipados. Ortega et al. [30] descreveram um sistema para detectar fraudes e abusos em guias médicas[1] de um plano de saúde privado chileno. He et al. [23] desenvolveram uma versão modificada do algoritmo KNN, que utiliza algoritmos genéticos com o objetivo de encontrar uma distância ótima não-euclidiana entre as instâncias de treinamento, para classificar profissionais médicos que realizam pedidos de exames desnecessários a pacientes. Araújo et al. [4] apresentaram um processo para realizar o aprendizado automático da regulação médica/odontológica de uma operadora privada de plano de saúde. Foi executado um conjunto de experimentos para avaliar a classificação de algoritmos de diferentes paradigmas de aprendizado. Cada um dos trabalhos, até agora apresentados, usou diferentes técnicas e algoritmos para solucionar problemas ligados ao contexto de planos de saúde; entretanto, não há em nenhum deles um objetivo, bem definido, visando entender as razões e motivos que levaram aos resultados da classificação. Esse melhor entendimento das razões e motivos, baseado na análise das árvores de decisão, foi um dos principais objetivos deste estudo.

Os trabalhos de Wojtusiak et al. [40] e Kumar et al. [26], apesar de experimentarem uma pequeno número de algoritmos, apresentaram um interesse em entender os comportamentos que influenciaram nos resultados obtidos. Wojtusiak et al. apresentaram um método que derivava regras para ajudar na preparação e investigação de guias médicas antes da submissão das mesma para as empresas responsáveis pelo pagamento, o que reduziu custos com erros e imperfeições na análise das guias. Kumar et al. 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.

Dentre os trabalhos encontrados na literatura, dois deles possuem o mesmo domínio de aplicação deste estudo: cancelamentos em planos de saúde. Su et al. [38] 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 presente estudo que analisa esse aspecto também sobre a ótica de interdependência entre os fatores. Para o treinamento/validação do classificador, todos os consumidores foram agrupados em dois grupos, por meio do algoritmo *k*-Means, e o grupo em que a taxa de cancelamento era maior foi escolhido. Goonetilleke and Caldera [19] 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.

1.1 Definição do Problema

Devido à margem estreita de lucro, as empresas de plano de saúde, 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 [26], detectar abusos em serviços requisitados

[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.

pelos médicos [23, 30] e reduzir custos com análises incorretas de guias [26, 40]. O foco deste trabalho é auxiliar a empresa, não de forma direta na redução dos gastos, mas na manutenção da receita esperada.

Como provedoras de seguro, as OPSs têm sua receita pautada no pagamento de um taxa periódica, geralmente mensal, por parte dos beneficiários do plano de saúde. Essa forma de receita se baseia na metodologia de gerenciamento de risco conhecida como *risk pool* ou grupo de risco [12]. Nessa metodologia 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 [29].

Visto que o valor pago de cada beneficiário é importante para o funcionamento estável do plano de saúde, o problema abordado por este trabalho é o cancelamento eletivo dos contratos, ou seja, quando o segurado decide, deliberadamente, cancelar seu vínculo com a empresa e assim encerrar sua parcela de aporte financeiro. É importante ressaltar que, apenas os contratos individuais, ligados diretamente a um segurado específico, fazem parte do escopo deste estudo. Contratos corporativos, que representam acordos entre o plano de saúde e outras empresas, são desconsiderados, pois acredita-se que, nesse caso, o cancelamento está mais vinculado à relação entre a seguradora de saúde e a empresa contratante, do que entre a seguradora de saúde e o beneficiário.

1.2 Objetivos

O objetivo principal deste estudo consistiu em desenvolver uma abordagem projetada para caracterizar o cancelamento eletivo de contratos em planos de saúde privados. Ressalta-se que o principal guia para essa caracterização foi a descoberta de conhecimento implícito na base de dados da OPS, pois, *a priori*, apesar de desconhecida para a gestão, essa conhecimento apresenta particularidades expressivas para o problema em foco.

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

- Classificar contratos ativos em contratos com características de contrato cancelado. Essa classificação foi realizada, principalmente, na etapa Reconhecimento de Contratos (fase de Mineração de Dados);
- Analisar ramos das árvores geradas por classificadores baseados em árvore de decisão, para identificar que padrões caracterizam os contratos rotulados com a classe "cancelado". Essa análise foi executada na etapa Identificação de Características (fase de Mineração de Dados);
- Estimar um grau de cancelamento para cada contrato rotulado como "cancelado". Esse grau de cancelamento foi utilizado como um fator de criticidade de forma a priorizar os contratos classificados. A estimação do grau de cancelamento e a priorização dos contratos foram realizados na fase de Priorização de Contratos;
- Realizar uma comparação entre diferentes classificadores para definir os paradigmas de aprendizado e os algoritmos que apresentam melhores resultados para o problema. Essa comparação foi realizada na etapa Reconhecimento de Contratos (fase de Mineração de Dados);
- Comparar diferentes soluções para resolver o problema do desbalanceamento de classes, analisando o impacto de cada uma das técnicas nos resultados obtidos. A comparação foi uma das ações da etapa Balanceamento de Classes (fase de Pré-Processamento).

1.3 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 do plano de saúde 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, é possível tomar medidas proativas sobre os contratos rotulados, com o objetivo de evitar que os beneficiários em questão realmente concretizem o cancelamento;
- Aferição de um grau de cancelamento para cada contrato classificado como "cancelado", baseando-se, para isso, no conhecimento do gestor/especialista. Ao se estimar um grau de cancelamento, obtém-se uma métrica que condensa, em um único valor, a criticidade de um determinado contrato. Além disso, ao se basear na expertise do gestor, essa aferição realizada acaba por expressar o conhecimento adquirido, pela empresa, sobre o cancelamento de contratos;
- Priorização dos contratos com a utilização do grau de cancelamento estimado. Por meio dessa priorização, amplia-se o horizonte de recursos da gestão, permitindo a seleção dos contratos de acordo com a capacidade de trabalho da empresa;

- Análise comparativa sobre o impacto, nas métricas de precisão e *recall*, do uso de diferentes técnicas que realizam o balanceamento de classes. Como o objetivo da classificação é maximizar tanto a precisão como o *recall*, é possível perceber, por meio da análise, que cada técnica representa um solução não-dominada. Desse modo, pode-se optar por técnicas diferentes de acordo com o objetivo da gestão, o que torna o processo, como um todo, flexível.

1.4 Estrutura do Trabalho

O restante deste trabalho está estruturado da seguinte forma: a Seção 2 detalha a estrutura da abordagem proposta e de cada uma das suas fases e etapas; a Seção 3 expõe e discute os resultados encontrados com a execução prática da abordagem em uma OPS real e a Seção 4 retrata as conclusões, limitações e pontos de continuidade da pesquisa desenvolvida neste estudo.

2. ABORDAGEM PROPOSTA

A abordagem proposta objetiva caracterizar o cancelamento eletivo de contratos em planos de saúde privados. Por caracterizar entende-se a capacidade de distinguir aspectos, padrões e propriedades que possam moldar, baseado em eventos passados, o perfil de um contrato cancelado. 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 proposta desenvolvida é baseada no modelo de descoberta de conhecimento em banco de dados proposto por Fayyad et al. [15]. A estrutura geral da abordagem é ilustrada na Figura 1, contemplando todas as fases e as respectivas etapas.

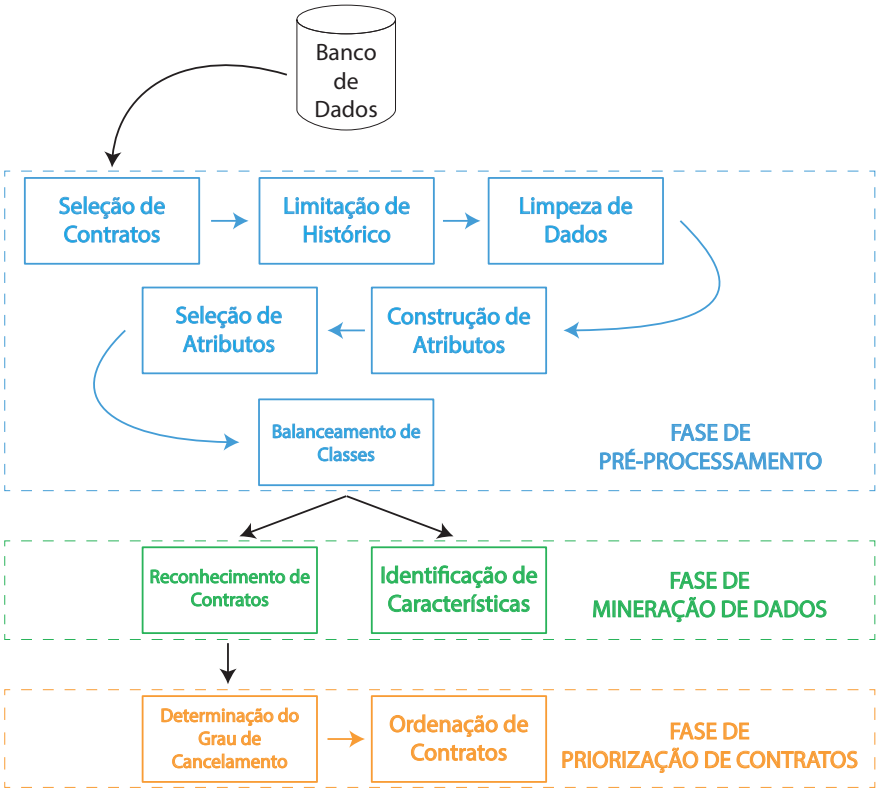

Figura 1: Estrutura da abordagem proposta.

