UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÉNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Mineração de Dados para Predição de Fracassos em Processos de Negócio

Pedro Otávio Teixeira Mello

Orientadores
Até o dia 15 de agosto de 2018: Kate Cerqueira Revoredo
Flávia Maria Santoro

Do dia 15 de agosto de 2018 até a defesa: Gleison dos Santos Souza

RIO DE JANEIRO, RJ - BRASIL
Outubro de 2018

Mineração de dados para predição de fracassos em processos de negócio

Pedro Otávio Teixeira Mello

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Gleison dos Santos Souza, D.Sc. - UNIRIO

Carlos Eduardo Ribeiro de Mello, Ph.D. - UNIRIO

Jonice de Oliveira Sampaio, D.Sc. - UFRJ

RIO DE JANEIRO, RJ - BRASIL

Outubro de 2018

Teixeira Mello, Pedro Otávio.
T527 Mineração de dados para predição de fracassos em processos de negócio.
/ Pedro Otávio Teixeira Mello. – Rio de Janeiro, 2018.
89 f.

Orientadora: Kate Cerqueira Revoredo.
Coorientadora: Flávia Maria Santoro.
Dissertação (Mestrado) – Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2018.

1. Mineração de Dados. 2. Predição. 3. Processos de Negócio. 4. Monitoramento de Processos. I. Revoredo, Kate Cerqueira, orient. II. Santoro, Flávia Maria. III. Título.

Em memória aos meus avós e pais
Doracy Mello dos Santos e Genil Dias Mello

Agradecimentos

Meus primeiros agradecimentos são destinados às minhas orientadoras Kate Revoredo e Flávia Santoro. Agradeço por toda a dedicação e por todo o aprendizado que este aluno recebeu. Imensurável é a gratidão que sinto por ter trilhado esta jornada com vocês.

Agradeço à professora Fernanda Baião por ter participado de todas as minhas apresentações durantes os seminários. Escutei, li, reli, guardei e transformei suas palavras em aprendizado e também em caminhos para a pesquisa.

Um agradecimento ao amigo que encontrei durante essa jornada, André de Souza Andrade que me ajudou muito com palavras de motivação e pelos momentos de construção filosófica acerca da existência e do ser. Também agradeço aos membros do grupo "Chama o Le-Boudec". Destaco que essas amizades há muito haviam se consumado, descrito em outras estrelas e rompendo o espaço-tempo até este plano e esta existência.

Agradeço ao meu também orientador Gleison Santos que, em meio ao meu próprio caos, encontrou algumas poucas boas versões de mim e me ajudou a chegar na defesa e nesta dissertação. Agradeço também aos membros da banca presentes. Em primeiro, agradeço à professora Jonice Oliveira pelas sábias palavras proferidas, me dando motivos e razões para eternizar este documento, onde, cujas palavras me fez lembrar a seguinte menção dada por Jorge Luis Borges: "... a biblioteca perdurará: iluminada, solitária, infinita, perfeitamente imóvel, armada de volumes preciosos, ...". Ao professor Carlos Eduardo Ribeiro de Mello, pela também presença na banca e também pela conversa e apoio que recebi alguns dias antes da defesa. Essa conversa me deu a confiança e a certeza que um bom trabalho havia se realizado.

Agradeço a CAPES pelo fomento a este trabalho de pesquisa.

Agradeço também aos amigos e familiares presentes e ausentes. Também deixo eternizado a minha saudade por aqueles que já não estão mais presentes e que ainda habitam nas minhas memórias, nas minhas lembranças, na minha história, ...

MELLO, PEDRO O. T. **Mineração de Dados para Predição de Fracassos em Processos de Negócio**. UNIRIO, 2018. 89 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

O monitoramento de processos de negócio emerge sob diversas necessidades e uma delas é garantir a execução dos processos de maneira a manter a confiabilidade de suas execuções. Entretanto, os processos de negócios estão sujeitos a uma natureza dinâmica o que dificulta a transição entre um cenário reativo para um cenário preditivo. Em um cenário reativo, as ações de mitigação de riscos só acontecem após a ocorrência dos fatos que colocam a execução do processo em risco. Por outro lado, um cenário preditivo pode estimar os riscos de uma execução de um processo. Dito isso, um cenário preditivo pode apoiar uma estratégia de mitigação de riscos de maneira antecipada. Para isso são necessárias algumas premissas tais como a identificação de situações e padrões em dados históricos dos processos de maneira a caracterizar o que determina o fracasso dos mesmos. Nesta dissertação, endereçamos o problema de como identificar e detectar de maneira preditiva os comportamentos que podem levar os processos ao fracasso. Neste trabalho, uma combinação de técnicas bem estabelecidas das áreas de Mineração de Dados e Mineração de Processos são utilizadas e aplicadas em um estudo de caso de um processo de gerenciamento de incidentes. A avaliação experimental da pesquisa é apresentada como uma forma de provar que a combinação de diferentes técnicas é aplicável ao estudo de caso de forma satisfatória considerando um cenário preditivo. As principais contribuições destacam-se em termos dos resultados alcançados a partir da combinação dessas técnicas somado as adaptações que foram feitas de maneira a se adequarem ao estudo de caso.

Palavras-chave: Mineração de Dados · Predição · Processos de Negócio · Monitoramento de Processos.

MELLO, PEDRO O. T. **Data Mining for Failure Prediction in Business Processes**. UNIRIO, 2018. 89 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

ABSTRACT

Business process monitoring emerges under various needs and one of them is to ensure the execution of processes in order to maintain the reliability of their executions. However, business processes are subject to a dynamic nature which hinders the transition from a reactive scenario to a predicted scenario. In a reactive scenario, risk mitigation actions only occur after the occurrence of the facts that put the execution of the process at risk. On the other hand, a predictive scenario can estimate the risks of process execution. In view of that, a predictive scenario can support a risk mitigation strategy in advance. For this, some premises are necessary such as the identification of situations and patterns in historical data of the processes in order to characterize what determines their failure. In this dissertation, we address the problem of how to identify and detect in a predictive way the behaviors that can lead the processes to a failure situation. In this master dissertation, a combination of well-established techniques from the Data Mining and Process Mining fields are used and applied in a case study of an incident management process. The experimental evaluation of the research is presented as a way to prove that the combination of different techniques is applicable to the case study in a satisfactory way considering a predictive scenario. The main contributions stand out in terms of the results obtained from the combination of these techniques added to the adaptations that were made so as to fit the case study.

Keywords: Data Mining · Prediction · Business Process · Process Monitoring

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

BP	BUSINESS PROCESS
BPM	BUSINESS PROCESS MANAGEMENT
DM	DATA MINING
DT	DECISION TREE
GBT	GRADIENT BOOSTING TREES
KDD	KNOWLEDGE DATABASE DISCOVERY
KPI	KEYS PERFORMANCE INDICATORS
NB	NAIVE BAYES
ML	MACHINE LEARNING
PFA	PROBABILISTIC FINIT AUTOMATON
PM	PROCESS MINING
PPI	PROCESS PERFORMANCE INDICATORS
QP	QUESTÃO DE PESQUISA
RF	RANDOM FOREST
SLA	SERVICE LEVEL AGREEMENT
TI	TECNOLOGIA DA INFORMAÇÃO

1. Introdução

Se você não pode medir, não pode gerenciar. É com esta citação a Peter Drucker que destaca-se a importância que a área de Mineração de Processos (PM, do inglês *Process Mining*) pode exercer sobre a área de Gerenciamento de Processos de Negócio (BPM, do inglês *Business Process Management*). A área de Mineração de Processos tem como seus principais objetivos descobrir, monitorar e aprimorar processos de negócio por meio da extração de conhecimento a partir de dados [50]. Os processos de negócio, por sua vez, visam ao alcance dos objetivos estratégicos e operacionais de um ambiente organizacional por meio de atividades coordenadas [51].

Neste contexto, algumas medidas podem desempenhar um papel importante para o controle e otimização desses objetivos e, por meio dessas medidas, é possível determinar se as instâncias[1] do processo se encontram em uma situação de fracasso, isto é, uma situação em que as instâncias não alcançam os objetivos almejados do processo e que podem apresentar risco ao negócio.

A dificuldade em controlar e gerenciar processos de negócio está associada a dois fatores: (*i*) a natureza dinâmica à qual os processos de negócio estão sujeitos, que resulta em um ambiente de alta variação; e (*ii*) o desconhecimento de situações, por parte de analistas e gestores, que exercem influências não só no desempenho do processo, mas também no fluxo de atividades. Esses dois fatores tornam as análises que visam a ações proativas e corretivas do processo um pouco mais complexas.

Em meio a essa complexidade, diversas técnicas de mineração de processos têm sido propostas pelo estado-da-arte para aprimorar o gerenciamento de processos, sejam em cenários que visam à melhoria de processos de negócio ou em

[1] As instâncias de processo também podem ser referenciadas como processos em execução ou já executados.

cenários de diagnóstico do processo. Cenários de diagnóstico do processo visam buscar explicações para a existência de determinadas características das instâncias do processo ou sejam em cenários preditivos onde a premissa é sair de uma abordagem reativa para uma abordagem proativa. A principal diferença entre as duas abordagens é que a abordagem reativa consiste na tomada de decisão e ações após a ocorrência de um fato. Por outro lado, em uma abordagem proativa, a tomada de decisão é feita de maneira antecipada ao fato, isto é, antes de sua ocorrência. Assim, uma abordagem proativa é possível por meio de estimativas de uma realidade empírica.

Este capítulo apresenta, na Seção 1.1, uma breve contextualização do tema de pesquisa e principais motivações. O problema e a questão de pesquisa formulada somado aos principais objetivos que guiou o processo evolutivo desta dissertação são expostos na Seção 1.2. Apresenta também a metodologia científica na Seção 1.3 e finaliza com a estrutura do documento apresentada de maneira geral na Seção 1.4.

1.1 Contextualização

Este trabalho de pesquisa está centrado na área de Mineração de Processos, que consiste em integrar técnicas e conceitos tanto da área de Gerenciamento de Processos de Negócio quanto da área de Mineração de Dados e Aprendizado de Máquina.

Tal como a Mineração de Dados, a área de Mineração de Processos consiste da extração de conhecimento a partir de dados e, neste caso, o ponto de partida de qualquer técnica de Mineração de Processos são os registros históricos de um processo de negócio denominado por *Event Log* [53]. Estes termos, além de outros, são tratadas com mais profundidade no Capítulo 2.

No ano de 2012, uma força tarefa sobre pesquisas referentes à Mineração de Processos identificou onze desafios em aberto e publicado em [50]. Esses desafios são:

1. **Encontrar, unificar e limpar.** Lida com problemas associados a variedade de fonte de dados e operações de união e junção de dados. Lida também com problemas relacionados a qualidade da informação presente nos dados, tais como falta de informação em determinados atributos e variáveis ou

informações erradas e confusas que não deveriam estar presentes;

2. **Dados históricos complexos e com diferentes características.** Lida com o volume de dados e variação que eles podem apresentar. Neste caso, os registros podem ter um grande volume ou pouco volume de informação associado a problemas de representatividade do comportamento do processo;

3. **Criação de marcas de referência (*Benchmarks*) representativas.** Lida com a falta de marcas de referência internacionalmente aceitas. Mineração de Processos ainda é considerada uma área emergente, diante disso ainda há uma carência por *benchmarks*;

4. **Mudança de conceito (*Concept Drift*).** Lida com a natureza dinâmica dos processos. Um processo de negócio pode ser impactado por fatores internos e externos que provocam mudanças na distribuição dos dados;

5. **Aprimoramento de representações dos modelos descobertos.** Lida com as dificuldades relacionadas às representações dos modelos descobertos. Em geral, as técnicas de descoberta de modelos de processo utilizam uma linguagem ou notação para representar a visualização do modelo. A escolha dessas formas de representações deve ser criteriosa, pois se um modelo não puder ser representado, então ele não poderá ser descoberto. Dito isso, a notação deve permitir formas de representar concorrência, múltiplas atividades, repetições de rótulos e nomes, encadeamento de atividades somado a outras características que um processo pode ter;

6. **Balanceamento entre critérios de qualidade.** Lida com as complicações relacionadas ao ajuste (*fitness*), simplicidade, precisão e generalização. Os históricos de eventos (*Event Logs*) estão longe de estar completos. Os modelos de processos tendem a um número muito grande, ou até mesmo infinito, de diferentes representações possíveis. Modelos bem ajustados implicam que o histórico pode ser repetível pelo modelo a partir de seu início. A simplicidade, por exemplo, torna-se eficiente para a compreensão do processo. A generalização surge a partir de um problema de precisão onde o modelo deve manter uma flexibilidade, pois um processo não pode ser muito restritivo enquanto também não deve permitir um comportamento excessivamente variado. A generalização visa não ser muito específica mas também não pode ser muito restritiva;

7. **Mineração e distinção de múltiplos ambientes organizacionais.** Lida com as multiplicidades de cenários que podem coexistir no histórico de eventos

ou estar fatiado em outros ambientes. Com o aprimoramento da tecnologia e das formas colaborativas de trabalho, um processo pode ser fatiado em partes e distribuído pela organização de maneira distinta, distribuindo até mesmo para outras organizações. Diante disso, o histórico de eventos pode estar incompleto, uma vez que outras partes do mesmo processo podem estar nas mãos de terceiros e não no mesmo histórico de eventos;

8. **Fornecimento de apoio operacional.** Lida com atividades relacionadas a detecção, predição e recomendação. Atividades de detecção podem identificar desvios de fluxo predefinido de um processo mapeado. Os registros históricos de um processo podem ser utilizados para a construção e obtenção de um modelo de predição que poderá levar abordagens reativas para abordagens proativas. Baseado nessas predições, a elaboração e construção de sistemas de recomendação tendem a propor ações de redução de custos ou redução de tempo baseando-se em uma otimização do fluxo operacional;

9. **Combinação das técnicas de Mineração de Processos com outros tipos de técnicas e análises.** Lida com a combinação de diferentes técnicas tanto de gerenciamento de processos quanto de mineração de dados visando aumentar a compreensão dos modelos e a capacidade de aprimoramento do negócio. Existe também uma necessidade de combinar essas técnicas com técnicas que aprimoram o raciocínio analítico, isto é, análise visual[2];

10. **Aprimoramento da usabilidade para não especialistas.** Linda com o aprimoramento entre as relações do modelo de processo com os dados. Por meio dos dados, as técnicas de mineração de processos criam uma projeção do comportamento do processo com o seu modelo. Atualmente, as técnicas aplicadas não são muito sugestivas para um usuário não especialista devido ao alto número de parâmetros e configurações adotadas aliado ao conhecimento *a priori* da técnica empregada;

11. **Aprimoramento da compreensão para não especialistas.** Lida com os problemas de compreensão do usuário final em relação aos resultados obtidos por meio das técnicas de mineração de processos. Busca-se por técnicas mais sofisticadas e representativas. Boas representações podem justificar determinadas conclusões e decisões tomadas.

[2]Este termo é mais comumente descrito em inglês como *Visual Analytics* e trata-se de uma área que visa aprimorar o raciocínio analítico por meio de gráficos e técnicas automatizadas de visualização e interação com os dados.

Durante o desenvolvimento desse trabalho de pesquisa, identificou-se que os onze desafios acima enumerados continuam em aberto. Entretanto, consideramos que muito já se evoluiu em termos de técnicas e aplicações. Tal evidência pode ser vista em diversos trabalhos encontrados em [9, 14, 40, 45, 58]. Apesar dessa evolução em termos de técnicas e aplicações, reconhecidas por meio dos trabalhos anteriormente citados, ainda há muito o que se explorar em torno dos onze desafios destacados por [50]. Muitas das técnicas aplicadas estão restritas a determinados domínios e cenários.

O domínio e principal objeto de estudo desta dissertação de mestrado foi o domínio de gerenciamento de incidentes, onde os processos destinam-se em restaurar e normalizar os serviços de *Tecnologia da Informação* (TI) o mais breve possível. Os processos de um cenário de gerenciamento de incidentes são tratados como o principal canal de comunicação entre diferentes setores de um ambiente organizacional com o setor de TI. Desta maneira, a premissa é que essa restauração rápida visa minimizar os impactos nas operações de negócio. Tal descrição é apresentada com maior profundidade na Seção 1.3 onde o cenário é descrito sob a forma de um estudo de caso. A próxima seção (Seção 1.2) destaca o problema de pesquisa que este trabalho se propõe a resolver em conjunto com os principais objetivos que guiou o processo evolutivo para o desenvolvimento do trabalho.

1.2 Problema e Objetivo de Pesquisa

O objetivo deste trabalho é a predição de fracassos em processos de negócio. Neste sentido, o fracasso é determinado por meio de uma situação que apresenta risco ao negócio. Essa situação é determinada pelo acordo a nível de serviço (SLA, do inglês *Service Level Agreement*). O SLA é um compromisso assumido por um prestador de serviços perante um cliente, uma vez que ele não é cumprido algumas penalidades podem ser aplicadas. Diante disso, o não cumprimento do SLA apresenta riscos ao negócio, podendo ocasionar em prejuízos financeiros para uma organização[3].

O estudo de predição aparece como um dos onze desafios descritos na Seção 1.1. Uma vez que a predição torna-se possível, assume-se que pode aumentar a confiabilidade dos sistemas de acordo com a dinamicidade requerida dos serviços prestados. Deste modo, lidar com o desafio que envolve a predição de alguma

[3]Em geral, a penalidade aplicada são multas que podem ser arcadas por meio de pagamentos a empresa lesada pelo não cumprimento do serviço ou redução do valor do contrato.

variável, que neste caso a variável considerada é o SLA, estamos defrontando com o aprimoramento de técnicas que lidam com as atividades de apoio operacional e possibilitando que abordagens proativas venham a surgir de maneira a mitigar riscos.

Este trabalho também pretende discutir a modelagem da predição, passando pelas variáveis consideradas e os critérios que levaram às respectivas escolhas até a aplicação do algoritmo de predição. Diante disso, o trabalho enfrenta também outros dos onze desafios, tais como o problema das fontes de dados e operações de união e junção de dados, as diferentes características do histórico de eventos, as marcas de referência, mudança de conceito (*Concept Drift*) e, por fim, a combinação de diferentes técnicas para o alcance do principal objetivo proposto que é a predição das situações de fracasso em processos de negócio.

Desta forma, este trabalho tem a seguinte **questão de pesquisa** (QP) a ser explorada:

> **[QP].** Como aprimorar e combinar as técnicas de predição e aplicar no domínio de gerenciamento de incidentes?

Nota-se que a questão de pesquisa formulada não exclui a existência de trabalhos que realizam predições em processos de negócio. Ao contrário, a questão de pesquisa considera propor uma melhoria, se necessário, dos modelos de predição atual por meio de possíveis combinações com outras técnicas.

Dito isso, entendemos que para alcançar o objetivo proposto este trabalho propõe utilizar e combinar diferentes técnicas de mineração de dados que foram aplicadas em outros domínios. Espera-se que essa combinação permita alcançar não só a predição das situações de fracasso mas também pode resultar em uma técnica presumivelmente melhor. De acordo com [55], uma técnica de pesquisa que visa a um resultado presumivelmente melhor trata-se de um estilo de pesquisa onde os resultados produzidos podem ser melhores do que os resultados descritos na literatura. As conclusões dessa melhoria se dá por meio de experimentos comparativos entre resultados e técnicas aplicadas. Este assunto é descrito na próxima seção pois trata da metodologia aplicada durante a pesquisa científica deste documento.

1.3 Metodologia de Pesquisa

Após a revisão da literatura e definição inicial do problema de pesquisa, definiu-se um domínio para aplicação do estudo de caso em que o método seria aplicado. Foi conduzido um estudo de caso exploratório em cima de dados reais de um processo de gerenciamento de incidentes. Os dados foram obtidos por meio de uma colaboração com o departamento de TI de uma empresa brasileira.

A premissa dos processos de gerenciamento de incidentes é a restauração de serviços de Tecnologia da Informação (TI) o mais breve possível, visando minimizar os impactos e efeitos negativos causados pela não oferta ou provimento de um determinado serviço, em geral, os processos de gerenciamento de incidentes são o principal canal de comunicação entre diversos setores e áreas distintas de um ambiente organizacional com o setor ou departamento de TI.

