Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Tese de Doutorado

Área: Computação Inteligente

Título: Modelos Inteligentes Explicáveis e Interpretáveis

Orientador(a): Roberta Andrade de A. Fagundes (roberta.fagundes@upe.br)

Descrição

A Inteligência Artificial (IA) [7] está se tornando cada vez mais presente em nossas vidas, impactando diversos setores da sociedade. No entanto, a falta de transparência e explicabilidade dos modelos de IA levanta preocupações sobre sua confiabilidade, justiça e equidade. É nesse contexto que a Inteligência Artificial Explicável (XAI) [1][2] surge como uma área crucial de pesquisa, buscando desenvolver métodos e ferramentas que permitam aos humanos entender como os sistemas de IA tomam decisões.

Esta proposta de doutorado visa desenvolver e avaliar métodos de XAI [3][4] para modelos de IA complexos, com foco em aplicações em áreas relevantes, como saúde, educação, finanças e justiça. O objetivo principal é contribuir para a construção de sistemas de IA mais confiáveis, transparentes e justos, que possam ser utilizados de forma responsável e ética.

Objetivos Específicos:

- **Revisar e analisar o estado da arte em XAI**, incluindo métodos, ferramentas e aplicações em diferentes áreas.
- **Desenvolver e implementar novos métodos de XAI** para modelos de IA complexos, como redes neurais profundas e modelos de aprendizado de máquina.
- **Avaliar o desempenho dos métodos de XAI** em termos de efetividade, eficiência e fidelidade às explicações.
- **Aplicar os métodos de XAI em casos de estudo** em áreas relevantes, como saúde, finanças e justiça.
- **Analisar os impactos éticos e sociais da XAI**, incluindo questões de transparência, responsabilidade e viés.

A pesquisa será conduzida em etapas, combinando pesquisa bibliográfica, desenvolvimento de métodos, implementação de software, avaliação experimental A pesquisa proposta contribuirá significativamente para o avanço da área de XAI [5][6], com o desenvolvimento de novos métodos, avaliação em casos de estudo e análise de impacto ético e social. Os resultados da pesquisa poderão auxiliar na construção de sistemas de IA mais confiáveis, transparentes e justos, beneficiando a sociedade como um todo e análise de casos de estudo.

Referências Bibliográficas:

[1] A. B. Arrieta, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58:82-115, 2020.

[2] Wojciech Samek, Gregor Weidmüller, Sebastian Kämecke, Tobias Bischoff, and Thomas Kohle. Explainable Artificial Intelligence: A Methodological and Conceptual Overview, 2019.

[3] Bohan Misra and Natalie Ruschmeier, Interpretable Machine Learning, 2022.

[4] F. Doshi-Velez and B. Kim. Towards a rigorous Science of interpretable machine learning, 2017.

[5] Riccardo Guidotti, Anna Maria Montagnani, Franco Tagliati, and Davide Pedreschi. A Survey of Explainable Machine Learning, 2018. (https://arxiv.org/abs/2011.07876)

[6] Sameer Singh, Yash Sharma, and Arijit Bose. Human-Centered AI Explainability: Practices, Challenges, and Promises, 2022.

[7] T. B. Ludermir. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. Estudos Avançados, v. 35, p. 85-94, 2021.