PROPOSTA PARA APLICAÇÃO DE CIÊNCIA DE DADOS E MINERAÇÃO DE PROCESSOS EM SERVIÇOS DE SAÚDE – Projeto de Pesquisa

MÁRCIA ITO[1]
[1]*Fatec SP - Departamento de Tecnologia da Informação*
e-mail marcia.ito01@fatec.sp.gov.br

Proposal for the application of data science and process mining in health services

Eixo Tecnológico: *Ambiente e Saúde*

Resumo

A tomada de decisão é feita com base na análise de informações, por outro lado a capacidade do ser humano na análise rápida e precisa das informações é limitada, ao contrário da tecnologia da informação em que o processamento é infinito. Na área da saúde as análises usando inteligência artificial tem auxiliado desde o diagnóstico até a melhoria no acesso a serviços de saúde. O objetivo desta pesquisa é aplicar as técnicas de ciência de dados e mineração de processos para analisar e apoiar a tomada de decisão em serviços de saúde. Os objetivos específicos são (a) entender o contexto e as necessidades na saúde; (b) coleta e pré-processamento dos dados; (c) desenvolvimento e avaliação de modelos e da descoberta dos processos reais; (d) desenvolvimento e validação de painéis interativos; (e) implantação e transferência da tecnologia. Esta pesquisa é do tipo exploratória, pois investiga como desenvolver modelos preditivos, usando técnicas de aprendizado de máquina; descobrir processos usando técnicas de mineração de processos e o desenvolvimento de painéis interativos. Avaliações de usabilidade, consistência e segurança serão feitos. Os resultados esperados são modelos e painéis que auxiliarão a tomada de decisão dos gestores e colaboradores dos serviços de saúde e por consequência, melhorando a oferta e qualidade dos serviços de saúde. Além disso a divulgação da pesquisa por meio de artigos científicos e palestras trarão possibilidades de disseminação do conhecimento para área de saúde digital.

Palavras-chave: *Saúde Digital, Ciência de Dados, Mineração de Processos. Serviço de Saúde, Gestão Pública.*

Abstract

Decision-making is based on the analysis of information. On the other hand, the capacity of human beings to analyze information quickly and accurately is limited, unlike information technology where processing is infinite. In the health sector, analysis using artificial intelligence has helped with everything from diagnosis to improving access to health services. The aim of this research is to apply data science and process mining techniques to analyze and support decision-making in health services. The specific objectives are (a) to understand the health context and needs; (b) data collection, pre-processing; (c) development and evaluation of models and discovery of real processes; (d) development and validation of interactive panels; (e) deployment and transfer of technology. This research is exploratory, as it investigates how to develop predictive models, using machine learning techniques; discover processes using process mining techniques and the development of interactive dashboards. Usability, consistency and security evaluations will be carried out. The expected results are models and dashboards that will help health service managers and employees make better decisions, improving the supply and quality of health services. In addition, the dissemination of the research through scientific articles and lectures will bring possibilities for disseminating knowledge in the area of digital health.

Key-words: *Digital health, Data Science, Process Minin,. Health services, Public management.*

1. Introdução

As transformações que a indústria e a sociedade tiveram a partir da era digital, afetam todas as áreas inclusive a da saúde e com isso surge o conceito de Saúde Digital que expande a definição da e-Saúde ao incluir os consumidores digitais que possuem os mais diversos dispositivos conectados e inteligentes. O aumento da popularidade das tecnologias de saúde digital e a necessidade de decisões baseadas em evidência tem

contribuído para a rápida expansão de um grande volume de dados nos últimos anos. Estudos de mercado projetam uma taxa de crescimento global de 19,2% de grande volume de dados (big data) na saúde digital nos próximos 10 anos, aumentando o seu valor para estimados US$ 39,7 bilhões só em 2022. As possibilidades de análises e processamento desses dados são inúmeras e todas podem trazer inovações que levarão a melhores resultados para a gestão de saúde, pacientes e a sociedade como um todo [1].