Os processos de gerenciamento de incidentes são conhecidos por serem processos flexíveis, isto é, tratam-se de processos de alta variação, o que implica que eles podem variar seu comportamento de acordo com alguma necessidade que venha a surgir. Tal necessidade pode acontecer durante a execução do processo e, neste caso, pode estar relacionada às decisões tomadas naquele instante.

O estudo de caso analisou um histórico de instâncias deste processo e, inicialmente, desejou-se descobrir as possíveis causas que podem caracterizar ou explicar o alto número de instâncias deste processo que estão diante de situações de fracasso, evidenciada pelo não cumprimento do SLA. Entretanto, considerando que o conteúdo que havia nos dados poderia não ser suficiente para descobrir as verdadeiras causas dos fracassos e o projeto deveria considerar uma análise qualitativa com os envolvidos do processo, o trabalho de pesquisa nesse cenário teria resultados explanativos e não preditivos. Sendo assim, a pesquisa é redirecionada com um enfoque para uma abordagem preditiva que envolve técnicas de exploração de dados, preparação dos dados e, por fim, a aplicação dos algoritmos de predição das situações de fracasso do processo de negócio.

Esta dissertação considera que não há *benchmarks* internacionalmente aceitos, como exposto na Seção 1.1. Dito isso, essa pesquisa testa abordagens da literatura com algumas adaptações para o estudo de caso. Em seguida, busca-se formas de aprimorar os resultados obtidos com o uso de técnicas combinadas. Ao final, os resultados são comparados em uma abordagem lado-a-lado.

Como contribuição, espera-se que o uso das técnicas combinadas possam melhorar os resultados da predição das situações de fracasso do processo de negócio além de abordar as adaptações realizadas dessas outras metodologias ao estudo de caso tratado por esta pesquisa.

1.4 Organização do trabalho

Por se tratar de termos técnicos da área de Mineração de Processos, foram utilizados os termos originais em inglês. Para instâncias de processos, foram utilizados *case* ou *process instance*. Para eventos identificados, para atributos de uma instância de processo que representa um conjunto finito e sequencial de eventos e para o registro histórico de eventos, foram utilizados respectivamente *Events*, *Traces* e *Event Logs*. Para o conjunto de transações utlizamos o termo *Transaction Dataset*.

Este trabalho está estruturado na seguinte forma:

- O Capítulo 2 introduz noções preliminares do principal problema de pesquisa. A área de Mineração de Dados e Mineração de Processos também são apresentadas.
- O Capítulo 3 descreve o cenário, o aparato experimental, a construção do modelo de predição e discute os resultados experimentais.
- O Capítulo 4 apresenta os trabalhos relacionados. Destacam-se trabalhos na área de monitoramento de processos que objetivaram seus estudos em predição de maneira a aumentar a confiabilidade de seus processos.
- O Capítulo 5 termina a dissertação com a conclusão e possíveis trabalhos futuros da pesquisa.

2. Fundamentação Teórica

Neste capítulo, discutimos alguns conceitos importantes e embasamentos teóricos necessários ao entendimento da pesquisa. Conceitos relacionados ao campo de estudo dedicado a resolução de problemas por meio de análise de dados são expostos aqui.

2.1 Mineração de Processos

Mineração de processos é uma disciplina relativamente nova e está situada entre as áreas de Aprendizado de Máquina (ML, do inglês *Machine Learning*) e Mineração de Dados (DM, do inglês *Data Mining*) por um lado e Modelagem e Análise de Processo por outro lado [50, 53].

De acordo com [37], aprendizado de máquina refere-se a algoritmos que dão a computadores e máquinas a capacidade de aprender sem a necessidade de serem explicitamente programados. Concentra-se em construir programas de computador que melhoram a partir de experiências passadas. Mineração de dados é a disciplina dedicada a resolução de problemas por meio da análise de dados, sendo uma parte do Processo de Descoberta de Conhecimento em Bases de Dados (KDD, do inglês *Knowledge Discovery in Database*). KDD tem como premissa a descoberta de conhecimento útil, válido, novo e potencialmente relevante sobre uma determinada atividade com o uso de algoritmos [5].

Os algoritmos de mineração de dados produzem resultados na forma de regras ou algum tipo de padrões [5]. Esses algoritmos podem ser aplicados sobre os mais diversos cenários e propósitos tais como classificação, regressão e clusterização. Para cada um deles, pode-se destacar um algoritmo por árvore de decisão para a resolução de problemas de classificação [36], redes neurais para regressão [24] e

k-means para clusterização [31].

Por outro lado, a modelagem e análise de processos estão situadas mais especificamente na área de Gerenciamento de Processos de Negócios que pode ser definida como uma área que combina conhecimentos de tecnologia da informação e as ciências de gerenciamento [51].

De acordo com [53], a área de Mineração de Processos pode ser subdivida em três áreas: (*i*) descoberta de processos, (*ii*) análise de conformidade e (*iii*) aprimoramento de processos.

Descoberta de processos refere-se à aplicação de métodos que visam à descoberta e aprendizado de modelos de processo a partir de dados. Nesta subárea, enfatiza-se a premissa que a descoberta se dá sem conhecimentos e informações *a-priori* [53]. Existem muitos algoritmos destinados a este propósito, tais como o algoritmo *α-algorithm* [54], *Fuzzy Miner* [20] e *Heuristic Miner* [57].

Análise ou verificação de conformidade é aplicada em cenários que visam à verificação entre o modelo e a realidade atual. Nesta subárea, o modelo de processo é comparado com os dados. Um algoritmo de análise de conformidade pode ser visto em [43].

Aprimoramento de processos destina-se à extensão ou melhorias de modelos de processos existentes. Essa melhoria se inicia a partir da análise de dados. A Figura 2.1 descreve as três subáreas, também referenciadas como tipos, de mineração de processos com diferentes entradas para aplicação das técnicas. Uma delas considera o histórico de eventos e as outras duas consideram uma combinação do histórico de eventos com o modelo.

De acordo com [35], definimos como modelo a representação de parte de uma realidade empírica. Dito isso, um modelo de processo é a representação dos objetivos estratégicos e operacionais de uma realidade empírica, isto é, do momento ao qual aquele processo foi idealizado. Essa idealização considera muitas perspectivas tais como perspectivas temporais (época de idealização ou momento pensado) e perspectivas estratégicas e operacionais. Essas últimas estão mais associadas ao negócio e as formas de operacionalização do negócio.

É comum encontrar o termo "modelo" associado aos resultados dos algoritmos de aprendizado da área de Aprendizado de Máquina. Neste caso, a definição é a mesma, uma vez que os modelos dessa área são formas matemáticas de

representações de uma parte da realidade [37].

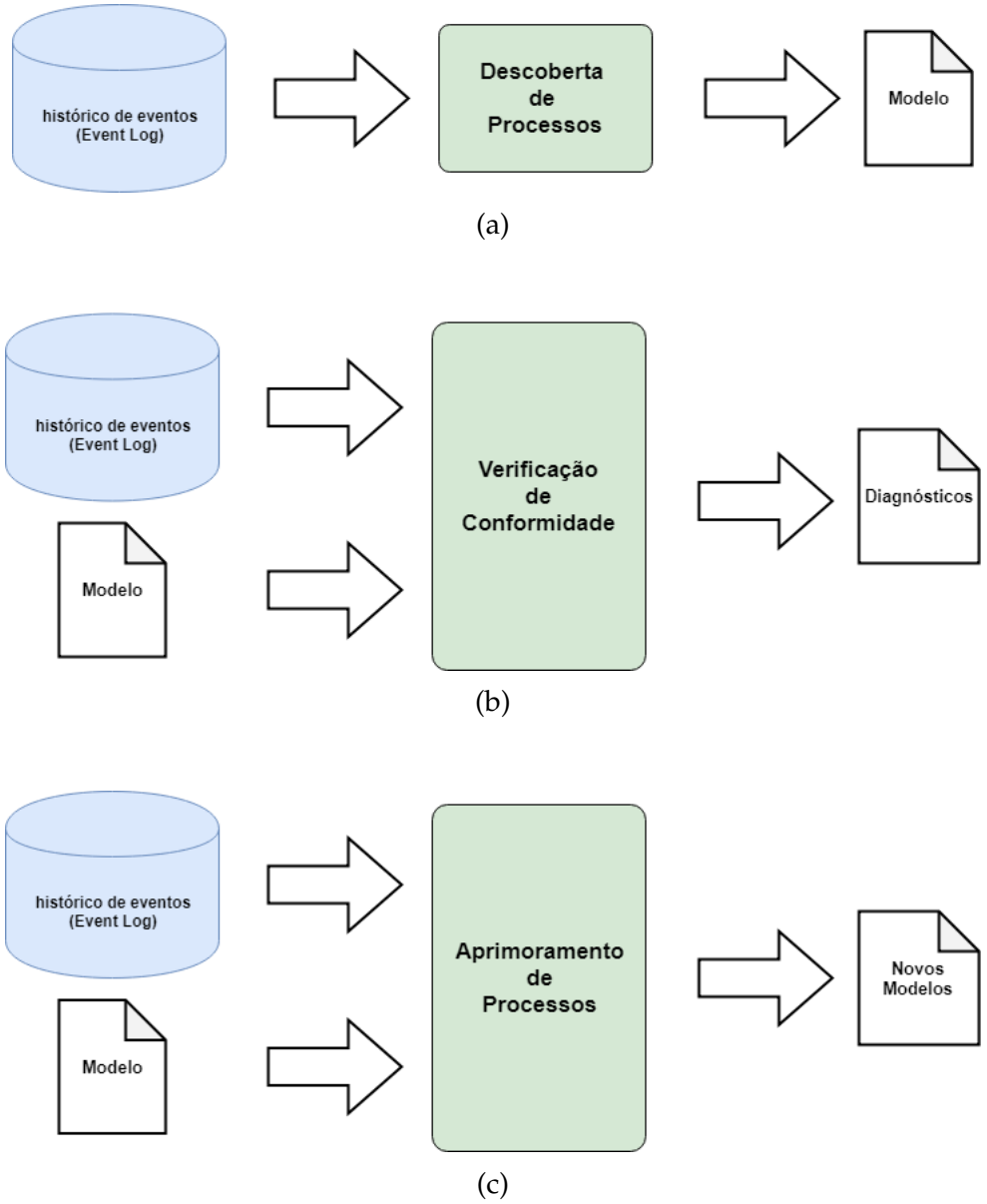


Figura 2.1: Mineração de Processos e suas subáreas. (a) Descoberta de processos (*Process Discovery*); (b) Análise ou Verificação de conformidade (*conformance checking*); (c) Aprimoramento de processos *enhancement*. Adaptado de [50].

Como apresentado na Figura 2.1, o ponto de entrada para quaisquer técnica de mineração de processos é um histórico de eventos (do inglês, *Event Log*). Esse histórico é uma sequência de eventos originados a partir de uma instância de processo. Cada evento possui um conjunto de atributos e são totalmente ordenados no tempo.

Cada evento pode conter propriedades adicionais. Uma definição formal para eventos e atributos é apresentada na Definição 1. Para *Cases*, *Trace* e *Event Log* destacamos a definição, apresentada por [26], e mostrada em Definição 2.

Definição 1 *(Event, Attribute). Seja $\mathcal{E}$ o conjunto de todos os eventos identificados e AN o conjunto de todos os atributos. Para um atributo $x \in AN$, permita que χ_x denote seu universo, i.e., o conjunto de todos o possíveis valores para x. Dado um $\mathcal{E}$ e um atributo $x \in AN$. Uma função dada por $\#x : \mathcal{E} \rightarrow \chi_x \cup \perp$ mapeia o valor do atributo x para qualquer evento $e \in \mathcal{E}$. Uma função dada por $\#_x(e) = \perp$ mapeia todos os atributos x não definidos para e.*

Definição 2 *(Case, Trace, Event Log). Seja $\mathscr{C}$ o conjunto de todas as instâncias de processo identificadas e AN o conjunto de todos os atributos. Para qualquer instância de processo $c \in \mathscr{C}$ e um atributo $x \in AN$. Uma função dada por $\widehat{\#_x} : \mathscr{C} \rightarrow \chi_x \cup \perp$ mapeia o valor do atributo x para uma instância de processo c (assim como eventos, instâncias de processo também possuem atributos). Um trace é um atributo obrigatório de uma instância de processo e representa uma sequencia finita de eventos $t \in \mathcal{E}$ de tal modo que não exista dois eventos iguais, i.e., para $1 \leq i < j \leq |t|, t(i) \neq t(j)$. Além disso, os eventos em um trace seguem uma ordem ascendente no tempo se o atributo relacionado ao tempo estiver presente, i.e., $1 \leq i < j \leq |t|, \#_{time}(t(i)) \leq \#_{time}(t(j))$.*

Para uma instância de processo c, $\mathcal{E}_c$ denota o conjunto de eventos que pertence a c, i.e., $\mathcal{E}_c = \widehat{\#_{trace}}(c)$. Um event log é um conjunto de instâncias de processos $\mathcal{L} \subseteq \mathscr{C}$ de modo que cada evento ocorre uma única vez, i.e., para duas instâncias de processos qualquer $c_1, c_2 \in \mathscr{C}, c_1 \neq c_2 : \mathcal{E}_{c1} \cap \mathcal{E}_{c1} = \emptyset$.

Nesta seção, vimos que a área de Mineração de Processos dividi-se em três subáreas e o principal propósito da área é aumentar a confiabilidade dos processos em termos gerenciais e operacionais. Do ponto de vista operacional, tem-se a atividade de monitoramento de processos que visa medir o desempenho do processo ao longo de suas atividades e execuções. Este assunto está mais relacionado a análise de conformidade de processo (Figura 2.1b). A próxima seção apresenta os conceitos da atividade de monitoramento de processos.

2.2 Monitoramento de Processos

As técnicas em Análise de Conformidade contribuem para o alinhamento de negócios e auditoria [53] e reúnem diversas técnicas de monitoramento de processos.

As técnicas de monitoramento de processos visam medir o desempenho de

processos e são empregadas sem a necessidade de construir ou utilizar um modelo de processo. Contribuem efetivamente para o mapeamento de situações que não estão em conformidade com o processo idealizado [53]. Situações que não estão em conformidade são chamadas de situações indesejadas ou até mesmo de fracassos do processo. Por exemplo, se o tempo total decorrido de uma instância de processo ultrapassa limiares predeterminados então pode-se inferir que a instância fracassou em termos de desempenho, uma vez que, o tempo total decorrido era uma premissa. Por outro lado, se o tempo decorrido está de acordo com o que se espera então é dito que instância segue em conformidade com o que se espera.

Dentre as atividades relacionadas ao monitoramento de processos, destacamos o monitoramento de indicadores chave de desempenho (KPI, do inglês *Key Performance Indicators*). Os indicadores chave de desempenho consiste de métricas importantes que podem ser utilizadas para indicar oportunidades de melhoria do processo ou indicar problemas que necessitam ações proativas ou corretivas [27, 51].

O monitoramento pode ser conduzido de duas maneiras, (*i*) análise qualitativa e (*ii*) análise quantitativa.

A análise qualitativa consiste de identificar etapas desnecessárias ao processo e então propor a retirada dessas atividades por meio da modelagem de outros processos.

A análise quantitativa de processos utiliza medidas como tempo total decorrido, custo e qualidade do processo. Cada medida pode ser refinado em um número de indicadores de desempenho.

De acordo com [39], indicadores de desempenho estão direcionados ao desempenho dos aspectos organizacionais que são cruciais para o atual e sucesso futuro da organização.

Como visto, os indicadores de desempenho podem formar um aspecto chave em avaliações de processos. Este trabalho de pesquisa foca-se em um tipo especial de indicador de desempenho. Este indicador especial é o indicador de desempenho de processos (PPI, do inglês *Process performance Indicator*). Os indicadores de desempenho de processo focam-se nos aspectos operacionais e são medidos diretamente a partir dos dados [11].

De maneira a citar exemplos, destacamos alguns tipos de indicadores encon-

trados na literatura [15,25,33,34,48], tais como:

- **Tempo.** É possível medir o desempenho do processo considerando o tempo decorrido das instancias do processo;
- **Frequência de eventos.** É possível medir o desempenho do processo considerando a frequência que determinados eventos ocorrem em uma instância do processo baseando-se na repetição dos mesmos;
- **Recursos ou envolvidos.** É possível medir o desempenho do processo considerando o total de participantes do processo ou nas instâncias do processo.
- **Custos.** É possível medir o desempenho do processo considerando o custo necessário de operacionalização do processo.
- **Qualidade.** É possível medir o desempenho do processo considerando fatores de qualidade dos entregáveis do processo, isto é, dos artefatos gerados ao final ou durante a execução do processo.

A literatura não se limita aos exemplos acima mencionados. Segundo [39], os indicadores de processos possui múltiplas perspectivas e são usados nos mais variados cenários.

A medição desses indicadores pode evidenciar mudanças de comportamento de processos. Existem muitos fatores que podem ser considerados para que essas mudanças ocorram tais como mudanças de regulamentos internos e externos, mudanças climáticas, mudanças geográficas, mudanças políticas e até mesmo mudanças estratégicas e operacionais. Diante disso, os processos de negócio estão sujeitos a uma natureza dinâmica que podem levar o processo a uma necessidade de adaptação. Este assunto é tratado na próxima seção.

2.3 Dinamicidade em Processos

Os processos de negócio estão sujeitos a uma natureza dinâmica, como visto nas seções anteriores. Segundo [53], os processos de negócios não são estáticos. Essa definição por si só já denota que mudanças podem ocorrer ao longo da execução dos processos.

Dito isso, os processos podem sofrer variações para manterem-se operantes a determinadas situações. Essas variações podem se manifestar perante a natureza

dos clientes, diferenças geográficas, mudanças climáticas além de outros exemplos que podem ser encontrados na literatura dada por [1]. Essas variações resultam na heterogeneidade de tipos de instâncias do processo.

A esta heterogeneidade de tipos chamaremos de variantes de processo (do inglês *process variants*). Segundo [29], uma variante de processo é uma sequência específica de atividades onde o caminho entre a primeira atividade e a última atividade formam um caminho único, isto é, um caminho sem ramificações.

Consideramos importante garantir essa heterogeneidade pois, por meio das variantes de processo, é possível identificar quais padrões em termos de fluxo operacional entregam bons resultados. Considerando o que está exposto em [17,29], a frequência das variantes de processo permite identificar padrões capaz de distinguir *outliers*[1], exceções[2], instâncias incompletas ou ainda em execução. Permitem também identificar pontos de customizações de um processo. A capacidade de flexibilização de um processo também pode ser medido e evidenciado por meio de análises das variantes de processo.

A Figura 2.2 expõe um exemplo de um processo e suas variantes.

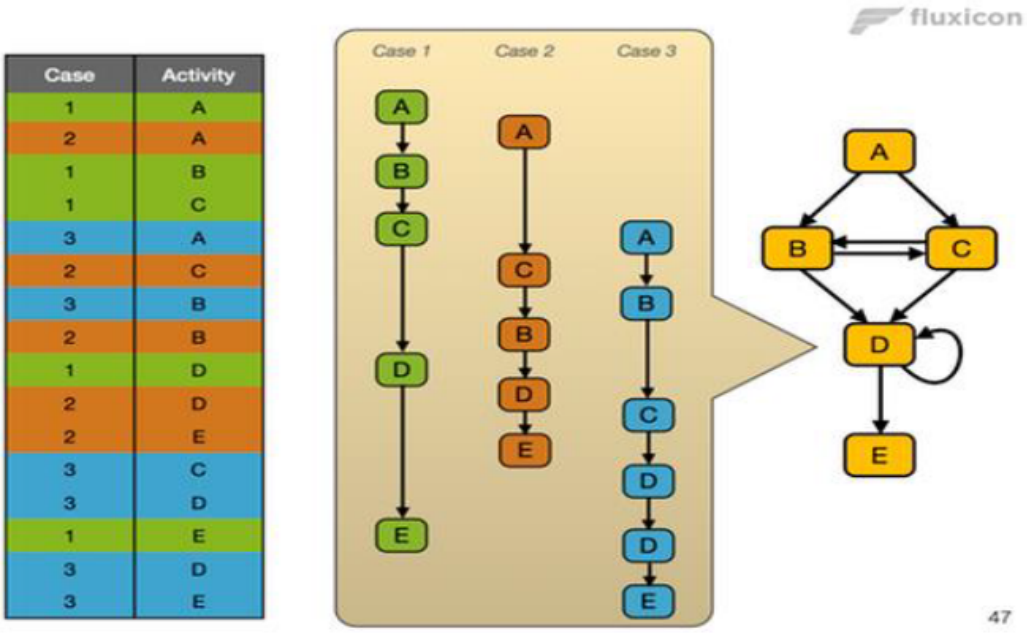


Figura 2.2: Caracterização das variantes de processo a partir de um histórico de eventos. Fonte: [17].

