INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Aplicabilidade da Mineração de Processos: um Caso de Estudo

ANA CAROLINA FERREIRA BARROS
Outubro de 2018

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Aplicabilidade da Mineração de Processos: um Caso de Estudo

Ana Carolina Ferreira Barros

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Engenharia de Software

Orientador: Isabel de Fátima Silva Azevedo
Co-Orientador: Nuno Alexandre Pinto da Silva

Júri:
Presidente:

Vogais:

Porto, outubro 2018

Dedicatória

Dedico esta dissertação aos meus pais pelo apoio incondicional ao longo desta jornada, e namorado, Manuel Correia, pelo amor, motivação e compreensão.

Resumo

A utilização de tecnologia tem evoluído de forma exponencial e também a necessidade de análise de dados para identificar padrões, obter métricas e descobrir tendências.

Há cada vez mais dados disponíveis sobre processos de negócios e a mineração de processos visa preencher a lacuna que se verifica entre modelos de processos tradicionais e técnicas de análises de dados. É usada para extrair conhecimento através de registos de eventos. As técnicas de mineração de processos permitem descobrir novos modelos de processos, monitorizar e melhorar os processos já existentes.

É na área da saúde que se enquadra o estudo de caso apresentado. Esta tem-se tornado um dos grandes focos de análise de processos de negócios, devido à crescente urgência em descobrir padrões, eliminar ineficiências de modo a melhorar a qualidade dos serviços, e, simultaneamente, reduzir custos financeiros e temporais. Devido à diversidade, dinamismo e complexidade dos processos médicos, torna-se difícil analisar os dados recolhidos e encontrar padrões. Grande parte dos dados recolhidos nos hospitais advêm do diagnóstico e tratamento dos pacientes. A aplicabilidade da mineração de processos na área da saúde auxilia à análise e monitorização dos percursos realizados por cada paciente, bem como, à deteção de anomalias, desvios e estrangulamentos nos processos existentes.

Esta dissertação apresenta uma abordagem capaz de determinar processos de negócio relacionados com tratamentos clínicos obtidos através de registos de eventos armazenados na base de dados *Medical Information Mart for Intensive Care III*.

Para atingir os objetivos propostos foi desenvolvido uma abordagem dividida por etapas utilizando a ferramenta *ProM*. Numa primeira fase os dados foram extraídos e mapeados para serem importados para a ferramenta ProM, o que realizado numa segunda fase, com algumas atividades de pré-processamento. Após importação, os dados foram transformados utilizando plugins específicos, e de seguida, aplicaram-se algoritmos de descoberta de processo. O resultado final são modelos de processo em notação *business process model and notation* ou *rede de petri*. Por fim, os modelos de processo gerados foram processados e avaliados de acordo com métricas definidas.

Palavras-chave: business process model and notation, descoberta de processos, mineração de processos, processos de negócio, rede de Petri, registo de eventos

Abstract

Technology as a whole has exponentially evolved throughout the last few decades. There is also a clear need for data analysis to indentify patterns, obtain metrics and discover tendencies.

Everyday, the number of data available about business processes is increasing. For this reason, the art of process mining aims to fil the gap between traditional process models and data analysis techniques. Therefore, it is used to extract knoweldge from event rigistries. Process mining techniques allow the discovery of new process models, monitoring and improvement of existent processes.

Health care has been one of the most important areas of activity today where there is a clear need for pattern discovery which permit the improvement of the overall quality of services and, at the same time, drastically reduce temporal and financial costs. Moreover, due to the highly diversified, dynamic and complex medical processes, it is substantially difficult to analyse collected data and find patterns in them. Almost the entirety of collected data in health centers such as hospitals come from the diagnosis and treatment of patients. The applicability of process mining in health care greatly helps the analysis and motorization of the overall paths taken by each patient in addition to the detection of anomalies, deviations and bottlenecks in existing processes.

This dissertation presents a possible approach capable of determining medical processes obtained from multiple event logs, which are stored in the Medical Information Mart for Intensive Care III database.

In order to achieve this dissertation objectives, a multiple phased approach was developed by using the *ProM* tool. In a first stage, the data provided by the medical center was extracted and mapped. At a second phase, the data was transformed using multiple specific plugins in addition to the application of knowledge process discovery algorithms. The final result is several process models in the Business Process Model and Notation and Petri Net notations. Finally, the generated process models were processed and evaluated according to multiple defined metrics.

Agradecimentos

Em primeiro lugar, quero agradecer aos meus pais por todos os sacrifícios que fizeram em prol da minha formação académica. Por acreditarem nas minhas capacidades e mostrarem que irão estar sempre por perto incondicionalmente.

À minha restante família por todas as mensagens de apoio e incentivos constantes.

Um especial agradecimento ao Manuel Correia, meu colega de faculdade e namorado, por todo o carinho, paciência e dedicação ao longo desta caminhada. Sempre acreditou e ajudou a alcançar os meus objetivos pessoais, académicos e professionais.

A todos os meus colegas de curso com quem tive oportunidade de trabalhar, um muito obrigado pelo companheirismo, interajuda e solidariedade.

Agradeço também a todos os docentes do Departamento de Engenharia Informática por terem contribuído para a minha formação académica e por terem conseguido incutir ainda mais o gosto pela tecnologia e pelo desenvolvimento de software.

Uma palavra de total agradecimento aos meus orientadores, Isabel Azevedo e Nuno Silva, pela disponibilidade, dedicação e conhecimento transmitido durante o desenvolvimento desta dissertação.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Acrónimos e Siglas

AHP	Analytic Hierarchy Process.
BPM	Business Process Management.
BPMN	Business Process Model and Notation.
CCU	Coronary Care Unit.
CMED	Cardiac Medical.
CPT	Current Procedural Terminology.
CSRU	Cardiac Surgery Recovery Unit.
CSURG	Cardiac Surgery.
EPC	Event-Driven Process Chain.
FFE	Fuzzy Front End.
GU	Genitourinary.
GYN	Gynecological.
HTTP	HyperText Transfer Protocol.
ICD	International Classification of Diseases.
IoC	Internet of Content.
IoE	Internet of Events.
IoL	Internet of Locations.
IoP	Internet of People.
IoT	Internet of Things.
MED	Medical.
MICU	Medical Intensive Care Unit.
MIMIC III	Medical Information Mart for Intensive Care IIII Information Mart for Intensive Care III.
MXML	Mining eXtensible Markup Language.
NCD	The New Concept Development.
NMED	Neurologic Medical.
NPD	New Product Development.
NSURG	Neurologic Surgical.
ORTHO	Orthopaedic.
SICU	Surgical Intensive Care Unit.

SSDI	Social Security Death Index.
SURG	Surgical.
TSICU	Trauma Surgical Intensive Care Unit.
TSURG	Thoracic Surgical.
UML	Unified Modeling Language.
VSURG	Vascular Surgical.
WFM	Workflow Management.
XES	Extensible Event Stream.
YAML	YAML Ain't Markup Language.

Capítulo 1

Introdução

Este capítulo introduz o contexto (secção 1.1), o problema (secção 1.2) e os objetivos do projeto descrito neste documento (secção 1.3). Complementarmente, descreve resumidamente a abordagem e o processo de desenvolvimento (secção 1.4). Por fim, apresenta a estrutura do documento, de modo a facilitar a sua compreensão (secção 1.5).

1.1 Contexto

Atualmente a maior parte das empresas necessitam e estão cada vez dependentes da tecnologia, mais concretamente, dos sistemas de informação (DataQlick 2015). Estes sistemas estão em constante evolução pois a competitividade de mercado, as necessidades do cliente e qualidade do produto/serviço são alguns dos fatores preponderantes na decisão de expandir o sistema.

A quantidade de dados gerados no dia-a-dia derivados dos eventos que cada pessoa executa é colossal. Realizar um pagamento de uma compra com o cartão de crédito, realizar uma chamada ou enviar um e-mail são apenas exemplos simples de atividades do quotidiano que levam à criação de dados de eventos. O termo Internet of Events (IoE) refere-se a todos os dados de eventos disponíveis. IoE é composto por: Internet of Content (IoC), Internet of People (IoP), Internet of Things (IoT) e Internet of Locations (IoL). Cada um representa a geração de informação a partir de fontes distintas:

- IoC: Representa a informação gerada por pessoas, de forma a aumentar o conhecimento sobre assuntos específicos. Exemplos são páginas web, artigos, *e-books* e *newsfeed*;
- IoP: Representa a informação gerada a partir da interação social (e-mail, redes sociais e fóruns);
- IoT: Representa a informação gerada a partir dos objetos físicos ligados a uma rede;
- IoL: Representa a informação gerada a partir de dispositivos com dimensão especial.

Os sistemas de informação das organizações lidam com grandes conjuntos de informação (Buhl et al. 2013) que geram grandes desafios às próprias organizações. Um dos desafios consiste em extrair informação através de registos de eventos, de forma a analisar, diagnosticar e melhorar os processos inerentes.

Mineração de processos (W. Van Der Aalst 2016a) é uma área de investigação que relaciona as áreas mineração de dados com técnicas de análise de processos.

Pertencente à área da ciência dos dados, aprendizagem automática (*machine learning*) relaciona-se com o estudo e desenvolvimento de algoritmos capazes de fornecer capacidade de aprendizagem às máquinas, de forma a não serem explicitamente programadas. A aprendizagem usa o poder das estatísticas e aprende com o conjunto de dados, usando regressões ou classificações. Dentro da área aprendizagem automática (*machine learning*), existem algoritmos que focam-se na descoberta de modelos, padrões e outras regularidades nos dados (Mitchell 2006).

O principal objetivo da área mineração de dados é descobrir padrões, relações e extrair informações a partir de grandes conjuntos de dados. Os dados são revistos pelos padrões e os critérios são aplicados para determinar quais são as relações mais frequentes (Witten et al. 2016). Durante o processo são usados algoritmos da área aprendizagem automática (*machine learning*).

Mineração de dados e aprendizagem automática (*machine learning*) são duas áreas em que o desenvolvimento é acelerado, devido aos avanços na pesquisa de análise de dados, crescimento das base de dados industriais, crescimento dos componentes computacionais e sobretudo às necessidades do mercado.

A área mineração de eventos carateriza-se por dispor técnicas de extração de conhecimento automáticas e eficientes a partir de registos de eventos (Tao Li 2015). Os eventos são temporais, podem ser definidos como ocorrências e habitualmente envolvem mudanças nos estados do sistema. Sistemas físicos, computacionais e sociais são exemplos de sistemas que distribuem eventos, dos quais, são armazenados como registos de eventos: registos de eventos de transações,registos de eventos de sensores, registos de eventos de pedidos HyperText Transfer Protocol (HTTP) e registos de eventos referentes a trafégo de rede.

Mineração de processos, área de investigação/conhecimento, visa preencher a lacuna que existe entre modelos de processos tradicionais e técnicas de análise de dados (Kantardzic 2011). As técnicas de mineração de processos assumem que é possível gravar sequencialmente eventos de tal forma que cada evento possa referenciar uma atividade, e dessa forma se torne possível (i) detetar desvios, (ii) verificar conformidade, (iii) analisar estrangulamentos, (iv) comparar variabilidades e propor melhorias nos processos.

O aumento da quantidade de dados provenientes de vários sistemas e de diversos domínios exponencia o crescimento de vários setores. Por exemplo, no campo da ciência, um dos maiores telescópios em operação desde 2000, *Sloan Digital Sky Survey*, gerou 150 terabytes de dados durante 15 anos de operação (Sloan Digital Sky Survey 2015). A utilização desses dados e os resultados obtidos substanciou mais de 5800 artigos científicos com cerca de 250 mil citações.

1.2 Definição do Problema

Processos são parte integrante do mundo de hoje, conduzem serviços e funcionalidades internas em organizações. No entanto, o software aplicacional das organizações nem sempre é desenvolvido considerando os processos de negócio. Durante a sua evolução nem sempre são acompanhados pelas aplicações, pelo que nem sempre os processos de negócio estão mapeados e alinhados com os processos disponibilizados pelas aplicações, o que causa problemas de eficácia, eficiência e usabilidade das aplicações.

Existe uma carência na definição de processos que definem e auxiliam os processos de negócio das organizações. Assim surge a necessidade de realizar a ponte entre a análise de dados de eventos e processos, dado que, as práticas atuais para monitorização e análise de execução de processos possuem falhas e defeitos (Park e Kang 2016).

Hoje em dia, é fundamental as empresas apresentarem fatores de distinção entre as restantes do mercado. Sentem a necessidade de aumentar a qualidade dos seus produtos e serviços, mantendo o foco na melhoria contínua. É essencial optimizar e identificar novos processos de négocio, quer a nível interno para uma melhor eficiência dos seus recursos, quer a nível externo, para uma melhor relação com os seus intervenientes.

1.3 Objetivos

O objetivo deste trabalho é desenvolver uma solução de software capaz de determinar os processos de negócio a partir de dados de eventos gerados pelo uso de aplicações usadas em várias organizações, e dessa forma determinar em que medida os dados dos eventos e o seu processamento pelos algoritmos de mineração de processos são capazes de gerar artefactos significativos para a análise, (re)definição e correção de processos de negócio e respetivas aplicações de software de suporte. Especificamente, pretende-se:

(i) Compreender e analisar os processos inerentes a uma determinada área;

(ii) Desenvolver e aplicar técnicas de extração de dados, recorrendo a algoritmos de descoberta e de verificação;

(iii) Desenvolver modelos de processos perceptíveis e coerentes com os registos de eventos;

(iv) Avaliar os modelos de processo de forma a encontrar e analisar possíveis desvios e anomalias.

1.4 Abordagem e Processo de Desenvolvimento

Adotar-se-á uma abordagem baseada em estudo de casos que permitirão o desenvolvimento e sistematização de métodos e ferramentas, que será complementado com uma análise pericial de relevância dos artefactos gerados por peritos nos domínios respetivos.

Os dados gerados pelos eventos não são habitualmente estruturados para serem usados conforme aparecem, ou seja, as atividades devem corresponder às modificações de estado e precisam estar relacionadas a casos. Para que os resultados sejam válidos, a qualidade dos dados a serem usados é de extrema importância. Por exemplo, os registos com informações ruidosas ou insuficientes colocam muitos desafios na descoberta de modelos completos ou altamente adequados. Sendo uma área relativamente nova, os conhecimentos adquiridos com a aplicação de técnicas de mineração de dados para conjuntos de dados reais são necessários para contribuir com sua maturidade.

1.5 Estrutura do Documento

O presente documento está estruturado em sete capítulos:

- **Introdução**: Apresenta o contexto onde o projeto se insere, definição do problema, objetivos propostos e por fim a abordagem proposta e processo de desenvolvimento;
- **Estado da Arte**: Apresenta o estado de arte relativo a abordagens já existentes. Expõe o conceito mineração de processos, algumas técnicas, tipos de formatos, algoritmos e ferramentas. Para terminar o capítulo, é realizada uma avaliação entre as ferramentas existentes e os algoritmos mais relevantes, apresentado as vantagens e desvantagens;
- **Análise de Valor**: Apresenta a análise de valor do caso de estudo, identifica o processo de negócio e inovação, o valor para o cliente, o valor percebido, proposta de valor e por fim, o método multicritério Analytic Hierarchy Process (AHP);
- **Caso de Estudo**: Apresenta a base de dados Medical Information Mart for Intensive Care IIII Information Mart for Intensive Care III (MIMIC III), suas características e estrutura. Por fim especifica os requisitos específicos para este caso de estudo;
- **Construção da Solução**: Apresenta o processo de implementação, descrição do fluxo de trabalho, seleção das ferramentas e algoritmos, a análise dos dados, e por fim, o processo desenvolvido para cada requisito;
- **Avaliação**: Apresenta a avaliação dos resultados obtidos. Analisa os modelos de processo gerados, dimensões de avaliação e possíveis desvios;
- **Conclusões**: Apresenta as conclusões finais desta dissertação, breve resumo do trabalho desenvolvido, limitações e trabalho futuro. Adicionalmente apresenta a apreciação final;

Sendo ainda complementado pelos seguintes anexos:

- **Apêndice A**: Apresenta o método Analytic Hierarchy Process (AHP) em linguagem R utilizado no capítulo 3:
- **Apêndice B**: Apresenta o ficheiro AHP em formato YAML Ain't Markup Language (YAML) utilizado no capítulo 3;
- **Apêndice C**: Apresenta o *script* utilizado para o requisito 1 presente no capítulo 5;
- **Apêndice D**: Apresenta o modelo do requisito 1 onde é aplicado o algoritmo *Heuristics Miner* e realizada a verificação de conformidade. Utilizado no capítulo 6;
- **Apêndice E**: Apresenta o modelo do requisito 2 onde é aplicado o algoritmo *Heuristics Miner* e realizada a verificação de conformidade. Utilizado no capítulo 6;

Capítulo 2

Estado da Arte

Este capítulo apresenta o conceito mineração de processos (secção 2.1) e mineração de processos na área médica (secção 2.2), dando a conhecer alguns conceitos e técnicas em maior detalhe.

De seguida, é descrito os tipos de mineração de processos (secção 2.3), registos de eventos (secção 2.4), modelos de representação de processo (secção 2.5), algoritmos existentes (secção 2.6) e por fim, as ferramentas disponíveis (secção 2.7).

2.1 Mineração de Processos

A evolução da era analógica para a era digital teve um grande impacto na evolução dos sistemas de informação. Os dados são recolhidos sobre qualquer dispositivo, a qualquer momento e em qualquer local. A quantidade de informação gerada é cada vez maior e consequentemente, é necessário um maior poder de armazenamento e tratamento dos dados.

O termo Big Data é usado para descrever quantidades volumosas de dados estruturados, semi-estruturados e desestruturados que podem ser extraídos para obter conhecimento. Big Data é caraterizado por 4Vs: volume extremo de dados, a grande diversidade de tipos de dados e a velocidade a que os dados devem ser processados e veracidade

A figura 1 representa as quatro dimensões que caraterizam os dados em Big Data (*The 4 V's of Big Data - Zarantech* 2016):

- **Volume** Por dia, em média gera-se petabytes de dados. Como existem diversos formatos de dados, o crescimento é exponencial;
- **Velocidade** O movimento dos dados é instantâneo e em tempo-real. Com a evolução da internet, os canais de comunicação mudaram a rapidez com que as notícias se difundem. Os dados estão em constante movimento e velocidade de propagação é cada vez maior;
- **Variedade** Os dados podem ser armazenados em vários formatos. Os dados mais comuns são dados estruturados, como textos, tweets, fotos, vídeos e folhas excel. No entanto, existem dados não estruturados como os e-mails, mensagens de voz, gravações de áudio. O desafio passa por classificar os dados em categorias e uniformizar;
- **Veracidade** Refere-se à confiabilidade dos dados. A veracidade na análise de dados é importante, visto que a fraca qualidade dos dados gera índices de incerteza por parte das organizações e custos inerentes.

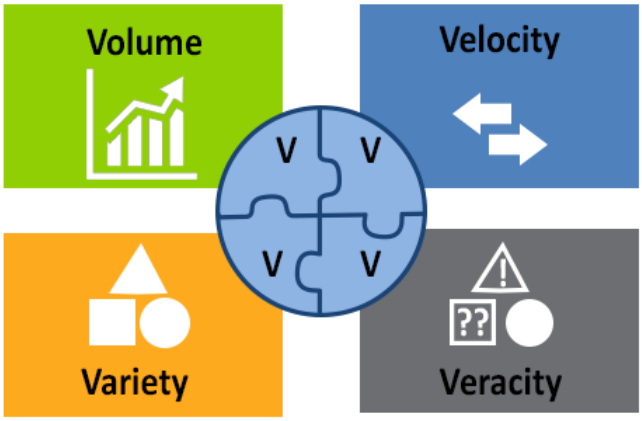


Figura 1: Big Data 4V

Para além de ser essencial guardar todos os acontecimentos, é fundamental, saber analisar e processar. Após a extração dos dados, é necessário limpá-los e estrutura-los em registos de eventos num formato específico. De seguida, é criado o modelo de processo. Os modelos de processo são descobertos a partir dos registos de eventos, são utilizados como modelos de referência e podem ser representados em várias notações.

Para apoiar estes processos, existem sistemas Workflow Management (WFM) e, mais recentemente sistemas Business Process Management (BPM) . Estes dois sistemas visam apoiar os processos inerentes ao fluxo de trabalho, sendo que, os sistemas BPM oferecem um conjunto mais alargado de ferramentas de análise e de suporte de gestão que o sistema WFM. Os sistemas BPM analisam a forma como os processos podem ser melhorados, geridos e monitorizados de uma forma contínua, enquanto que o fluxo de trabalho é focado numa tarefa/objetivo e nas etapas que são necessárias para conseguir alcançá-lo.

Na figura 2 pode-se observar os componentes e relações presentes na área de mineração de processos. Os sistemas de informação suportam, registam e analisam os acontecimentos do quotidiano. Simultaneamente são gerados registos de eventos que, através de algoritmos específicos, produzem modelos de processos.

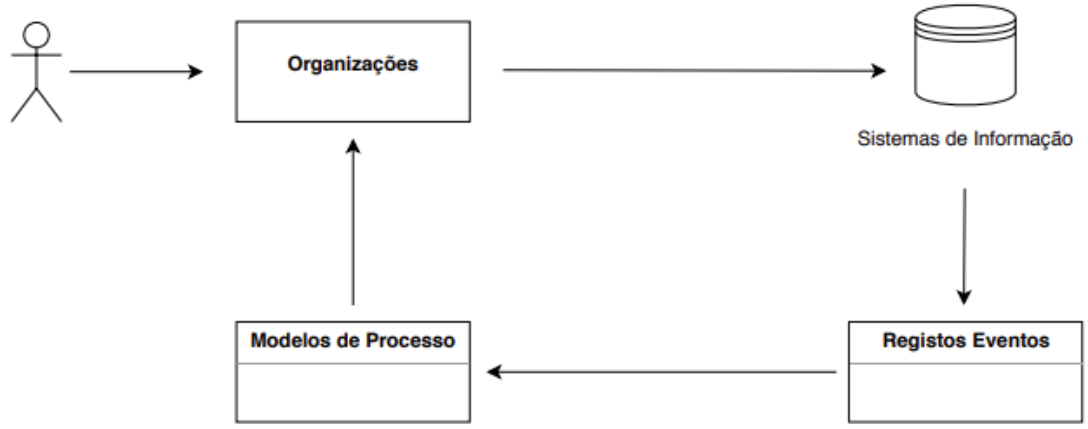


Figura 2: Resumo Mineração de Processos

Um dos objetivos da mineração de processos é recolher informações dos registos de eventos tendo em conta várias perspetivas, de forma a enriquecer o modelo de processo. (W. Van Der Aalst 2016b) afirma que a mineração de processos inclui diferentes perspetivas: **Controlo de Fluxo**, **Organizacional**, **Caso** e **Temporal**:

(i) **Controlo de fluxo**: foca-se na ordenação das atividades. Tem como principal objetivo encontrar uma boa caraterização de todos os caminhos possíveis dentro de um processo;

(ii) **Organizacional**: foca-se na descoberta de interações e colaborações entre participantes (pessoas, funções, departamentos, sistemas, entre outros). Tem como principais objetivos descobrir o papel dos autores na estrutura organizacional e responsabilidades na execução das atividades;

(iii) **Caso**: foca-se nas propriedades de instância de um único processo (caso). Um caso pode ser caraterizado pelo seu caminho no processo, pelos atores que executam ou pelos valores especificados nos dados. Um exemplo pode ser um caso que representa uma ordem de stock, em que será importante saber quem é o fornecedor e/ou o número de produtos encomendados;

(iv) **Temporal**: foca-se na calendarização e na frequência dos eventos. Dado que os eventos possuem um *timestamp* associado, é possível descobrir pontos de estrangulamento, medir níveis de serviço, monitorizar a utilização dos recursos e realizar uma previsão do tempo de processamento para os restantes casos.

As perspetivas aumentam o conhecimento e enriquecem o modelo de processo. A perspetiva organizacional ajuda a descobrir as funções na organização e quais são os recursos que realizam atividades. A perspetiva do caso auxilia a descoberta das caraterísticas de um caso que influenciam uma decisão e a perspetiva temporal facilita na descoberta dos estrangulamentos no processo.

2.2 Mineração de Processos na Área Médica

O envelhecimento da população é cada vez mais uma problemática das sociedades atuais. Com o aumento da longevidade, diminuição da taxa de natalidade e aumento dos custos, os sistemas de saúde enfrentam muitos desafios. Tem que se adaptar de forma a responder às diferentes necessidades de cada faixa etária, com qualidade, eficiência, eficácia e com custos cada vez mais baixos.

Um processo médico é um conjunto de atividades destinadas ao diagnóstico, tratamento e prevenção de qualquer doença, a fim de melhorar a saúde de um paciente. O processo consiste em atividades clínicas executadas por vários tipos de recursos (médicos, especialistas, enfermeiros e técnicos). As melhorias dos processos são fundamentais para alcançar um aumento da qualidade dos serviços prestados e menores custos ao mesmo tempo. Consequentemente pode ter um impacto significativo na qualidade de vida dos pacientes.

As organizações de saúde armazenam grandes quantidades de registos que estão associadas à área clínica (diagnóstico, tratamento e prevenção de doenças dos pacientes) e à área administrativa. Na área da saúde, a área de mineração de processos consegue explorar todos os registos, fornecendo uma análise precisa sobre os processos existentes, e ao mesmo tempo, a descoberta de possíveis desvios e estrangulamentos.

A obtenção de novos processos não é trivial. O ambiente da saúde tem caraterísticas muito peculiares, desde ao seu grau de dinamismo, complexidade e natureza multidisciplinar (Poulymenopoulou, Malamateniou e Vassilacopoulos 2003). Os processos médicos possuem caraterísticas específicas, tais como:

(i) **Dinamismo**: as mudanças no processo podem ocorrer devido a novos procedimentos administrativos, desenvolvimentos tecnológicos, descoberta de novos medicamentos, evolução dos tratamentos, procedimentos e diagnósticos. O aparecimento de novas doenças também contribui para constantes alterações nos processos (Cardoso, Miller e Kochut 2003);

(ii) **Complexidade**: o processo de decisão médica é realizado pela interpretação de dados específicos do paciente bem como o conhecimento do próprio médico. Este processo de decisão é díficil de capturar pois o conhecimento médico varia de acordo as diretizes médicas e pela experiência individual de cada médico. Conclui-se que as decisões médicas, resultados do tratamento e reações do paciente caraterizam-se por serem imprevisíveis e complexas (Rojas et al. 2016);

(iii) **Multidisciplinar**: as organizações de saúde são caraterizadas por um nível crescente de departamentos especializados. Os processos são executados por atividades distribuídas, realizadas pelo esforco colaborativo de vários professionais com diferentes habilidades e conhecimentos (Rojas et al. 2016);

(iv) **Ad-hoc**: os cuidados médicos dependem da colaboração dos professionais. O professional toma decisões de acordo com o seu conhecimento e experiência. Por vezes precisam de desviar-se de determinadas diretizes para lidar com situações excepcionais. O resultado final origina processos com alto grau de variabilidade e a ordem de execução nao é constante (Rojas et al. 2016).

2.3 Tipos Mineração de Processos

2.3.1 Descoberta do Processo

O primeiro tipo da área mineração de processos é a descoberta. Através dos conjuntos dos registos de eventos é produzido um modelo de processos que representa o comportamento do evento.Este tipo não necessita de qualquer outra informação ou recurso para construir o modelo de processos.

De forma a atingir os objetivos desta primeira fase, é necessário o uso de algumas técnicas para a análise dos registos de eventos. Existem algoritmos que auxiliam à extração da informação e podem ser agrupados mediante a sua perspetiva. Exemplos como o *Alpha Miner*, *Alpha+*, Alpha++, *Alpha#*, *Fuzzy miner* e *Heuristic miner* são abordados na secção 2.4.

Um dos aspetos importantes da área mineração de processos é o ênfase em esecer uma forte relação entre o modelo de processo e a realidade representada no registo de evento. Usando a terminologia usada por (Daniel Larel 2003), no contexto *Live Sequence Charts*, usa-se o termo *Play-In* para representar a descoberta do processo. *Play-In* (figura 3) consiste em, a partir de um registo de eventos, é gerado o comportamento e apresentado num modelo com uma notação específica (exemplo: Rede de *Petri*) (W. Van Der Aalst 2016c).

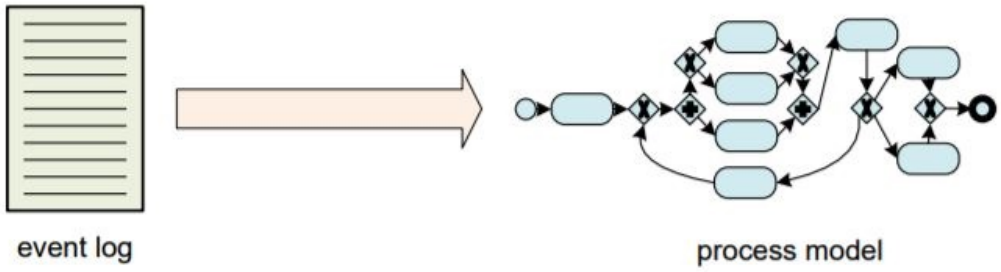


Figura 3: Play-In: Relação entre o registo de evento e o modelo de processo

2.3.2 Verificação de Conformidade

O segundo tipo da área mineração de processos denomina-se por verificação de conformidade. É realizado uma comparação entre os modelos de processos e os registos de eventos do mesmo processo. O objetivo principal é encontrar pontos comuns e/ou discrepâncias entre o comportamento do modelo (modelos de processo) e o comportamento observado (registo de eventos).

Verificação de conformidade pode ser usada para detetar, localizar, explicar desvios presentes nos processos atuais, e, numa segunda instância, medir a gravidade e o impacto desses mesmos desvios. Na realidade, verifica se os eventos presentes no registos estão em conformidade com o modelo ou vice-versa.

Usando a terminologia usada por (Daniel Larel 2003), no contexto *Live Sequence Charts*, usa-se o termo *Play-Out* para representar a verificação do desempenho. *Play-Out* (figura 4) consiste em, a partir de um modelo de processo gerar o comportamento (W. Van Der Aalst 2016c).

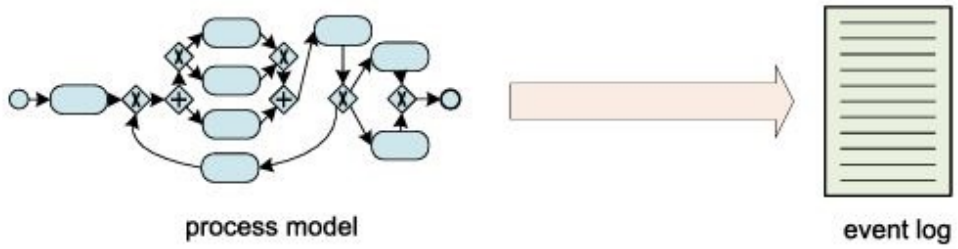


Figura 4: Play-Out: Relação entre o modelo de processo e o registo de evento

2.3.3 Aprimoramento

O terceiro tipo da área mineração de processos denomina-se por aprimoramento. O conceito é estender ou melhorar o modelo de processo existente usando informação acerca do processo atual gravado no registo de eventos. Este tipo visa a mudança e/ou estende o modelo

a-priori. Ao passo que a verificação de conformidade determina o posicionamento entre o modelo de processo e a realidade, o aprimoramento tem como alvo complementar ou aperfeiçoar o modelo fornecido à priori.

Com este tipo é possível realizar uma correção a um modelo, mas também adicionar uma nova perspetiva enriquecendo o registo de eventos com dados. O 'reparo' é um tipo de aprimoramento que altera o modelo para refletir a realidade, com maior exatidão.

Usando a terminologia usada por (Daniel Larel 2003), no contexto *Live Sequence Charts*, usa-se o termo *Replay* para representar o aprimoramento. *Replay* (figura 5) consiste em usar um registo de eventos e um modelo de processo como input. O registo de eventos é repetido por cima do modelo de processo de forma a que o modelo seja melhorado e/ou estendido (W. Van Der Aalst 2016c).