2.1 Base de Dados

A base de dados utilizada neste estudo foi fornecida por uma OPS sediada no estado do Piauí, Brasil. A base possui informações de 22.542 beneficiários, com contratos estabelecidos entre Março de 2005 e Julho de 2015, presentes em 5 tabelas, que juntas totalizam 230 colunas/atributos. Para facilitar a compreensão das fases e etapas da abordagem, é primordial entender os seguintes conceitos comuns ao contexto de planos de saúde:

- 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.

2.2 Pré-Processamento

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, devido à grande quantidade de dados presentes, muitos deles irrelevantes ou prejudiciais ao entendimento do processo. Essa fase é constituída das seguintes etapas: Seleção de Contratos, Limitação de Histórico, Limpeza de Dados, Seleção de Atributos, Construção de Atributos e Balanceamento de Classes.

2.2.1 Seleção de Contratos

Na etapa Seleção de Contratos, há uma primeira redução na quantidade de contratos processados pela abordagem. São selecionados apenas os contratos dos titulares, ou seja, os contratos que representam apenas dependentes não são resgatados da base de dados. Optou-se por selecionar somente 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 relevantes para o problema 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 estudo 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 entre 01 de Janeiro de 2013 e 31 de Dezembro de 2014. Essas datas não foram escolhidas 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. A data final de seleção foi estabelecida como o final de 2014, porque, dessa forma, são representados dois anos inteiros, de janeiro a dezembro, de contratos selecionados, já que na base de dados original só constam registros até julho de 2015.

Uma restrição adicional, nessa etapa de Seleção de Contratos, foi adicionada devido a uma política interna da OPS. Essa política oferece como benefício aos funcionários internos da empresa um plano de saúde gratuito, ou seja, os funcionários da OPS são considerados titulares dos contratos, porém, como exceção à regra geral, não são responsáveis pelo pagamento de suas mensalidades.

2.2.2 Limitação de Histórico

O objetivo dessa etapa é limitar o histórico dos contratos selecionados na etapa Seleção de Contratos, pois existe uma diferença entre beneficiários que começaram o plano de saúde no início de 2013 e os que começaram no final de 2014. Essa diferença é motivada pelo fato de que os beneficiários, cujos contratos iniciaram em 2013, têm um horizonte maior de tempo e insatisfação para cancelar seus contratos, enquanto outros beneficiários, com menos tempo de plano de saúde, provavelmente tiveram menos experiências para avaliar os serviços da OPS. Desse ponto de vista, a probabilidade de um contrato, cuja data de adesão é Janeiro de 2013, estar com o estado de "cancelado", é maior que a probabilidade de um contrato com data de adesão em Dezembro de 2014.

Decidiu-se limitar o histórico de contrato para no máximo 6 meses depois da data de adesão, ou seja, apenas é considerado o último estado associado ao contrato depois de 6 meses da entrada do beneficiário no plano de saúde. A Figura 2 mostra a saída da etapa de Limitação de Histórico ao receber dois contratos de exemplo: 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 seus histórico limitado. O Contrato *A* passou a ser considerado "não cancelado", pois houve um diferença de 13 meses entre a adesão e o cancelamento, diferença superior a 6 meses, implicando que para as próximas etapas e fases da abordagem, esse contrato não será mais considerado "cancelado".

O período específico de 6 meses foi escolhido baseado na análise da quantidade de contratos cancelados entre 2013 e 2014. A Figura 3 apresenta a percentagem acumulada do total de contratos cancelados, de acordo com o número de meses após a data de adesão. É possível notar que 39% dos cancelamentos ocorreram em até 6 meses desde a entrada do beneficiário no plano de saúde, e que, a partir desse período, a taxa de crescimento do número de contratos cancelados começa a diminuir. Escolher um período inferior a 6 meses, além de representar menos tempo de histórico do beneficiário, equivaleria a no máximo 26% dos cancelamentos, caso se optasse por 5 meses. Escolher um período superior englobaria mais histórico e abrangeria mais cancelamentos, porém, quanto maior o período, maior será o tempo de atraso para aferir se um contrato irá cancelar ou não. Por

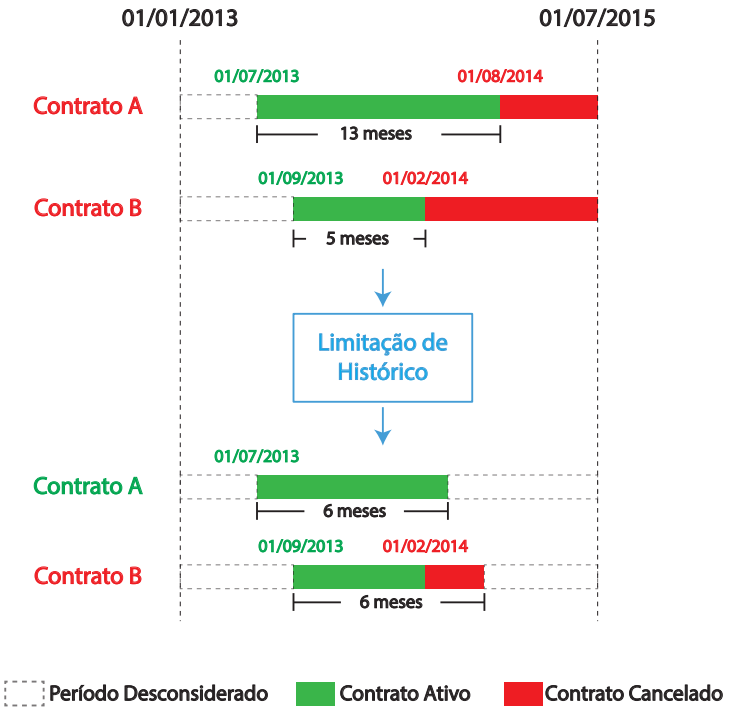

Figura 2: Exemplificação do processo executado pela etapa Limitação de Histórico.

exemplo, se o período escolhido fosse 12 meses, os contratos mais adequados a essa escolha teriam que estar há pelo menos um ano ativos na base de dados da OPS.

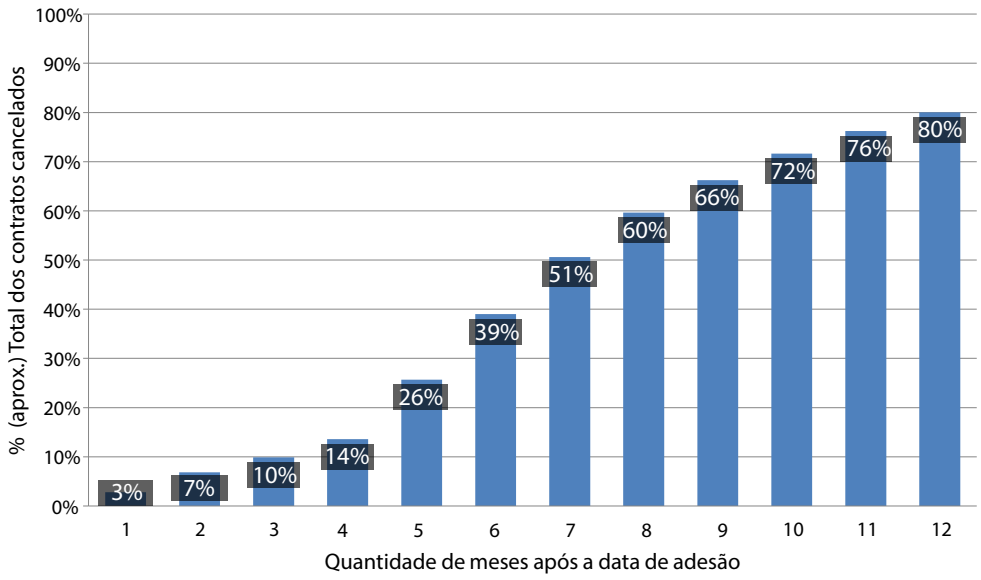

Figura 3: Percentagem acumulada do total de contratos cancelados após a data de adesão.

2.2.3 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 5 grupos de atributos a serem removidos:

- Atributos Poluídos: representam atributos que são preenchidos sem seguir um padrão estabelecido, por exemplo: um atributo que deveria representar uma idade ser preenchido com valores negativos.

- Atributos Preenchidos com Valor Padrão: representam atributos que, por algum motivo, não estão sendo utilizadas pelo sistema durante algum processo; assim, assumem o valor padrão para o tipo de dados representando.
- Atributos Duplicados ou Redundantes: representam atributos que contêm informações já presentes em outros atributos.
- Atributos Irrelevantes: representam atributos que contêm informações não consideradas úteis para a resolução do problema. Essa atribuição de relevância está relacionada a uma tarefa conjunta com um especialista da OPS.
- 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.

2.2.4 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. A criação de um atributo "idade", por exemplo, pode ser mais representativa para a fase de Mineração de Dados do que utilizar a data de nascimento com valores do tipo data. Para a maioria dos algoritmos classificadores é mais simples trabalhar com valores numéricos do que com valores que representam datas.

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 criado para expandir a informação contida no contrato original, pois esse dado não está presente de forma direta na base de dados cedida.

2.2.5 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 [7], o papel da etapa Seleção de Atributos é reduzir a dimensionalidade do problema. Para isso é aplicado uma estratégia de seleção de atributos do tipo filtro, em que 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.

2.2.6 Balanceamento de Classes

Ao final das etapas anteriores, os contratos possuem dois tipos de estados: contrato já cancelado e contrato não cancelado. Como a frequência de contratos ativos é maior que a de contratos cancelados, a quantidade entre essas duas classes possui diferenças significativas. Visto que o desbalanceamento entre as classes pode influenciar em uma performance menor dos algoritmos classificadores [17], o objetivo da etapa de Balanceamento de Classes é definir uma estratégia para balancear os contratos.

