PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Bacharelado em Ciência da Computação

Alice Pereira de Aguilar Penido

Visualização Interativa para Modelos de Inteligência Artificial Explicáveis

Belo Horizonte

2024

Alice Pereira de Aguilar Penido

Visualização Interativa para Modelos de Inteligência Artificial Explicáveis

Projeto de Pesquisa apresentado na disciplina Trabalho Interdisciplinar III - Pesquisa Aplicada do curso de Ciência da Computação da Pontifícia Universidade Católica de Minas Gerais.
Professor: Leonardo Vilela

Belo Horizonte

2024

RESUMO

Este projeto tem como objetivo o desenvolvimento de uma plataforma de visualização interativa para tornar modelos de *Artificial Inteligence* (AI) mais compreensíveis e confiáveis para os usuários. Com a utilização de metodologias de *Explainable Artificial Inteligence* (XAI), como o *SHapley Additive exPlanations* (SHAP), a plataforma permitirá que os usuários manipulem variáveis e visualizem como essas mudanças impactam as previsões dos modelos, promovendo transparência e facilitando a compreensão dos processos decisórios automatizados.

Palavras-chave: *Explainable Artificial Inteligence* (XAI), *Machine Learning* (ML), *SHapley Additive exPlanations* (SHAP), Visualização Interativa, Interação Humano-Computador (IHC), Acessibilidade, Transparência de Modelos, Manipulação de Variáveis-

LISTA DE ABREVIATURAS E SIGLAS

AI – *Artificial Inteligence*

DSR – *Design Science Research*

IHC – Interação Humano-Computador

LIME – *Local Interpretable Model-agnostic Explanations*

ML – *Machine Learning*

NLP – *Natural Language Processing*

SHAP – *SHapley Additive exPlanations*

XAI – *Explainable Artificial Inteligence*

SUMÁRIO

1 INTRODUÇÃO

A crescente complexidade dos modelos de AI e o aumento de seu uso em diversas áreas, como saúde, finanças e indústria, levantam a necessidade de tornar esses modelos mais compreensíveis para os usuários. A XAI busca resolver essa questão ao oferecer técnicas que tornam os modelos mais transparentes, permitindo que os usuários humanos compreendam e confiem nas decisões automatizadas. Entre essas técnicas, destaca-se o SHAP, que utiliza a teoria dos jogos para explicar as previsões dos modelos, atribuindo a cada variável uma contribuição para o resultado final.

Entretanto, uma das principais limitações dessas abordagens é a dificuldade de interação e exploração dos dados pelos usuários, o que impede uma compreensão profunda dos processos decisórios subjacentes. Assim, a criação de plataformas de visualização interativas torna-se essencial, permitindo aos usuários não apenas observar, mas também interagir com os dados e os modelos, escolhendo variáveis e vendo como suas alterações afetam os resultados. Isso melhora a confiança nas decisões geradas pela AI e facilita a compreensão dos mecanismos dos modelos, tornando-os mais acessíveis e transparentes.

Este projeto visa desenvolver uma plataforma de visualização interativa para modelos de AI explicáveis, utilizando metodologias como o SHAP para ilustrar os impactos das variáveis nas previsões, permitindo uma exploração intuitiva e visual dos modelos e seus processos decisórios.

Este trabalho está organizado da seguinte forma. A **Seção 1.1** apresenta os objetivos gerais e específicos do projeto, destacando as metas a serem alcançadas. O **Capítulo 2** apresenta o referencial teórico utilizado neste trabalho, discutindo conceitos fundamentais como XAI e técnicas de explicabilidade como SHAP e LIME.

O **Capítulo 3** apresenta os trabalhos relacionados, discutindo as principais contribuições na área de visualização interativa para modelos de inteligência artificial explicáveis. São abordados estudos que exploram técnicas como SHAP, LIME e plataformas interativas, além de aplicações práticas com diferentes focos. Esses trabalhos fundamentam a proposta deste projeto ao identificar lacunas e oportunidades para o desenvolvimento de soluções inovadoras.

O **Capítulo 4** descreve os procedimentos metodológicos adotados, incluindo as

etapas de desenvolvimento do artefato e as estratégias de validação utilizadas e, em seguida, o cronograma da pesquisa.

Ao final, encontram-se as referências.

1.1 Objetivos

O objetivo geral deste projeto é desenvolver uma plataforma de visualização interativa que permita a explicação de modelos de AI, utilizando técnicas como SHAP para melhorar a interpretabilidade e acessibilidade dos modelos, possibilitando que usuários explorem e compreendam os impactos das variáveis nos resultados.