Considerando a Figura 2.2, tem-se ao lado esquerdo um exemplo de um

[1]Padrões, instâncias ou amostras que diferem de maneira considerável de todas as outras.
[2]Dada a dinâmica dos processos, situações de exceção podem surgir para atender determinadas necessidades.

histórico de eventos onde a coluna *Case* identifica a instância de processo e a coluna *Activity*, o nome de uma atividade. Cada linha representa o registro de um evento. Ao centro, têm-se as variantes de processo consideradas e seus respectivos fluxos. Ao lado direito, tem-se o processo descoberto a partir do histórico de eventos utilizando algoritmos de mineração.

Pelas definições e exemplos apresentados, pode ser percebido que as variantes de processo são importantes no sentido de contribuições à atividade de monitoramento de processos. Por outro lado, essa heterogeneidade pode aumentar a complexidade de gerenciamento do processo por partes de analistas e gestores [25], uma vez que mudanças podem levar o processo a estados indesejados de maneira que o torne inoperante. A dinamicidade à qual os processos de negócio estão sujeitos não estão relacionadas aos seus fluxos apenas, mas também a todo um contexto operacional. As mudanças podem ser provocadas com a entrada de novos clientes, recursos e até mesmo por erros operacionais. Dito isso é fácil ver que os processos de negócio estão sujeitos a um problema bem conhecido da área de Ciência de Dados, termo guarda-chuva ao qual a área de Mineração de Processos se encaixa [53]: a mudança de conceito (*concept drift*). Este é o principal assunto da próxima seção.

2.4 Mudança de Conceito em Processos

A mudança de conceito (do inglês *concept drift*), em análise de dados, refere-se às situações de mudança das características dos dados. Essas mudanças podem estar relacionadas à rotulagem de dados ou quando a distribuição de dados muda [18, 46]. Uma mudança de conceito pode acontecer por diversos motivos, por exemplo, fatores sazonais, mudança de regulamentos internos e externos, situações de calamidade e desastres.

De acordo com [56], a mudança de conceito é um dos principais responsáveis pela degradação de modelos. Um modelo reúne princípios específicos que visam representar parte de uma realidade empírica sobre uma forma matemática [35,38]. A degradação do modelo refere-se ao momento em que este modelo deixa de alcançar um desempenho aceitável.

A justificativa por esta degradação se dá pelo fato que o mundo em que vivemos é dinâmico e está em constante mudança, o que vai de acordo com a dinamicidade requerida pelos processos. Diante disso, novas situações podem

surgir de acordo com alguma necessidade, algumas delas já mencionadas em tópicos anteriores.

O estudo e aprofundamento da mudança de conceito ainda é um tópico recente em Mineração de Processos [4]. Diversos trabalhos destacam a importância do estudo e criação de métodos que visam a detecção da mudança de conceito, tais como [4, 18, 46, 52, 56].

De acordo com [4], é necessário considerar três desafios ao lidar com a mudança de conceito em mineração de processos.

- **Detecção de ponto de mudança.** Consiste em detectar o ponto de mudança e quando este ponto acontece.
- **Caracterização da mudança** Consiste em caracterizar a natureza da mudança e identificar as regiões no tempo em que elas ocorrem.
- **Captura e Predição de Modelo.** Consiste em capturar um comportamento irregular.

O estado-da-arte destaca quatro diferentes tipos de mudanças de conceito: repentinos, graduais, recorrentes e incremental. Cada um deles é exposto na Figura 2.3.

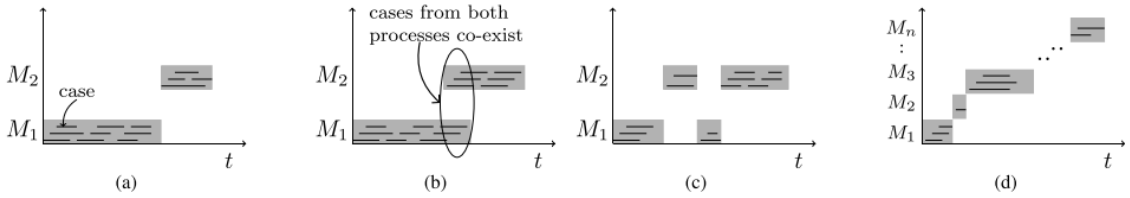


Figura 2.3: Diferentes tipos de mudança de conceito. eixo-x: tempo. eixo-y: variantes de processo. Retângulos: instancias de processo. (a) mudança repentina. (b) mudança gradual. (c) mudança recorrente. (d) mudança incremental. Fonte: [4].

Considerando a Figura 2.3, tem-se:

- **Mudança repentina.** Trata-se da substituição de um processo de maneira abrupta.
- **Mudança gradual.** Trata-se da substituição de um processo de maneira gradual, isto é, mantém-se o processo anterior a substituição em operação

até o término de suas instâncias. Novas instâncias são consideradas apenas para o novo processo. Os processos coexistem temporariamente.

- **Mudança repentina.** Trata-se de processos recorrentes. Neste caso, a substituição pode não ser observada, os processos existem na linha temporal, mas não há evidência de coexistência. Novas instâncias podem surgir para ambos os processos.

- **Mudança incremental.** Trata-se da substituição de um processo de maneira gradual. Novas instâncias vão surgindo para diversos processos até se consolidar em um processo específico que se perdurará por um tempo até que novas mudanças venham a surgir.

Diferentes técnicas têm sido aplicadas no estudo do *concept drift* de maneira a aprimorar a percepção e compreensão deste fenômeno comportamental dos processos. Essas técnicas visam aumentar a capacidade de análise do processo por partes de analistas e gestores.

A ideia básica para a aplicação das técnicas de detecção do *concept drift* segue os seguintes passos:

- Ordenar o conjunto de dados de maneira análoga a uma série temporal, isto é, o conjunto de dados ou conjunto de observações são ordenados sequencialmente no tempo;

- Segmentar o conjunto de dados em amostras, isto é, em subconjuntos.

- Aplicar algum algoritmo de detecção da mudança de conceito.

Para o primeiro ponto, tem-se a ordenação do histórico de eventos a partir do atributo que demarca o momento de registro do evento. Esse atributo certamente estará presente em um histórico de eventos $\mathcal{L}$ por ser uma premissa do mesmo, como visto a Definição 2. No segundo ponto, tem-se a segmentação do histórico de eventos $\mathcal{L}$ resultando em subconjuntos $\mathcal{L}_1$ e $\mathcal{L}_2$. Por fim, a aplicação dos algoritmos de detecção que caracterizam-se em comparar a distribuição de dados dos subconjuntos obtidos a partir da segmentação.

As próximas seções destacam três técnicas que já foram utilizadas em diferentes trabalhos na área de Mineração de Processos.

2.4.1 Mineração por Regras de Associação

Um dos métodos para detectar não apenas a mudança de conceito, mas também desvios de processo, é a Mineração por Regras de Associação [12].

A Mineração por Regras de Associação visa encontrar regras que correspondem aos limiares pré-determinados para o suporte (*support*) e confiança (*confidence*). Por um lado, o suporte é o indicador para um número de instâncias de processo para o qual uma determinada regra se aplica. Por outro lado, a confiança indica de quantas maneiras as regras estão atribuídas corretamente as instâncias de processo. Uma regra é uma estrutura lógica condicional dada por:

$$X \rightarrow Y \tag{2.1}$$

Mineração por Regras de Associação são aplicadas em um conjunto de dados transacional, também referenciado como *Transaction Dataset*. Formalmente, baseado em [22], definimos um *Transaction Dataset* como:

Definição 3 *(Transaction Dataset). Seja $I = a_1, a_2, \ldots, a_n$ um conjunto de itens, e um conjunto de transações $D = \langle T_1, T_2, \ldots, T_n \rangle$ onde $T_i (i \in [1..n])$ é uma transação que contém um conjunto de itens em I.*

Por fim, suporte (Eq. 2.2) e confiança (Eq. 2.3) estão definidos como seguem:

Seja D um conjunto de transações, de modo que:

$$support(X \rightarrow Y) = \frac{freq(X \cup Y)}{|D|} \tag{2.2}$$

$$confidence(X \rightarrow Y) = \frac{support(X \rightarrow Y)}{support(X)} \tag{2.3}$$

Para X e Y conjuntos não vazios em D.

A Equação 2.2 apresenta a probabilidade de uma transação $X \cup Y$ ocorrer. A Equação 2.3 apresenta a confiança, dada pela divisão do suporte de $X \rightarrow Y$ pelo suporte do antecedente X. Em outras palavras, a confiança é obtida por meio da estimativa da probabilidade (Pr) dada por $\Pr(Y \mid X)$.

2.4.2 Mineração por Padrões Frequentes

A técnica de Mineração por Padrões Frequentes consiste da primeira etapa de Mineração por Regras de Associação, i.e., não considera a confiança. Esta técnica consiste em descobrir e identificar uma coleção de um ou mais conjuntos de itens (em inglês, *itemsets*) frequentes. Se a ocorrência deste conjunto de itens for maior ou igual que determinados limiares configurados como suporte mínimo (em inglês, *minimum support*), então é dito que este conjunto de itens é um padrão frequente (em inglês, *frequent pattern*). As definições para suporte e padrão frequente estão conforme apresentadas em [21].

Definição 4 *(Support, frequent pattern). O suporte (ou ocorrência frequente) de um padrão A, que é um conjunto de itens I, é o número de transações que contém I em D. A é um padrão frequente (frequent pattern) se a ocorrência de A é maior ou igual a um valor mínimo definido para o suporte ϵ.*

Sua utilização pode ser vista no trabalho de [8].

2.4.3 Kolmogorov-Smirnov Test

Um outro método, e possivelmente um dos mais utilizados na detecção de mudanças de conceito em Mineração de Processos [9, 14, 45, 53, 61], é a utilização do teste estatístico Kolmogorov-Smirnov (KS-Test).

KS-Test é um teste estatístico não-paramétrico que detecta diferenças na distribuição de dados por meio da comparação entre duas amostras. O teste estatístico está baseado na função de distribuição empírica dada por:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{2.4}$$

onde $X_1, \ldots, X_n$ denota uma sequencia de variáveis aleatórias e $I(e)$ a função indicadora do evento $e \in \mathcal{E}$.

Por fim, o teste estatístico não-paramétrico é dado por:

$$D = \max_{1 \leq i \leq N} (F(Y_i) - \frac{(i-1)}{N}, \frac{i}{N} - F(Y_i)) \tag{2.5}$$

2.5 Predição em Processos

As técnicas que visam alcançar cenários preditivos em processos de negócios emergem sob diversas necessidades. Destacamos a necessidade pela transição entre um cenário costumeiramente reativo para um cenário proativo. Em um cenário reativo, um fato indesejado de uma instância de um processo é conhecida após a sua ocorrência. Um cenário proativo permitiria a tomada de decisão de maneira antecipada ao fato, podendo inclusive impedir a ocorrência de algo indesejado ou minimizar os possíveis riscos que a ocorrência do cenário causaria ao negócio.

Se uma instância de processo não alcança os objetivos ao qual o processo foi idealizado, então essa instância de processo fracassou. Por outro lado, se a instância manteve-se operante em todos os níveis de medidas adotadas e todas consideradas aceitáveis, então a instância teve um bom desempenho e não fracassou. Se fosse possível predizer o fracasso das instâncias de processo, então se poderia supor que ações poderiam ser tomadas de maneira a impedir a existência desse fracasso e manter a confiabilidade dos processos.

Deste modo, a área de Mineração de Processos tem concentrado esforços na consolidação e aprimoramento das técnicas que visam o alcance de predições em processos de negócio nos mais variados cenários. Esses cenários podem se dar na predição dos indicadores de desempenho, predição dos próximos eventos, predição de riscos e predição de determinados atributos [33].

De acordo com [33], as técnicas de predição em processos dividem-se em dois tipos: (*i*) técnicas que seguem métodos cientes de processos (do inglês *process-aware methods*) e (*ii*) técnicas que seguem métodos não cientes de processos (do inglês *non process-aware methods*).

Para o primeiro tipo, as técnicas consideram a estrutura dos processos [33], isto é, consideram os caminhos, fluxos e as transições entre as atividades. Um exemplo é o uso de Autômatos Finitos Probabilísticos (PFA, do inglês *Probabilistic Finit Automaton*) para a predição dos próximos eventos de uma instância do processo [7].

Para o segundo tipo, as técnicas de predição não utilizam as representações explicitas dos processos [33], isto é, não consideram os caminhos, fluxos e as transições das atividades do processo. Neste caso, essas técnicas tornam-se de-

pendentes de uma atividade de sumarização das informações dos processos. Um exemplo é o uso de Árvores de Decisão (DT, do inglês *Decision Tree*) para a predição dos riscos de um processo e também do SLA. Sendo uma técnica oriunda da área de Mineração de Dados, DTs tem alcançado bons resultados em diferentes domínios [10, 32].

Esta dissertação utiliza o segundo tipo devido ao principal objetivo da pesquisa que é a predição de fracassos de um processo de gerenciamento de incidentes evidenciado pelo não cumprimento do SLA. Esta evidência se dá por meio do cálculo de uma variável categórica. Este assunto é explorado em detalhes no Capítulo 3.

A atividade de sumarização visa obter uma representação vetorial do histórico de eventos também conhecida por *Event Encode* [33]. Existem diferentes técnicas que podem ser aplicadas na atividade de sumarização dos eventos do processo, ou mais especificamente no *Event Encoding*. A próxima seção destaca três atividades já citadas pela literatura e utilizadas nesta pesquisa durante a aplicação do estudo empírico.

2.5.1 Sumarização dos Eventos e o Vetor de Variáveis

Destacamos algumas técnicas dessa representação sumarizada, tais como (*i*) *boolean encoding* (Tabela 2.1a); (*ii*) *frequency-based encoding* (Tabela 2.1b) e (*iii*) *index-based encoding* (Tabela 2.1c). A Tabela 2.1 apresenta exemplos dessas representações. Essas representações são definidas em [30].

Considerando a Tabela 2.1, as formas de *encoding* diferem entre si. Para *boolean encoding* e *frequency-based encoding*, tem-se uma sequencia de eventos descrita em um vetor de variáveis (do inglês *feature vectors*) onde cada variável corresponde a uma classe de eventos (ou atividades) do processo que está registrado no histórico de eventos, isto é, cada variável é dada por $e_1, e_2, \ldots, e_n \in \mathcal{E}$. Os valores presentes nas variáveis do *boolean encoding* são representados, em termos gerais, por:

Seja $e_1, e_2, \ldots, e_n \in \mathcal{E}$ para um *case* c, com:

σ_i = uma instância de processo de índice i

v_i = vetor de variáveis

Tabela 2.1: Exemplos de representações sumarizadas do histórico de eventos, também chamada de *Event Encode*. Adaptado de [30].

(a) *boolean encoding*

case	manutenção	instalação	email	$\cdots$	label
σ_1	1	0	0	$\cdots$	false
σ_2	0	0	1	$\cdots$	true
σ_3	1	1	0	$\cdots$	true
$\vdots$					
σ_k	0	1	0	$\cdots$	false

(b) *frequency-based encoding*

case	manutenção	instalação	email	$\cdots$	label
σ_1	1	0	0	$\cdots$	false
σ_2	0	0	3	$\cdots$	true
σ_3	1	1	2	$\cdots$	true
$\vdots$					
σ_k	0	1	1	$\cdots$	false

(c) *index-based encoding*

case	$evento_1$	$\cdots$	$evento_m$	$recurso_1$	$\cdots$	$recurso_m$	label
σ_1	manutenção	$\cdots$	email	Anna	$\cdots$	Anna	false
σ_2	manutenção	$\cdots$	instalação	Anna	$\cdots$	John	false
σ_3	rede	$\cdots$	email	John	$\cdots$	Paul	false
$\vdots$							
σ_k	manutenção	$\cdots$	email	Anna	$\cdots$	Anna	false

De modo que,

$$v_{ij} = \begin{cases} 1, & \text{se } e \text{ está presente em } \sigma_i \\ 0, & \text{caso contrário.} \end{cases} \tag{2.6}$$

Onde,

$$v_i = [v_{i1}, v_{i2}, \ldots, v_{ij}]$$

$$v_{ij} = \text{ corresponde à uma classe de evento } e \in \mathcal{E}$$

Desta maneira, define-se que cada evento ou atividade é representada por meio de uma relação de pertinência no vetor de variáveis. Para *frequency-based encoding* tem-se o fluxo representado por meio da frequencia de cada evento e não por meio de uma simples relação de pertinência com apenas dois valores possíveis (1 ou 0).

Considerando *index-based encoding*, têm-se dados associados aos eventos divididos em informações estáticas e dinâmicas. As informações estáticas são as mesmas informações associadas para todos os eventos e as informações dinâmicas são informações que são diferentes para cada evento.

Durante a realização desse trabalho de pesquisa, entendemos que as informações estáticas se tratam das informações relacionadas aos atributos da instância do processo e as informações dinâmicas são as informações relacionadas aos atributos dos eventos do processo. No exemplo apresentado na Tabela 2.1c, os atributos dinâmicos são as variáveis que apresentam os recursos, isto é, os envolvidos das atividades e m caracteriza o índice da atividade no fluxo operacional do processo. Este fluxo operacional do processo é representado por uma sequência finita de eventos $t \in \mathcal{E}$ de tal modo que $1 \leq i < j \leq |t|, t(i) \neq t(j)$.

Como visto até aqui, as técnicas de predição, que são originadas da área de Mineração de Dados, utilizam como dados de entrada o *event encoding* obtido por meio da preparação de dados até um formato ideal como *boolean encoding*, *frequency-based encoding* ou *index-based encoding*. A próxima seção apresenta algumas técnicas de predição já utilizadas em alguns trabalhos.

2.5.2 Algoritmos de Classificação e a Predição

Como exposto até aqui, a predição em processos permitiria a transição entre um cenário reativo para um cenário proativo. Este último cenário possibilitaria a tomada de decisão de maneira a mitigar riscos que uma falha operacional de um processo pode apresentar aos negócios. Dito isso, predições em processos enriquecem e apoiam de maneira operacional o gerenciamento dos processos.

Considerando que a área de Mineração de Processos faz parte do termo guarda-chuva da Ciência de Dados que consiste da resolução de problemas por meio da análise de dados [53], a utilização de algoritmos e técnicas bem estabelecidas das áreas de Aprendizado de Máquina e Mineração de Dados segue de maneira natural e aplicável. Diante disso, é necessário definir a forma de aprendizado que será adotada para alcançar a predição em processos.

A literatura [35, 44] destaca três formas de aprendizado: (*i*) supervisionado; (*ii*) não supervisionado e (*iii*) reforçado.

O aprendizado supervisionado, como destacado por [44], utiliza pares de entrada e saída das amostras ou de um conjunto de exemplos. Este conjunto de

exemplos é o que caracteriza o aprendizado por dados ou aprendizado por meio de exemplos [35]. Os pares de entrada e saída são utilizados para determinar se as respostas esperadas (saída) seguem de maneira correta para cada entrada [44]. Deste modo, o comportamento dessa forma de aprendizado utiliza as saídas esperadas como um gabarito durante o aprendizado onde um especialista pode decidir se o aprendizado é suficiente por meio da verificação das saídas ou se o aprendizado deve ser refeito considerando outras perspectivas, tais como a adoção de outros algoritmos e modelos, novos parâmetros ou até mesmo novos conjuntos de dados.