Como técnica padrão para solucionar o desbalanceamento, optou-se pelo *undersampling* aleatório, por ser considerado uma das técnicas/estratégias mais efetivas [18]. Porém, devido a diversidade de técnicas existentes, com distintas abordagens para atacar o problema, é realizada também uma comparação entre os resultados obtidos ao se utilizar cada uma das seguintes estratégias: *undersampling* aleatório, *oversampling* aleatório, SMOTE e SMOTE+*undersampling*. A comparação é realizada de acordo com as métricas de precisão e *recall* alcançadas por cada técnica.

2.3 Mineração de Dados

A fase de Mineração de Dados explora os dados pré-processados com o objetivo de descobrir novos conhecimentos, ou seja, padrões, relacionamento entre atributos, e tendências ainda não conhecidos pela gestão do plano de saúde. Essa fase consiste das seguintes etapas: Reconhecimento de Contratos e Identificação de Características.

2.3.1 Reconhecimento de Contratos

Essa etapa visa reconhecer contratos ativos que possuem características de contratos anteriormente cancelados. Para executar essa tarefa de reconhecimento, são utilizados algoritmos classificadores, com o objetivo de aprender um modelo capaz de rotular um contrato ainda ativo com a classe "cancelado" caso haja valores de atributos comuns a um contrato cancelado.

Para definir qual o algoritmo mais adequado à abordagem proposta, é realizada uma comparação entre diversos paradigmas de aprendizado. A avaliação das técnicas é dada a partir das métricas: área sob a curva ROC[2] [14], *recall* e taxa de falso positivos (TFP). Uma seleção foi realizada com algoritmos conhecidos na literatura oriundos dos paradigmas: "Baseado em Árvore de Decisão", "Bayesiano", "Baseado em Exemplos" e "Conexionista". Os algoritmos escolhidos foram:

- Paradigma Baseado em Árvore de Decisão: C4.5[35], RandomTree[2] e CART[6];

[2] A curva ROC apresenta no seu gráfico a relação entre a taxa de *recall* e a taxa de falsos positivos ao se variar o limiar de discriminação na classificação.

- Paradigma Bayesiano: Naive Bayes[24] e BayesNet[8];
- Paradigma Baseado em Exemplos: KNN[16] e K*[10];
- Paradigma Conexionista: MLP (*Multilayer Perceptron*) [3] e SVM (*Support Vector Machines*)[11].

2.3.2 Identificação de Características

Essa etapa, paralela a etapa de Reconhecimento de Contratos, visa identificar quais ações, comportamentos ou padrões, levam o beneficiário a cancelar seu contrato. Essa identificação é guiada pela análise dos ramos gerados por classificadores baseados em árvores de decisão.

Além de serem computacionalmente mais econômicas que outros tipos de classificadores, as árvores de decisão podem ter seu modelo de aprendizado facilmente interpretado, diferente de redes neurais artificiais, por exemplo, que são técnicas tradicionalmente empregadas em problemas de classificação [22]. 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 analisados a fim de investigar o que cada atributo, contido em um desses ramos da árvore de decisão, representa no contexto do cancelamento.

2.4 Priorização de Contratos

Por fim, a fase de Priorização de Contratos objetiva priorizar os contratos classificados como "cancelado", de acordo com o risco de cancelamento associado a cada contrato. Para isso, são desenvolvidas as seguintes etapas: Determinação do Grau de Cancelamento e Ordenação de Contratos.

2.4.1 Determinação do Grau de Cancelamento

Após a etapa de Reconhecimento de Contrato, os contratos ativos que foram rotulados como "cancelado" possuem o mesmo nível de criticidade, ou seja, não se tem informações suficientes que indiquem quais contratos são mais suscetíveis a serem realmente cancelados. Visando resolver essa falta de informação sobre a risco de cancelamento do contrato, a etapa de Determinação do Grau de Cancelamento tem como objetivo atribuir um valor numérico, para cada contrato classificado como "cancelado", que indique sua respectiva criticidade.

Para estimar esse valor numérico, é utilizado um sistema de inferência *fuzzy* (SIF) de Mamdani [33], com duas entradas e uma saída, modelado por meio da ferramenta jFuzzyLogic [9]. A primeira entrada é o nível de certeza (NC) da classificação do contrato, ou seja, a probabilidade estimada pelo classificador de que um determinado contrato pertença a classe "cancelado". A segunda entrada é o valor total (VT) pago mensalmente pelo contrato, incluindo o valor de cada dependente associado. Como saída do sistema obtém-se o objetivo da etapa, o grau de cancelamento (GC) do contrato. A Figura 4 representa a estrutura do SIF.

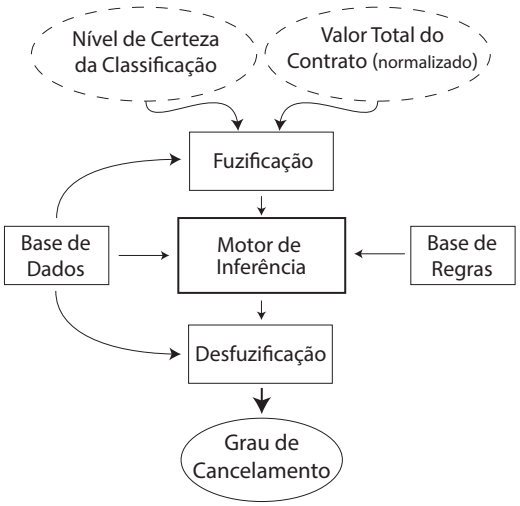

Figura 4: Estrutura do sistema de inferência *fuzzy* utilizado para estimar o grau de cancelamento.

São definidos para cada variável de entrada três termos linguísticos: baixo, médio e alto. Para a variável de saída são definidos cinco termos: muito baixo, baixo, médio, alto e muito alto. A função de pertinência definida para cada termo linguístico, seja pertencente a uma variável de entrada ou saída, obedece a forma triangular. Não se possuem informações claras de como deve ser a distribuição de pertinência para essas variáveis, por isso, e por sua simplicidade, foi escolhida a forma triangular para todos os termos linguísticos [32]. Todos os elementos de entrada e saída do SIF são normalizados para um valor entre 0 e 1. Como o nível

de certeza já está compreendido entre 0 e 1, apenas o valor total do contrato precisou ser normalizado. O valor total normalizado é igual a razão entre o valor de um determinado contrato e o maior valor encontrado presente na base de dados.

A base contendo as regras para inferência é definida de acordo com o conhecimento do gestor do plano de saúde. Foi pedido ao gestor que indicasse o grau de cancelamento mais adequado (muito baixo, baixo, médio, alto e muito alto) para cada combinação possível entre as variáveis de entrada, nível de certeza e valor total. Como existem três termos linguísticos para cada uma das variáveis de entrada, a base de regras contempla nove regras. Essas regras, definidas pelo gestor, são as seguintes:

1. se **NC é BAIXO** e **VT é BAIXO**, então **GC é MUITO BAIXO**;
2. se **NC é BAIXO** e **VT é MÉDIO**, então **GC é BAIXO**;
3. se **NC é BAIXO** e **VT é ALTO**, então **GC é MÉDIO**;
4. se **NC é MÉDIO** e **VT é BAIXO**, então **GC é MÉDIO**;
5. se **NC é MÉDIO** e **VT é MÉDIO**, então **GC é MÉDIO**;
6. se **NC é MÉDIO** e **VT é ALTO**, então **GC é ALTO**;
7. se **NC é ALTO** e **VT é BAIXO**, então **GC é ALTO**;
8. se **NC é ALTO** e **VT é MÉDIO**, então **GC é MUITO ALTO**;
9. se **NC é ALTO** e **VT é ALTO**, então **GC é MUITO ALTO**.

2.4.2 Ordenação de Contratos

Nessa etapa, os contratos rotulados como "cancelado" são ordenados de acordo com o grau de cancelamento estimado pela etapa Determinação do Grau de Cancelamento. A ordenação é realizada do maior para o menor GC, permitindo a priorização dos contratos de acordo com a sua criticidade.

Apesar de ser uma etapa simples, a etapa de Ordenação de Contratos realiza um papel importante na entrega de informação à gestão, pois se houvesse apenas a estimação do grau de cancelamento o gestor ainda ficaria encarregado de organizar os contratos de acordo com a sua gravidade. Outro fator importante é que há um desacoplamento entre as etapas da fase de Priorização de Contratos, o que permite, mais facilmente, a adição de outras informações para realizar a priorização. Essas informações adicionais poderiam representar, por exemplo, a quantidade máxima de contratos a serem selecionados depois da ordenação.

3. RESULTADOS E DISCUSSÕES

Nessa seção são demonstrados os resultados obtidos com a aplicação prática da abordagem proposta. Para a execução dos algoritmos relacionados ao aprendizado de máquina, é utilizada a ferramenta WEKA (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 [39]. 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, devido à compatibilidade com a ferramenta WEKA. Adicionalmente é utilizado o Eclipse como plataforma de desenvolvimento e o repositório Gitlab para o controle de versão dos elementos produzidos.

3.1 Fase de Pré-Processamento

Antes de iniciar a fase de Pré-Processamento, a base de dados apresenta um total de 82.222 tuplas que representam contratos; todas essas tuplas são entradas para a etapa Seleção de Contratos. A Figura 5 mostra a redução na quantidade de contratos realizada nessa etapa, de acordo com cada restrição descrita na Seção 2. Ao final, a quantidade de contratos é limitada a 9.814 elementos.