Ciência de dados é uma área interdisciplinar que extrai conhecimento e ideias de dados estruturados, semiestruturados e não estruturados usando métodos científicos, técnicas de mineração de dados, algoritmos de aprendizado de máquina e *big data*. Na área saúde os dados vêm em várias formas e fontes como os registros eletrônicos do paciente, prescrições, laudos de exames, sinistros, redes sociais, dispositivos IoT, entre outros. A integração e análise detalhada e o entendimento dos padrões podem ajudar a melhorar a tomada de decisão resultando no aumento da qualidade da assistência à saúde. De acordo com Subrahmanya et. al [2], 77% são dados gerados por profissionais da saúde e pacientes/cuidadores, 11% por cidadão por meio das redes sociais e 12% por sensores. A qualidade do dado gerado por humanos deve ter especial atenção, pois erros podem ocorrer com maior frequência fazendo com que a fase de pré-processamento como por exemplo a limpeza dos dados seja crucial para um adequado resultado na análise dos dados. As aplicações são inúmeras como a vigilância em saúde, assistência, entre outros. Como este projeto relaciona-se com os serviços de saúde, dar-se-á a partir deste momento, maior ênfase na aplicação da ciência de dados e mineração de processos na gestão em saúde.

Na gestão de saúde os dados primários gerados estão relacionados com o processo de assistência, de pagamento, de execução dos serviços prestados e da experiência do usuário. Além disso, o objetivo de um serviço de saúde é executá-lo de uma forma a trazer uma relação custo-benefício adequado sem perder a qualidade do serviço e dando uma boa experiência ao paciente e familiares. Assim, tem-se que os dados de uso primário, coletados não suprem a necessidade de informações que ajudem na tomada de decisão de gestores. Por isso é necessário fazer o processamento, armazenamento e organização adequado dos dados secundários integrados aos primários para aproveitar ao máximo essas informações para auxiliar os gestores na tomada de decisão. Ao aplicar as técnicas de ciência de dados é possível ter informações que permitam otimizar o uso e alocação dos recursos de saúde, melhorando a qualidade e experiência dos usuários. Políticas de saúde e programas de promoção e prevenção a saúde podem ser instituídos ao desenvolver modelos preditivos usando algoritmos de aprendizado de máquina [1] [3].

No serviço de saúde os processos para a assistência é um ponto delicado, pois com o aumento das doenças crônicas houve um aumento no uso do serviço de saúde. A gestão destes pacientes no serviço se torna complexo, pois além da complexidade de lidar com várias doenças há o problema da intensidade de intervenções. Portanto, os serviços de saúde precisam se organizar de forma a comportar a intensidade e a integração necessária para estes pacientes. A análise do processo é importante para que o planejamento não seja feito de modo empírico. Portanto, o estudo dos processos de atendimento permite customizar o tempo de atendimento do paciente de acordo com a sua necessidade, assim como, gerenciar e analisar as informações ao longo do processo de assistência permitem melhorar a assistência, otimizar o uso dos recursos e reduzir os custos operacionais [4] [5].

Assim, o objetivo desta proposta é aplicar as técnicas de ciência de dados para analisar e apoiar a tomada de decisão em serviços de saúde. Além da mineração de dados fará a aplicação de técnicas de mineração de processo a fim de analisar e avaliar os processos

dos serviços de saúde. Também será possível fazer análises preditivas com o uso de técnicas de aprendizado de máquina com a finalidade de apoiar gestores e colaboradores na sua tomada de decisão.

Os objetivos específicos para alcançar as metas estabelecidas são:

- Mapeamento sistemático da literatura na aplicação de técnicas de ciências de dados e mineração de processos em serviços públicos de saúde;
- Entrevistas semiestruturadas e sessões de *design thinking* com os colaboradores e gestores das organizações parceiras;
- Coleta, organização, anonimização, pré-processamento e análise preliminar dos dados das instituições parceiras;
- Desenvolvimento de modelos preditivos para os gestores e colaboradores de acordo com a necessidade coletada durante as entrevistas semiestruturadas;
- Descoberta do processo real do atendimento ao Infarto Agudo do Miocárdio, análise de sua aderência ao protocolo de regulação;
- Desenvolvimento e avaliação de usabilidade, confiabilidade e segurança dos painéis interativos com as análises preliminares dos dados e os modelos criados;
- Transferência da tecnologia utilizada e implantação dos painéis.