O aprendizado não supervisionado não possui um *feedback* tão explícito como a forma supervisionada. Ao contrário, os conceitos são construídos gradualmente. Uma tarefa de aprendizado não supervisionado é a tarefa de agrupamentos ou *clustering* que consiste em detectar potenciais grupos distintos que existem em um conjunto de amostras [44]. Dito isso, o conjunto de saídas esperadas para cada entrada não são conhecidas, como, por exemplo: de que forma cada pessoa definiria e diferenciaria um dia ruim de um dia bom?

O aprendizado reforçado, como definido por [44], se dá por meio de uma série de reforços. Tais reforços são distribuídos entre punições e recompensas. Desta maneira, o aprendizado reforçado é capaz de mensurar determinadas decisões tomadas e utilizar tais informações como conhecimento prévio, por exemplo: Um motorista, ao final de um percurso, pode receber algumas dicas de otimização do percurso e utilizar essas informações para a próxima vez que realizar o mesmo percurso.

Após determinar a forma de aprendizado, um algoritmo pode ser aplicado sobre um determinado propósito tais como classificação, regressão e outros já expostos na Seção 2.1.

Os algoritmos de classificação possuem como saída esperada um conjunto finito de valores. De acordo com [44], uma saída esperada y pode ter um conjunto finito de valores tais como ensolarado, nublado e chuvoso[3]. Por outro lado, se y é um número contínuo e não discreto (como por exemplo a temperatura) então o algoritmo é um algoritmo de regressão.

Considerando a literatura dado por [9, 14, 40, 45, 58], destacamos os seguintes algoritmos de aprendizado supervisionado *non process-aware* que classificam uma

[3]Um exemplo de classificação de dados climáticos.

variável dependente em termos de variáveis independentes que foram aplicados em diferentes domínios com resultados consideráveis: (*i*) *Naive Bayes* (NB), (*ii*) *Decision Tree* (DT), (*iii*) *Random Forest* (RF) e (*iv*) *Gradient Boosted Tree* (GBT). Esses algoritmos podem ser vistos em detalhes de maneira bem aprofundada em [3, 5, 6, 23, 41].

2.6 Considerações Finais

Neste capítulo, apresentamos a área de Mineração de Processos e definições formais de suas características e os problemas que a área se propõe em resolver. Foi visto também que as atividades de monitoramento de processos permitem que analistas e gestores entenda o comportamento de seus processos. O estudo do *concept drift* apresenta-se de maneira enriquecedora para a compreensão do comportamento do processo em diferentes perspectivas. Por fim, destacamos a importância da aplicação das técnicas de classificação visando à predição, permitindo a transição entre um cenário reativo para um cenário proativo por meio da predição de eventos, riscos, SLA ou outras variáveis consideradas.

Desta maneira, contextualizamos o cenário que esta pesquisa se insere no âmbito científico.

3. Predição de Fracassos em Processos de Negócio

O principal objetivo deste trabalho de pesquisa é a predição de fracassos em processos de negócio. Os fracassos dos processos são obtidos por meio do cálculo de um indicador de desempenho e referenciado como um atributo categórico das instâncias de processos. Posteriormente, utiliza-se algoritmos de classificação de maneira a obter o indicador de desempenho anteriormente calculado em um cenário preditivo. Este trabalho realiza um estudo de caso em um processo de gerenciamento de incidentes com dados reais, isto é, não são tratados dados sintéticos e simulados de um processo e sim dados extraídos de um ambiente corporativo.

Este capítulo apresenta os passos propostos para a modelagem da predição das situações de fracassos. Esses passos envolvem a combinação de diferentes técnicas oriundas das áreas de Mineração de Dados e Mineração de Processos, envolvem também técnicas de pré-processamento que visa à obtenção de um conjunto final de dados a partir de um conjunto inicial.

As técnicas de pré-processamento são importantes durante a modelagem preditiva pois, por meio delas, é possível obter novas representações que melhor descrevem uma realidade empírica e servem como entrada para os algoritmos de classificação. A limpeza, o tratamento e até mesmo a análise exploratória dos dados podem ser realizadas durante a etapa de pré-processamento. O aparato experimental e os resultados obtidos também são discutidos neste capítulo.

3.1 Projeto Experimental

Dado a **questão de pesquisa** formulada na Seção 1.2 do Capítulo 1, o projeto experimental foi definido de maneira que permita a combinação de diferentes

técnicas das áreas de Mineração de Dados e Mineração de Processos. No início deste capítulo, foi visto o principal objetivo e os resultados ou saídas esperadas da pesquisa. Diante disso, alguns passos devem ser seguidos para o alcance do objetivo proposto. Esses passos são apresentados em detalhes nas Seções 3.1.1, 3.1.2, 3.1.3, 3.1.4 e 3.1.5.

3.1.1 Histórico de Eventos

A premissa para a aplicação das técnicas de mineração de processos é o histórico de eventos. Um histórico de eventos é gerado a partir da captura das informações relativas ao fluxo de atividades do processo. Cada atividade realizada é registrada por um sistema de informação e deve compor as seguintes características:

- Cada evento refere-se a uma tarefa específica;
- Cada evento está relacionado a uma instância de processo;
- Os eventos possuem uma ordenação topológica.

As características acima mencionadas caracterizam os dados como um *Event Log*, como apresentado por [49]. Deste modo, é fácil ver a presença das propriedades apresentadas nas Definições 1 e 2.

Não é o foco desta dissertação abordar estratégias de captura dos eventos de um processo com a finalidade de gerar o histórico de eventos. Toda a pesquisa foi realizada a partir de um histórico de eventos já mapeado por um sistema de informação. Diante disso, para garantir a reprodutividade deste trabalho pressupõe-se a existência de um histórico de eventos *a priori* para a adoção das estratégias de mineração de dados e mineração de processos aqui expostos.

3.1.2 Mineração de Processos

Há muitos cenários em que os modelos de processos não são conhecidos. Nesses cenários, o que existe é uma sensação por parte dos envolvidos no processo de como o processo está sendo executado. Este é um problema já conhecido e um dos principais tópicos abordados pela área de Mineração de Processos, como exposto no Capítulo 2.

Nesta etapa do projeto experimental, o principal objetivo é aplicar técnicas de

área de Mineração de Processos, visando especificamente à descoberta de modelos de processos, de maneira a suprir a falta de um modelo ou esclarecer algumas divergências entre o processo idealizado e o processo que está em execução.

Não é foco desta dissertação criar um algoritmo de descoberta de modelos de processos. Ao contrário, a pesquisa experimental utilizou técnicas bem estabelecidas pela literatura e, inclusive, recomenda seu uso.

Um dos principais objetivos desta etapa é utilizar o conhecimento adquirido como resultante da aplicação dessas técnicas de maneira a guiar as possíveis transformações que o conjunto de dados inicial deve ser submetido. Deste modo, dependendo da resultante das respectivas técnicas aplicadas, tem-se a obtenção de um novo conjunto de dados a partir do conjunto inicial de dados.

Destacamos três ferramentas bem estabelecidas pela indústria e também pela literatura que concentram diversos algoritmos que podem ser utilizados durante a descoberta de modelos de processo. Essas ferramentas são *ProM*, *Apromore* e *Disco*.

Este trabalho de pesquisa fez o uso do *ProM* e do *Disco*. Entretanto, a exportação dos dados em um formato tabular[1] dos modelos de processo foi possível apenas com este último. O algoritmo de mineração de processos utilizado é o *Fuzzy Miner* implementado pela equipe de desenvolvedores do *Disco*[2].

A Figura 3.1 apresenta o fluxo adotado durante a experimentação para esta etapa do trabalho de pesquisa realizado.

A Figura 3.1 apresenta como entrada o histórico de eventos que deve ser carregado pelo *software* e finalizando com uma tarefa de exportação da resultante da aplicação. Durante essa etapa é possível fazer uma análise exploratória dos processos descobertos. Após a exportação das informações referentes aos modelos de processo descoberto no histórico de eventos, neste caso a resultante, tem-se uma nova fonte de dados caracterizada pela seguinte estrutura:

- **Case**. Trata-se de uma coluna que identifica o incidente tratado.
- **Events**. Trata-se de uma coluna que soma o número de eventos que a instância de processo teve.

[1] Informação apresentada como uma tabela, contendo linhas e colunas.

[2] Essa informação pode ser vista em `https://fluxicon.com/blog/2012/05/say-hello-to-disco/`

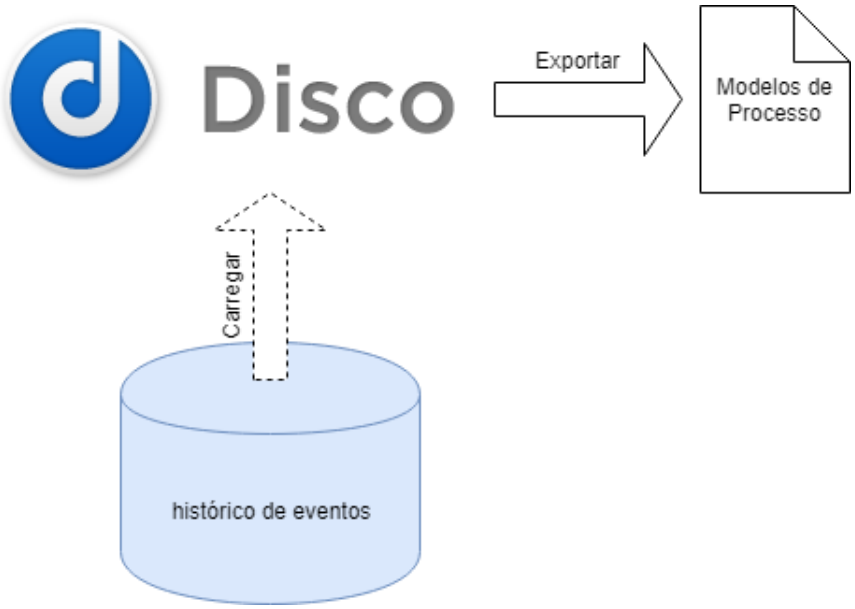


Figura 3.1: Fluxo adotado durante a obtenção e descoberta de modelos de processo.

- **Variant**. Trata-se de uma coluna que identifica uma variante de processo.
- **Started**. Trata-se de uma coluna que demarca o primeiro momento que um evento de uma determinada instância foi registrado.
- **Finished**. Trata-se de uma coluna que demarca o último momento que um determinado evento de uma instância foi registrada.
- **Duration**. Trata-se de uma coluna que calcula o tempo decorrido de uma determinada instância do processo.

A exportação dessas informações podem ser utilizadas em um cenário de enriquecimento de dados por meio de técnicas de preparação de dados sendo uma tarefa extremamente relevante para a construção do modelo de predição.

A construção do modelo de predição envolve duas etapas, que são:

- **Preparação de Dados e Encoding**. A preparação de dados envolve a obtenção de um conjunto final de dados a partir de um conjunto inicial. Esta etapa visa realizar operações de limpeza e transformação de dados. A transformação de dados consiste em ajustar os dados para um formato que seja apropriado para uma específica tarefa de mineração de dados que, segundo [19] e [5], essas tarefas podem ser classificação, regressão e clusterização.

- **Construção do Modelo**. A aplicação do modelo consiste da utilização de algoritmos de aprendizado de máquina de forma a descobrir padrões e construir representações de parte de uma realidade empírica.

Esse assunto é apresentado com mais detalhes na próxima seção.

3.1.3 Preparação de Dados e Encoding

A construção do modelo de predição envolve uma fase de pré-processamento com algumas tarefas de preparação de dados e *encoding*. Entendemos que *encoding* é a obtenção de um vetor de variáveis capaz de ser interpretado por algoritmos preditivos [33] enquanto que a preparação de dados envolve a transformação e obtenção de variáveis, podendo envolver a captura de informações em outras fontes de dados, e também operações de limpeza e remoção de dados. Essa obtenção de um vetor de variáveis se dá por meio de técnicas de *encoding*, algumas técnicas e definições foram expostas na Seção 2.5.1 do Capítulo 2.

Para o algoritmo de predição alcançar um bom desempenho, é importante que o vetor de variáveis esteja bem representado e com informações que de fato possuem relevância para a predição [13]. O *encoding* visa principalmente à obtenção dessas variáveis, sejam por meio de transformações realizadas nos dados originais ou por meio de fontes externas. A preparação de dados é crucial no intuito de obter bons resultados [13].

O projeto experimental dessa dissertação foi realizado com algoritmos de aprendizado supervisionado. A escolha de algoritmos de aprendizado supervisionado se dá pela definição da situação de fracasso que foi utilizada como a variável resposta no estudo de caso. Essa variável pôde ser obtida por meio da verificação da diferença entre o momento da abertura do chamado e o momento do encerramento do chamado. Deste modo, durante o pre-processamento, essa variável é calculada por meio de duas etapas: (*i*) Tempo decorrido entre o primeiro evento registrado e o evento da iteração; e (*ii*) testar se tempo decorrido é maior que 720 minutos[3]. Essas duas etapas estão presentes no Algoritmo 1. Além disso, todas as etapas de descritas nessa seção seguiram de acordo com essa escolha.

O Algoritmo 1 assume a existência de duas funções (*i*) *agrupaOrdena* e (*ii*)

[3]Essa definição foi dada pela organização que nos enviou o conjunto de dados para estudo e pesquisa.

Algoritmo 1 Calcular o PPI.

Entrada: Histórico de eventos $\mathcal{L}$
Saída: $\mathcal{L}'$

1: $\mathcal{L}' \leftarrow agrupaOrdena(\mathcal{L})$
2: **Para cada** $c \in \mathcal{L}'$ **faça**
3: **Para cada** $e \in \mathcal{E}_c$ **faça**
4: $\#_{elapsedTime}(e) \leftarrow converteMinutos(\#_{time}(e) - \#_{time}(e(1))$
5: **Se** $\#_{elapsedTime}(e) \leq 720$ **então**
6: $\#_y(e) \leftarrow 0$
7: **Se não**
8: $\#_y(e) \leftarrow 1$
9: **fim Se**
10: **fim Para cada**
11: **fim Para cada**

converteMinutos. A primeira função tem como objetivo agrupar e ordenar o conjunto de todos os eventos identificados $\mathcal{E}$ por instância c do histórico de eventos $\mathcal{L}$. A segunda função faz uma conversão da diferença do tempo entre o evento da iteração e o primeiro evento da instância do processo e a retorna com a unidade de tempo representada em minutos. Em seguida, os atributos do tempo decorrido e o atributo que identifica o momento do fracasso são mapeados por meio da função de mapeamento representada por $\#_{elapsedTime}(e)$ e $\#_y(e)$ respectivamente. Por fim, o algoritmo tem como saída um histórico de eventos derivado ($\mathcal{L}'$) a partir do histórico de eventos inicial ($\mathcal{L}$).

Para recapitular, as instâncias de processo também possuem seus respectivos atributos (Definição 2). O Algoritmo 2 adiciona o atributo de fracasso na instância do processo. Deste modo, tem-se atributos de fracasso nos eventos e também nas instâncias. Seguimos desta maneira para evidenciar o momento exato que a instância obteve seu valor de fracasso diferenciando dos demais eventos. O Algoritmo 2 tem como saída um histórico de eventos derivado ($\mathcal{L}''$).

Para que abordagem siga de maneira preditiva, é necessário que a modelagem seja construída de uma maneira antecipada ao evento que demarca a situação de fracasso. Dado que a situação de fracasso torna-se conhecida por meio de um novo atributo construído pelo Algoritmo 1, sugerimos um corte de eventos a, no mínimo, um evento de antecedência ao evento que possui o atributo de fracasso valorado em 1, i.e., $\#_y(e) = 1$. A Figura 3.2 exemplifica a etapa de corte dos eventos em uma instância de processo.

Considerando a Figura 3.2, os eventos são representados pelos retângulos A,

Algoritmo 2 Obter um atributo indicador de fracasso da instância e atribuir a própria instância.

Entrada: Histórico de eventos $\mathcal{L}'$
Saída: $\mathcal{L}''$

1: $\mathcal{L}'' \leftarrow \mathcal{L}'$
2: **Para cada** $c \in \mathcal{L}''$ **faça**
3: $\widehat{\#_y(c)} \leftarrow 0$
4: **Para cada** $e \in \mathcal{E}_c$ **faça**
5: **Se** $\#_y(e) = 1$ **então**
6: $\widehat{\#_y(c)} \leftarrow 1$
7: **interrompa**
8: **fim Se**
9: **fim Para cada**
10: **fim Para cada**

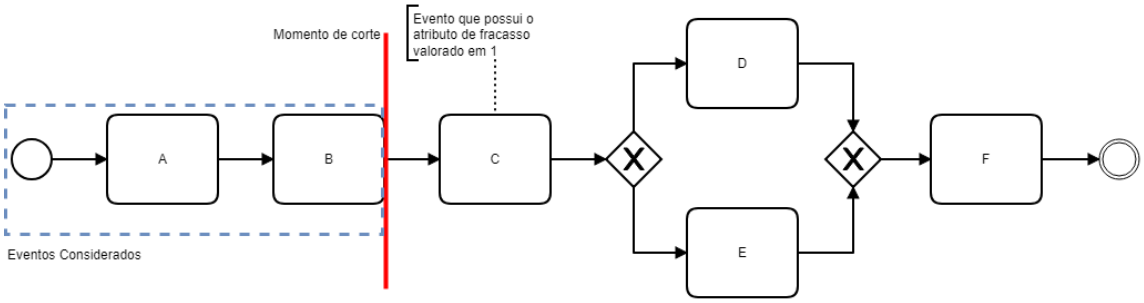


Figura 3.2: Momento de corte a um evento de antecedência a uma situação de fracasso.

B, C, D, E e F. Os eventos A e B antecedem um evento que possua o atributo da situação de fracasso da instância de processo e são os eventos considerados, representado por meio do retângulo pontilhado. Todos os outros eventos são removidos durante essa preparação pois precedem o momento de corte, representado pela barra de cor vermelha. Mediante a situação de fracasso, os próximos eventos já não possuem relevância para a predição pois a situação de fracasso já ocorreu. Sendo assim, não é necessário considerar os eventos posteriores na modelagem para o estudo de caso e cenário considerados.

Ao final, têm-se mais um histórico de eventos derivado $\mathcal{L}'''$ a partir de $\mathcal{L}''$. Essa abordagem não é apresentada na forma de algoritmos, como em 1 e 2, pois é possível utilizar o *DISCO* ou até mesmo o *ProM* para realizar o corte de eventos. As instâncias de processo que não possuem $\widehat{\#_y(c)} = 1$ não passam por cortes.

Após a remoção dos eventos que possuem $\#_y(e) = 1$, as instâncias de processo se encontrarão com números de eventos diferentes e o último evento de cada uma delas também será diferente, visto que, o fracasso das instâncias podem acontecer

em momentos distintos. Dito isso, o número de variantes de processo irá mudar em relação as variantes de processo do conjunto de dados inicial ($\mathcal{L}$).

Se utilizar o conjunto de dados com características muito distintas entre as instâncias que possuem fracassos e as instâncias que não possuem fracassos, o modelo preditivo poderá resultar em sobreajuste (*overfitting*)[4]. Um modelo em *overfit* apresentará uma precisão muito alta mas não estará com uma boa representação da realidade ao qual foi modelado [13]. Apesar de ser uma preocupação da construção do modelo de predição, esse problema deve ser resolvido durante a preparação dos dados.

Para evitar esse problema, sugerimos os seguintes passos:

1. Filtrar as instâncias de modo que $\widehat{\#_y(c)} = 1$;
2. Para cada instância, valorar o atributo $\#_y(e)$ do último evento em 1;
3. Aplicar a técnica *event window* para o conjunto de dados $\widehat{\#_y(c)} = 0$ pois $\widehat{\#_y(c)} = 1$ já passou por uma etapa de cortes. Outras técnicas também podem ser aplicadas como por exemplo a técnica *sliding window* para todo o conjunto de dados $\widehat{\#_y(c)} = 0$ e $\widehat{\#_y(c)} = 1$;
4. Obter um novo conjunto de dados a partir da técnica *event window*.