Após a etapa Limitação de Histórico, são removidos dos contratos os atributos que podem prejudicar a classificação. Cada contrato pode ser visto como uma longa tupla de colunas, totalizando 230 atributos. A Figura 6 mostra a redução na quantidade de atributos após a remoção dos cinco grupos estabelecidos pela etapa Limpeza de Dados.

Os contratos, após a limpeza de dados, tiveram 211 atributos removidos, restando para as etapas posteriores um total de 19 atributos. Nota-se que a maior parte dos atributos retirados pertence ao grupo de atributos irrelevantes, devido principalmente às 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 uma série de mudanças estruturais no banco de dados, tornando obsoletas diversas tabelas e colunas. Os atributos restantes são os seguintes:

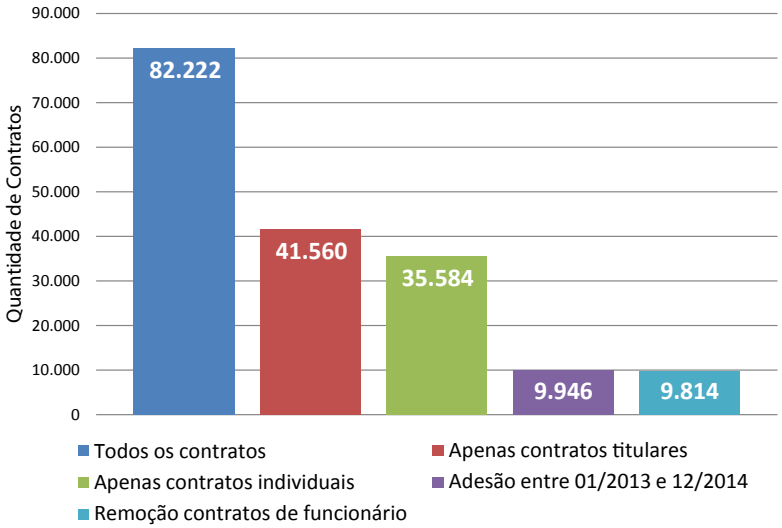

Figura 5: Redução na quantidade de contratos efetuada pela etapa Seleção de Contratos.

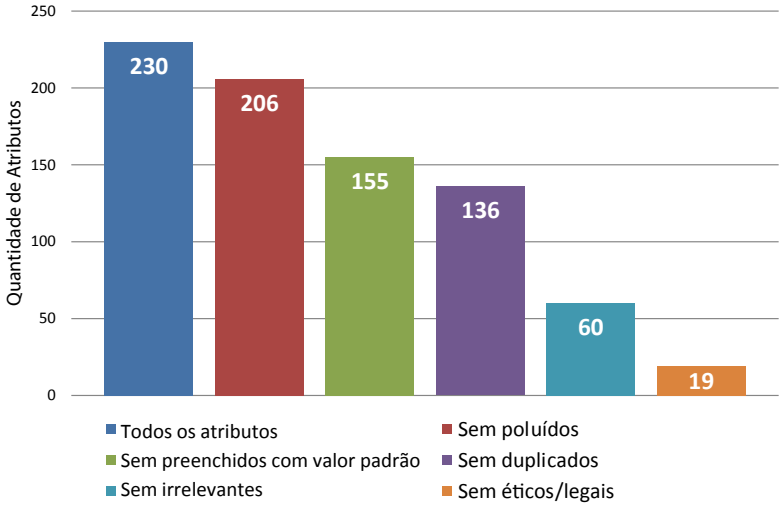

Figura 6: Redução na quantidade de atributos efetuada pela etapa Limpeza de Dados.

1. **“beneficiario”:** [nominal] representa se o titular do contrato é um beneficiário do plano de saúde. Pode assumir os valores: “SIM” e “NAO”;
2. **“desconto”:** [numérico] representa um possível desconto dado ao contrato;
3. **“diabase”:** [numérico] representa qual o dia do mês foi escolhido pelo titular para realizar o pagamento da mensalidade;
4. **“estadocivil”:** [nominal] representa qual o estado civil do titular do contrato. Pode assumir os valores: “VIUVO”, “SOLTEIRO”, “CASADO”, “SEPARADO” e “OUTRO”;
5. **“faixapagamento”:** [nominal] representa em qual faixa de pagamento se encaixa o contrato. Uma faixa de pagamento representa um valor de referência que deve pago por um beneficiário de acordo com uma faixa de idade. Pode assumir os valores: “0”, “1”, “2” e “3”;
6. **“iddescontocontrato”:** [nominal] representa qual tipo de desconto está associado ao contrato. Pode assumir os valores: “1”, “2”, “3”, “4”, “5”, “6” e “7”;
7. **“idempresaterceirizada”:** [nominal] representa qual setor interno foi responsável pelo contrato. Pode assumir os valores: “1”, “2”, “3”, “4”, “5”, “6”, “7”, “8” e “9”;
8. **“idproduto”:** [nominal] representa qual o conjunto de coberturas está associado ao contrato. Pode assumir os valores: “1”, “25” e “26”;

9. **“quantpagamentosinsuficienciafundos”:** [numérico] representa a quantidade de tentativas de pagamento onde não foi possível efetivar a transação, ou seja, o pagamento da mensalidade não foi efetivado por algum motivo, como problemas no cartão de crédito e cheque sem fundos.

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

11. **“servidorpublico”:** [nominal] representa se o titular é um servidor público. Pode assumir os valores: “SIM” e “NAO”;

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

13. **“shift”:** [nominal] representa se já foram realizados acordos financeiros (reajustes para diminuir o valor da mensalidade) entre o titular do contrato e o plano de saúde. Pode assumir os valores: “SIM” e “NAO”;

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

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

16. **“valoradicional”:** [numérico] representa um possível valor extra cobrado no contrato;

17. **“valororiginal”:** [numérico] representa o valor inicialmente combinado para o contrato;

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

19. **“vencimentonodiabase”:** [nominal] representa se o vencimento do contrato ocorre no dia escolhido para o pagamento da mensalidade. Pode assumir os valores: “SIM” e “NAO”.

Após a etapa Limpeza de Dados, 10 novos atributos foram adicionados ao contrato na etapa Construção de Atributos. Os elementos adicionados são os seguintes:

1. **“idade”:** [numérico] representa a idade do titular. Essa idade é calculada a partir do atributo que representa a data de nascimento do titular. Vale ressaltar que a data de nascimento não faz parte dos atributos selecionados;

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.

5. **“tem_atendimentos”:** [nominal] representa se a “qtd_atendimento” é maior do que zero. Pode assumir os valores: “SIM” e “NAO”;

6. **“tem_dependentes”:** [nominal] representa se a “qtd_dependentes” é maior do que zero. Pode assumir os valores: “SIM” e “NAO”;

7. **“tem_ocorrencias”:** [nominal] representa se a “qtd_ocorrencias” é maior do que zero. Pode assumir os valores: “SIM” e “NAO”;

8. **“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”;

9. **“ultima_situacao”:** [nominal] representa última situação do contrato antes do cancelamento. Pode assumir os valores: “Cadastrado”, “Suspenso” e “Ativo”;

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

A quantidade de atributos após as etapas Limpeza de Dados e Construção de Atributos totaliza 29 atributos. Desses atributos, 4 foram removidos na etapa Seleção de Atributos por possuírem valor de razão de ganho igual a zero, sendo eles: “idade”, “valoradicional”, “diabase‘” e “desconto”.

Após a consolidação dos atributos que devem representar os contratos, realizada nas etapas anteriores, é possível executar a etapa Balanceamento de Contratos e comparar as técnicas de balanceamento de classes. A partir desse ponto da abordagem, a base de dados é dividida em 4 grupos de contratos, representando 4 semestres entre 2013 e 2014. Os grupos são os seguintes:

- "2013_1": Contratos cuja data de adesão está situada no primeiro semestre de 2013;
- "2013_2": Contratos cuja data de adesão está situada no segundo semestre de 2013;
- "2014_1": Contratos cuja data de adesão está situada no primeiro semestre de 2014;
- "2014_2": Contratos cuja data de adesão está situada no segundo semestre de 2014.

Para realizar a comparação das técnicas de balanceamento, optou-se por utilizar o algoritmo C4.5 como classificador. São executados dois experimentos, Experimento A1 e Experimento B1. Os detalhes de qual base de dados foi utilizada para treinamento e teste são mostrados na Tabela 1. Os resultados obtidos pelos Experimento A e Experimento B são representados na Figura 7 e Figura 8, respectivamente.

Tabela 1: Parâmetros utilizados nos experimentos da etapa Balanceamento de Classes.

Experimento	Base de Treinamento	Base de Teste
A1	"2013_1"	"2013_2"
B1	"2014_1"	"2014_2"

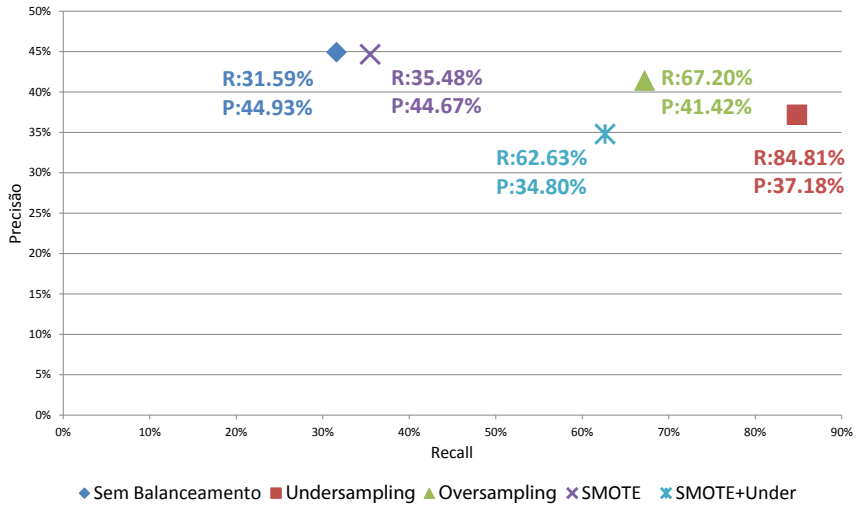

Figura 7: Resultados obtidos pelo Experimento A1.

