



Programa Interdisciplinar de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Ricardo Gerhardt

Um estudo para identificar fatores que conduzem ao atraso no processo de reembolso de contas hospitalares via mineração de processos e mineração de dados

São Leopoldo, 2018

UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO
EM COMPUTAÇÃO APLICADA
NÍVEL MESTRADO

RICARDO GERHARDT

UM ESTUDO PARA IDENTIFICAR FATORES QUE CONDUZEM AO
ATRASO NO PROCESSO DE REEMBOLSO DE CONTAS HOSPITALARES
VIA MINERAÇÃO DE PROCESSOS E MINERAÇÃO DE DADOS

São Leopoldo

2018

G368e Gerhardt, Ricardo.

Um estudo para identificar fatores que conduzem ao atraso no processo de reembolso de contas hospitalares via mineração de processos e mineração de dados / Ricardo Gerhardt. – 2018.

83 f. : il. color. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2018.

1. Mineração de dados (Computação). 2. Cuidados médicos - Custos. 3. Mineração de processo. I. Título.

CDU 004.65

Dados Internacionais de Catalogação na Publicação (CIP)

(Bibliotecária: Bruna Sant'Anna – CRB 10/2360)

Ricardo Gerhardt

Um estudo para identificar fatores que conduzem ao atraso no processo de reembolso de contas hospitalares via mineração de processos e mineração de dados

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre, pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos – UNISINOS

Aprovado em 28 Março de 2018

BANCA EXAMINADORA

Prof. Dr. José Vicente Canto dos Santos – UNISINOS

Prof. Dr. Sandro José Rigo - UNISINOS

Prof. Dr. Silvio Cesar Cazella - UFCSPA

Prof. Dr. José Vicente Canto dos Santos – UNISINOS

Visto e permitida a impressão

São Leopoldo, 2018

Prof. Dr. Rodrigo da Rosa Righi
Coordenador PPG em Computação Aplicada

Dedico este trabalho aos meus pais, Paulo e Lia.

AGRADECIMENTOS

Agradeço primeiramente a Deus, aos meus pais, Paulo e Lia, e a toda a minha família, que, com muito carinho e apoio, não mediram esforços para que eu concluísse mais esta etapa da minha vida.

Agradeço ao Hospital Bruno Born, em especial ao Guilherme Redeker pelo apoio e pela disponibilização do material necessário para a realização deste projeto.

A todos os professores e colegas do curso, especialmente ao professor Dr. João F. Valiati, pela grande dedicação, conhecimentos e atenção que foram fundamentais durante o desenvolvimento deste projeto.

RESUMO

O impacto do processo de reembolso das despesas médico-hospitalares das prestadoras de serviço de saúde tem sido enorme. Com o surgimento de novos procedimentos clínicos, mudanças em regulamentações e políticas há uma elevação da complexidade do processo de reembolso e conseqüentemente a sua duração e seus custos. Desse modo, métodos de análise de processos têm sido empregados como estratégia básica para melhorar a eficácia organizacional de instituições hospitalares. Perante a isso, o presente trabalho investiga fatores que levam ao atraso da submissão das contas hospitalares às respectivas seguradoras de saúde no sentido de reduzir seu tempo de faturamento. A abordagem proposta constitui-se em combinar técnicas da Mineração de Processos e Mineração de Dados com o intuito de identificar fatores que contribuem para o atraso do processo de reembolso. A Mineração de Processos permite vislumbrar detalhadamente o impacto causado pela realização de atividades durante a execução de processos, bem como a ocorrência de gargalos que podem indicar a necessidade de uma investigação mais apurada para detectar as suas prováveis causas. Nesse ponto, a Mineração de Dados pode ser empregada através de técnicas, como as regras associativas que possibilitam identificar relacionamentos não tão evidentes. Desta forma, este estudo investigativo demonstra sobre um caso real os benefícios do emprego da Mineração de Processos e da Mineração de Dados objetivando fornecer suporte as atividades de auditoria e de faturamento do processo de reembolso. A avaliação subjetiva das regras associativas mostrou que quase 45% das regras associativas geradas foram consideradas relevantes ou muito relevantes para a identificação de fatores que contribuem para o atraso no processo de reembolso de contas hospitalares.

Palavras-Chave: Reembolso de contas hospitalares. Mineração de Processo. Mineração de Dados.

ABSTRACT

The healthcare reimbursement process impact has been enormous for the healthcare providers and the economy. The arising of new clinical procedures, changes in regulations and policies have been increasing the complexity of the reimbursement process and consequently its duration and costs. Therefore, methods of process analysis have been used as a basic strategy to improve the organizational effectiveness of healthcare institutions. In this context, the present study investigates factors that cause delays in the reimbursement process. The proposed approach aims to combine Process Mining and Data Mining techniques to identify factors that can explain the reimbursement process delay. Process Mining techniques allow exploring in detail how activities can impact the process execution, as well as the occurrence of bottlenecks that may indicate the need for a systematic investigation to detect its root causes. Considering this, Data Mining can be employed through techniques, such as associative rules that can be used to identify unknown relationships. Hence, this study demonstrates through a real case the benefits that the combination of Process Mining and Data Mining techniques to support the audit and billing activities of the reimbursement process. A subjective evaluation of the mined rules showed that almost 45% of them were considered relevant or very relevant for the identification of factors that can lead to delay in the reimbursement process.

Keywords: Reimbursement Process. Process Mining. Data Mining.

LISTA DE FIGURAS

Figura 1: Ciclo de vida do BPM.....	30
Figura 2: Modelo de processo de requisição de compras.....	34
Figura 3: Etapas do processo de descoberta de conhecimento.....	37
Figura 4: Geração de itens frequentes e candidatos com suporte mínimo igual a 2.....	39
Figura 5: Árvore de Decisão.....	40
Figura 6: Metodologia Proposta.....	51
Figura 7: Duração do processo.....	58
Figura 8: Duração do processo em função do mês de início.....	59
Figura 9: Duração do processo em função da especialidade médica (esquerda) e do tipo do procedimento e caráter de internação (direita).....	59
Figura 10: Casos iniciados e encerrados.....	60
Figura 11: Modelos do convênio A enriquecidos com a frequência (esquerda) e desempenho (direita).....	62
Figura 12: Resultado do questionário.....	66

LISTA DE TABELAS

Tabela 1: Exemplo Log de Eventos.....	35
Tabela 2: Conjunto de transações.....	38
Tabela 3: Atividades presente no log de eventos e seus atributos.....	56
Tabela 4: Atributos do caso e sua descrição.....	57
Tabela 5: Atributos numéricos e categóricos.....	58
Tabela 6: Importância dos atributos.....	64
Tabela 7: Regras geradas a partir do conjunto A.....	65
Tabela 8: Regras geradas a partir do conjunto B.....	65

LISTA DE QUADROS

Quadro 1: Algoritmo Random Forest.....	42
--	----

LISTA DE SIGLAS

AIH	Autorização de Internação Hospitalar
BPM	<i>Business Process Management</i>
CART	<i>Classification and Regression Trees</i>
CSV	<i>Comma-Separated Values</i>
KDD	<i>Knowledge Discovery in Databases</i>
MD	Mineração de Dados
MP	Mineração de Processos
SUS	Sistema Único de Saúde
TI	Tecnologia da Informação
WFM	Workflow Management System

SUMÁRIO

1 INTRODUÇÃO	26
1.1 Objetivos	27
1.2 Organização do trabalho	28
2 FUNDAMENTAÇÃO TEÓRICA	30
2.1 Gerenciamento de Processos de Negócio	30
2.2 Processos Hospitalares.....	31
2.2.1 Processo de reembolso despesas médico-hospitalares	32
2.3 Mineração de Processos	33
2.3.1 Tarefas da Mineração de Processos	33
2.3.2 Log de Eventos	34
2.3.3 Fuzzy Miner.....	36
2.4 Descoberta de Conhecimento em Bases de Dados	36
2.5 Mineração de Regras de Associação.....	38
2.6 Classificação de Dados	39
2.6.1 Árvores de Decisão	40
2.6.2 <i>Random Forest</i>	41
3 TRABALHOS RELACIONADOS	44
3.1 Metodologia de pesquisa bibliográfica.....	44
3.2 Análise de processos hospitalares através de MP.....	44
3.3 Uso de abordagens da MD para a análise de processos hospitalares.....	46
3.4 Integração de técnicas da MP e MD	47
3.5 Considerações.....	48
4 METODOLOGIA	51
4.1 Considerações.....	52
5 RESULTADOS	54
5.1 Ferramentas e bibliotecas utilizadas	54
5.2 Entendimento do Negócio.....	54
5.3 Seleção e Exploração	56
5.4 Mineração de Processos	60
5.5 Pré-processamento e Transformação de Dados	63
5.6 Mineração de Dados	64
5.7 Avaliação.....	65
5.8 Considerações.....	67
6 CONCLUSÃO	69
REFERÊNCIAS	71
ANEXO A CONJUNTO DE REGRAS A	76
ANEXO B CONJUNTO DE REGRAS B	78
ANEXO C QUESTIONÁRIO	80

1 INTRODUÇÃO

O aumento contínuo de gastos com serviços de saúde e cuidados médicos, combinados com a complexidade crescente do processo de reembolso de contas médico-hospitalares, vem gerando desafios operacionais para as instituições de saúde (BRILL, 2015). Somente nos Estados Unidos foram gastos com saúde o equivalente a 3,4 trilhões de dólares, 4,8% a mais que no ano de 2015 (Keehan et al., 2017). No Brasil em 2015, os gastos com saúde somaram 388,7 bilhões de reais, ou 6,4 % do PIB. Deste total, 204,4 bilhões foram gastos pelas famílias com serviços de saúde privados (incluindo planos de saúde) e 184,3 bilhões foram gastos pelo governo (IBGE, 2015). Além disso, o surgimento de novos procedimentos clínicos, mudanças constantes em regulamentações e políticas por parte de seguradoras de saúde privadas e do governo vêm aumentando a duração, os custos, e os erros operacionais ocorridos durante a execução do processo de reembolso (LEE et al., 2016).

Problemas de gestão operacional com o processo de reembolso podem afetar tanto a saúde financeira da instituição como também a prestação de cuidados médicos (LEE et al., 2016). Por exemplo, atrasos no pagamento das despesas por parte das seguradoras podem postergar ou interromper tratamentos médicos. Ou então, em casos extremos, quando seguradora e prestadora discordam sobre o tratamento empregado, o paciente pode até perder a cobertura do seguro. Além disso, conforme American Hospital Association – AHA (2016), o processo de contestação de contas médicas com alguma inconsistência¹ pode durar mais que 90 dias, quando não há um acordo entre as partes (AHA, 2016). Somente em 2015, 12,1% ou 43,3 bilhões de dólares de contas pagas pelo Medicare, seguradora de saúde do governo norte americano, foram contestadas (HHS, 2016).

Além desses problemas, comuns a muitas instituições Lee et al. (2016), instituições de saúde públicas e sem fins lucrativos brasileiras passam por outras dificuldades, que estão diretamente relacionadas ao processo de reembolso. Valores reembolsados pelo SUS (Sistema Único de Saúde) pelos serviços de saúde fornecidos de forma gratuita a população, raramente cobrem os seus custos. Ademais, a contribuição financeira feita pelo governo para as instituições públicas e sem fins lucrativos já atingiu um nível de esgotamento, comprometendo a manutenção (GRAGNOLATI et al, 2013). Isso tornou os hospitais brasileiros dependentes de contratos com seguradoras de saúde privadas, não só pela falta de apoio financeiro dado pelo governo (COSTA, 2017), mas também porque os gastos por pessoas com serviços de saúde privado são três vezes maiores do que os gastos com saúde pública (MARTEN et al, 2014).

Neste contexto, as instituições de saúde vêm buscando novas estratégias para melhorar o desempenho de seus processos. Porém, essa não é uma tarefa simples, pois sempre existirá a necessidade de se reduzir custos e, ao mesmo tempo, aumentar a capacidade de atendimento para suprir a crescente demanda (WILLIAMS, 2017; KEEHAN et al, 2017). Assim, métodos de análise de processos têm sido empregados como estratégia básica para melhorar a eficiência organizacional de instituições hospitalares, seja por meio da criação de novos modelos de processos ou então através do reprojeto de processos existentes (REBUGUE; FERREIRA, 2012; ROJAS, 2016).

¹ Contas com problemas de superfaturamento, subfaturamento, com serviços desnecessários ou não prestados, com erros de codificação ou então que não estavam de acordo com a regulamentação (HSS, 2016).

De acordo com Rojas (2016), a Mineração de Processos (MP) é um método que tem se mostrado promissor para a análise de processos do âmbito hospitalar. Ele tem como premissa básica extrair conhecimentos de informações armazenadas em base de dados que foram geradas durante a execução de um dado processo. Dessa forma, com as técnicas providas pela MP é possível gerar modelos que reflitam a realidade dos processos existentes e assim descobrir o seu real funcionamento. Também é factível verificar se um determinado processo está sendo executado conforme um modelo existente e com isso identificar desvios ou falhas durante a sua execução. Sendo possível, ainda, enriquecer os modelos de processos gerados com informações relativas ao desempenho do processo, a utilização de recursos e a frequência da execução das atividades (AALST, 2016).

Além disso, objetivando uma análise mais apurada dos processos, pode-se combinar as técnicas de MP com as da Mineração de Dados (MD) (AALST, 2016). Permitindo, por exemplo: prever e enriquecer um modelo de processo com o tempo necessário para a sua conclusão, dado que este tempo é obtido com o uso de técnicas de predição (AALST; SCHONENBERG; SONG, 2011); segregar um modelo complexo em menores com o uso de técnicas de agrupamento, facilitando assim a sua análise (DELIAS et al., 2015); descobrir os prováveis motivos que levaram a tomada de uma decisão em uma determinada etapa do processo, através de árvores de decisão (AALST, 2016).

Assim, a MP tem sido empregada, em conjunto ou não com técnicas de MD, para a resolução de diversos problemas através da análise de processos administrativos e clínicos, como por exemplo: melhorar o gerenciamento de instituições hospitalares; identificar e entender o real comportamento de recursos e pacientes; analisar o desempenho e desse modo reduzir filas de esperas e tempos de serviços; prever o comportamento de pacientes com base em dados históricos; identificar as causas de gargalos em processos, entre outros (MANS; AALST; VANWERSCH, 2015; ROJAS, 2016).

1.1 Objetivos

O objetivo principal deste trabalho é aplicar e investigar métodos da MP no sentido de identificar um conjunto de casos e características que promovam o atraso da submissão de contas hospitalares a seguradoras de saúde. Uma vez identificado um conjunto de casos, em função de intervalos de atrasos de interesse, aplicar a MD, por meio do uso de regras associativas, objetivando encontrar fatores que promovem o atraso no processo de reembolso de valores.

Para alcançar este objetivo foram definidos os seguintes objetivos específicos:

- Propor uma metodologia que combine as técnicas da Mineração de Processos e Mineração de Dados;
- Validar o conhecimento gerado de forma automática com especialistas da área de atuação;
- Repassar o conhecimento gerado para o hospital com o intuito de auxiliar a tomada de decisão.

1.2 Organização do trabalho

O presente trabalho está estruturado da seguinte forma: o segundo capítulo contextualiza os temas gerenciamento de processos de negócio, processos hospitalares e o processo de reembolso, bem como os temas MP e MD; no terceiro capítulo são expostos os trabalhos relacionados selecionados, os quais apresentam similaridades e diferenças com o presente trabalho; no quarto capítulo é apresentada a metodologia empregada; no quinto capítulo são apresentadas a descrição do processo analisado, as técnicas utilizadas, os resultados obtidos e considerações sobre os mesmos; o sexto e último traz a conclusão do estudo em função dos resultados obtidos.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo traz uma contextualização dos termos e técnicas empregadas no decorrer deste trabalho. Inicialmente, na seção 2.1 é apresentado a área de gerenciamento de processos de negócio, na sequência, seção 2.2, as peculiaridades dos processos hospitalares e do processo de reembolso são apresentadas.

O conjunto de seções 2.3 à 2.6 aborda uma série de técnicas que podem ser empregadas à resolução de problemas em processos hospitalares. Primeiramente a Mineração de Processos (MP) é apresentada seguida pela descoberta de conhecimento em base de dados. Por fim, são apresentadas técnicas de geração de regras de associação e classificadores, os quais podem ser combinados com as técnicas de Mineração de Dados (MD).

2.1 Gerenciamento de Processos de Negócio

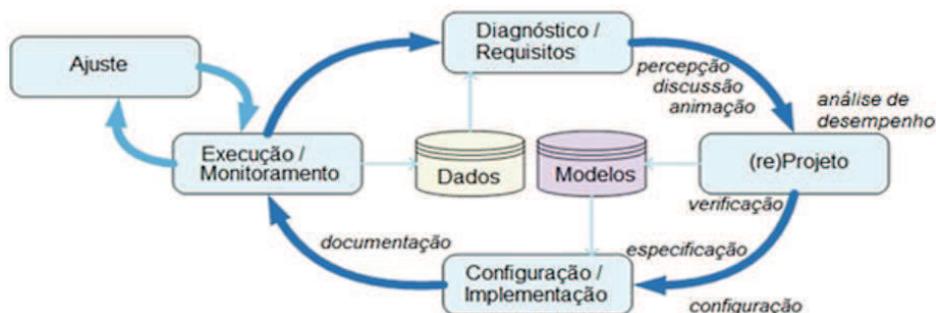
Processos de Negócios consistem de um ou mais subprocessos, tarefas ou atividades, realizadas de forma manual ou automatizada, que se relacionam para alcançar um determinado objetivo do negócio. Cada processo de negócio é inato de uma organização, mas pode interagir com processos de negócios de outras organizações (WESKE, 2012).

O Gerenciamento de Processos de Negócios (*Business Process Management - BPM*) é um mecanismo que visa aumentar a produtividade e reduzir os custos de uma organização, combinando e aplicando os conhecimentos de Tecnologia da Informação (TI) e de gestão aos processos de negócio. Para isto, o BPM inclui métodos, técnicas e ferramentas para apoiar a concepção, a representação, a gestão e a análise de processos de negócios (AALST, 2013).

Segundo Aalst (2013), o BPM também pode ser visto como sendo uma extensão de sistemas e abordagens clássicas de gestão de processos como o gerenciamento de *workflow* (*Workflow Management - WFM*). Enquanto o WFM tem como foco principal a automatização de processos de forma mecanizada, o BPM tem um alcance ainda maior, que vai desde a automação e a análise de processos até a gestão de processos e o trabalho organizacional.

O ciclo de vida do BPM, apresentado na Figura 1, é composto por quatro estágios: (i) Projeto, (ii) Configuração/Implementação (iii) Execução/Monitoramento e (iv) Diagnóstico/Requisitos (AALST, 2013). No estágio de Projeto, os processos de negócios são identificados e representados através de um padrão de modelagem definido pelo BPM. No estágio de Configuração/Implementação, a infraestrutura necessária (sistemas e serviços) é criada para implementar e executar os processos de negócios modelados.

Figura 1: Ciclo de vida do BPM



Fonte: Adaptado de Aalst (2013).

No terceiro estágio, o processo é executado e monitorado para averiguar a necessidade de pequenos ajustes, que não acarretam no reprojeção do modelo. Já no último estágio do ciclo de vida do BPM, o de Diagnósticos/Requisitos, o processo é avaliado e, se houver necessidade, decorrente de novas demandas, mudanças em políticas internas ou devido a gargalos, um novo ciclo de vida que pode levar ao reprojeção do modelo existente (AALST, 2013).

Sendo assim, o BPM vem provendo métodos, técnicas e ferramentas para a melhora contínua dos processos hospitalares. E dessa forma, permitindo que as organizações sejam mais eficientes, assertivas e adaptáveis a mudanças do que aquelas com foco funcional, com abordagem de gerenciamento tradicional hierárquico (REICHERT, 2011).