2. Materiais e métodos

2.1. Materiais

Os dados são o material mais importante desta proposta de projeto, assim foram selecionados dois grupos de dados que serão utilizados nesta proposta:

- de dengue do departamento de vigilância sanitária da Secretaria de Saúde do Município de São José do Rio Preto - Projeto aprovado na FAPESP número 2023100803 – os dados virão do Sistema SISAWEB e SINAM;
- de regulação de acesso dos casos de infarto agudo do miocárdio do Centro de Regulação de Sistemas de Saúde (CROSS) da Secretaria de Saúde do Estado de São Paulo (SES-SP) – Projeto em fase de aprovação pela SES-SP - os dados serão coletados dos sistemas que registram os dados das fichas de urgência, horários de agendamento e listagem do Cadastro de Demanda por Recurso.

Para o estado da arte, mapeamentos sistemáticos da literatura serão realizados e que usarão o fluxograma PRISMA, PARSIFAL (software colaborativo para o registro do protocolo), ferramentas de busca (PubMed e Scopus), Zotero (software de gerenciamento de referências), VosViewer (software para análise léxica) e RAYAN para a seleção dos artigos.

As entrevistas serão baseadas no protocolo COREQ (*Consolidated Criteria for Reporting Qualitative Research*) e o *desing thinking* no Framework HCD (*Human-Centered Design*). Nas entrevistas semiestrutura um questionário, ficha com o perfil dos entrevistados, formulário de consentimento informado, roteiro de entrevistas serão elaboradoras. O Trint (software de transcrição) será utilizado para a transcrição das entrevistas. Além disso, as atividades vão precisar do Canvas de Empatia, Mapas de Jornada do Usuário, Brainstorming, Miro (software colaborativo).

Para as análises e desenvolvimento dos modelos as ferramentas Python e suas bibliotecas serão utilizadas, assim como para o desenvolvimento dos paineis. Em alguns casos para a descoberta dos processos a ferramenta ProM poderá ser usada.

2.2. Metodologia

Esta pesquisa é do tipo exploratória ao investigar como desenvolver modelos preditivos, descobrir processos e construir painéis interativos para análise de dados de serviços de saúde. Adotará uma abordagem mista, combinando métodos quantitativos e qualitativos para a criação de painéis que auxiliem na tomada de decisão de gestores e colaboradores das organizações parceiras.

O mapeamento sistemático da literatura na aplicação de técnicas de ciências de dados e mineração de processos em serviços públicos de saúde será feita. O Protocolo do mapeamento sistemático será baseado em Kitchenham [6]. As pessoas envolvidas na tarefa é a pesquisadora, pesquisadores do LNCC, mestrandos do programa de pós do CPS e pesquisadores do UFSJ, UFF e UFJF.

Para o trabalho com os dados do CROSS-SES-SP é preciso definir as necessidades na tomada de decisão dos gestores e colaboradores e para este fim serão aplicadas entrevistas semiestruturadas e sessões de *design thinking*. As entrevistas serão gravadas e transcrita e análise de conteúdo serão feitas, para definir os modelos e painéis relevantes que serão desenvolvidas no projeto. Para este trabalho ter-se-ia um grupo composto pela pesquisadora, mestrandos do programa de pós do CPS, alunos de TCC da Fatec-SP e alunos e professores da disciplina do projeto Integrador I do curso de Ciência de Dados da Fatec Cotia.

Com as necessidades das organizações parceiras serão definidos os dados necessários para desenvolver os modelos do projeto. Os dados coletados serão do tipo estruturado, semiestruturado e não estruturados. Após coleta, o pré-processamento dos dados serão realizados, como por exemplo a anonimização e a normalização dos dados. Após processados, será feita uma análise exploratória com uso de métodos de estatísticas descritivas e visualizações (histogramas, boxplots, etc.). Os softwares para armazenamento e organização dos dados serão definidos após o pré-processamento dos dados, pois o dimensionamento da base somente poderá ser feito após esta análise inicial. Para esta fase as pessoas envolvidas é um grupo composto pela pesquisadora, pesquisador da UFES, mestrandos do programa de pós do CPS, alunos de TCC da Fatec-SP e alunos do curso de Ciência de Dados da Fatec-Cotia.