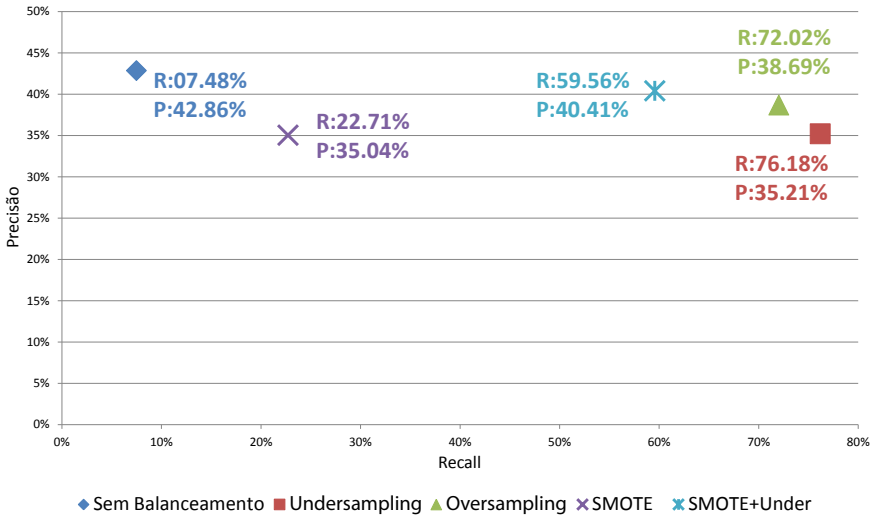

Figura 8: Resultados obtidos pelo Experimento B1.

É possível verificar que, em ambos os experimentos, a técnica de *undersampling* aleatório obteve as maiores taxas de *recall*, ou seja, ao utilizar essa técnica para o balanceamento de classes o algoritmo classificou, corretamente, o maior número de

contratos que deveriam ser rotulados como "cancelado". Porém, a mesma técnica de *undersampling* obteve as menores taxas de precisão, cerca de 36%, significando que mais de 60% dos contratos classificados como "cancelado" pertencem, na verdade, a classe "não cancelado".

A técnica *oversampling* aleatório apresenta resultados bem próximos da técnica *undersampling*, mas nota-se que há um redução na taxa de *recall* e um aumento na taxa de precisão. Isso significa que, ao utilizar o *oversampling*, o classificador "erra" menos ao rotular contratos como "cancelado"; porém, isso é obtido ao custo de classificar uma quantidade menor dos contratos.

Nota-se que quando não há a utilização de técnicas de balanceamento de classes, obtêm-se os piores resultados para a métrica *recall*. Dessa forma, a utilização de qualquer das estratégias listadas eleva a quantidade de contratos classificados que deveriam ter sido rotulados como "cancelado". Por se apresentarem como um conjunto de soluções não-dominadas, no qual uma métrica não sobrepuja outra, a comparação realizada permite a escolha da técnica mais adequada às necessidades da gestão. A análise das métricas de precisão e *recall* está diretamente relacionada às restrições impostas ao gestor do plano de saúde. Por exemplo, caso a política da empresa para evitar a perda do contrato "cancelado" seja oferecer descontos para o titular, dois possíveis cenários são os seguintes:

- **Cenário 1.** A empresa passa por uma ótima fase financeira e pode ofertar uma grande quantidade de descontos;
- **Cenário 2.** A empresa está em crise, reduzindo despesas, e só pode oferecer uma pequena quantidade de descontos.

No Cenário 1, como há um boa disponibilidade de recursos, o ideal é alcançar o maior número possível de contratos que tem tendência a serem cancelados no futuro, ou seja, o objetivo nesse caso é que a classificação obtenha a maior taxa de *recall* possível, mesmo que a precisão seja baixa e boa parte dos contratos classificados sejam, de fato, falsos positivos. Porém, no Cenário 2, a situação da empresa é mais delicada, e oferecer descontos para contratos que não irão cancelar, falso positivos, representaria um investimento talvez desnecessário. O Cenário 2 é um exemplo de cenário no qual o objetivo do classificador é aumentar a precisão, pois dessa forma, mesmo que poucos contratos sejam retornados, haverá uma certeza maior de que o rótulo "cancelado" está correto.

3.2 Fase de Mineração de Dados

Para a etapa de Reconhecimento de Contratos, foram definidos 3 experimentos: Experimento A2, Experimento B2 e Experimento C2. Para essa definição, foi utilizado um formato no qual um semestre serviu como base de treinamento e o semestre seguinte serviu como base de testes. Como a base de dados foi dividida em quatro grupos ("2013_1", "2013_2", "2014_1", "2014_2"), cada grupo representando um semestre, gerou-se os três experimentos detalhados na Tabela 2. Para cada métrica definida, área sob a curva ROC, *recall* e taxa de falso positivos (TFP), foi gerado um intervalo de confiança de 95% baseado na distribuição t-Student [20], pois os dados seguem, com 95% de confiança, a distribuição normal de acordo com o teste de Shapiro-Wilk [37]. Para cada experimento, as Tabelas 3, 4 e 5 mostram os intervalos de confiança para as métricas escolhidas. As Figuras 9, 10 e 11 representam a curva ROC do melhor algoritmo de cada paradigma de aprendizado[3] utilizado.

Tabela 2: Parâmetros utilizados nos experimentos da etapa Reconhecimento de Contratos.

Experimento	Base de Treinamento	Base de Teste
A2	"2013_1"	"2013_2"
B2	"2013_2"	"2014_1"
C2	"2014_1"	"2014_2"

Tabela 3: Resultados comparativos do Experimento A2 para as métricas: área sob curva ROC, *recall* e TFP.

Algoritmo	**ROC**	***Recall***	**TFP**
Bayes Net	**0.721 ∼ 0.821**	**60.0% ∼ 61.3%**	25.2% ∼ 26.7%
Naive Bayes	0.738 ∼ 0.742	51.9% ∼ 52.6%	**22.7% ∼ 23.0%**
KNN (*N*=1)	0.653 ∼ 0.658	64.7% ∼ 65.4%	34.0% ∼ 34.5%
KNN (*N*=3)	0.696 ∼ 0.699	64.5% ∼ 65.0%	34.7% ∼ 35.2%
KNN (*N*=5)	0.708 ∼ 0.712	64.8% ∼ 65.4%	33.7% ∼ 34.3%
K*	**0.757 ∼ 0.760**	**65.4% ∼ 65.9%**	**28.4% ∼ 29.1%**
MLP	**0.719 ∼ 0.726**	61.0% ∼ 62.3%	29.0% ∼ 30.6%
SVM	0.695 ∼ 0.699	**65.8% ∼ 67.2%**	**26.8% ∼ 27.3%**
C4.5	0.767 ∼ 0.779	70.5% ∼ 71.6%	**24.9% ∼ 25.9%**
RandomTree	0.702 ∼ 0.715	67.0% ∼ 68.6%	29.1% ∼ 30.5%
CART	**0.792 ∼ 0.801**	**73.3% ∼ 74.3%**	25.6% ∼ 27.4%

[3]Baseado em Exemplos, Bayesiano, Baseado em Árvore de Decisão e Conexionista

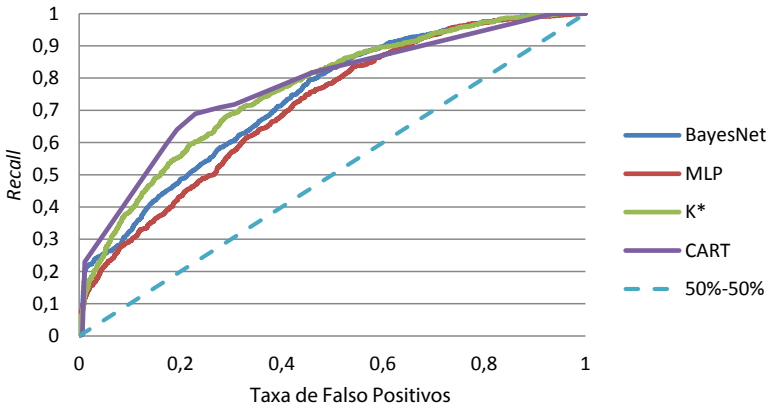

Figura 9: Gráfico da curva ROC dos melhores algoritmos de cada paradigma de aprendizado para o Experimento A2.

Tabela 4: Resultados comparativos do Experimento B2 para as métricas: área sob curva ROC, *recall* e TFP.