Os passos 1 e 2 estão presentes no Algoritmo 3.

Algoritmo 3 Valorar o ultimo evento de cada instância com fracasso.

Entrada: Histórico de eventos $\mathcal{L}'''$
Saída: $\mathcal{L}''''$

1: $\mathcal{L}'''' \leftarrow \mathcal{L}'''$
2: **Para cada** $c \in \mathcal{L}''''$ **faça**
3: **Se** $\widehat{\#_y(c)} = 1$ **então**
4: assuma que $t \in \mathcal{E}_c$ de tal modo que $1 \leq i < j \leq |t|, t(i) \neq t(j)$.
5: $\#_y(|t|) \leftarrow 1$
6: **fim Se**
7: **fim Para cada**

O Algoritmo 3 atribui ao ultimo evento de uma instância de processo testada por meio do mapeamento $\widehat{\#_y(c)} = 1$. A atribuição de valor ao último evento se dá por meio do mapeamento da cardinalidade dos eventos da instância do

[4]Mais informações podem ser encontradas no trabalho de [13], onde apresenta as muitas facetas do clássico problema dos modelos sobreajustados.

processo dado por $\#_y(|t|) \leftarrow 1$. Destacamos essa atribuição apenas para manter um indicador no nível do evento e também possibilitaria cortes mais distantes da verdadeira atribuição de fracasso. O Algoritmo 3 tem como saída um histórico de eventos $\mathcal{L}''''$ derivado a partir de $\mathcal{L}'''$.

Os passos 3 e 4 considera uma janela de tempo no intuito de equilibrar todo conjunto de dados em termos de quantidade de eventos e retorna um novo conjunto de dados.

Após a etapa de cortes, convém aplicar a descoberta de modelos de processo, como visto no fluxo apresentado na Seção 3.1.2, e realizar algumas operações de união entre o histórico de eventos $\mathcal{L}''''$ e o conjunto de dados que contém os modelos de processo com o intuito de enriquecer $\mathcal{L}''''$ com as informações dos modelos de processo.

Algoritmo 4 Unir instancias de processo com os modelos de processo descobertos

Entrada: Histórico de eventos $\mathcal{L}''''$
Saída: $\mathcal{L}'''''$
1: $\mathcal{L}''''' \leftarrow \mathcal{L}''''$
2: **Para cada** $c \in \mathcal{L}'''''$ **faça**
3: $\quad \widehat{\#_{variant}}(c) \leftarrow retornaVariante(\widehat{\#_{case}}(c))$
4: **fim Para cada**

O Algoritmo 4 assume a existência de uma função *retornaVariante* que consulta o conjunto de dados extraído a partir do *DISCO* utilizando como parâmetro de busca o atributo *case* e retorna a variante de processo referente a instância. Ao final, obtém-se um histórico de eventos $\mathcal{L}'''''$ derivado de $\mathcal{L}''''$ com o atributo mapeado por $\widehat{\#_{variant}}(c)$ referindo-se à variante que a instância de processo se caracteriza.

As variáveis relacionadas as variantes de processo podem estar mais relacionadas a variável que identifica os clientes por exemplo. Sendo assim, isso poderia explicar o motivo de tantas variantes de processo do estudo empírico. Além disso, essas variáveis permitem sumarizar e agrupar as instâncias do processo segundo algumas características relacionadas aos seus fluxos e, desta maneira, obter variáveis representativas dos respectivos valores médios do tempo decorrido dessas instâncias e que podem também ser de grande importância para o algoritmo de predição. A literatura chama essas variáveis que influenciam os resultados da predição de **variáveis previsoras** (*predictors variables*) [53].

Ao final das etapas apresentadas por meios dos algoritmos e outros passos descritos até aqui, espera-se obter um conjunto de dados ($\mathcal{L}'''''$) minimamente

representado com as seguintes estruturas dadas por (*i*) Atributos da instância do processo e (*ii*) Atributos dos eventos do processo.

Atributos da instância do processo:

- **Case**. Trata-se de um atributo que identifica o incidente tratado.
- **Events**. Trata-se de um atributo que soma o número de eventos que a instância de processo teve. Neste caso, este número terá o tamanho préfixado do tamanho da janela considerada. Este atributo pode ser obtido por meio das operações de união entre $\mathcal{L}''''$ e o conjunto de dados com os modelos de processo obtido por meio do algoritmo *Fuzzy Miner*;
- **Variant**. Trata-se de um atributo que identifica uma variante de processo. Este atributo trata-se do atributo construído $\#_{\widehat{v}ariant}(c)$.
- **Started**. Trata-se de um atributo que demarca o primeiro momento que um evento de uma determinada instância foi registrado. Este atributo pode ser obtido por meio das operações de união entre $\mathcal{L}''''$ e o conjunto de dados com os modelos de processo obtido por meio do algoritmo *Fuzzy Miner*;
- **Finished**. Trata-se de um atributo que demarca o último momento que um determinado evento de uma instância foi registrada. Este atributo foi obtido por meio das operações de união entre $\mathcal{L}''''$ e o conjunto de dados com os modelos de processo obtido por meio do algoritmo *Fuzzy Miner*;
- **Duration**. Trata-se de um atributo que calcula o tempo decorrido de uma determinada instância do processo. Este atributo foi obtido por meio das operações de união entre $\mathcal{L}''''$ e o conjunto de dados com os modelos de processo obtido por meio do algoritmo *Fuzzy Miner*;
- **Mean Duration**. Trata-se de um atributo que calcula o tempo médio de duração de todas as instâncias de processo por variantes de processo. Este atributo foi obtido por meio das operações de união entre $\mathcal{L}''''$ e o conjunto de dados com os modelos de processo obtido por meio do algoritmo *Fuzzy Miner*;
- **Y**. Trata-se de um atributo que representa se a instância de processo fracassou ou não. Este atributo pode variar entre $\{0, 1\}$. Este atributo trata-se do atributo construído $\widehat{\#_y}(c)$.

Atributos dos eventos do processo:

- **TimeStamp**. Trata-se de um atributo que identifica o momento que o evento foi registrado.
- **Activity**. Trata-se de um atributo que identifica a atividade que está relacionada ao evento;
- **ElapsedTime**. Trata-se de um atributo que identifica o tempo decorrido do processo até o evento registrado daquele instante, isto é, $\Delta = Started - TimeStamp$. Este atributo trata-se do atributo construído $\#_{elapsedTime}(e)$;
- **Y**. Trata-se de um atributo que representa se o evento viola as regras do SLA anteriormente definido. Este atributo pode variar entre $\{0, 1\}$. Este atributo trata-se do atributo construído $\#_y(e)$.

Os atributos acima mencionados tratam-se de atributos gerais. Os atributos podem apresentar alguma variação de acordo com o domínio de aplicação. Na Seção 3.2 apresentamos em detalhes os atributos do estudo de caso.

Ao final, o *encoding* consiste em representar o histórico de eventos final sobre um formato encapsulado em um vetor de variáveis. Entendemos que esse formato encapsulado é uma forma compacta e representativa das características do histórico de eventos final que contém as informações das instâncias do processo e também dos eventos. Essas informações são indicadas por propriedades ou variáveis. A este formato encapsulado como vetor de variáveis daremos o nome de *Event Encoding*, seguindo a definição dada por [33].

Para obter essa representação final, é preciso tratar os atributos dos eventos que apresentam variações entre um evento para o outro de uma mesma instância do processo. Diferentes técnicas de *encoding* podem ser aplicadas, tais como *boolean encoding*, *frequency-based encoding* e *index-based encoding*. Um exemplo do *Event Encoding* esperado é apresentado em 3.1.

$$v_i = [case, events, variant, started, finished, duration, mean, attr_1, attr_2, \cdots, Y] \quad (3.1)$$

Considerando *Event Encoding* apresentado em 3.1 como um vetor de variáveis, tem-se os atributos da instância e os atributos dos eventos evidenciado por $attr_1$,

$attr_2$ e ao final o indicador de fracasso calculado dado por Y. Este vetor é representado por v_i de maneira a permitir que cada índice i de v representa uma instância de processo sumarizada.

Com o *Event Encoding* gerado, inicia-se a construção do modelo de predição. A próxima seção apresenta a construção do modelo por meio de um fluxo de atividades.

3.1.4 Construção do modelo

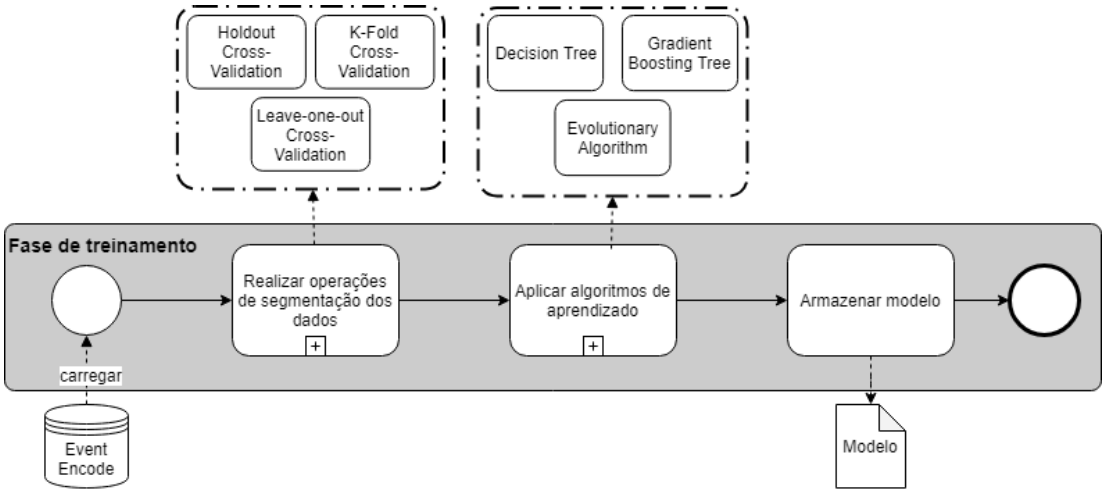
A construção do modelo de predição se dá por meio de um fluxo de atividades, também chamado de *pipeline* de processamento, que produz como resultado final a variável construída referente a instância do processo. O *pipeline* de processamento é apresentado na Figura 3.3 em duas fases (a) Pipeline referente a fase de treinamento e (b) Pipeline referente a fase de predição, avaliação e teste.

Considerando a Figura 3.3a, tem-se o fluxo de atividades da fase de treinamento dado por:

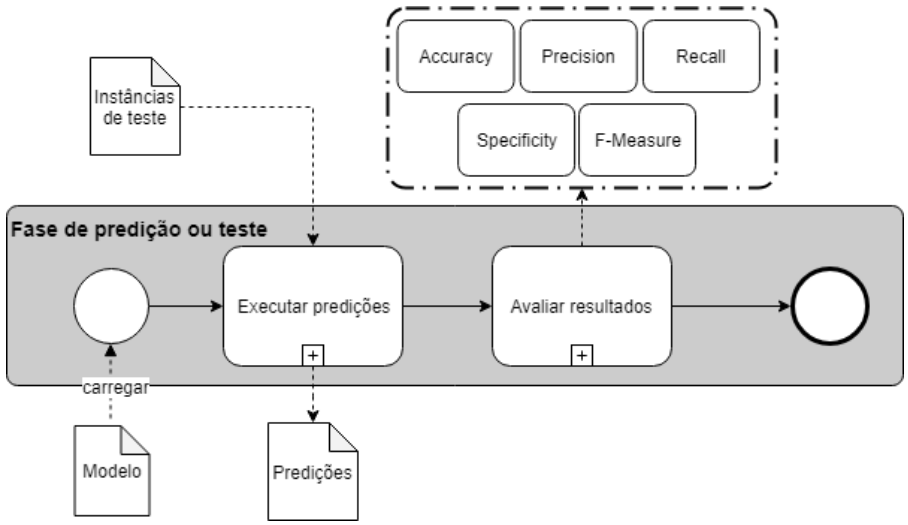
- **Realizar operações de segmentação dos dados.** Nesta atividade, a segmentação segue uma estratégia de separação do conjunto de dados entre instâncias de treinamento e instâncias de teste. Algumas estratégias destacadas são *Holdout Cross-Validation*, *K-Fold Cross-Validation* e *Leave-one-out Cross-Validation*, mais informações podem ser encontradas em [28].;
- **Aplicar algoritmos de aprendizado.** Nesta atividade, aplicam-se os algoritmos de aprendizado de máquina. Esses algoritmos tem como entrada o conjunto de instâncias de treinamento, obtidas por meio da atividade anterior. Alguns algoritmos destacados são *Decision Tree*, *Gradient Boosting Tree* e *Evolutionary Algorithm*;
- **Armazenar modelo.** Nesta atividade, o modelo treinado é guardado para avaliações futuras. Dependendo da estratégia de segmentação adotada, diversos modelos são gerados.

Na segunda fase da construção do modelo de predição (Figura 3.3b), tem-se um outro fluxo de atividades dado por:

- **Executar predições.** Nesta atividade, as predições são realizadas no conjunto de dados de teste obtido por meio da estratégia de segmentação ado-



(a)



(b)

Figura 3.3: Pipeline de construção do modelo de predição em duas fases. (a) Pipeline referente a fase de treinamento, (b) Pipeline referente a fase de predição, avaliação e teste.

tada produzindo os resultados, referentes a variável que indica o fracasso, que o modelo estima ser como assertivo;

- **Avaliar resultados.** Nesta atividade, a avaliação é realizada comparado os resultados que o modelo estimou com os resultados reais das instâncias de teste. Algumas estratégias de avaliação são *Accuracy*, *Precision*, *Recall*, *Specificity*

Algumas ferramentas computacionais podem ser utilizadas como auxílio para o cumprimento das atividades das fases da construção do modelo. Por exemplo,

ferramentas como o *WEKA*[5], *RapidMiner*[6] e *Orange*[7] são alguns exemplos de ferramentas que implementam os algoritmos de aprendizado de máquina tanto quanto os algoritmos que permitem fazer a segmentação dos dados. Também é possível utilizar linguagens de programação em conjunto com bibliotecas que podem ser importadas ao código fonte de maneira a seguir o *pipeline* sugerido utilizando programação tais como *scikit-learn*[8] ou *SparkMLlib*[9].

Em relação à avaliação da predição de fracassos aplicado ao estudo de caso, esse trabalho segue de maneira pragmática no que diz respeito à avaliação de métodos supervisionados. A próxima seção apresenta o projeto de avaliação em detalhes com suas devidas justificativas.

3.1.5 Método de Avaliação

A avaliação dos resultados será conduzida por meio de uma análise quantitativa. Segundo [42], o método quantitativo descreve um conjunto de técnicas que visa responder questionamentos de pesquisa sob uma enfase numérica. A escolha do método quantitativo por meio de experimentos se deve ao fato da pesquisa apoiar-se principalmente em dados e está destinada em examinar as relações de causa e efeito dos experimentos.

Para avaliar a confiança dos resultados obtidos, a pesquisa experimental adotará as seguintes medidas: Acurácia (*Accuracy*) [60], Precisão (*Precision*) [16], Sensibilidade ou Revocação (*Recall*) [16], Especificidade (*Specificity*) [16] e Medida F (*F Measure*) [60], definidas pelas equações 3.2, 3.3, 3.4, 3.5 e 3.6 respectivamente.

$$Accuracy = \frac{VP + VN}{VP + FN + FP + FN} \tag{3.2}$$

$$Precision = \frac{VP}{VP + FP} \tag{3.3}$$

[5]Mais informações podem ser vistas em `https://www.cs.waikato.ac.nz/ml/weka/`
[6]Mais informações podem ser vistas em `https://rapidminer.com/`
[7]Mais informações podem ser vistas em `https://orange.biolab.si/`
[8]Mais informações podem ser vistas em `https://scikit-learn.org/stable/`
[9]Mais informações podem ser vistas em `https://spark.apache.org/docs/latest/ml-guide.html`

$$Recall = \frac{VP}{VP + FN} \tag{3.4}$$

$$Specificity = \frac{VN}{VN + FP} \tag{3.5}$$

$$Fmeasure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.6}$$

Onde:

VP = Verdadeiro Positivo,

VN = Verdadeiro Negativo,

FP = Falso Positivo,

FN = Falso Negativo.

Essas medidas são amplamente utilizadas para avaliar técnicas de aprendizado de máquina pela literatura [33] e são balanceadas em termos comparativos entre as taxas de verdadeiros positivos e as taxas de falso positivo. Esse balanceamento encontra-se também para as taxas verdadeiro negativo e falso negativo. Dito isso, o método de predição adotado por este trabalho consiste da classificação de uma variável booleana (binária) que denota uma situação de fracasso ou de sucesso e, sendo assim, a escolha por essas medidas seria natural diante do contexto apresentado.

Considerando as equações 3.2, 3.3, 3.4, 3.5 e 3.6 tem-se:

- *Accuracy*. O número de instâncias corretamente classificadas na proporção do total de instâncias;
- *Precision*. A razão entre o número de instâncias verdadeiras corretamente preditas e o total de instâncias verdadeiras corretamente preditas e instâncias falsas incorretamente preditas;
- *Recall*. A razão entre o número de instâncias verdadeiras corretamente preditas e o total de instâncias verdadeiras corretamente preditas e incorretamente preditas;

- *Specificity*. A razão entre o número de instâncias falsas corretamente preditas e o total de instâncias falsas corretamente preditas e incorretamente preditas;

- *F Measure*. A média ponderada entre a precisão e a revocação de cada categoria, no caso do domínio tratam-se das situações de fracasso.

O uso de uma matriz de confusão também permite observar a taxa de erros do modelo de forma intuitiva. A principal ideia de uma matriz de confusão é comparar os valores preditivos com valores reais do modelo. Dito isso, com o uso da matriz de confusão é possível observar se a diferença entre os valores esperados e os valores reais ultrapassam limites aceitáveis.

3.2 Descrição dos Dados

O histórico de eventos abordado por esta pesquisa são os incidentes registrados pelo departamento de tecnologia da informação de uma empresa brasileira entre os anos de 2015 e início de 2016. Os incidentes são definidos por qualquer falha operacional de um serviço de TI, isto é, qualquer falha que envolva *softwares*, *hardwares* e quaisquer outros periféricos tais como servidores, impressoras e telefonia.

Neste estudo de caso, considera-se a existência de um acordo a nível de serviço (SLA) com tempo de resolução de problemas fixado em 720 minutos, i.e., cada instância de processo com o tempo decorrido desde o momento da sua abertura até o momento de fechamento maior que 720 minutos é considerada uma instância com uma situação de fracasso. O fracasso das instâncias do processo podem apresentar riscos ao negócio pois punições poderiam ser aplicáveis mediante tal evidência. Os fracassos das instâncias também podem evidenciar um momento que o modelo de processo idealizado não reflete mais a realidade atual da prestação dos serviços. Por meio dos fracassos evidenciados é possível mensurar o desacordo entre a realidade em que o modelo de processo foi idealizado com a realidade atual do negócio e prestação dos serviços.

O cenário considerado possui um total de 294628 eventos e 7259 instâncias de processo. O uso do *software* DISCO[10] foi considerado durante a análise exploratória dos dados. O histórico de eventos possui os seguintes atributos:

[10]Mais informações podem ser vistas em `https://fluxicon.com/disco/`

Atributos da instância do processo:

- **Case**. Trata-se de um atributo que identifica o incidente tratado.
- **OpenTime**. Trata-se de um atributo do registro o momento da abertura do incidente.
- **CloseTime**. Trata-se de um atributo que registra a data de encerramento do incidente.
- **Customer**. Trata-se de um atributo que identifica o cliente ou o responsável por abrir o incidente frente ao sistema.
- **Service**. Trata-se de um atributo que identifica se o incidente está relacionado a instalação e manutenção de softwares, hardwares ou manutenção da rede local.