1.1.1 Objetivos específicos

Os objetivos específicos deste projeto são:

1. Implementar visualizações gráficas interativas que permitam a exploração dos impactos das variáveis nos modelos de AI.
2. Integrar a metodologia SHAP à plataforma para explicações baseadas em teoria dos jogos.
3. Facilitar a seleção e manipulação de dados pelos usuários na plataforma.
4. Avaliar a usabilidade da plataforma por meio de testes com usuários de diferentes áreas.
5. Validar a plataforma utilizando modelos de AI aplicados a diferentes setores, como saúde, computação e finanças.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta definições necessárias para compreender o tema do trabalho, bem como a solução proposta.

2.1 AI

Campo da ciência da computação que trata da criação de sistemas ou máquinas capazes de realizar tarefas que normalmente exigem inteligência humana (RUSSELL; NORVIG, 2016). Esses sistemas são projetados para simular capacidades cognitivas como aprendizagem, raciocínio, tomada de decisão, reconhecimento de padrões e resolução de problemas. A AI inclui subcampos como ML, *Natural Language Processing* (NLP), visão computacional e robótica.

2.2 XAI

A XAI refere-se a métodos que tornam os modelos de aprendizado de máquina mais interpretáveis, tanto para desenvolvedores quanto para usuários finais. Técnicas como LIME fornecem explicações locais, explicando previsões individuais ao destacar a influência de variáveis específicas (DWIVEDI et al., 2023). Por outro lado, SHAP fornece uma abordagem consistente baseada na teoria dos jogos, atribuindo valores *Shapley* a cada variável e medindo sua contribuição para a previsão final (MARCíLIO; ELER, 2020).

2.3 LIME

O LIME é uma técnica de explicabilidade de AI que visa interpretar modelos complexos (conhecidos como "caixa-preta"). Ele gera explicações locais, ou seja, explica previsões individuais ao aproximar o comportamento do modelo com um modelo mais simples (como uma regressão linear) em torno de um ponto específico. Isso ajuda a entender como cada variável influencia uma previsão específica, sem depender da arquitetura do modelo.

2.4 SHAP

O SHAP, introduzido por Lundberg e Lee (2017), é amplamente utilizado para fornecer explicações locais e globais em modelos complexos, como redes neurais e árvores de decisão (MARCíLIO; ELER, 2020). A técnica permite uma explicação robusta e detalhada, baseada em teoria dos jogos, sobre como cada variável contribui para a saída do modelo. Aplicações de SHAP são observadas em setores como a saúde, onde a compreensão das variáveis de um paciente é crucial para prever diagnósticos. (LUNDBERG; ERION; LEE, 2020)

2.5 Visualização Interativa em XAI

Ferramentas de visualização interativa têm sido desenvolvidas para facilitar a compreensão de modelos explicáveis de AI. Interfaces que permitem a manipulação de variáveis e observação de seus impactos em tempo real sobre os resultados do modelo são essenciais para explorar e entender o comportamento do sistema (OOGE; VERBERT, 2022; ZöLLER et al., 2023).

2.6 Proveniência de Dados e Transparência

A proveniência de dados é um elemento importante na explicação de modelos de AI. A rastreabilidade do ciclo de vida dos dados, desde sua origem até o uso no modelo, garante transparência e permite auditoria dos resultados (WERDER; RAMESH; ZHANG, 2022). A integração de dados de proveniência com XAI oferece uma visão mais ampla dos processos decisórios e possibilita maior confiança nos modelos.

2.7 Seleção de Características com SHAP

Além de fornecer explicações sobre as previsões, SHAP tem sido utilizado como uma ferramenta eficaz para a seleção de características. Marcílio e Eler (2020) demonstraram que os valores SHAP podem ser usados para identificar as variáveis mais relevantes de um modelo, otimizando seu desempenho. Essa técnica tem aplicações práticas em cenários com grandes volumes de dados, onde é necessário identificar e utilizar apenas as variáveis mais importantes.

2.8 IHC

A IHC é fundamental no desenvolvimento de plataformas de XAI. A usabilidade e acessibilidade das interfaces são essenciais para que usuários possam explorar os modelos de AI e entender seus processos decisórios (NAZAR et al., 2021). Em setores como a saúde, sistemas de IHC que integram explicações claras e acessíveis são especialmente importantes para a adoção de AI em ambientes de tomada de decisão (NAZAR et al., 2021).