2.2 Processos Hospitalares

Segundo Rebugue e Ferreira (2012), os processos hospitalares são considerados não triviais. Pois além de serem executados em um ambiente imprevisível e dinâmico, muitas vezes, eles envolvem profissionais de diferentes especialidades, que precisam colaborar entre si.

Em estudo realizado por Rojas et al. (2016), com o objetivo revisar a literatura sobre processos hospitalares e mineração de processos, é possível definir que os processos hospitalares podem ser classificados em dois tipos: os processos clínicos e os processos administrativos.

Em síntese, os processos clínicos estão diretamente ligados ao paciente e são executados de acordo com o seu quadro clínico. Eles compreendem a observação, o raciocínio clínico e a execução. Os processos clínicos dependem muito do conhecimento de profissionais da saúde para lidar com decisões específicas que devem ser tomadas de acordo com a evolução clínica de cada paciente. Já os processos administrativos, como por exemplo, a requisição de exames e o agendamento de consultas, são processos genéricos que, geralmente, dão apoio aos processos clínicos. Estes processos, que não estão ligados a uma situação específica como os processos clínicos, têm como objetivo coordenar o tratamento médico entre diferentes pessoas e unidades organizacionais (ROJAS et al., 2016).

De acordo com Rebugue e Ferreira (2012) e Lismont et al. (2016), o ambiente hospitalar e seus processos possuem características peculiares em relação ao seu:

- **Dinamismo:** processos hospitalares mudam constantemente devido a uma variedade de razões incluindo a introdução de novos procedimentos e tratamentos, desenvolvimento tecnológico ou a descoberta de novos medicamentos;
- **Complexidade:** devido ao processo de decisão médica, da quantidade de informações trocadas entre profissionais e a imprevisibilidade do quadro clínico dos pacientes;
- **Multidisciplinariedade:** organizações hospitalares contam um número crescente de departamentos especializados e especialidades médicas;
- **Cooperativismo:** processos hospitalares dependem muito das informações trocadas entre os seus participantes, além disso, muitos deles, têm o conhecimento necessário e a autonomia para definir a sua forma de trabalho;
- **Imprevisibilidade:** visto que cada paciente pode seguir um processo distinto durante a sua permanência em uma intuição de saúde;

- Variabilidade: uma vez que existe uma grande diversidade de atividades realizadas em um único processo clínico.

2.2.1 Processo de reembolso despesas médico-hospitalares

O processo de reembolso é um processo administrativo que tem como objetivo cobrar, pelos produtos e serviços utilizados durante o tratamento dos pacientes dentro e fora do hospital, dos próprios pacientes, das seguradoras de saúde ou então de programas financiados pelo governo (FERENC, 2013). Segundo Porter e Kaplan (2014), existem diversos métodos diferentes de reembolso como, o pagamento por serviço (*Fee-For-Service* - FFS), o per capita (*Capitation*), o por desempenho (*Bundled services*) e por pacote (*Diagnosis Related Groups* - DRG). Dentre estes, destaca-se o por serviço, pois o mesmo é o modelo mais empregado no nos EUA e em muitos países (PORTER; KAPLAN, 2014).

No Brasil, os métodos de reembolso mais presentes são o por serviço e o por pacote. O método por pacote é um modelo similar a Autorização de Internação Hospitalar (AIH), o qual é adotado pelo programa público de saúde SUS (Sistema Único de Saúde), que é oferecido pelo governo (GIOVANELLA et al., 2012). Já o modelo por serviço é o método mais utilizado pelas seguradoras de saúde privadas do país (CARDOSO et al., 2017).

O pagamento por serviço consiste em cobrar por cada procedimento, tratamento, serviço e insumo utilizado. Este método pode variar de país para país e até mesmo dentro do mesmo país. Geralmente ele envolve três etapas distintas: a codificação, o faturamento e o pagamento. Durante a etapa de codificação, que acontece após a alta do paciente, o médico responsável ou outro profissional habilitado, preenche o registro médico do paciente com os serviços prestados e os insumos utilizados durante o tratamento do mesmo. Após isso, este registro é encaminhado para outro profissional, o qual é responsável por traduzir os serviços e insumos do registro do paciente de acordo com um conjunto de códigos padronizado. Na próxima etapa, a de faturamento, um profissional cria uma fatura através dos códigos selecionados na etapa anterior. Por fim, na etapa de pagamento, a fatura é submetida para a fonte pagadora (seguradora pública ou privada). Contudo, a fatura pode ser recusada pela fonte pagadora, devido a inconsistências ou a divergências entre prestadora e seguradora sobre o tratamento dado ao paciente, o que implica na geração e submissão de uma nova fatura (LEE et al. 2016).

Mesmo sendo o método mais empregado, o pagamento por serviço, apresenta diversos problemas como, por exemplo: a remuneração não leva em conta a qualidade dos serviços prestados, pois a mesma é feita com base na quantidade de serviços prestados e insumos utilizados; o aumento do custo administrativo de certos tratamentos, como no caso do câncer, onde um grande volume de serviços e insumos são prestados e utilizados; pode influenciar as escolhas médicas, já que as seguradoras tendem a negar contas contendo tratamentos onerosos (PORTER; KAPLAN, 2014).

No método por pacote, as prestadoras de saúde recebem um pagamento único, baseado na média das contas históricas, por todos os serviços prestados de acordo com o diagnóstico do paciente. Apesar de ser mais simples que o pagamento por serviço, ele também possui problemas. Neste método as prestadoras assumem o risco por eventuais gastos acima da média, proporcionados pelas variabilidades inerentes ao processo de tratamento. Além disso, este método desconsidera o ciclo completo de cuidados médicos para o tratamento de doenças, o que exclui serviços como a reabilitação por exemplo. Outro problema desse método é realizar

pagamentos separados para as prestadoras e profissionais, o que pode descoordenar o serviço, já que os mesmos trabalham em conjunto (PORTER; KAPLAN, 2014).

De acordo com Lee et al. (2016) o processo de reembolso é de extrema importância para instituições, visto que o seu baixo desempenho pode acarretar em problemas como: o adiamento ou a interrupção de tratamentos médicos, devido a atrasos no pagamento das despesas por parte das seguradoras; perda da cobertura do seguro por parte do paciente, quando seguradora e prestadora discordam sobre o tratamento empregado; atrasos no pagamento de apólices de seguro para a prestadora, mediante a constatação de superfaturamento, subfaturamento, erros de codificação ou então quando a conta não estava de acordo com a regulamentação vigente; rompimento de contrato entre prestadora e seguradora, devido a pagamentos fraudulentos ou abusivos. Sendo assim, evitar problemas com o processo de reembolso é de extrema importância para os consumidores, as instituições e as seguradoras de saúde.

2.3 Mineração de Processos

A Mineração de Processos (MP), ou, em inglês, *Process Mining* foi desenvolvida em resposta às necessidades que as empresas têm em conhecer como os seus processos realmente funcionam. Ela descreve um conjunto de métodos e abordagens que combinam técnicas de Mineração de Dados (MD) com técnicas de modelagem e análises de processos de negócios para prover novos meios de análise de processos dos mais diversos domínios. Desse modo, a MP permite que as empresas diagnostiquem problemas existentes em seus processos tendo como base modelos que representam o seu real comportamento (AALST, 2016).

2.3.1 Tarefas da Mineração de Processos

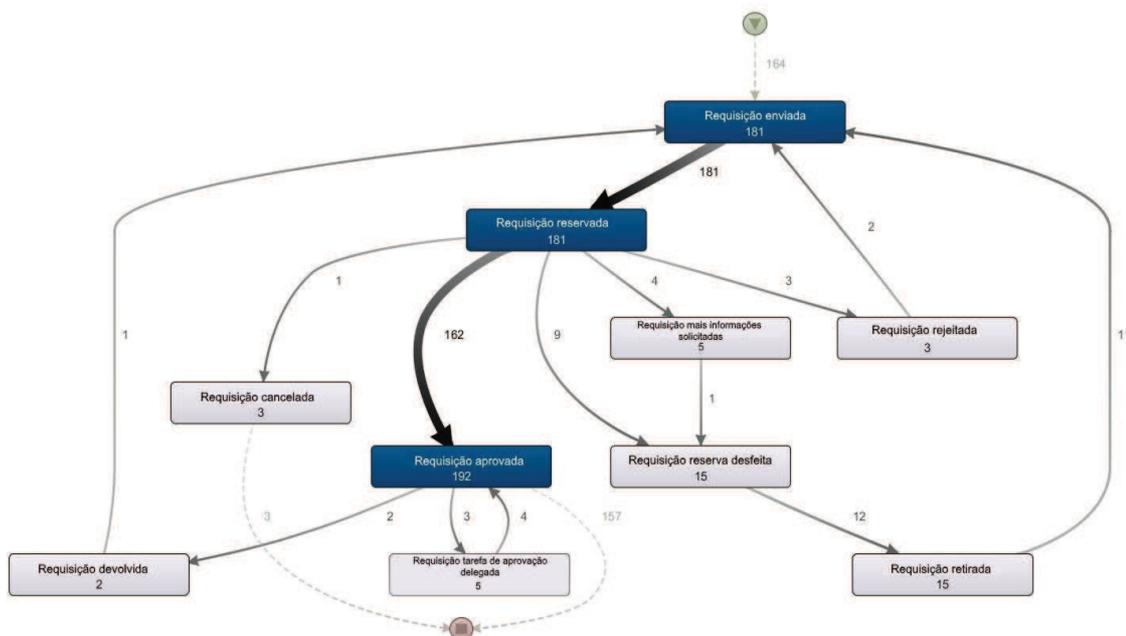
Aalst (2016) define que a MP pode ser dividida em três tarefas: a descoberta de processos, a verificação de conformidade de processos, e o aprimoramento de processos. A descoberta de processos tem como objetivo transformar um *log* de eventos num modelo de processo. Para isto, algoritmos de descoberta de processos recebem como entrada um *log* de eventos e retornam como saída um modelo de processo. Desse modo, pode-se gerar modelos que sejam uma representação precisa do comportamento observado no *log* de eventos.

Já a verificação de conformidade busca comparar um modelo existente com um modelo descoberto através de um *log* de eventos. Desta forma, a verificação de conformidade pode ser utilizada para averiguar se realmente os eventos de *log* registrados estão de acordo com o modelo existente ou vice e versa. Por fim, o aprimoramento do modelo tem o intuito de estender ou melhorar o modelo descoberto usando as informações contidas no *log* de eventos. Com isso, é possível gerar modelos com informações sobre desempenho, utilização de recursos, métricas de qualidade, entre outros.

Na Figura 2 é possível observar um modelo de processo descoberto e aprimorado gerado com o algoritmo Fuzzy Miner (GÜNTHER; AALST, 2007). O modelo apresentado refere-se a um processo de requisição de compras por parte de funcionários de uma empresa. Em resumo, este processo começa com o envio de uma requisição de compra para análise. Após o envio, a requisição é reservada e analisada. Caso ela apresente problemas, ela é enviada de volta para o solicitante, caso contrário, a mesma pode ser aprovada dando fim ao processo. No modelo apresentado é possível observar as atividades mais frequentes (nodos azuis), que neste caso são,

“Requisição enviada”, “Requisição reservada” e “Requisição aprovada”. Também é possível observar as relações de precedência mais frequentes (arestas mais espessas) formadas pelas atividades “Requisição enviada” e “Requisição reservada”, bem como as atividades “Requisição reservada” e “Requisição aprovada”.

Figura 2: Modelo de processo de requisição de compras



Fonte: Adaptado de Fluxicon (2015).

De acordo com Weske (2012), cada uma das atividades da MP pode ser associada a diferentes fases do ciclo de vida de BPM. Os modelos descobertos podem ser utilizados na etapa de (re)Projeto. Já a verificação de conformidade pode ser utilizada na etapa de Execução/Monitoramento. Além de que, modelos aprimorados podem ser utilizados na fase de Diagnóstico/Requisitos.

2.3.2 Log de Eventos

Os *logs* de eventos são considerados o ponto de partida para a MP, visto que eles contêm informações históricas sobre a execução de processos. Contudo, segundo Aalst (2016), para os mesmos serem utilizados é necessário que os *logs* de eventos conttenham informações adequadas sobre as atividades executadas no processo, uma vez que elas serão utilizadas por algoritmos durante a geração de modelos. Sendo assim, as seguintes características devem ser consideradas durante a extração de informações de *log*:

- Um *log* de eventos deve ser constituído por *cases*. Estes *cases*, por sua vez, necessitam ser formados por uma sequência de eventos (atividades ou etapas) pré ou não definida. Sendo que, estes eventos, devem obrigatoriamente estar relacionados a um único *caso*;
- Eventos podem ter atributos como: atividade, tempo e recurso. O atributo atividade descreve a ação realizada. Já o atributo tempo, registra data e/ou hora do evento. Por fim, o atributo recurso descreve o responsável pela execução;

- Eventos de um mesmo caso devem estar ordenados de maneira crescente, por exemplo, por ordem de ocorrência. Pois, caso contrário, podem ocorrer mudanças entre as dependências causais dos eventos, alterando os resultados dos algoritmos.

A Tabela 1 traz um exemplo de um *log* de eventos, relacionado a um processo de vendas de produtos, que apresenta essas características. O processo, apresentado neste exemplo, contém três *cases* (*Case 1, 2 e 3*) e cada *case* contém conjuntos de eventos que, por sua vez, são compostos por um identificador e seus atributos (*Timestamp*, *Atividade*, *Recurso* e *Custo*). Também é possível observar que os eventos estão ordenados de forma crescente por meio do atributo *Timestamp*, que registra o tempo da ocorrência de cada evento.

Tabela 1: Exemplo Log de Eventos

Case	Evento				
	Identificador	Timestamp	Atividade	Recurso	Custo
1	111	03-03-2015:12.03	Registro de mercadoria	João	60
	112	04-03-2015:13.03	Negociação com cliente	Pedro	160
	113	05-03-2015:14.03	Efetivar venda	Maria	100
2	211	03-04-2015:12.03	Negociação com cliente	Pedro	160
	212	03-05-2015:13.03	Efetivar venda	Maria	100
	213	03-06-2015:14.03	Registro de devolução	José	200
	214	05-06-2015:14.03	Efetivação de devolução	Pedro	120
	215	05-07-2015:14.03	Recompensação de cliente	Maria	50
3	311	03-04-2015:12.03	Registro de mercadoria	João	60
	312	04-04-2015:13.03	Negociação com cliente	Pedro	160
	313	05-04-2015:14.03	Efetivar venda	Maria	100
	314	06-04-2015:14.03	Pedido de troca de produto	José	100
	315	07-04-2015:15.03	Efetivação de troca	Pedro	160

Fonte: Adaptado de Aalst (2016).

Mesmo que o *log* de eventos possua as características necessárias para a extração de informações, ele pode conter problemas relacionados a sua qualidade. Segundo Bose, Mans e Aalst (2013), os problemas de qualidade podem ser agrupados em quatro categorias distintas:

- Dados incompletos: corresponde ao cenário onde diferentes tipos de informações estão faltando. Por exemplo, atividades que são executadas, mas não são constam no *log* de eventos;
- Dados incorretos: cenário onde apesar de que os dados estão presentes no *log* de eventos, os mesmos podem estar incorretos. Nestes casos, por exemplo, as informações registradas não correspondem a data real de execução;
- Dados imprecisos: as entradas registradas são muito genéricas e podem conduzir a uma perda de precisão. Por exemplo, a ordenação de eventos poderia ser prejudica

nos casos em que o momento da execução for registrado em dias, já que mais de um evento poderia acontecer no mesmo dia;

- d) Dados irrelevantes: cenário onde é necessária uma filtragem e/ou transformação (abstração ou agrupamento) de registros que contenham eventos irrelevantes para uma determinada análise, por exemplo.

2.3.3 Fuzzy Miner

Fuzzy Miner é um algoritmo de descoberta de modelos proposto por Günther e Aalst (2007), que tem como principal característica a simplificação de modelos de processos através de agregações e abstrações. Tais simplificações, permitem a análise de modelos do tipo “spagetti”, ou seja, modelos de processos complexos, que apresentam um comportamento variável, que são difíceis de serem interpretados e/ou analisados.

Para simplificar modelos complexos, o algoritmo Fuzzy Miner usa como critério de decisão duas métricas, a significância e a correlação. A significância pode ser determinada tanto para os eventos (atividades) quanto para relações de precedência (arestas). Com isso, é possível construir modelos com base no nível de interesse que há nos eventos ou nas relações de precedência. Um exemplo de unidade de medida para a significância é a frequência, onde os eventos ou relações de precedência observados com maior frequência são considerados mais significantes.

Já a correlação só é relevante para as relações precedência, ela mede o quanto são relacionados dois eventos seguidos. Eventos são considerados correlacionados quando compartilham uma grande quantidade de dados, ou têm semelhança expressa em seus nomes (por exemplo, "verificar a aprovação do pedido do cliente " e "aprovar o pedido do cliente").

De acordo com Günther e Aalst (2007), através da utilização das duas métricas é possível obter um modelo de processo simplificado onde: o comportamento altamente significativo é preservado; o comportamento menos significativo, mas altamente correlacionado é agregado, ou seja, escondido em agrupamentos; o comportamento menos significativo e menos correlacionado é removido.

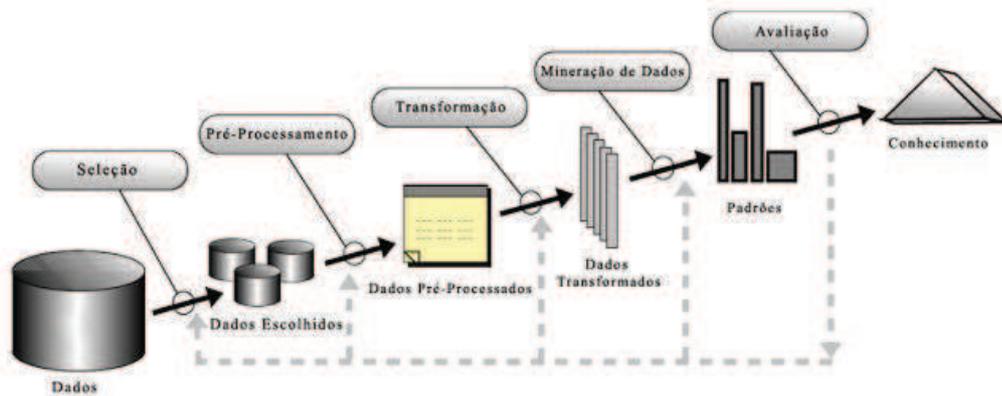
Em resumo, a abordagem proposta por Günther e Aalst (2007), pode reduzir e concentrar o comportamento exibido no modelo do processo, através dos conceitos de agregação e abstração. Isso torna o algoritmo Fuzzy Miner adequado para minerar *logs* de eventos com muitos ruídos, ou com um comportamento muito variável. Contudo, com este algoritmo, não é possível gerar modelos de processos que contenham decisões como, por exemplo, se atividade X, então a próxima atividade é Y, senão a próxima atividade é Z.

2.4 Descoberta de Conhecimento em Bases de Dados

A Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases - KDD*) (FAYYAD et al., 1996) preocupa-se em desenvolver métodos e técnicas que possibilitam inferir novos conhecimentos a partir de informações extraídas de bases de dados. Isto é, ela tem como objetivo mapear grandes volumes de dados em uma forma mais compacta (um pequeno relatório), abstrata (uma descrição aproximada dos dados), ou então, em algo que gere novos conhecimentos (um modelo preditivo que possa estimar casos futuros).

Segundo Fayyad et al. (1996), o KDD é um processo não trivial de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em banco de dados. Esse processo é considerado interativo e iterativo, pois envolve decisões feitas pelo usuário e ciclos de repetição de etapas do processo.

Figura 3: Etapas do processo de descoberta de conhecimento



Fonte: Adaptado de Fayyad et al. (1996).