Atributos dos eventos do processo:

- **EventId**. Trata-se de um atributo que identifica o evento em um nível transacional.
- **EventName**. Trata-se de um atributo para distinguir os nomes dos eventos ou atividades do processo.
- **Article**. Trata-se de um atributo que identifica se houve trocas de mensagens entre os envolvidos do incidente, i.e., entre quem abriu o incidente com o técnico responsável por atender o incidente.
- **Priority**. Trata-se de um atributo numérico que pode variar de 1 à 5.
- **TimeStamp**. Trata-se das informações do dia e a hora que o evento foi registrado pelo sistema. Um processo é caracterizado por uma cadeia de eventos, atividades e decisões. Este atributo é responsável por registrar o evento no sistema e por meio dele, tem-se a ordenação topológica da cadeia de eventos de uma instância de processo.

Por meio dos atributos **TimeStamp**, **Case** e **EventName** tem-se as propriedades que caracterizam um histórico de eventos (*Event Log*) como visto na Seção 3.1.1 e também em acordo com a Definição 2 apresentada na Seção 2.1 do Capítulo 2.

Após utilizar o DISCO e levantar informações do cenário e do comportamento do processo, foram identificadas 4047 variantes de processo e um total de 32 atividades distintas. A Tabela 3.1 apresenta os nomes das atividades e suas respectivas frequências e frequências relativas.

Tabela 3.1: Atividades registradas e respectivas frequências.

Atividade ou EventName	**Frequência**	**Frequência Relativa**
SendAgentNotification	47657	16,18%
Misc	38689	13,13%
TicketDynamicFieldUpdate	35430	12,03%
TimeAccounting	20380	6,92%
OwnerUpdate	17412	5,91%
AddNote	14509	4,92%
StateUpdate	13047	4,43%
SendAnswer	12826	4,35%
SLAUpdate	11853	4,02%
Unlock	11214	3,81%
Lock	11210	3,8%
CustomerUpdate	7949	2,7%
NewTicket	7259	2,46%
EmailCustomer	5562	1,89%
SendAutoReply	5292	1,8%
TypeUpdate	4533	1,54%
Move	4359	1,48%
PriorityUpdate	3156	1,07%
SetPendingTime	2558	0,87%
PhoneCallCustomer	1882	0,64%
FollowUp	1714	0,58%
TicketLinkAdd	1045	0,35%
EmailAgent	907	0,31%
Merged	851	0,29%
Forward	532	0,18%
Subscribe	446	0,15%
Unsubscribe	436	0,15%
PhoneCallAgent	55	0,02%
TicketLinkDelete	8	0%
WebRequestCustomer	4	0%
SendCustomerNotification	1	0%

Como se pode notar na Tabela 3.1, as atividades das instâncias de processo não seguem uma mesma distribuição. Dado que o histórico de eventos contempla apenas instâncias finalizadas, essa diferença entre as distribuições das atividades das instâncias de processo se dá pelo fato que algumas das atividades aparecem apenas para determinadas variantes de processo.

Durante a análise descritiva obteve-se também o modelo de processo por meio

do algoritmo de mineração de processos *Fuzzy Miner*. O modelo obtido é confuso e de difícil compreensão. Essa dificuldade está associada ao segundo desafio apresentado na Seção 1.1 do Capítulo 1.

Para ilustrar a discrepância entre o que de fato foi capturado pelo sistema de informação e o processo que se espera, tem-se na Figura 3.4 uma abstração de um processo de gerenciamento de incidentes, como entendido a partir do trabalho apresentado em [2] e, na Figura 3.5, tem-se o modelo descoberto com o uso do algoritmo de descoberta de modelos de processos implementado pelo *software* DISCO.

Os itens abaixo destacados descrevem as atividades do modelo de processo apresentado na Figura 3.4.

- **Registrar o incidente**. O registro do incidente acontece no primeiro nível do processo. Um agente é responsável por registrar alguns detalhes do incidente, trata-se de uma atividade caracterizada como o primeiro contato registrado de um incidente por intermédio de um agente.
- **Classificar o incidente**. A classificação do incidente precede a atividade do registro. Dependendo dos detalhes obtidos do primeiro contato, uma classificação é necessária para estabelecer alguma ordem de priorização de atendimento pelas equipes especializadas.
- **Realizar um diagnóstico inicial**. O diagnóstico tem como finalidade a obtenção maiores detalhes do incidente e possíveis resoluções do problema.
- **Realizar a Designação e Escalação funcional**. Após o diagnóstico, uma equipe mais especializada pode ser designada. A designação e escalação funcional de equipes mais especializadas, ou equipes de segundo nível, pode ser necessária caso a equipe de primeiro nível não consiga identificar o tipo de incidente e possíveis resoluções imediatas.
- **Investigar o incidente**. A investigação do incidente por meio de uma equipe de segundo nível envolve um maior aprofundamento técnico. Em seguida, a equipe de segundo nível reporta a solução para uma equipe de primeiro nível.
- **Solucionar o incidente e restaurar os serviços impactados**. Uma vez que a solução é conhecida, inicia-se a atividade de restauração dos serviços impactados.

- **Encerrar o Incidente**. O incidente é encerrado por meio de registros e comunicação entre os envolvidos. A equipe de primeiro nível registra a solução e as estratégias de restauração adotadas para uso futuro, formando uma base de conhecimento.

A Tabela 3.2 destaca três variantes do processo e, para exemplificar, apresenta as atividades agrupadas por variante e de maneira ordenada em relação a sua ocorrência como registrada pelo histórico de eventos.

Tabela 3.2: Atividades do processo agrupadas por variantes de processo. A tabela considera três variantes de processo selecionadas de maneira aleatória com o objetivo de exemplificar seu fluxo de atividades.

Variante 1	**Variante 2**	**Variante 3**
NewTicket	NewTicket	NewTicket
ServiceUpdate	ServiceUpdate	ServiceUpdate
SLAUpdate	SLAUpdate	SLAUpdate
CustomerUpdate	CustomerUpdate	CustomerUpdate
EmailCustomer	TicketDynamicFieldUpdate	EmailCustomer
SendAutoReply	TicketDynamicFieldUpdate	SendAutoReply
SendAgentNotification	TicketDynamicFieldUpdate	SendAgentNotification
SendAgentNotification	PhoneCallCustomer	Lock
Lock	OwnerUpdate	Misc
Misc	Lock	OwnerUpdate
OwnerUpdate	Misc	TypeUpdate
TypeUpdate	TimeAccounting	ServiceUpdate
ServiceUpdate	SendAnswer	SLAUpdate
SLAUpdate	Misc	AddNote
AddNote	TimeAccounting	TimeAccounting
TimeAccounting	TicketDynamicFieldUpdate	TicketDynamicFieldUpdate
TicketDynamicFieldUpdate	TicketDynamicFieldUpdate	TicketDynamicFieldUpdate
TicketDynamicFieldUpdate	StateUpdate	AddNote
PriorityUpdate	Unlock	TimeAccounting
SendAnswer	Misc	TicketDynamicFieldUpdate
Misc		TicketDynamicFieldUpdate
TimeAccounting		TicketDynamicFieldUpdate
TicketDynamicFieldUpdate		StateUpdate
TicketDynamicFieldUpdate		Unlock
TicketDynamicFieldUpdate		
StateUpdate		
Misc		
SendAnswer		
Misc		
TimeAccounting		
StateUpdate		
Unlock		
Misc		

As discrepâncias entre os modelos existem para manter o processo operante mediante a dinamicidade requerida do negócio. Também considera-se o fato

que os sistemas de TI têm diferentes formas de registrar e representar eventos históricos de um determinado processo [2,53]. Diante disso, guiar-se por meio da abstração de um modelo pode contribuir para a compreensão do domínio e dos objetivos estratégicos e operacionais do processo.

Dado que o cenário contempla os incidentes registrados dentro de um período de aproximadamente um ano (2016 - 2017), consideramos uma análise trimestral de modo a obter a distribuição dos incidentes agrupados pelo total de variantes de processo identificadas, total de incidentes abertos e também por total de clientes que registraram incidentes. Esta distribuição não considerou a data de encerramento do incidente. Deste modo, a Tabela 3.3 apresenta o detalhamento da distribuição dos incidentes.

Tabela 3.3: Distribuição trimestral dos incidentes registrados.

Trimestres	Primeiro	Segundo	Terceiro	Quarto
Total de variantes	1038	1037	1220	922
Total de incidentes abertos	1612	1585	2079	1981
Total de clientes	158	166	205	234

Outros agrupamentos poderiam ser incluídos na distribuição, tais como o tipo de serviço e os níveis de prioridade. Entretanto, não foram considerados pela Tabela 3.3 devido à impossibilidade de caracterizar os atributos. Por exemplo, não foi possível dizer se a prioridade de número 5 equivale a uma priorização de urgência ou de priorização leve. Diante disso, consideramos que o conteúdo exposto pela Tabela 3.3 é suficiente para descrever o cenário e o conjunto de dados inicial, isto é, sem envolver técnicas sofisticadas de pré-processamento além do agrupamento por trimestre.

Também é possível notar que as situações de fracasso não foram apresentadas em nenhum aspecto quantitativo. Neste caso, as situações de fracasso não estão evidentes de maneira direta pelo conjunto de dados inicial, isto é, faz-se a necessidade em calcular as situações de fracasso em termos de outros atributos. Sendo assim, iremos abordar abordar como resultados obtidos da preparação de dados e discutido na próxima seção.

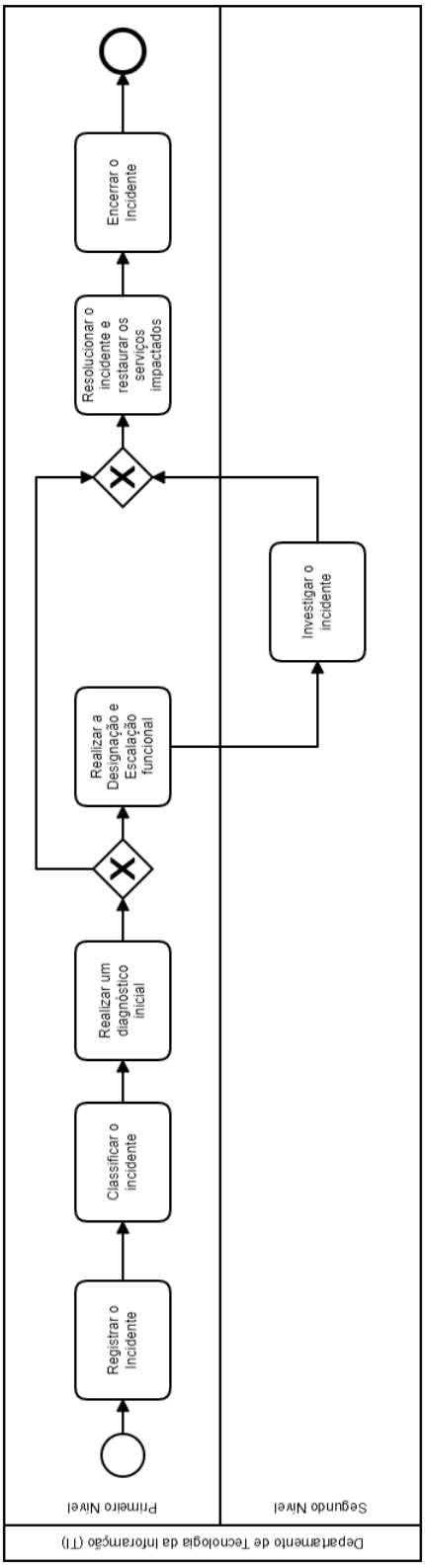


Figura 3.4: Abstração do modelo do processo de gerenciamento de incidentes. Um formato de comportamento esperado e modelado sobre a notação BPMN. Modelo adaptado de [2].

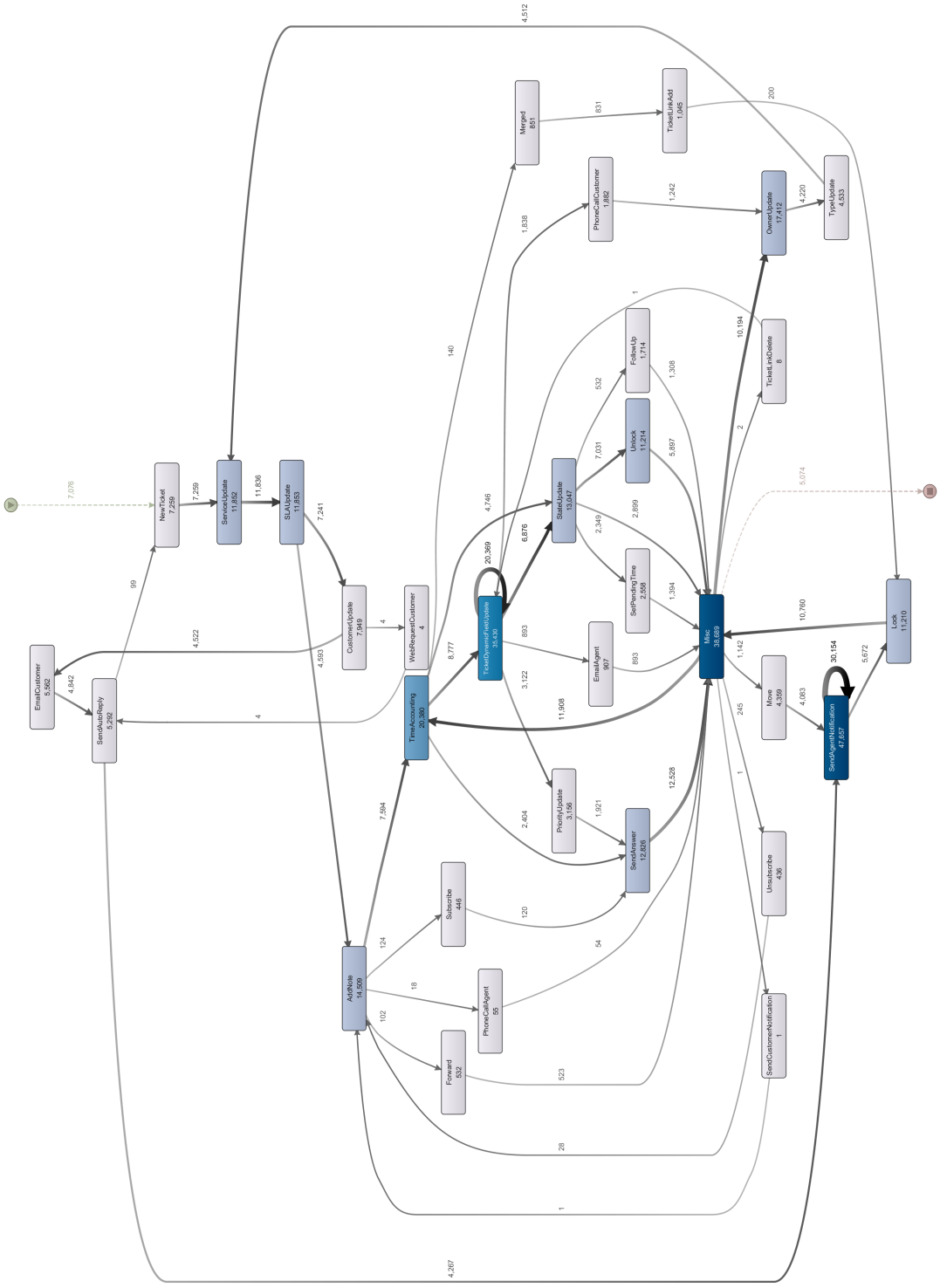


Figura 3.5: Modelo de processo obtido com o uso do DISCO.

3.3 Resultados Experimentais

Os resultados experimentais divide-se em duas seções. A Seção 3.3.1 apresenta o que nomeamos como resultados explanativos. Esses resultados referem-se as resultantes da Preparação de Dados e *Encoding* pois aumentam o poder discriminatório do histórico de eventos por meio da criação de novos atributos. A Seção 3.3.2 apresenta os resultados preditivos considerando o atributos que denota o fracasso das instâncias do processo e apresenta-os seguindo as medidas apresentadas por meio das equações 3.2, 3.3, 3.4, 3.5 e 3.6.

3.3.1 Avaliação dos Resultados Explanativos

Esta seção dedica-se em apresentar os resultados referentes aos atributos obtidos que o histórico de eventos inicial obteve ao longo das diversas transformações. O principal objetivo dos resultados explanativos é verificar que os atributos ou variáveis construídas podem ser variáveis previsoras do estudo de caso.

A Tabela 3.4 apresenta a análise trimestral do histórico de eventos seguindo de maneira análoga a Tabela 3.3, apresentada na Seção 3.2, com a diferença que as instâncias que fracassaram são conhecidas e evidenciadas por meio do atributo mapeado por $\widehat{\#y(c)}$. Os resultados apresentados não consideram a etapa de corte de eventos posteriores ao fracasso das instâncias.

Tabela 3.4: Distribuição trimestral dos incidentes registrados utilizando histórico de eventos derivado $\mathcal{L}''''''$.

Trimestres	Primeiro	Segundo	Terceiro	Quarto
Total de variantes	1038	1037	1220	922
Total de incidentes abertos	1612	1585	2079	1981
Total de clientes	158	166	205	234
Total de instâncias com fracasso	852	919	1079	966

Considerando a Tabela 3.4, é possível verificar que há trimestres que as instâncias que fracassaram em termos de objetivos estratégicos e operacionais do negócio ultrapassam em mais de 50% do total de instâncias do trimestre, isto é, do total de incidentes registrados.

De maneira a verificar que os atributos que indicam a variante de processo podem ser consideradas variáveis previsoras, apresentamos um gráfico de dis-

persão na Figura 3.6 onde é possível observar que algumas variantes de processo fracassaram com mais frequência. Desta maneira, algumas variantes são mais suscetíveis ao fracasso. O resultado apresentado não considera a etapa de corte de eventos posteriores ao fracasso das instâncias.

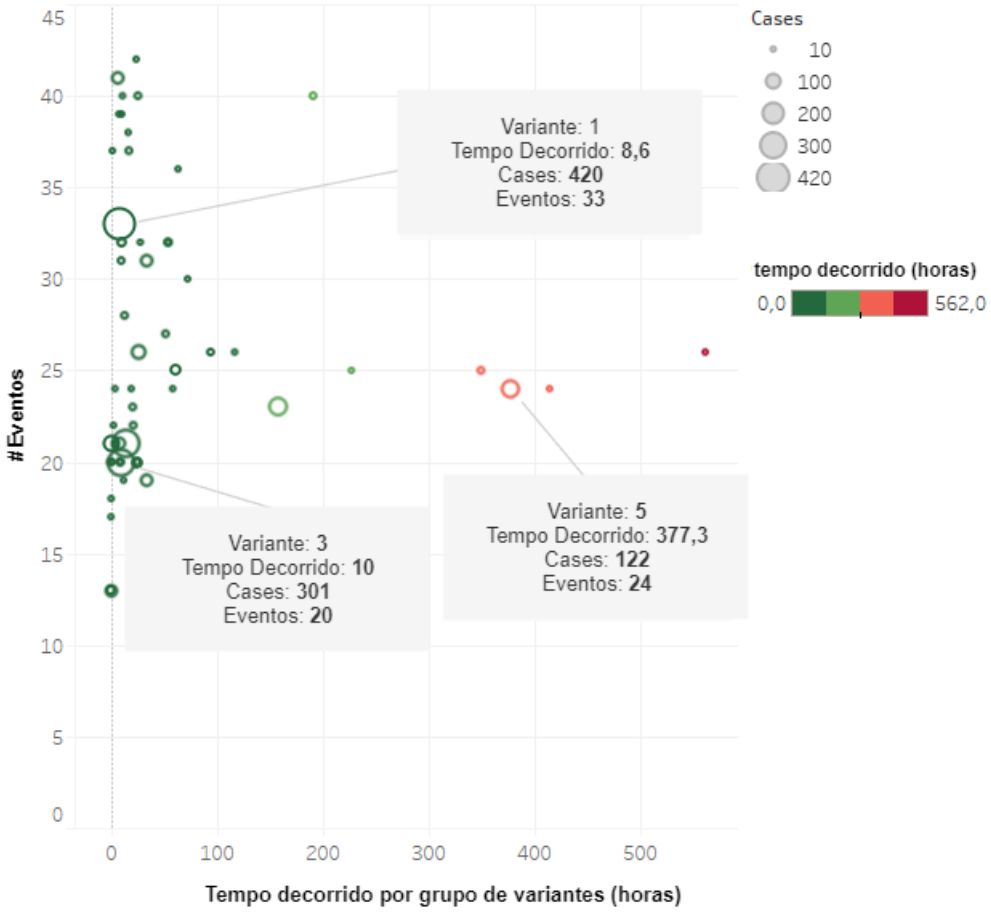


Figura 3.6: Gráfico de dispersão utilizado como verificação da capacidade de discriminação do conjunto de dados por meio das variantes de processo. No eixo-x: total de eventos que uma variante de processo possui; eixo-y: Tempo médio decorrido.