Algoritmo	ROC	*Recall*	TFP
Bayes Net	0.659 ∼ 0.671	**73.5% ∼ 74.4%**	50.7% ∼ 52.0%
Naive Bayes	**0.736 ∼ 0.739**	63.8% ∼ 64.5%	**30.7% ∼ 32.2%**
KNN (*N*=1)	0.606 ∼ 0.616	64.9% ∼ 66.6%	43.2% ∼ 44.2%
KNN (*N*=3)	0.666 ∼ 0.675	68.5% ∼ 69.6%	41.5% ∼ 42.9%
KNN (*N*=5)	**0.693 ∼ 0.699**	**69.8% ∼ 70.4%**	**40.1% ∼ 41.2%**
K*	0.662 ∼ 0.671	65.2% ∼ 66.7%	42.4% ∼ 43.7%
MLP	**0.696 ∼ 0.717**	69.9% ∼ 72.4%	**44.3% ∼ 46.8%**
SVM	0.624 ∼ 0.634	**74.9% ∼ 75.9%**	48.4% ∼ 50.9%
C4.5	**0.695 ∼ 0.721**	71.4% ∼ 73.3%	**38.7% ∼ 42.1%**
RandomTree	0.573 ∼ 0.600	66.6% ∼ 69.5%	49.4% ∼ 52.5%
CART	0.600 ∼ 0.611	**71.7% ∼ 74.4%**	49.3% ∼ 51.5%

Tabela 5: Resultados comparativos do Experimento C2 para as métricas: área sob curva ROC, *recall* e TFP.

Algoritmo	ROC	*Recall*	TFP
Bayes Net	0.663 ∼ 0.678	**71.4% ∼ 72.1%**	50.8% ∼ 52.5%
Naive Bayes	**0.685 ∼ 0.688**	71.4% ∼ 71.8%	**46.8% ∼ 47.7%**
KNN (*N*=1)	0.589 ∼ 0.597	64.5% ∼ 66.3%	46.1% ∼ 47.1%
KNN (*N*=3)	0.623 ∼ 0.630	64.2% ∼ 66.2%	45.0% ∼ 46.8%
KNN (*N*=5)	**0.637 ∼ 0.645**	**67.0% ∼ 69.0%**	**44.2% ∼ 46.2%**
K*	0.622 ∼ 0.626	65.6% ∼ 67.0%	46.2% ∼ 47.1%
MLP	**0.679 ∼ 0.692**	67.1% ∼ 70.5%	42.3% ∼ 45.8%
SVM	0.654 ∼ 0.660	**73.1% ∼ 73.8%**	**41.3% ∼ 42.7%**
C4.5	**0.677 ∼ 0.687**	71.5% ∼ 73.7%	**41.9% ∼ 43.2%**
RandomTree	0.613 ∼ 0.635	64.8% ∼ 70.2%	44.1% ∼ 47.6%
CART	0.654 ∼ 0.674	**72.2% ∼ 74.3%**	45.7% ∼ 47.6%

Nota-se, pelas métricas avaliadas que, na maioria dos experimentos, os paradigmas de aprendizado bayesiano e baseado em árvores de decisão obtiveram os maiores valores para a área sob a curva ROC e os menores valores para a taxa de falso positivos. Apenas no Experimento C2, o paradigma conexionista obtém melhores resultados, principalmente apresentando menores valores para TFP. E por fim, aparece o paradigma baseado em exemplos, cujos resultados para os algoritmos utilizados não tiveram destaque quando comparados aos demais paradigmas. Ressalta-se que os parâmetros de configuração, para cada algoritmo utilizado, foram os padrões estabelecidos pela ferramenta WEKA.

Ao se analisar a curva ROC do melhor algoritmo de cada paradigma, observa-se uma superioridade do C4.5 e CART para taxas de falso positivos inferiores a algo em torno de 40%, como pode ser notado na Figura 9 e Figura 10, implicando em um maior valor para a métrica de *recall* quando o valor de TFP é baixo. Isso significa que esses algoritmos são mais indicados caso se queira entregar informações mais confiáveis ao gestor, mesmo que isso custe ignorar informações de outros contratos que poderiam ser analisados. Por esse motivo, escolheu-se o C4.5 como técnica para realizar a classificação dos contratos ativos. Vale ressaltar que utilizar mais de um algoritmo e combinar diversos classificadores é uma estratégia com possibilidade de obter melhores resultados para as métricas definidas, mas que não foi realizada neste trabalho. Um exemplo básico dessa combinação poderia ser uma votação por média dos resultados encontrados pelos classificadores mais promissores, como o C4.5, CART e NaiveBayes.

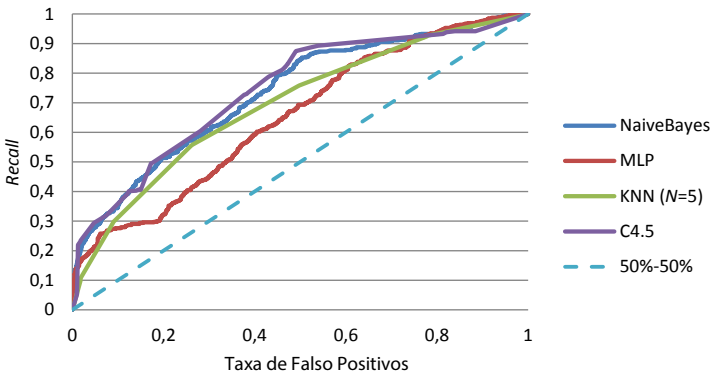

Figura 10: Gráfico da curva ROC dos melhores algoritmos de cada paradigma de aprendizado para o Experimento B2.

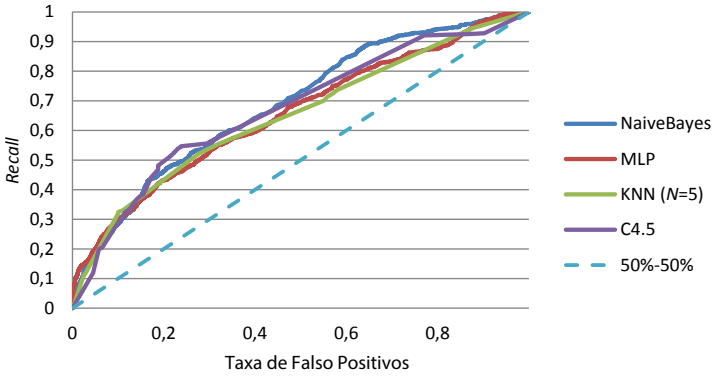

Figura 11: Gráfico da curva ROC dos melhores algoritmos de cada paradigma de aprendizado para o Experimento C2.

Outro ponto que pode ajudar a entender a dinâmica do plano de saúde é uma melhor definição do formato dos experimentos. Para o formato escolhido, as bases de treinamento e de testes foram definidas como semestres exatos, o que pode comprometer os resultados encontrados na classificação. Uma estratégia seria utilizar uma definição dinâmica dos conjuntos de treinamento e teste, de forma a descobrir o formato mais adequado ao contexto do cancelamento de contratos. Importante ressaltar que essa definição dinâmica pode ser ainda mais eficaz se também houver mudanças na etapa Limitação de Histórico, pois é de acordo com o período escolhido por essa etapa que se define o tempo necessário para avaliar o histórico de um contrato.

A partir da definição do C4.5 como classificador padrão, da etapa Reconhecimento de Contratos, é possível executar a classificação e encaminhar os contratos rotulados como "cancelado" para a fase de Priorização de Contratos. Porém, antes de um maior detalhamento dessa fase são apresentados os resultados obtidos pela etapa Identificação de Características, por meio da análise das árvores de decisão geradas pelos algoritmos C4.5 e CART. Para criação do modelo desses classificadores, é utilizada a base "2013_1" como base de treinamento e, para a avaliação das taxas de acerto, é utilizada a base "2013_2" como base de testes. Os 5 ramos com as melhores taxas de acerto, foram:

- **Ramo 1:** "Valor Total $<$ RS 62" $\rightarrow$ "Qtd. Tratamentos $>$ 1" $\rightarrow$ "Último Tratamento $<$ 75 dias" $\rightarrow$ **"Não Cancelado (100%)"**.

  O Ramo 1 mostra, com uma taxa de acerto igual a 100%, que o titular não cancela o contrato quando o valor total é menor do que R$ 62,00, a quantidade de tratamentos é maior do que 1 e o último tratamento realizado foi até 75 dias antes da data limite do contrato. Com uma taxa de acerto bastante expressiva, esse ramo evidencia que, quando o titular paga um contrato relativamente barato e usou recentemente a cobertura do plano, ele parece não ter interesse em cancelar o seu vínculo. Isso poderia guiar o gestor a criar políticas de incentivo ao uso do plano de saúde, principalmente para contratos mais baratos, com o objetivo de aproximar o titular e fortalecer a ideia de que o investimento em saúde é importante.

- **Ramo 2:** "Tem Tratamento = SIM" $\rightarrow$ "Qtd. Ocorrências $>$ 7" $\rightarrow$ "Último Tratamento $>$ 94 dias" $\rightarrow$ **"Cancelado (87,5%)"**.

  O Ramo 2 mostra, com um taxa de acerto igual a 87,5%, que o titular cancela o contrato quando já fez algum tratamento, teve registrado mais de 7 ocorrências e o último tratamento foi realizado pelo menos 94 dias antes da data limite do contrato. Esse ramo apresenta um informação interessante, pois o titular do plano tende a cancelar seu contrato quando há um número de ocorrências relativamente alto e o último tratamento realizado não é recente. Essa quantidade de ligações pode representar dúvidas e/ou reclamações que o titular ou dependentes estão tendo sobre o plano de saúde. Uma política

para contratos que se encaixem no Ramo 2 poderia ser a definição de contatos de aproximação com os beneficiários do contrato, visando descobrir se há dúvidas, reclamações e/ou sugestões a respeito do atendimento oferecido pela empresa.

- **Ramo 3:** "Tem Tratamento = SIM" → "Qtd. Tratamentos > 7" → "Último Tratamento <= 94 dias" → "Pagamento <= Boleto ou Conta Corrente" → **"Não Cancelado (94,4%)"**.