Na Figura 3 são apresentadas as cinco etapas do processo de KDD, especificadas da seguinte forma: na etapa de **Seleção** busca-se ter um entendimento inicial do domínio da aplicação e dos objetivos desejados para então selecionar o conjunto de dados pertencente ao domínio, que contém todas as possíveis variáveis (também conhecido por características ou atributos) e registros (também denominados de casos ou amostras) que farão parte da análise (FAYYAD et al., 1996).

A etapa de **Pré-processamento** visa melhorar a qualidade dos dados através do tratamento ou remoção das amostras com dados ausentes e/ou ruidosos e também a partir da seleção de atributos. O tratamento dos dados ausentes pode ser feito preenchendo-se os valores ausentes pela média aritmética das variáveis, uma constante ou então manualmente. Já os dados ruidosos, que são erros aleatórios ou uma variância anormal dos dados, podem ser tratados usando-se técnicas como regressão ou agrupamentos.

Já a seleção de atributos busca escolher um subconjunto de atributos ou criar outros que possam substituir o conjunto original a fim de reduzir a sua dimensão. Tal redução, pode diminuir a complexidade do conjunto de dados e assim o tempo necessário para extração de algum conhecimento do mesmo. Além disso, pode-se remover atributos desnecessários, os quais podem provocar algum ruído no resultado final. Portanto, o pré-processamento dos dados é fundamental, pois a qualidade dos mesmos influenciará diretamente na eficiência dos algoritmos de mineração (HAN; KAMBER; PEI, 2011).

Na etapa de **Transformação**, o conjunto de dados é preparado de acordo com as suas características e das peculiaridades dos algoritmos que serão utilizados na etapa de mineração. Nesta etapa, são empregadas técnicas como: **(a) agregação**: onde são gerados totalizadores levando em conta determinados atributos ou instâncias; **(b) generalização**: na qual dados são substituídos por conceitos de mais alto nível; **(c) normalização**: onde atributos são escalonados para uma faixa específica como, por exemplo de -1,0 a 1,0 ou de 0,0 a 1,0; **(d) construção de atributos**: no qual atributos são construídos a partir de informações pré-existent (HAN; KAMBER; PEI, 2011).

A etapa **Mineração de Dados**, que segundo Fayyad et al. (1996) consiste na aplicação de algoritmos específicos, sob alguma limitação aceitável de eficiência computacional, para

produzir uma lista particular de padrões sobre os dados. Estes algoritmos compõem técnicas que são escolhidas e aplicadas conforme o objetivo e a tarefa desejada como, por exemplo, a geração de regras de associação, agrupamentos, classificação, entre outros. As técnicas de interesse do presente trabalho são detalhadas nas seções seguintes (WITTEN; FRANK; HALL, 2011).

Por fim, a etapa **Avaliação** consiste na avaliação e interpretação dos resultados obtidos durante a etapa anterior. A avaliação dos resultados, pode ser feita em conjunto com especialistas no domínio e/ou então através de um conjunto métricas específicas. Desse modo, é possível mensurar a qualidade da técnica de MD empregada, bem como a do processo antecessor a ela (HAN; KAMBER; PEI, 2011).

2.5 Mineração de Regras de Associação

A mineração de regras de associação tem como premissa básica descobrir relações e ou associações entre itens pertencentes a uma transação. Possibilitando assim, a tomada de decisões através das novas regras e associações descobertas (ZAKI; MEIRA; 2014).

A partir de um conjunto de transações, como pode observado na Tabela 2, onde registram-se os itens adquiridos por clientes de um supermercado, poderia se gerar a seguinte regra de associação: $\{leite, pão\} \rightarrow \{ovos\}$, a qual demonstra que, com um determinado grau de certeza, clientes que adquirem leite e pão também podem vir a adquirir ovos (HAN; KAMBER; PEI, 2011).

Tabela 2: Conjunto de transações

Transação	Carne	Leite	Pão	Ovos	Vegetais
1	1	1	1	1	0
2	1	0	0	0	1
3	1	1	1	0	1
4	0	1	1	1	0
5	1	0	1	0	0

Fonte: Adaptado de Lacose (2014)

De acordo com Agrawal, Imieliński e Swami (1993), a tarefa da descoberta de regras de associação pode ser decomposta em duas partes: encontrar o conjunto de itens frequentes e gerar as regras de associação a partir deles.

Os conjuntos de itens frequentes são determinados a partir das transações que contenham um limite mínimo de suporte pré-estabelecido. Sendo que, o suporte de uma regra $X \Rightarrow Y$, onde X e Y são conjuntos de itens, é dado pela seguinte fórmula:

$$\text{Suporte}(X \Rightarrow Y) = P(XY) \quad (2.1)$$

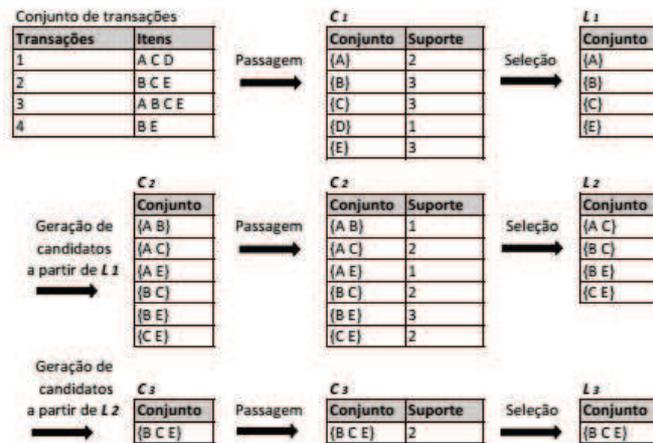
A geração das regras de associação é feita a partir da seleção dos conjuntos de itens frequentes. A seleção das regras é feita conforme um grau de confiança mínimo, o qual é definido pela seguinte equação:

$$\text{Confiança}(X \Rightarrow Y) = \frac{P(XY)}{P(X)} \quad (2.2)$$

O suporte (equação 2.1) pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto X quanto Y , enquanto que a confiança (equação 2.2) é a probabilidade de que uma transação satisfaça Y , dado que ela satisfaz X (ZAKI; MEIRA; 2014).

Um dos algoritmos mais conhecidos para a mineração de regras de associação é denominado Apriori. Este algoritmo, proposto por Agrawal e Swami (1994), emprega múltiplas passagens sobre um conjunto de transações. Na primeira passagem, o suporte de cada item é calculado e os itens com suporte maior que o suporte mínimo, denominados de itens frequentes (L_k), são selecionados. Em cada uma das passagens subsequentes, dado que k é o número da passagem, os itens frequentes da passagem anterior ($k-1$) são agrupados em conjuntos de k itens denominados de itens candidatos (C_k). Caso o suporte dos itens candidatos for maior que o suporte mínimo, estes são considerados frequentes. O processo continua até que o conjunto de itens frequentes seja um conjunto vazio, como pode ser observado na Figura 4. Por fim, as regras são geradas a partir dos conjuntos de itens frequentes que possuam um grau de confiança maior que o mínimo preestabelecido.

Figura 4: Geração de itens frequentes e candidatos com suporte mínimo igual a 2



Fonte: Adaptado de Han, Kamber e Pei (2011).

Muitas vezes, o algoritmo Apriori pode gerar um número grande de regras sem relevância, mesmo ajustando-se os valores do suporte e a confiança. Nestes casos, é possível utilizar a métrica denominada de *lift*, apresentada na equação (2.3), a qual pode mensurar o grau de independência do suporte de $X \Rightarrow Y$ (LAROSE, 2014).

$$lift(X \Rightarrow Y) = \frac{P(XY)}{P(X) \cdot P(Y)} \quad (2.3)$$

Por fim, é possível afirmar que para uma regra ser interessante, ela precisa ser de fácil compreensão, ser válida quando aplicada a novos dados ou dados de testes, ser potencialmente útil e, por fim, ser nova, ou seja, regras que já são conhecidas não são interessantes (HAN; KAMBER; PEI, 2011).

2.6 Classificação de Dados

A classificação de dados visa identificar a qual classe uma determinada amostra pertence. Esta tarefa é baseada no aprendizado supervisionado, no qual um modelo é criado a partir de um conjunto de dados, onde as classes das amostras são conhecidas, com o intuito de

“aprender” a como classificar uma amostra desconhecida. Por exemplo, clientes de um banco são categorizados como sendo: bons ou maus pagadores. Então, a partir das amostras analisadas o modelo é capaz de produzir uma resposta, como uma margem de erro, em qual categoria um novo cliente se encaixaria (HAN; KAMBER; PEI, 2011).

Sendo assim, a classificação de dados pode ser dividida em duas etapas:

- Na primeira etapa, amostras, onde suas classes são conhecidas, são apresentadas para um algoritmo, para que o mesmo aprenda e identifique padrões. Esta etapa é denominada de aprendizagem ou treinamento;
- Na segunda etapa, novas amostras de dados são apresentadas ao modelo construído na etapa anterior e então, uma saída é obtida do modelo, representado as classes definidas por ele para as amostras apresentadas.

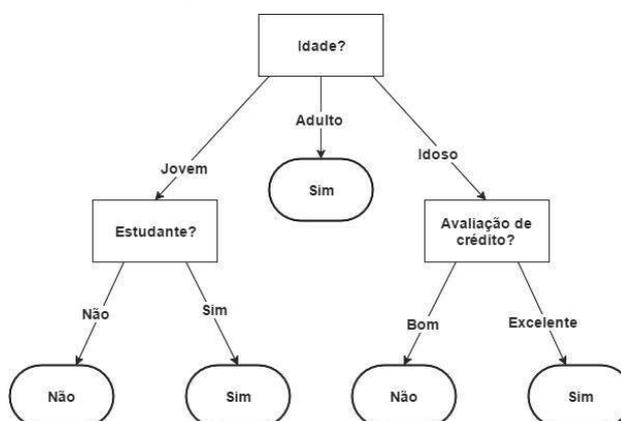
Nas subseções seguintes são apresentadas as Árvores de Decisão, um dos métodos mais populares de se construir classificadores dentre vários métodos existentes pois, eles podem gerar regras de fácil entendimento, manusear dados de altas dimensões e por serem fáceis de construir e avaliar. Assim como, serão apresentadas técnicas utilizadas para a interpretação e avaliação de classificadores.

2.6.1 Árvores de Decisão

As Árvores de Decisão são estruturas hierárquicas composta por nós, arestas e folhas. Onde, cada nó representa um teste de um atributo; cada aresta uma saída do teste e cada folha uma classe. A classificação de uma amostra se dá percorrendo esta árvore testando os atributos da amostra de entrada contra as condições dos nodos e prosseguindo pelo ramo correto até chegar na folha que indica a classe da amostra.

A Figura 5, apresentada a seguir, mostra uma Árvore de Decisão capaz de determinar se um dado cliente irá ou não adquirir um computador. No diagrama, os nodos são representados por retângulos e as folhas por elipses. Na árvore, é possível observar que, a idade é o atributo mais representativo e que todos os adultos, independentemente de serem estudantes ou terem excelentes avaliações de crédito, tem uma tendência maior de adquirir um computador.

Figura 5: Árvore de Decisão



Fonte: Adaptado de Han, Kamber e Pei (2012).

A construção de uma Árvore de Decisão, como a apresentada anteriormente, é realizada por recursivas divisões do conjunto de dados, durante o processo de construção. Estas divisões, podem ser realizadas a partir de critérios como o ganho de informação, a razão do ganho, o índice gini, entre outros (WITTEN; FRANK; HALL, 2011).

Porém, estas divisões recursivas, podem gerar nodos que reflitam anomalias, como ruídos e/ou *outliers*², que podem estar presentes no conjunto de dados de treinamento. Levando, desta forma, a geração de árvores com problemas de *overfitting*, ou seja, modelos que não são genéricos e que acabam apresentando uma taxa de acerto baixa para novas amostras (HAN; KAMBER; PEI, 2011).

O problema de *overfitting*, geralmente, é resolvido com o emprego de métodos de pré-poda e pós-poda aplicados respectivamente durante ou após a construção da árvore. Estes métodos utilizam medidas estatísticas a fim de remover as arestas com menor relevância. Dessa forma, as árvores podadas, tendem a ser menores e menos complexas e conseqüentemente mais fáceis de se compreender, elas também podem ser mais rápidas e eficientes, em termos de classificação, se comparadas às árvores não podadas (WITTEN; FRANK; HALL, 2011).

Existem vários algoritmos que permitem a construção de Árvores de Decisão. Porém, é muito difícil de se determinar qual é o melhor, pois dependendo do problema, um algoritmo pode ser mais eficiente que outro. Entre estes algoritmos destacam-se o CART (*Classification and Regression Trees*) (BREEMAN et al. 1984) e o C4.5 (QUILAN, 1993).

O algoritmo CART, tem como característica principal, construir Árvores de Decisão e/ou regressão através do particionamento recursivo binário do conjunto de dados. No CART, cada folha da árvore é relacionada a uma classe, no caso das Árvores de Decisão, ou então, a um valor contínuo no caso das Árvores de Regressão. Outra característica deste algoritmo é a habilidade de se trabalhar com atributos numéricos ou nominais sem a necessidade de transformá-los (LAROSE, 2014).

Já o algoritmo C4.5, sucessor do ID3 (QUILAN, 1986), proposto pelo mesmo autor, adota uma estratégia gulosa para a geração da árvore, a qual é baseada no ganho de informação ou na redução da entropia para particionar o conjunto de dados em dois ou mais conjuntos. Neste algoritmo, os dados numéricos precisam ser transformados em nominais antes de serem aplicados no classificador. Por fim, este algoritmo apresenta um método de pós-poda, que visa transformar em nós folha os ramos que não apresentam nenhum ganho significativo (LAROSE, 2014).

2.6.2 *Random Forest*

Random Forest consiste em um conjunto de Árvores de Decisão ou Árvores de Regressão, onde cada árvore é construída a partir de um conjunto de dados diferente e um número aleatório de variáveis. Dessa forma, busca-se reduzir a variância e a correlação entre as árvores, e conseqüentemente reduzindo as chances de *overfitting*. (BREIMAN, 2001).

Os conjuntos de dados utilizados para a construção das árvores são gerados através de amostragem aleatória com reposição (*bootstrap*) do conjunto de dados original, sendo que cada conjunto gerado tem o mesmo tamanho do conjunto original. Além disso, para cada conjunto

² *Outliers*: amostras que não estão em conformidade com o comportamento geral dos dados (HAN; KAMBER; PEI, 2011)

de treinamento, cerca de 1/3 das amostras não são utilizadas. Tais amostras, denominadas de *out-of-bag*, são utilizadas no processo de construção para determinar a taxa de erro de predição de cada árvore, bem como para medir a força de predição de cada variável (BREIMAN, 2001).

A taxa de erro de predição é determinada através da apresentação das amostras *out-of-bag* para a árvore recém construída. A força de predição de cada variável é determinada através da redução média da taxa de erro de predição que, por sua vez, é obtida permutando os valores de cada variável e computando uma nova taxa de erro e , finalmente, comparando a nova taxa de erro com a taxa de erro inicial (HASTIE et al., 2009).

O algoritmo apresentado no quadro a seguir descreve de forma simplificada a construção de um modelo com o algoritmo *Random Forest*.

Quadro 1: Algoritmo Random Forest.

Entrada: T - Conjunto de treinamento original

N - Número de árvores

Método:

1. **para cada** $arvore_i$ **N faça**
2. Gere um novo conjunto de treinamento t_i utilizando amostragem por *bootstrap* do conjunto de treinamento original T;
3. Construa a $arvore_i$, sem poda, utilizando o conjunto de treinamento t_i ;
4. Para cada nó da $arvore_i$, escolha aleatoriamente F atributos e determine a melhor divisão do nó utilizando os F atributos e o conjunto t_i ;
5. Apresente o conjunto *out-of-bag* para a $arvore_i$;
6. **fim**

Fonte: Elaborado pelo autor com base em (BREIMAN, 2001).

Quando a técnica *Random Forest* é usada para a classificação, as predições são feitas através de uma votação entre as árvores do conjunto, onde a classe de uma amostra é determinada a partir do voto majoritário. Por outro lado, quando usada para regressão, as previsões são feitas através do cálculo da média da previsão de cada árvore do conjunto (HASTIE et al., 2009).

3 TRABALHOS RELACIONADOS

Este capítulo objetiva situar a proposta perante a literatura. Mediante isso, são apresentados estudos que visam aprimorar processos hospitalares. Na seção 3.1 a metodologia de pesquisa utilizada para a seleção dos estudos é apresentada. Já nas seções 3.2 e 3.3 são apresentados cinco estudos que demonstram como as técnicas da MP e também da MD podem ser aplicadas para a análise de processos hospitalares.

Além desses estudos, na seção 3.3 são apresentados quatro estudos que mostram como a MP pode ser empregada em conjunto da MD para obtenção de modelos de processo mais compreensíveis e também para detectar problemas em processos. Por fim, a seção 3.4 traz considerações e comentários sobre os principais estudos, posicionando esta proposta frente à literatura.

3.1 Metodologia de pesquisa bibliográfica

A pesquisa bibliográfica foi realizada utilizando o portal Google Scholar, em função do mesmo indexar diversos periódicos como, por exemplo, Elsevier, ACM Digital Library, IEEE-Xplorer e MEDLINE. A pesquisa levou em consideração trabalhos com temas relacionados ao aprimoramento de processos hospitalares através de técnicas da MD, da MP ou então com ambas. Vale ressaltar que, durante a pesquisa, foi dada uma maior ênfase para trabalhos relacionados a processos hospitalares administrativos já que, o presente estudo se enquadra neste tópico. Como critério de busca, foram utilizadas palavras chaves como, process mining, data mining, machine learning, healthcare, healthcare administration, healthcare management, healthcare operations, reimbursement, insurance, audit, billing, decision making, business process management. As palavras chaves foram combinadas com operadores lógicos (AND e/ou OR) como, por exemplo, process mining AND (healthcare management OR reimbursement) com o intuito de segregar os trabalhos em três grupos distintos conforme a abordagem empregada pelo estudo (Mineração de Dados, Mineração de Processos ou então a combinação de ambas). As próximas seções apresentam os três grupos de estudos selecionados.

3.2 Análise de processos hospitalares através de MP

O trabalho de Yoo et al. (2016) busca averiguar o impacto das alterações estruturais, decorrentes da construção de uma nova edificação e da realocação de unidades de tratamento, nos processos clínicos e administrativos de um hospital especializado.

Para isso, foram coletados dados de atendimentos feitos nas unidades de tratamento contra o câncer e de neurologia realizados respectivamente um mês antes e um mês após as alterações estruturais. No total, foram coletados dados de 7.126 atendimentos, sendo que 2.337 são anteriores as alterações e 4.789 se referem após a construção.

A partir desses dados, foi criado um *log* de eventos contendo 15 atividades distintas. Basicamente, cada *log* é composto de um identificador do caso, a data de término da atividade, o nome da atividade, o recurso envolvido e o código do departamento, com exceção das atividades tratamento e impressão da prescrição. Algumas atividades continham informações extras como por exemplo, o tipo do exame e o tipo do paciente no caso das atividades

relacionadas respectivamente com os exames e as consultas. Eventos repetidos que ocorreram em sequência e dentro de três minutos ou menos foram removidas do *log*. Além disso, a atividade tratamento foi considerada como a última etapa, já que a mesma só possuía a data de início.

Com isso, os autores usaram algoritmos e ferramentas de estatística e de MP como Heuristic Miner (WEIJTERS; AALST; MEDEIROS, 2006) para a descoberta do modelo, Dotted Chart Analysis (SONG; AALST, 2007) para mostrar a distribuição dos eventos em função da frequência e/ou tempo e o software ProM (DONGEN et al., 2005) para análise do desempenho dos processos, bem como testes de significância estatística a fim de avaliar se houve uma mudança significativa nos processos. A comparação dos processos foi realizada em função de três métricas: (i) os tempos de espera, do processo como um todo e também da realização de consultas e exames; (ii) os tempos de execução, com ênfase nos processos que apresentam similaridades e distorções em função do tempo total de execução; (iii) e os padrões dos fluxos mais frequentes do processo.

