Explainable AI: Explicando as predições e obtendo insights dos modelos de Machine Learning

Milton Gama Neto

Um pouco sobre mim

Milton Gama Neto

Recifense

com muito orgulho <3

Formado em Ciência da Computação na UFPE

curso que escolhi fazer desde criança (melhor escolha)

Estudante de mestrado da UFPE

pesquisando sobre Explainable AI

Data Scientist e Tech Lead no CESAR

há um ano trabalhando remotamente

Apaixonado por inovação, tecnologia e dados

gosto muito de participar de desafios e eventos de tech

Machine Learning

UM BREVE RESUMO

- Subárea da Inteligência Artificial
- Os programas aprendem comportamentos sem precisar ser explicitamente programados
- Padrões identificados através de dados

ML de forma resumida

$$f(X) = \hat{y}$$

f é uma função (modelo)
X são os dados
$\hat{y}$ são as predições

Modelos Black-box

São sistemas que não revelam seus mecanismos internos. Em ML, os modelos black-box (BB) são complexos e não fornecem explicações de suas decisões e dos parâmetros e estruturas aprendidas.

Entretanto costumam ter alto poder de predição.

Exemplos de Black-box:

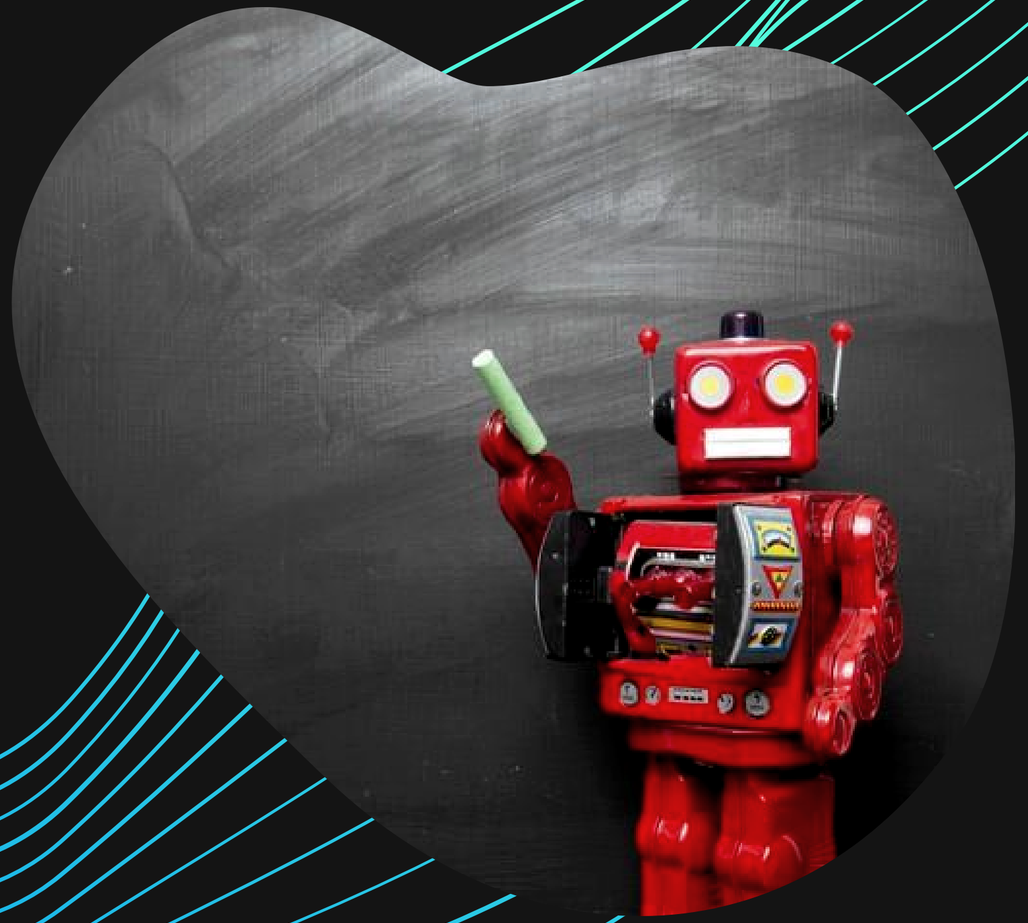
- Redes Neurais Artificiais
- Deep Learning
- Grandient Boosting
- Ensembles