3 TRABALHOS RELACIONADOS

Nesta seção, discutimos os principais trabalhos relacionados à visualização interativa aplicada à explicabilidade de modelos de AI. A visualização interativa em AI explicável tem como objetivo tornar os modelos de aprendizado de máquina mais acessíveis, permitindo que os usuários explorem, compreendam e ajustem os modelos de maneira intuitiva.

3.1 ExplAIner: Visualização Interativa e Explicações de IA

Spinner et al. (2019), em seu estudo, apresentam o *framework ExplAIner*, uma ferramenta de *visual analytics* que visa fornecer explicações interativas para modelos de aprendizado de máquina. A plataforma permite que os usuários explorem diferentes aspectos dos modelos, incluindo a importância das variáveis e o impacto de cada variável sobre as previsões. O *framework* utiliza gráficos interativos que facilitam a exploração dos resultados, tornando mais fácil para o usuário identificar as variáveis que mais influenciam o desempenho do modelo. A visualização é um elemento-chave para tornar os modelos explicáveis e acessíveis a públicos com diferentes níveis de experiência técnica. Sua robustez na apresentação de explicações interativas é altamente aproveitável. No entanto, ele não aborda profundamente a necessidade de personalização para públicos distintos e contextos variados.

3.2 Aplicações em Saúde com Visualização Interativa

Ferramentas de visualização interativa têm encontrado aplicações em áreas como a saúde, onde a explicabilidade de AI é fundamental para garantir a confiança nas decisões automáticas. O estudo de Lundberg, Erion e Lee (2020) aplica SHAP em um sistema clínico, utilizando visualizações interativas para explicar as previsões de um modelo que monitora pacientes durante a cirurgia. A visualização do impacto das variáveis foi crucial para que os profissionais de saúde pudessem entender como o modelo tomava decisões e para aumentar a confiança nas previsões de risco. A combinação de visualização interativa com técnicas de explicabilidade, como o SHAP, demonstra o valor dessas ferramentas em

cenários sensíveis. Todavia, ele não aborda interações diretas com os modelos para ajustes em tempo real.

3.3 Personalização e Flexibilidade nas Visualizações

O trabalho de Ooge e Verbert (2022) também destaca a importância da personalização em interfaces de visualização interativa. Ferramentas que permitem ajustar a interface às necessidades do usuário podem melhorar significativamente a acessibilidade e a utilidade dos modelos de AI. A flexibilidade nas visualizações permite que usuários de diferentes contextos e níveis de expertise, como cientistas de dados ou médicos, explorem os modelos de maneira personalizada, adaptando as explicações ao seu nível de entendimento e ao seu contexto de decisão. Contudo, ele não explora a possibilidade de interação entre usuários com diferentes níveis.

3.4 Visualização Interativa e Explicabilidade

O desenvolvimento de ferramentas de visualização interativa tem avançado significativamente, com a intenção de melhorar a compreensão dos usuários sobre os modelos de AI. Ooge e Verbert (2022) apresentam uma plataforma de visualização que oferece explicações personalizadas de modelos de AI, adaptando-se ao público-alvo e ao contexto de uso. A pesquisa sugere que interfaces visuais customizadas podem facilitar a compreensão de modelos complexos por diferentes tipos de usuários, como especialistas e leigos, ao fornecer *insights* sobre como o modelo toma suas decisões. Apesar de relevante, a pesquisa não se aprofunda na adaptação de explicações para públicos não técnicos em cenários de alta complexidade.

3.5 XAutoML: Visual Analytics para AutoML

Zöller et al. (2023), em seu trabalho, introduzem a plataforma *XAutoML*, que combina análise visual interativa com técnicas de *AutoML* (*Automated Machine Learning*). A ferramenta permite que os usuários visualizem e ajustem os parâmetros dos modelos de aprendizado de máquina ao longo do pipeline. O *XAutoML* é particularmente relevante para a explicabilidade, pois oferece uma maneira de visualizar cada etapa do processo de construção de modelos, permitindo uma exploração detalhada de como as variáveis e configurações impactam o desempenho do modelo. A integração de visualizações interativas com *AutoML* possibilita que os usuários explorem modelos complexos sem a necessidade de conhecimentos profundos em aprendizado de máquina. Ainda assim, o foco excessivo em pipelines automatizados pode limitar o controle de variáveis individuais pelos usuários.