Na Figura 3.6, destacam-se três variantes (1, 3 e 5) em termos de volume de instâncias de processo onde as variantes 1 e 3 possuem o tempo médio decorrido inferior à 720 minutos, isto é, inferior à 12 horas. Por outro lado, a variante de processo identificada pelo número 5 é caracterizada por possuir um volume considerável de instâncias de processo (122 instâncias) que possuem tempo superior à 377 horas. Há muitas outras variantes de processo que poderiam ser consideradas como *outliers* do histórico de eventos por possuírem baixo volume de instâncias de processo.

Em seguida, utilizamos as três variantes de processo destacadas na Figura 3.6

e verificamos seu comportamento ao longo do tempo em uma série temporal para verificar se há alguma evidência de *Concept Drift*. A Figura 3.7 apresenta as três variantes (1, 3 e 5) em uma série temporal.

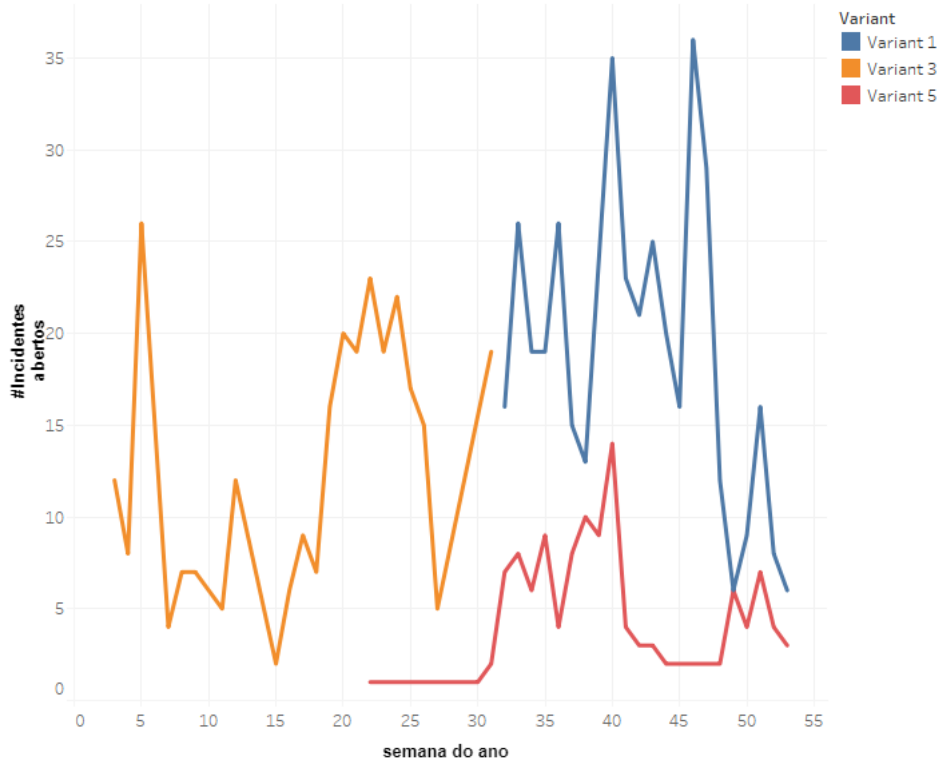


Figura 3.7: Variantes 1, 3 e 5 ao longo do período de observação. No eixo-x: total de incidentes abertos; eixo-y: semana do ano.

Considerando a Figura 3.7, é possível notar que a variante de processo identificada por 1 surge após a semana 30 do período observado. Por outro lado, não há mais registros da variante de processo 3. A variante de processo 5 coocorre com a variante de processo 1. Poder-se-ia formular uma hipótese que a variante de processo 3 foi substituída pela variante de processo 1 como se uma melhoria do processo de negócio tivesse sido implementada. Entretanto, essa afirmação só poderia vir diante de uma análise qualitativa. Diante de tal cenário, conclui-se que há evidência do *Concept Drift*, uma vez que, o comportamento do processo muda em termos de distribuição dos dados ao longo do período observado.

Uma vez que notamos evidências do *Concept Drift*, pode-se utilizar as técnicas dessa área para compreender melhor a mudança da distribuição de dados. Utilizamos a Mineração por Padrões Frequentes no histórico de eventos dividido por amostras de maneira a verificar se variantes de processo podem estar mais relacionadas a clientes específicos ou a outros atributos e como as mudanças se

comportam em diferentes partes do histórico de eventos. Essa escolha se dá diante do cenário multivariado, isto é, buscamos captar mudanças nas distribuições dos dados em atributos diferentes e coocorrentes.

Para alcançar tal objetivo, consideramos o histórico de eventos $\mathcal{L}''''$ e também o *pipeline* de construção do modelo, adotando como algoritmo a Mineração por Padrões Frequentes na atividade de treinamento. A estratégia de segmentação dos dados se deu por meio do particionamento do histórico de eventos contendo até 2000 instâncias de processo. Poderíamos ter seguido esse particionamento a nível mais granular, mas consideramos que o volume é suficiente para particionar em até 3 partes o histórico de eventos, número que consideramos capaz de detectar alguma mudança na distribuição dos dados. A Tabela 3.5 apresenta os resultados obtidos por meio da Mineração por Padrões Frequentes com suporte mínimo com os limiares configurados entre 10% e 70%.

Tabela 3.5: Resultados da detecção da mudança da distribuição de dados em partições do histórico de eventos.

Partição $0 - 2000$	**Partição** $2000 - 4000$	**Partição a partir de** 4000
VARIANT=3	VARIANT=1, CUSTOMER=40	VARIANT=1, CUSTOMER=40
CUSTOMER=16	VARIANT=1	VARIANT=2, CUSTOMER=16
CUSTOMER=40	VARIANT=2	VARIANT=1
CUSTOMER=57	CUSTOMER=5	VARIANT=2
CUSTOMER=85	CUSTOMER=16	CUSTOMER=16
	CUSTOMER=57	CUSTOMER=40
	CUSTOMER=85	CUSTOMER=57
		CUSTOMER=85

Considerando a Tabela 3.5, notamos a coocorrência do atributo que permite identificar a variante de processo com o atributo que permite identificar o cliente (*customer*). Deste modo, verificamos que algumas variantes de processo estão mais relacionadas com determinados clientes. Também é possível verificar outros comportamentos, tais como a variante de processo 2 (*VARIANT=2*) se ajustando ao cliente identificado por 16 (*CUSTOMER=16*) na última partição. Percebemos isso por meio da segunda partição onde a variante 2 e o cliente 16 aparecem mas não coocorrem com frequência. Diante disso, torna-se evidente a natureza dinâmica do histórico de eventos com mudanças em outros cenários na distribuição de dados, tais como o ajuste de determinados clientes em determinadas variantes de processo.

Nesta seção, verificamos que as variáveis construídas podem exercer grande influência na predição dos indicadores de fracasso da instância do processo. A próxima seção apresenta os resultados referentes a predição desses indicadores.

3.3.2 Avaliação dos Resultados Preditivos

A etapa de corte resultou em diversas variantes de processo cujo tamanho mínimo era de um total de 8 eventos. Isto implica que após o oitavo evento muitas instâncias fracassaram. Deste modo, utilizamos este resultado para delimitar o tamanho do *event window* considerado. Sendo assim, todas as instâncias de processo foram delimitadas para o total de 8 eventos a partir da abertura do incidente. Em seguida, inicia-se o *Event Encoding*.

A forma de *Event Encoding* considerada se deu por meio da combinação de duas técnicas vistas na Seção 2.5 do Capítulo 2. A primeira é aplicar o *boolean encoding* para os atributos dinâmicos e em seguida, para cada evento de uma instância do processo aplicar o *frequency encoding* nas resultantes do *boolean encoding*. Desta maneira, não se perdeu as informações estáticas e nem as informações dinâmicas e, por meio do *frequency encoding*, é possível identificar se há mudanças de um mesmo atributo em tempos distintos de uma instância do processo, sendo uma forma de lidar com os atributos dinâmicos. Após aplicar as técnicas de *encoding* um agrupamento foi considerado de maneira a obter um vetor de variáveis para cada instância de processo.

Ao obter o vetor de variáveis dado pelo *Event Encoding*, os seguintes algoritmos de classificação foram aplicados: *Naive Bayes* (NB), *Decision Tree* (DT), *Random Forest* (RF) e *Gradient Boosted Tree* (GBT)[11]. Para esta dissertação, a forma de segmentação entre os conjuntos de treinamento e testes da predição se deu por meio da técnica *k-fold cross-validation* que consiste em segmentar em k partições de tamanhos iguais. Uma das partes é reservada para o conjunto de teste e as outras $k-1$ partes são utilizadas na fase de treinamento. O particionamento de k partições se dá aleatoriamente. Por fim, a validação segue uma apuração da média das métricas formuladas na Seção 3.1.5 de k iterações [28].

A Tabela 3.6 mostra os parâmetros utilizados para cada um dos algoritmos aplicados exceto para NB cujas estimativas se dão por meio da máximo verossimilhança, mais informações podem ser encontradas em [23, 44].

[11] Mais informações sobre esses algoritmos podem ser vistas em [3, 5, 6, 23, 41]

Tabela 3.6: Parâmetros de treinamento dos algoritmos utilizados no *pipeline* de predição.

Decision Tree	
criterion	information gain
maximal depth	10
confidence	0.25
minimal gain	0.01
minimal leaf size	2
minimal size for split	4
minimal leaf size	2
number of prepruning	3
Random Forest	
criterion	information gain
number of trees	20
maximal depth	10
min rows	1.0
number of bins	20
learning rate	0.1
sample rate	1.0
minimal leaf size	2
number of prepruning	3
Gradient Boosted Tree	
number of trees	20
maximal depth	10
min rows	1.0
number of bins	20
learning rate	0.1
sample rate	1.0
minimal leaf size	2
number of prepruning	3

Tabela 3.7: Apuração média dos resultados da predição de fracassos.

	accuracy	*precision*	*recall*	*specificity*	*Fmeasure*
NB	63,02%	58,65%	87,14%	39,11%	70,11%
DT	64,40%	69,32%	51,10%	77,58%	58,83%
RF	66,95%	66,15%	68,84%	65,08%	67,46%
GBT	70,39%	68,44%	75,18%	65,65%	71,66%

Os resultados considerados na Tabela 3.7 referem-se à apuração média para

10-fold *cross-validation*. Como se pode notar, o algoritmo GBT se destaca em dois resultados: *accuracy*, *Fmeasure* enquanto que para *recall* obtem-se um valor acima de 75%. O algoritmo NB possui o pior resultado para *specificity* e também para *precision*.

A Tabela 3.8 apresenta a matriz de confusão obtida para cada um dos algoritmos de classificação da experimentação. A matriz é organizada com os dados pré-conhecidos obtidos por meio do cálculo do indicador de fracasso e com os valores resultantes da predição.

Dado os resultados apresentados, consideramos também outras experimentações com diferentes parâmetros para verificar se é possível alcançar melhores resultados. Neste caso, os algoritmos considerados são DT, RF e GBT. A Tabela 3.9 apresenta as novas parametrizações adotadas e os respectivos resultados nas Tabelas 3.10 e 3.11.

3.4 Considerações Finais

Este capítulo descreveu a proposta utilizada para o estudo de caso; apresentou os algoritmos e as técnicas adotadas em cada etapa; e, por fim, discutiu o método de avaliação adotado por meio de comparação entre os valores preditos e os valores reais.

Os resultados expostos no capítulo são capazes de responder à **QP** de uma maneira que evidencie a complementaridade de diferentes técnicas, não só de predição, mas também de toda uma densa fase de preparação de dados resultando em derivadas do histórico de eventos inicial. Diferentes técnicas combinadas tornaram-se complementares umas às outras ao longo do estudo empírico resultando em uma predição dos fracassos das instâncias do processo com os melhores resultados para o algoritmo GBT, onde todos os seus resultados foram acima de 65%. Também foi visto que é possível encontrar resultado melhores por meio da parametrização dos algoritmos aplicados.

A resultante do *pipeline* mostrou-se flexível o suficiente ao ponto de aplicar três algoritmos distintos de classificação sem que houvesse adaptações ao *Event Encode* obtido ao final.

Tabela 3.8: Matriz de confusão referente aos resultados de (a) NB; (b) DT; (c) RF; (d) GBT.

(a)

		Pré-conhecido		
		Positivo	Negativo	Total
Predição	Positivo	485	158	77,43%
	Negativo	755	1071	58,65%
	Total	39,11%	87,14%	

(b)

		Pré-conhecido		
		Positivo	Negativo	Total
Predição	Positivo	962	601	61,55%
	Negativo	278	628	69,32%
	Total	77,58%	51,10%	

(c)

		Pré-conhecido		
		Positivo	Negativo	Total
Predição	Positivo	807	383	67,82%
	Negativo	433	846	66,15%
	Total	65,08%	68,84%	

(d)

		Pré-conhecido		
		Positivo	Negativo	Total
Predição	Positivo	814	305	72,74%
	Negativo	426	924	68,44%
	Total	65,65%	75,18%	

Tabela 3.9: Parâmetros de treinamento dos algoritmos utilizados no *pipeline* de predição. Experimentação 2.

Decision Tree	
criterion	gini index
maximal depth	10
confidence	0.25
minimal gain	0.01
minimal leaf size	2
minimal size for split	4
minimal leaf size	2
number of prepruning	3
Random Forest	
criterion	gini index
number of trees	100
maximal depth	10
min rows	1.0
number of bins	20
learning rate	0.1
sample rate	1.0
minimal leaf size	2
number of prepruning	3
Gradient Boosted Tree	
number of trees	100
maximal depth	10
min rows	1.0
number of bins	20
learning rate	0.1
sample rate	1.0
minimal leaf size	2
number of prepruning	3

Tabela 3.10: Apuração média dos resultados da predição de fracassos. Experimentação 2.

	accuracy	*precision*	*recall*	*specificity*	*Fmeasure*
DT	65,17%	71,04%	50,69%	79,52%	59,16%
RF	67,07%	66,15%	69,32%	64,84%	67,70%
GBT	72,22%	73,07%	69,98%	74,44%	71,49%

Tabela 3.11: Matriz de confusão referente aos resultados de (a) DT; (b) RF; (c) GBT. Experimentação 2.

(a)

		Pré-conhecido		
		Positivo	Negativo	Total
Predição	Positivo	986	606	61,93%
	Negativo	254	623	71,04%
	Total	79,52%	50,69%	

(b)

		Pré-conhecido		
		Positivo	Negativo	Total
Predição	Positivo	804	377	68,08%
	Negativo	436	852	66,15%
	Total	64,84%	69,32%	

(c)

		Pré-conhecido		
		Positivo	Negativo	Total
Predição	Positivo	923	369	71,44%
	Negativo	317	860	73,07%
	Total	74,44%	69,98%	

4. Trabalhos Relacionados

Este capítulo destaca alguns trabalhos que visam à predição dentro de um contexto de monitoramento de processos de negócio encontrados em [9,14,40,45, 58] e também considera os trabalhos de [34] e [7]. Dado que a pesquisa realizada por esta dissertação de mestrado considera um estudo de caso que difere da literatura acima exposta, alguns comparativos são realizados considerando as seguintes perspectivas: (*i*) Event Encoding; (*ii*) Volume de instâncias de processo e total de eventos; (*iii*) Volume de variantes de processo; (*iv*) Forma de aprendizado adotada; (*v*) Tipo e objetivo do modelo adotado e (*vi*) Forma de avaliação. Uma breve síntese dos trabalhos é apresentada antes da comparação entre os diferentes trabalhos.

Em [32], os autores destacam que o principal problema relacionado a tarefas de monitoramento de processos está no fato que as abordagens reativas só identificam falhas operacionais no processo após sua ocorrência. Como solução, os autores sugerem que uma abordagem preditiva pode atuar na prevenção das falhas operacionais. A solução proposta pelos autores é treinar uma Árvore de Decisão utilizando um conjunto de variáveis a partir dos dados históricos. A solução consiste da predição das violações das restrições de negócio e é avaliada em um estudo de caso cujo domínio trata de um processo de diagnósticos médicos. A avaliação se dá por meio das métricas mais comuns dentro da área de aprendizado de máquina tais como *Accuracy*, *Precision*, *Recall*, *Specificity* e a forma de segmentação e separação entre conjunto de dados de treinamento e teste segue a proporção 80% – 20% respectivamente. Os autores também utilizam 10-*fold cross-validation*. A principal contribuição científica dos autores é um *framework* capaz de estimar a probabilidade do cumprimento do processo.

Em [47], os autores destacam que as pesquisas na área de Mineração de Processos têm focado em soluções de apoio operacional cujas análises se dão por meio

de métodos quantitativos. Destacam também que uma maneira de aprimorar esse apoio operacional é por meio da mineração de determinadas perspectivas, tais como a identificação de gargalos no fluxo operacional do processo que podem resultar em atrasos na execução e cumprimento do mesmo. Desta maneira, os autores endereçam que o problema relacionado ao atraso da execução das instâncias de processo se dá por meio desses gargalos. Como solução, os autores propõem a predição do tempo de execução do processo considerando informações que identificam possíveis gargalos. A solução é avaliada em um estudo de caso cujo domínio trata de uma central de atendimento a clientes de um banco ou instituição financeira. O método de avaliação adotado pelos autores se deu por meio de duas medidas: *absolute bias* para *accuracy* e *root mean squared error* (RMSE) para *precision*. Ao final, destacamos como principal contribuição uma metodologia de mineração de processos que considera outras perspectivas de análise que vai além das instâncias de processos. Essa perspectiva de análise é capaz de mensurar os atrasos por meio da observação da coocorrência de eventos de instâncias distintas. Os resultados possuem uma acurácia entre 30% – 40% em média.

Em [34], os autores também destacam a dificuldade em gerenciar processos de maneira proativa. Dito isso, a solução proposta pelos autores é descobrir um conjunto de padrões adotando uma janela de eventos (*Event Window*). A avaliação se dá por meio do estudo de caso de dois processos de gerenciamento de incidentes. Os autores também utilizam as medidas expostas na Seção 3.1.5 do Capítulo 3. O algoritmo adotado trata-se de um algoritmo genético cujo objetivo é a classificação do SLA considerando um cenário preditivo de modo a aprimorar ou fornecer o apoio operacional em estratégias de mitigação de riscos. Os autores conseguem resultados superiores frente a outras técnicas

Em [7], os autores destacam a necessidade que a área de Mineração de Processos tem por técnicas de monitoramento preditivo das atividades do processo. Uma vez que esta predição seja possível, um sistema de informação pode emitir alertas sobre eventos indesejáveis que podem ocorrer no futuro. O principal objetivo é desenvolver uma técnica de modelagem de processos de maneira preditiva baseada em técnicas oriundas da área de Mineração de Processos e Inferência Gramatical, mais informações acerca desta última podem ser encontradas em [59]. Como principal contribuição científica, destacamos um artefato produzido pelos autores que faz a predição de modelos de processo, isto é, de eventos de um processo. A forma de avaliação se dá também por meio das medidas apresentadas na Seção 3.1.5 do Capítulo 3.