Desta maneira, os autores constataram que, em geral, a construção de um novo prédio e a realocação de unidades de tratamento contribuiu para a melhoria no desempenho do processo. Mesmo que, algumas atividades administrativas como o recebimento de prescrições e o pagamento, passaram a demandar mais tempo após as mudanças estruturais.

O trabalho de Partington et al. (2015), realiza uma análise comparativa, entre quatro hospitais distintos, objetivando identificar as diferenças e semelhanças durante a prestação de serviços de cuidados médicos em pacientes diagnosticados preliminarmente com dores no peito.

Para realizar a comparação, os autores criaram um *log* de eventos com a união do conjunto de dados pertencentes a quatro hospitais. Os atributos dos *logs* foram selecionados de acordo com protocolos de tratamento e também com base na literatura sobre processos cardíacos. Alguns atributos como o identificador do hospital e categoria da triagem foram incorporados ao *log* a fim de auxiliar a visualização da trajetória do paciente através do hospital.

Com o intuito de minimizar a presença de casos incompletos e variáveis inconsistentes, os *logs* foram filtrados em função de seus atributos e os tempos de início e fim de cada caso. Além disso, um subconjunto dos *logs* foi selecionado com base nos pacientes que foram diagnosticados com dores no peito no departamento de emergência. Sendo assim, dos 27.773 casos pertencentes aos quatro hospitais, 9.713 casos foram selecionados após o pré-processamento dos *logs*.

Para investigar os elementos que impactam a eficiência da prestação de cuidados médicos entre os quatro hospitais. Os autores avaliaram métricas como a proporção de pacientes internados, o tempo entre a apresentação e a internação, a frequência de procedimentos realizados e o tempo total de permanência dos pacientes através do uso de técnicas de MP.

Inicialmente os autores, aplicaram o algoritmo Heuristic Miner (WEIJTERS; AALST; MEDEIROS, 2006) para a descoberta de um modelo do processo e *plugins* da ferramenta ProM (DONGEN et al., 2005) para a análise de desempenho. Após esta etapa, um conjunto de atividades comuns aos quatro hospitais foi abstraída do modelo inicial e representada em quatro modelos BPMN (WHITE, 2004), um para cada hospital. Estes modelos foram aprimorados com informações de desempenho e com as métricas mencionadas para a realização da análise comparativa. Além disso, os autores criaram modelos utilizando o algoritmo Fuzzy Miner (GÜNTHER; AALST, 2007) para visualizar os caminhos mais frequentes percorridos pelos pacientes entre as quatro instituições.

Dessa forma, foi possível identificar particularidades nas práticas utilizadas entre as quatro instituições para o tratamento de pacientes diagnosticados com dores no peito. Embora algumas dificuldades tenham sido encontradas pelo autor durante a comparação visual dos modelos, devido à ausência de ferramentas apropriadas.

3.3 Uso de abordagens da MD para a análise de processos hospitalares

O estudo realizado por Tsumoto et al. (2014) tem como objetivo analisar aspectos ligados ao gerenciamento de instituições hospitalares como a qualidade dos serviços prestados aos pacientes, a qualidade das práticas médicas adotadas e a eficácia das práticas administrativas adotadas para o gerenciamento de instituições como um todo.

Para tanto, os autores utilizaram dados em formatos de “ordens”, que representam as informações, mensagens e/ou documentos trocados durante a execução das atividades dos *workflows* hospitalares.

Através desses dados, os autores obtiveram uma visão geral sobre a execução dos *workflows* hospitalares com o uso da estatística descritiva. Além disso, técnicas da MD, como Árvores de Decisão (QUILAN, 1993) e Trajectory Mining (ZHENG, 2015) foram aplicadas para identificar as características dos processos clínicos e os padrões temporais dos fluxos dos *workflows*. Dessa maneira, os autores obtiveram uma visão geral sobre o comportamento dos processos hospitalares.

O trabalho de Spruit et al. (2014) explorou técnicas de descoberta do conhecimento em base de dados para melhorar a qualidade dos serviços médicos e também para diminuir os gastos financeiros de tratamentos clínicos prolongados. Para tal, os autores em conjunto com gerentes, médicos e outros profissionais de uma instituição de saúde holandesa identificaram uma série de pontos a serem analisados.

Tais pontos incluem: (i) a ocorrência de incidentes em setores específicos e os danos físicos e mentais causados aos pacientes; (ii) identificação de padrões em falhas cometidas nos processos clínicos; (iii) a relação entre os incidentes e falhas; (iv) o tempo de permanência dos pacientes na instituição; (v) o aumento da demanda de atendimentos nas unidades de tratamento intensivo. Os pontos (i), (iii), (iv) foram analisados com o uso de ferramentas da estatística descritiva, no ponto (ii) o algoritmo Apriori (AGRAWAL; SWAMI, 1994) foi aplicado sobre o conjunto de dados contendo falhas nos processos e na investigação do ponto (v) foram utilizados modelos preditivos estatísticos. Como resultado da análise, os autores identificaram problemas e oportunidades para melhorar o gerenciamento da instituição estudada.

Um outro estudo relevante é o de Huang et al. (2011), o qual propõe um algoritmo baseado no Apriori (AGRAWAL; SWAMI, 1994) para a geração de regras associativas com o intuito de servir de apoio à alocação de recursos humanos. Em resumo, o algoritmo proposto pelos autores consiste em gerar regras associativas tendo como base um *log* de eventos. O grande diferencial deste algoritmo reside em considerar a ordem cronológica da execução das atividades do processo, o que segundo os autores, possibilitaria gerar um conjunto de regras menor e mais representativo.

Os autores, usaram como estudo de caso um processo de radiologia, que abrange atividades clínicas e administrativas, para avaliar a qualidade das regras geradas, bem como o desempenho do algoritmo proposto em relação ao Apriori. Em geral, o algoritmo proposto

obteve melhores resultados do que o Apriori em relação ao número de regras geradas e ao tempo de processamento.

3.4 Integração de técnicas da MP e MD

O trabalho elaborado por Delias et al. (2015) apresenta uma abordagem para representar modelos de processos complexos de uma forma compacta e compressível, que desconsidera os fluxos infrequentes.

Para isso, os autores utilizaram a técnica de agrupamento denominada de Spectral Clustering (LUXBURG, 2007) para gerar subprocessos a partir modelos criados com o algoritmo Fuzzy Miner (GÜNTHER; AALST, 2007). A definição dos agrupamentos foi realizada através de uma métrica de similaridade, proposta pelos autores, que leva em consideração as atividades realizadas e as transições do processo. A escolha do Fuzzy Miner deu-se devido a sua capacidade de gerar modelos que apresentem todos os fluxos do processo e não somente os mais frequentes, como no caso dos demais algoritmos usados para a descoberta de modelos.

Com o intuito de avaliar a abordagem proposta, os autores compararam a sua abordagem com outras três outras através de métricas como: a complexidade do fluxo de controle, que considera a quantidade de decisões presentes nos fluxos; a estruturação, que atribui pesos às estruturas básicas de um processo como (sequências, escolhas e iterações); e a densidade, ou seja, o número de interconexões existentes no modelo. O resultado da comparação mostrou que, em geral, o método proposto se saiu melhor que os demais pois, somente na métrica estruturação os autores não obtiveram o melhor resultado.

Outro estudo relevante, que integra técnicas da MP e MD, é o trabalho apresentado por Suriadi et al. (2013). Neste estudo, os autores buscaram identificar os motivos que levam ao atraso na conclusão de aprovação de contas médicas e hospitalares consideradas (simples e demoradas) por parte de uma seguradora de saúde.

Para atingir este objetivo, primeiramente os autores em conjunto com a seguradora criaram um *log* de eventos referente ao processo de aprovação. Após esta etapa, o *log* de eventos foi subdividido manualmente de acordo com quatro classes distintas baseadas em dois atributos: complexidade (simples e complexo) e duração (rápido e demorado). A partir disso, modelos de processos foram criados e enriquecidos com a frequência de execução das atividades, para possibilitar a identificação de diferenças e similaridade entre os processos das classes (simples e rápido) e (simples e demorado). Com o resultado da análise, os autores, aplicaram técnicas de classificação como Árvores de Decisão (QUILAN, 1993) sobre o *log* de eventos, com o objetivo de gerar regras passíveis de explicar o porquê do atraso da aprovação das contas. Os resultados em relação à identificação de possíveis causas com a aplicação de Árvores de Decisão foram positivos segundo os autores, apesar das dificuldades na identificação de variáveis independentes.

O trabalho de Lismont et al. (2016) apresenta uma metodologia para a análise de processos que leva em consideração as peculiaridades do âmbito hospitalar. A metodologia foi baseada em técnicas da MP porém, a mesma emprega algumas técnicas da MD para resolver alguns problemas comuns em processos do âmbito hospitalar.

Em resumo, a metodologia é composta por seis etapas distintas: (i) a preparação do *log* de eventos, onde o *log* de evento é criado a partir de uma ou mais fontes; (ii) a inspeção do *log*,

na qual realiza-se uma análise inicial dos dados através de estatística descritiva; (iii) abstração e seleção, na qual detalhes irrelevantes são removidos e um subconjunto dos dados é selecionado; (iv) agrupamento, onde são criados *sub-logs* de eventos de acordo com métricas pré-definidas; (v) MP, que consiste na aplicação de técnicas da Mineração de Processos para a descoberta, conformidade e enriquecimento de modelos de processos; e (vi) validação, onde os resultados são apresentados e validados por especialistas. A aplicabilidade da metodologia foi demonstrada através da análise do processo de tratamento de pacientes com diabetes do tipo 2. Nesta análise, os autores demonstraram a efetividade da metodologia apresentada perante as peculiaridades dos processos hospitalares.

Por fim, o estudo de Djenouri et al. (2018), apresenta uma abordagem para extração de conhecimento através da extração de conjuntos de itens frequentes de *logs* de eventos de processos de negócio.

Em síntese, a abordagem envolve três etapas distintas: pré-processamento, mineração do conjunto de itens frequentes e poda. O primeiro passo envolve a criação de atributos categóricos com base nos atributos existentes, bem como a criação de várias bases de transações, uma em função do *log* de eventos como um todo e outras em função de cada atributo presente no *log* de eventos para fornecer múltiplas perspectivas sobre os dados. O segundo passo consiste na extração conjuntos de itens frequentes das diferentes bases de transações criadas na etapa anterior através de um algoritmo proposto pelos próprios autores. O último passo consiste em filtrar os conjuntos de itens frequentes com o intuito de obter-se um número menor, porém mais relevante, de itens frequentes. Para este fim, os autores utilizaram uma função de poda que consiste em selecionar um pequeno grupo de itens que estão presentes em um número máximo de eventos. O desempenho em função do tempo de processamento e a utilização de recursos computacionais foi avaliada analiticamente com registros de eventos reais. A avaliação mostrou que a abordagem possibilitou a descoberta proativa de conhecimento em *log* de eventos, as quais podem conduzir novas investigações para identificar as melhores práticas ou então redesenhar o processo.

3.5 Considerações

Os estudos de Yoo et al. (2016) e Partington et al. (2015), podem ser considerados similares, pois os dois trazem uma análise comparativa de processos hospitalares. O primeiro comparou processos antes e após mudanças estruturais, já o segundo comparou processos de instituições distintas.

No trabalho de Yoo et al. (2016), o autor não desfrutou da quantidade de informações adicionais presentes no *log* de eventos, para por exemplo: enriquecer o modelo do processo; explorar técnicas da MD a fim de identificar possíveis causas que poderiam explicar alguns problemas no processo; identificar similaridade e diferenças dos processos em relação aos atributos existentes nos *logs* de eventos. Já no trabalho de Partington et al. (2015), o autor não avaliou se existe alguma significância estatística entre os dados das quatro instituições, de forma análoga ao trabalho de Yoo et al. (2016). Além disso, os resultados da análise do trabalho de Partington et al. (2015), podem ter sido prejudicados devido às dificuldades que o autor encontrou para unir os diferentes conjuntos de dados, bem como pela ausência de ferramentas para comparar diferentes processos de forma visual.

O estudo de Tsumoto et al. (2014) fez o uso de técnicas provindas da MD como, Árvores de Decisão e Trajectory Mining, porém a análise proposta pelo autor poderia ser beneficiada

pelas técnicas da MP. Pois os dados utilizados foram extraídos de um sistema de *workflow*, o que facilitaria o uso das técnicas da MP, e também porque ele poderia ter criado modelos com o intuito de obter uma visualização mais apropriada dos processos se comparado aos resultados obtidos pelo Trajectory Mining.

O trabalho de Spruit et al. (2014) fez um uso mais intenso da estatística descritiva do que da MD. Contudo, a aplicação da técnica Apriori para a identificação de regras associativas se mostrou eficaz. Os autores Huang et al. (2011), que apresentaram uma variação do Apriori para geração de regras a partir de *log* de eventos, não compararam o desempenho do algoritmo proposto em relação ao tempo de execução e uso de recursos computacionais de outros algoritmos presentes na literatura, de forma análoga ao trabalho de Djenouri et al. (2018), que apresentou uma abordagem para extrair itens frequentes de *log* de eventos. Vale ressaltar que, ambos autores, Huang et al. (2011) e Djenouri et al. (2018), não buscaram o auxílio de especialistas no domínio para avaliar a representatividade das regras associadas e itens frequentes obtidos em relação ao problema abordado.

Os estudos dos autores Delias et al. (2015), Suriadi et al. (2013) e Lismont et al. (2016) fazem a integração de técnicas da MP e MD para a análise de processos. Delias et al. (2015), entretanto, não demonstrou como os dados utilizados foram pré-processados, fato que gera incertezas sobre os resultados obtidos, já que os mesmos foram coletados manualmente.

O trabalho de Suriadi et al. (2013) é o único que não faz a análise de processos realizados no âmbito hospitalar, mesmo assim o seu caso de uso está relacionado, de certa forma, com o processo de reembolso de contas. Um fato que não ficou claro neste trabalho foi o porquê da não utilização de técnicas como abstração e agrupamento, uma vez que, o autor encontrou dificuldades para a geração de um modelo do processo mais simples e compreensível.

Já o trabalho Lismont et al. (2016), que propôs uma metodologia para análise de processos hospitalares, não empregou técnicas da MD para identificar fatores que possam ser usados para, por exemplo, diagnosticar problemas em processos como em Suriadi et al. (2013) ou então para a tomada de decisão como no trabalho Huang et al. (2011).

Por fim, perante aos trabalhos expostos, percebe-se que a literatura deixa lacunas para estudos que comparem processos hospitalares de forma análoga ao apresentado por Yoo et al. (2016) e Partington et al. (2015), porém incorporando técnicas da MD para a identificação de diferenças entre processos em função das informações geradas durante a execução dos mesmos.

Assim como estudos, que façam uso da mineração de regras associativas sobre *log* de eventos, contudo avaliando aspectos subjetivos das regras obtidas como, por exemplo, a relevância das mesmas em função do problema abordado. Diferentemente do que foi apresentado no trabalho de Huang et al. (2011), onde os autores avaliaram o tempo e o uso de recursos computacionais usados para obtenção de regras associativas.

Ou então estudos, que utilizem de forma mais intensiva as técnicas da MP em conjunto com as da MD, uma vez que os diferentes métodos podem se beneficiar um do outro. De modo distinto aos trabalhos de Djenouri et al. (2018) e Huang et al. (2011), nos quais os autores fizeram o uso de técnicas para a mineração de itens frequentes e para mineração de regras associativas sobre *log* de eventos, sem levar em consideração as técnicas da MP para analisar os dados de uma outra perspectiva.

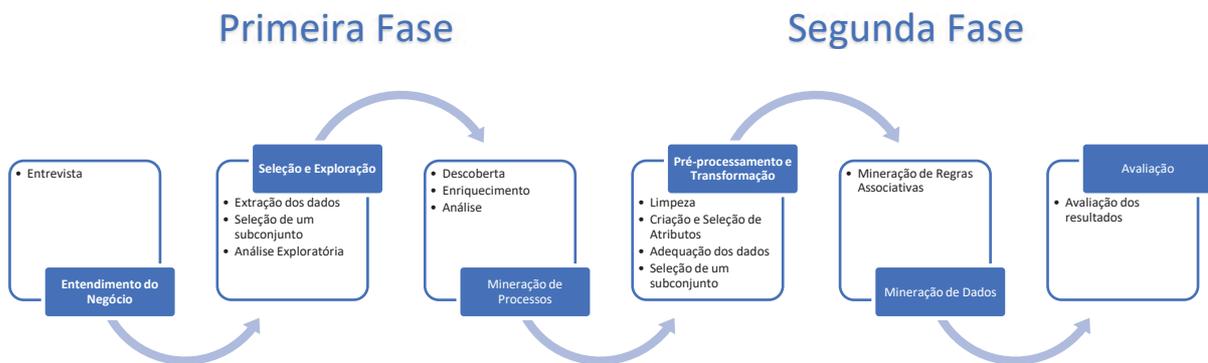
4 METODOLOGIA

Este capítulo, apresenta a metodologia utilizada para investigar a efetividade de técnicas da MD em encontrar fatores que levam ao atraso da submissão das contas hospitalares às respectivas seguradoras de saúde, no sentido de reduzir seu tempo de faturamento, por meio do mapeamento do processo de reembolso das despesas médico-hospitalares através da MP.

A metodologia empregada é formada por duas fases distintas cada uma composta por diferentes etapas, a qual está fundamentada na metodologia apresentada por Lismont et al. (2016) para a análise de processos hospitalares, bem como no Processo de Descoberta do Conhecimento em Bases de Dados.

Assim, a primeira fase busca, através do emprego das técnicas da MP, um entendimento inicial sobre o funcionamento do processo de reembolso da instituição a ser estudada, bem como a identificação de possíveis gargalos no processo. Já a segunda fase contempla o uso de técnicas da MD como, Árvores de Decisão e/ou o Apriori, para a geração de regras com o intuito de identificar fatores que levam ao atraso da submissão das contas hospitalares às respectivas seguradoras de saúde. A Figura 6, apresentada a seguir, demonstra a metodologia empregada, destacando-se os procedimentos ou técnicas empregadas em cada uma das fases e suas etapas.

Figura 6: Metodologia Proposta



Fonte: Elaborado pelo autor.

Conforme pode ser observado na Figura 6, a **Primeira Fase** é composta por três etapas cada uma com um objetivo distinto. Na etapa de **Entendimento do Negócio**, busca-se obter um entendimento inicial do funcionamento do processo de reembolso da instituição, já que o mesmo pode variar de instituição para instituição. Para tanto, foi conduzida uma entrevista com o gerente de operações da instituição.

Já a etapa de **Seleção e Exploração** contempla: a extração dos dados em conjunto com a instituição; a seleção de um subconjunto de interesse da instituição; e a análise exploratória do subconjunto de dados. Por fim, a etapa de **Mineração de Processos** envolve: a descoberta de modelos e o enriquecimento do mesmo com informações referentes ao desempenho e a frequência de execução das atividades; e por fim a análise dos resultados, onde espera-se identificar gargalos no processo através da análise do modelo descoberto e enriquecido.

Após a conclusão da análise do modelo de processo, inicia-se a **Segunda Fase** da metodologia proposta. A primeira etapa da segunda fase, a de **Pré-processamento e Transformação** inclui: a limpeza dos dados em razão dos problemas detectados na fase anterior; a criação de novos atributos a partir das informações existentes; a seleção automática dos atributos mais importantes; a adequação dos dados em função da técnica de mineração empregada; e por fim seleção a criação de dois subconjuntos de dados de interesse.