4 METODOLOGIA

Este capítulo apresenta a classificação da pesquisa e as etapas metodológicas que norteiam o desenvolvimento do projeto, conforme a figura 1. A metodologia adotada neste projeto segue a abordagem de *Design Science Research* (DSR), detalhada em Peffers et al. (2007), que integra rigor científico e relevância prática para a construção e avaliação de artefatos tecnológicos. Este modelo metodológico é particularmente adequado, pois visa resolver problemas reais, promovendo a criação de soluções inovadoras fundamentadas em bases teóricas sólidas.

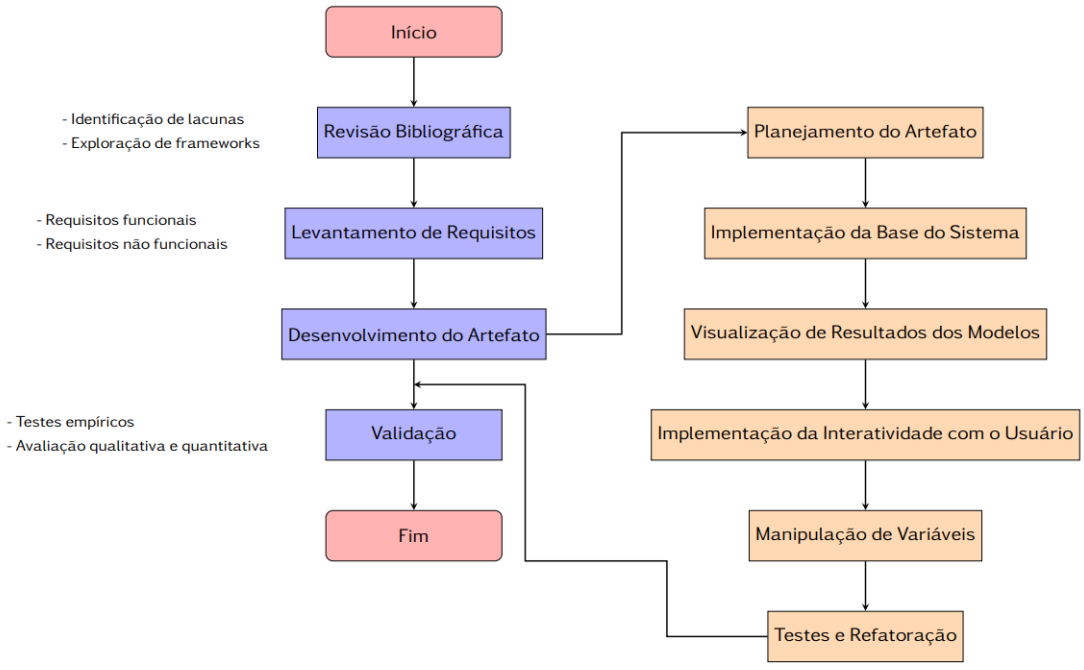

Figura 1 – Metodologia

A presente pesquisa é classificada como aplicada, uma vez que tem como objetivo o desenvolvimento de uma solução prática e inovadora para explicabilidade de modelos de AI. Seu caráter é explicativo, pois busca investigar como a visualização interativa e a manipulação de variáveis podem contribuir para a compreensão e confiança em modelos de AI. Quanto aos procedimentos metodológicos, ela combina revisão bibliográfica, para identificação de lacunas no estado da arte e análise de técnicas existentes, com aborda-

gem experimental, para validação do artefato em cenários práticos. A pesquisa utiliza métodos predominantemente quantitativos, como métricas de compreensão e confiança, mas também incorpora elementos qualitativos, como *feedback* dos participantes.

4.1 Atividades a serem realizadas

A execução do projeto segue um conjunto de etapas estruturadas, conforme descrito a seguir:

4.1.1 Atividade 1: Revisão Bibliográfica

A primeira etapa envolve a investigação das principais técnicas de XAI, como SHAP e LIME, além de *frameworks* de visualização interativa. Esta revisão visa identificar lacunas teóricas e práticas no estado da arte, fundamentando o desenvolvimento do artefato proposto.

4.1.2 Atividade 2: Levantamento de Requisitos

Nesta etapa, serão definidos os requisitos funcionais e não funcionais do sistema. Os requisitos funcionais incluem a visualização dos resultados dos modelos de AI e a manipulação de variáveis pelos usuários. Requisitos não funcionais abordarão aspectos como desempenho, acessibilidade e usabilidade, visando garantir uma experiência robusta e eficiente para os usuários. Com base nos requisitos levantados, será realizado um planejamento detalhado. Serão especificados os componentes do sistema, as ferramentas e linguagens a serem utilizadas, e os cronogramas de desenvolvimento. Esta etapa assegura o alinhamento entre os objetivos do projeto e a execução técnica.