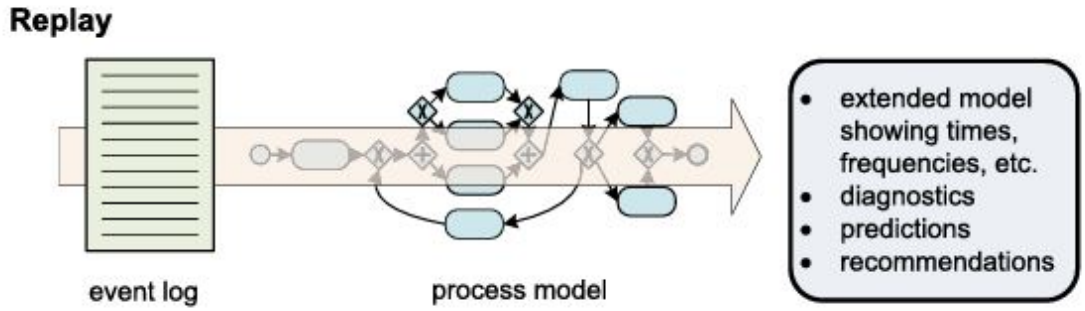


Figura 5: Replay: Relação entre e o registo de evento e o modelo de processo repetidamente

2.4 Registo de Eventos

É possível gravar sequencialmente eventos de tal forma que cada evento representa uma atividade e está relacionado com um caso particular. Através dos registos de eventos disponibilizados pelos sistemas de informação, consegue-se extrair conhecimento. As atividades realizadas pelas pessoas, funcionamento de máquinas e de software são identificadas como registos de eventos.

A escolha das propriedades dos eventos é uma etapa importante. As propriedades irão influenciar a avaliação final, de acordo com os objetivos traçados. Caso uma organização defina como objetivo avaliar o desempenho dos seus recursos num determinado fuso horário, é importante que para cada entrada no registo de eventos exista uma referência temporal.

A tabela 1 representa um exemplo simples de um fragmento de um registo de eventos relacionado com o tratamento de pedidos de compensação (W. Van Der Aalst 2016d). Na tabela 1, os eventos referem-se a atividades como o pedido de registo ou verificação do bilhete. A atividade "pedido de registo" foi realizada pelo *Pete* no dia 30/12/2010 às 11:02, teve um custo de 50 e pertence ao primeiro caso (id=1).

Tabela 1: Registo Eventos

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request	Pete	50	...
	35654522	30-12-2010:15.06	examine casually	Mike	400	...
	35654524	30-12-2010:16.34	check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	decide	Sara	200	...
	35654526	06-01-2011 12.18	reinitiate request	Sata	200	...
	35654527	06-01-2011 13.06	examine thoroughly	Sean	400	...
	35654530	08-01-2011 11.43	check ticket	Pete	100	...
	35654531	09-01-2011 09.55	decide	Sara	200	...
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	...
4	35654641	06-01-2011:15.02	register request	Pete	50	...
	35654643	07-01-2011:12.06	check ticket	Mike	100	...
	35654644	08-01-2011::14.43	examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02:	decide	Sara	200	...
	35654647	12-01-2011:15.44:	reject request	Ellen	200	...

Numa outra perspetiva, a figura 6 representa a estrutura de um registo de eventos. Um processo pode conter vários casos e um caso pode conter vários eventos (W. Van Der Aalst 2016e). Cada evento possui atributos tais como o nome da atividade, data, custo e recurso. Os eventos de cada caso são ordenados.

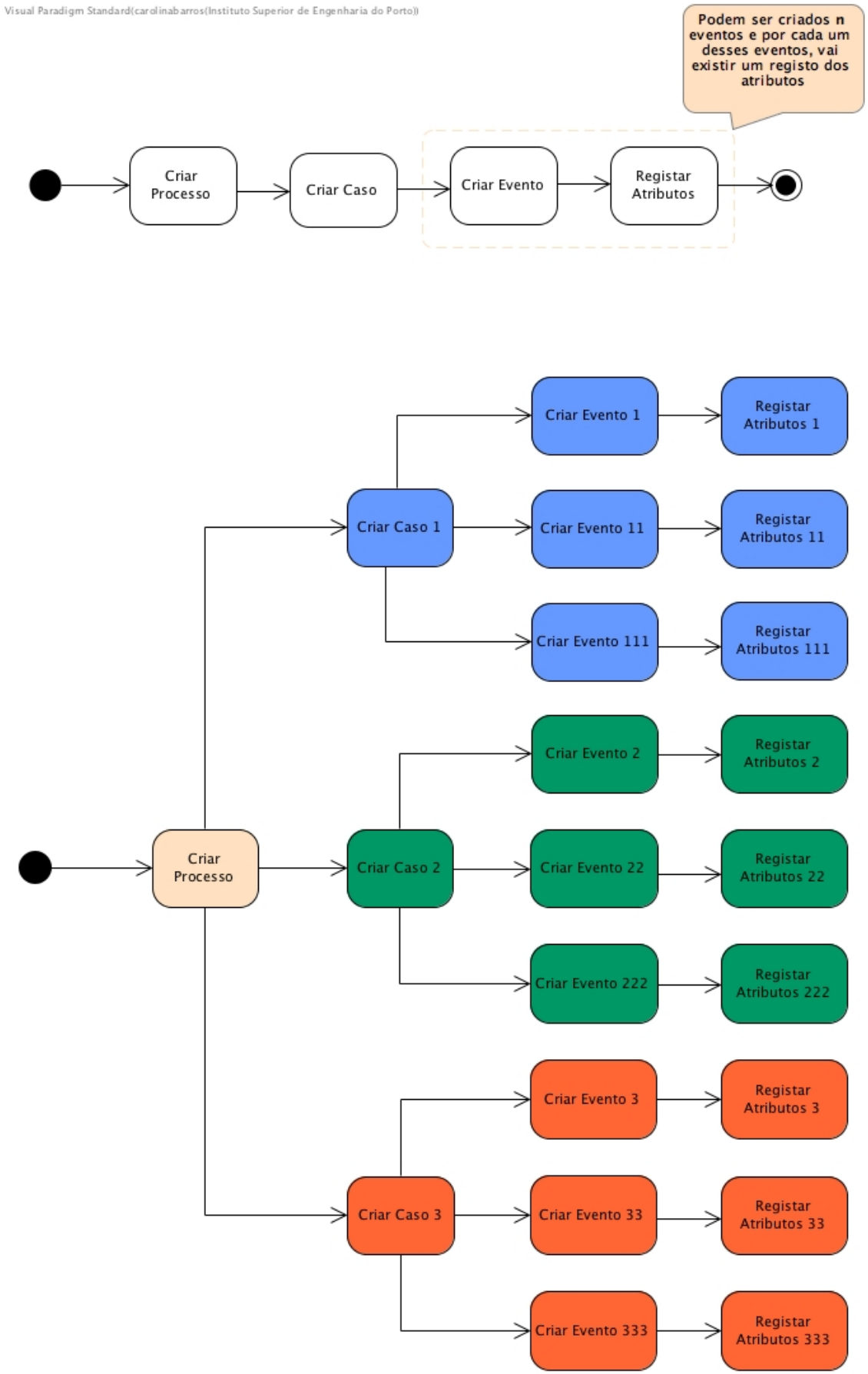


Figura 6: Estrutura Registo Eventos

Os registos de eventos podem ser utilizados de forma a conduzir os três tipos que existem: descoberta do processo (*process discovery*), verificação de conformidade (*performance checking*), e aprimoramento (*enhancement*). Estes tipos irão ser abordadas com maior profundidade na subsecção 2.3.1, subsecção 2.3.2 e subsecção2.3.3.

A figura 7 apresenta as *inputs* que são válidos (registo de eventos e modelos), e de acordo com o tipo escolhido, o output é gerado (modelo, diagnóstico e modelo melhorado) (Rudnitckaia 2015).

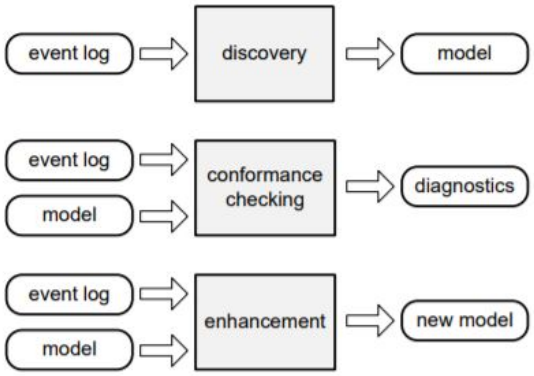


Figura 7: Tipos Mineração de Processos

2.4.1 Formatos

Mining eXtensible Markup Language (MXML) foi desenvolvido pela *framework ProM* e criado em 2005 por *Van Dongen* e *Van der Aalst*. Emergiu em 2013 (W. Van Der Aalst 2016f) e é suportado por várias ferramentas de análise de mineração de processos, como é o caso da ferramenta *ProM*.

Este formato tem como foco standarizar o modo como os registos de eventos são guardados no sistema. É extensível e baseia-se em formatos genéricos XML para armazenar registos de eventos. Após a conversão dos registos, estes são guardados e alterados no sentido de serem acessíveis para análise.

Um documento em MXML possui uma estrutura hierárquica em árvore definida. Como nodo raiz tem um registo de eventos, denominado por *WorkflowLog*. Cada registo pode conter um elemento *source*, que é usado para descrever o sistema que forneceu o registo de eventos, e poderá também conter um número arbitrário de processos (elementos filho).

Em novembro de 2016, o formato **Extensible Event Stream (XES)** foi adotado pela *IEEE Task Force on Process Mining* (*IEEE 1849-2016 XES Standard* 2016) e tornou-se o formato padrão para alcançar a invariabilidade entre os registos de eventos da área mineração de processos.

O formato *XES* foi o sucessor do formato MXML e possui caraterísticas bem definidas:

- Simplicidade, dado que usa uma representação simples para representar informações;
- Flexibilidade, dado que é possível extrair registos de eventos de diferentes domínios;
- Extensibilidade, dado que é possível adicionar uma norma no futuro;

- Expressividade, dado que o processo de serialização dos registos de eventos possuem pouca perda de informação.

A figura 8 apresenta o metamodelo do *XES*. Um documento em formato XES não descreve um conjunto fixo de atributos para cada elemento (W. Van Der Aalst 2016e). Um registo de eventos pode conter vários casos, cada caso pode conter vários eventos e todos têm atributos. Para fornecer semântica aos atributos, os registos de eventos realizam referências a extensões. Uma extensão fornece semântica a atributos específicos.

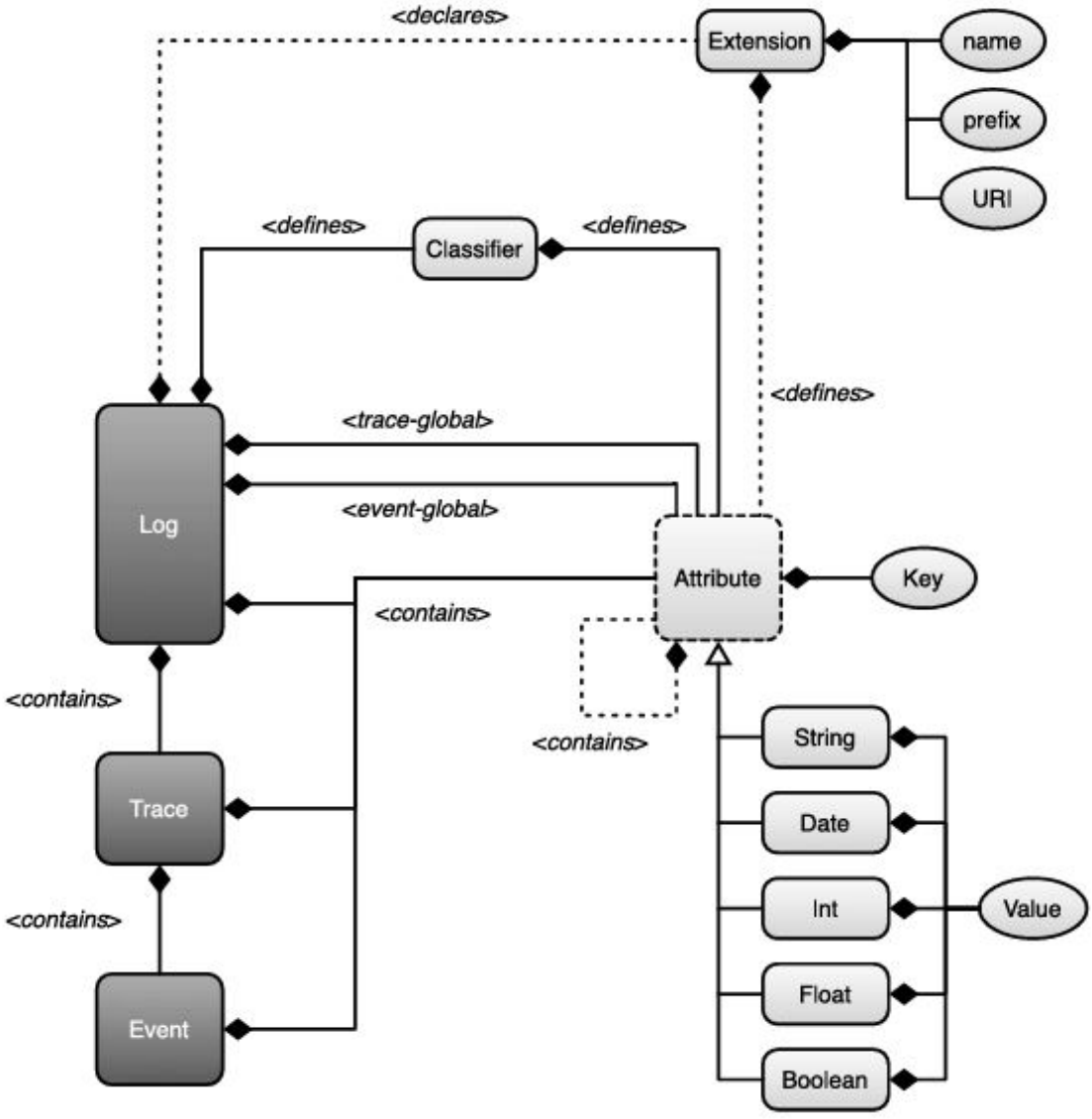


Figura 8: Metamodelo Extensible Event Stream (XES)

2.4.2 Problemas Comuns

Um dos grandes problemas relativos à análise de registos de eventos devem-se à qualidade dos mesmos. Os problemas mais recorrentes surgem devido a:

(i) **Dados em falta** No registo de eventos pode faltar diferentes tipos de informação. Pode faltar um evento, um atributo ou valor. Os dados em falta refletem um problema no processo de registo (Jagadeesh et al. 2012);

(ii) **Dados incorretos/incompletos**: No registo de eventos podem estar presentes todos os dados, embora incorretos ou incompletos. Os dados podem ser introduzidos de forma incorreta (Jagadeesh et al. 2012);

(iii) **Dados genéricos**: No registo de eventos podem estar presentes dados demasiado genéricos, originando uma falta de precisão. Os dados genéricos afetam os resultados e podem impedir a realização de certos técnicas de análise. Um exemplo concreto trata-se de verificar quais as horas com maior afluência na recepção de uma empresa. Se os dados temporais conterem apenas o dia da realização dos eventos, torna-se difícil obter um resultado preciso. É necessário os dados conterem outras referências temporais mais específicas (hora de entrada e hora de saída) (Jagadeesh et al. 2012);

(iv) **Dados irrelevantes**: No registo de eventos podem estar presentes dados que contém conteúdo irrelevante para a análise pretendida. Em algumas ferramentas, tais como o ProM e Disco, existem filtros de limpeza, no entanto, esses filtros não são suficientemente eficazes em alguns casos (Jagadeesh et al. 2012).

2.5 Modelos de Representação de Processo

Os modelos de processos têm como objetivo decidir quais são as atividades que irão ser executadas e respetiva ordem de execução. As atividades podem ser executadas de forma sequencial, paralela e a execução da atividade pode ser repetida.

Os modelos de processo podem ser representados por diferentes notações, sendo alguns deles: sistema transacional, rede de *Petri*, Business Process Model and Notation (BPMN), Event-Driven Process Chain (EPC) ou diagrama de atividade de Unified Modeling Language (UML).

Esta subsecção apresenta três notações de modelos de representação de processos mais comuns e suas caraterísticas.

2.5.1 Sistema Transacional

Sistema transacional é considerado a notação mais básica e simples. Um sistema de transição é composto por estados e transições, em que os estados são representados por círculos negros e possuem uma identificação exclusiva. As transições são representadas por arcos, conectando dois estados e são identificadas com nomes de atividades/eventos.

A figura 9 representa um modelo de processo com notação sistema transacional, em que possui um único estado inicial(s1) e um final(s7) (W. Van Der Aalst 2016g). Contém sete estados e um total de onze transições. Exemplo retirado de (W. Van Der Aalst 2016g).

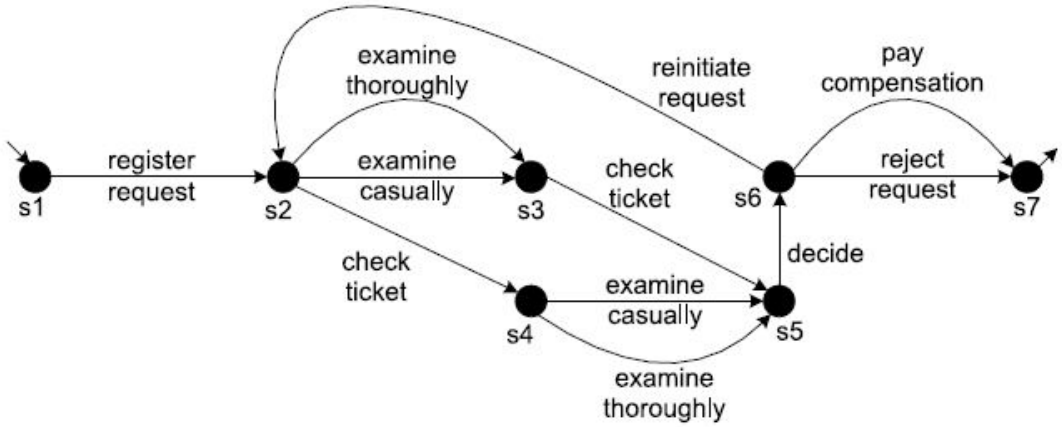


Figura 9: Modelo de processo com notação sistema transacional

Dada a definição de sistema transacional (W. Van Der Aalst 2016g), pode-se formalizar:

S = {s1, s2, s3, s4, s5, s6, s7}, S contém o conjunto de todos os estados;

S^{start} = {s1}, S contém o estado inicial;

S^{end} = {s7}, S contém o estado final;

A = {register request, examine thoroughly, examine casually, check ticket, decide, reinitiate request, reject request, pay compensation}, A contém o conjunto de todas as atividades;

T = {(s1, register request, s2), (s2, examine casually, s3), (s2, examine thoroughly, s3), (s2, check ticket, s4), (s3, check ticket, s5), (s4, examine casually, s5), (s4, examine thoroughly, s5), (s5, decide, s6), (s6, reinitiate request, s2), (s6, pay compensation, s7), (s6, reject request, s7)}, T representa a união do conjunto S e A. Contém todos as transições (estado origem, atividade, estado destino).

A notação sistemas transacionais é simples mas não é suficientemente eficiente quando o modelo de processo exige representar casos de concorrência paralela.

2.5.2 Rede de Petri

Rede de *Petri* é a notação mais conhecida e consistente da área, pois permite especificar modelos de processos de casos bem definidos e pouco ambíguos. Consegue suportar concorrência. Esta notação apresenta um grafo composto por nós de posição, nós de transição e arcos de ligação. Os nós de posição são representados por círculos, as transições por quadrados e os arcos por setas direcionais. O início da atividade é marcado quando é espoletado um *token* num nó de posição.

A figura 10 representa um modelo de processo com notação rede *Petri*, em que possui sete nós de posição, oito nós de transição e dezanove arcos de ligação. Exemplo retirado de (W. Van Der Aalst 2016h).

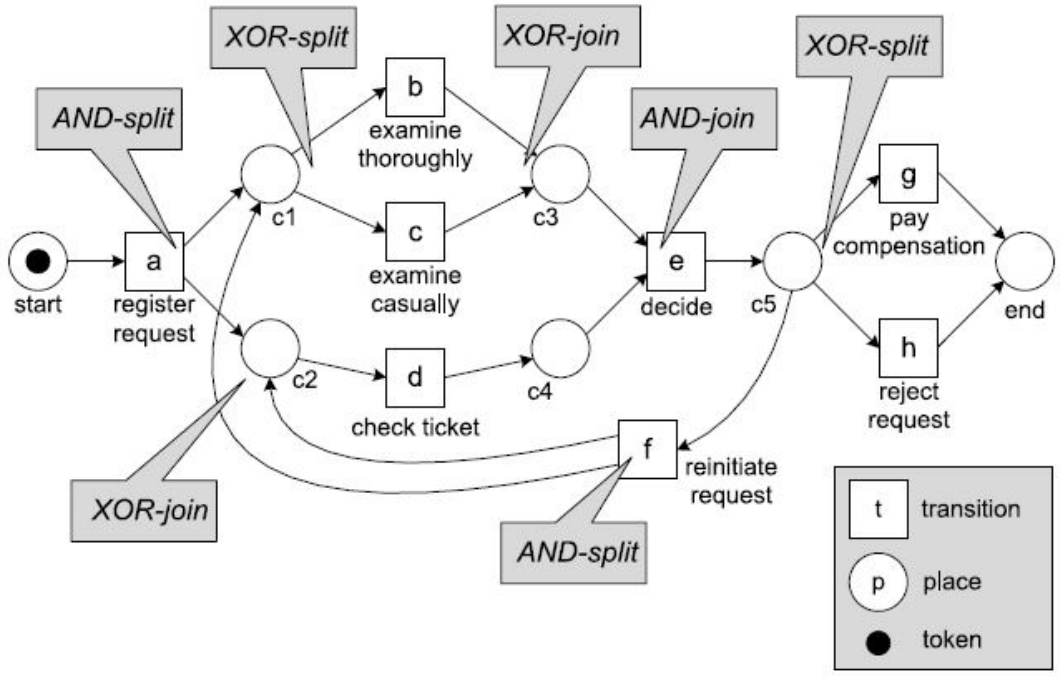


Figura 10: Modelo de processo com notação rede *Petri*

A partir da definição de rede *Petri* (W. Van Der Aalst 2016h), pode-se formalizar:

P = {start, c1, c2, c3,c4, c5, end}, P contém todos os *places*.

T = {a, b, c, d, e, f, g,h}, T contém as transições.

F = {(start, a), (a, c1), (a, c2), (c1, b), (c1, c), (c2, d), (b, c3), (c, c3), (d, c4), (c3, e), (c4, e), (e, c5), (c5,f), (f, c1), (f, c2), (c5, g), (c5, h), (g, end), (h, end)}, F contém todos os arcos, denominada relação de *flow*.

2.5.3 Business Process Modeling Notation

Business Process Modeling Notation tornou-se uma das notações mais utilizadas na área de modelação de processos. Fornece uma notação gráfica para a especificação de processo de negócio baseada em técnicas de fluxogramas. Esta notação é intuitiva, faz uso de um conjunto de ícones padrão para representar os diversos fluxos, auxiliando o entendimento de utilizadores de diferentes áreas e/ou competências (Group 2006).

A figura 11 representa um modelo BPMN que retrata uma parte do processo de negócio relativo a um pedido de encomenda de uma pizza (Incubator 2009). O processo é composto por uma *pool* com quatro *lanes* (*call center*, cozinha, gestão de entrega, condutor). O pedido é iniciado através do *call center*, que depois de confirmado, dá ordem na cozinha para preparar a pizza. Depois da fase de preparação, cozinha e empacotamento, o processo é reencaminhado para as atividades do condutor: entrega da pizza e receção do pagamento.

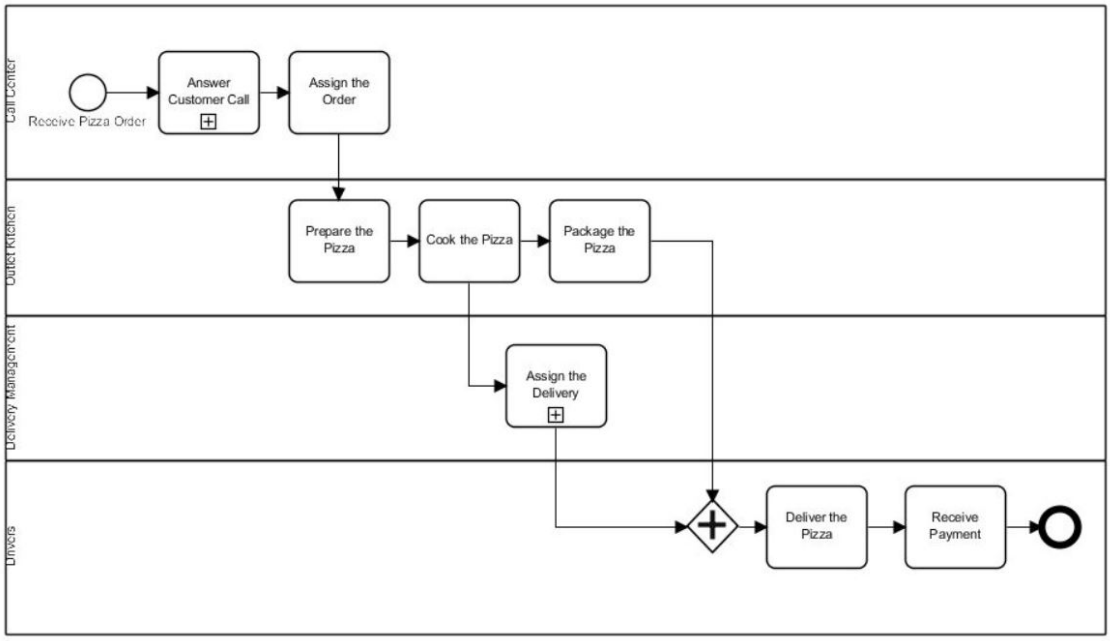


Figura 11: Modelo de processo com notação BPMN

2.5.4 Problemas Comuns

Vários são os problemas comuns relativos às características dos processos, nomeadamente, a granularidade de eventos e a heterogeneidade.

(i) **Granularidade de eventos**: Alguns processos são definidos por um elevado número de atividades muito granulares; As técnicas da área mineração de processos têm dificuldade em lidar com registos de eventos granulares. Os modelos gerados são do tipo *spaghetii*, e tornam-se imperceptíveis ao olho humano; De forma a contornar o problema, é preferível usar um nível maior de abstração, evitando os eventos muitos granulares presentes nos registos de eventos, facilitanto a leitura do modelo e análise (Jagadeesh et al. 2012);

(ii) **Heterogeneidade**: Existem dois cenários de possível heterogeneidade. No primeiro caso verifica-se que os registos de eventos podem possuir diversos cenários heterogéneos com comportamentos distintos e desestruturados. O segundo caso diz respeito às constantes mudanças temporais dos processos, ou seja, com o passar do tempo é espectável que os processos operacionais sofram melhorias e que se adaptem à mudança (exemplo: nova legislação, variações externas e efeitos sazonais). Por exemplo, os sistemas da área da saude suportam procedimentos médicos, em que estes possuem centenas de variações que criam heterogeneidade nos registos de eventos. Quando estes registos são analisados por algoritmos, geralmente os modelos de processo produzidos são incompreensíveis (tipo *spaghetti*). De forma a contornar o problema, é preferível que exista uma maior objetividade nos registos de eventos (Jagadeesh et al. 2012).

2.6 Algoritmos

Na área de mineração de processos os algoritmos podem ser agrupados mediante a perspetiva. Na perspetiva de fluxo de controlo, destacam-se os algoritmos *Alpha Miner* (W. Van Der Aalst 2016i), *Heuristic Miner* (W. Van Der Aalst 2016j), *Inductive Miner* (W. Van Der Aalst 2016k), *Fuzzy Miner* (Günther et al. 2007) e *Genetic Miner* (Aalst, Medeiros e Weijters 2005).

Na perspetiva organizacional tem-se como referência o *Social Network Miner* e o *Organizational Model Miner* (Song e Wil M P Van Der Aalst s.d.). Na perspetiva organizacional o algoritmo *Social Network Miner* analisa o registo de eventos e determina a rede social dos participantes no processo, permitindo a identificação de papéis e interações dentro da organização. Já o algoritmo *Organizational Model Miner* extrai a informação do registo de eventos que contém informação do criador de cada evento e apresenta um gráfico que associa atividades a utilizadores.

Esta subsecção apresenta três algoritmos na perspetiva de fluxo de controlo, abordando as suas principais caraterísticas, vantagens e algumas limitações. Como ponto final apresenta uma avaliação comparativa entre as ferramentas.

2.6.1 Algoritmo Alpha Miner

O algoritmo *Alpha Miner* apresentado na lista de algoritmos **1** foi um dos primeiros capazes de suportar concorrência nos registos de eventos mas por outro lado, não tem em consideração a frequência das relações, verificando apenas se estas existem ou não (W. Van Der Aalst 2016l). Não é aplicável a situações reais pois não consegue suportar situações complexas, incompletas ou ruidosas. Dado um registo de eventos, o algoritmo *Alpha Miner* gera um modelo de processo em notação rede de *Petri* (W. Van Der Aalst 2016i).

As primeiras 3 linhas do algoritmo apresentam os três conjuntos de tarefas a determinar. 'TW' representa o conjunto de tarefas que ocorrem pelo menos uma vez, 'TI' representa o conjunto de tarefas que ocorrem numa fase inicial e 'TO' representa o conjunto de tarefas que ocorrem numa fase final. O algoritmo procura padrões de execução. Caso uma atividade seja imediatamente seguida por outra, assume-se que existe uma dependência causal entre elas (linhas: 4 ('XW'), 5 ('YW'), 6 ('PW'), 7 ('FW').

Algoritmo 1 Alpha Miner

1: $TW = \{t \in T | \exists \sigma \in W t \in \sigma\}$,
2: $TI = \{t \in T | \exists \sigma \in W t = first(\sigma)\}$,
3: $TO = \{t \in T | \exists \sigma \in W t = last(\sigma)\}$,
4: $XW = \{(A, B) | A \subseteq TW \wedge A \neq \varnothing \wedge B \subseteq TW \wedge B \neq \varnothing \wedge \forall a \in A \forall b \in Ba \rightarrow Wb \wedge \forall a1, a2 \in Aa1 \# Wa2 \wedge \forall b1, b2 \in Bb1 \# Wb2\}$,
5: $YW = \{(A, B) \in X | \forall (A', B') \in XA \subseteq A' \wedge B \subseteq B' \Rightarrow (A, B) = (A', B')\}$,
6: $PW = \{p(A, B) | (A, B) \in YW\} \cup \{iW, oW\}$,
7: $FW = \{(a, p(A, B)) | (A, B) \in YW \wedge a \in A\} \cup \{(p(A, B), b) | (A, B) \in YW \wedge b \in B\} \cup \{(iW, t) | t \in TI\} \cup \{(t, oW) | t \in TO\}$,
8: $\alpha(W) = (PW, TW, FW)$.

O algoritmo está centralizado no controlo do fluxo (ordem das atividades), ignorando os recursos, comportamento, *timestamps* e outro qualquer tipo de dados.

Numa primeira fase, os registos são lidos de forma a ordenar relações entre as atividades. A partir das relações é construída a *matrix footprint* que irá gerar um modelo de processo em notação rede de *Petri*. Existem quatro relações de ordenação que o algoritmo consegue detetar:

(i) Sucessão direta: $x > y$ (Se x é seguido diretamente por y);

(ii) Sequência: $x \rightarrow y$ (Se $x > y$ e $\neg y > a$);

(iii) Paralelo: $x||y$ (Se $x > y$ e $y > a$);

(iv) Não tem relação direta: $x\#y$.