Dando continuidade, a etapa de **Mineração de Dados** tem como principal objetivo gerar regras, que possam explicar possíveis causas de problemas nos processos por meio de dados extraídos do *log* de evento e de métricas do processo (frequência de execução e desempenho). Nesta etapa, a geração de regras, é feita através do algoritmo Apriori, o qual é capaz de descobrir relações ou associações em um conjunto de dados. Por fim, cabe à etapa de **Análise de Resultados** avaliar a efetividade dos resultados obtidos na etapa de Mineração de Dados. Esta avaliação é feita através da aplicação de um questionário, onde um conjunto de especialistas avalia cada regra descoberta por meio de uma escala de Likert.

4.1 Considerações

Em resumo, a metodologia proposta possui duas fases distintas, cada uma com propósitos diferentes. A primeira fase tem como propósito explorar o processo através de técnicas da MP como a descoberta e enriquecimento de modelos de processos. Com isso, espera-se obter um entendimento inicial do processo, identificar gargalos no processo e também detectar desvios no fluxo do processo. Ressalta-se que, a exploração do processo também pode ser realizada exclusivamente com as técnicas da MD, porém com certas restrições. A principal delas, está relacionada com ausência de técnicas que possam ser usadas para mapear processos de forma automática.

Já a segunda fase, tem como propósito analisar o processo sobre a perspectiva dos dados gerados durante a sua execução por meio de técnicas da MD como, a mineração de regras associativas. Tal técnica, permite encontrar relações ou padrões entre variáveis pertencentes a um mesmo conjunto de dados, tarefa que não é possível fazer automaticamente por meio das técnicas da MP.

Concluindo, mediante dos argumentos expostos, bem como dos resultados obtidos pelos estudos apresentados na seção 3.4, os quais integram as técnicas de ambas disciplinas. Evidencia-se que a combinação das técnicas da MP e MD possa gerar resultados superiores, já que uma técnica complementa a outra, e vice-versa.

5 RESULTADOS

Este capítulo apresenta a aplicação da metodologia para a obtenção dos resultados. Inicia-se indicando as ferramentas e bibliotecas de software utilizadas para a implementação dos experimentos, passando para a etapa de seleção de dados, juntamente com a descrição dos mesmos, após o pré-processamento e transformação dos dados. E finalizando com a etapa de mineração de processos.

5.1 Ferramentas e bibliotecas utilizadas

O desenvolvimento dos experimentos foi realizado com o uso do software Disco (FLUXICON PROCESS LABORATORIES, 2012) e com *scripts* escritos com a linguagem Python (PYTHON SOFTWARE FOUNDATION, 2017) e R (R FOUNDATION, 2017)

Em resumo, os *scripts* escritos em Python foram usados com a finalidade de fazer o pré-processamento, a divisão do conjunto de dados, assim como a análise exploratória inicial dos dados. Sendo que, o desenvolvimento dos mesmos foi realizado com o auxílio da biblioteca Pandas (PYDATA DEVELOPMENT TEAM, 2017). Já a seleção automática de atributos e a geração de regras associativas, foi realizada respectivamente através das bibliotecas Party e Arules (HAHSLER et al, 2005) pertencentes à linguagem R.

Já o software Disco foi usado para a geração e enriquecimento de processos, bem como para a exploração do modelo e das métricas geradas. A geração dos processos é feita por padrão através do algoritmo Fuzzy Miner, que é implementado no software, já o enriquecimento dos processos inclui o cálculo de métricas referentes ao desempenho e a frequência da execução das atividades. Este software foi escolhido devido a sua facilidade de uso e porque o mesmo possibilita a criação de filtros que permitem a exploração de métricas de desempenho e dos dados presentes no *log* de eventos, funcionalidade esta que não é acessível no software de código aberto denominado de ProM Tools³. Vale ressaltar que, a utilização do Disco foi possível através de uma iniciativa, disponibilizado pelos seus desenvolvedores, que concede licenças para o uso acadêmico com certas restrições. Tais restrições, como a quantidade máxima de registros em um *log* de eventos, 5 milhões neste caso, bem como a proibição do uso do *software* para fins comerciais, não geram nenhum impedimento para a realização deste estudo.

5.2 Entendimento do Negócio

Os dados utilizados para o desenvolvimento deste estudo, foram coletados de um hospital filantrópico situado no sul do Brasil. Atualmente, esta instituição disponibiliza à comunidade 297 médicos e 421 enfermeiros(as), bem como 188 leitos, 157 referentes ao centro de internação e 31 a UTI, que estão distribuídos em uma área construída de 28 mil m². Sendo que, o centro de internação é composto pelas unidades de tratamento, clínico médico, cirúrgico, pediátrico e obstétrico e a UTI pelas unidades adulta, neonatal e pediátrica.

³ ProM Tools (www.promtools.org)

A instituição ainda conta com um centro cirúrgico, uma unidade de emergência e uma unidade de pronto atendimento 24 horas. O centro cirúrgico, possui 15 salas de cirurgias e procedimentos, que contam com 41 leitos e 10 poltronas de recuperação pós-anestésica. A unidade de emergência e atendimento 24 horas são compostos por salas de urgência, para politraumatizados, de observação com 15 leitos, para atendimento de queimados, suturas, gesso, e entre outros procedimentos.

Para se ter uma ideia da importância da instituição na sua região de atuação. Nos anos de 2015 e 2016 foram realizadas aproximadamente: 25 mil internações, 123 mil atendimentos nas unidades de emergência e pronto atendimento, assim como 140 mil atendimentos que incluem consultas e tratamentos de radioterapia e quimioterapia.

Para a realização da extração de dados, um projeto de pesquisa precisou ser submetido à apreciação do Centro de Ensino e Pesquisa (CENEPE) da instituição. O CENEPE é um setor responsável por aprovar e gerir todas as pesquisas de cunho acadêmico realizadas na instituição. Então, para dar início à extração de dados, o projeto de pesquisa teve que ser aprovado pelo comitê do CENEPE, o qual é formado por diferentes profissionais da instituição.

Assim, após a devida aprovação do projeto, foi conduzida uma entrevista com o gerente de operações com o objetivo de entender melhor o processo de reembolso do hospital. O processo de reembolso começa logo após a alta do paciente com a atividade de ajuste. Esta atividade ocorre na unidade onde o paciente recebeu o tratamento. Nela, pequenos ajustes são feitos no registro médico do paciente como, por exemplo, o registro dos materiais usados e medicamentos administrados, assim como a coleta de assinaturas de autorizações dos procedimentos médicos que não foram realizadas até o momento.

Após a etapa de ajustes, o registro médico é transferido para a auditoria. Nesta etapa, o registro médico do paciente é analisado. A análise visa verificar se todos os itens que devem ser cobrados foram lançados devidamente na conta do paciente, através da análise dos registros médicos e de enfermagem tais como, prescrições, evoluções, descrições cirúrgicas, checagem de medicamentos. Além disso, são observadas regras definidas pelos convênios por exemplo, quantidade máxima utilizada de um material/medicamento diário por paciente. Caso os auditores encontrem algum problema, o registro médico é retornado para a unidade onde o paciente foi tratado, para regularização.

Posteriormente, inicia-se a etapa de faturamento. Nesta etapa é realizada uma análise dos procedimentos e honorários médicos da conta, visando cobrar todos os procedimentos, exames e honorários executados no atendimento de acordo com as tabelas de codificação de procedimentos definidas pelos convênios. Muitas vezes, o registro médico ainda pode ter um problema por causa da ausência de uma informação requerida pela seguradora. Nesses casos, o registro médico é encaminhado para regularização, como feito na etapa da auditoria. A etapa final consiste em encaminhar a faturada para a seguradora de saúde.

A maioria das etapas são executadas durante os cinco dias da semana de trabalho, exceto pela etapa de ajuste que também pode ser executada nos sete dias da semana e inclusive nos feriados. Vale ressaltar que, conforme o aumento da demanda, a etapa de auditoria pode ser realizada aos sábados.

De acordo com o gerente de operações, as faturas relacionadas aos atendimentos são enviadas para as respectivas seguradoras em torno de 30 dias após a alta do paciente, o que vem gerando problemas financeiros para a instituição e afetando a qualidade dos serviços prestados. Sendo assim, a necessidade de se identificar fatores que levem ao atraso na submissão das

faturas é de extrema importância para a instituição hospitalar e para a população que depende de seus serviços.

5.3 Seleção e Exploração

Esta etapa envolve, a extração de um conjunto de dados com o auxílio do departamento de Tecnologia de Informação (TI) do hospital, a seleção de um subconjunto de interesse, e pôr fim a análise exploratória do subconjunto de dados.

A extração do conjunto de dados foi realizada pelo departamento de TI do hospital após uma reunião com a pessoa responsável pelo mesmo. Nesta reunião foram definidos quais dados seriam extraídos e o formato do conjunto de dados inicial. Os dados históricos contêm informações relacionadas à execução das principais atividades do processo de reembolso, que foram identificadas durante a entrevista realizada com o gerente de operações do hospital. Já o formato do conjunto de dados foi determinado segundo os requisitos necessários para a criação de um *log* de eventos, apresentados na seção 2.3.2. Vale ressaltar que a equipe de TI do hospital foi responsável pela anonimização dos dados, dessa forma nenhuma informação pessoal que poderia identificar o paciente foi utilizada neste estudo.

O conjunto de dados inicial contempla dados históricos relacionados ao processo de reembolso de 20.032 internações realizadas no hospital nos anos de 2015 e 2016. Destes casos, 65% estão associados a seguradora pública do país e o restante 35% estão associados a 12 seguradoras de saúde privadas. As principais atividades relacionadas ao processo de reembolso presentes no *log* de eventos podem ser visualizadas na Tabela 4. Algumas delas, contêm informações adicionais como o nome do setor onde a mesma foi executada e/ou ainda o tipo da pendência que deve ser corrigida. A atividade recepção, que não foi descrita na entrevista, compreende interações com agentes externos, como por exemplo, o recebimento de um documento ausente.

Tabela 3: Atividades presente no log de eventos e seus atributos

Atividade	Atributos
Alta do Paciente	ID do caso, Data da alta;
Ajuste	ID do caso; Data de início; Setor
Recepção	ID do caso; Data de início;
Auditoria	ID do caso; Data de início
Auditoria – Pendência	ID do caso; Data de início; Setor; Tipo da pendência;
Faturamento Central	ID do caso; Data de início
Faturamento - Pendência	ID do caso; Data de início; Setor; Tipo da pendência;
Inserção Protocolo Envio	ID do caso; Data de início
Envio Seguradora	ID do caso; Data de início

Fonte: Elaborado pelo autor.

Além disso, o *log* de eventos contém atributos que são comuns a todas as atividades, ou seja, eles são atributos pertencentes a um processo de reembolso específico (caso). Estes atributos e sua descrição podem ser visualizados na

Tabela 4 apresentada a seguir.

Tabela 4: Atributos do caso e sua descrição

Atributo	Descrição
Caráter da internação	Caráter da internação (urgente ou programada)
Especialidade Médica	Especialidade Médica
Seguradora de Saúde	Seguradora utilizada pelo paciente
Idade	Idade do paciente
Responsável	Médico responsável pela conta
Sexo	Sexo do paciente
Valor da fatura	Valor total da fatura
Tempo de internação	Tempo total de internação
Tipo do procedimento	Clínico, cirúrgico, ou ambos

Fonte: Elaborado pelo autor.

Do conjunto de dados inicial, foram selecionados somente os casos de uma seguradora de saúde privada que representa 25% dos casos. Esta seguradora, que utiliza o método pagamento por serviço para o cálculo do reembolso, é a mais representativa em função dos valores reembolsados para o hospital. Os dados da seguradora de saúde pública não foram selecionados devido a uma restrição contratual, que limita a quantidade de contas que podem ser enviadas e a frequência de envio das mesmas, uma vez ao mês neste caso. Restrições que inviabilizam o objetivo de identificar fatores que expliquem o atraso na submissão das faturas.

Após a definição do subconjunto, buscou-se um entendimento inicial do processo de reembolso através da análise exploratória dos dados. Durante a análise foram utilizadas técnicas de estatística descritiva, como por exemplo a análise univariada e bivariada de variáveis, para descrever os dados dos 5.008 casos e 39.589 eventos presentes no subconjunto de dados.

A Tabela 5 apresenta a sumarização de todos atributos numéricos e categóricos relacionados com os 5.008 casos. Nela, destaca-se: o número elevado de especialidade médicas (19) e profissionais responsáveis (108) envolvidos, o que ressalta a complexidade dos processos hospitalares; que a maioria dos tratamentos são cirúrgicos, considerando os atributos especialidade médica e tipo de procedimento; o número de internações programadas e urgentes é parecido; por fim, a dispersão dos valores dos atributos valor da fatura e o tempo de internação, as quais segundo o gerente de operações, podem influenciar no tempo total do processo de reembolso. Vale ressaltar que somente os cinco profissionais que atuam com maior frequência nas internações foram apresentados.

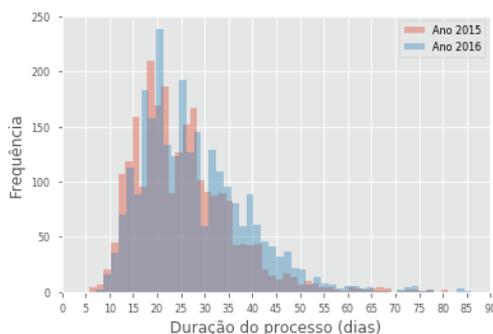
Tabela 5: Atributos numéricos e categóricos

Atributos Numéricos									
Nome	Valor mínimo	Valor Máximo	Média	Desvio Padrão					
Idade	0	102	48,56	23,12					
Valor da Fatura	59	261.671	4.215,42	8.619,06					
Tempo de Internação	1	30	4,53	4,23					
Atributos Categóricos									
Nome	Categoria	Freq.	Freq. (%)	Nome	Categoria	Freq.	Freq. (%)		
Especialidade Médica	Obstetrica/Ginecologia	946	18,89	Caráter da Internação	Programada	2.597	51,86		
	Cirurgia Geral	945	18,87		Urgente	2.411	48,14		
	Ortopedia/Traumatologia	612	12,22		Responsável	878	288	5,75	
	Cardiologia	602	12,02			134	286	5,71	
	Clínica Médica	482	9,62			67408	219	4,37	
	Neurologia	280	5,59			137	219	4,37	
	Pediatria	220	4,39		620486	179	3,57		
	Urologia	182	3,63		Sexo	Feminino	3.098	61,86	
	Otorrinolaringologia	137	2,74			Masculino	1.910	38,14	
	Oncologia	108	2,16		Tipo do Procedimento	Cirúrgico	1.925	38,44	
	Proctology	99	1,98			Clínico & Cirúrgico	1.590	31,75	
	Psiquiatria	96	1,92			Clínico	1.493	29,81	
	Pneumologia	89	1,78						
	Gastroenterologia	79	1,58						
	Nefrologia	43	0,86						
	Deformidades Lábio-Palatais	25	0,50						
Cirurgia Bucomaxilofacial	25	0,50							
Cirurgia Plástica	20	0,40							
Hematologia	18	0,36							

Fonte: Elaborado pelo autor.

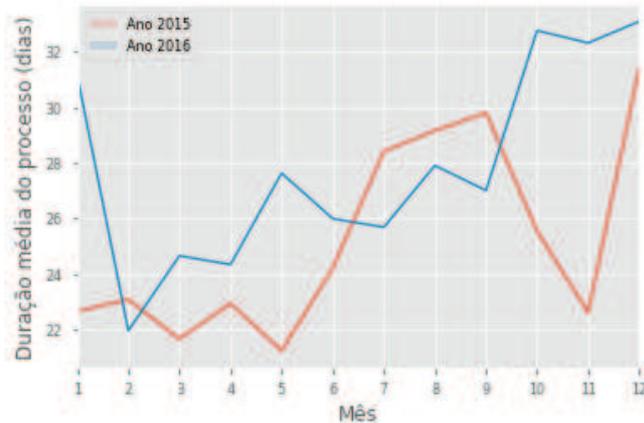
Analisando a duração do processo através da Figura 7, é possível observar que a houve uma variação pequena entre os anos 2015 ($\mu = 25,42$; $\sigma = 10,33$) e 2016 ($\mu = 27,78$; $\sigma = 11,17$). No entanto, ao comparar a duração do processo em função do mês de início do processo (Figura 8), é possível perceber que existe uma diferença entre os dois anos, nos meses de janeiro, junho, outubro e novembro.

Figura 7: Duração do processo



Fonte: Elaborado pelo autor.

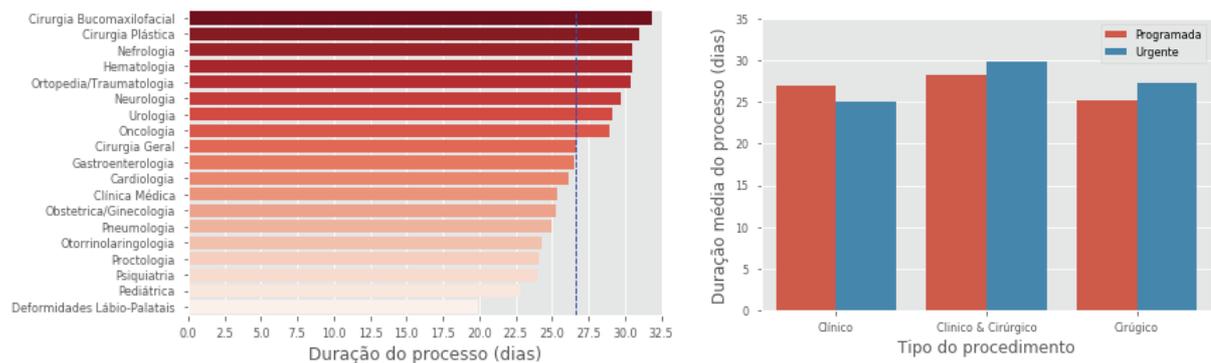
Figura 8: Duração do processo em função do mês de início



Fonte: Elaborado pelo autor.

Já na Figura 9, é possível visualizar a duração do processo em função da especialidade médica, e do tipo de procedimento e caráter de internação. Em relação à especialidade médica, percebe-se que o tempo de duração média de quase a metade das especialidades é maior que 26.64 dias (média global representada pela linha azul tracejada). Já em relação ao caráter de internação e tipo de procedimento é possível notar que os atendimentos programados relacionados aos procedimentos clínicos demoram em média mais que os atendimentos urgentes, diferentemente dos casos que envolvem procedimentos cirúrgicos e clínicos ou somente cirúrgicos.

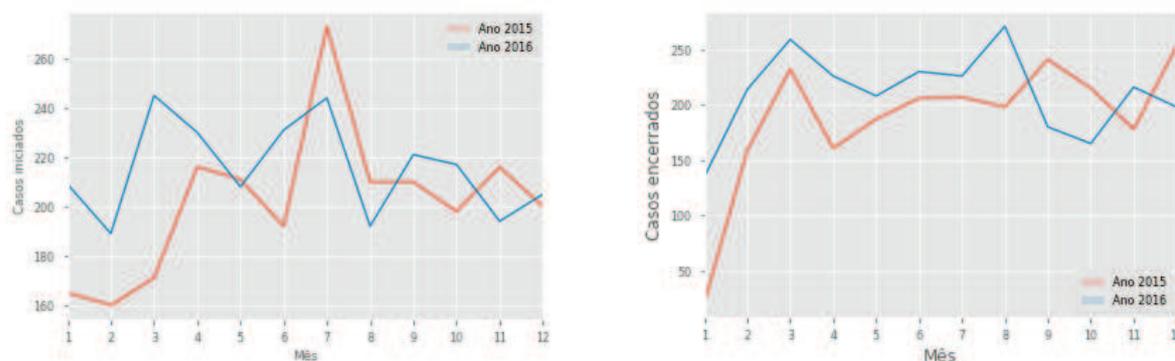
Figura 9: Duração do processo em função da especialidade médica (esquerda) e do tipo do procedimento e caráter de internação (direita)



Fonte: Elaborado pelo autor.