  O Ramo 3 mostra, com uma taxa de acerto igual a 94,4%, que o titular não cancela o contrato quando já fez algum tratamento, teve registrado mais de 7 ocorrências, o último tratamento realizado foi no máximo 94 dias antes da data limite do contrato e o tipo de pagamento escolhido foi boleto ou conta corrente. Esse ramo deixa mais evidente que a utilização recente do plano de saúde é importante para que o titular mantenha seu contrato ativo, pois percebe-se, ao se comparar com o Ramo 2, que a diferença básica entre os dois ramos é justamente o tempo em que foi realizado o último tratamento. O Ramo 3 reforça as potenciais políticas discutidas no Ramo 1 e no Ramo 2.

- **Ramo 4:** "Tem Tratamento = SIM" → "Qtd. Ocorrências <= 7" → **"Não Cancelado (93,8%)"**.

  O Ramo 4 mostra, com uma taxa de acerto igual a 93,8%, que o titular não cancela o contrato quando já fez algum tratamento e teve registrado menos de 7 ocorrências. Dessa baixa quantidade de ocorrências, pode-se entender que o titular do plano e seus dependentes parecem estar satisfeitos com o plano de saúde, pois já foram atendidos pelo menos uma vez e não entram muito em contato para esclarecer dúvidas e realizar reclamações por telefone. Nesse caso, o gestor poderia definir políticas visando ratificar essa tendência para o não cancelamento do contrato, como ligações periódicas para os integrantes do contrato.

- **Ramo 5:** "Última Situação = SUSPENSO" → **"Cancelado (98,8%)"**.

  O Ramo 5 mostra, com uma taxa de acerto igual a 98,8%, que o titular cancela quando a última situação do contrato é "suspenso". Esse ramo poderia representar uma informação nova para a gestão, mas na verdade representa uma informação valiosa para a abordagem proposta. Ao discutir o Ramo 5 com um especialista do plano do saúde, notou-se que a situação "suspenso" é automaticamente atribuída a um contrato que passa uma determinada quantidade de tempo sem realizar o pagamento das mensalidades, e também, automaticamente, esse contrato passa a ser considerado cancelado caso a situação permaneça. Dessa forma, a gestão já percebe um contrato suspenso por falta de pagamento como um contrato cancelado. Conclui-se, então, que os contratos suspensos não devem ser selecionados na etapa Seleção de Contratos da fase de Pré-Processamento, pois a adição desses contratos influencia em resultados melhores para a classificação realizada na etapa Reconhecimento de Contratos, visto que "suspenso" e "cancelado" podem ser considerados sinônimos na base de dados analisada.

A Figura 12 representa o Experimento A3, no qual são comparadas as curvas ROC do algoritmo C4.5 aplicado em duas situações: com contratos suspensos e sem contratos suspensos. A base "2013_1" foi utilizada para treinamento e a base "2013_2" como teste. É possível perceber que há um diferença significativa na parte inicial das curva ROC, pois para um menor taxa de falso positivos, a inclusão dos contratos suspensos favorece a classificação, pois, como já discutido, o nível de certeza é alto quando se classifica um contrato suspenso com o rótulo "cancelado".

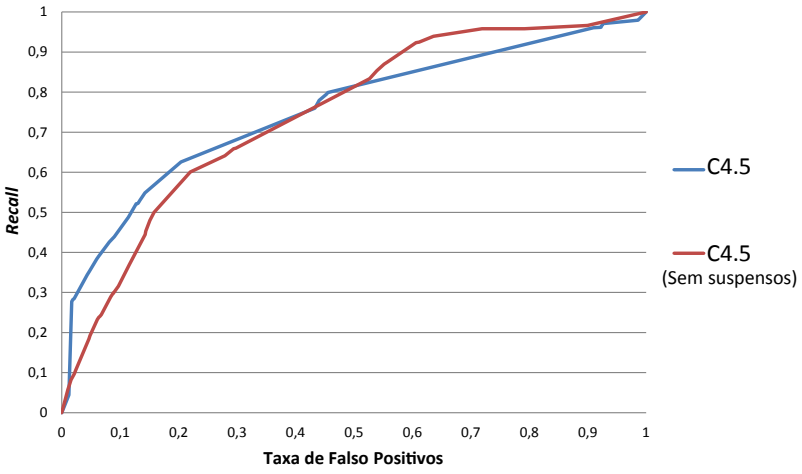

Figura 12: Curvas ROC obtidas no Experimento A3.

3.3 Fase de Priorização de Contratos

Na fase de Priorização de Contratos, são utilizados os contratos classificados como "cancelado" provenientes da etapa Reconhecimento de Contratos. O nível de certeza dessa classificação e o valor total do contrato são entradas para o sistema de

inferência *fuzzy* utilizado na etapa Determinação do Grau de Cancelamento. Após esse processo, a etapa Ordenação de Contratos organiza os contratos de acordo com o grau de cancelamento estimado, visando ordenar do contrato mais crítico (maior grau de cancelamento) ao contrato menos crítico (menor grau de cancelamento).

A Tabela 6 mostra o nível de certeza, o valor total normalizado[4], o valor total em reais e o grau de cancelamento dos 5 (cinco) contratos mais críticos processados pela fase de Priorização de Contratos. Esses contratos fazem parte da base "2013_2", utilizada como base de testes no Experimento A2. Pode-se notar, pelos resultados, que o nível de certeza influencia mais no grau de cancelamento do que o valor total do contrato; isso é um reflexo da base de regras definida pelo especialista e exposta na Seção 2.4.

Tabela 6: Os 5 contratos mais críticos resultantes da fase de Priorização de Contratos.

#	Nível de Certeza	Valor Total (norm.)	Valor Total	Grau de Cancelamento
1	1,0 (100%)	1,0	R$ 617,72	0,9163
2	1,0 (100%)	0,5120	R$ 269,96	0.9163
3	1,0 (100%)	0,6115	R$ 322,44	0,9130
4	1,0 (100%)	0,7992	R$ 421,43	0,9068
5	1,0 (100%)	0,7217	R$ 380,57	0,9050

Ressalta-se que as definições adotadas para o sistema de inferência ainda podem ser bastante melhoradas. É possível perceber que a ordem do 2º e do 3º contratos mais críticos parece estar invertida, pois apesar de ambos possuírem o mesmo nível de certeza, o 3º contrato representa um valor total superior ao 2º. Isso deve-se ao fato de a normalização colocar o 2º contrato com uma alta pertinência para o termo linguístico "médio", o que influencia diretamente na Regra 8[5] da base de regras e induz a um valor alto para o grau de cancelamento.

4. 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 mais de 88%, o que permitiu a análise de uma conjunto mais específico de elementos. Além disso, houve uma redução ainda maior na dimensionalidade do problema, pois menos de 10% dos 230 atributos disponíveis foram selecionados. Por se tratar de um conjunto de dados no qual há um desbalanceamento de classes, realizou-se uma análise comparativa entre diferentes técnicas de balanceamento, visando entender a influência da utilização de cada uma sobre a classificação dos contratos. Confirmou-se o *undersampling* aleatório como uma técnica efetiva, pois, nos experimentos realizados, ela obteve os melhores resultados para a métrica *recall*. Porém, explanou-se por meio de exemplos que, dependendo da situação política/financeira da empresa, o *undersampling* pode não ser a técnica mais indicada, pois técnicas com melhores taxas de precisão do que de *recall* influenciariam em resultados mais apropriados às necessidades da gestão.

Na fase de Mineração de Dados, realizou-se uma comparação entre classificadores de diferentes paradigmas de aprendizado, o que permitiu, de forma experimental, definir o algoritmo mais adequado para um melhor funcionamento da abordagem proposta. Apesar do resultado da classificação dos contratos não ter sido elevado, com valores de *recall* entre 70% e 75%, considera-se um resultado promissor por estar significantemente acima de 50%, resultado esse que poderia ser produto do acaso. 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 87%, notou-se características que têm influência direta na rotulação de um contrato como "cancelado" ou "não cancelado", informação que 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 se mostrou como uma fase importante na apreensão do conhecimento tácito do gestor sobre a criticidade dos contratos. Por meio da utilização de um sistema de inferência *fuzzy*, foi possível condensar o conhecimento do especialista em uma base de regras utilizada no processo de estimar um grau de cancelamento para cada contrato classificado. A partir desse grau de cancelamento, foi possível não só apresentar os contratos rotulados como "cancelado", mas também atribuí-los uma métrica de criticidade e organizá-los de forma a facilitar a priorização dos contratos com maiores chances de cancelamento.

Conclui-se com este trabalho que a abordagem proposta promoveu resultados promissores para a classificação dos contratos ativos e identificação de características relevantes, para contratos cancelados e não cancelados, que podem auxiliar o gestor na tomada de decisão, no que diz respeito ao problema do cancelamento eletivo de contratos. Ratifica-se, então, o uso de técnicas de

[4]O valor total normalizado é igual ao valor total de um contrato específico pelo maior valor encontrado entre os contratos da base de dados, com o resultado dessa divisão limitado ao valor máximo de 1

[5]se NC é ALTO e VT é **MÉDIO**, então GC é MUITO ALTO

mineração de dados e descoberta de conhecimento em bancos de dados como estratégias aliadas no entendimento de problemas ligados à saúde suplementar.