A tabela 2 apresenta um registo de eventos que será tomado como exemplo para demonstração do algoritmo *Alpha Miner*.O registo contém cinco diferentes atividades (reservar voo, comprar seguro, reservar hotel, realizar pagamento, confirmar pagamento). O primeiro passo é converter o registo de eventos em conjuntos de traces (sequência de atividades) tendo em atenção a sua ordem.

Tabela 2: Registo de Eventos Processo de Reserva

Id Caso	**Atividade**	**Timestamp**	**Preço**	**IP Cliente**
1	reservar voo	2014-12-24 08:30:00:232	145	188:44:22:45
1	comprar seguro	2014-12-24 08:31:05:171	23	188:44:22:45
2	reservar voo	2014-12-24 08:31:08:543	94	93.180.0.62
1	reservar hotel	2014-12-24 08:32:08:703	295	188.44.42.45
3	reservar voo	2014-12-24 08:32:11:534	192	188.44.50.103
1	realizar pagamento	2014-12-24 08:34:17:456	463	188:44:22:45
1	confirmar pagamento	2014-12-24 08:35:17:537	463	188:44:22:45

O conjunto L representa o número de ocorrências para cada sequência de atividades. A sequência de atividades que possui maior número de ocrrências é: reservar voo, comprar seguro, reservar hotel, realizar pagamento e confirmar pagamento. O número no fim de cada sequência representa o factor exponencial, ou seja, a primeira sequência do conjunto L ocorreu 25 vezes.

Conjunto L =
[<reservar voo, comprar seguro,reservar hotel,realizar pagamento,confirmar pagamento>^{5},
<reservar voo,reservar hotel,comprar seguro,realizar pagamento,confirmar pagamento>^{4},
<reservar hotel,reservar voo,comprar seguro,realizar pagamento,confirmar pagamento>^{4},
<reservar hotel,comprar seguro,reservar voo,realizar pagamento,confirmar pagamento>^{3},
<comprar seguro,reservar hotel,reservar voo,realizar pagamento,confirmar pagamento>^{1},
<comprar seguro,reservar voo,reservar hotel,realizar pagamento,confirmar pagamento>^{1}]

O segundo passo passa pela aplicação do algoritmo *Alpha Miner* ao conjunto L, gerando um modelo de processo. A figura 12 apresenta o modelo de processo gerado em notação Rede *Petri*.

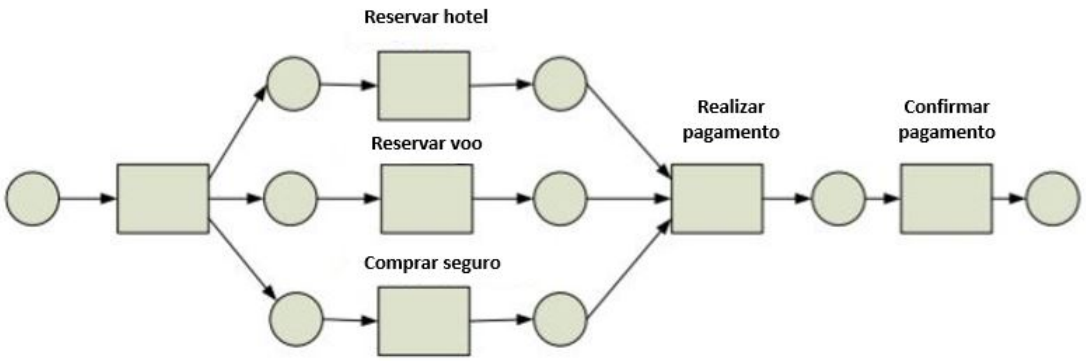


Figura 12: Modelo de processo gerado pelo Alpha Miner com notação Rede Petri

O algoritmo *Alpha Miner* é considerado um bom algoritmo a nível teórico para iniciação à área mineração de dados. Apesar de conseguir encontrar ciclos e situações de concorrência não deve ser tomado como referência, pois apresenta muitas fragilidades, nomeadamente (Banerjee e Gupta 2015):

- Não é tolerante ao ruído, ou seja, não tolera dados incorretos;
- Não aplicável nos registos de eventos reais;
- Não tolera tarefas duplicadas.

2.6.2 Algoritmo Heuristic Miner

Heuristic Miner, apresentado na lista de algoritmos 2, é outro algoritmo na área de mineração de dados, seguido do *Alpha Miner*. Foi desenvolvido por *Ton Weijters* (*ProM Tips — Which Mining Algorithm Should You Use? — Flux Capacitor* 2018) que utilizou uma abordagem heurística de forma a conseguir resolver falhas detetadas no algoritmo *Alpha Miner*.

Heuristic Miner apresenta o comportamento mais frequente do processo, considera a ordem dos registos de eventos e consegue filtrar comportamentos com ruído ou comportamento pouco frequente. Os caminhos que não são frequentes não devem ser incorporados no modelo de processo. Este algoritmo é classificado como robusto face às suas caraterísticas.

Identifica-se as seguintes melhorias em comparação com o algoritmo *Alpha Miner*:

- Tolera o ruído;
- Considera a frequência;
- Deteta ciclos pequenos;
- Permite ignorar atividades.

Algoritmo 2 Heuristic Miner

1: $a > w$ b iff there is a trace $\sigma = t1t2t3 \ldots t_n$ and $i \in \{1, \ldots, n-1\}$ such that $\sigma \in W$ and $t_i = a$ and $t_{i+1} = b$,
2: $a \rightarrow w$ b iff $a > w$ b and $b \not> w$ a,
3: $a \# w$ b iff $a \not> w$ b and $b \not> w$ a, and
4: $a || w$ b iff $a > w$ b and $b > w$ a,
5: $a >> w$ b iff there is a trace $\sigma = t1t2t3 \ldots t_n$ and $i \in \{1, \ldots, n-2\}$ such that $\sigma \in W$ and $t_i = a$ and $t_{i+1} = b$ and $t_i + 2 = a$,
6: $a >>> w$ b iff there is a trace $\sigma = t1t2t3 \ldots t_n$ and $i < j$ and $i, j \in \{1, \ldots, n\}$ such that $\sigma \in W$ and $t_i = a$ and $t_j = b$.

Numa primeira fase, os registos são lidos de acordo com a sua frequência e a matriz é construída. Dados os dados de frequência apresentados na tabela 3 avançamos para a construção da matriz representada na tabela4. A partir da matriz, conclui-se que 'a' foi seguido diretamente por 'b' cinquenta e seis vezes (1º registo 6x + 2º registo 38x + 3º registo 12x).

Tabela 3: Dados de frequência

Repetição	**Dados**
6x	<a,b,c,d,e,g>
38x	<a,b,c,d,f,g>
2x	<a,c,d,b,f,g>
12x	<a,b,d,c,e,g>
4x	<a,d,c,b,f,g>

Tabela 4: Matriz de frequência - *Heuristic Miner*

|>|	**a**	**b**	**c**	**d**	**e**	**f**	**g**
a	-	56	2	4	-	-	-
b	-	-	44	12	-	6	-
c	-	4	-	46	12	-	-
d	-	2	4	-	18	38	-
e	-	-	-	-	-	-	18
f	-	-	-	-	-	-	44
g	-	-	-	-	-	-	-

Numa segunda fase, é necessário construir a matriz de dependência apresentada na tabela (5). Esta matriz utiliza a matriz de frequência em conjunto com a seguinte formula. Os valores obtidos variam entre -1 e 1, valores negativos significam relações mais fracas (frequência baixa) e vice-versa.

$$| => | = \frac{|a>b| - |b>a|}{|a>b| + |b>a| + 1}$$

Tabela 5: Matriz de dependência - *Heuristic Miner*

|=>|	a	b	c	d	e	f	g
a	-	.98	.67	.80	-	-	-
b	-9.8	-	.82	.67	-	.86	-
c	-.67	-.82	-	.90	.92	-	-
d	-.80	-.67	-.90	-	.95	.97	-
e	-	-	-.92	-.95	-	-	.95
f	-	-.86	-	-.97	-	-	.98
g	-	-	-	-	-.95	-.98	-

Numa terceira fase, após construir as duas matrizes (frequência e dependência), aplica-se o algoritmo *Heuristic Miner* e obtém-se o modelo de processo em notação rede de *Petri*, representado na figura 13. O modelo de processo gerado tem lacunas, contém um nível de simplicidade médio, não possui ruído, nível de precisão aceitável e baixa generalização. Este algoritmo é mais completo que o *Alpha Miner* porém, apresenta falhas na descoberta de modelos de processo com níveis altos de qualidade.

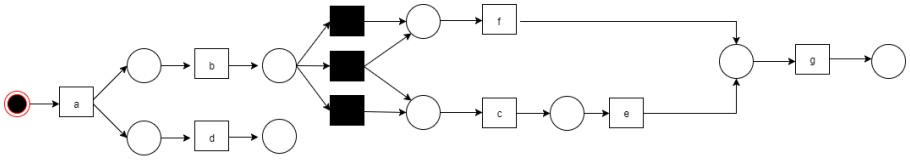


Figura 13: Modelo de processo gerado pelo Heuristic Miner com notação Rede Petri

2.6.3 Algoritmo Inductive Miner

O algoritmo *Inductive Miner* reune um conjunto de melhorias face aos algoritmos *Alpha Miner* e *Heuristic Miner*. O algoritmo usa uma abordagem *Divide&Conquer* que através do registo de eventos, divide as atividades que o constituem, seleciona partes importantes do processo e divide o registo de eventos repetidamente até obter um caso suficientemente robusto que ilustre o workflow do processo. Internamente este algoritmo não utiliza a notação rede de Petri, mas sim, árvores de processo.

A partir dos dados apresentados abaixo, é aplicado o algoritmo *Inductive Miner* e é construído a árvore de processo representada na figura 14.

Dividindo o conjunto de dados em duas partes distintas, pode-se constatar que na esquerda começa sempre por 'a' e na direita por 'g'. Assim sendo, no primeiro nível da árvore de processo, as extremidades serão constituidas por 'a' e 'g'. Observando a quinta coluna no conjunto de dados, observa-se que pode acontecer 'e' ou 'f', logo no segundo nível da árvore coloca-se uma variável 'X' que no terceiro nível da árvore irá conter 'e' e 'f'. Para finalizar, no conjunto de dados, a segunda, terceira e quarta coluna pode acontecer 'b', 'c' ou 'd'. Assim sendo no segundo nível da árvore coloca-se o símbolo '$\wedge$' e no terceiro nível coloca-se 'b', 'c' e 'd'.

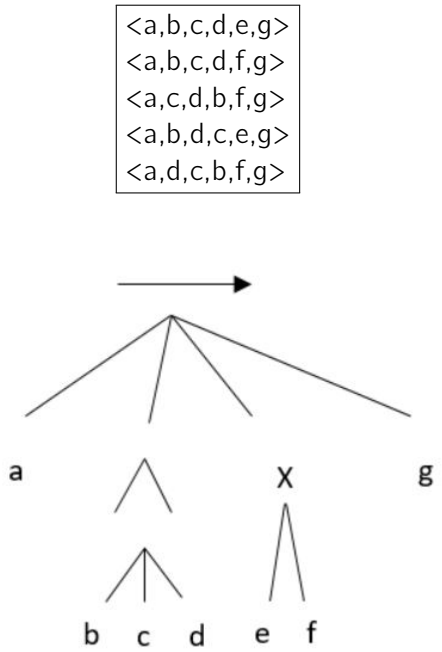


Figura 14: Árvore de Processo

Numa segunda fase, após construir a árvore de processo, é possivel construir o modelo de processo em notação rede de *Petri*, representado na figura 15.

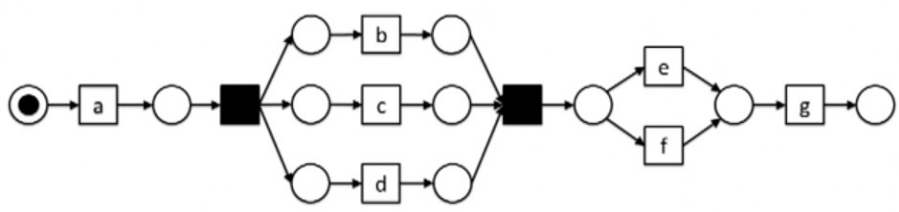


Figura 15: Modelo de processo com notação Rede Petri

2.6.4 Comparação Algoritmos

A tabela 6 apresenta a comparação realizada entre cinco algoritmos da perspetiva de controlo de fluxo (*Alpha Miner*, *Heuristic Miner*, *Inductive Miner*, *Fuzzy Miner* e *Genetic Miner*). Para cada algoritmo é descrito qual é a perspetiva, quando é que deve ser usado, o seu comportamento, *output* final e problemas identificados.

Tabela 6: Comparação Algoritmos

	Características				
Algoritmo	Perspetiva	Quando Usado	Comportamento	Output	Problemas
Alpha Miner	Controlo	Registo de eventos simples; Efeitos teóricos	Baseia-se em relações binárias presentes no registo de eventos	Rede de *petri*	Não tolera ruído; Não tolera tarefas duplicadas; Não aplicável em registo de eventos reais
Heuristic Miner	Controlo	Registo de eventos que não possuem diferentes tipos de eventos	Consideram a frequência dos eventos	Heuristic net	Não representa os detalhes e exceções presentes no registo de eventos
Indutive Miner	Controlo	Registo de eventos complexo (vários tipos de eventos); Modelo de processo preciso e simples	Abordagem *Divide & Conquer:* divide repetidamente o registo de eventos até obter um caso robusto	Árvore de processo	-
Fuzzy Miner	Controlo	Regista eventos complexos e desestruturados	Utiliza a representação do grafo de dependência	Modelo *fuzzy*	Não converte para a notação de *petri*
Generic Miner	Controlo	Gera modelos de processos aleatórios de forma a encontrar uma solução satisfatória	Simula a evolução natural	Rede de *petri*	Requer muitos recursos computacionais

Em suma, o algoritmo *Inductive Miner* é considerado o mais completo face aos algoritmos *Alpha Miner* e *Heuristic Miner*. Os registos de eventos podem ser mais complexos, possuir

tarefas de vários tipos e conter duplicações. O ouput gerado é uma árvore de processo que facilmente é transformada num modelo de processo em notação rede de *petri*.

Já o algoritmo *Fuzzy Miner* suporta registos de eventos desestruturados e com comportamentos irregulares. Este algoritmo utiliza métricas de significância/correlação para simplificar de forma interativa o modelo de processo. O *output* gerado (modelo *fuzzy*) não pode ser convertido noutras notações de modelos de processo.

O algoritmo *Genetic Miner* é capaz de detectar padrões não locais no registo de eventos. O principal objetivo é obter uma rede heurística que descreva o melhor possível o registo de eventos.

Os algoritmos *Alpha Miner*, *Heuristic Miner* e *Inductive Miner* são mais utilizados em abordagens locais, enquanto que os algoritmos *Fuzzy Miner* e *Genetic Miner* para abordagens globais (Karla e Alves De Medeiros De Medeiros s.d.).

2.7 Ferramentas

Dada a evolução da área mineração de processos, várias ferramentas foram surgindo com diversas funcionalidades. Porque atualmente existe uma grande oferta, as ferramentas são agrupadas por categoria (open-source, académicas e comerciais) e é necessário realizar um estudo avaliando as vantagens e desvantagens de cada ferramenta.

Esta secção apresenta uma análise entre as ferramentas ProM e Disco. As duas são muito completas, extensíveis e com um número considerável de plugins disponíveis. Também apresenta um estudo comparativo entre outras ferramentas disponíveis no mercado, apresentando para cada uma delas, as suas forças e fraquezas.

A tabela 7 apresenta as ferramentas da área mineração de processos agrupadas por três caraterísticas: *open-source*, académicas e comerciais (W. Van Der Aalst 2016m).

Tabela 7: Ferramentas disponíveis

Software	**Versão**	**Categoria**	**Organização**
ARIS Process Perfomance Manager	2.4	Comercial	Software AG
Celonis Process Mining	4.0	Comercial	Celonis GmbH
Disco	2.1.0	Comercial	Fluxicon
Enterprise Visualization Suite	-	Comercial	Businesscape
Fujitsu Interstage Business Process Manager Analytics	-	Comercial	Fujitsu Ltd
Minit	-	Comercial e Académica	Gradient ECM
Perceptive Process Mining	2.7	Comercial	LexMark
Petrify	4.2	Académica	Universidade Politécnica da Cataluna
Process Gold	-	Comercial	ProcessGold AG
ProM	6.7	*open-source*	Process Mining Group

2.7.1 ProM

Inserida na categoria open-source, *ProM* é considerada a ferramenta de eleição para a área mineração de processos. Foi desenvolvida pela Universidade Tecnológica de Eindhoven, e tem como foco adicionar plugins sem modificar o código já existente, suportando uma grande variedade de técnicas de mineração de dados em forma de plugins. Por outras palavras, os algoritmos a aplicar encontram-se disponíveis em forma de plugins, o que torna o processo customizável e mais rápido.

A figura 16 apresenta os cinco tipos de plugins que a ferramenta ProM disponibiliza. Os plugins do tipo *mining* são responsáveis pela integração dos algoritmos da área mineração de processos, os plugins de exportação são responsáveis por exportar os resultados (dados/gráficos/modelos) na extensão pretendida, os plugins de importação são responsáveis por realizar o *upload* dos diferentes tipos de dados ou objetos, os plugins de análise são responsáveis por analisar os resultados obtidos após aplicação do algoritmo e os plugins de conversão são responsáveis por converter os dados em qualquer formato.

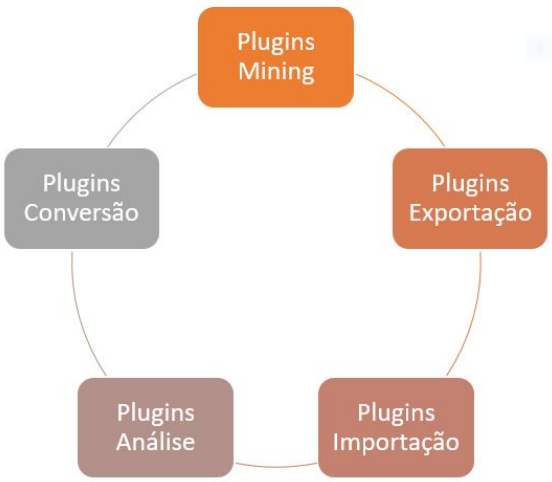


Figura 16: Tipos de Plugins ProM

Em 2004 (W. Van Der Aalst 2016n) foi lançada, a primeira versão do *ProM* (*ProM* 1.1), continha vinte e nove plugins: seis de mineração (*Alpha Miner*, *Tshinghua Miner*, *Genetic Miner*, *Multi-Phase Miner*, *Social Network Miner*, *Case Data Extraction Miner*), sete de análise, quatro de importação, nove de exportação e três de conversão de modelos.

Em Novembro de 2010, foi lançada a versão 6.0, com uma nova arquitetura e suporte para o formato *XES*. Atualmente mantém-se na versão *ProM* 6.7 e dispõe mais de mil e quinhentos plugins. Além de conseguir importar registos de eventos em formato *XES*, *MXML* e *CSV*, os resultados gerados pelos algoritmos podem ser convertidos em notação rede de *Petri*, *EPC*, *statecharts* e modelos *BPMN*.

As principais caraterísticas desta ferramenta são:

- Descobre o fluxo de controlo do processo;
- Analisa a perspetiva de recursos e desempenho do processo;
- Suporta notações como: rede de *Petri*, BPMN e modelos difusos;
- Descobre eventos com base em regras de decisão;
- Exporta resultados em diversos formatos (*CSV*, *PNG*, entre outros);
- Importa registos de eventos em diversos formatos (*XES*, *MXML* e *CSV*).

A figura 17 apresenta um screenshot do *ProM* na versão 6.5, após selecionado um registo de eventos (W. Van Der Aalst 2016o). No lado esquerdo permanece o registo de eventos previamente carregado, o pop-up central disponibiliza uma lista de algoritmos disponíveis, e após selecionar um deles, os resultados gerados são disponibilizados no painel direito.

Figura 17: Screenshot ProM 6.5

Em suma, pode-se concluir que a ferramenta *ProM* oferece uma grande variedade de opções para o tratamento dos registos de eventos, um mercado de plugins diversificado capaz de lidar com grandes quantidades de informação, produzir gráficos interessantes e robustos. Como ponto fraco, esta ferramenta apresenta uma interface ao utilizador complexa, sendo necessário dispender algum tempo a estudar todas as suas potencialidades.

2.7.2 Disco

Desenvolvida pela Fluxicon (Devi 2017), empresa sediada em Eindhoven, a ferramenta Disco insere-se na categoria comercial das ferramentas da área de mineração de processos. Tem como principal foco o alto desempenho, possui uma interface intuitiva e o esforço de aprendizagem é menor face à ferramenta *ProM*.

Tem como principais características:

- Suporta diversos formatos de registos de eventos (*XES*, *MXML*, *CSV*, *FXL Disco Logs* e *DSC Disco*);
- Foco na descoberta e análise de desempenho;
- Filtros eficientes na comparação entre processos;
- Utiliza uma variante de modelos difusos (*fuzzy models*);
- Não suporta verificação de conformidade nem suporte operacional;
- Fácil utilização, mesmo por utilizadores inexperientes ou de outras áreas.

A figura 18 apresenta um screenshot realizado da interface da ferramenta Disco, após a importação do registo de eventos e respetivo modelo de processo. A figura 19 apresenta outro screenshot, exibindo um painel de estatísticas onde se aplicou alguns filtros.

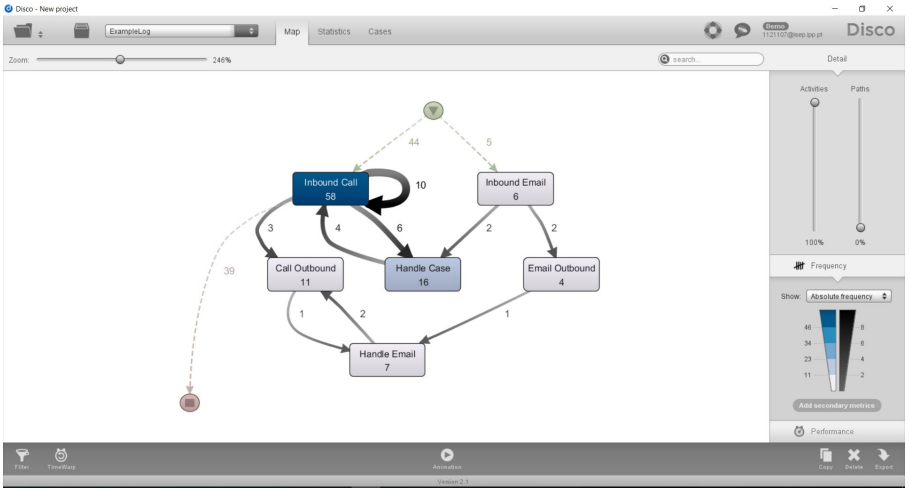


Figura 18: Screenshot Disco versão 2.1.0 - Após importação do registo de eventos

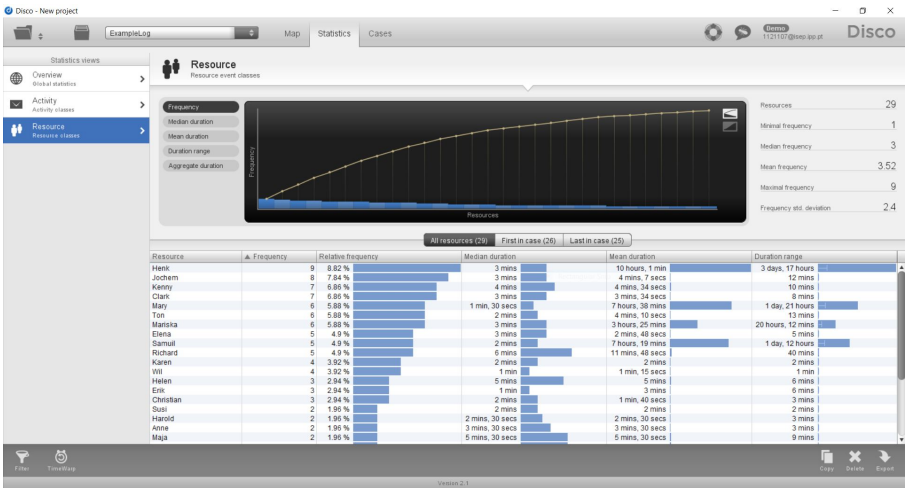


Figura 19: Screenshot Disco versão 2.1.0 - Painel Estatístico

2.7.3 Comparação Ferramentas

Estão disponíveis várias ferramentas na área mineração de processos que estão inseridas em diferentes categorias, possuem diferentes ambientes e caraterísticas.

A tabela 8 apresenta um sumário das caraterísticas das ferramentas ProM, Disco, Minit, uma comercial, uma *open-source* e uma comercial e académica.

Tabela 8: Caraterísticas Ferramentas

Caraterísticas	**ProM**	**Disco**	**Minit**
Tipos de Importação	MXML, XES	CSV, XLS/XLSX, MXML, XES, FXL	CSV, XES, MXML, SQL Server, XLS/XLSX, ODBC, Access
Tamanho registo de eventos	ilimitado	5 milhões de eventos	Sem informação
Notações	BPMN, WF, rede de *petri*, EPC, sistemas transacionais	Modelo difuso, rede de *petri*	rede de *petri*
Plataforma	Desktop	Desktop	Desktop
Filtro de dados	Sim	Sim	Sim
Tipo descoberta do processo	Sim	Sim	Sim
Tipo verificação de conformidade	Sim	Não	Não
Visualização do processo	Sim	Sim	Sim
Regras de decisão	Sim	Não	Não
Relatório de performance	Sim	Sim	Sim
Animação do processo	Não	Sim	Sim

Disco é uma ferramenta que foca-se no alto desempenho, trata de elevados conjuntos de dados. A nível de tipos de ficheiros para importação, é a ferramenta que possui maior número de formatos suportados, mas em contra-partida tem um limite máximo de dados (5 milhões). Como ponto fraco, esta ferramenta é comercial, não possuindo versão académica. Não suporta o tipo verificação de conformidade e não possui regras de decisão nos processos.

Minit (*Product - Minit Process Intelligence Software* 2018) é ferramenta recente no mercado, que se carateriza pela sua robustez e alta performance. Das três ferramentas em análise, é a que possui maior número de tipos de formatos de importação. Como ponto fraco, esta ferramenta mantém-se na categoria comercial, não possuindo versão académica.

Por último, ProM destaca-se por ser uma ferramenta *open-source*, conhecida pela comunidade da área mineração de processos e por existir documentação oficial para sua utilização. É um software completo, possui uma variedade de plugins em comparação com outras ferramentas, suporta o tipo descoberta do processo, verificação de conformidade e aprimoramento. Como pontos fracos, destacam-se a falta de animações nos processos (fluxo de dados) e os poucos formatos suportados aquando da importação dos registos de eventos.

Capítulo 3

Análise de Valor

Neste capítulo é apresentado o processo de negócio e inovação (secção 3.1) detalhado através do modelo de Peter Koen New Concept Development Model, bem como o valor para o cliente (secção 3.2) e proposta de valor (secção 3.3). É ainda apresentado o método AHP, ferramenta de apoio à tomada de decisão multicritério (secção 3.5).

3.1 Processo de Negócio e Inovação

A inovação é hoje umas das mais importantes fontes de vantagem competitiva em economias avançadas. De acordo com Peter Koen, o processo de inovação pode ser dividido em três partes distintas: Fuzzy Front End (FFE), o desenvolvimento de novos produtos New Product Development (NPD) e a comercialização. A figura 20 (Koen et al. 2011) representa o processo, envolvendo todas as partes.

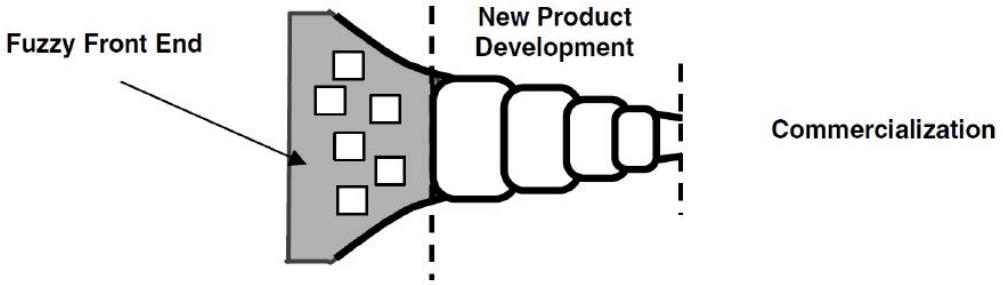


Figura 20: Processo de Inovação

- **FFE (Fuzzy Front End)**: Considerado como uma das maiores oportunidades de melhoria do processo geral de inovação. Desenvolve-se e define-se actividades que são apresentadas antes da estrutura formal NPD;
- **(NPD) Desenvolvimento novo produto** : Estruturado formalmente com um conjunto de atividades e perguntas;
- **Comercialização**: conjunto de etapas onde ocorrem a produção e lançamento do produto/serviço no mercado.

3.1.1 The New Concept Development Model (NCD)

De acordo com Peter Koen, o modelo teórico The New Concept Development (NCD) fornece uma linguagem comum para a definição das componentes chave do *Front End of Innovation* e é dividido em três partes (figura 21). O modelo apresenta uma forma circular, os elementos interagem entre si independentemente da sua ordem. Possui duas entradas e uma saída.

- **Motor**: Representa a estratégia da organização(liderança, cultura e negócio). Impulsiona os cinco-elementos-chave que são orientados pela organização;
- **Área Interna**: Define os cinco elementos-chave do modelo (identificação da oportunidade (*Opportunity Identification*), análise da oportunidade (*Opportunity Analysis*), geração da ideia (*Idea Generation & Enrichment*), seleção da ideia (*Idea Selection*) e definição de conceito (*Concept Definition*);
- **Fatores Influenciadores**: Fatores que a organização não consegue controlar, tais como: canais de distribuição, leis, política governamental, clientes, potenciais concorrentes e clima politico e ecónomico.

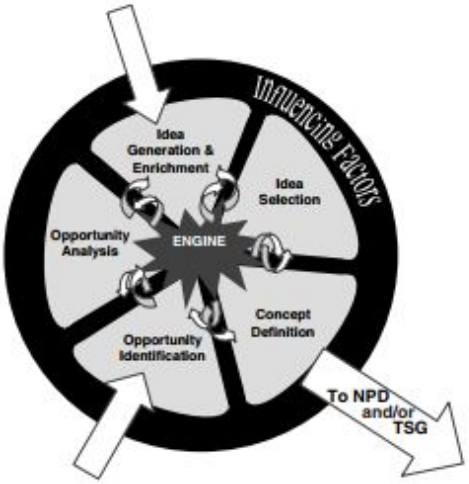


Figura 21: The New Concept Development Model (NCD)

O próximo passo descreve os cinco elementos chave neste caso de estudo, inserido na área mineração de processos.

Identificação da Oportunidade

A identificação de oportunidade surgiu no decorrer de um estudo sobre a quantidade de informação que as organizações atualmente geram no seu dia-a-dia em conjunto com a evolução da tecnologia.

Em 2011 preveram que em 2015, o número de dispositivos interligados seria duas vezes maior que a população global (Florence Hudson 2016). Como casos exemplos, o Youtube processa *uploads* de 300 horas de vídeo por minuto (Kit Smith 2017), a Google processa mais de quarenta mil pesquisas por segundo (*Google Search Statistics - Internet Live Stats* 2018). Estas ações geram dados, mais concreatamente, registos de eventos.

A quantidade de informação que as empresas armazenam continua a crescer exponencialmente, porém, as empresas não conseguem analisar toda a informação de forma eficaz.