Por fim, na Figura 10, é possível perceber uma relação entre a quantidade de casos iniciados e encerrados nos anos de 2015 e 2016. Por exemplo, na figura, nota-se um aumento de casos iniciados entre os meses de julho e agosto em ambos os anos, o que provavelmente provocou um aumento de casos encerrados nos meses de agosto de 2015 e setembro de 2016. O número de casos encerrados no mês de janeiro do ano de 2015 é menor pois os novos casos que iniciaram no final do ano de 2014 não foram considerados.

Figura 10: Casos iniciados e encerrados



Fonte: Elaborado pelo autor.

5.4 Mineração de Processos

A Mineração de Processos incluiu as atividades de descoberta e enriquecimento de modelos de processos realizadas com o software Disco. A opção desse software deu-se devido, primeiro a facilidade de uso em relação à visualização do processo e a exploração do mesmo através de filtros e dados estatísticos. Em segundo lugar, pois o mesmo implementa o algoritmo Fuzzy Miner, o qual segundo Lismont et al. (2016) é particularmente útil para lidar com processos hospitalares devido a sua capacidade de abstração e agrupamento de atividades. Por último, devido à capacidade do Disco em agrupar atividades e/ou então fluxos em função da frequência, assim é possível analisar tanto o comportamento frequente e infrequente do processo.

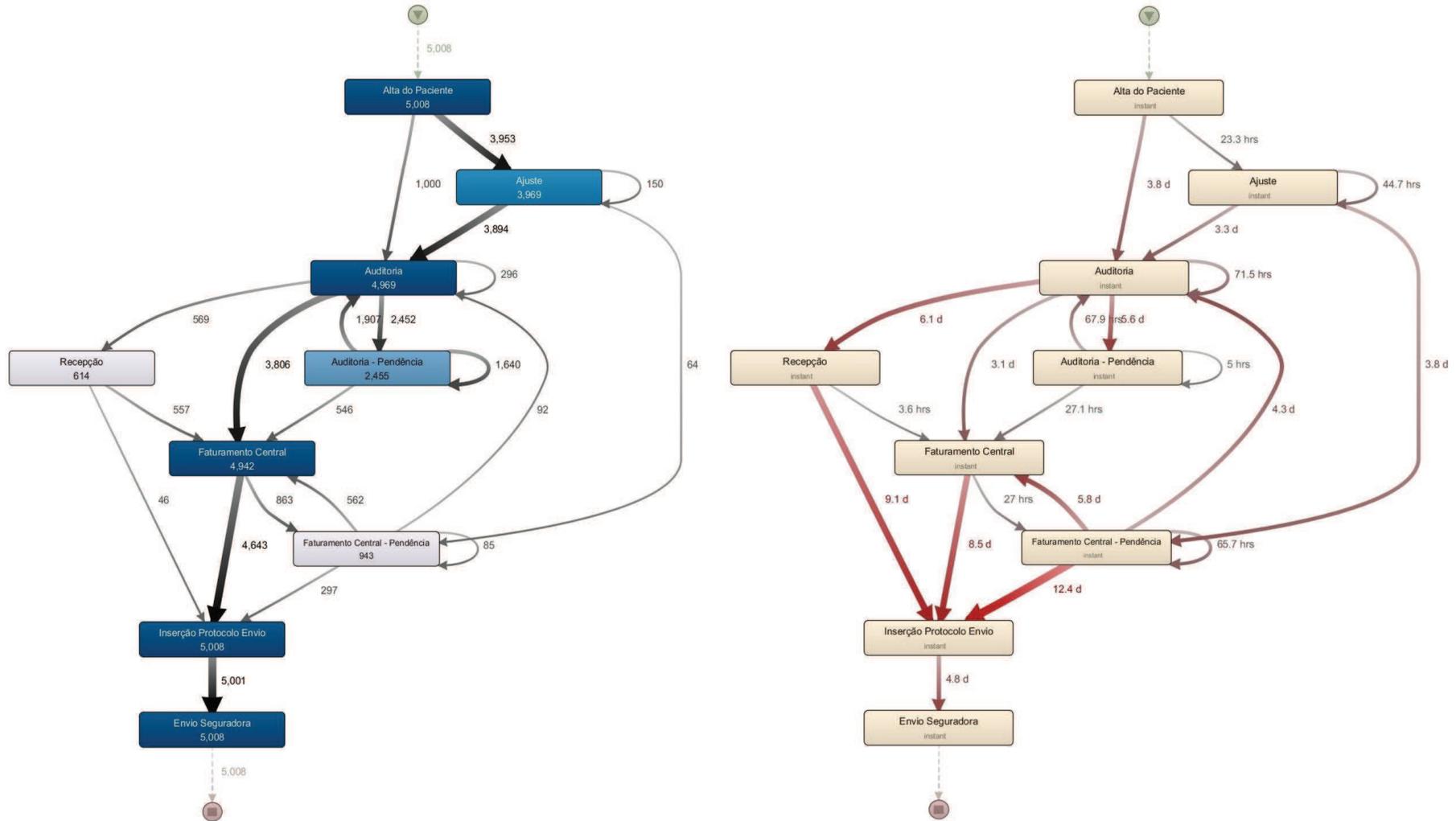
Através do software Disco foi gerado um modelo de processo com base no subconjunto de dados selecionado na etapa anterior. Por padrão, o Disco automaticamente enriquece o processo com métricas relativas ao tempo e a frequência de execução de cada atividade. Sendo que, as métricas de tempo são formadas pela duração média, mediana, mínima, máxima de cada atividade, bem como o tempo total de cada atividade levando em consideração todos os registros. Já as métricas de frequência são compostas pela frequência de execução das atividades absoluta, frequência de execução das atividades em relação ao caso, bem como pelo número total de repetições de uma atividade em relação ao caso.

O modelo do processo gerado com Disco pode ser observado na Figura 11, onde são apresentados os modelos, em forma de grafos, enriquecidos em razão da frequência de execução das atividades (esquerda) e do desempenho das atividades (direita). Na figura da esquerda, é possível identificar os fluxos mais percorridos e atividades executadas, através da espessura das arestas (menor menos frequente e maior mais frequente) e da cor dos nodos (azul claro menos frequente e azul escuro mais frequente). Já, na figura da direita, é possível visualizar as atividades que levaram mais tempo para serem executadas, por meio da espessura das arestas, neste caso quanto menor a espessura menor o tempo e quanto maior a espessura maior o maior o tempo de execução.

Em relação à frequência, nota-se a grande quantidade de casos onde foram detectados problemas durante a fase de auditoria. Desses casos, 1.640 possuíam mais de um problema, o

que pode ser observado pelo *loop* formado na atividade “Auditoria - Pendência”. Durante a fase de faturamento também é possível perceber uma grande quantidade de problemas detectados, porém em menor escala se comparado com a fase de autoria. Fatores que podem estar associados à demora no processo de reembolso. Por fim, constata-se desvios no processo, entre as atividades “Ajuste” e “Faturamento Central – Pendência”, bem como entre as atividades “Recepção” e “Inserção Protocolo Envio”, o que indica que o processo não está tão bem estruturado como foi demonstrado durante a entrevista com o gerente.

Figura 11: Modelos do convênio A enriquecidos com a frequência (esquerda) e desempenho (direita)



Fonte: Elaborado pelo autor.

Em relação ao desempenho das atividades, é possível notar que problemas estavam sendo reconhecidos após cinco dias em média na etapa de auditoria. No entanto, esses problemas foram resolvidos dentro de 68 horas em média, como pode ser visto nas arestas que ligam as atividades “Auditoria” e “Auditoria – Pendência”. Por outro lado, durante o faturamento os problemas são identificados em menos tempo, porém eles levam mais tempo para serem resolvidos, conforme o tempo médio entre as atividades “Faturamento” e “Faturamento – Pendência”. Finalmente, pode-se notar que a maior parte do tempo gasto no processo envolve as etapas de auditoria e faturamento, como destacado na Figura 11 pela espessura das arestas vermelhas.

5.5 Pré-processamento e Transformação de Dados

Esta fase inclui as seguintes atividades: limpeza de dados com o objetivo melhorar a qualidade do conjunto de dados; a criação e seleção de atributos; a transformação do conjunto de dados conforme os requisitos da técnica de mineração e finalmente a seleção de dados de acordo com o interesse da instituição.

A limpeza de dados foi realizada com o software Disco e com o auxílio de *scripts*, através da filtragem de casos que apresentem alguma anomalia. Esta filtragem de dados contempla: casos contendo algum desvio, como os que foram apresentados na etapa de Mineração de Processo; eventos com alguma data de execução inválida ou ausente; eventos relacionados às atividades “Ajuste”, “Recepção” e “Faturamento Central” que ocorreram em sequência, nestes casos somente a primeira ocorrência do evento foi mantida. Como resultado da limpeza, restaram 4.816 casos e 38.605 eventos no conjunto de dados.

Após a limpeza, foram criados os seguintes atributos, “tempo-auditoria”, “tempo-faturamento”, “mês-de-início”, “setores” e “pendências” a partir das informações existentes no conjunto de dados. Os atributos “tempo-auditoria” e “tempo-faturamento”, que representam as principais atividades do processo de reembolso, foram criados usando-se o tempo entre as atividades “Auditoria” e “Faturamento Central”, e “Faturamento Central” e “Inserção Protocolo Envio”. Já o atributo “mês-de-início” foi criado a partir da data de execução da atividade “Alta do Paciente”, com o objetivo de detectar algum padrão temporal apresentado na análise exploratória dos dados. Por fim, os atributos “unidades” e “pendências” formam dois conjuntos contendo os setores em que o registro de paciente passou e as pendências detectadas durante a etapa de auditoria e faturamento.

Para determinar quais os atributos mais relevantes para o problema, utilizou-se o algoritmo *Random Forest* disponível na biblioteca Party da linguagem R. Como dito anteriormente, este algoritmo é capaz de medir a força de predição de cada atributo em função da variável dependente, representada pela duração total do processo de reembolso neste caso. Para a execução do algoritmo foram utilizados os parâmetros pré-definidos da biblioteca (número de árvores=500; número de variáveis de entrada = $n^{\circ} \text{ total de variáveis} / 3$). O resultado da execução do algoritmo, pode ser observado na Tabela 6, onde a importância é determinada através da permutação dos valores de cada atributo e a mensuração de quanto a mesma diminui o erro médio do modelo. A partir dos resultados obtidos, os três atributos menos importantes foram removidos, “Valor da fatura”, “Idade do paciente” e “Sexo”.

Tabela 6: Importância dos atributos

Posição	Atributo	Importância
1	Tempo Faturamento	69,23
2	Tempo Auditoria	55,99
3	Setores	8,61
4	Mês de Início	1,48
5	Especialidade médica	1,16
6	Responsável	1,05
7	Tipo do procedimento	1,01
8	Pendências	0,76
9	Tempo de internação	0,73
10	Caráter da internação	0,60
11	Valor da fatura	0,50
12	Idade do paciente	0,45
13	Sexo	0,14

Fonte: Elaborado pelo autor.

Após a seleção dos atributos, os dados foram transformados de acordo com os requisitos do algoritmo Apriori e dois subconjuntos foram criados com base no interesse da instituição. Primeiramente, os atributos “Tempo Auditoria”, “Tempo Faturamento”, “Tempo de internação” foram discretizados manualmente, com o auxílio do gerente de operações. Após isso, o conjunto de dados estruturado segundo o formato CSV (*Comma-separated values*), foi transformado em um conjunto de transações, onde cada linha do arquivo CSV passou a representar uma transação no conjunto. Finalmente, foram criados dois subconjuntos transações, denominados de A e B, a partir do conjunto de transações com base em o interesse da instituição. O conjunto A contempla 1.962 transações e o conjunto B contempla 1.055 transações, os quais estão relacionados respectivamente com processos que levaram entre 20-30 e 30-40 dias para terminar.

5.6 Mineração de Dados

Esta seção apresenta os detalhes de como o algoritmo Apriori foi empregado com o intuito de encontrar fatores que levam ao atraso da submissão das contas hospitalares às respectivas seguradoras de saúde. Desse modo, espera-se encontrar relações e ou associações entre itens pertencentes aos conjuntos A e B que possam explicar o tempo elevado do processo de reembolso.

Para a geração de regras, a biblioteca Arules disponível no pacote de software RStudio foi utilizada. A definição do suporte mínimo foi feita através de uma série de experimentos, que visaram a obtenção de 40 regras para cada conjunto. Desse modo, somente os itens mais frequentes foram usados para a geração de regras. No final dos experimentos, os suportes mínimos de 0,056 e 0,043 foram atribuídos respectivamente para a geração dos conjuntos A e B. Já para o nível de confiança mínima foi atribuído o valor de 0,75. Por fim, o número de itens máximo de uma regra foi limitado para 3, com o objetivo de reduzir a complexidade das regras geradas.

As Tabela 7 e Tabela 8 apresentam algumas regras geradas respectivamente através dos conjuntos A e B, sendo que as demais podem ser visualizadas nos anexos A e B. Em relação as regras geradas do conjunto de dados A, que compreende os processos que duraram de 20 até 30 dias, pode-se perceber problemas relacionados às internações programadas. Nesses casos, o processo não deveria levar muito tempo para ser concluído já que o hospital está ciente dos materiais, medicamentos, procedimentos e exames necessários para o tratamento do paciente. Isto fica evidente na regra 3, onde problemas ocorridos no setor Centro Cirúrgico foram identificados na fase de auditoria. Por fim, nota-se que mesmo quando o tempo de permanência do paciente, o que sugere menos itens para serem auditados, for inferior ao tempo médio (4,53 dias) a etapa de auditoria pode levar de 10 à 15 dias para ser concluída, como mostrado na regra 5 na Tabela 7.

Tabela 7: Regras geradas a partir do conjunto A

#	Regra	Suporte (%)	Confiança (%)
1	Tipo do procedimento=Clinico Então Caráter da internação=Urgente	0,252	0,893
2	Tipo do procedimento=Cirúrgico E Tempo Auditoria=5-10 Então Caráter da internação=Programada	0,142	0,904
3	Tipo da Pendência=Pendência bloco E Setor=Centro Cirúrgico Então Caráter da internação=Programada	0,125	0,926
4	Tempo Auditoria=0-5 E Tempo de Internação=2 Então Tipo do procedimento=Cirúrgico	0,116	0,776
5	Tempo Auditoria=5-10 E Tempo de Internação=2 Então Caráter da internação=Programada	0,108	0,821

Fonte: Elaborado pelo autor.

Tabela 8: Regras geradas a partir do conjunto B

#	Regra	Suporte (%)	Confiança (%)
1	Tempo de Internação=5-10 Então Caráter da internação=Urgente	0,110	0,771
2	Tipo da Pendência=Assinatura Ausente E Caráter da internação=Programada Então Setor=Centro Cirúrgico	0,092	0,758
3	Especialidade Médica=Cirurgia Geral E Setor=H6 Então Caráter da internação=Programada	0,086	0,766
4	Especialidade Médica= Ortopedia/Traumatologia E Tipo do Procedimento=Cirúrgico Então Caráter da internação=Programada	0,083	0,872
5	Tipo do Procedimento=Cirúrgico E Tempo Auditoria=5-10 Então Caráter da internação=Programada	0,077	0,864

Fonte: Elaborado pelo autor.

No que tange as regras criadas a partir do conjunto de dados B. Pode-se observar o mesmo problema com internações programadas presentes nas regras do conjunto A. Além disso, a regra 1 da Tabela 8 sugere que um longo período de permanência do paciente pode influenciar no tempo total do processo quando a internação do paciente for urgente, ou seja, quando a equipe do hospital desconhece os procedimentos necessários para o tratamento do paciente. Por fim, também é possível observar diversas regras relacionadas a procedimentos cirúrgicos nos processos de reembolso que levaram de 30 à 40 dias para terminar, fato que pode estar relacionado com a complexidade desse tipo de procedimento.

5.7 Avaliação

A avaliação das 80 regras geradas a partir dos conjuntos A e B, foi realizada através da aplicação de um questionário, que pode ser observado no anexo C, a dois grupos de participantes. A aplicação deste questionário teve como objetivo averiguar a efetividade das

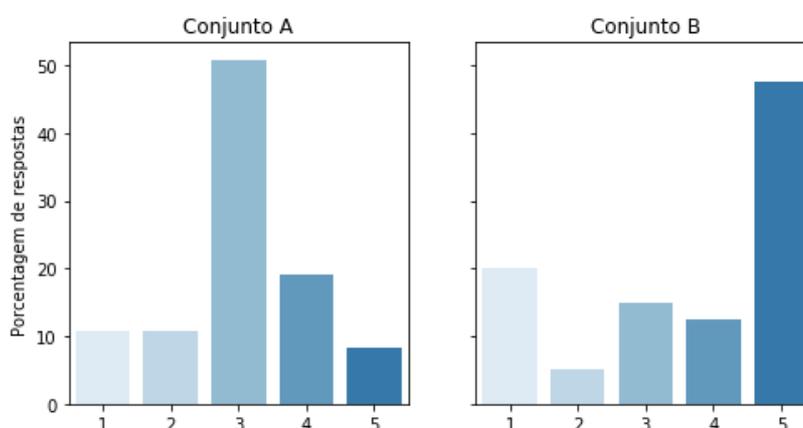
regras geradas em função do problema apresentado, e não do tempo necessário e os recursos computacionais usados para a geração das regras, como apresentado nos trabalhos de Huang et al. (2011) e Djenouri et al. (2018).

No questionário desenvolvido, os participantes avaliaram a efetividade das 80 regras através de uma escala Likert, composta por cinco critérios (1 - Nenhuma Relevância; 2 - Pouca Relevância; 3 - Média Relevância; 4 - Relevante; 5 - Muita Relevância). Além disso, também foi verificado, com duas perguntas abertas, se alguma regra em função de cada conjunto era desconhecida pelos participantes. Por fim, através de outra pergunta aberta, os participantes puderam apresentar sua opinião pessoal sobre a efetividade das regras para encontrar fatores e reduzir o tempo do processo de reembolso do hospital.

O questionário foi aplicado aos dois grupos por intermédio do gerente de operações do hospital, nos setores onde os funcionários trabalham atualmente. Estes dois grupos formados por 4 e 3 profissionais que trabalham, em média a 10 anos no hospital, nos setores de auditoria e faturamento respectivamente, estes responderam às perguntas de forma separada. Os funcionários do faturamento avaliaram a relevância de todas as regras, porém não responderam às perguntas abertas. Já os funcionários da auditoria, por algum motivo, desconhecido até o momento, iniciaram a avaliação, rasuraram as respostas marcadas, e em seguida recomeçaram o questionário marcando todas as regras como não relevantes. Uma hipótese levantada sobre o acontecimento foi que os funcionários da auditoria começaram a responder em um dia e por falta de tempo continuaram no próximo dia útil. Devido a isso, as respostas do grupo de funcionários da auditoria foram descartadas.

O resultado da avaliação da efetividade das regras geradas através dos conjuntos A e B é apresentado na Figura 12. Na figura, nota-se que mais de 50% das regras do conjunto de dados A foi considerado como sendo de média relevância. Já no conjunto de dados B, mais de 40% das regras foram consideradas muito relevantes, contudo neste conjunto um maior número de regras foi considerada não relevante. Considerando os dois conjuntos, 45% das 80 regras foram consideradas como relevantes ou então muito relevantes, o que demonstra a aplicabilidade das técnicas empregadas.

Figura 12: Resultado do questionário



Fonte: Elaborado pelo autor.

5.8 Considerações

Este capítulo apresentou os resultados obtidos através da metodologia proposta. Como apresentado anteriormente, com as técnicas da MP foi possível detectar diversos problemas no processo de reembolso da instituição como, por exemplo, o tempo gasto na etapa da auditoria para a resolução das pendências encontradas, o número alto de problemas detectados com os registros médicos dos pacientes, bem como desvios na execução do processo.