4.1 Limitações

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

- A limitação de histórico dos contratos em 6 meses influencia diretamente na classificação dos contratos. Maiores ou menores períodos podem retratar aspectos diferentes do cancelamento dos contratos. 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;
- Não foi desenvolvida uma estratégia para combinar diferentes algoritmos para realizar a rotulação dos contratos. A união entre diferentes classificadores pode melhorar o resultado da classificação;
- Os ramos analisados na etapa Reconhecimento de Características foram definidos de forma manual, baseando-se na taxa de acerto de cada um. Essa taxa de acerto é baseada na frequência de elementos que foram classificados no ramo, ou seja, um determinado ramo pode ter taxa de acerto de 100% se apenas um elemento for identificado, enquanto outro ramo pode ter 99% de taxa de acerto se 99 elementos forem identificados corretos e apenas um for identificado erroneamente. Outras métricas mais completas podem ser utilizadas para identificar a corretude dos ramos, assim como formas automáticas para definir os mais importantes;
- O sistema de inferência *fuzzy* utiliza apenas duas entradas para a estimação do grau de cancelamento; todavia, outras entradas podem ser importantes na determinação dessa saída. Os resultados obtidos não representaram tão bem o conceito de criticidade, uma melhor definição da base de regras e do formato das funções de pertinência devem ser estudados;
- Os algoritmos foram utilizados com a parametrização padrão oferecida pela ferramenta WEKA. Uma definição mais sistemática desses parâmetros pode implicar em melhores resultados para a classificação dos contratos;
- 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.

4.2 Continuidade da Pesquisa

Baseando-se nas limitações encontradas e em ideias para complementação deste trabalho, as seguintes linhas são consideradas pontos de continuidade da pesquisa:

- Melhorar a etapa Limitação de Histórico, realizando experimentos para que seja possível comparar diversos tamanhos de período, não somente 6 meses, e detectar o mais adequado para analisar a base de dados;
- Adicionar novas técnicas à etapa Limpeza de Dados, como a remoção de atributos numéricos a partir de um filtro por variância. Alguns dos atributos numéricos selecionados possuem uma distribuição e dispersão pouco representativas;
- Utilizar outros tipos de técnicas para mitigar o problema do desbalanceamento de classes, como seleção unilateral [25] e limpeza de vizinhança [27];
- Avaliar a combinação entre os melhores classificadores por meio de estratégias de classificação que levem em consideração a execução de cada algoritmo. Além disso, avaliar também, uma melhoria na execução de cada algoritmo de forma individual, utilizando técnicas como *bagging* e *boosting* [13];
- Utilizar uma métrica melhor para a taxa de acerto dos ramos da árvore de decisão, como a correção de Laplace [31, 34] para a taxa de acerto baseada apenas na frequência. Por meio de uma métrica melhorada, também se objetiva automaticamente detectar os melhores ramos a serem analisados, levando em consideração, de forma complementar, outros elementos como tamanho do ramo e relação entre ramos que são semelhantes mas classificam de forma diferente;
- Melhorar o sistema de inferência *fuzzy* utilizado, desde a definição das entradas às definições da funções de pertinência;
- Explorar os parâmetros dos algoritmos classificadores visando obter configurações mais adequadas ao conjunto de dados disponível. Essa exploração será feita de forma automática, utilizando meta-heurísticas para realizar o *tuning* dos parâmetros;
- 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, principalmente;

- Adicionar à abordagem proposta, na fase Mineração de Dados, uma etapa que execute uma regressão para estimar o tempo até o cancelamento dos contratos rotulados como "cancelado". Essa estimação, além de prover uma informação adicional à gestão, também será utilizada na etapa Determinação do Grau de Cancelamento, pois está relacionada à definição da criticidade de um contrato.

REFERÊNCIAS

[1] Ans, 2015. URL `http://www.ans.gov.br/perfil-do-setor/dados-e-indicadores-do-setor`.

[2] David Aldous. The continuum random tree. i. *The Annals of Probability*, 19(1):1–28, jan 1991. doi: 10.1214/aop/1176990534.

[3] Shunichi Amari. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, EC-16(3):299–307, June 1967. ISSN 0367-7508. doi: 10.1109/PGEC.1967.264666.

[4] Flávio Henrique Duarte de Araújo, Santana André Macedo, and Pedro de Alcântara Santos Neto. An Approach Influenced to Pre-processing for Learning Medical Claim Process. *Journal of Health Informatics*, 7(1):8–15, 2015.

[5] Sanjay Basu, Jason Andrews, Sandeep Kishore, Rajesh Panjabi, and David Stuckler. Comparative performance of private and public healthcare systems in low- and middle-income countries: A systematic review. *PLoS Medicine*, 9(6):e1001244, jun 2012. doi: 10.1371/journal.pmed.1001244.

[6] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984. ISBN 9780412048418.

[7] Stephen Chen, James Montgomery, and Antonio Bolufé-Röhler. Measuring the curse of dimensionality and its effects on particle swarm optimization and differential evolution. *Applied Intelligence*, 42(3):514–526, nov 2015. doi: 10.1007/s10489-014-0613-2.

[8] Jie Cheng and Russell Greiner. Comparing bayesian network classifiers. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 101–108, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-614-9.

[9] Pablo Cingolani and Jesus Alcala-Fdez. jFuzzyLogic: a robust and flexible fuzzy-logic inference system language implementation. Institute of Electrical & Electronics Engineers (IEEE), jun 2012.

[10] John G. Cleary and Leonard E. Trigg. K*: An instance-based learner using an entropic distance measure. In *International Conference on Machine Learning*, pages 108–114, 1995.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411.

[12] David M. Cutler and Richard J. Zeckhauser. Chapter 11 the anatomy of health insurance. In *Handbook of Health Economics*, pages 563–643. Elsevier BV, 2000. doi: 10.1016/s1574-0064(00)80170-5.

[13] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-67704-8.

[14] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, jun 2006.

[15] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. ISBN 0-262-56097-6.

[16] Evelyn Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238, dec 1989. doi: 10.2307/1403797.

[17] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera. 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*, 42(4):463–484, July 2012. ISSN 1094-6977. doi: 10.1109/TSMCC.2011.2161285.

[18] V. García, J.S. Sánchez, and R. A. Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2012. ISSN 09507051.

[19] T. L. Oshini Goonetilleke and H. a. Caldera. Mining Life Insurance Data for Customer Attrition Analysis. *Journal of Industrial and Intelligent Information*, 1(1):52–58, 2013.

[20] W. S. Gosset. The Probable error of a mean. *Biometrika*, 6(1):1–25, mar 1908. doi: 10.1093/biomet/6.1.1.

[21] Catarina Goulão. Voluntary public health insurance. *Public Choice*, 162(1-2):135–157, 2014. ISSN 0048-5829. doi: 10.1007/s11127-014-0207-x.

[22] Daniela Grigori, Fabio Casati, Malu Castellanos, Umeshwar Dayal, Mehmet Sayal, and Ming-Chien Shan. Business process intelligence. *Computuers in Industry*, 53(3):321–343, April 2004. ISSN 0166-3615. doi: 10.1016/j.compind.2003.10.007.

[23] Hongxing He, Warwick Graco, and Xin Yao. Application of genetic algorithm and k-nearest neighbour method in medical fraud detection. In Bob McKay, Xin Yao, CharlesS. Newton, Jong-Hwan Kim, and Takeshi Furuhashi, editors, *Simulated Evolution and Learning*, volume 1585 of *Lecture Notes in Computer Science*, pages 74–81. Springer Berlin Heidelberg, 1999. ISBN 978-3-540-65907-5.

[24] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-385-9.

[25] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.

[26] Mohit Kumar, Rayid Ghani, and Zhu-Song Mei. Data mining to predict and prevent errors in health insurance claims processing. In *SIGKDD international conference on Knowledge discovery and data mining*, pages 65–74, 2010. ISBN 978-1-4503-0055-1. doi: 10.1145/1835804.1835816.

[27] Jorma Laurikkala. Improving identification of difficult small classes by balancing class distribution. In *Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine*, AIME '01, pages 63–66, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-42294-3.

[28] Wendy K Mariner. Health Insurance Is Dead; Long Live Health Insurance. *American Journal of Law & Medicine*, 40: 195–214, 2014.

[29] Charles Normand. *Social health insurance a guidebook for planning*. VAS, Bad Homburg v.d.H, 2009. ISBN 978-3-88864-491-7.

[30] Pedro A Ortega, Gonzalo A Ruz, and Cristian J. Figueroa. A Medical Claim Fraud / Abuse Detection System based on Data Mining: A Case Study in Chile. *Proceedings of International Conference of Data Mining*, 2006. doi: 10.1.1.176.796.

[31] Michael Pazzani, Christopher Merz, Patrick Murphy, Kamal Ali, Timothy Hume, and Clifford Brunk. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 217–225, 1994.

[32] Witold Pedrycz. Why triangular membership functions? *Fuzzy Sets and Systems*, 64(1):21–30, may 1994. doi: 10.1016/0165-0114(94)90003-5.

[33] Witold Pedrycz. *Fuzzy systems engineering toward human-centric computing*. John Wiley IEEE, Hoboken, N.J, 2007. ISBN 978-0471788577.

[34] Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52(3):199–215, 2003. ISSN 0885-6125. doi: 10.1023/A:1024099825458.

[35] J.R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, sep 1987. doi: 10.1016/s0020-7373(87)80053-6.

[36] Neelam Sekhri and William Savedoff. Policy and Practice Private health insurance : implications for developing countries. *Bulletin of the World Health Organization*, 010611(03):127–134, 2005. ISSN 00429686.

[37] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591, dec 1965. doi: 10.2307/2333709.

[38] Jin Su, Kimberly Cooper, Tina Robinson, and Brad Jordan. Customer Retention Predictive Modeling in HealthCare Insurance Industry. In *SESUG Southeast SAS Users Group*, pages 1–8, 2009.

[39] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 2011. ISBN 978-0-12-374856-0.

[40] Janusz Wojtusiak, Che Ngufor, John Shiver, and Ronald Ewald. Rule-based prediction of medical claims' payments: A method and initial application to medicaid data. *Proceedings of International Conference on Machine Learning and Applications*, 2:162–167, 2011. doi: 10.1109/ICMLA.2011.126.