Análise da Oportunidade Os processos organizacionais são fundamentais para o correto funcionamento e consequente crescimento a nível organizacional. A mineração de processos fornece técnicas, tipos e ferramentas para descobrir e definir estruturas de processo sustentáveis. Atualmente o mercado é muito diferenciado, dividido por várias áreas de negócio e é necessário colmatar as necessidades que as organizações encontram.

A maioria dos sistemas de informação que suportam execução de processos são inflexíveis, uniformes e são poucos os que conseguem detetar desvios e sugerir melhorias.

Este caso de estudo, possibilita a utilização de uma abordagem diferente, usando tipos diferentes e algoritmos capazes de realizar análises profundas com um nível de detalhe considerável. Irá permitir encontrar falhas, desvios e identificar oportunidades de melhorias.

Geração da Ideia

Após análise da oportunidade, surgiram três ideias:

(i) Desenvolver métodos de análise a fim de comparar processos;

(ii) Utilizar um conjunto de dados (*datasets*) reais, provenientes de um projeto que envolve o Instituto Superior de Engenharia do Porto;

(iii) Utilizar um conjunto de dados (*datasets*) *on-line*, disponíveis pelo site oficial da área mineração de processos.

Seleção da Ideia

Todas as ideias foram consideradas e analisadas. A segunda ideia, utilizar um conjunto de dados (*datasets*) reais, provenientes de um projeto que envolve o Instituto Superior de Engenharia do Porto foi a escolhida. O caso de estudo usando dados reais, torna o projeto mais interessante, e sobretudo mais vantajoso.

Definição de Conceito

O último elemento do modelo NCD permite a evolução do processo para a fase de desenvolvimento de um novo produto. Não irá ser criado um novo produto ou serviço. Irá ser desenvolvido uma solução de software capaz de determinar processos de negócio a partir de dados de eventos gerados por várias aplicações de várias organizações.

3.2 Valor para o Cliente

Criação de valor é um conceito difícil de atingir, compreender, modelar e/ou contextualizar. Alguns autores considerem a criação de valor um *trade-off* entre os benefícios e sacrifícios (Lancaster 2000). Para outros autores o valor é definido como: necessidade, desejo, interesse, crença, atitudes e preferência (Nicola, E. P. Ferreira e J. J. P. Ferreira 2012).

A solução de software a desenvolver irá dar resposta a vários clientes de vários sectores. Num primeiro nível a solução irá ser comercializada para um único cliente que mais tarde irá ficar responsável por comercializar para outros clientes. Dado que este trabalho é um caso de estudo na área mineração de processos, todo o trabalho desenvolvido irá ajudar no

crescimento do estudo nesta área, mas também na contribuição de trabalhos em casos reais, dentro da comunidade.

3.3 Valor Percebido

Segundo Lindgreen e Wynstra (Lindgreen e Wynstra 2005), o valor percebido do ponto de vista do produtor pode ser diferente do ponto de vista do cliente. Enquanto que o produtor é menos sensível ao preço do produto/serviço, o cliente é mais sensível à qualidade do produto.

Podem existir várias percepções para o mesmo produto/serviço, para diferentes clientes. O valor percebido para o cliente é composto pela diferença entre os benefícios e os sacrifícios. Por outras palavras, os ganhos têm que compensar as perdas. O valor percebido é representado pela fórmula:

$$Valor = Beneficios - Sacrificios > 0$$

A tabela 9 apresenta os benefícios e sacrifícios associados ao valor criado para o cliente com o desenvolvimento do caso de estudo, segundo "Conceptualising Value for the Customer"(Woodall 2003).

Tabela 9: Benefícios e Sacrifícios (Woodall 2003)

Benefícios		**Sacrifícios**
Atributos	**Outcomes**	
Independente do tipo de dados recolhidos	Redução dos custos nos processos	Esforço de aprendizagem
Performance	Prevenção de falhas e desvios	
Tempo de execução	Eficiência na análise dos dados	
Independente das restrições tecnológicas da organização	Redução de recursos na localização e resolução de estrangulamentos	Energia
Processos evolutivos	Visualização do desempenho da organização em tempo real	Tempo
Agnóstico à ferramenta	Previsão e simulação do comportamento futuro do processo	Investimento em ferramentas pagas
Compreensão do processo de negócio		

3.4 Proposta de Valor

A proposta de valor define uma estratégia específica de modo a competir por novos clientes (Jalili and Rezaie, 2010). A criação de valor é fundamental para a sustentabilidade e

crescimento de uma organização.

A proposta de valor desta tese de mestrado é especificar, analisar e melhorar os processos existentes numa organização. A solução a desenvolver irá apresentar valor para qualquer organização que pretenda analisar e melhorar os seus processos através dos dados registados diariamente pelos sistemas de informação.

3.5 Método AHP

Analytic Hierarchy Process (AHP), método multicritério de auxílio à tomada de decisão foi desenvolvido na década de 80 por Thomas L. Saaty (Saaty 2008).

Este método permite o uso de critérios qualitativos e quantitativos, e foca-se por dividir o problema de decisão em níveis hierárquicos. A figura 22 representa a decomposição da estrutura hierárquica do modelo AHP.

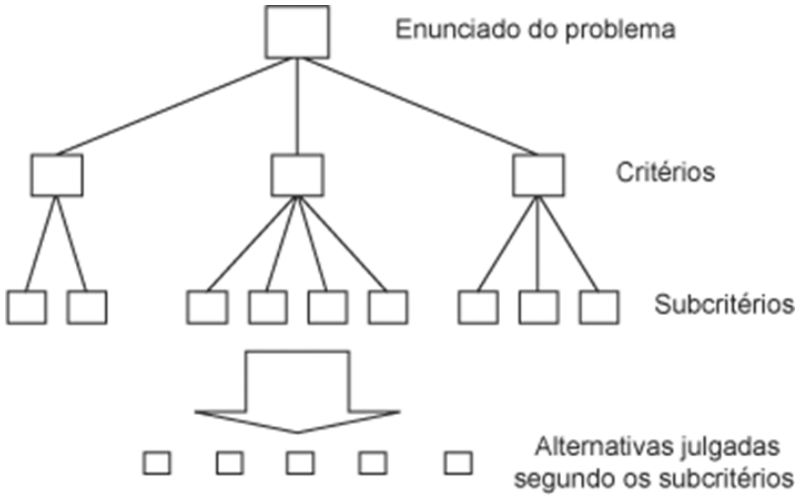


Figura 22: Decomposição Estrutura Hierárquica

É necessário decompor o método em AHP em etapas, nomeadamente:

- Definir o problema e construir árvore de hierarquia de decisão;
- Comparar alternativas e critérios;
- Definir prioridades relativas de cada critério;
- Avaliar a consistência das prioridades relativas;
- Construir a matriz de comparação paritária para cada critério;
- Obter a prioridade composta para as alternativas;
- Escolher a alternativa;

3.5.1 Construção Árvore de Hierarquia de Decisão

O objetivo é aplicar o método AHP a fim de determinar a **melhor ferramenta para a mineração de processos** baseando nos seguintes critérios:

(i) **Categoria**: As ferramentas estão agrupadas por categorias (*opensource*, comerciais, académicas);

(ii) **Formatos Importação**: Existem vários tipos formatos de importação que as ferramentas podem suportar (MXML, XES, CSV, XLS/XLSX, FXL, *logs* ODBC, *logs* Access e *logs* SQL Server);

(iii) **Tipos**: Na área mineração de processos, existem três tipos, nomeadamente, descoberta do processo, verificação de conformidade e aprimoramento. Cada tipo possui um objetivo distinto e uma abordagem específica. Os três tipos encontram-se detalhadas na subsecção 2.3.1, subsecção 2.3.2 e subsecção 2.3.3;

(iv) **Notações**: Existem várias as notações que são usadas pelos modelos de processo e suportadas pelas ferramentas. Exemplos de notações: notação rede de *petri*, sistema transacional, BPMN, WF e EPC;

(v) **Plugins**: Existem vários plugins disponíveis, que podem ser integrados nas ferramentas de forma a facilitar vários procedimentos do processo, tais como: limpar e análisar registos de eventos, aplicar algoritmos, gerar resultados aplicando filtros, exportar para determinados formatos, entre outros.

A figura 23 representa a árvore hierárquica de decisão. No primeiro nível encontra-se o objetivo geral de decisão, no segundo nível, os críterios associados ao problema de decisão e por último (terceiro nível) encontram-se as alternativas disponíveis e mais adequadas.

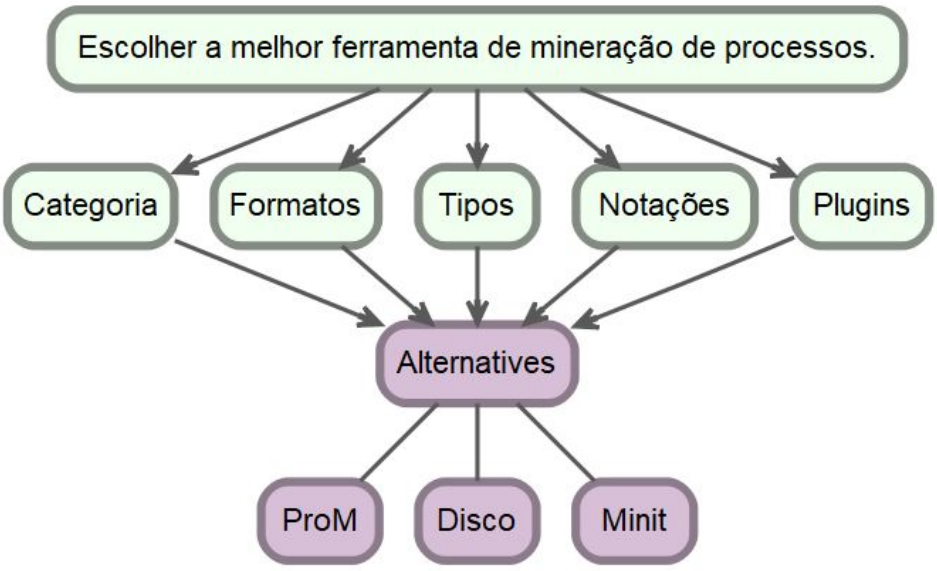


Figura 23: Árvore de hierarquia

3.5.2 Comparação entre os Elementos da Hierárquia

Nesta fase estabelece-se prioridades entre os elementos de cada nível da hierarquia, através de uma matriz de comparação. É necessário determinar uma escala de valores para comparação, que não deve exceder um total de nove fatores, a fim de, manter a matriz consistente. A tabela 10, definida por Saaty, representa a escala fundamental que irá ser usada para atribuir os pesos (Saaty 2008).

Neste caso, definiu-se três alternativas (ProM, Disco e Minit). Cada uma das alternativas irá ser comparada com as outras duas: **ProM vs Disco**, **ProM vs Minit**, **Disco vs Minit**. Para cada comparação será determinado qual é a alternativa mais fraca em relação à outra, no critério de avaliação específico.

Tabela 10: Escala Fundamental - Níveis de importância de comparações

Intensity of Importance	Definition	Explanation
1	Equal Importance	Two activities contribute equally to the objective
2	Week or slight	
3	Moderate importance	Experience and judgement slightly favor one activity over another
4	Moderate plus	
5	Strong importance	Experience and judgement strongly favor one activity over another
6	Strong plus	
7	Very strong or demonstrated importance	An activity is favored very strongly over another; its dominance demonstrated 5 in practice
8	Very, very strong	
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation
Reciprocals of above	If activity i has one of the above non-zero numbers assigned to it when compared with activity j, then j has the reciprocal value when compared with i	A reasonable assumption
1.1-1.9	If activities are very close	May be difficult to assign the best value but when compared with other contrasting activities the size of the small numbers would not be too noticeable, yet they can still indicate the relative importance of the activities.

3.5.3 Comparação Alternativas vs Critério

Critério Categoria

A tabela 11 apresenta as comparações para cada par de alternativas, para o critério categoria. Os pesos foram atribuídos de acordo com a tabela 10.

Tabela 11: Matriz comparação par-a-par: Critério Categoria

Alternativas - Comparação Categoria			
ProM	7	Disco	3
ProM	9	Minit	1
Disco	5	Minit	2

ProM possui uma grande vantagem face à ferramenta Disco e Minit, no sentido de ser uma ferramenta totalmente *opensource* e gratuita a todos os utilizadores. A ferramenta Disco é categorizada por ser comercial e académica, pelo que é necessário licença para sua utilização. Em relação à ferramenta Minit, esta categorizar-se por uma ferramenta somente comercial. Detém desvantagem sobre a ferramenta Disco, por não abranger a comunidade académica.

A tabela 12 apresenta a matriz de pesos para o critério categoria.

Tabela 12: Matriz de pesos: Critério Categoria

Categoria	ProM	Disco	Minit
Prom	1	7	9
Disco	1/7	1	5
Minit	1/9	1/5	1

Critério Formatos Importação

A tabela 13 apresenta as comparações para cada par de alternativas, para o critério formatos de importação. Os pesos foram atribuídos de acordo com a tabela 10.

Tabela 13: Matriz comparação par-a-par: Critério Formatos Importação

Alternativas - Comparação Formatos Importação			
ProM	3	Disco	5
ProM	3	Minit	9
Disco	6	Minit	9

ProM possui desvantagem sobre a ferramenta Disco, dado que, suporta apenas dois formatos de importação (MXML e XES) enquanto que Disco, suporta cinco formatos (CSV, XLS/XLSX, MXML, XES e FXL). A ferramenta Minit é a mais completa, possui uma grande vantagem sobre as restantes dado que suporta sete formatos de importação.

A tabela 14 apresenta a matriz de pesos para o critério formatos de importação.

Tabela 14: Matriz de pesos: Critério Formatos Importação

Formatos Importação	ProM	Disco	Minit
Prom	1	1/5	1/9
Disco	5	1	1/9
Minit	9	9	1

Critério Tipos

A tabela 15 apresenta as comparações para cada par de alternativas, para o critério tipos. Os pesos foram atribuídos de acordo com a tabela 10.

Tabela 15: Matriz comparação par-a-par: Critério Tipos

Alternativas - Comparação Tipos			
ProM	7	Disco	3
ProM	7	Minit	3
Disco	1	Minit	1

ProM possui vantagem sobre as outras duas ferramentas, pois suporta todas os tipos (descoberta do processo, verificação de conformidade e aprimoramento). A vantagem entre a ferramenta Disco e Minit é nula, dado que as duas suportam o tipo (descberta do processo).

A tabela 16 apresenta a matriz de pesos para o critério tipos.

Tabela 16: Matriz de pesos: Critério Tipos

Tipos	ProM	Disco	Minit
Prom	1	7	7
Disco	1/7	1	1/9
Minit	1/7	1	1

Critério Notações

A tabela 17 apresenta as comparações para cada par de alternativas, para o critério notações. Os pesos foram atribuídos de acordo com a tabela 10.

Tabela 17: Matriz comparação par-a-par: Critério Notações

Alternativas - Comparação Notações			
ProM	8	Disco	4
ProM	8	Minit	3
Disco	4	Minit	3

ProM apresenta maior vantagem face às ferramentas Disco e Minit, dado que suporta cinco notações (BPMN, WF, rede de *petri*, EPC e sistemas transacionais). Disco possui uma vantagem sobre Minit, pois suporta mais um formato (modelo difuso). Minit suporta apenas a notação rede de *petri*, comum às restantes.

A tabela 18 apresenta a matriz de pesos para o critério notações.

Tabela 18: Matriz de pesos: Critério Notações

Notações	ProM	Disco	Minit
Prom	1	8	8
Disco	1/8	1	4
Minit	1/8	1/4	1

Critério Plugins

A tabela 19 apresenta as comparações para cada par de alternativas, para o critério plugins. Os pesos foram atribuídos de acordo com a tabela 10.

Tabela 19: Matriz comparação par-a-par: Critério Plugins

Alternativas - Comparação Plugins			
ProM	9	Disco	3
ProM	9	Minit	3
Disco	1	Minit	1

ProM apresenta maior vantagem face às ferramentas Disco e Minit, dado que possui uma estrutura extensa que suporta uma grande variedade de plugins (*minig*, conversão, exportação, análise e importação). A vantagem entre a ferramenta Disco e Minit é nula, dado que as duas suportam o mesmo número de plugins (básicos).

A tabela 20 apresenta a matriz de pesos para o critério plugins.

Tabela 20: Matriz de pesos: Critério Plugins

Plugins	ProM	Disco	Minit
Prom	1	9	9
Disco	1/9	1	1
Minit	1/9	1	1

3.5.4 Comparação Critérios vs Objetivo

Após analisar as alternativas com os critérios, é também necessário avaliar os critérios face ao objetivo definido. Na tabela 21 é possivel observar a comparação par-a-par, dando um peso relativo para cada critério.

Tabela 21: Comparação critério vs objectivo par-a-par

Critério	Categoria	Formatos Importação	Tipos	Notações	Plugins
Categoria	1	2	1	2	2
Formatos Importação	1/2	1	1/2	2	1/2
Tipos	1	2	1	2	1/2
Notações	1/2	1/2	1/2	1	1/2
Plugins	1/2	2	2	2	1

3.5.5 Ferramenta RStudio

De modo a facilitar o processo da escolha da alternativa, fez-se uso da ferramenta *opensource* RStudio (*RStudio* s.d.) em conjunto com o repositório público (*GitHub Analytical Hierarchy Process (AHP) with R* s.d.), de forma a aplicar o método AHP na linguagem R.

Após instalação da ferramenta RStudio, foi necessário instalar o package *devtools* e compilar o código fonte do ficheiro "ahp.R"(repositório GitHub). O código fonte utilizado encontra-se no Apêndice A.

O resultado da compilação do código fonte gera uma interface independente que pode ser acedida pelo browser. Na interface é necessário realizar o *upload* do ficheiro ahp em formato YAML. A figura 24 apresenta a estrutura exigida para o ficheiro AHP. O ficheiro AHP utilizado para o caso em estudo, encontra-se no Apêndice B.

```
Version
Alternatives
  alternative 1
    property 1 (optional)
    property 2 (o)
    ...
  alternative 2
    property 1 (o)
    property 2 (o)
  ...
Goal
  decision-makers (o)
  preferences
    decision maker 1 (o)
      scoreFunction or
      score or
      pairwiseFunction or
      pairwise or
      priority
    decision maker 2
      ...
  children
    criteria 1
      preferences
      childrend
        sub-criteria 1.1
        sub-criteria 1.2
          children: *alternatives
      ...
    criteria 2
      ...
```

Figura 24: Estrutura Genérica Ficheiro AHP

a

3.5.6 Escolha da Alternativa

No sentido de encontrar a melhor ferramenta para a área mineração de processos, a alternativa eleita foi a ferramenta ProM. O conjunto de matrizes de pesos de todos os critérios foram escritos na linguagem R e aplicados na ferramenta RStudio. O resultado obtido está representado na figura 25.

	Weight	ProM	Minit	Disco	Inconsistency
Escolher a melhor ferramenta de mineracao de processos.	100.0%	68.3%	18.1%	13.6%	4.8%
Categoria	28.7%	22.2%	1.6%	5.0%	18.0%
Plugins	25.2%	20.7%	2.3%	2.3%	0.0%
Tecnicas	21.5%	16.7%	2.4%	2.4%	0.0%
Formatos	14.0%	0.7%	11.1%	2.1%	25.4%
Notacoes	10.6%	8.1%	0.7%	1.8%	18.7%

Figura 25: Resultados Finais

Para os cinco critérios foram atribuídos pesos diferentes: categoria 28.7%, plugins 25.2%, tipos 21.5%, formatos 14.0% e notações 10.6%.

A ferramenta **ProM** foi eleita com **68.3%**, enquanto que a ferramenta Disco e Minit alcançaram 18.1% e 13.6% respetivamente.

É de salientar que a percentagem do critério mais baixo da ferramenta ProM foi os formatos de importações (**0.7%**) e a percentagem do critério mais alto foi a categoria (**22.2%**). A ferramenta Minit ultrapassou a ferramenta Disco dado que, no critério formatos de importação atingiu 11.1% face a 2.1%.

Capítulo 4

Caso de Estudo

Este capítulo apresenta a base de dados MIMIC III que é usada neste estudo, sua estrutura, principais caraterísticas e modelo de dados (secção 4.1). Posteriormente, são apresentados os requisitos específicos para este caso de estudo (secção 4.2).

4.1 Medical Information Mart for Intensive Care III

MIMIC III é uma base de dados *open-source*, disponível gratuitamente para a comunidade, que inclui diversos dados médicos associados a mais de quarenta mil pacientes que permaneceram em unidades de cuidados intensivos do centro médico *Beth Israel Deaconess*, entre os anos 2001 e 2012. Este centro médico encontra-se localizado em Boston, Massachusetts (Johnson et al. 2016a).

A base de dados possui, entre outros, dados demográficos, sinais vitais, resultados de exames laboratoriais, procedimentos, medicamentos, anotações médicas, relatórios de imagem, tempos de permanência nas unidades e dados sobre mortalidade. A tabela 22 apresenta uma visão geral dos tipos de dados disponíveis (Johnson et al. 2016c).

Tabela 22: Tipos de dados disponíveis

Área	**Relacionado com**
Faturação	Faturação e administração
Entradas/Saídas	Entradas e saídas do hospital
Exames	Exames e intervenções
Laboratórios	Dados recolhidos em laboratório (análises de sangue, urina e hematologia)
Medicação	Dados de administração de medicação por via intravenosa e pedidos de medicação
Altas	Progresso do paciente e resumos de altas hospitalares
Fisiologia	Sinais vitais (frequência cardíaca e pressão arterial)
Relatórios	Relatórios de texto livre (estudos de imagem, electrocardiogramas e ressonências)

MIMIC III destaca-se por vários fatores:

(i) : Está disponível gratuitamente para investigadores de todo o mundo;

(ii) : Abrange vários grupos de pacientes;

(iii) : Contém dados com mais de uma década, com dados detalhados sobre o atendimento individual do paciente;

(iv) : Existe grande suporte a nível documental.

MIMIC III contém dados de cinco unidades de cuidados intensivos: Coronary Care Unit (CCU), Cardiac Surgery Recovery Unit (CSRU), Medical Intensive Care Unit (MICU), Surgical Intensive Care Unit (SICU) e Trauma Surgical Intensive Care Unit (TSICU).

Estão registadas cinquenta e três mil e quatrocentos e vinte e três (53.423) admissões, dos quais trinta e oito mil e quinhentos e noventa e sete (38.597) são admissões de pacientes adultos. Adicionalmente a idade média do paciente é 65.8, e a taxa de pacientes do sexo masculino é 55.9%. A tabela 23 apresenta alguns dados estatísticos correspondentes às cinco unidades de cuidados intensivos (Johnson et al. 2016b).

Tabela 23: Detalhes por unidade de cuidados intensivos

Unidade	**CCU**	**CSRU**	**MICU**	**SICU**	**TSICU**	**Total**
Nº pacientes distintos (% admissões)	5,674 (14,7%)	8,091 (20,9%)	13,649 (35,4%)	6,372 (16,5%)	4,811 (12,5%)	38,597 (100%)
Nº admissões hospital (% admissões)	7,258 (14,6%)	9,156 (18,4%)	19,770 (39,7%)	8,110 (16,3%)	5,491 (11,0%)	49,785 (100%)
Idade média (%)	70,1	67,6	64,9	63,6	59,9	65,8
Tempo médio de permanência ICU (dias)	2,2	2,2	2,1	2,3	2,1	2,1
Tempo médio de permanência hospital (dias)	5,8	7,4	6,4	7,9	7,4	6,9
Taxa de Mortalidade ICU	8,9%	3,6%	10,5%	9,1%	8,4%	8,5%
Taxa de Mortalidade hospital	11,3%	4,6%	14,5%	12,6%	11,4%	11,5%

A base de dados MIMIC III é composta por 26 tabelas sendo que 16 contém dados temporais.

Cada tabela possui o identificador do paciente ("subject_id") e o identificador de admissão no hospital ("hadm_id"). As tabelas que possuem o sufixo "ID"no identificador da tabela, são tabelas que contêm dados relativos ao paciente e ao seu percurso. Já as tabelas que possuem o prefixo "D_"são tabelas que fornecem definições para identificadores.

A tabela 24 apresenta informações das dezasseis tabelas existentes na base de dados MIMIC III com alguma referência temporal.

Tabela 24: Tabelas existentes em MIMIC III com referência temporal

Nome	**Relacionado com**	**Nº registos**	**Nº Atividades**	**Nº pacientes**
admissions	Admissão do paciente no hospital	58976	8	7361
Continua na página seguinte				

Tabela 24 continuada da página anterior

Nome	Relacionado com	N° registos	N° Atividades	N° pacientes
callout	Momento quando um paciente está pronto para receber alta do ICU	34499	6	4197
chartevents	Dados do paciente durante a permanência no hospital (sinais vitais, configurações do ventilador, valores laboratoriais, estado mental)	33766594	2580	7359
cptevents	Códigos de procedimento. Facilitam a faturação dos procedimentos realizados em pacientes	573146	1	3327
datetimeevents	Dados recolhidos nos cuidados intensivos	4485937	148	3327
icustays	Dados sobre a permanência no ICU	61532	2	7345
inputevents_cv	Pacientes cujos dados foram originalmente armazenados na base de dados CareVue	17527935	756	3924
inputevents_mv	Pacientes cujos dados foram originalmente armazenados na base de dados MetaVision	3618991	251	3850
labevents	Eventos relacionados com testes laboratoriais	27854055	556	7351
microbiologyevents	Eventos relacionados com testes microbiológicos	631726	47	3457
noteevents	Notas associadas às permanências hospitalares	2083180	584	5351

Continua na página seguinte

Tabela 24 continuada da página anterior

Nome	Relacionado com	Nº registos	Nº Atividades	Nº pacientes
outputevents	Outputs registados durante a permanência no ICU	4349218	415	7278
prescriptions	Pedidos relacionados com prescrições médicas	4156450	2697	6900
procedureevents_mv	Procedimentos (início/fim) que foram registados para os pacientes da base de dados MetaVision	258066	114	3853
services	Serviços hospitalares do paciente internado	73343	18	7357
transfers	Localização do paciente durante o internamento	261897	8	6902

A tabela 25 apresenta informações das dez tabelas existentes na base de dados MIMIC III que não possuem qualquer referência temporal.

As cinco tabelas *'patients'*, *'admissions'*, *'icustays'*, *'services'* e *'transfers'* são usadas para definir e monitorizar o internamento do paciente no hospital. Todas as tabelas que possuem o prefixo 'd_' apenas definem códigos de referência para tipos de doenças, estados, entre outros.

Tabela 25: Tabelas existentes em MIMIC III sem referência temporal

Nome tabela	Relacionado com	Nº registos
caregivers	Prestadores de cuidados durante o internamento (médico, enfermeiro)	7567
d_cpt	Códigos procedimentais Current Procedural Terminology (CPT)	134
d_icd_diagnoses	Códigos da classificação internacional de doenças (ICD-9) para diagnóstico	14710
d_icd_procedures	Códigos da classificação internacional de doenças (ICD-9) para procedimentos	3898
d_items	Items da base de dados CareVue e MetaVision	12487
d_labitems	Items relacionados com os laboratórios	753
diagnoses_icd	Diagnósticos relacionados à internação hospitalar	651047
drgcodes	Códigos de grupos relacionados ao dignóstico de pacientes	125557
patients	Pacientes admitidos	46520
procedures_icd	Procedimentos para os pacientes (procedimentos ICD-9)	240095

Na figura 26 é possível visualizar o modelo de dados geral do sistema MIMIC III. Dado o número elevado de atributos presentes nas tabelas, o modelo encontra-se simplificado, contendo os atributos e ligações mais significativos.

As tabelas com maior nível de ligações são *'patients'*, *'admissions'* e *'icustays'*. O atributo *'subject_id'* está presente na tabela 'patients' como identificador único que especifica o paciente. Já o atributo *'hadm_id'* está presente na tabela 'admissions' e representa a admissão de cada paciente no hospital. Por fim, o atributo *'icustays_id'* está presente na tabela 'icustays' como identificador único que representa a permanência do paciente na unidade de cuidados intensivos.

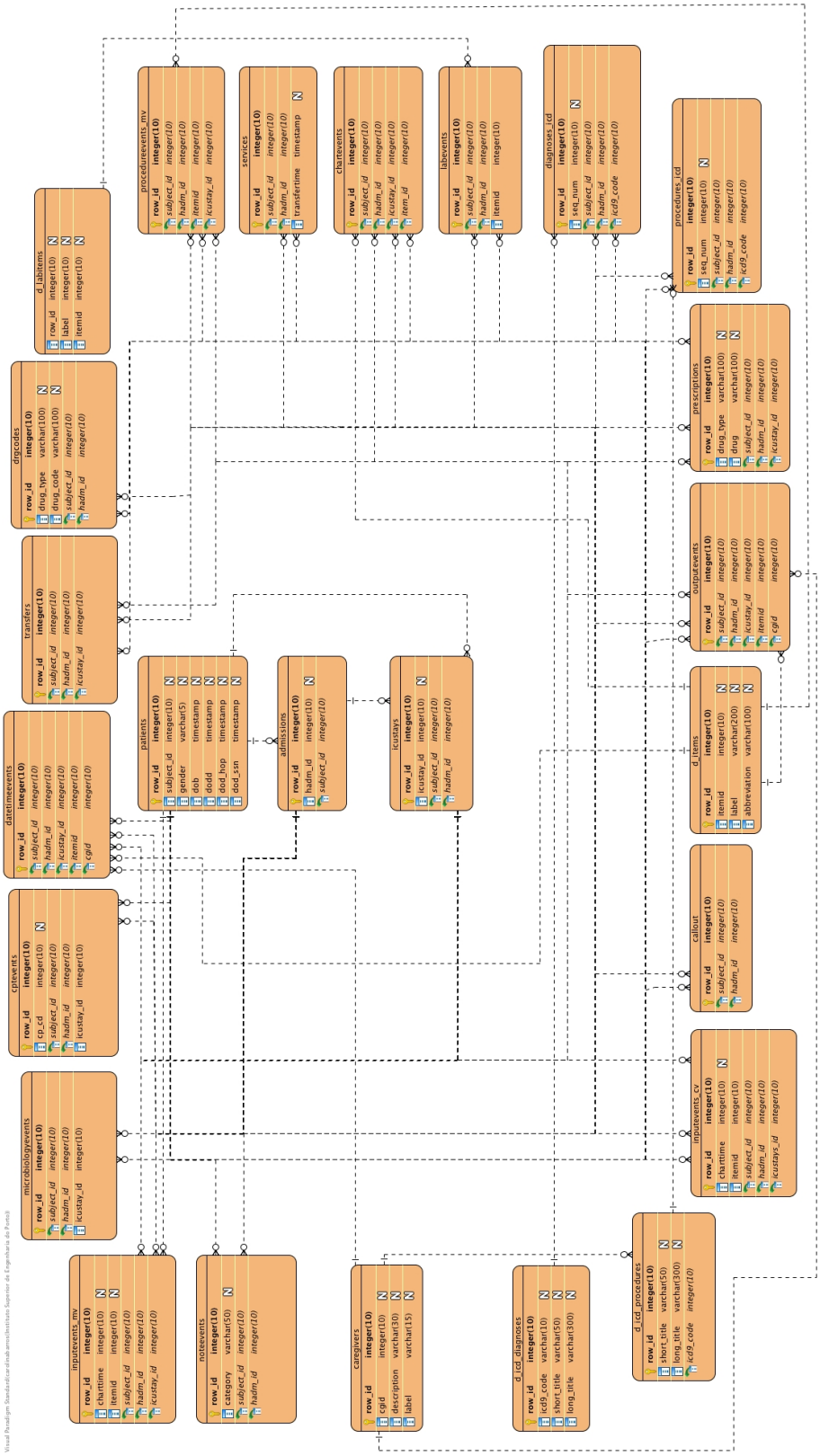


Tabela 26: Modelo de dados do sistema MIMIC III

4.2 Requisitos

O levantamento de requisitos é uma das etapas fundamentais para o desenvolvimento de um sistema. Após identificar problemas e analisar necessidades, procede-se à definição dos requisitos funcionais e não funcionais.