4.1.3 Atividade 4: Desenvolvimento do Artefato

O desenvolvimento do artefato será dividido em subetapas:

- **Planejamento do Artefato:** Organização das atividades de desenvolvimento, incluindo a arquitetura do sistema e os recursos necessários.
- **Implementação da Base do Sistema:** Criação da infraestrutura básica, com estruturas de dados e mecanismos essenciais para integração com modelos de AI.
- **Visualização de Resultados dos Modelos:** Implementação de ferramentas para apresentar, de forma clara e interativa, o impacto das variáveis nos resultados dos modelos.

- **Interatividade com o Usuário:** Desenvolvimento de funcionalidades para manipulação de variáveis, permitindo aos usuários observar em tempo real os efeitos de suas mudanças sobre os resultados.
- **Manipulação de Variáveis:** Implementação de um módulo para que os usuários possam alterar parâmetros do modelo e compreender melhor seus impactos.
- **Testes e Refatoração:** Ao final do desenvolvimento, os componentes serão submetidos a testes empíricos para avaliar funcionalidade, desempenho e usabilidade. Caso sejam identificadas limitações, o sistema será refinado para atender aos requisitos estabelecidos.

4.1.4 Atividade 6: Validação

A etapa final consiste na validação do artefato em cenários práticos, com usuários de diferentes áreas, como saúde, finanças e computação. A validação incluirá métricas quantitativas, como taxa de compreensão e eficácia do modelo, além de análises qualitativas baseadas no *feedback* dos participantes.

4.2 Cronograma

Esta seção apresenta o cronograma das atividades planejadas para o projeto (Tabela 1).

Tabela 1 – Cronograma de Atividades do Projeto

Etapas	**Mês 1-2**	**Mês 3-4**	**Mês 5-6**	**Mês 7-8**	**Mês 9-10**	**Mês 11-12**
Revisão Bibliográfica	X					
Levantamento de Requisitos		X				
Desenvolvimento do Artefato		X	X			
Testes e Validação			X	X		
Comparação com o Estado da Arte				X	X	
Redação e Divulgação dos Resultados					X	X

REFERÊNCIAS

DWIVEDI, R. et al. Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys, v. 55, n. 9, p. Article 194, 33 pages, Sep. 2023.

LUNDBERG, S.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. 2017. Disponível em: <https://arxiv.org/abs/1705.07874>.

LUNDBERG, S. M.; ERION, G. G.; LEE, S. I. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, p. 1–17, 2020.

MARCíLIO, W. E.; ELER, D. M. From explanations to feature selection: Assessing shap values as feature selection mechanism. In: 33RD SIBGRAPI CONFERENCE ON GRAPHICS, PATTERNS AND IMAGES, 33., 2020, Porto de Galinhas, Brazil. **Proceedings...** Porto de Galinhas: IEEE, 2020. p. 340–347.

NAZAR, M. et al. A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access, v. 9, p. 153316–153348, Nov. 2021.

OOGE, J.; VERBERT, K. Explaining artificial intelligence with tailored interactive visualisations. In: 27TH INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES (IUI '22 COMPANION), 27., 2022, New York, USA. **Proceedings...** New York: ACM, 2022. p. 120–123.

PEFFERS, K. et al. A design science research methodology for information systems research. Journal of management information systems, Taylor & Francis, v. 24, n. 3, p. 45–77, 2007.

RUSSELL, S. J.; NORVIG, P. Artificial Intelligence: A Modern Approach. 3rd. ed. Upper Saddle River, NJ: Pearson Education, 2016.

SPINNER, T. et al. explainer: A visual analytics framework for interactive and explainable machine learning. IEEE Transactions on Visualization and Computer Graphics, 2019.

WERDER, K.; RAMESH, B.; ZHANG, R. Establishing data provenance for responsible artificial intelligence systems. ACM Transactions on Management Information Systems, v. 13, n. 2, p. Article 22, 23 pages, Jun. 2022.

ZöLLER, M.-A. et al. Xautoml: A visual analytics tool for understanding and validating automated machine learning. ACM Transactions on Interactive Intelligent Systems, v. 13, n. 4, p. Article 28, 39 pages, Dec. 2023.