Exemplos de White-Box:

- *Regressão Linear*
- *Árvore de Decisão*

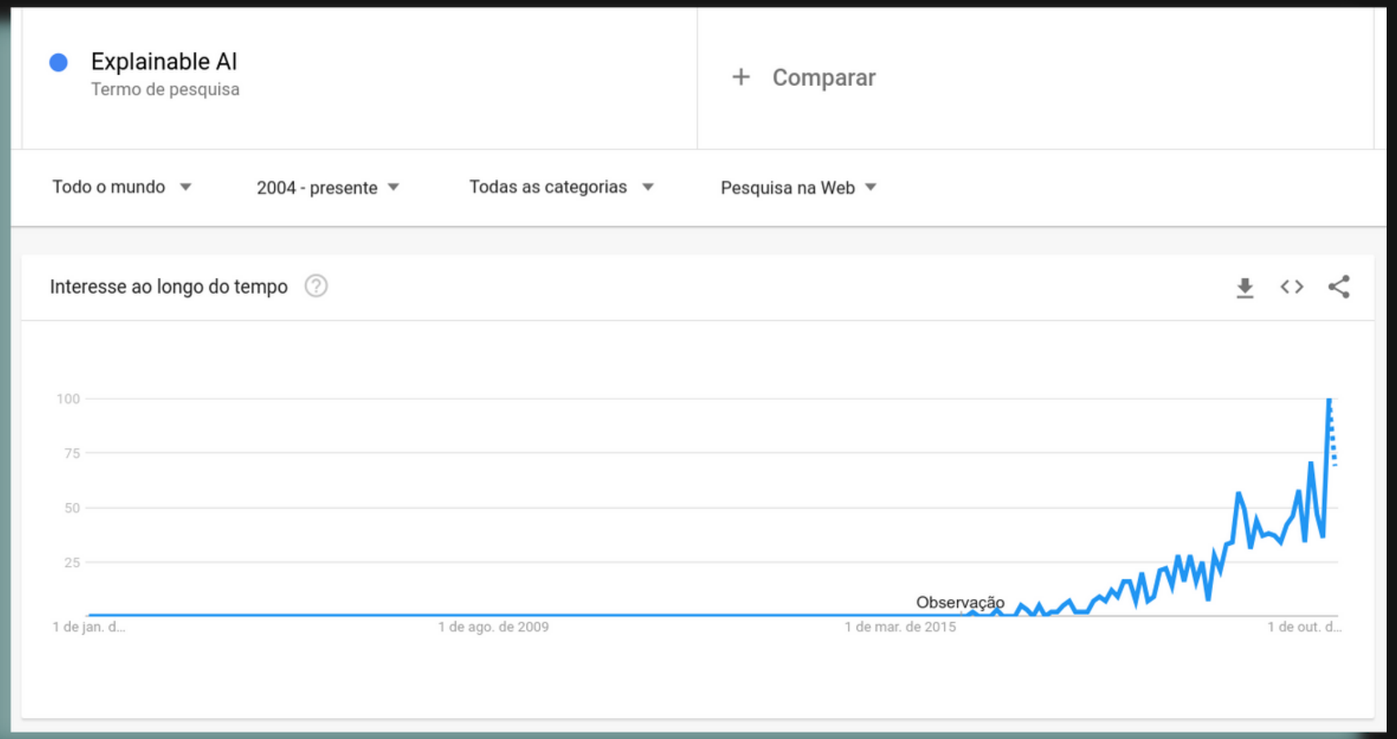
Explainable AI

eXplainable Artificial Intelligence (XAI) é um campo de pesquisa que visa tornar os resultados dos sistemas de IA mais compreensíveis para humanos, tornar os mecanismos internos mais transparentes, provendo explicações para suas decisões em alguns níveis de detalhes.