A área de mineração de processos visa melhorar a extração de conhecimento a partir de registos de eventos estruturados, não estruturados ou semiestruturados, fornecendo técnicas e ferramentas ao longo de todo o processo. Este caso de estudo é aplicado à área da saúde e utiliza dados médicos e administrativos do centro médico *Beth Israel Deaconess* (4.1) apresentados na secção 4.1 . O foco concentra-se na descoberta e verificação de percursos oncológicos, identificando possíveis ineficiências, comportamentos e situações que possam estar a comprometer sucessos no tratamento de doenças oncológicas.

São milhares os dados disponíveis na base de dados MIMIC III. Possui informação sobre 40000 pacientes, por um período de 12 anos. As possibilidades para este caso de estudo são muitas e houve a necessidade de definir um escopo específico, identificando algumas questões para este caso.

No total existem 7361 pacientes que foram diagnosticados com pelo menos um tipo de cancro (um paciente pode ser diagnosticado com mais que um tipo de cancro). Ao todo estão registados 13 tipos de cancro e 10857 admissões no centro médico.

Os tipos de cancro com maior número de admissões foram o *"Malignant neoplasm of other and unspecified sites"*, com o código compreendido entre 190-199, de acordo com a classificação internacional de doenças (Centers for Disease Control and Prevention 2015). Todos os códigos estão presentes na tabela de dados 'diagnoses_icd'.

Os pacientes diagnosticados com o tipo de cancro *"Malignant neoplasm of other and unspecified sites"*, código 190-199 foram o nicho de pacientes escolhidos para o caso de estudo, uma vez que apresentam maior número de admissões (3950) e pacientes (2846). O código 190-199 (International Classification of Diseases (ICD)) integra 10 tipos de cancro (*Online ICD9/ICD9CM codes* 2018) :

- **190** *Malignant neoplasm of eye*;
- **191** *Malignant neoplasm of brain*;
- **192** *Malignant neoplasm of other and unspecified parts of nervous system*;
- **193** *Malignant neoplasm of thyroid gland*;
- **194** *Malignant neoplasm of other endocrine glands and related structures*;
- **195** *Malignant neoplasm of other and ill-defined sites*;
- **196** *Secondary and unspecified malignant neoplasm of lymph nodes*;
- **197** *Secondary malignant neoplasm of respiratory and digestive systems*;
- **198** *Secondary malignant neoplasm of other specified sites*;
- **199** *Malignant neoplasm without specification of site*.

Após selecionado o alvo de pacientes, identifica-se os seguintes requisitos específicos para o caso de estudo:

(i) Qual é o percurso mais comum com taxa de insucesso?;

(ii) Qual é o percurso mais comum com taxa de sucesso?;

(iii) Qual é o percurso mais curto (número de atividades) com taxa de sucesso?;

(iv) Qual é o percurso mais curto (tempo) com taxa de sucesso?.

Dado os requisitos gerais definidos na secção 1.2 e os requisitos específicos definidos nesta secção, identifica-se na tabela 27 a possibilidade de relação entre o primeiro requisito geral com o primeiro requisito específico, o mesmo para os restantes. Se houver relação entre o requisito geral e o requisito específico, é colocado o símbolo 'X'. Se não houver qualquer relação é colocado o símbolo '-'. A sigla **RG1** significa requisito geral e o número 1 corresponde ao índice do requisito na lista respectiva.

Tabela 27: Relação entre requisitos gerais e requisitos específicos

Gerais/Específicos	**RE1**	**RE2**	**RE3**	**RE4**
RG1	X	X	X	X
RG2	X	X	X	X
RG3	X	X	X	X
RG4	-	-	-	-

Em suma, todos os requisitos específicos não devem atingir o requisito geral nº4 ("Avaliar os modelos de processo de forma a encontrar e analisar possíveis desvios e anomalias"), dado que, para encontrar uma anomalia ou estrangulamento no processo seria necessário a análise de um perito da área. Uma vez que o autor não possui conhecimento técnico na área médica, o requisito geral nº4 não será alcançado. Os desvios poderão ser encontrados mas não analisados com detalhe.

Capítulo 5

Construção da Solução

Este capítulo apresenta o processo de implementação, incluindo a descrição do fluxo de trabalho e apresentação da estrutura que constitui os componentes (secção 5.1). Adicionalmente é apresentado a seleção dos algoritmos utilizados e ferramentas (secção 5.2), análise e preparação dos dados (secção 5.3), e por fim a mineração dos processos (secção 5.4). Esta última secção é composta por cinco subseções que descrevem o processo e resultados obtidos para cada um dos requisitos.

5.1 Estrutura

Na figura 26 é possível visualizar a arquitetura do sistema. É composta por seis componentes, sendo que o componente Hospital agrega cinco subsistemas, nomeadamente:

(i) Unidade de cuidados intensivos (ICU): composta por cinco unidades CCU, CSRU, MICU, SICU e TSICU, responsável pela monitorização e gestão das mesmas unidades (alarmes, sinais vitais, medicação, progressos do paciente, entre outros);

(ii) Financeiro: responsável por gerir a área financeira do hospital e procedimentos a realizar;

(iii) Demográfico: responsável por gerir os dados demográficos do paciente, bem como as admissões e altas realizadas dentro do hospital;

(iv) Documentação: responsável por gerir os relatórios médicos, laboratoriais e sumários de alta médica.

A base de dados Social Security Death Index (SSDI) é um componente externo que contém todos os registos de óbitos desde 1936 nos Estados Unidos da América. É frequentemente utilizada pelos hospitais nacionais e por investigadores da área genealógica.

O componente **Hospital** e a base de dados **SSDI** alimentam a base de dados do hospital. Este componente é responsável por realizar a persistência de todos os dados gerados, assim como, converter formatos de dados e realizar mapeamentos. Os dados tratamentos são persistidos na base de dados final **MIMIC III**.

O componente **Sistema de Análise e Processamento** é responsável por extrair os dados disponíveis na base de dados **MIMIC III**, processar os dados, realizar ações de limpeza, filtro e transformação, aplicar algoritmos, e por fim gerar modelos de processo.

O componente **Modelo de Avaliação de Processos** é responsável por avaliar os modelos de processo gerados, discutir resultados e realizar propostas de melhoria.

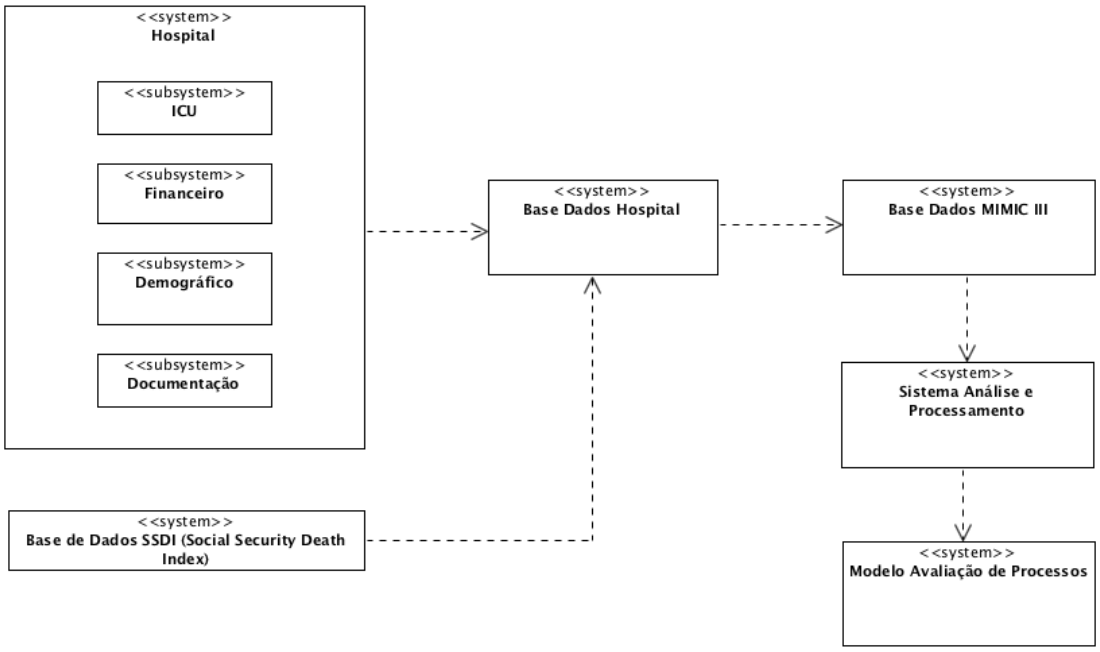


Figura 26: Diagrama de Arquitetura

A figura 27 apresenta o fluxo de trabalho a seguir, contendo as etapas que constituem o processo de desenvolvimento:

(i) **Planeamento**: definição do problema e dos objetivos a alcançar. Especificar os requisitos e o nicho de pacientes que serão abrangidos;

(ii) **Extração dados**: extração, seleção, mapeamento e processamento dos dados. Necessário estudo profundo dos dados existentes e de que forma estes podem ser interligados;

(iii) **Importação dos dados para a ferramenta**: seleção e aplicação de plugins de conversão;

(iv) **Transformação dados**: seleção e aplicação de plugins de limpeza e enriquecimento;

(v) **Aplicação algoritmos**: seleção e aplicação de plugins que executam o algoritmo no registo de eventos de forma a gerarem modelos de processo;

(vi) **Análise e processamento dos modelos de processo gerados**: apresentação dos modelos de processo gerados pela ferramenta e realização de uma análise detalhada dos resultados obtidos;

(vii) **Avaliação do processo**: avaliação dos modelos de processo gerados de acordo com métricas de avaliação.

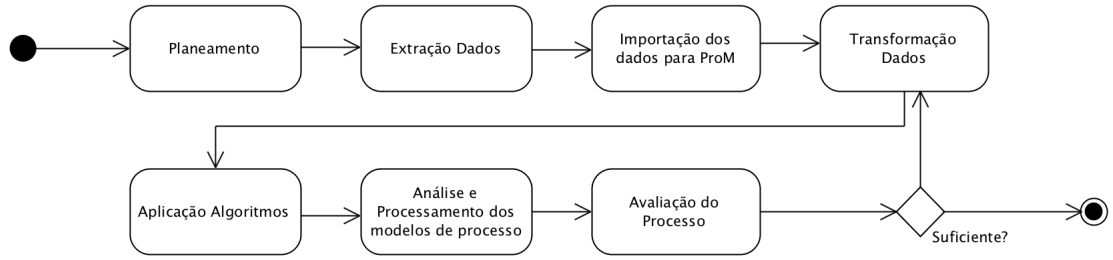


Figura 27: Fluxo de trabalho

O planeamento encontra-se descrito na secção 4.2 do capítulo 4. A implementação descrita na secção 5 apresenta as fases de extração de dados, importação dos dados para a ferramenta ProM, transformação de dados, aplicação dos algoritmos e análise e processamento dos modelos de processo. A avaliação do processo encontra-se descrito no capítulo 6.

5.2 Seleção Algoritmos e Ferramentas

Para persistência de dados é utilizado o PostgreSQL, sistema de gestão de base de dados relacional (DBMS) (PostgreSQL 2018). A decisão de de utilização toma em consideração o facto de ser *open-source*, possuir documentação variada, ser extensível e o excelente desempenho com grande volume de dados. Para a análise, mapeamento e gestão de dados foi utilizado o pgAdmin versão 4, interface gráfica de gestão e desenvolvimento para PostgreSQL (Pgadmin 2018).

Para o desenvolvido da mineração de processos a ferramenta escolhida foi a ferramenta *open-source* ProM. A decisão da escolha encontra-se descrita na subsecção 3.5.6.

Dos algoritmos descritos na secção 2.6 são utilizados os algoritmos Heuristic Miner (subsecção 2.6.2) e Inductive Miner (subsecção 2.6.3). Ambos os algoritmos pertencem à perspetiva de controlo, conseguem lidar com registos de eventos complexos e os modelos de processo gerados podem ser convertidos para outras notações. Foram escolhidos estes dois algoritmos porque apresentam caraterísticas que se adequam à estrutura dos registos de eventos e por serem dos mais utilizados na área de mineração de processos.

O algoritmo Fuzzy Miner é considerado um dos algoritmos mais poderosos dado que apresenta melhorias face aos algoritmos anteriores. Porém nao é utilizado uma vez que o *output* final é um modelo *Fuzzy* e este não pode ser convertido para outros tipos de notações. A conversão é imprescindível dado que, na etapa de avaliação é necessário converter os modelos de processo para a notação Rede de *Petri*. Este algoritmo não se encontra disponível na ferramenta ProM, apenas nas ferramentas Disco e Celonis.

5.3 Análise e Preparação dos Dados

Numa primeira fase são analisados e recolhidos todos os dados de todos os pacientes com pelo menos um tipo de cancro diagnosticado. Os dados são mapeados de forma a obter um registo de eventos que dê resposta aos requisitos definidos. O *script* a utilizar dá origem a um

registo de eventos composto por 298111 registos, com tamanho total de 13.9 megabytes. A seleçao dos registos de eventos é realizada na própria ferramenta com o auxílio dos plugins.

Após a importação do ficheiro resultante na ferramenta ProM não é possível dar seguimento ao processo. Uma das caraterísticas da área mineração de processos é o grande poder computacional que exige na descoberta de conhecimento. Após várias tentativas falhadas na análise e na descoberta do processo conclui-se que o registo de eventos exige recursos computacionais que a máquina não possui. A ferramenta não consegue processar o registo de eventos e é necessário encontrar uma alternativa que dê continuidade ao processo e que responda aos requisitos inicialmente definidos.

A alternativa encontrada é dividir o registo de eventos para cada requisito e reduzir o nicho de pacientes diagnosticados com uma doença oncólogica (código 190-199). No total são criados quatro *scripts*, um para cada requisito. Cada script dá origem a um ficheiro com aproximadamente metade dos registos do ficheiro inicial. Esta alternativa favorece a diminuição do tempo de processamento bem como o tempo de análise. O script construído para responder ao requisito 1 encontra-se no Apêndice C. A estrutura dos *scripts* para os restantes requisitos é semelhante.

5.4 Mineração dos Processos

Esta secção apresenta de forma aprofundada todas as etapas desenvolvidas durante o processo de desenvolvimento. A secção está dividida em quatro subseções (5.4.1, 5.4.2, 5.4.3 e 5.4.4), uma para cada requisito. As subsecções descrevem o fluxo de trabalho realizado, enumerando as decisões tomadas e respetivos detalhes técnicos.

Plugins Utilizados

Durante o processo são utilizados vários plugins comuns entre os requisitos. A tabela 28 apresenta os plugins utilizados de acordo com a sua categoria, autor e uma breve descrição.

Tabela 28: Plugins utilizados - Processo

Plugin	Categoria	Autor	Descrição
Convert CSV to XES	Analítica	F. Mannhardt, N.Tax, D.M.M Schunselaar	Realiza a conversão de um ficheiro em formato .csv para um objeto em formato XES
Add Artificial Events	Analítica	J. Claes	Adiciona evento inicial ou final caso seja necessário
Remove Duplicate Attribute Values	Filtragem	F. Mannhardt	X Remove registos duplicados em que os atributos não foram alterados desde o último evento
Remove all attributes with value 'NULL'	Filtragem	F. Mannhardt	Remove registos que possuem atributos com valores nulos
Filter Log using Simple Heuristics	Filtragem	H.M.W. Verbeek	Filtra o registo de eventos de acordo com atributos especificados pelo utilizador
BPMN Miner	Algorítmica	R. Conforti M. Dumas, L. Garcia-Banuelos, M. La Rosa	Disponibiliza uma lista de algoritmos (Heuristics Miner, Alpha Miner, Inductive Miner, ILP Miner) e aplica o selecionado de acordo com uma lista de parâmetros. Gera um modelo de processo em notação BPMN.
Convert BPMN to PetriNet	Analítica	R. Conforti	Converte modelo de processo de notação BPMN para rede de *Petri*
PetriNet Analysis	Analítica	F. Mannhardt	Disponibiliza uma lista de parâmetros e apresenta detalhadamente uma análise ao modelo de processo

5.4.1 Requisito 1: Qual é o percurso mais comum com taxa de insucesso?

Esta subsecção descreve o processo desenvolvido para o requisito 1, apresentando o trabalho de mapeamento (subsecção 5.4.1.1), procedimentos na ferramenta (subsecção 5.4.1.2), configurações de plugins e resultados obtidos (subsecção 5.4.1.3).

5.4.1.1 Mapeamento

Para o **requisito 1 - Qual é o percurso mais comum com taxa de insucesso?**, são selecionados os pacientes que foram diagnosticados com o tipo de cancro "Malignant neoplasm of other and unspecified sites", código 190-199. Estes registos encontram-se na tabela

diagnoses_icd, atributo *'icd9_code'*. Também são selecionados os pacientes do género masculino e feminino sem restrição de idade.

O registo de eventos a importar para a ferramenta ProM possui uma estrutura específica com 4 campos: 'subject_id', 'hadm_id', 'activity' e 'charttime'. De forma a construir este registo de eventos é necessário o mapeamento dos dados das tabelas **admissions**, **services**, **icustays** e **diagnoses_icd**.

Na tabela **admissions** são usados os campos: *'subject_id'*, *'hadm_id'*, *'admittime'*, *'deathtime'*, *'edregtime'* e *'edouttime'*. Os tipos de atividades registados são: 'Admission', 'Death', 'Emergency Department Registration', 'Emergency Department Exit'. Para cada atividade é associado a respetiva referência temporal. Estas atividades e referências temporais ficam associadas aos novos campos *'activity'* e *'charttime'* respetivamente.

Na tabela **services** são usados os campos: *'subject_id'*, *'hadm_id'*, *'transfertime'* e *'curr_service'*. No total registam-se 11 tipos de serviços: Cardiac Medical (CMED), Cardiac Surgery (CSURG), Medical (MED), Neurologic Medical (NMED), Genitourinary (GU), Gynecological (GYN), Neurologic Surgical (NSURG), Surgical (SURG), Thoracic Surgical (TSURG), Orthopaedic (ORTHO) e Vascular Surgical (VSURG). Estes serviços representam tipos de cirurgias e tipos de consultas médicas que foram realizadas ao paciente. Os serviços são mapeados como atividades. Para cada atividade é associado a respetiva referência temporal. Estas atividades e referências temporais ficam associadas aos novos campos *'activity'* e *'charttime'* respetivamente.

Na tabela **icustays** são usados os campos: *'subject_id'*, *'hadm_id'*, *'ontime'* e *'first_careunit'*. O campo *"first_careunit'* representa a primeira unidade de cuidados intensivos no qual o paciente esteve internado. As unidades de cuidados intensivos são mapeadas como atividades e a sua referência temporal está presente no campo *"ontime'*.

Na tabela **diagnoses_icd** são usados os campos: *'hadm_id'* e *'icd9_code'*. São selecionadas as admissões de pacientes que possuem o tipo de cancro "Malignant neoplasm of other and unspecified sites"com o código(icd9_code) compreendido entre 190 e 199.

O *script* final contendo todas as *query's* realizadas encontra-se no Apêndice C.

A figura 28 ilustra um excerto do ficheiro final com os dados estruturados em formato .csv, preparado para importar para a ferramenta ProM. O ficheiro contém 18794 registos em que a primeira coluna representa o identificador do paciente, a segunda coluna representa o identificador de admissão no hospital, terceira coluna representa o tipo de atividade e a quarta coluna representa a referência temporal (data a hora) da respetiva atividade.

171	28666,151351,Death,2009-11-23 13:15:00		
172	22312,177653,Care Unit MICU,2004-12-06 22:35:41		
173	6489,168342,Care Unit TSICU,2003-04-11 19:10:53		
174	5008,126625,Discharge,2002-04-24 14:45:00		
175	16913,120594,Admission,2008-05-09 21:55:00		
176	52077,163736,CMED,2007-09-17 17:52:41		
177	27430,149122,Emergency Department Registration,2003-04-08 10:35:00		

Figura 28: Excerto do ficheiro de dados - Requisito 1

5.4.1.2 Ferramenta

Após a importação do ficheiro na ferramenta ProM, é necessário utilizar o plugin *CSV to XES*, de modo a converter o ficheiro de dados para o formato XES.

De seguida procede-se a fase de filtragem. São eliminados os registos duplicados usando o plugin *Remove duplicate attribute values*. Também é usado o plugin *Remove all attributes with value 'NULL'* para remover os atributos com valores a *NULL* presentes nos registos. Por último aplica-se o plugin *Filter Log using Simple Heuristics* para filtrar os eventos de entrada e os de saída. Neste caso os eventos de entrada são 2: 'Admission ' e 'Emergency Department Registration' e o evento de saída é 'Death'. O uso destes 3 plugins da categoria de filtro diminuem significativamente o consumo de memória.

Após a fase de filtragem é apresentado um *dashboard* que fornece a informação quantitativa do número de processos, número de casos e número de eventos. Graficamente apresenta o valor mínimo, valor máximo e a média da quantidade de eventos por caso. A figura 29 apresenta um *screenshot* do *dashboard* de informação.

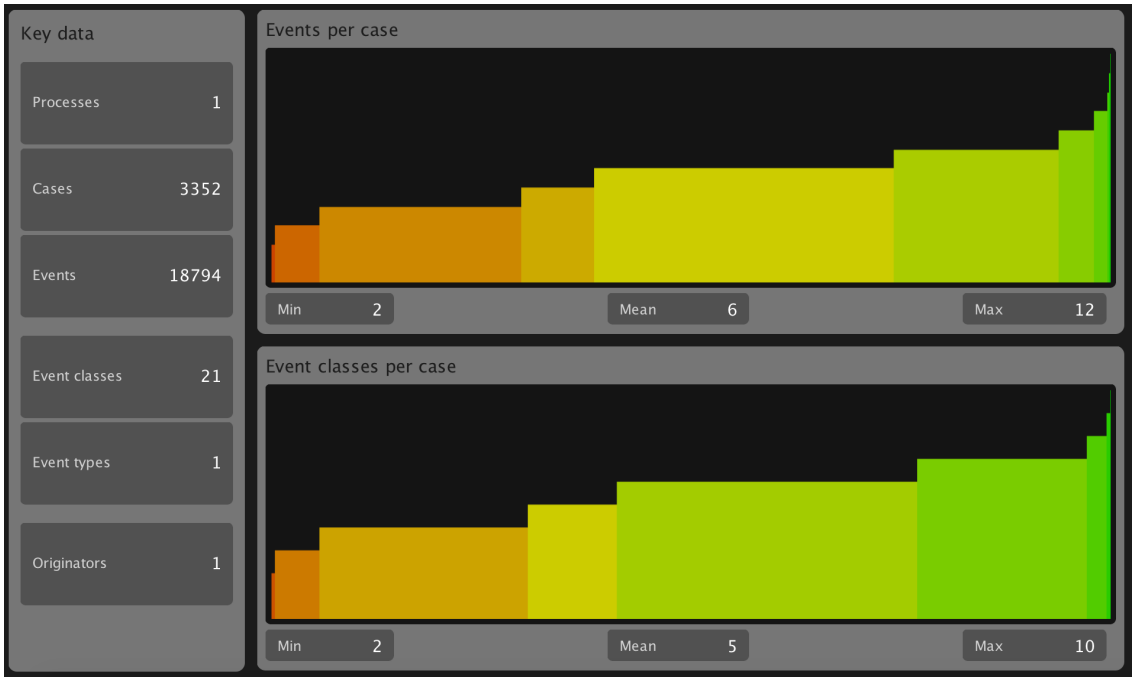


Figura 29: Dashboard - Requisito 1

A tabela 29 apresenta o número de ocorrências por atividade e respetiva taxa de ocorrência relativa. A figura 30 ilustra os cinco percursos com maior número de ocorrências de acordo com o requisito 1. O percurso mais seguido, registo de 56 ocorrências, é composto pelas atividades: (1) 'Emergency Department Registration', (2) 'Admission', (3) 'Care Unit MICU', (4) 'MED', (5) 'Emergency Department Exit', (6) 'Discharge' e (7) 'Death'.

Tabela 29: Sumário Ocorrências - Requisito 1

Atividade	**Ocorrências (absoluto)**	**Ocorrências (relativo)**
Discharge	3352	17.385%
Admission	3351	17.83%
Emergency Department Registration	1976	10.514%
Emergency Department Exit	1976	10.514%
MED	1915	10.189%
Care Unit MICU	1737	9.242%
Care Unit SICU	949	5.049%
NSURG	688	3.661%
Death	682	3.629%
SURG	467	2.485%
Care Unit TSICU	391	2.08%
NMED	261	1.389%
Care Unit CCU	229	1.218%
TSURG	209	1.112%
CMED	179	0.952%
Care Unit CSRU	133	0.708%
GYN	95	0.505%
GU	72	0.383%
ORTHO	66	0.351%
CSURG	54	0.287%
VSURG	12	0.064%
TOTAL	**18794**	**100%**

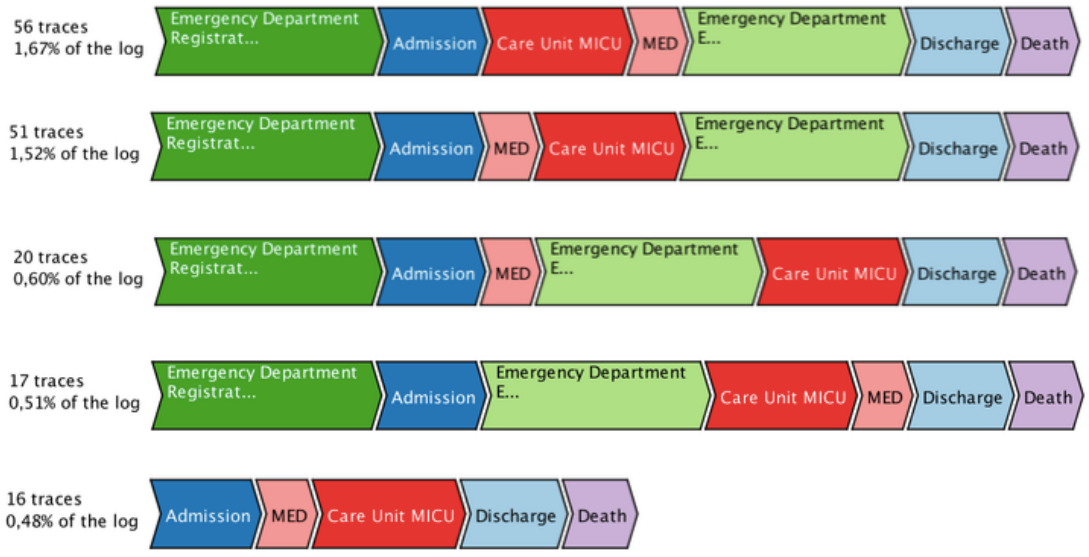


Figura 30: Top 5 percursos com maior número de ocorrências - Requisito 1

5.4.1.3 Configurações e Resultados

A próxima fase é a aplicação dos algoritmos *Inductive Miner* e *Heuristics Miner*. É utilizado o plugin *BPMN Miner* que disponibiliza vários algoritmos de descoberta do processo.

Após a escolha do algoritmo, o plugin apresenta um *pop-up* com a possibilidade de ajustar parâmetros de alguns atributos (multi-instância, 'noise' e 'timer event'). A figura 31 apresenta o *screenshot* do pop-up para o algoritmo Heuristics Miner. Os valores utilizados são os pré-definidos pela ferramenta. Também é possível selecionar as atividades de entrada e de saída. Para ambos os algoritmos não são especificadas atividades de entrada nem de saída.

Após a parametrização é gerado o modelo de processo em notação BPMN.

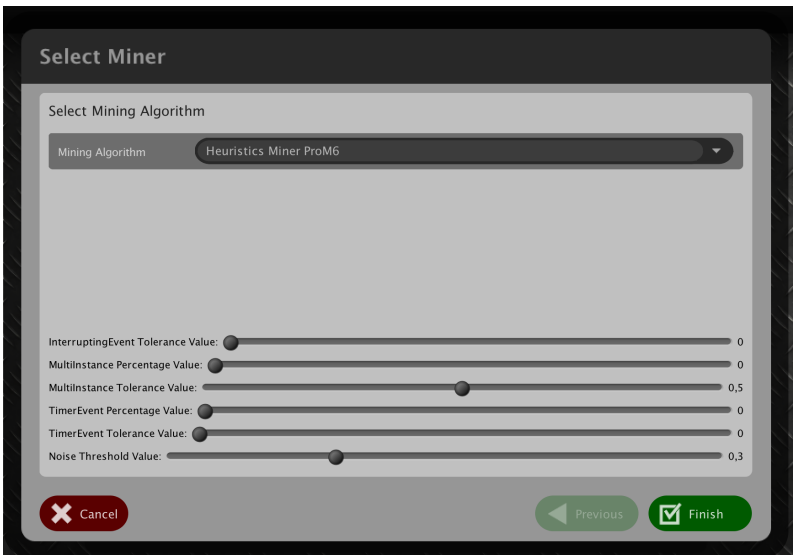


Figura 31: Configuração de parâmetros Plugin 'BPMN Miner'

Algoritmo Heuristics Miner

O modelo de processo resultante por este algoritmo é apresentando pela figura 6.2.1. É possível observar que o modelo de processo gerado pelo algoritmo *Heuristics Miner* é completo, envolve todas as atividades presentes no registo de eventos e apresenta diversas sequências de fluxo. A atividade 'Discharge' e 'Death' são as atividades finais, prosseguindo pelo término do processo.

Algoritmo Inductive Miner

O modelo de processo resultante por este algoritmo é apresentando pela figura 6.2.1.

Já o modelo de processo gerado pelo algoritmo *Inductive Miner* apresenta o percurso em que a atividade final é necessariamente 'Death' ou 'Discharge'. O registo de entrada no hospital é realizado por duas vias, pelo departamento de emergência (Emergency Department Registration) ou por admissão (Admission). Caso seja realizada a admissão, o paciente pode:

(i) Prosseguir para a unidade de cuidados intensivos SICU;

(ii) Realizar cirurgia (VSURG, TSURG, CSURG, SURG, NSURG);

(iii) Realizar consulta (NMED, ORTHO, GU, GYN);

(iv) Prosseguir para outra unidade de cuidados intensivos(MICU, TSICU).