Em [30], os autores destacam a dificuldade da predição dos possíveis resultados e saídas esperadas de um processo em execução, isto é, de processos incompletos. Como solução, os autores sugerem diferentes formas de *event encoding* de maneira a possibilitar o uso de algoritmos de classificação em tempo de execução do processo. A proposta realiza o *encoding* por meio da enriquecimento do *index-based encoding* com informações probabilísticas dos eventos. A metodologia aplicada se dá por meio de estudo de caso. São considerados dois processos reais. O primeiro processo consiste do registro histórico de pacientes de um hospital e o segundo de uma companhia de seguros. Os autores utilizam a área sob a curva ROC definida a partir da matriz de confusão como método de avaliação da proposta. A principal contribuição científica dos autores é mostrar que o enriquecimento das formas de *encoding* com informações probabilísticas melhora em alguns casos a confiabilidade da predição.

Considerando as perspectivas enumeradas por i, ii, iii, iv, v e vi (expostas no início deste capítulo), têm-se o seguinte quadro comparativo entre essas técnicas com a pesquisa exposta nesta dissertação de mestrado:

Tabela 4.1: Quadro comparativo de diferentes técnicas de encoding.

	Event Encoding			
Trabalhos	*boolean encoding*	*frequency-based encoding*	*index-based encoding*	*process-aware*
[32]			✓[1]	
[47]				✓
[34]			✓	
[7]				✓
[30]	✓	✓	✓	✓
Esta pesquisa	✓	✓	✓[2]	

A Tabela 4.1 apresenta uma coluna que demarca que a forma de *encoding* seguiu uma abordagem ciente de processos onde considera a transição entre as atividades do processo e o encadeamento entre elas. Na Tabela 4.2, os trabalhos que apresentam múltiplos valores implicam que houve mais de um estudo de caso envolvido durante a experimentação cientifica das propostas. Os valores representados por "–" significam que a informação não está presente nos artigos

[1]Essa informação não é explícita. Entretanto, os autores destacam a separação entre atributos estáticos e dinâmicos e também relatam a ocorrência de 623 atividades possíveis. Desta forma, se os autores utilizassem *boolean encoding* por exemplo, isto resultaria em mais de 623 variáveis.

[2]Consideramos que *index-based encoding* está indiretamente presente na abordagem por meio da variável que identifica a variante de processo.

Tabela 4.2: Quadro comparativo do total de registros. Consideram-se as instâncias, eventos e variantes de processo.

	Quantidade Total		
Trabalhos	*Instâncias*	*Eventos*	*Variantes*
[32]	1143	150291	–
[47]	7000	879591	–
[34]	7554 e –	67543 e 1055128	–
[7]	13087 e 3777	262000 e 36730	– e –
[30]	1143 e 1065	150291 e 16869	– e –
Esta pesquisa	7259	294628	4047

e trabalhos relatados.

As abordagens destacadas acima não consideraram o uso das informações das variantes de processos em suas abordagens e nem avaliaram o que elas poderiam agregar em um contexto de predição. A forma de aprendizado, tipo e objetivo do modelo adotado e, por fim, a forma de avaliação foram destacadas durante a síntese dos trabalhos relacionados.

5. Conclusão

No presente trabalho de pesquisa, apresentamos um estudo empírico que aplica diferentes técnicas da área de Mineração de Dados e Mineração de Processos de maneira complementar. Essa complementaridade se dá por meio da combinação dessas técnicas com o objetivo de preencher determinadas lacunas em alguns passos que a predição de uma variável ou atributo do processo necessita em termos de dependência. Por vezes, esse atributo é referenciado como indicador de desempenho (PPI) da instância do processo. É referenciado desta maneira pois ele indica se o processo possui o desempenho esperado em termos do alcance dos objetivos estratégicos e operacionais do processo. A motivação para a escolha do problema é a busca, através de monitoramento, por possibilidades de transição entre um cenário reativo para um cenário preditivo.

Este capítulo apresenta as principais contribuições da pesquisa (Seção 5.1), as limitações em termos de aplicabilidade (Seção 5.2) e, por fim, trabalhos futuros e potenciais caminhos que a pesquisa pode seguir a partir daqui (Seção 5.3).

5.1 Contribuições do Trabalho de Pesquisa

A abordagem adotada por esta dissertação de mestrado é diferente de outras abordagens utilizadas na literatura em alguns pontos. Portanto, este trabalho oferece uma contribuição para a área de Mineração de Processos ao utilizar técnicas e algoritmos combinados do estado-da-arte em um cenário cujo processo apresenta alta variação.

Consideramos que o estudo de caso considerado apresenta uma complexidade em alguns pontos, tais como:

- O volume de instâncias de processos e eventos considerados no estudo de caso é maior que a maioria dos trabalhos encontrados em [9,14,40,45,58];
- Não encontramos nenhum estudo de caso, considerando a literatura [9,14, 40,45,58], que apresente um número de variantes de processo tão alto como o que foi exposto durante este trabalho de pesquisa.

Diante dos pontos acima mencionados, consideramos que este trabalho de dissertação caminha por um cenário complexo que a literatura talvez ainda não tenha lidado em termos de volume de instâncias de processo, variantes de processo e outras características como a evidência do *Concept Drift* no histórico de eventos. Este excesso de variantes de processo está mais relacionado ao problema de balanceamento entre critérios de qualidade, como visto na Seção 1.1 do Capítulo 1.

Consideramos também como contribuição, o forte detalhamento e as técnicas apresentadas durante a preparação dos dados no sentido de obter novos atributos ou variáveis que, algumas vezes, são tratados como engenharia de variáveis (*Feature Engineering*). Utilizamos técnicas que detecta a mudança da distribuição de dados de maneira enriquecedora em uma etapa explanativa. O *Event Encoding* considerado segue uma abordagem especial, a sumarização das atividades se dá por meio dos resultados do *Fuzzy Miner* onde não seria necessário aplicar a forma *index-based encoding*, uma vez que, o fluxo está representado diretamente nas informações de variantes de processo. As formas *boolean encoding* e *frequency-based encoding* são utilizados de maneira complementar nos atributos dinâmicos.

Em relação à predição dos fracassos das instâncias do processo, foi verificado o potencial de quatro algoritmos de classificação de maneira a lidar com a falta de *benchmarks* internacionalmente aceitos.

Por fim, destacamos também o *pipeline* de processamento sugerido que mostrou-se bem flexível para a aplicação das diferentes técnicas e algoritmos. Concluímos que o trabalho passou por diversos desafios ainda em aberto da área de Mineração de Processos para alcançar a predição, tais como os desafios enumerados por 1, 2, 3, 4, 6,7, 8 e 9 (vistos na Seção 1.1 do Capítulo 1).

5.2 Limitações da Proposta

As principais dificuldades que surgiram no trabalho estão relacionadas à limpeza e pré-processamento de dados, quais sejam:

- **Ciclos ou laços.** Os ciclos em um processo são de interesse para um analista de mineração de processos. No entanto, este trabalho não verifica e não faz nenhum tipo de tratamento especial aos ciclos, que podem ser uma explicação para a grande quantidade de variantes de processo.
- **Clusterização.** Este trabalho não usou clusterização ou agrupamento de caminhos de execução similares, o que poderia melhorar o modelo criado e reduzir a grande quantidade de variantes de processo. Ao contrário, a única forma de agrupamento considerada se deu por meio da aplicação do algoritmo *Fuzzy Miner* que agrupa por variantes de processo.
- **Informações das variantes de processo.** Neste trabalho, informações relacionadas às variantes do processo foram obtidas usando ferramentas que aplicam os algoritmos de Mineração de Processos. Contudo, tal abordagem para a obtenção dessas informações ocorreu de maneira não automatizada, sendo necessário a intervenção manual.
- **Segmentação.** Existe uma grande quantidade de operações de segmentação, o que pode inviabilizar a execução da proposta de solução em tempo real, dado o alto custo computacional dessas operações.
- **Redução de dimensionalidade.** Este trabalho considerou apenas a técnica *event window* como uma técnica de redução de dimensionalidade pois um cumprimento máximo de eventos por instância de processo foi delimitado. No entanto, não houve uma revisão da literatura de outras técnicas que poderiam ser mais apropriadas ao estudo de caso.
- **Generalização do modelo proposto.** Consideramos que, certamente, o modelo proposto (o *pipeline* de execução) e assim como as técnicas de pré-processamento ainda carecem de adaptações para que sejam utilizadas em outros domínios.

5.3 Trabalhos Futuros

Após apresentar as principais contribuições e limitações deste trabalho de pesquisa, apresentamos alguns direcionamentos possíveis para pesquisas futuras.

- **Software ProM.** Durante a experimentação, foram utilizadas diferentes ferramentas de software, tais como o software Tableau para a geração de gráficos e *scripts* na linguagem de programação Python em cada etapa do processo proposto. Como um trabalho futuro, as funcionalidades destas ferramentas de software utilizadas na presente pesquisa podem ser consolidadas em uma implementação de um *plugin* para o software ProM. Dessa forma, a comunidade científica da área de mineração de processos poderá usar a abordagem consolidada em um *plugin*.

- **Descoberta de modelos de processo.** Destacamos como principal limitação, a utilização de algoritmos de descoberta de modelos de processo por meio de ferramentas externas. Diante disso, visamos a uma implementação dos algoritmos de descoberta de modelos de processo com a finalidade de garantir a automação de todo o pré-processamento.

- **Encoding.** Adotamos duas técnicas de *Event Encoding* de maneira complementar (*boolean encoding* e *frequency-based encoding*). No entanto, consideramos utilizar a técnica *index-based encoding* e comparar os resultados obtidos. Essa técnica também possibilita o uso de modelos *markovianos* escondidos (HMM) na etapa de predição [7, 30].

- **Concept Drift.** Também consideramos aprofundar os estudos relacionados ao Concept Drift e verificar se houve influência dos mesmos nos erros de predição dos algoritmos utilizados. Desta maneira, esperamos entender o comportamento para criar um ciclo evolutivo e de operacionalização das fases de treinamento dos modelos, considerando um descarte e um retreino para obter as representações de partes até então desconhecidas de uma realidade empírica.

- **Diferentes domínios.** Também consideramos aplicar os métodos apresentados em outros domínios e obter novos *benchmarks*. Consolidando desta maneira a pesquisa em um cenário preditivo aplicado em processos de negócio.

- **Otimização de Parâmetros.** Dado os resultados apresentados, observamos a possibilidade de resultados melhores ao configurar novos parâmetros dos algoritmos adotados. Dito isso, consideramos também um estudo aprofundado para obter os melhores parâmetros de um domínio.

Referências

[1] Business process comparison: A methodology and case study. In *Lecture Notes in Business Information Processing* (2017), vol. 288, pp. 253–267.

[2] Baier, T., Di Ciccio, C., Mendling, J., and Weske, M. Matching events and activities by integrating behavioral aspects and label analysis. *Software and Systems Modeling 17*, 2 (2018), 573–598.

[3] Bishop, C. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York* (2007).

[4] Bose, R. P. C., Van Der Aalst, W. M., Zliobaite, I., and Pechenizkiy, M. Dealing with concept drifts in process mining. *IEEE Transactions on Neural Networks and Learning Systems 25*, 1 (2014), 154–171.

[5] Bramer, M. Data for data mining. In *Principles of data mining*. Springer, 2016, pp. 9–19.

[6] Breiman, L. *Classification and regression trees*. Routledge, 2017.

[7] Breuker, D., Matzner, M., Delfmann, P., and Becker, J. Comprehensible predictive models for business processes. *MIS Quarterly 40*, 4 (2016), 1009–1034.

[8] Carvalho, J., Santoro, F. M., and Revoredo, K. A method to infer the need to update situations in business process adaptation. *Computers in Industry 71* (2015), 128 – 143.

[9] Casati, F., Ilnicki, S., Jin, L., Krishnamoorthy, V., and Shan, M.-c. Advanced Information Systems Engineering. In *CAiSE 2017* (2017), E. Dubois and K. Pohl, Eds., vol. 10253 of *Lecture Notes in Computer Science*, Springer International Publishing, pp. 13–31.

[10] Conforti, R., de Leoni, M., Rosa, M. L., van der Aalst, W. M., and ter Hofstede, A. H. A recommendation system for predicting risks across multiple business process instances. *Decision Support Systems 69* (2015), 1 – 19.

[11] del Río-Ortega, A., García, F., Resinas, M., Weber, E., Ruiz, F., and Ruiz-Cortés, A. *Enriching Decision Making with Data-Based Thresholds of Process-Related KPIs*. Springer International Publishing, Cham, 2017, pp. 193–209.

[12] Depaire, B., Swinnen, J., Jans, M., and Vanhoof, K. A process deviation analysis framework. *Lecture Notes in Business Information Processing 132 LNBIP* (2013), 701–706.

[13] Domingos, P. A few useful things to know about machine learning. *Communications of the ACM 55*, 10 (oct 2012), 78.

[14] Eds, P. J., Conference, I., and Hutchison, D. Advanced Information Systems Engineering. In *CAiSE 2015* (2015), J. Zdravkovic, M. Kirikova, and P. Johannesson, Eds., vol. 9097 of *Lecture Notes in Computer Science*, Springer International Publishing.

[15] Ensslin, L., Ensslin, S. R., Dutra, A., Nunes, N. A., and Reis, C. Bpm governance: a literature analysis of performance evaluation. *Business Process Management Journal 23*, 1 (2017), 71–86.

[16] Fawcett, T. An introduction to roc analysis. *Pattern recognition letters 27*, 8 (2006), 861–874.

[17] Fluxicon. How to understand the variants in your process, 2012.

[18] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM Computing Surveys 46*, 4 (2014), 1–37.

[19] García, S., Luengo, J., and Herrera, F. *Data Preprocessing in Data Mining*, vol. 72. 2015.

[20] Günther, C., and van der Aalst, W. Fuzzy mining–adaptive process simplification based on multi-perspective metrics. *Business Process Management* (2007), 328–343.

[21] Han, J., Cheng, H., Xin, D., and Yan, X. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery 15*, 1 (2007), 55–86.

[22] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. *SIGMOD Rec. 29*, 2 (May 2000), 1–12.

[23] Hastie, T., Tibshirani, R., and Friedman, J. The Elements of Statistical Learning. *Bayesian Forecasting and Dynamic Models 1* (2009), 1–694.

[24] Haykin, S. *Neural Networks: A Comprehensive Foundation*, 1st ed. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1994.

[25] Hompes, B. F. A., Maaradji, A., La Rosa, M., Dumas, M., Buijs, J. C. A. M., and van der Aalst, W. M. P. Discovering Causal Factors Explaining Business Process Performance Variation. In *CAiSE*, vol. 7328. 2017, pp. 177–192.

[26] Jagadeesh Chandra Bose, R. *Process mining in the large : preprocessing, discovery, and diagnostics*. PhD thesis, TUE : Department of Mathematics and Computer Science, 2012.

[27] Janiesch, C., Matzner, M., and Müller, O. Beyond process monitoring: A proof-of-concept of event-driven business activity management. *Business Process Management Journal 18*, 4 (2012), 625–643.

[28] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.

[29] La Rosa, M., van der Aalst, W. M. P., Dumas, M., and Milani, F. P. Business process variability modeling: A survey. 1–45.

[30] Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., and Maggi, F. M. Complex symbolic sequence encodings for predictive monitoring of business processes. In *International Conference on Business Process Management* (2015), Springer, pp. 297–313.

[31] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (Berkeley, Calif., 1967), University of California Press, pp. 281–297.

[32] Maggi, F. M., Di Francescomarino, C., Dumas, M., and Ghidini, C. *Predictive Monitoring of Business Processes*. Springer International Publishing, Cham, 2014, pp. 457–472.

[33] Marquez-Chamorro, A. E., Resinas, M., and Ruiz-Cortes, A. Predictive monitoring of business processes: a survey, 2017.

[34] Márquez-Chamorro, A. E., Resinas, M., Ruiz-Cortés, A., and Toro, M. Run-time prediction of business process indicators using evolutionary decision rules. *Expert Systems with Applications 87* (2017), 1–14.

[35] Maruster, L. *A machine learning approach to understand business processes*. PhD thesis, Technische Universiteit Eindhoven, 2003.

[36] Mitchell, T. M. Decision Tree Learning, 1997.

[37] Mitchell, T. M. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill 45*, 37 (1997), 870–877.

[38] Moulines, C. U. Introduction: Structuralism as a program for modelling theoretical science. *Synthese* (2002).

[39] Parmenter, D. *Key Performance Indicators: Developing, Implementing, ans Using Winning KPIs*. No. 3. 2015.

[40] Pernici, B., and Weske, M. *Business Process Management*, vol. 9850 of *Lecture Notes in Computer Science*. Springer International Publishing, 2016.

[41] Quinlan, J. R. Induction of Decision Trees. *Machine Learning* (1986).

[42] Recker, J. *Scientific Research in Information Systems: A Beginner's Guide*. Springer Publishing Company, Incorporated, 2012.

[43] Rozinat, A., and Van der Aalst, W. M. Conformance checking of processes based on monitoring real behavior. *Information Systems 33*, 1 (2008), 64–95.

[44] Russell, S. J., and Norvig, P. *Artificial Intelligence: A Modern Approach*, third edit ed. 2010.

[45] Scekic, O., Truong, H. L., and Dustdar, S. Advanced Information Systems Engineering. In *CAiSE 2016* (2016), S. Nurcan, P. Soffer, M. Bajec, and J. Eder, Eds., vol. 9694 of *Lecture Notes in Computer Science*, Springer International Publishing, pp. 688–703.

[46] Schlimmer, J. C., and Granger, R. H. Incremental Learning from Noisy Data. *Machine Learning* (1986).

[47] Senderovich, A., Weidlich, M., Gal, A., and Mandelbaum, A. Queue mining–predicting delays in service processes. In *International Conference on Advanced Information Systems Engineering* (2014), Springer, pp. 42–57.

[48] van der Aa, H., del Río-Ortega, A., Resinas, M., Leopold, H., Ruiz-Cortés, A., Mendling, J., and Reijers, H. A. Narrowing the business-it gap in process performance measurement. In *International Conference on Advanced Information Systems Engineering* (2016), Springer, pp. 543–557.

[49] van der Aalst, W. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, 1 ed. Springer-Verlag Berlin Heidelberg, 2011.

[50] Van Der Aalst, W., et al. Process mining manifesto. In *Lecture Notes in Business Information Processing* (2012), vol. 99 LNBIP, pp. 169–194.

[51] Van Der Aalst, W. M. Business process management: a comprehensive survey. *ISRN Software Engineering* (2013).

[52] van der Aalst, W. M. Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining. *Asia Pacific Business Process Management: First Asia Pacific Conference, AP-BPM 2013, Beijing, China, August 29-30, 2013. Selected Papers 159* (2013), 1–22.

[53] van der Aalst, W. M. *Process Mining: Data Science in Action*, 2 ed. Springer-Verlag Berlin Heidelberg, 2016.

[54] Van der Aalst, W. M., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G., and Weijters, A. J. Workflow mining: A survey of issues and approaches. *Data & knowledge engineering 47*, 2 (2003), 237–267.

[55] Wazlawick, R. *Metodologia de pesquisa para ciência da computação*, vol. 2. Elsevier Brasil, 2017.

[56] Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., and Petitjean, F. Characterizing concept drift. *Data Mining and Knowledge Discovery 30*, 4 (2016), 964–994.

[57] Weijters, A., van Der Aalst, W. M., and De Medeiros, A. A. Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven, Tech. Rep. WP 166* (2006), 1–34.

[58] Weske, M. Business Process Management. In *Business Process Management* (Cham, 2015), H. R. Motahari-Nezhad, J. Recker, and M. Weidlich, Eds., vol. 9253 of *Lecture Notes in Computer Science*, Springer International Publishing, pp. 5–11.

[59] Wieczorek, W. Grammatical inference: Algorithms, routines and applications. sci, vol. 673, 2017.

[60] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[61] Zhao, D., Bu, L., Alippi, C., and Wei, Q. A Kolmogorov-Smirnov Test to Detect Changes in Stationarity in Big Data. *IFAC-PapersOnLine 50*, 1 (2017), 14260–14265.