

Nota: Este material complementar, disponível em https://www.rettore.com.br/public_data/lectures/ representa uma cópia resumida de conteúdos bibliográficos disponíveis gratuitamente na Internet.

Inteligência Artificial Explicável

[Introdução](#)

[Conceitos](#)

[Ferramentas e Técnicas](#)

[Referências](#)

Introdução

A **Inteligência Artificial Explicável (XAI)** é um campo que busca criar sistemas de IA cujas decisões possam ser compreendidas por humanos. Em muitas aplicações de IA, como modelos de aprendizado profundo ou redes neurais, as decisões podem parecer como "caixas-pretas", difíceis de entender e interpretar. No entanto, para garantir a confiança, a ética e a adoção ampla dessas tecnologias, é essencial que os modelos sejam transparentes e explicáveis.

Conceitos

1. **Transparência:** Refere-se à clareza de como um sistema de IA funciona. Um modelo transparente é aquele cujas operações internas podem ser facilmente inspecionadas e compreendidas.
2. **Interpretabilidade:** É a capacidade de entender a lógica por trás da decisão de um modelo. Modelos interpretáveis permitem que humanos sigam o raciocínio utilizado pela IA e identifiquem os fatores que levaram a uma determinada saída.
3. **Explicabilidade:** Vai além da simples interpretação. Ela envolve fornecer explicações compreensíveis para os usuários sobre como e por que a IA chegou a uma conclusão específica. Essas explicações devem ser adaptadas ao nível de entendimento do público-alvo, seja um especialista técnico ou um usuário final.

Ferramentas e Técnicas

- **LIME (Local Interpretable Model-agnostic Explanations):** Uma técnica que gera explicações locais para modelos de IA complexos, aproximando-os de modelos simples, como regressões lineares, para facilitar o entendimento de uma decisão específica.
- **SHAP (SHapley Additive exPlanations):** Baseado na teoria dos jogos, o SHAP atribui valores (ou pesos) a cada característica de entrada do modelo, mostrando quanto cada uma delas contribuiu para a decisão final. É uma ferramenta muito útil para entender a importância de variáveis em um modelo complexo.
- **Redes Neurais Interpretáveis:** São arquiteturas de redes neurais desenvolvidas para que suas decisões sejam mais compreensíveis. Um exemplo são as redes neurais modulares ou redes simplificadas, que buscam melhorar a transparência sem sacrificar a precisão.

Referências

1. Da “Caixa-Preta” à “Caixa de Vidro”: o Uso da Explainable Artificial Intelligence (XAI) para Reduzir a Opacidade e Enfrentar o Enviesamento em Modelos Algorítmicos
<https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973/pdf>
2. Inteligência Artificial Uma Abordagem Moderna - 4ª Edição (Versão Inglês)
3. Explaining Explanations: An Overview of Interpretability of Machine Learning.
<https://arxiv.org/abs/1806.00069>
4. Conteúdo gratuito disponível na Internet:
https://www.youtube.com/watch?v=LCJ_PfmeO5Q;
<https://www.youtube.com/watch?v=OS5iRU0ZhWY>

Isenção de Responsabilidade:

Os autores deste documento não reivindicam a autoria do conteúdo original compilado das fontes mencionadas. Este documento foi elaborado para fins educativos e de referência, e todos os créditos foram devidamente atribuídos aos respectivos autores e fontes originais.

Qualquer utilização comercial ou distribuição do conteúdo aqui compilado deve ser feita com a devida autorização dos detentores dos direitos autorais originais. Os compiladores deste documento não assumem qualquer responsabilidade por eventuais violações de direitos autorais ou por quaisquer danos decorrentes do uso indevido das informações contidas neste documento.

Ao utilizar este documento, o usuário concorda em respeitar os direitos autorais dos autores originais e isenta os compiladores de qualquer responsabilidade relacionada ao conteúdo aqui apresentado.