PRADO VIDIGAL

Discriminação Racial e Inteligência Artificial: como mitigar vieses algorítmicos na prática

21 DE MARÇO

DIA INTERNACIONAL PELA ELIMINAÇÃO
DA DISCRIMINAÇÃO RACIAL

A ORIGEM DESTA DATA

Sharpeville, 1960: uma data que transformou a luta por direitos humanos

Em 21 de março de 1960, o massacre de Sharpeville, na África do Sul, deixou 69 mortos durante um protesto contra o apartheid. O episódio levou a ONU a instituir a data como o **Dia Internacional pela Eliminação da Discriminação Racial**

Décadas depois, a luta por equidade racial também passa pela **Inteligência Artificial**. Evitar vieses exige:

- Governança
- Dados representativos
- Monitoramento contínuo

Bibliotecas open-source para auditar e corrigir viés em modelos de ML

FAIRNESS END-TO-END

AI Fairness 360 ▶ IBM | Linux Foundation AI

70+ métricas de fairness e 10+ algoritmos de mitigação em 3 estágios: pré-processamento (rebalanceamento de dados), in-processing (regularização do modelo durante o treinamento) e pós-processamento (ajuste de limiares por grupo demográfico). Compatível com scikit-learn e TensorFlow.

MITIGAÇÃO COM RESTRIÇÕES DE EQUIDADE

Fairlearn ▶ Microsoft | Open Source

O algoritmo ExponentiatedGradient treina modelos com restrições de equidade (demographic parity, equalized odds). O ThresholdOptimizer ajusta limiares de decisão pós-treinamento. Inclui dashboard interativo para visualização de disparidades entre grupos demográficos.

AUDITORIA DE VIÉS

Aequitas ▸ University of Chicago | Center for Data Science and Public Policy

Framework de auditoria com interface web e Python. Avalia disparidades por grupo em taxas de falsos positivos e falsos negativos. O módulo Aequitas Flow integra fairness diretamente em pipelines de ML, automatizando a avaliação contínua ao longo do ciclo de vida do modelo.

Transparência nas decisões e controle de modelos de linguagem

EXPLICABILIDADE

SHAP + LIME ▸ Open Source

SHAP (Shapley Additive Explanations): atribui a contribuição de cada variável a cada predição individual, com fundamentação na teoria dos jogos. LIME (Local Interpretable Model-Agnostic Explanations): gera explicações locais para qualquer modelo. Combinados, permitem comparar explicações entre grupos demográficos e identificar proxies raciais.

DOCUMENTAÇÃO TRANSPARENTE

Model Cards + Datasheets ▸ Google | Gebru et al.

Model Cards (Mitchell et al., 2019): documentam uso pretendido, métricas desagregadas por grupo demográfico e limitações éticas. Datasheets for Datasets (Gebru et al., 2021): protocolo com 7 seções sobre composição, coleta e potenciais vieses do dataset. Instrumentos essenciais de accountability.

GOVERNANÇA DE LLMS

NeMo Guardrails + Garak ▸ NVIDIA

NeMo Guardrails: utiliza a linguagem Colang para definir rails programáveis — moderação de conteúdo, controle de tópicos e prevenção de outputs discriminatórios. Garak (v0.10): scanner de vulnerabilidades com probes estáticos, dinâmicos e adaptativos para testar viés e toxicidade em Large Language Models.

Diversidade e inovação **avançam juntas**

PRADO VIDIGAL