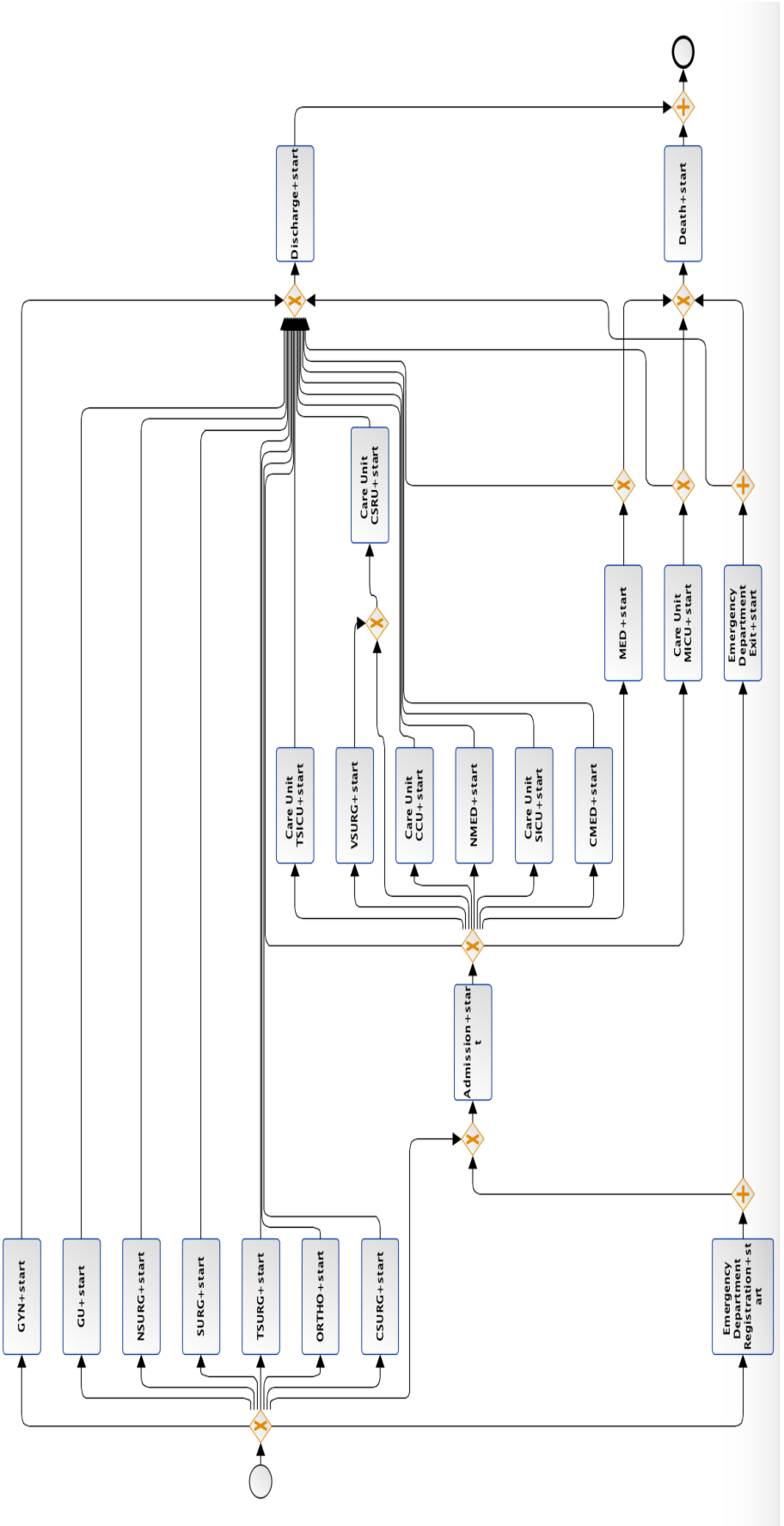


Tabela 30: Modelo de Processo Requisito 1 - Algoritmo Heuristics Miner

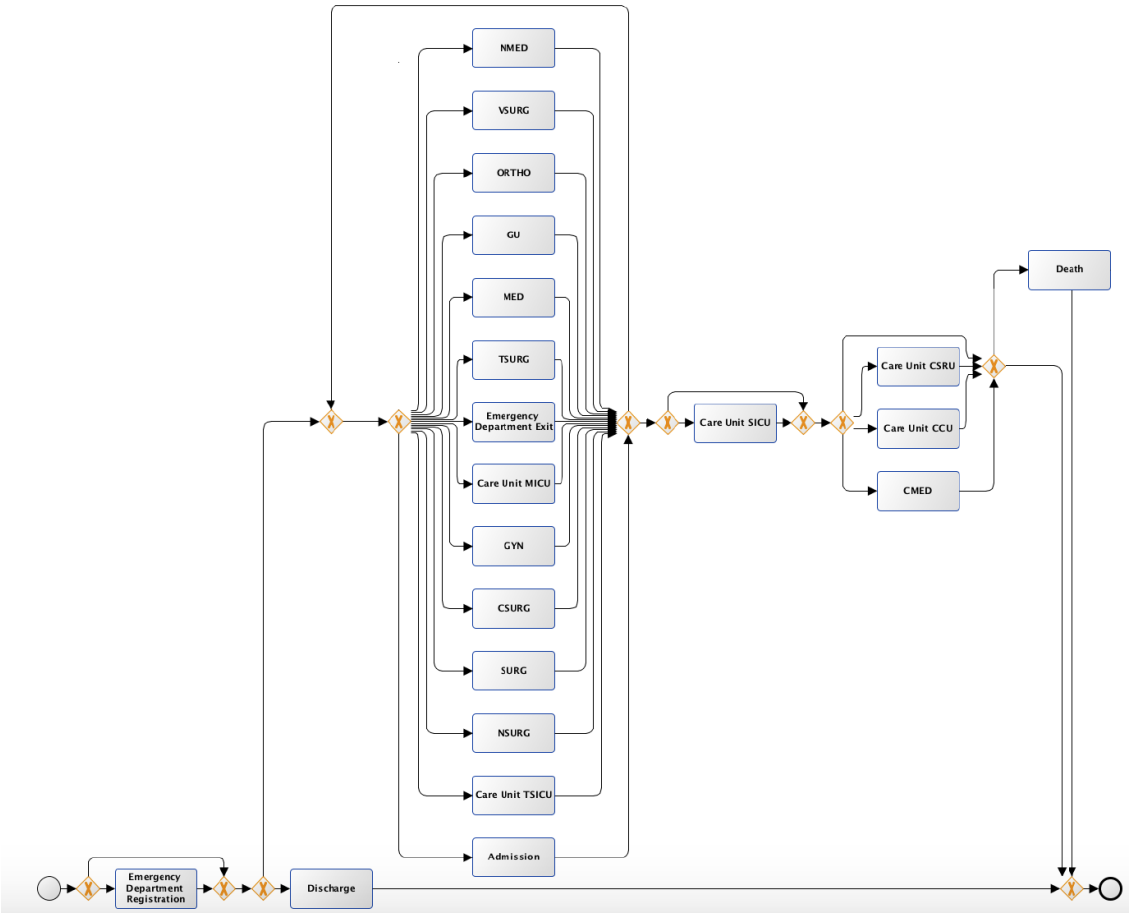


Figura 32: Modelo de Processo Requisito 1 - Algoritmo Inductive Miner

5.4.2 Requisito 2: Qual é o percurso mais comum com taxa de sucesso?

Esta subsecção descreve o processo desenvolvido para o requisito 2, apresentando o trabalho de mapeamento (subsecção 5.4.2.1), procedimentos na ferramenta (subsecção 5.4.2.2), configurações de plugins e resultados obtidos (subsecção 5.4.2.3).

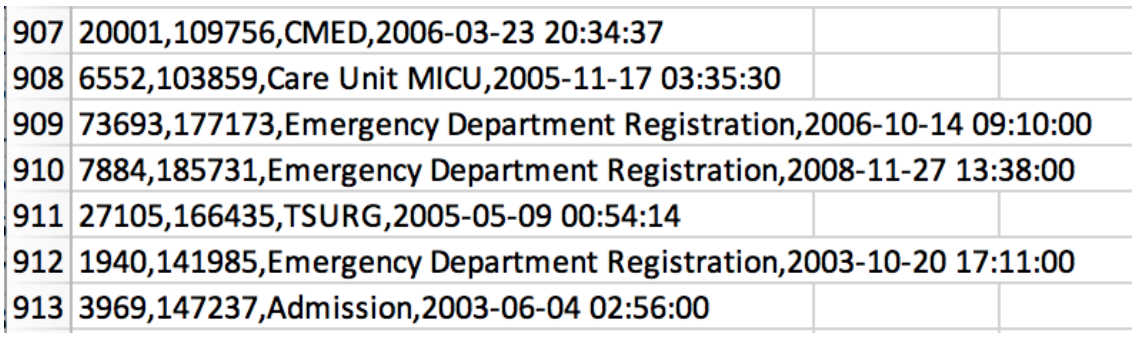
5.4.2.1 Mapeamento

Assume-se que um paciente obteve sucesso quando não existe referência temporal do atributo 'deathtime' da tabela **admissions**, por outras palavras, o valor do atributo 'deathtime' é nulo ou vazio.

Para o **requisito 2 - Qual é o percurso mais comum com taxa de sucesso?**, são selecionados os pacientes que foram diagnosticados com o tipo de cancro "Malignant neoplasm of other and unspecified sites", código 190-199. Estes registos encontram-se na tabela **diagnoses_icd**, atributo *'icd9_code'*. Também são selecionados os pacientes do género masculino e feminino sem restrição de idade.

O processo para este requisito é muito similar ao requisito 1. A alteração a realizar é apenas à condição do valor do atributo 'deathtime'. Para este caso pretende-se que o valor seja nulo ou vazio.

A figura 33 ilustra um excerto do ficheiro final com os dados estruturados em formato .csv, preparado para importar para a ferramenta ProM. O ficheiro contém 18112 registos em que a primeira coluna representa o identificador do paciente, a segunda coluna representa o identificador de admissão no hospital, terceira coluna representa o tipo de atividade e a quarta coluna representa a referência temporal (data a hora) da respetiva atividade.



907	20001,109756,CMED,2006-03-23 20:34:37
908	6552,103859,Care Unit MICU,2005-11-17 03:35:30
909	73693,177173,Emergency Department Registration,2006-10-14 09:10:00
910	7884,185731,Emergency Department Registration,2008-11-27 13:38:00
911	27105,166435,TSURG,2005-05-09 00:54:14
912	1940,141985,Emergency Department Registration,2003-10-20 17:11:00
913	3969,147237,Admission,2003-06-04 02:56:00

Figura 33: Excerto do ficheiro de dados - Requisito 2

5.4.2.2 Ferramenta

Após importação do ficheiro, procede-se a fase de filtragem. São eliminados os registos duplicados usando o plugin *Remove duplicate attribute values*. Este plugin remove os valores de atributos duplicados dos eventos, criando um novo registo de eventos mais pequeno. Também é usado o plugin *Remove all attributes with value 'NULL'* para remover os atributos com valores a *NULL* presentes nos registos. Por último aplica-se o plugin *Filter Log using Simple Heuristics* para filtrar os eventos de entrada e os de saída. Neste caso os eventos de entrada são 2: 'Admission ' e 'Emergency Department Registration' e o evento de saída é 'Discharge'. O uso destes 3 plugins da categoria de filtro diminuem significativamente o consumo de memória.

Após a fase de filtragem é apresentado um *dashboard* que fornece a informação quantitativa do número de processos, número de casos e número de eventos. Graficamente apresenta o valor mínimo, valor máximo e a média da quantidade de eventos por caso. A figura 34 apresenta um *screenshot* ao *dashboard* de informação.

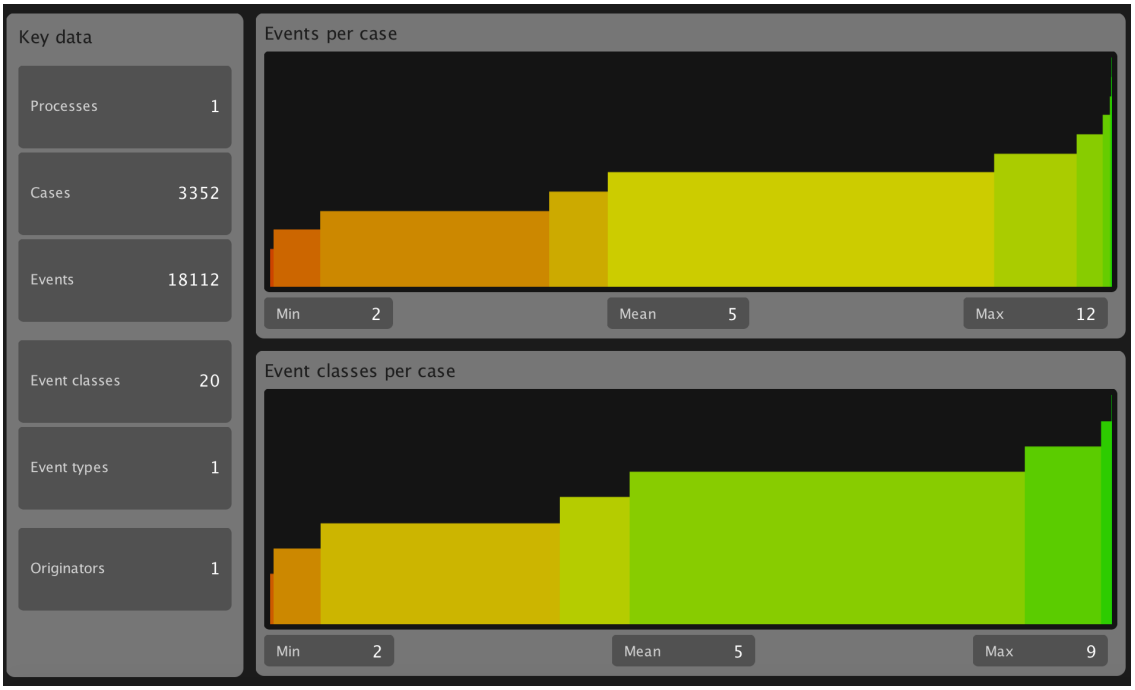


Figura 34: Dashboard - Requisito 2

A figura 35 ilustra os cinco percursos com maior número de ocorrências de acordo com o requisito 2. O percurso mais seguido, registo de 56 ocorrências, é composto pelas atividades: (1) 'Emergency Department Registration', (2) 'Admission', (3) 'Care Unit SICU, (4) 'NSURG', (5) 'Emergency Department Exit' e (6) 'Discharge'.

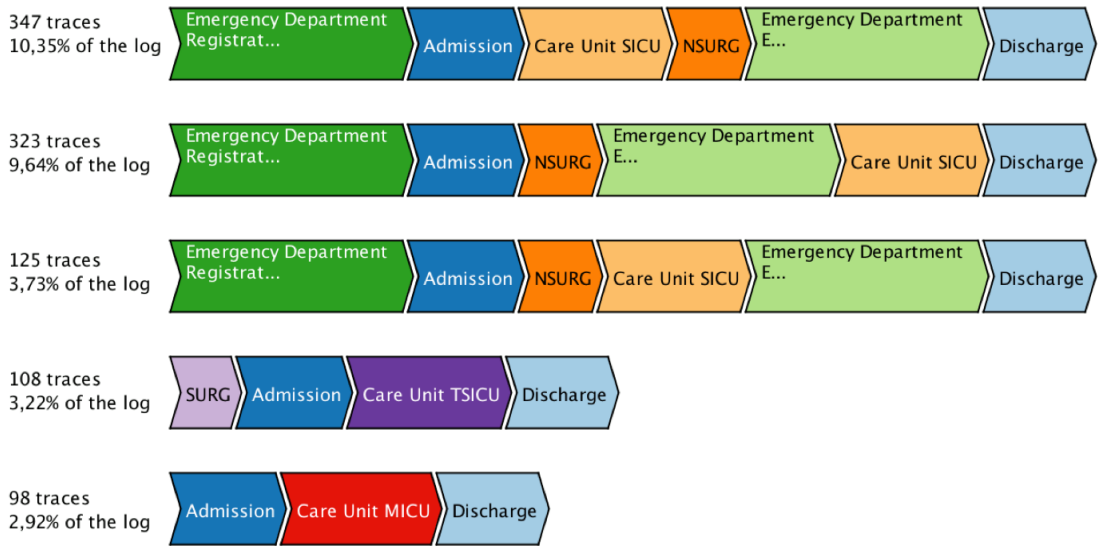


Figura 35: Top 5 percursos com maior número de ocorrências - Requisito 2

5.4.2.3 Configurações e Resultados

A próxima fase é a aplicação dos algoritmos *Inductive Miner* e *Heuristics Miner*. É utilizado o plugin *BPMN Miner*.

Após a escolha do algoritmo, o plugin apresenta um pop-up com a possibilidade de ajustar parâmetros de alguns atributos (multi-instância, 'noise' e 'timer event'). O procedimento foi semelhante e apresenta-se descrito na subsecção acima. Relativamente às atividades de entrada e de saída, é selecionada a atividade 'Discharge' para saída. Em ambos não há restrições para as atividades de entrada.

Algoritmo Heuristics Miner

A 6.2.2 apresenta o modelo de processo gerado pelo algoritmo *Heuristics Miner*.

É possível observar que o modelo de processo gerado pelo algoritmo *Heuristics Miner* é idêntico ao modelo de processo do requisito 1. Tal como no modelo de processo do requisito 1, o registo do paciente no hospital pode ser realizado pelo departamento de emergência ou por admissão. Numa primeira fase o paciente pode ser submetido a uma das seguintes cirurgias ou consultas:

(i) Consulta do tipo ORTHO (Ortopedia);

(ii) Consulta do tipo GU (Urogenital);

(iii) Cirurgia do tipo TSURG (Cardiotorácica);

(iv) Cirurgia do tipo SURG (Geral);

(v) Cirurgia do tipo SURG (Geral);

(vi) Cirurgia do tipo CSURG (Cardiovascular);

(vii) Consulta do tipo GYN (Ginecologia).

Algoritmo Inductive Miner

A figura 6.2.2 apresenta o modelo de processo gerado pelo algoritmo *Inductive Miner*.

O registo de entrada no hospital é realizado por duas vias, pelo departamento de emergência (Emergency Department Registration) ou por admissão ('Admission'). Caso o registo seja feito pelo departamento de emergência médica, o paciente pode dar imediatamente saída do hospital (fim do fluxo) ou ir a uma consulta neurológica ('NMED').

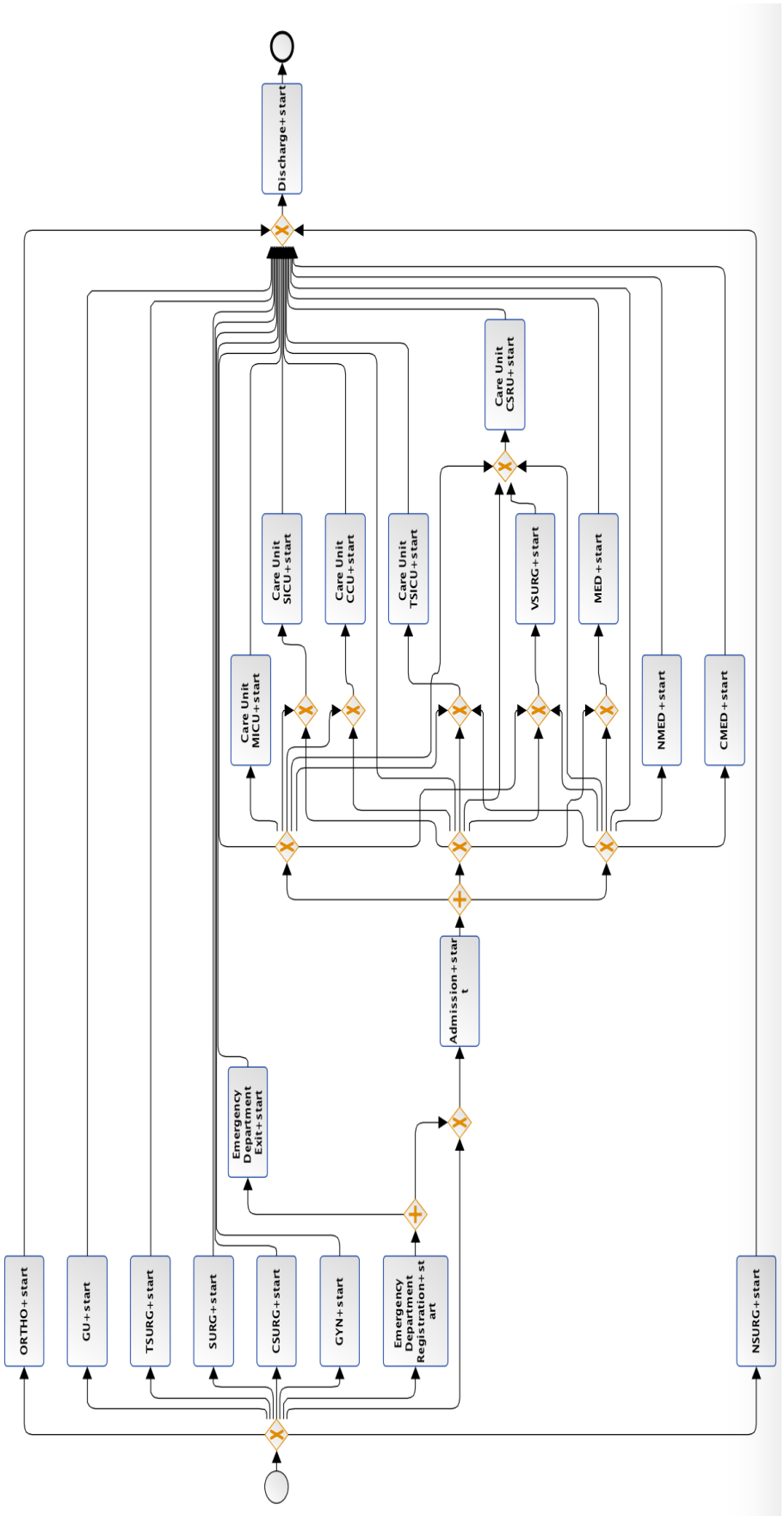


Tabela 31: Modelo de Processo Requisito 2 - Algoritmo Heuristics Miner

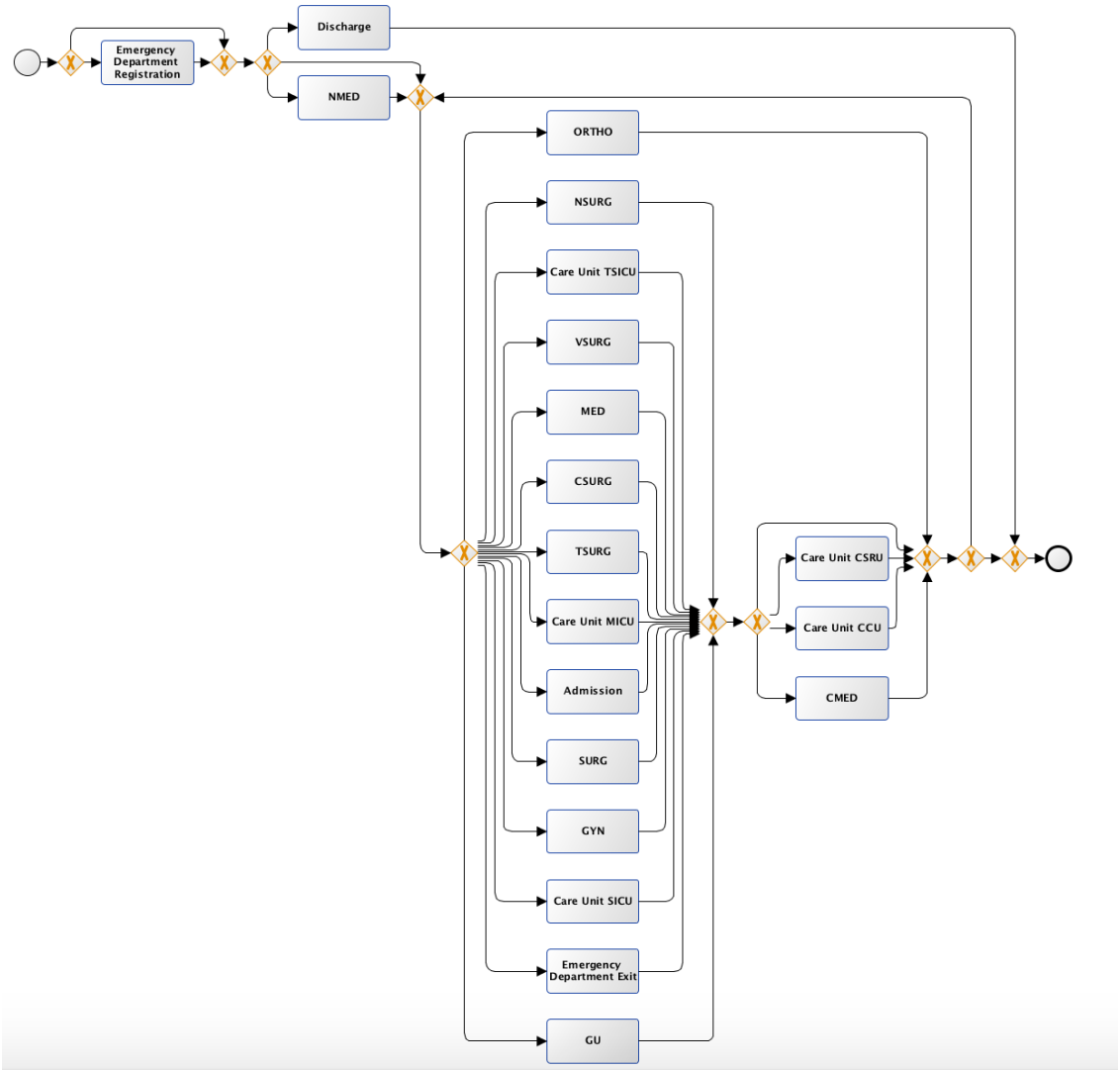


Figura 36: Modelo de Processo Requisito 2 - Algoritmo Inductive Miner

5.4.3 Requisito 3: Qual é o percurso mais curto (número de atividades) com taxa de sucesso?

A partir do processo desenvolvido na subsecção 5.4.2 consegue-se extrair o percurso mais curto com taxa de sucesso. Assume-se que o percurso mais curto é o que contém o menor número de atividades até atingir a atividade final 'Discharge'.

A figura 37 ilustra o percurso mais curto com 98 ocorrências registadas (2.92 % do registo de eventos). O paciente dá entrada no hospital pelo departamento de emergência médica (Emergency Department Registration), de seguida realiza a admissão (Admission) e fica em permanência na unidade de cuidados intensivos SICU. Esta unidade fornece cuidados intensivos e tratamento especializado aos pacientes, monitorizando de perto os pacientes cujas condições são instáveis (pressão arterial, função renal, ritmo cardíaco, funcionamento dos orgãos de suporte, entre outros). Após saída da unidade SICU o paciente dá saída do hospital (Discharge).

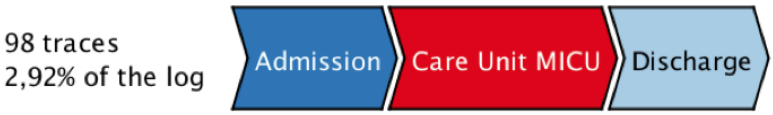


Figura 37: Percurso mais curto Requisito 2

A partir do percurso mais curto ilustrado na figura 37 é possível obter as ocorrências com maior nível de detalhe. A figura 38 ilustra cinco exemplos de ocorrências de diferentes pacientes e as respetivas referências temporais de cada atividade.

A primeira ocorrência é do paciente com o id 10397. A paciente é do sexo feminino, realiza a admissão no hospital no dia 12 de setembro de 2004 às 17 horas e 41 minutos. Após 5 dias dá entrada na unidade de cuidados intensivos MICU e permanece durante 8 dias até sair do hospital. A data de saída é no dia 25 de setembro de 2009 às 17 horas e 20 minutos. No total o paciente permanece 13 dias no hospital.

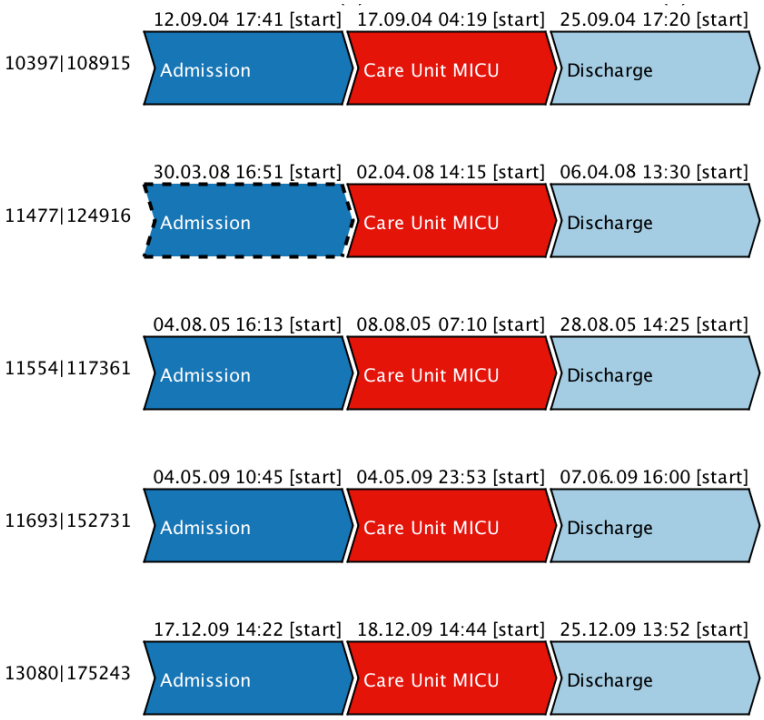


Figura 38: Ocorrências - percurso mais curto (número de atividades)

5.4.4 Requisito 4: Qual é o percurso mais curto (tempo) com taxa de sucesso?

Assume-se que o percurso mais curto em tempo possua o menor tempo entre a atividade inicial e a atividade final (Discharge).

A partir do modelo de processo gerado em notação BPMN (6.2.2), aplica-se o algoritmo 'Convert BPMN to PetriNet'. Após obter o modelo de processo com a notação Rede de Petri, aplica-se o algoritmo 'Petri Net Analysis' e obtém-se o percurso com o menor tempo registado.

A figura 39 ilustra o percurso mais curto a nível temporal. O tempo de permanência no hospital é de aproximadamente 2 dias, 9 horas e 41 minutos.

O paciente dá entrada pelo departamento de emergência no dia 12 de setembro de 2009 às 05:40h e permanece durante 3 horas e 23 minutos. às 09:12h realiza a admissão no hospital e entra na unidade de cuidados intensivos SICU às 16:09. Após 32 minutos é submetido a uma cirurgia do tipo NSURG e sai do departamento de emergência às 19:00h do mesmo dia. O paciente sai do hospital no dia 14 de setembro de 2009 às 15:30h.

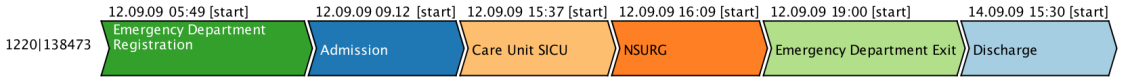


Figura 39: Ocorrência - percurso mais curto (tempo)

Capítulo 6

Avaliação

Este capítulo apresenta as métricas de avaliação utilizadas na área de mineração de processos (6.1), incluindo a análise e avaliação dos processos (secção 6.2) obtidos na secção 5.4. É ainda apresentando os desvios encontrados no processo (secção 6.3).

6.1 Dimensões de Avaliação

Determinar a qualidade do modelo de processo é caraterizada por várias dimensões. A verificação de conformidade, descrita na subsecção 2.3.2, fornece quatro dimensões de avaliação: *fitness*, precisão, simplicidade e generalização. Para cada dimensão existe uma métrica específica (Rozinat et al. 2008).

As métricas analisam e avaliam o modelo e o registo de eventos de forma a apresentarem diagnósticos e percentagens estatísticas. Estas são calculadas a partir da reprodução dos registos e analisam a estrutura tendo em conta o tamanho e formato:

(i) **Fitness**: Avalia o comportamento capturado pelo modelo de processo que está presente no registo de eventos inicial.

O modelo é reproduzido através da utilização de tokens. A partir do estado inicial os tokens são produzidos e consumidos de acordo com o registo de entrada. Após executar todos os registos, o total de tokens consumidos são contabilizados. A fórmula utilizada para o cálculo de *fitness* é a seguinte (Buijs, Van Dongen e W M P Van Der Aalst s.d.):

$$F = 1 - \frac{custo\ alinhamento\ entre\ modelo\ e\ registo\ de\ eventos}{custo\ minimo\ para\ alinhar\ o\ registo\ de\ eventos\ no\ modelo\ e\ vice-versa} \tag{6.1}$$

O valor obtido fica compreendido entre 0 e 1. Quanto mais próximo do valor 1, mais registos foram executados corretamente.

(ii) **Precisão**: Avalia o grau de precisão do comportamento capturado pelo modelo de processo.

Durante a reprodução do modelo, compara-se o espaçamento entre estados de execução e são contabilizados os ramos não visitados (*escaping edges*). Os ramos não visitados são decisões que são possíveis no modelo e não no registo de eventos. Se não existirem, a precisão é perfeita.

A fórmula utilizada para o cálculo da precisão é a seguinte (Buijs, Van Dongen e W M P Van Der Aalst s.d.):

$$P = 1 - \frac{\sum_x visitas\ marcadas\ +\ visitas\ *\ \frac{arestas\ -\ arestas\ visitadas}{vertices}}{total\ de\ visitas\ marcadas} \tag{6.2}$$

(iii) **Generalização**: Avalia a qualidade do comportamento não capturado. Procura generalizar sequências.