Com as regras da associação foi possível identificar fatores que levam ao atraso da submissão das contas hospitalares às respectivas seguradoras de saúde. Por exemplo, a regra número 2 da Tabela 8 (Tipo da Pendência=Assinatura Ausente E Caráter da internação=Programada **Então** Setor=Centro Cirúrgico) indica que, um profissional da saúde não assinou uma autorização para a realização de um procedimento programado no Centro Cirúrgico. Por se tratar de um procedimento programado, a equipe administrativa poderia enviar uma mensagem para os profissionais envolvidos com o intuito de reforçar a necessidade de assinar autorizações. Esta simples ação poderia minimizar o esforço para encontrar o profissional responsável para assinar a autorização. Em uma análise foi verificado que esse tipo de situação aconteceu em 250 casos aproximadamente.

Por fim, destaca-se o resultado positivo obtido com avaliação das regras feita por profissionais envolvidos diretamente com o processo de reembolso. Nesta avaliação, quase 45% das 80 regras associativas, geradas através dos dados dos processos que duraram de 20 à 30 dias e 30 à 40 dias, foram consideradas relevantes ou muito relevantes.

6 CONCLUSÃO

O processo de reembolso de despesas médico-hospitalares consiste em cobrar por todos os serviços de saúde e cuidados médicos prestados por instituições de saúde. Assim, melhorar o processo de reembolso torna-se uma tarefa decisiva para todas as partes envolvidas na prestação de serviços de saúde e cuidados médicos. Pois atrasos no recebimento do reembolso podem levar a problemas financeiros e, conseqüentemente, afetar a qualidade dos serviços prestados à população.

Neste sentido, o presente trabalho tem como objetivo investigar a aplicabilidade de métodos da MP e MD, para encontrar fatores que levam ao atraso da submissão das contas hospitalares às respectivas seguradoras de saúde. Para isso, uma metodologia foi proposta para combinar os métodos de ambas disciplinas. Em linhas gerais, a metodologia contempla duas fases distintas, a primeira onde foram empregados métodos da MP, com o objetivo de fazer um mapeamento do processo de reembolso e a segunda onde foram utilizados métodos da MD para a descoberta de padrões no conjunto de dados de forma automática.

Durante a primeira fase, por meio da análise dos modelos de processos descobertos e enriquecidos, foi possível identificar fatores que podem levar ao atraso no processo como: o número alto de problemas detectados com a conta do paciente durante as fases da auditoria e faturamento; desvios no processo que indicam que o processo não está tão bem estruturado; o tempo elevado para a detecção de problemas na etapa da auditoria; e o tempo elevado para a resolução de problemas na etapa do faturamento.

Já na segunda fase, foi possível encontrar padrões e relacionamentos entre as variáveis do conjunto de dados de forma automática, através da mineração de regras associativas. Com as regras obtidas foi possível identificar fatores como, por exemplo, a ausência de assinaturas de autorização em procedimentos previamente agendados e realizados no centro cirúrgico da instituição. Problema que, caso evitado, por meio de uma notificação, poderia reduzir o tempo total do processo de reembolso.

Assim, evidencia-se que os métodos da MP e MD se complementam, no sentido que um pode ser usado para analisar o processo visualmente por meio de modelos e o outro pode ser usado para analisar o processo em função dos seus dados.

Na literatura, diversos estudos aplicaram técnicas da MP em conjunto com técnicas de MD. No estudo apresentado por Delias et al. (2015), os autores empregaram técnicas de agrupamento, para a geração de modelos de processos mais simples. Já no trabalho de Suriadi et al. (2013), os autores utilizaram Árvores de Decisão, com o intuito de gerar regras que possam explicar atrasos nos processos.

O presente estudo propôs uma metodologia que combina a descoberta e enriquecimento de modelos de processo com a mineração de regras de associação para a análise de processo do âmbito hospitalar. De maneira diferente do que foi apresentado nos estudos de Huang et al. (2011) e Djenouri et al. (2018) onde, os autores, fizeram uso da mineração de regras de associativas e itens frequentes sobre *log* de eventos, porém sem empregar os métodos do MP. Também foi avaliado a subjetividade das regras, ou seja, o grau de relevância das regras em função do problema, e não somente o tempo necessário e os recursos computacionais usados para a geração das regras, como apresentado no trabalho de Huang et al. (2011) e Djenouri et al. (2018).

A principal limitação deste estudo decorre da análise de apenas um método de pagamento de reembolso e em uma única instituição, uma vez que existem diferentes métodos e os mesmos podem variar de instituição para instituição. No entanto, devido aos resultados obtidos, a abordagem proposta pode ser aplicada a outros cenários semelhantes, porém com certas adaptações referentes às características do ambiente. Por exemplo, as atividades podem ser adicionadas ou removidas no processo de acordo com o método de pagamento ou ambiente, e a importância de diferentes atributos pode ser testada antes da mineração de regras associativas. Enfim, a metodologia proposta pode ser empregada como ponto de partida para melhorar o processo de reembolso baseados em outros métodos e em outras instituições, dado que existe uma carência de estudos entorno do tópico de reembolso de contas médico-hospitalares (LEE et al., 2016).

Estudos futuros, podem vir ao encontro das limitações apresentadas, ou seja, dar suporte a diferentes métodos de reembolso através da extensão da metodologia proposta. Além disso, a aplicabilidade de outras técnicas da MP como, a verificação de conformidade de processos e *handover of work* poderiam ser testadas, o intuito de verificar, de forma automática, desvios presentes no processo e como os diferentes profissionais se relacionam. Por fim, um conjunto maior de dados e um número maior de profissionais da auditoria e faturamento poderiam ser usados com o intuito de validar uma possível extensão da metodologia.

REFERÊNCIAS

- AALST, W. van der. Business Process Management: a comprehensive survey. **ISRN Software Engineering**, [S.l.], v. 2013, p. 1–37, 2013.
- AALST, W. van der. **Process Mining: data science in action**. [S.l.]: Springer Berlin Heidelberg, 2016.
- AALST, W. van der; SCHONENBERG, M.; SONG, M. Time prediction based on process mining. **Information Systems**, [S.l.], v. 36, n. 2, p. 450-475, apr 2011
- AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining Association Rules Between Sets of Items in Large Databases. **SIGMOD Rec.**, New York, NY, USA, v. 22, n. 2, p. 207–216, June 1993.
- AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994, San Francisco, CA, USA. **Proceedings**. Morgan Kaufmann Publishers Inc., 1994. p. 487-499. (VLDB '94).
- AHA. **Exploring the Impact of the RAC Program on Hospitals Nationwide - Results of AHA RACTRAC Survey, 4th Quarter 2015**. Disponível em: <http://www.aha.org/content/16/15q4ractracresults.pdf>. Acesso em: 2 janeiro 2017
- BOSE, R. J. C.; MANS, R. S.; AALST, W. M. van der. Wanna improve process mining results? In: IEEE SYMPOSIUM ON COMPUTATIONAL INTELLIGENCE AND DATA MINING (CIDM), 2013., 2013. **Anais**. Institute of Electrical and Electronics Engineers (IEEE), 2013.
- BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.
- BRILL, S. **America's Bitter Pill: Money, Politics, Backroom Deals, and the Fight to Fix Our Broken Healthcare System**. [S.l.] Random House Publishing Group, 2015.
- CARDOSO, M.; FERLA, A.; OLIVEIRA, P.; NUNES, N.; O mix público e privado no sistema de saúde brasileiro: coexistência em evidência. **Saúde em Redes**, vol. 3, n.2, p. 107-118, 2017.
- COSTA, N. Austeridade, predominância privada e falha de governo na saúde. **Ciência & Saúde Coletiva**, v. 22, n. 4, p. 1065-1074, 2017.
- DELIAS, P.; DOUMPOS, M.; GRIGOROUDIS, E.; MANOLITZAS, P.; MATSATSINIS, N. Supporting healthcare management decisions via robust clustering of event logs. **Knowledge-Based Systems**, v. 84, p. 203-213, 2015.
- DJENOURI, Y.; BELHADI, A.; FOURNIER-VIGER, P. Extracting useful knowledge from event logs: A frequent itemset mining approach. **Knowledge-Based Systems**, v. 139, p. 132-148, 2018.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, [S.l.], v. 17, p. 37–54, 1996.

FERENC, D. **Understanding Hospital Billing and Coding**. [S.l.]: Elsevier Health Sciences, 2013.

FLUXICON. **Generate Your Own Event Log From Oracle E-Business Suite**. Disponível em: <<https://fluxicon.com/blog/2015/04/generate-your-own-event-log-from-oracle-e-business-suite/>>. Acesso em: 22 novembro 2017.

FLUXICON FOUNDATION. **Process Mining and Process Analysis - Fluxicon**. Disponível em: <https://fluxicon.com/>. Acesso em: 15 setembro 2017.

GIOVANELLA, L.; ESCOREL, S.; VASCONCELOS COSTA LOBATO, L.; CARVALHO NORONHA, J.; CARVALHO, A. **Políticas e sistema de saúde no Brasil**. [S.l.]: SciELO - Editora FIOCRUZ, 2012.

GRAGNOLATI, M.; LINDELÖW, M.; COUTTOLENC, B. **Twenty years of health system reform in Brazil: an assessment of the Sistema Único de Saúde**. Washington, D.C.: The World Bank, 2013.

GÜNTHER, C. W.; AALST, W. M. P. van der. Fuzzy Mining - Adaptive Process Simplification Based on Multi-perspective Metrics. In: **Lecture Notes in Computer Science**. [S.l.]: Springer Berlin Heidelberg, 2007. p. 328–343.

HAHSLER, M.; GRÜN, B.; HORNIK, K. arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. **Journal of Statistical Software**, v. 14, n. 15, 2005.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining: concepts and techniques**. [S.l.]: Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems).

HAPFELMEIER, A; HOTHORN, T; ULM, K; STROBL, C. A new variable importance measure for random forests with missing data. **Statistics and Computing**, v. 24, n. 1, p. 21-34, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: data mining, inference, and prediction**. [S.l.]: Springer New York, 2009.

HHS. **FY 2016 HHS Agency Financial Report**. Disponível em: <https://www.hhs.gov/sites/default/files/fy-2016-hhs-agency-financial-report.pdf>. Acesso em: 1 dezembro 2017.

HUANG, Z.; LU, X.; DUAN, H. Mining association rules to support resource allocation in business process management. **Expert Systems with Applications**, [S.l.], v. 38, n. 8, p. 9483–9490, aug 2011.

IBGE. **Conta-Satélite de Saúde 2010-2015: em 2015, 9,1% do PIB foram gastos no consumo de bens e serviços de saúde**. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/18915-conta-satelite-de-saude-release.html>. Acesso em: 10 janeiro 2018.

KEEHAN, S. P.; STONE, D. A.; POISAL, J. A.; CUCKLER, G. A.; SISKI, A. M.; SMITH, S. D.; MADISON, A. J.; WOLFE, C. J.; LIZONITZ, J. M. National Health Expenditure

Projections, 2016-25: price increases, aging push sector to 20 percent of economy. **Health Affairs**, [S.l.], v. 36, n. 3, p. 553–563, 2017.

LAROSE, D. **Discovering Knowledge in Data**: an introduction to data mining. [S.l.]: Wiley, 2014. (Wiley Series on Methods and Applications in Data Mining).

LEE, S. J.; ABBEY, J. D.; HEIM, G. R.; ABBEY, D. C. Seeing the forest for the trees: institutional environment impacts on reimbursement processes and healthcare operations. **Journal of Operations Management**, [S.l.], v. 47-48, p. 71-79, 2016.

LISMONT, J.; JANSSENS, A.-S.; ODNOLETKOVA, I.; BROUCKE, S. vanden; CARON, F.; VANTHIENEN, J. A guide for the application of analytics on healthcare processes: a dynamic view on patient pathways. **Computers in Biology and Medicine**, [S.l.], v. 77, p. 125–134, oct 2016.

LUXBURG, U. von. A tutorial on spectral clustering. **Statistics and Computing**, [S.l.], v. 17, n. 4, p. 395–416, 2007.

MANS, R. S.; AALST, W. M. P. van der; VANWERSCH, R. J. B. **Process Mining in Healthcare**. [S.l.]: Springer International Publishing, 2015.

MARTEN, R. et al. An assessment of progress towards universal health coverage in Brazil, Russia, India, China, and South Africa (BRICS). *The Lancet*, v. 384, n. 9960, p. 2164-2171, 2014.

PARTINGTON, A.; WYNN, M.; SURIADI, S.; OUYANG, C.; KARNON, J. Process Mining for Clinical Processes. **ACM Transactions on Management Information Systems**, [S.l.], v. 5, n. 4, p. 1-18, 2015.

PORTER, M. E.; KAPLAN, R. S. **How should we pay for health care?** Working Paper No 15-041. Harvard Business School, 2015.

PYTHON FOUNDATION. **Python Software Foundation**. Disponível em: <https://www.python.org/psf/>. Acesso em: 15 setembro 2017.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, [S.l.], v. 1, n. 1, p. 81-106, 1986.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Mateo, CA: Morgan Kaufmann, 1993.

R FOUNDATION. **R: The R Project for Statistical Computing**. Disponível em: <https://www.r-project.org/>. Acesso em: 1 novembro 2017.

REBUGE, Á.; FERREIRA, D. R. Business process analysis in healthcare environments: a methodology based on process mining. **Information Systems**, [S.l.], v. 37, n. 2, p. 99-116, 2012.

ROJAS, E.; MUNOZ-GAMA, J.; SEPÚLVEDA, M.; CAPURRO, D. Process mining in healthcare: a literature review. **Journal of Biomedical Informatics**, [S.l.], v. 61, p. 224-236, 2016.

SONG, M.; AALST, W. M. van der. Supporting process mining by showing events at a glance. In: ANNUAL WORKSHOP ON INFORMATION TECHNOLOGIES AND SYSTEMS (WITS), 17., 2007. **Proceedings**. [S.l.: s.n.], 2007. p. 139-145.

SPRUIT, M.; VROON, R.; BATENBURG, R. Towards healthcare business intelligence in long-term care. **Computers in Human Behavior**, [S.l.], v. 30, p. 698-707, 2014.

SURIADI, S.; WYNN, M. T.; OUYANG, C.; HOFSTEDDE, A. H. M. ter; DIJK, N. J. van. Understanding Process Behaviours in a Large Insurance Company in Australia: a case study. In: **Advanced Information Systems Engineering**. [S.l.]: Springer Berlin Heidelberg, 2013. p. 449-464.

TSUMOTO, S.; IWATA, H.; HIRANO, S.; TSUMOTO, Y. Similarity-based behavior and process mining of medical practices. **Future Generation Computer Systems**, [S.l.], v. 33, p. 21-31, 2014.

VEIGA, G. M.; FERREIRA, D. R. Understanding Spaghetti Models with Sequence Clustering for ProM. In: BUSINESS PROCESS MANAGEMENT WORKSHOPS, 2010, Berlin, Heidelberg. **Anais...** Springer Berlin Heidelberg, 2010. p. 92–103.

WEERDT, D. J.; VANDEN BROUCKE, S. K.; VANTHIENEN, J.; BAESENS, B. Leveraging process discovery with trace clustering and text mining for intelligent analysis of incident management processes. In: EVOLUTIONARY COMPUTATION (CEC), 2012 IEEE CONGRESS ON, 2012. **Anais...** [S.l.: s.n.], 2012. p. 1-8.

WEIJTERS, A.; DER AALST, W. M. van; DE MEDEIROS, A. A. Process mining with the heuristics miner-algorithm. **Technische Universiteit Eindhoven, Tech. Rep. WP**, [S.l.], v. 166, p. 1-34, 2006.

WESKE, M. **Business Process Management: concepts, languages, architectures**. [S.l.]: Springer Berlin Heidelberg, 2012.

WHITE, S. A. Introduction to BPMN. **IBM Cooperation**, [S.l.], v. 2, n. 0, p. 0, 2004.

WILLIAMS, S. J. **Improving Healthcare Operations**. [S.l.]: Springer International Publishing, 2017.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Burlington, MA: Morgan Kaufmann, 2011.

YOO, S.; CHO, M.; KIM, E.; KIM, S.; SIM, Y.; YOO, D.; HWANG, H.; SONG, M. Assessment of hospital processes using a process mining technique: outpatient process analysis at a tertiary hospital. **International Journal of Medical Informatics**, [S.l.], v. 88, p. 34-43, 2016.

ZAKI, M.; MEIRA, W. **Data Mining and Analysis: fundamental concepts and algorithms**. [S.l.]: Cambridge University Press, 2014.

ZHENG, Y. Trajectory Data Mining. **ACM Transactions on Intelligent Systems and Technology**, [S.l.], v. 6, n. 3, p. 1-41, 2015.

ANEXO A CONJUNTO DE REGRAS A

#	Regra	Então	Suporte	Confiança
1	Se Pendência=Pendências bloco	Então Setor=Centro Cirúrgico	0,153	0,945
2	Se Tipo do procedimento=Clínico E Setor=H5	Então Caráter da internação=Urgente	0,133	0,937
3	Se Especialidade Médica=Obstétrica/Ginecologia E Setor=H3 - Maternidade	Então Tipo do procedimento=Clínico&Cirúrgico	0,125	0,930
4	Se Setor=Centro Cirúrgico E Tempo internação=2	Então Tipo do procedimento=Cirúrgico	0,125	0,926
5	Se Tipo do procedimento=Cirúrgico E Tempo internação=2	Então Caráter da internação=Programada	0,250	0,912
6	Se Especialidade Médica=Obstétrica/Ginecologia E Tempo internação=3	Então Tipo do procedimento=Clínico&Cirúrgico	0,104	0,912
7	Se Tipo do procedimento=Cirúrgico E Tempo auditoria=5-10	Então Caráter da internação=Programada	0,142	0,904
8	Se Tipo do procedimento=Clínico E Tempo faturamento=5-10	Então Caráter da internação=Urgente	0,115	0,900
9	Se Tipo do procedimento=Cirúrgico E Setor=H6	Então Caráter da internação=Programada	0,274	0,896
10	Se Tipo do procedimento=Clínico	Então Caráter da internação=Urgente	0,252	0,893
11	Se Setor=Centro Cirúrgico E Tempo internação=2	Então Caráter da internação=Programada	0,121	0,893
12	Se Tipo do procedimento=Cirúrgico E Setor=Centro Cirúrgico	Então Caráter da internação=Programada	0,167	0,893
13	Se Tipo do procedimento=Cirúrgico	Então Caráter da internação=Programada	0,362	0,891
14	Se Setor=H6 E Tempo internação=2	Então Caráter da internação=Programada	0,206	0,891
15	Se Especialidade Médica=Ortopedia/Traumatologia	Então Caráter da internação=Programada	0,107	0,884
16	Se Setor=H6 E Tempo internação=2	Então Tipo do procedimento=Cirúrgico	0,202	0,877
17	Se Setor=Centro Cirúrgico E Setor=H6	Então Caráter da internação=Programada	0,173	0,866
18	Se Setor=H5	Então Caráter da internação=Urgente	0,168	0,837
19	Se Pendência=Pendências bloco E Setor=Centro Cirúrgico	Então Caráter da internação=Programada	0,128	0,836
20	Se Tempo auditoria=5-10 E Setor=H6	Então Caráter da internação=Programada	0,139	0,833
21	Se Tempo auditoria=5-10 E Tempo internação=2	Então Caráter da internação=Programada	0,108	0,821
22	Se Setor=Centro Cirúrgico	Então Caráter da internação=Programada	0,225	0,821
23	Se Especialidade Médica=Cirurgia Geral E Caráter da internação=Programada	Então Setor=H6	0,120	0,821
24	Se Tempo auditoria=5-10 E Tempo internação=2	Então Tipo procedimento=Cirúrgico	0,107	0,817
25	Se Pendência=Pendências bloco	Então Caráter da internação=Programada	0,132	0,814