Tendência

CRESCIMENTO ACELERADO



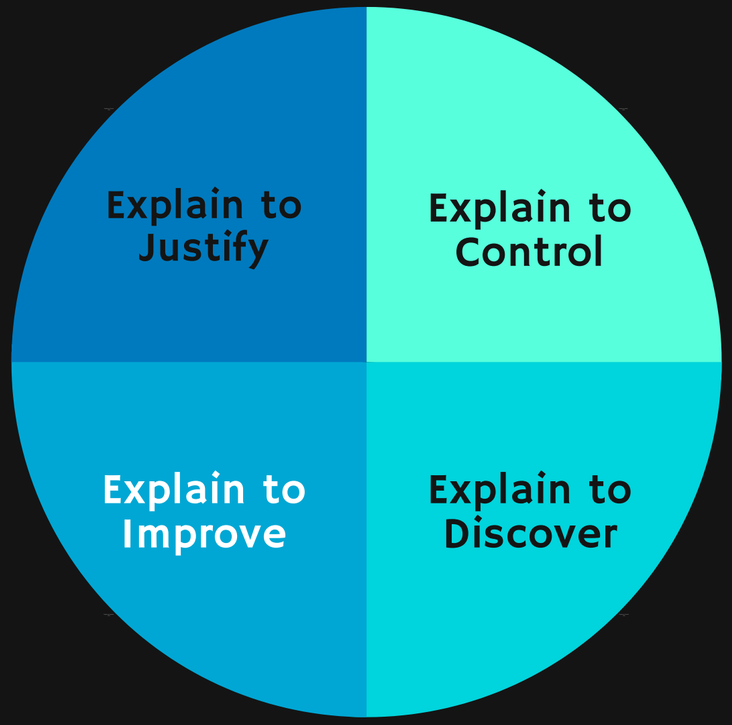
Popularidade no Google Trends

XAI word cloud [1]

Motivação

IMPORTÂNCIA E NECESSIDADE

- Detectar vieses e preconceito;
- Garantir a atuação de variáveis úteis;
- Entender o comportamento do modelos e das predições;
- Atender os domínios que necessitam de explicações das decisões;
- Atender questões regulatórias (LGPD, GDPR, etc);
- Utilizar modelos complexos e conseguir alta interpretabilidade;
- Aumentar robustez e confiança das soluções;



Razões para XAI [1]

[1] Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6 (2018): 52138-52160.

Abordagens

Local x Global

White (intrínseca) x Post-hoc

Specific model x Agnostic-model

Data types

Abordagens

LOCAL X GLOBAL

Local

- Explica o comportamento da decisão para uma instância individualmente. Ou seja, para cada predição.
- Ajuda a entender os fatores que levaram a predizer determinado valor.
- Entendimento mais específico.

Global

- Explica o comportamento geral do modelo todo, a lógica que ele segue.
- Ajudar a entender os principais fatores que influenciam as decisões do modelo.
- Entendimento mais genérico.

Abordagens

SPECIFIC MODEL X AGNOSTIC-MODEL

Specific model

- Funciona para algum modelo específico ou família de modelos (exemplo: para redes neurais artificiais).
- Voltado para alguma estrutura específica.

Agnostic-model

- Funciona para qualquer modelo preditivo.
- Voltado para as saídas dos modelos, utilizando os dados de entrada e/ou funções de predições.

Abordagens

WHITE (INTRÍNSECA) X POST-HOC

White (intrínseca)

- Objetivo de criar modelos interpretáveis e de fácil entendimento.
- Conhecido como White-Box por gerar modelos transparentes.

Post-hoc

- Objetivo de explicar as decisões de um modelo já treinado.
- Técnicas para explicar os modelos.

Prática

Exemplos com notebooks

Obrigado!

Milton Gama Neto

EMAIL

miltongamaneto@gmail.com

LINKEDIN

linkedin.com/in/milton-gama-neto/

GITHUB

github.com/miltongneto
github.com/miltongneto/Explainable-AI

MEDIUM

medium.com/@miltongamaneto