Durante a reprodução do modelo é considerada a frequência com que cada nó do modelo precisa de ser visitado. Se houver nós no modelo que são pouco visitados, considera-se que a generalização é baixa.

A fórmula utilizada para o cálculo da generalização é a seguinte (Buijs, Van Dongen e W M P Van Der Aalst s.d.):

$$G = 1 - \frac{\sum_x \text{vertices } (\sqrt{\text{execucoes}})^{-1}}{total\ \text{vertices}} \tag{6.3}$$

(iv) **Simplicidade**: Avalia a complexidade do modelo de processo. Compara o tamanho total do modelo com o número de atividades presente no registo de eventos.

A fórmula utilizada para o cálculo da simplicidade é a seguinte (Buijs, Van Dongen e W M P Van Der Aalst s.d.):

$$S = 1 - \frac{atividades\ duplicadas\ +\ atividade\ em\ falta}{vertices} \tag{6.4}$$

A figura 40 apresenta um exemplo de aplicação de métricas de avaliação num registo de eventos com 5 instâncias e 9 atividades (A, B, C, D, E, F, G, H e I). Exemplo adaptado de (Rozinat et al. 2008).

O modelo (b) é considerado um bom modelo dado que possui valores elevados em todas as métricas e consegue reproduzir todas as sequências possíveis. Já o modelo (c) é fraco na métrica *fitness* porque apenas consegue reproduzir uma sequência 'A-B-D-E-I' de cinco possíveis. O modelo (d) apresenta um modelo pobre em precisão dado que consegue executar sequências de atividades com qualquer ordem. Por fim, o modelo (e) é fraco na métrica generalização porque só consegue executar exatamente as cinco sequências presentes no registo de eventos.

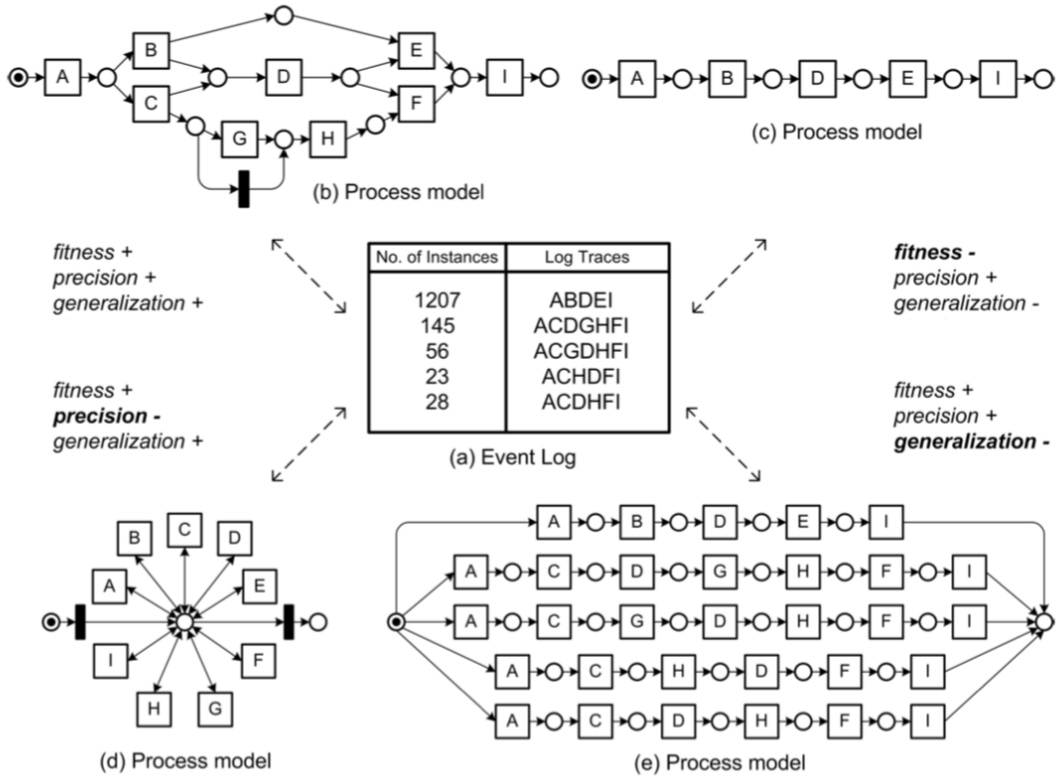


Figura 40: Exemplo de utilização de métricas de avaliação

6.2 Análise e Avaliação

Esta secção apresenta a análise e avaliação dos modelos de processo obtidos nas subseções 5.4.1 e 5.4.2.

São utilizados plugins disponíveis na ferramenta ProM para realizar a verificação de conformidade entre o modelo e o registo de eventos inicial.

Plugins Utilizados

Durante o processo de avaliação são utilizados 3 plugins. A tabela 32 apresenta os plugins utilizados de acordo com a sua categoria, autor e uma breve descrição.

Tabela 32: Plugins utilizados - Avaliação

Plugin	Categoria	Autor	Descrição
Convert BPMN to PetriNet	Analítica	R. Conforti	Converte modelo de processo de notação BPMN para rede de *Petri*
Replay a Log on Petri Net for Comformance Analysis	Verificação de Conformidade	A. Adriansyah	Realiza comparações entre o modelo de processo e o registo de eventos inicial
Check Conformance using ETConformance	Analítica	J. Munoz-Gama e J.Carmona	Calcula a percentagem de *fitness* e precisão

Numa primeira fase todos os modelos obtidos são convertidos para a notação rede de *Petri*, utilizando o plugin *Convert BPMN to PetriNet*. De seguida é realizada a verificação de conformidade entre o registo de eventos inicial e o modelo, utilizando o plugin *Replay a Log on Petri Net for Conformance Analysis*. Para utilizar este plugin é necessário dois ficheiros de entrada, o registo de eventos e o modelo. O ficheiro de saída é um modelo de processo na mesma notação.

A figura 41 apresenta um *screenshot* realizando na ferramenta ProM, no momento da escolha do plugin *Replay a Log on Petri Net for Conformance Analysis*.

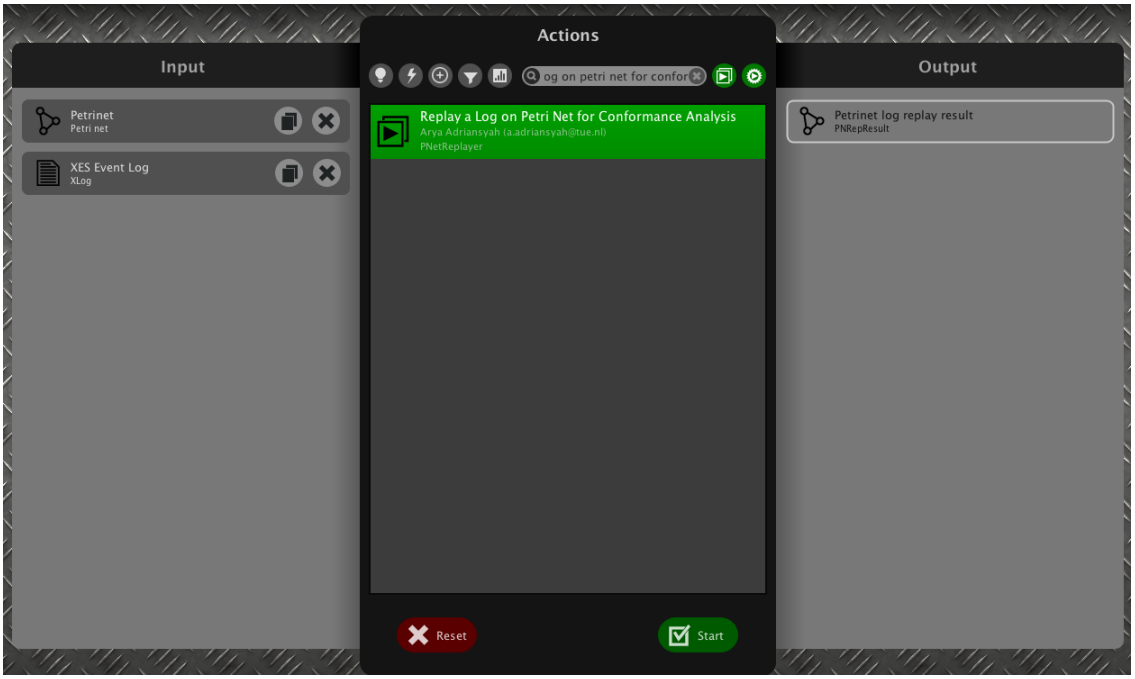


Figura 41: Screenshot Escolha do Plugin 'Replay a Log on Petri Net for Conformance Analysis'

6.2.1 Avaliação Modelos - Requisito 1

A figura 33 apresenta um excerto do modelo obtido pelo plugin de verificação de conformidade. Os ficheiros de entrada foram o modelo de processo obtido pelo algoritmo *Heuristics*

Miner (figura da subsecção 5.4.1) e o registo de eventos inicial no formato XES.

O modelo completo encontra-se disponível no Apêndice D.

Atividades com maior frequência são representadas com a cor azul mais escura. Estas atividades são: 'Emergency Departament Registration', 'Emergency Department Exit', 'Discharge' e 'Death'. As atividades com borda de cor vermelha e borda inferior de cor verde apresentam maior detalhe.

O número apresentado à esquerda representa o número total de ocorrências assíncronas e o número apresentado à direita representa as ocorrências descobertas pelo modelo. Por exemplo, a atividade 'Emergency Department Registration' foi registada 1974 vezes e 1378 foram descobertas pelo modelo. No total, existem 596 ocorrências que não se encontram no registo de eventos inicial.

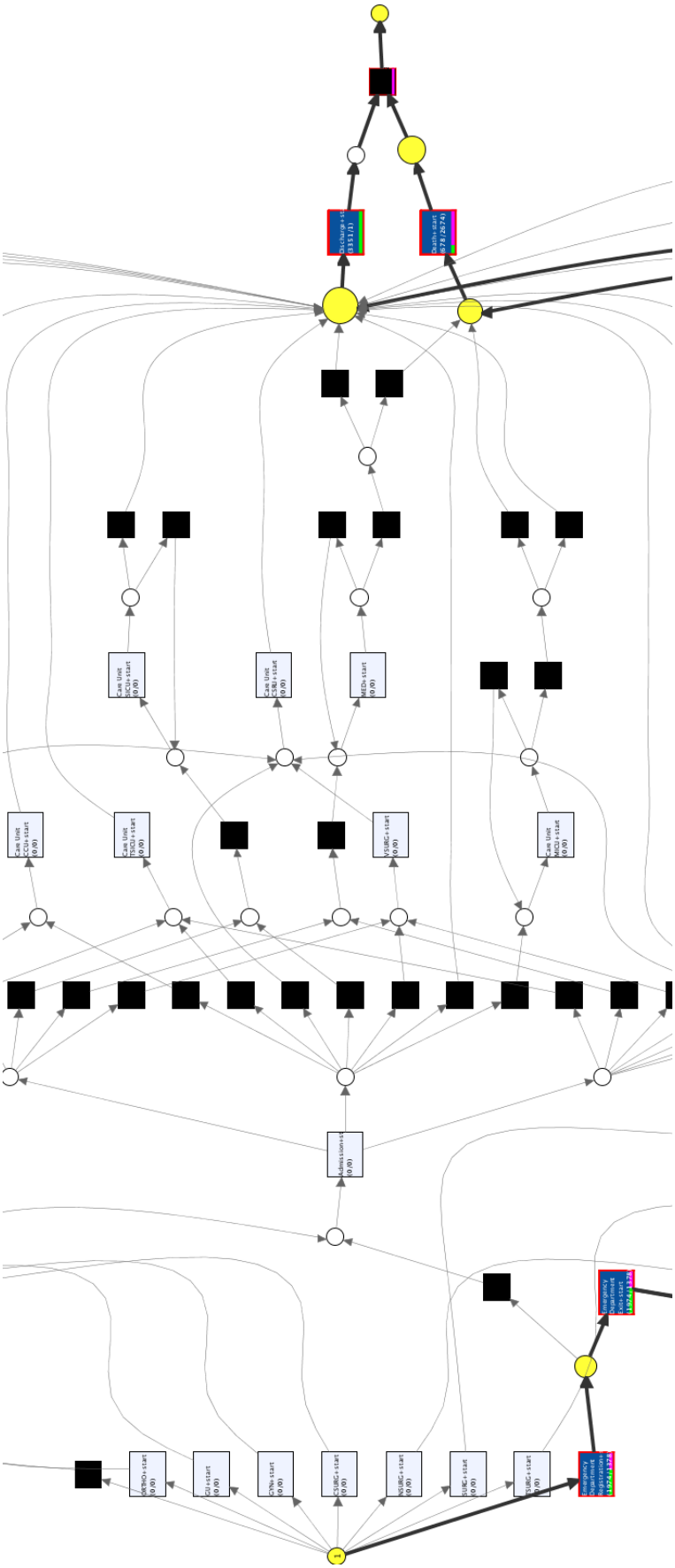


Tabela 33: Verificação de Conformidade - Modelo Ampliado Requisito 1 (Heuristics Miner)

Já a figura 34 apresenta o modelo obtido pelo plugin de verificação de conformidade. Os ficheiros de entrada foram o modelo de processo obtido pelo algoritmo *Inductive Miner* (figura da subsecção 5.4.1) e o registo de eventos inicial no formato XES.

As atividades possuem uma cor azul homogénea, o que significa que não existe elevado grau de descrepância entre as atividades.

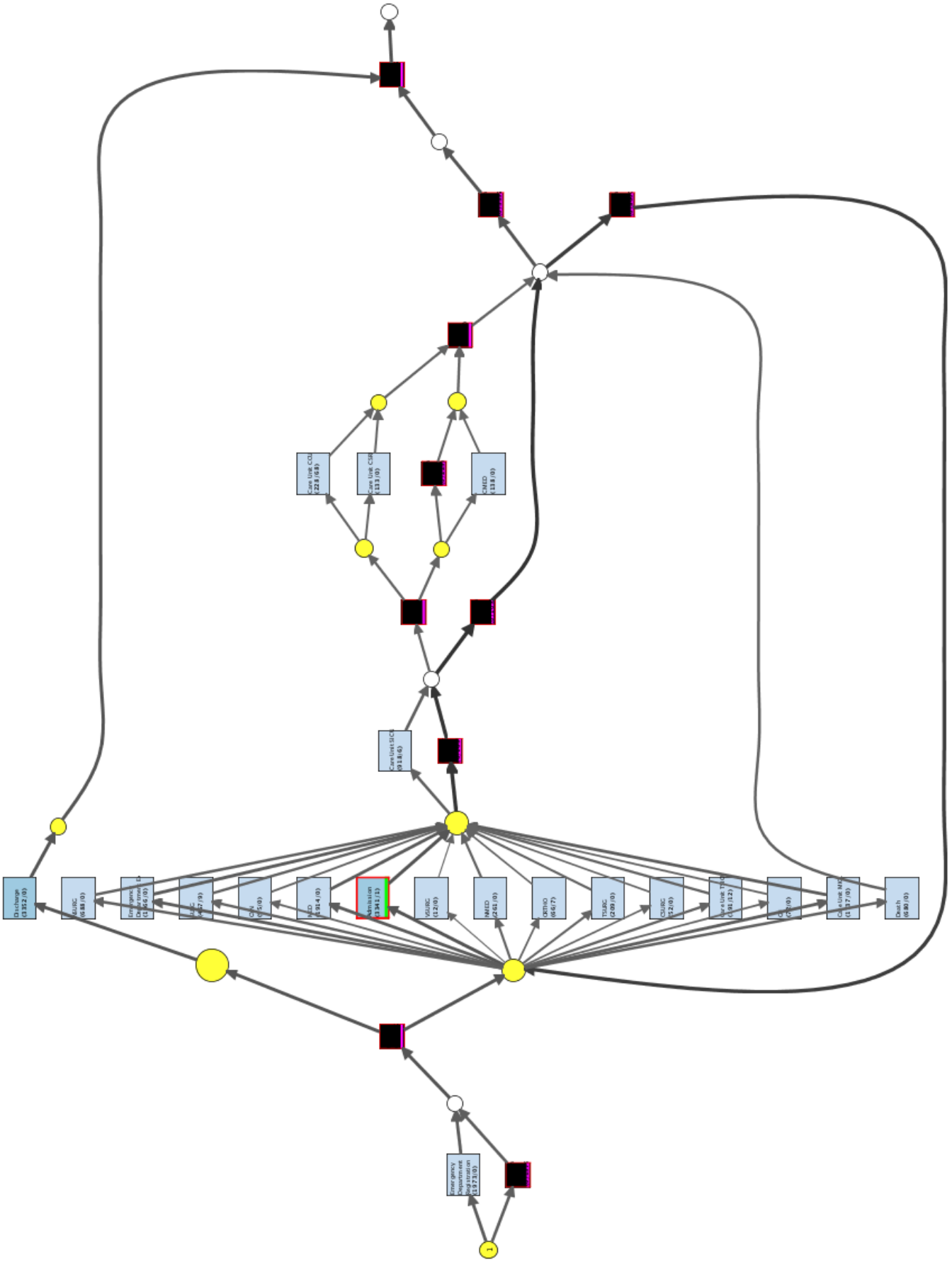


Tabela 34: Verificação de Conformidade - Modelo Requisito 1 (Inductive Miner)

Por último é utilizado o plugin *Check Conformance using ETConformance* que fornece a percentagem de *fitness* e precisão do modelo. Não é possível calcular a simplicidade e a generalização devido à falta de plugins incorporados na ferramenta ProM.

A tabela 35 apresenta os resultados obtidos pelos dois modelos. Comparando os resultados de ambos os modelos, verifica-se que a percentagem de todos as propriedades do modelo onde é aplicado o algoritmo Inductive Miner (1) é muito superior aos do modelo onde é aplicado o algoritmo Heuristics Miner (2). Esta diferença deve-se ao facto do modelo (1) ter descoberto mais conhecimento face ao modelo (2) e consequentemente ter tido um custo de alinhamento maior.

Tabela 35: Resultados Obtidos - Modelos Requisito 1

Propriedade	**Modelo Requisito 1 (Heuristics Miner)**	**Modelo Requisito 1 (Inductive Miner)**
Fitness: Modelo - Registo de Eventos	0.4059	0.9954
Fitness: Trace	0.4760	0.9964
Fitness: Registos de Eventos - Model	0.5949	0.9999
Precisão	0.6032	0.386

6.2.2 Avaliação Modelos - Requisito 2

A figura 36 apresenta um excerto do modelo obtido pelo plugin de verificação de conformidade. Os ficheiros de entrada são o modelo de processo obtido pelo algoritmo *Heuristics Miner* (figura da subsecção 5.4.2) e o registo de eventos inicial no formato XES.

O modelo completo encontra-se disponível no Apêndice E.

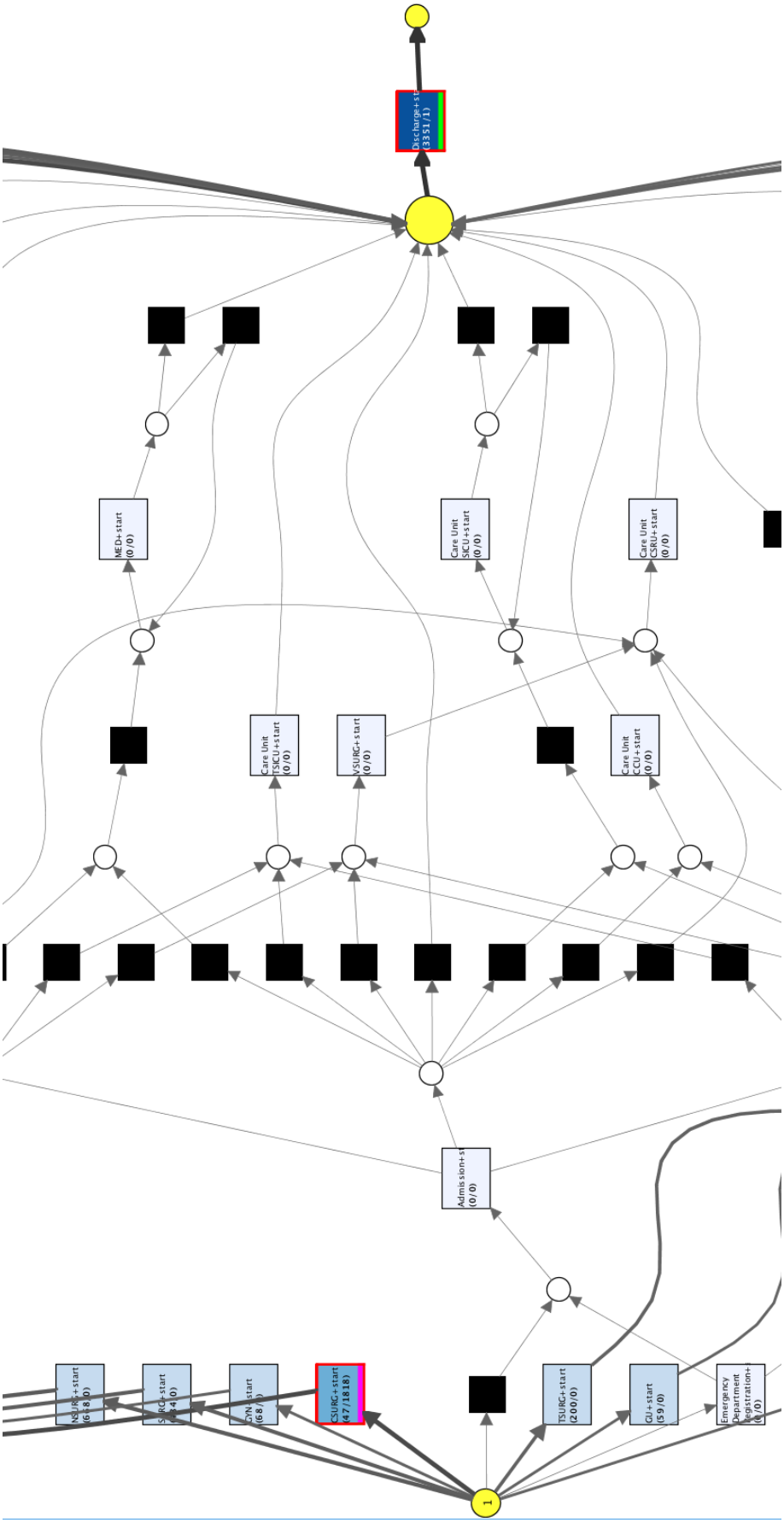


Tabela 36: Verificação de Conformidade - Modelo Ampliado Requisito 2 (Heuristics Miner)

Já a figura 37 apresenta o modelo obtido pelo plugin de verificação de conformidade. Os ficheiros de entrada são o modelo de processo obtido pelo algoritmo *Inductive Miner* (figura da subsecção 5.4.2) e o registo de eventos inicial no formato XES.

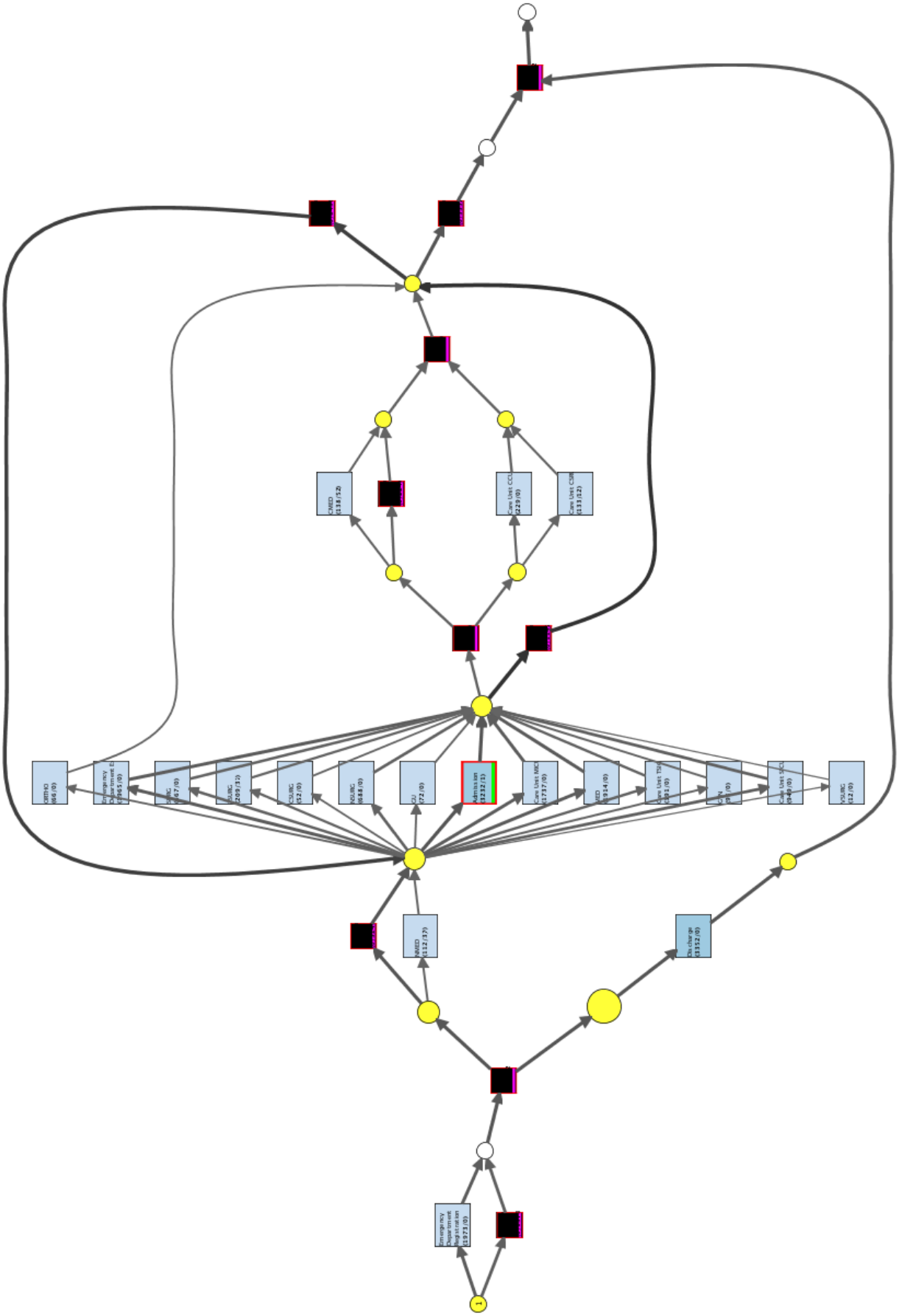


Tabela 37: Verificação de Conformidade - Modelo Requisito 2 (Inductive Miner)

Por último é novamente utilizado o plugin *Check Conformance using ETConformance*.

A tabela 38 apresenta os resultados obtidos pelos dois modelos. Os resultados comparativos entre os dois modelos são semelhantes face aos resultados obtidos entre os modelos do requisito 1.

Comparando os resultados de ambos os modelos, verifica-se que a percentagem de quase todas as propriedades do modelo onde é aplicado o algoritmo Inductive Miner (1) é muito superior aos do modelo onde é aplicado o algoritmo Heuristics Miner (2). Apenas a percentagem de precisão apresenta um valor maior no primeiro modelo. Esta diferença deve-se ao facto do modelo (1) ter descoberto mais conhecimento face ao modelo (2) e consequentemente ter tido um custo de alinhamento maior.

Tabela 38: Resultados Obtidos - Modelos Requisito 2

Propriedade	**Modelo Requisito 2 (Heuristics Miner)**	**Modelo Requisito 2 (Inductive Miner)**
Fitness: Modelo - Registo de Eventos	0.2934	0.9851
Fitness: Trace	0.4120	0.9886
Fitness: Registos de Eventos - Model	0.7286	0.9988
Precisão	0.7812	0.4521

6.3 Descoberta de Desvios

A análise e descoberta de desvios nos processos são parte fundamental da avaliação. Mostram com precisão os excertos do modelo que desviam do registo de eventos inicial (Leemans, Fahland e Wil M P Van Der Aalst 2014). Existem dois tipos de desvios:

(i) **Log Move**: Quando uma sequência contém um evento que não é permitido pelo modelo;

(ii) **Model Move**: Quando o modelo requer um evento que não está presente na sequência.

A figura apresenta um excerto de um modelo ilustrado com os dois tipos de desvios. A linha tracejada mais à esquerda representa o tipo **model move** e a linha mais à direita (circular) representa o tipo **log move** (Leemans, Fahland e Wil M P Van Der Aalst 2014).

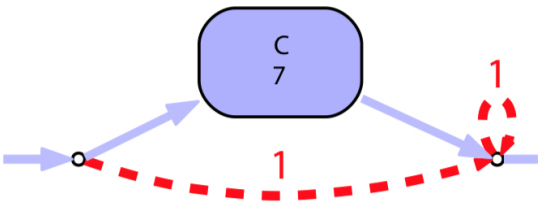


Figura 42: Tipos de Desvios

Foi utilizado o plugin *Inductive Visual Miner* disponibilizado pela ferramenta ProM. Este plugin aplica o algoritmo *Inductive Miner* e apresenta um modelo de processo interativo onde facilmente consegue-se observar o fluxo e os respetivos desvios, caso existam.

A figura 39 apresenta um *screenshot* realizado durante o fluxo do processo. Os pontos amarelos são dinâmicos e representam o comportamento de cada sequência no processo.

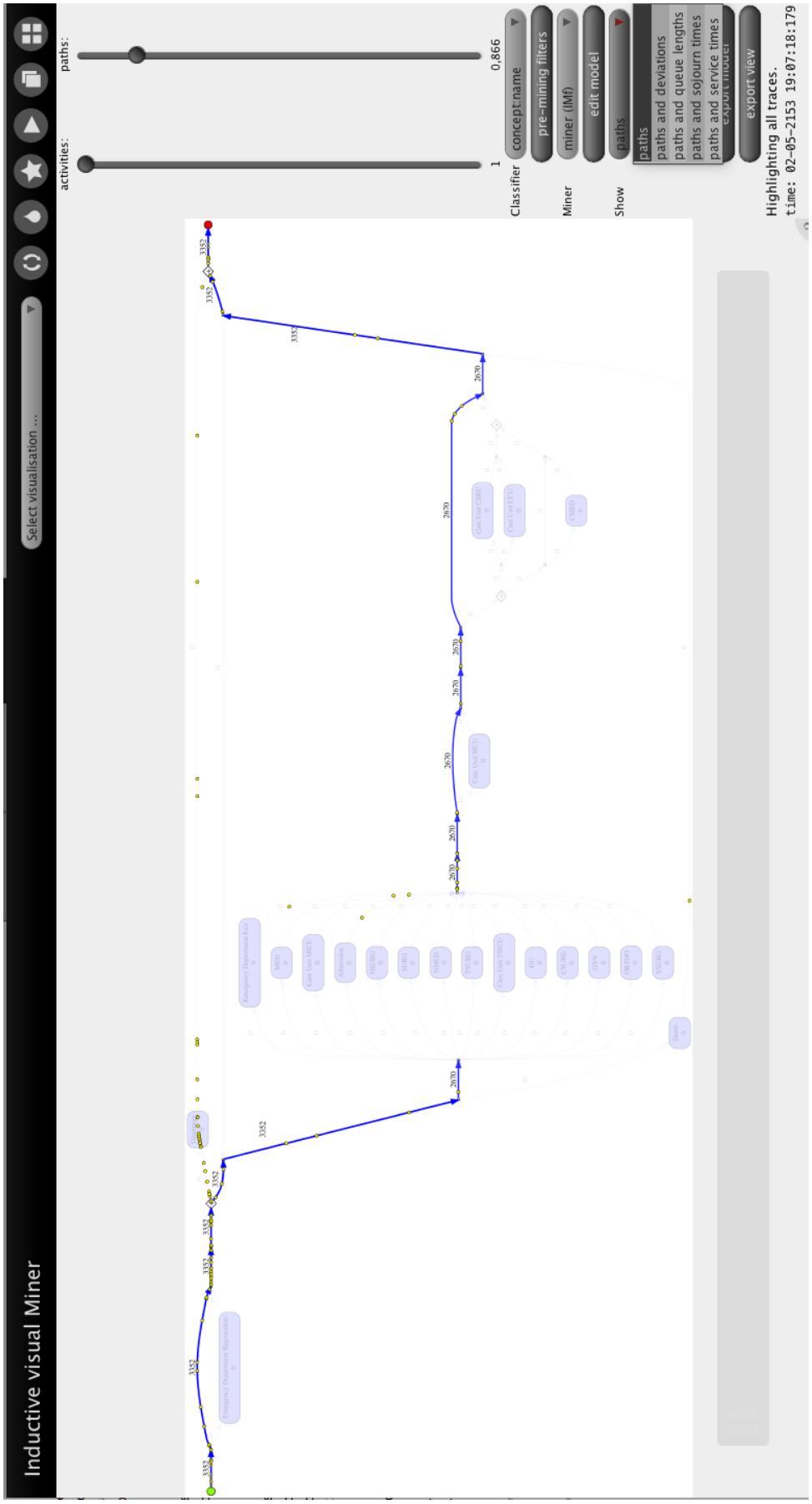


Tabela 39: Modelo Requisito 1 - Algoritmo Inductive Visual Miner

O plugin disponibiliza opções de filtro caso o utilizador pretenda visualizar apenas os desvios. A figura 40 apresenta um *screenshot* realizado durante o fluxo do processo com os desvios encontrados. Os ramos tracejados a vermelho representam a existência de desvios. O número em cima do ramo representa o número de vezes que esse ramo não foi executado.