Modelos baseados em aprendizado de máquina (regressão linear, árvore de decisão, etc.) serão desenvolvidos a fim de atender as necessidades das organizações parceiras. A avaliação dos modelos será feita usando métricas de desempenho (acurácia, precisão, recall, F1-score, etc.) e Curvas ROC. Avaliações comparando o resultado do modelo com o mundo real serão feitos sempre que possível, os desenhos das avaliações serão feitos conforme a necessidade de cada caso. Como os modelos não estão relacionados com diagnósticos e tratamento não há razão para avaliação do tipo ensaio clínico. No projeto da FAPESP a pesquisadora faz parte de um grupo que irá desenvolver os modelos, neste grupo encontram-se pesquisadores da USP, LNCC, UFJF e UFSJ. No projeto do IAM/CROSS a pesquisadora irá desenvolver os modelos com mestrandos do programa de pós do CPS, pesquisador do UFES e UFF e alunos de TCC e IC do curso de ADS da FATEC-SP. A descoberta do processo real em relação ao atendimento do IAM será feito usando uma técnica baseada em grafos multiaspectos [7]. Para validar o processo obtido um estudo comparativo com outras técnicas de mineração de processos como o *heuristic miner* [8] serão realizados. Uma vez validado que o processo encontrado é o real, será feito uma análise de sua aderência ao protocolo de regulação do CROSS. Os envolvidos

nas atividades serão feitas pela pesquisadora com mestrandos do programa de pós do CPS, pesquisador da UFES e UFF, alunos de TCC e IC do curso de ADS da FATEC-SP.

Tendo as análises estatísticas, os modelos preditivos e a análise dos processos, painéis interativos serão desenvolvidos a partir das necessidades e dos perfis dos tomadores de decisão. Testes de consistências, segurança, usabilidade e experiência dos usuários baseados em protocolos reconhecidos serão realizados nos painéis e analisados, como por exemplo o *System Usability Scale* [9] ou o *Technology Acceptance Model* [10]. No projeto do IAM no CROSS, esta fase será feita pela pesquisadora, mestrandos do programa de pós do CPS, alunos de TCC da Fatec-SP e alunos e professores da disciplina do projeto Integrador I do curso de Ciência de Dados da Fatec Cotia.

Em seguida a implantação e transferência do conhecimento para as organizações parceiras serão planejadas e realizadas com a produção de material instrucional e workshops. Os pesquisadores ficarão responsáveis pelas reuniões de planejamento e desenvolvimento de material serão realizados com os envolvidos.

Como informações da área de saúde com dados sensíveis estarão presentes, será necessário obter a aprovação do Comitê de Ética em ambos os casos, assim como nas entrevistas semiestruturadas. Em ambos os projetos serão considerados os possíveis impactos sociais e éticos das decisões automatizadas geradas pelos modelos. No Projeto FAPESP, já foi obtido a aprovação pelo CEP (CAAE: 77937424.0.0000.5421). No projeto CROSS-SES-SP o projeto foi submetido ao comitê de ética da plataforma Brasil em abril e ainda está em análise. As entrevistas semiestruturadas serão submetidas como adendo ao CEP principal.

3. Resultados esperados e obtidos até o momento

Como o projeto iniciou-se em 2025, os resultados obtidos até o momento são iniciais e, portanto, insuficientes para que uma discussão ampla sobre a proposta e metas alcançadas seja produzida. Pode-se dizer que as atividades estão obedecendo o cronograma inicial estabelecido e que mesmo o projeto do CROSS-SES-SP estando em análise pelo CEP da Plataforma Brasil, o grupo tem trabalhado em simulações e experimentos que permitirão, assim que os dados forem liberados fazer uso deles rapidamente, pois já se sabe o que e como desenvolver os modelos.

Assim, sendo um dos resultados a formalização de acordo de colaborações com outras instituições de pesquisa, até o momento tem-se declarações de participações de pesquisadores de outras instituições de pesquisa e iniciou-se a formalização do acordo de colaboração com a UFES, para depois ser replicada nas outras instituições.

Para o resultado da revisão da literatura, o mapeamento sistemático do uso de técnicas de sistemas complexos em dengue foi feito e o artigo está em desenvolvimento. O mapeamento da aplicação de técnicas de mineração de processos em cardiologia foi desenvolvido e publicado no 1º. semestre de 2025. Atualmente está em andamento um mapeamento sistemático de otimização e mineração de processos na área da saúde.