26	Se	Tempo auditoria=0-5 E Tempo internação=2	Então	Caráter da internação=Programada	0,120	0,802
27	Se	Setor=H6	Então	Caráter da internação=Programada	0,361	0,798
28	Se	Tempo faturamento=5-10 E Tempo internação=2	Então	Tipo procedimento=Cirúrgico	0,129	0,797
29	Se	Tempo faturamento=5-10 E Tempo internação=2	Então	Caráter da internação=Programada	0,129	0,797
30	Se	Tempo internação=2	Então	Caráter da internação=Programada	0,280	0,793
31	Se	Especialidade Médica=Cirurgia Geral E Tipo do procedimento=Cirúrgico	Então	Setor=H6	0,104	0,788
32	Se	Especialidade Médica=Obstetrícia/Ginecologia	Então	Tipo do procedimento=Clínico&Cirúrgico	0,143	0,786
33	Se	Tempo auditoria=0-5 E Tempo internação=2	Então	Tipo do procedimento=Cirúrgico	0,116	0,776
34	Se	Tempo internação=2	Então	Tipo do procedimento=Cirúrgico	0,274	0,776
35	Se	Pendência=Pendências bloco E Setor=Centro Cirúrgico	Então	Setor=H6	0,118	0,774
36	Se	Tipo do procedimento=Cirúrgico E Setor=Centro Cirúrgico	Então	Setor=H6	0,144	0,773
37	Se	Especialidade Médica=Cirurgia Geral	Então	Setor=H6	0,152	0,760
38	Se	Tipo do procedimento=Cirúrgico E Tempo auditoria=5-10	Então	Setor=H6	0,119	0,759
39	Se	Tipo do procedimento=Cirúrgico	Então	Setor=H6	0,306	0,754
40	Se	Pendência=Pendências bloco	Então	Setor=H6	0,122	0,752

ANEXO B CONJUNTO DE REGRAS B

#	Regra		Suporte	Confiança
1	Se Tipo do procedimento=Cirúrgico E Pendência=Pendências bloco	Então Setor=Centro Cirúrgico	0,090	0,978
2	Se Tipo do procedimento=Clínico E Setor=H5	Então Caráter da internação=Urgente	0,147	0,967
3	Se Especialidade Médica=Obstetrícia/Ginecologia E Tempo auditoria=20-30	Então Tipo do procedimento=Clínico&Cirugico	0,075	0,961
4	Se Setor=H3 - Maternidade E Tempo internação=3	Então Tipo do procedimento=Clínico&Cirugico	0,117	0,958
5	Se Especialidade Médica=Obstetrícia/Ginecologia E Tempo internação=3	Então Tipo do procedimento=Clínico&Cirugico	0,125	0,953
6	Se Setor=H3 - Maternidade E Tempo internação=3	Então Especialidade Médica=Obstetrica/Ginecologia	0,116	0,950
7	Se Especialidade Médica=Obstetrícia/Ginecologia E Setor=H3 - Maternidade	Então Tipo do procedimento=Clínico&Cirugico	0,162	0,947
8	Se Especialidade Médica=Obstetrícia/Ginecologia E Tempo faturamento=5-10	Então Tipo do procedimento=Clínico&Cirugico	0,089	0,946
9	Se Pendência=Pendências bloco E Caráter da internação=Programada	Então Setor=Centro Cirúrgico	0,109	0,930
10	Se Tipo do procedimento=Clínico	Então Caráter da internação=Urgente	0,246	0,910
11	Se Setor=Centro Cirúrgico E Tempo internação=2	Então Tipo do procedimento=Cirurgico	0,090	0,908
12	Se Especialidade Médica=Obstetrícia/Ginecologia E Caráter da internação=Urgente	Então Tipo do procedimento=Clínico&Cirugico	0,110	0,893
13	Se Tempo faturamento=5-10 E Setor=H3 - Maternidade	Então Especialidade Médica=Obstetrica/Ginecologia	0,082	0,890
14	Se Tempo faturamento=5-10 E Setor=H3 - Maternidade	Então Tipo do procedimento=Clínico&Cirugico	0,082	0,890
15	Se Pendência=Pendências bloco E Setor=H6	Então Setor=Centro Cirúrgico	0,110	0,885
16	Se Pendência=Assinatura E Setor=H3 - Maternidade	Então Especialidade Médica=Obstetrica/Ginecologia	0,073	0,878
17	Se Pendência=Assinatura E Setor=H3 - Maternidade	Então Tipo do procedimento=Clínico&Cirugico	0,073	0,878
18	Se Especialidade Médica=Ortopedia/Traumatologia E Tipo do procedimento=Cirúrgico	Então Caráter da internação=Programada	0,083	0,872
19	Se Tipo do procedimento=Cirúrgico E Setor=H6	Então Caráter da internação=Programada	0,196	0,869
20	Se Tipo do procedimento=Cirúrgico E Pendência=Assinatura	Então Setor=Centro Cirúrgico	0,072	0,866
21	Se Pendência=Pendências bloco	Então Setor=Centro Cirúrgico	0,155	0,864
22	Se Tipo do procedimento=Cirúrgico E Tempo internação=2	Então Caráter da internação=Programada	0,148	0,864
23	Se Tipo do procedimento=Cirúrgico E Tempo auditoria=5-10	Então Caráter da internação=Programada	0,077	0,864
24	Se Especialidade Médica=obstétrica/Ginecologia E Caráter da internação=Urgente	Então Setor=H3 - Maternidade	0,105	0,860
25	Se Setor=H6 E Tempo internação=2	Então Tipo do procedimento=Cirurgico	0,120	0,855

26	Se	Especialidade Médica=obstétrica/Ginecologia	Então	Tipo do procedimento=Clínico&Cirugico	0,181	0,852
27	Se	Setor=H3 - Maternidade	Então	Tipo do procedimento=Clínico&Cirugico	0,172	0,850
28	Se	Setor=H3 - Maternidade	Então	Especialidade Médica=Obstetrícia/Ginecologia	0,171	0,845
29	Se	Tipo do procedimento=Cirúrgico	Então	Caráter da internação=Programada	0,257	0,841
30	Se	Setor=H6 E Tempo internação=2	Então	Caráter da internação=Programada	0,117	0,833
31	Se	Setor=H5	Então	Caráter da internação=Urgente	0,189	0,827
32	Se	Setor=Centro Cirúrgico E Tempo internação=2	Então	Caráter da internação=Programada	0,081	0,816
33	Se	Especialidade Médica=Ortopedia/Traumatologia E Setor=H6	Então	Caráter da internação=Programada	0,091	0,811
34	Se	Especialidade Médica=Ortopedia/Traumatologia	Então	Caráter da internação=Programada	0,122	0,784
35	Se	Especialidade Médica=Ortopedia/Traumatologia E Tipo do procedimento=Cirúrgico	Então	Setor=H6	0,074	0,777
36	Se	Tempo internação=5-10	Então	Caráter da internação=Urgente	0,110	0,771
37	Se	Especialidade Médica=Cirurgia Geral E Setor=H6	Então	Caráter da internação=Programada	0,086	0,766
38	Se	Pendência=Assinatura E Caráter da internação=Programada	Então	Setor=Centro Cirúrgico	0,092	0,758
39	Se	Setor=Centro Cirúrgico E Setor=H6	Então	Caráter da internação=Programada	0,149	0,758
40	Se	Pendência=Assinatura E Tempo internação=3	Então	Tipo do procedimento=Clínico&Cirugico	0,073	0,750

ANEXO C QUESTIONÁRIO

Questionário de avaliação:

O presente questionário visa avaliar um conjunto de regras geradas a partir de dados históricos do processo de reembolso das contas dos pacientes do hospital. Espera-se que, com o conjunto de regras, seja possível identificar fatores que contribuam para o atraso no processo de reembolso.

As regras estão apresentadas no seguinte formato, “**Se** Pendência=Pendências bloco **Então** Setor=Centro Cirúrgico”. Esta regra evidencia que uma pendência denominada de “Pendências bloco” ocorre, com certa frequência, no Centro Cirúrgico.

Os dados usados para a geração das regras são referentes ao processo de reembolso de pacientes internados nos anos de 2015 e 2016 por meio do convênio da Unimed. Já o tempo total do processo foi calculado usando a data da alta do paciente (início do processo) e a data envio da conta para o respectivo convênio (fim do processo).

1) Em qual setor do hospital você trabalha atualmente e a quanto tempo aproximadamente?

- Auditoria
 Faturamento

Tempo (anos): _____

2) Escolha o grau de relevância de cada regra que foi gerada a partir dos dados dos processos que duraram de 20 até 30 dias?

*Assinale um valor para cada regra, onde:

1 - Nenhuma Relevância; 2 - Pouca Relevância; 3 - Média Relevância; 4 – Relevante; 5 - Muito relevante;

#	Regra	Então	Relevância	1	2	3	4	5
1	Se Pendência=Pendências bloco	Então Setor=Centro Cirúrgico						
2	Se Tipo do procedimento=Clinico E Setor=H5	Então Caráter da internação=Urgente						
3	Se Especialidade Médica=Obstétrica/Ginecologia E Setor=H3 - Maternidade	Então Tipo do procedimento=Clinico&Cirúrgico						
4	Se Setor=Centro Cirúrgico E Tempo internação=2	Então Tipo do procedimento=Cirúrgico						
5	Se Tipo do procedimento=Cirúrgico E Tempo internação=2	Então Caráter da internação=Programada						
6	Se Especialidade Médica=Obstétrica/Ginecologia E Tempo internação=3	Então Tipo do procedimento=Clinico&Cirúrgico						
7	Se Tipo do procedimento=Cirúrgico E Tempo auditoria=5-10	Então Caráter da internação=Programada						
8	Se Tipo do procedimento=Clinico E Tempo faturamento=5-10	Então Caráter da internação=Urgente						
9	Se Tipo do procedimento=Cirúrgico E Setor=H6	Então Caráter da internação=Programada						
10	Se Tipo do procedimento=Clinico	Então Caráter da internação=Urgente						
11	Se Setor=Centro Cirúrgico E Tempo internação=2	Então Caráter da internação=Programada						
12	Se Tipo do procedimento=Cirúrgico E Setor=Centro Cirúrgico	Então Caráter da internação=Programada						
13	Se Tipo do procedimento=Cirúrgico	Então Caráter da internação=Programada						
14	Se Setor=H6 E Tempo internação=2	Então Caráter da internação=Programada						
15	Se Especialidade Médica=Ortopedia/Traumatologia	Então Caráter da internação=Programada						
16	Se Setor=H6 E Tempo internação=2	Então Tipo do procedimento=Cirúrgico						
17	Se Setor=Centro Cirúrgico E Setor=H6	Então Caráter da internação=Programada						

18	Se	Setor=H5	Então	Caráter da internação=Urgente	1	2	3	4	5
19	Se	Pendência=Pendências bloco E Setor=Centro Cirúrgico	Então	Caráter da internação=Programada	1	2	3	4	5
20	Se	Tempo auditoria=5-10 E Setor=H6	Então	Caráter da internação=Programada	1	2	3	4	5
21	Se	Tempo auditoria=5-10 E Tempo internação=2	Então	Caráter da internação=Programada	1	2	3	4	5
22	Se	Setor=Centro Cirúrgico	Então	Caráter da internação=Programada	1	2	3	4	5
23	Se	Especialidade Médica=Cirurgia Geral E Caráter da internação=Programada	Então	Setor=H6	1	2	3	4	5
24	Se	Tempo auditoria=5-10 E Tempo internação=2	Então	Tipo procedimento=Cirúrgico	1	2	3	4	5
25	Se	Pendência=Pendências bloco	Então	Caráter da internação=Programada	1	2	3	4	5
26	Se	Tempo auditoria=0-5 E Tempo internação=2	Então	Caráter da internação=Programada	1	2	3	4	5
27	Se	Setor=H6	Então	Caráter da internação=Programada	1	2	3	4	5
28	Se	Tempo faturamento=5-10 E Tempo internação=2	Então	Tipo procedimento=Cirúrgico	1	2	3	4	5
29	Se	Tempo faturamento=5-10 E Tempo internação=2	Então	Caráter da internação=Programada	1	2	3	4	5
30	Se	Tempo internação=2	Então	Caráter da internação=Programada	1	2	3	4	5
31	Se	Especialidade Médica=Cirurgia Geral E Tipo do procedimento=Cirúrgico	Então	Setor=H6	1	2	3	4	5
32	Se	Especialidade Médica=Obstetria/Ginecologia	Então	Tipo do procedimento=Clínico&Cirúrgico	1	2	3	4	5
33	Se	Tempo auditoria=0-5 E Tempo internação=2	Então	Tipo do procedimento=Cirúrgico	1	2	3	4	5
34	Se	Tempo internação=2	Então	Tipo do procedimento=Cirúrgico	1	2	3	4	5
35	Se	Pendência=Pendências bloco E Setor=Centro Cirúrgico	Então	Setor=H6	1	2	3	4	5
36	Se	Tipo do procedimento=Cirúrgico E Setor=Centro Cirúrgico	Então	Setor=H6	1	2	3	4	5
37	Se	Especialidade Médica=Cirurgia Geral	Então	Setor=H6	1	2	3	4	5
38	Se	Tipo do procedimento=Cirúrgico E Tempo auditoria=5-10	Então	Setor=H6	1	2	3	4	5
39	Se	Tipo do procedimento=Cirúrgico	Então	Setor=H6	1	2	3	4	5
40	Se	Pendência=Pendências bloco	Então	Setor=H6	1	2	3	4	5

3) Entre as regras definidas como Relevante ou Muito Relevante na questão anterior, alguma delas representa um fato desconhecido, se sim quais?

4) Escolha o grau de relevância de cada regra que foi gerada a partir dos dados dos processos que duraram de 30 até 40 dias?

*Assinale um valor para cada regra, onde:

1 - Nenhuma Relevância; 2 - Pouca Relevância; 3 - Média Relevância; 4 – Relevante; 5 - Muito relevante;

#	Regra				Relevância				
	Se	Então	Caráter da internação	Setor	1	2	3	4	5
1	Se	Tipo do procedimento=Cirúrgico E Pendência=Pendências bloco	Então	Setor=Centro Cirúrgico	1	2	3	4	5
2	Se	Tipo do procedimento=Clínico E Setor=H5	Então	Caráter da internação=Urgente	1	2	3	4	5
3	Se	Especialidade Médica=Obstetria/Ginecologia E Tempo auditoria=20-30	Então	Tipo do procedimento=Clínico&Cirúrgico	1	2	3	4	5
4	Se	Setor=H3 - Maternidade E Tempo internação=3	Então	Tipo do procedimento=Clínico&Cirúrgico	1	2	3	4	5
5	Se	Especialidade Médica=Obstetria/Ginecologia E Tempo internação=3	Então	Tipo do procedimento=Clínico&Cirúrgico	1	2	3	4	5
6	Se	Setor=H3 - Maternidade E Tempo internação=3	Então	Especialidade Médica=Obstetria/Ginecologia	1	2	3	4	5
7	Se	Especialidade Médica=Obstetria/Ginecologia E Setor=H3 - Maternidade	Então	Tipo do procedimento=Clínico&Cirúrgico	1	2	3	4	5
8	Se	Especialidade Médica=Obstetria/Ginecologia E Tempo faturamento=5-10	Então	Tipo do procedimento=Clínico&Cirúrgico	1	2	3	4	5
9	Se	Pendência=Pendências bloco E Caráter da internação=Programada	Então	Setor=Centro Cirúrgico	1	2	3	4	5
10	Se	Tipo do procedimento=Clínico	Então	Caráter da internação=Urgente	1	2	3	4	5

11	Se	Setor=Centro Cirúrgico E Tempo internação=2	Então	Tipo do procedimento=Cirurgico	1	2	3	4	5
12	Se	Especialidade Médica=Obstetria/Ginecologia E Caráter da internação=Urgente	Então	Tipo do procedimento=Clínico&Cirurgico	1	2	3	4	5
13	Se	Tempo faturamento=5-10 E Setor=H3 - Maternidade	Então	Especialidade Médica=Obstetria/Ginecologia	1	2	3	4	5
14	Se	Tempo faturamento=5-10 E Setor=H3 - Maternidade	Então	Tipo do procedimento=Clínico&Cirurgico	1	2	3	4	5
15	Se	Pendência=Pendências bloco E Setor=H6	Então	Setor=Centro Cirúrgico	1	2	3	4	5
16	Se	Pendência=Assinatura E Setor=H3 - Maternidade	Então	Especialidade Médica=Obstetria/Ginecologia	1	2	3	4	5
17	Se	Pendência=Assinatura E Setor=H3 - Maternidade	Então	Tipo do procedimento=Clínico&Cirurgico	1	2	3	4	5
18	Se	Especialidade Médica=Ortopedia/Traumatologia E Tipo do procedimento=Cirurgico	Então	Caráter da internação=Programada	1	2	3	4	5
19	Se	Tipo do procedimento=Cirurgico E Setor=H6	Então	Caráter da internação=Programada	1	2	3	4	5
20	Se	Tipo do procedimento=Cirurgico E Pendência=Assinatura	Então	Setor=Centro Cirúrgico	1	2	3	4	5
21	Se	Pendência=Pendências bloco	Então	Setor=Centro Cirúrgico	1	2	3	4	5
22	Se	Tipo do procedimento=Cirurgico E Tempo internação=2	Então	Caráter da internação=Programada	1	2	3	4	5
23	Se	Tipo do procedimento=Cirurgico E Tempo auditoria=5-10	Então	Caráter da internação=Programada	1	2	3	4	5
24	Se	Especialidade Médica=obstétrica/Ginecologia E Caráter da internação=Urgente	Então	Setor=H3 - Maternidade	1	2	3	4	5
25	Se	Setor=H6 E Tempo internação=2	Então	Tipo do procedimento=Cirurgico	1	2	3	4	5
26	Se	Especialidade Médica=obstétrica/Ginecologia	Então	Tipo do procedimento=Clínico&Cirurgico	1	2	3	4	5
27	Se	Setor=H3 - Maternidade	Então	Tipo do procedimento=Clínico&Cirurgico	1	2	3	4	5
28	Se	Setor=H3 - Maternidade	Então	Especialidade Médica=Obstetria/Ginecologia	1	2	3	4	5
29	Se	Tipo do procedimento=Cirurgico	Então	Caráter da internação=Programada	1	2	3	4	5
30	Se	Setor=H6 E Tempo internação=2	Então	Caráter da internação=Programada	1	2	3	4	5
31	Se	Setor=H5	Então	Caráter da internação=Urgente	1	2	3	4	5
32	Se	Setor=Centro Cirúrgico E Tempo internação=2	Então	Caráter da internação=Programada	1	2	3	4	5
33	Se	Especialidade Médica=Ortopedia/Traumatologia E Setor=H6	Então	Caráter da internação=Programada	1	2	3	4	5
34	Se	Especialidade Médica=Ortopedia/Traumatologia	Então	Caráter da internação=Programada	1	2	3	4	5
35	Se	Especialidade Médica=Ortopedia/Traumatologia E Tipo do procedimento=Cirurgico	Então	Setor=H6	1	2	3	4	5
36	Se	Tempo internação=5-10	Então	Caráter da internação=Urgente	1	2	3	4	5
37	Se	Especialidade Médica=Cirurgia Geral E Setor=H6	Então	Caráter da internação=Programada	1	2	3	4	5
38	Se	Pendência=Assinatura E Caráter da internação=Programada	Então	Setor=Centro Cirúrgico	1	2	3	4	5
39	Se	Setor=Centro Cirúrgico E Setor=H6	Então	Caráter da internação=Programada	1	2	3	4	5
40	Se	Pendência=Assinatura E Tempo internação=3	Então	Tipo do- procedimento=Clínico&Cirurgico	1	2	3	4	5

5) Entre as regras definidas como Relevante ou Muito Relevante na questão anterior, alguma delas representa um fato desconhecido, se sim quais?

6) Na sua opinião as regras possam ser usadas na tomada de decisão com o objetivo de melhorar a eficiência do processo de reembolso e porquê?