O maior desvio encontrado foi entre a atividade 'Discharge' e o término do processo. No total foram 3352 as vezes que esse ramo não foi executado.

Tabela 40: Modelo Requisito 1 - Desvios Encontrados

Capítulo 7

Conclusões

Este capítulo resume o trabalho desenvolvido (secção 7.1) abordando alguns aspetos importantes de cada capítulo. Adicionalmente são descritos os objetos realizados (secção 7.2), bem como as limitações, dificuldades e trabalho futuro (secção 7.3). Por fim é apresentado os contributos (secção 7.4).

7.1 Resumo

A mineração de processos emergiu como uma nova área de investigação que se foca na extração de conhecimento a partir de registos de eventos. O seu objetivo é analisar, processar e produzir modelos de processo detalhados que ilustrem a realidade. Os sistemas de informação capturam um grande volume de dados das mais diversas fontes, que na sua maioridade não são analisados nem processados. Dada a evolução e integração dos sistemas de informação, os desafios cada vez são maiores. Os registos capturados possuem eventos de fina granularidade, heterogeneidade, volume e referências temporais incorretas.

Esta dissertação teve como objetivo desenvolver uma solução de software capaz de determinar processos a partir de dados de eventos, e entender em que medida os dados e o seu processamento são capazes de gerar artefactos de relevância.

No capítulo 1 foi identificado o contexto, definição do problema, objetivos a realizar e abordagem a seguir. Posteriomente, no capítulo 2 foi apresentado o estado de arte. Verificou-se que existem tipos e perspetivas que são aplicados em diferentes contextos com objetivos específicos, assim sendo, existem atualmente vários algoritmos e ferramentas que conseguem corresponder às necessidades e requisitos encontrados. Durante a análise das ferramentas possíveis destacaram-se três, especialmente pelo resultados positivos em trabalhos desta área. No capítulo 3 foi apresentado a análise de valor e com auxílio do método AHP, foi decidido a ferramenta a utilizar - ProM.

De seguida, no capítulo 4, foi abordado o caso de estudo, procurando por descrever detalhamente a base de dados a utilizar, sua estrutura e principais caraterísticas. Os requisitos levantados tomaram em consideração algumas questões que são muito questionadas na área oncológica e que ajudam a encontrar padrões ou desvios, mais especificamente, nos processos de tratamento de pacientes oncológicos.

Posteriomente no capítulo 5 foi apresentado a construção da solução. Durante o desenvolvimento de mineração de processos foram identificados problemas relacionados com a performance da ferramenta, quer a nível de processamento, quer a nível de memória. Para

que os problemas tivessem sido ultrapassados, foi necessário reestruturar o registo de eventos, dividindo-o em registos mais curtos e com apenas a informação necessária para cada um dos requisitos.

Por fim, no capítulo 6, os modelos obtidos foram avaliados tendo em conta as dimensões e métricas existentes. Os resultados obtidos foram satisfatórios, principalmente nos modelos onde foram aplicados o algoritmo *Inductive Miner*, tendo-se obtido valores de *fitness* muito próximos de 1. Nos modelos foram detectados alguns desvios, principalmente entre a atividade de *'Discharge'* e o término do processo. Para uma análise mais profunda dos processos médicos seria necessário a colaboração de um perito da área de modo a avaliar e identificar possíveis anomalias.

7.2 Objetivos Realizados

Na secção 1.3 foram definidos os requisitos gerais e na secção 4.2 os requisitos específicos. Ao longo deste caso de estudo, foi descrito todo o percurso que permitiu alcançar a maioria dos objetivos. A tabela 41 apresenta os requisitos definidos e a percentagem de conclusão respetiva.

Tabela 41: Objetivos Realizados e Percentagem de Conclusão

Objetivos Gerais	**Percentagem de Conclusão**
Compreender e analisar os processos inerentes a uma determinada área	100%
Desenvolver e aplicar técnicas de extração de dados, recorrendo a algoritmos de descoberta e de verificação	100%
Desenvolver modelos de processos perceptíveis e coerentes com os registos de eventos	100%
Avaliar os modelos de processo de forma a encontrar e analisar possíveis desvios e anomalias	20%
Objetivos Específicos	**Percentagem de Conclusão**
Percurso mais comum com taxa de insucesso	100%
Percurso mais comum com taxa de sucesso	100%
Percurso mais curto (número de atividades) com taxa de sucesso	100%
Percurso mais curto (tempo) com taxa de sucesso	100%

7.3 Limitações e Trabalho Futuro

Durante a fase de desenvolvimento e avaliação do caso de estudo, verificaram-se algumas limitações que deverão ser alvo de reparo em trabalho futuro.

O trabalho realizado não determinou o tempo de intervalo entre a saída do paciente do hospital e a sua morte, o que seria um tópico relevante de estudo. É importante realçar que foi inferido que a morte do paciente foi causada pelo tipo de cancro diagnosticado, tendo em conta o registo na base de dados **Social Security Death Index**. Esta base de dados

apenas contém os registos de óbitos, não tendo qualquer informação sobre o(s) motivo(s) do óbito.

Relativamente ao trabalho a realizar no futuro, é importante a melhoria contínua dos registos de eventos de forma a melhorar a qualidade e conformidade dos resultados. Apesar de apenas ter sido usada a perspetiva de controlo de fluxo, deve ser incluído o estudo para as restantes perspetivas: organizacional, caso e temporal. Os modelos são enriquecidos e conseguem envolver outros contextos (performance, interações e colaborações entre pessoas, entre outros).

Foram aplicados dois algoritmos para a descoberta de conhecimento (*Heuristics Miner* e *Inductive Miner*). Considera-se relevante a aplicação de outros algoritmos, nomeadamente: *Fuzzy Miner*, *Genetic Miner* e *Social Network Miner*.

7.4 Contributos

Além dos contributos inerentes ao trabalho e aos seus objetivos, é também de ser referido a própria tarefa de extração de dados, amplamente documentada e pode auxiliar outros investigadores na obtenção de dados de eventos para outros trabalhos e noutras áreas.

A extração de dados é uma tarefa morosa, sendo um desafio a localização dos dados relevantes, e o propósito e abrangência dos mesmos. É impossível a consideração de todos os dados, com conjuntos de dados de grande dimensão, e têm de ser considerados quais os dados específicos com interesse e como se relacionam, muitas vezes com diversas tabelas a serem consultadas para a satisfação de dois requisitos: eventos ordenados temporalmente, e correlacionados, com cada evento relacionado com um caso particular.

A reflexão e análise aquando da construção dos registos de eventos é fundantal para alcançar resultados de qualidade. "The quality of a process mining result heavily depends on the input. Therefore, event logs should be treated as first-class citizens in the information systems supporting the processes to be analyzed"(W. Van Der Aalst et al. 2012).

Bibliografia

Aalst, W. M. P. van der, A. K. Alves de Medeiros e A. J. M. M. Weijters (2005). «Genetic Process Mining». Em: Springer, Berlin, Heidelberg, pp. 48–69. doi: 10.1007/11494744_5. url: http://link.springer.com/10.1007/11494744%7B%5C_%7D5.

Banerjee, Anoopam e Preeti Gupta (2015). «Extension to Alpha Algorithm for Process Mining». Em: *nternational Journal Of Engineering And Computer Science* 4.9. url: https://www.ijecs.in/index.php/ijecs/article/view/3168/2931.

Buhl, Hans Ulrich et al. (abr. de 2013). «Big Data». Em: *Business & Information Systems Engineering* 5.2, pp. 65–69. issn: 1867-0202. doi: 10.1007/s12599-013-0249-5. url: http://link.springer.com/10.1007/s12599-013-0249-5.

Buijs, J C A M, B F Van Dongen e W M P Van Der Aalst (s.d.). *On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery*. Rel. téc. url: https://pdfs.semanticscholar.org/2ee1/507c7cffdadd7228bce1bb697bad8b7d63f0.pdf.

Cardoso, J, J A Miller e K J Kochut (2003). *Healthcare Enterprise Process Development and Integration*. Rel. téc. 2, pp. 83–98. url: https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=1211&context=knoesis.

Centers for Disease Control and Prevention (2015). *ICD-9 - International Classification of Diseases, Ninth Revision*. url: https://www.cdc.gov/nchs/icd/icd9cm.htm.

Daniel Larel, R.Marelly (2003). *Come, Let's Play: Scenario-Based Programming Using LSCs and the Play-Engine*. Ed. por Springer. Berlim.

DataQlick (2015). *Businesses and their Growing Dependency on Technology | DataQlick - Dashboardstream Software*. url: http://dashboardstream.com/businesses-growing-dependency-technology/ (acedido em 04/02/2018).

Devi, Aruna T (2017). «An Informative and Comparative Study of Process Mining Tools». Em: *International Journal of Scientific & Engineering Research* 8.5. issn: 2229-5518. url: http://www.ijser.org.

Florence Hudson (2016). *The Internet of Things Is Here | EDUCAUSE*. url: https://er.educause.edu/articles/2016/6/the-internet-of-things-is-here (acedido em 13/02/2018).

GitHub Analytical Hierarchy Process (AHP) with R (s.d.). url: https://github.com/gluc/ahp.

Google Search Statistics - Internet Live Stats (2018). url: http://www.internetlivestats.com/google-search-statistics/ (acedido em 13/02/2018).

Group, Object Management (2006). *BPMN Specification - Business Process Model and Notation*. url: http://www.bpmn.org/.

Günther, C W ; et al. (2007). «Fuzzy mining -adaptive process simplification based on multi-perspective metrics Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics». Em: *Lecture Notes in Computer Science* 4714, pp. 328–343. doi: 10.1007/978-3-540-75183-0_24. url: https://pure.tue.nl/ws/files/2094639/Metis210572.pdf.

IEEE 1849-2016 XES Standard (2016). url: http://www.xes-standard.org/%7B%5C#%7Dieee%7B%5C_%7D1849-2016%7B%5C_%7Dxes%7B%5C_%7Dstandard (acedido em 08/01/2018).

Incubator, Business Process (2009). *Pizza Co. Delivery Process - BPI - The destination for everything process related*. url: https://www.businessprocessincubator.com/content/pizza-co-delivery-process/ (acedido em 15/01/2018).

Jagadeesh, R P et al. (2012). «Wanna Improve Process Mining Results? It's High Time We Consider Data Quality Issues Seriously». Em: url: https://pdfs.semanticscholar.org/89a5/f6d2efa1836f6d0cbe35813cd52a6f33eb34.pdf.

Johnson, Alistair E.W. et al. (mai. de 2016a). «MIMIC-III, a freely accessible critical care database». Em: *Scientific Data* 3, p. 160035. issn: 2052-4463. doi: 10.1038/sdata.2016.35. url: http://www.nature.com/articles/sdata201635.

– (mai. de 2016b). «Table 1: Details of the MIMIC-III patient population by first critical care unit on hospital admission for patients aged 16 years and above.» Em: *Scientific Data*, p. 160035. issn: 2052-4463. doi: 10.1038/sdata.2016.35. url: https://www.nature.com/articles/sdata201635/tables/1.

– (mai. de 2016c). «Table 3: Classes of data available in the MIMIC-III critical care database.» Em: *Scientific Data*, p. 160035. issn: 2052-4463. doi: 10.1038/sdata.2016.35. url: https://www.nature.com/articles/sdata201635/tables/3.

Kantardzic, Mehmed; (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*. 2ª ed. Louisville: Wiley, p. 10. isbn: 978-1-118-02914-5. url: https://books.google.pt/books?hl=pt-PT%7B%5C&%7Dlr=%7B%5C&%7Did=ZZ7l6v0CvRMC%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PA1%7B%5C&%7Ddq=process+mining+gap+between+process+models+and+data%7B%5C&%7Dots=pOumunoMEe%7B%5C&%7Dsig=yhNuEZV3OIm-oQLURiObleDTgX0%7B%5C&%7Dredir%7B%5C_%7Desc=y%7B%5C#%7Dv=onepage%7B%5C&%7Dq=gap%7B%5C&%7Df=false.

Karla, Ana e Alves Alves De Medeiros De Medeiros (s.d.). «/faculteit technologie management Process Mining: Control Process Mining: Control - -Flow Flow Mining Algorithms Mining Algorithms». Em: (). url: http://www.processmining.org/%7B%5C_%7Dmedia/courses/processmining/lecture3%7B%5C_%7Dcontrolflowminingalgorithms.pdf.

Kit Smith (2017). *39 Fascinating and Incredible YouTube Statistics | Brandwatch*. url: https://www.brandwatch.com/blog/39-youtube-stats/ (acedido em 13/02/2018).

Koen, Peter A et al. (2011). «FuzzyFrontEnd: Effective Methods, Tools, and Techniques LITERATURE REVIEW AND RATIONALE FOR DEVELOPING THE NCD MODEL». Em: url: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.320.852%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

Leemans, Sander J J, Dirk Fahland e Wil M P Van Der Aalst (2014). *Process and Deviation Exploration with Inductive visual Miner*. Rel. téc. url: http://vimeo.com/user29103154/.

Lindgreen, A. e J.Y.F. Wynstra (2005). *Publications TU/e*. url: http://repository.tue.nl/611542.

Mitchell, Tom M (2006). «The Discipline of Machine Learning». Em: url: http://www-cgi.cs.cmu.edu/%7B~%7Dtom/pubs/MachineLearningTR.pdf.

Nicola, Susana, Eduarda Pinto Ferreira e J. J. Pinto Ferreira (2012). «A Novel Framework For Modeling Value For The Customer, An Essay On Negotiation». Em: *International Journal of Information Technology & Decision Making* 11.03, pp. 661–703. doi: 10.1142/s0219622012500162.

Online ICD9/ICD9CM codes (2018). url: http://icd9cm.chrisendres.com/index.php?action=child%7B%5C&%7Drecordid=1375 (acedido em 27/08/2018).

Park, Sungbum e Young Sik Kang (jan. de 2016). «A Study of Process Mining-based Business Process Innovation». Em: *Procedia Computer Science* 91, pp. 734–743. issn: 1877-0509. doi: 10.1016/J.PROCS.2016.07.066. url: https://www.sciencedirect.com/science/article/pii/S1877050916312492.

Pgadmin (2018). *pgAdmin 4*. url: https://www.pgadmin.org/docs/pgadmin4/dev/.

PostgreSQL (2018). *PostgreSQL: About*. url: https://www.postgresql.org/about/.

Poulymenopoulou, M., F. Malamateniou e G. Vassilacopoulos (2003). «Specifying Workflow Process Requirements for an Emergency Medical Service». Em: *Journal of Medical Systems* 27.4, pp. 325–335. doi: 10.1023/A:1023701219563. url: http://link.springer.com/10.1023/A:1023701219563.

Product - Minit Process Intelligence Software (2018). url: https://www.minit.io/product%7B%5C#%7Doptions (acedido em 12/02/2018).

ProM Tips — Which Mining Algorithm Should You Use? — Flux Capacitor (2018). url: https://fluxicon.com/blog/2010/10/prom-tips-mining-algorithm/ (acedido em 11/02/2018).

Rojas, Eric et al. (jun. de 2016). «Process mining in healthcare: A literature review». Em: *Journal of Biomedical Informatics* 61, pp. 224–236. issn: 1532-0464. doi: 10.1016/J.JBI.2016.04.007. url: https://www.sciencedirect.com/science/article/pii/S1532046416300296.

Rozinat, A et al. (2008). *Towards an Evaluation Framework for Process Mining Algorithms*. Rel. téc. url: http://www.processmining.org/%7B%5C_%7Dmedia/publications/bpm-07-06.pdf.

RStudio (s.d.). url: https://www.rstudio.com/products/rstudio/.

Rudnitckaia, Julia (2015). «Process Mining. Data science in action». Em: p. 4. url: http://www.fit.vutbr.cz/study/courses/TJD/public/1415TJD-Rudnickaia.pdf.

Saaty, Thomas L (2008). «Decision making with the analytic hierarchy process». Em: *Int. J. Services Sciences* 1.1, pp. 83–98. url: http://www.rafikulislam.com/uploads/resourses/197245512559a37aadea6d.pdf.

Sloan Digital Sky Survey (2015). *The Sloan Digital Sky Survey Opens a New Public View of the Sky*. url: http://www.sdss.org/releases/the-sloan-digital-sky-survey-opens-a-new-public-view-of-the-sky/ (acedido em 06/01/2018).

Song, Minseok e Wil M P Van Der Aalst (s.d.). «Towards Comprehensive Support for Organizational Mining». Em: (). url: http://www.pads.rwth-aachen.de/wvdaalst/publications/p484.pdf.

Tao Li (2015). *Event Mining - Algorithms and Applications*. url: https://dl.acm.org/citation.cfm?id=2994430 (acedido em 05/02/2018).

The 4 V's of Big Data - Zarantech (2016). url: http://www.zarantech.com/blog/the-4-vs-of-big-data/ (acedido em 10/02/2018).

Van Der Aalst, Wil (2016a). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. Process Mining: The Missing Link, p. 25. isbn: 978-3-662-49850-7.

– (2016b). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven, p. 34. isbn: 978-3-662-49850-7.

– (2016c). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven, p. 42. isbn: 978-3-662-49850-7.

– (2016d). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven, p. 129. isbn: 978-3-662-49850-7.

– (2016e). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven, p. 139. isbn: 978-3-662-49850-7.

Van Der Aalst, Wil (2016f). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven, p. 138. isbn: 978-3-662-49850-7.
– (2016g). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. Process Modeling and Analysis, p. 58. isbn: 978-3-662-49850-7.
– (2016h). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. Process Modeling and Analysis, p. 60. isbn: 978-3-662-49850-7.
– (2016i). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. A Simple Algorithm For Process Discovery, p. 167. isbn: 978-3-662-49850-7.
– (2016j). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. Heuristic Mining, p. 201. isbn: 978-3-662-49850-7.
– (2016k). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. Advanced Process Discovery Techniques, p. 334. isbn: 978-3-662-49850-7.
– (2016l). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. A Simple Algorithm For Process Discovery, p. 171. isbn: 978-3-662-49850-7.
– (2016m). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. Process Mining Software, p. 340. isbn: 978-3-662-49850-7.
– (2016n). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. ProM: An Open-Source Process Mining Platform, p. 331. isbn: 978-3-662-49850-7.
– (2016o). «Process Mining Data Science in Action». Em: ed. por Springer. 2ª ed. Eindhoven. Cap. ProM: An Open-Source Process Mining Platform, p. 334. isbn: 978-3-662-49850-7.
Van Der Aalst, Wil et al. (2012). *Process Mining Manifesto*. Rel. téc., pp. 169–194. url: https://link.springer.com/content/pdf/10.1007/978-3-642-28108-2%7B%5C_%7D19.pdf.
Witten, Ian H. et al. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. 4ª ed. Morgan Kaufmann. Cap. What's its all about?, p. 8. isbn: 978-0-12-804291-5.
Woodall, Tony (jan. de 2003). «Conceptualising 'Value for the Customer': An Attributional, Structural and Dispositional Analysis». Em: 12.

Apêndice A

Metódo AHP em R

```
1
2  # The following code is an exact copy of the code available at:
       https://github.com/gluc/ahp
3
4  devtools::install_github("gluc/ahp", build_vignettes = TRUE)
5  vignette("car-example", package = "ahp")
6  vignette("multiple-decisionmakers", package = "ahp")
7
8  # run analysis
9  library(ahp)
10 ahpFile <- system.file("extdata", "car.ahp", package = "ahp")
11 carAhp <- Load(ahpFile)
12 Calculate(carAhp)
13 Visualize(carAhp)
14 Analyze(carAhp)
15 AnalyzeTable(carAhp)
16
17
18 # looking at the vacation example, a multi-decision-maker model
19 ahpFile <- system.file("extdata", "vacation.ahp", package = "ahp")
20 vacationAhp <- Load(ahpFile)
21 Calculate(vacationAhp)
22 Analyze(vacationAhp, decisionMaker = "Dad")
23 AnalyzeTable(vacationAhp, decisionMaker = "Mom")
24 AnalyzeTable(vacationAhp)
25 RunGUI()
```

Apêndice B

Ficheiro AHP em formato YAML

```
1
2  Version: 2.0
3
4  ##########################
5  # Alternatives Section
6  #
7
8  Alternatives: &alternatives
9
10 # For this reason, here the alternatives are specified
       without any attributes.
11 ProM:
12 Disco:
13 Minit:
14
15 #
16 # End of Alternatives Section
17 ####################################
18
19 ####################################
20 # Goal Section
21 #
22
23
24 Goal:
25 name: Escolher a melhor ferramenta de mineracao de
       processos.
26 description: >
27 Este e um programa simples de decision making usando AHP.
       O objetivo e escolher a melhor ferramenta de mineracao
       de processos usando uma serie de criterios definidos.
28 author: Ana Carolina Ferreira Barros
29 preferences:
30 # preferences are typically defined pairwise
31 # 1 means: A is equal to B
32 # 9 means: A is highly preferrable to B
33 # 1/9 means: B is highly preferrable to A
```

```
34 pairwise:
35
36 - [Categoria, Formatos, 2/1]
37 - [Categoria, Tipos, 1]
38 - [Categoria, Notacoes, 2/1]
39 - [Categoria, Plugins, 2]
40 - [Formatos, Tipos, 1/2]
41 - [Formatos, Notacoes, 2/1]
42 - [Formatos, Plugins, 1/2]
43 - [Tipos, Notacoes, 2/1]
44 - [Tipos, Plugins, 1/2]
45 - [Notacoes, Plugins, 1/2]
46
47
48 children:
49 Categoria:
50 preferences:
51 pairwise:
52 - [ProM, Disco, 7]
53 - [ProM, Minit, 9]
54 - [Minit, Disco, 1/5]
55 children: *alternatives
56 Formatos:
57 preferences:
58 pairwise:
59 - [ProM, Disco, 1/5]
60 - [ProM, Minit, 1/9]
61 - [Minit, Disco, 9]
62 children: *alternatives
63 Tipos:
64 preferences:
65 pairwise:
66 - [ProM, Disco, 7]
67 - [ProM, Minit, 7]
68 - [Minit, Disco, 1]
69 children: *alternatives
70 Notacoes:
71 preferences:
72 pairwise:
73 - [ProM, Disco, 7]
74 - [ProM, Minit, 7]
75 - [Minit, Disco, 1/4]
76 children: *alternatives
77 Plugins:
78 preferences:
79 pairwise:
80 - [ProM, Disco, 9]
81 - [ProM, Minit, 9]
82 - [Minit, Disco, 1]
```

```
83 children: *alternatives
84
85 #
86 # End of Goal Section
87 ######################################
```

Apêndice C

Script Utilizado - Requisito 1

```
CREATE TABLE mimiciii.table1 AS
(SELECT DISTINCT
admissions.subject_id,
admissions.hadm_id,
'Admission' AS activity,
admissions.admittime AS charttime
FROM
mimiciii.admissions
WHERE admittime IS NOT NULL
UNION ALL
SELECT DISTINCT
admissions.subject_id,
admissions.hadm_id,
'Death' AS activity,
admissions.deathtime AS charttime
FROM
mimiciii.admissions
WHERE deathtime IS NOT NULL
UNION ALL
SELECT DISTINCT
admissions.subject_id,
admissions.hadm_id,
'Discharge',
admissions.dischtime
FROM
```

```
mimiciii.admissions
WHERE dischtime IS NOT NULL
UNION ALL
SELECT DISTINCT
admissions.subject_id,
admissions.hadm_id,
'Emergency Department Registration' AS activity,
admissions.edregtime AS charttime
FROM
mimiciii.admissions
WHERE edregtime IS NOT NULL
UNION ALL
SELECT DISTINCT
admissions.subject_id,
admissions.hadm_id,
'Emergency Department Exit' AS activity,
admissions.edouttime AS charttime
FROM
mimiciii.admissions
WHERE edouttime IS NOT NULL
UNION ALL
SELECT DISTINCT
icustays.subject_id,
icustays.hadm_id,
'Care Unit CCU' AS activity,
icustays.intime AS charttime
FROM
mimiciii.icustays
WHERE intime IS NOT NULL AND first_careunit='CCU'
UNION ALL
SELECT DISTINCT
icustays.subject_id,
icustays.hadm_id,
```

```
'Care Unit CSRU' AS activity,
icustays.intime AS charttime
FROM
mimiciii.icustays
WHERE intime IS NOT NULL AND first_careunit='CSRU'
UNION ALL
SELECT DISTINCT
icustays.subject_id,
icustays.hadm_id,
'Care Unit MICU' AS activity,
icustays.intime AS charttime
FROM
mimiciii.icustays
WHERE intime IS NOT NULL AND first_careunit='MICU'
UNION ALL
SELECT DISTINCT
icustays.subject_id,
icustays.hadm_id,
'Care Unit SICU' AS activity,
icustays.intime AS charttime
FROM
mimiciii.icustays
WHERE intime IS NOT NULL AND first_careunit='SICU'
UNION ALL
SELECT DISTINCT
icustays.subject_id,
icustays.hadm_id,
'Care Unit TSICU' AS activity,
icustays.intime AS charttime
FROM
mimiciii.icustays
WHERE intime IS NOT NULL AND first_careunit='TSICU'
UNION ALL
```

```
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'CMED' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='CMED'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'CSURG' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='CSURG'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'MED' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='MED'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'NMED' AS activity,
services.transfertime AS charttime
FROM
```

```
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='NMED'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'GU' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='GU'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'GYN' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='GYN'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'NSURG' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='NSURG'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
```

```
'SURG' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='SURG'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'TSURG' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='TSURG'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'ORTHO' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='ORTHO'
UNION ALL
SELECT DISTINCT
services.subject_id,
services.hadm_id,
'VSURG' AS activity,
services.transfertime AS charttime
FROM
mimiciii.services
WHERE transfertime IS NOT NULL AND curr_service='VSURG')
ORDER BY subject_id, hadm_id, activity, charttime;
```

```
CREATE TABLE mimiciii.table1_1 AS
SELECT DISTINCT
table1.subject_id,
table1.hadm_id,
table1.activity,
table1.charttime
FROM
mimiciii.table1,
mimiciii.diagnoses_icd
WHERE
table1.hadm_id = diagnoses_icd.hadm_id AND
diagnoses_icd.icd9_code BETWEEN '190%' AND '199%';
```

Apêndice D

Verificação de Conformidade - Modelo Completo Requisito 1 (Heuristics Miner)

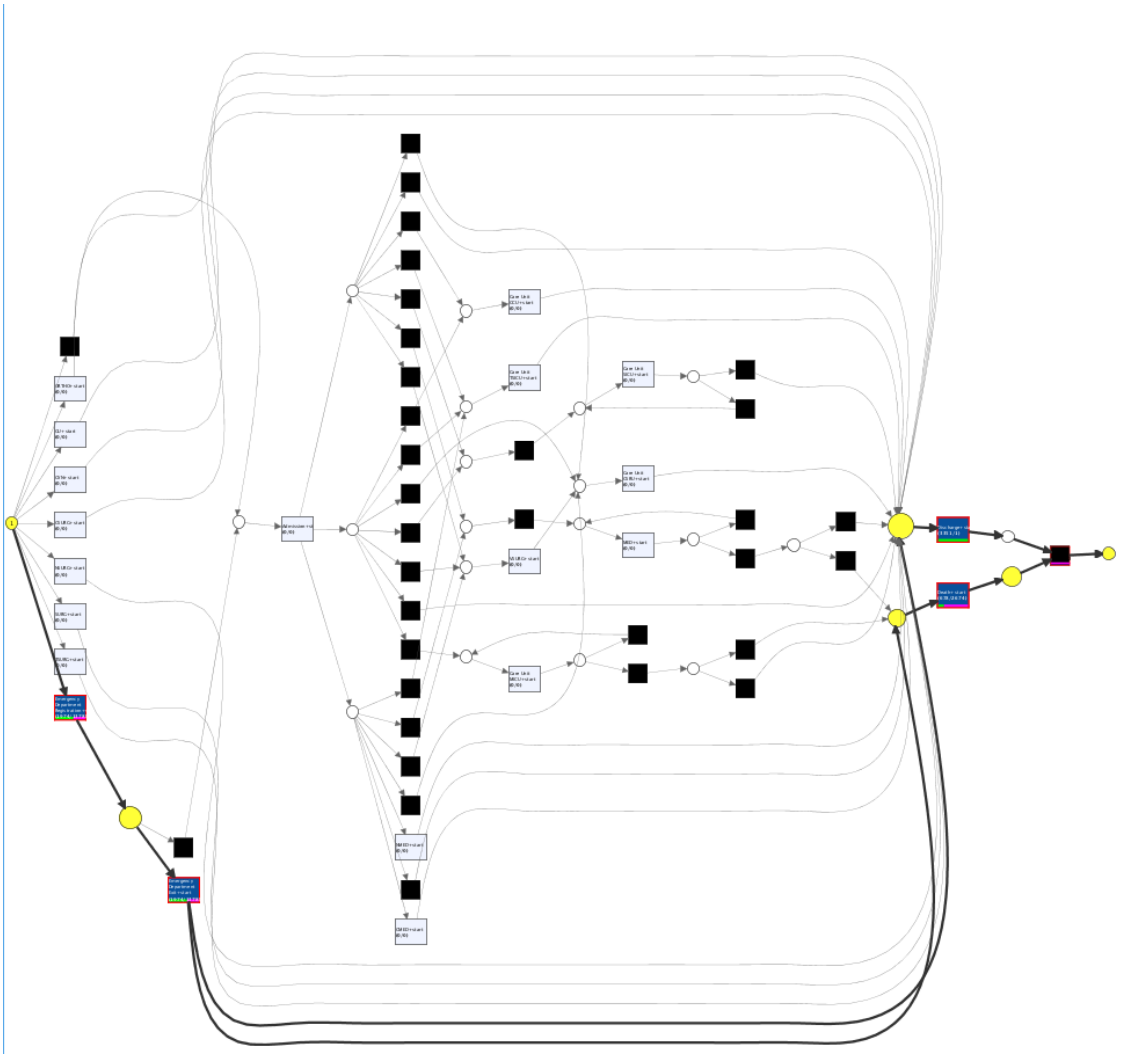


Figura 43: Verificação de Conformidade - Modelo Completo Requisito 1 (Heuristics Miner)

Apêndice E

Verificação de Conformidade - Modelo Completo Requisito 2 (Heuristics Miner)