O relatório de um ano de projeto da FAPESP foi entregue em março de 2025 na qual o grupo de modelagem com técnicas de ciências de redes contribuiu com os achados no estado da arte e simulações sobre possíveis soluções. Além disso, uma proposta de patente está em andamento com a orientação do INOVA-CPS.

Para o ambiente e o modelo de arquitetura de dados está em desenvolvimento uma POC para validar o modelo de arquitetura de dados proposto para o projeto.

Outro resultado esperado são os modelos do processo real do atendimento ao Infarto Agudo do Miocárdio que permita análises de conformidade e otimização. Com a ausência dos dados do CROSS, estamos realizando experimentos com dados de outro projeto. E espera-se ter um artigo para submeter ao Congresso Internacional de Mineração de Processos que ocorrerá em outubro no Uruguai.

Por fim painéis interativos desenvolvidos e validados para uso dos gestores e colaboradores das organizações parceiras; manuais de uso, implantação e manutenção dos painéis e modelos são resultados esperados nesta proposta. Até o momento estamos desenvolvendo os roteiros e as entrevistas para aplicar no projeto.

4. Considerações finais

Com os resultados obtidos até o momento pode-se concluir que a proposta é viável e será possível alcançar os objetivos propostos, pois as revisões e mapeamentos sistemáticos da literatura demonstram que é factível aplicar as técnicas de ciências de dados e mineração de processos nos grupos de dados propostos para este projeto. Um risco considerado neste projeto é a demora na aprovação pelo CEP da Plataforma Brasil que pode comprometer o cronograma do projeto.

Agradecimentos

Parte desta proposta foi realizado com o apoio da FAPESP através do Projeto aprovado sob o número 2023100803. Agradecimentos também ao CPRJI que ao aprovar o projeto da pesquisadora para o regime de jornada integral (RJI) está permitindo a execução deste projeto proposto e relatado neste resumo expandido.

Referências

[1] GOYAL, P.; MALVIYA, R. Challenges and opportunities of big data analytics in healthcare. **Health Care Science**, v. 2, n. 5, p. 328–338, out. 2023.

[2] SUBRAHMANYA, S. V. G. et al. The role of data science in healthcare advancements: applications, benefits, and future prospects. **Irish Journal of Medical Science (1971 -)**, v. 191, n. 4, p. 1473–1483, ago. 2022.

[3] ZHOU, S. et al. A novel framework for bringing smart big data to proactive decision making in healthcare. **Health Informatics Journal**, v. 27, n. 2, p. 146045822110246, abr. 2021.

[4] ITO, M. Chapter 6 - patient-centered care. In Gogia, S., editor, **Fundamentals of Telemedicine and Telehealth**, pages 115–126. Academic Press. 2020.

[5] ROSA, C.O.C.S., ITO, M., VIEIRA, A.B., GOMES, T.A. Modelagem, Mineração e Análise de Jornadas/Trajetórias de Pacientes. In Oliveira, L.F. e de Araújo, F.H.D., editores, **Minicursos do XXII Simpósio Brasileiro de Computação Aplicada à Saúde**. Sociedade Brasileira de Computação. 2022. DOI: https://doi.org/10.5753/sbc.10508.8.3

[6] KITCHENHAM, B. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. Keele University, 2007. Disponível em: <https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf>. Acesso em: 12/04/2025

[7] ROSA, C. DE O. C. S. **Complex networks to model and mine patient pathways**. Tese—Petrópolis: Laboratório Nacional de Computação Científica, 2024.

[8] NAMAKI ARAGHI, S. et al. Stable Heuristic Miner 2: Evaluating the Statistical Stability in Event Logs to Discover Business Processes. **Human-Centric Intelligent Systems**, v. 4, n. 2, p. 256–277, 15 mar. 2024.

[9] BANGOR, A.; KORTUM, P. T.; MILLER, J. T. An Empirical Evaluation of the System Usability Scale. **International Journal of Human-Computer Interaction**, v. 24, n. 6, p. 574–594, 29 jul. 2008.

[10] HOLDEN, R. J.; KARSH, B.-T. The Technology Acceptance Model: Its past and its future in health care. **Journal of Biomedical Informatics**, v. 43, n. 1, p. 159–172, fev. 2010.