APRENDIZADO DE MÁQUINA PARA SAÚDE

Ética, Explicabilidade e Privacidade na IA em Saúde

Como auditar modelos, reduzir injustiças e proteger dados clínicos

APOSTILA TÉCNICA · ENCONTRO 14

Prof. Flávio Luiz Seixas
Pós-graduação em Computação

Sumário

Introdução

A avaliação de IA em saúde não termina na acurácia. É preciso compreender para quem o sistema funciona, que explicação oferece, como trata dados sensíveis e quais danos podem emergir da interação entre modelo, instituição e usuário. [1]

O encontro conecta explicabilidade, equidade, privacidade e segurança como dimensões de governança durante todo o ciclo de vida. [1]

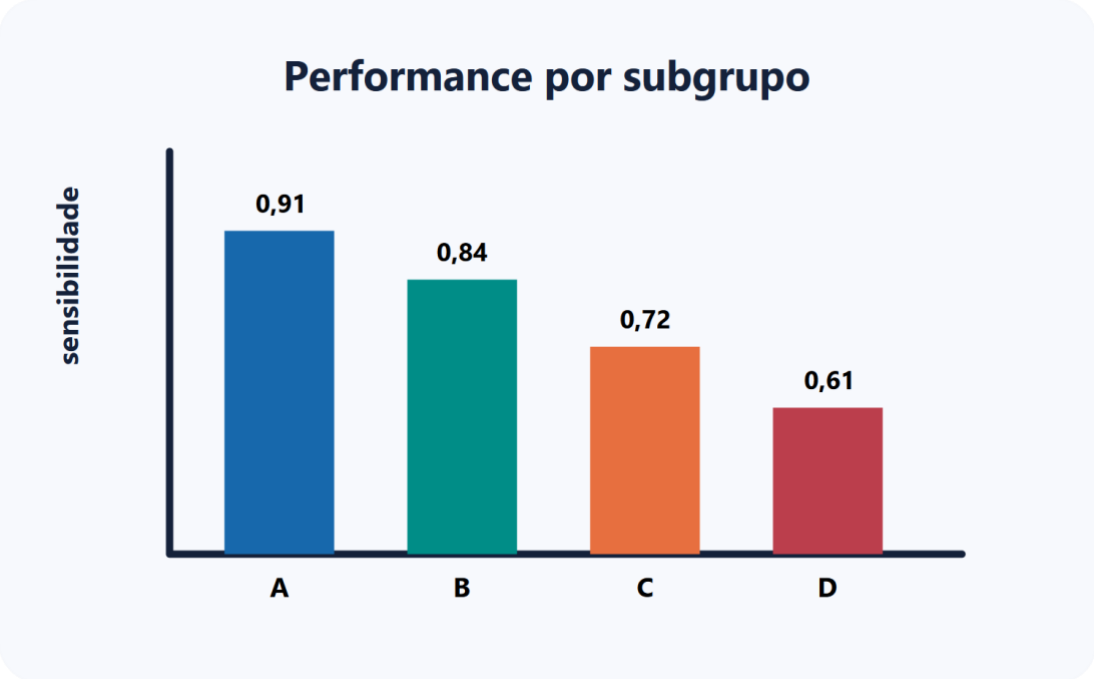

Figura 1 — Auditoria de desempenho por subgrupos. Fonte: elaboração própria com base em [2].

1. Interpretabilidade e explicabilidade

Modelos complexos mapeiam preditores para saídas por relações difíceis de representar em raciocínio humano. Explicabilidade reúne métodos para tornar aspectos desse processo acessíveis. A finalidade deve ser definida: compreensão global, justificativa local, depuração, auditoria ou comunicação clínica exigem evidências diferentes [1].

Métodos locais atribuem contribuição a variáveis para uma previsão; mapas de saliência destacam regiões de imagem associadas à saída. Essas representações não provam causalidade nem garantem que o modelo use mecanismos clínicos válidos. Precisam ser avaliadas quanto a fidelidade, estabilidade e utilidade [1].

CONCEITO-CHAVE
Modelos complexos mapeiam preditores para saídas por relações difíceis de representar em raciocínio humano.

2. Equidade e fontes de viés

Viés pode ser introduzido na formulação, seleção de dados, construção e implantação, além da interpretação humana. Equidade não equivale a repartir erros de forma mecanicamente igual: requer explicitar benefícios, danos, grupos afetados e contexto [2].

Auditoria deve desagregar desempenho e calibração, verificar cobertura, qualidade de rótulo e proxies. Uma variável aparentemente neutra pode representar acesso, utilização ou prática institucional e reproduzir desigualdade quando tomada como necessidade clínica [2].

3. Privacidade, confidencialidade e segurança

Privacidade diz respeito ao controle e ao uso apropriado de informações pessoais; confidencialidade envolve obrigações de quem recebe os dados; segurança reúne controles que protegem confidencialidade, integridade e disponibilidade [3].

Remover identificadores diretos não elimina todo risco de reidentificação. Governança deve combinar minimização, autorização, controle de acesso, registro, retenção e resposta a incidentes. Segurança cibernética afeta segurança do paciente quando disponibilidade ou integridade de sistemas clínicos é comprometida [3].

4. Colaboração distribuída e vigilância

Aprendizado federado mantém dados localmente e coordena atualização de um modelo comum. Essa arquitetura reduz centralização, mas ainda requer avaliar vazamento por atualizações, heterogeneidade entre instituições e governança [4].

A vigilância pós-implantação deve observar desempenho, subgrupos, incidentes e mudanças de uso. Efeitos éticos não são propriedades fixas do arquivo do modelo; emergem no sistema sociotécnico e podem mudar com o tempo [2].

5. Governança de IA generativa

IA generativa acrescenta riscos de alucinação, perda de contexto e comportamento mais autônomo. Um checklist recente organiza salvaguardas por pré-desenvolvimento, desenvolvimento, pós-implantação e temas transversais, incluindo controle de acesso, verificação especializada, robustez a prompts, limites de decisão, humano-no-loop para alto risco e plano de monitoramento contínuo [5].

Responsabilidade precisa ser distribuída antes do incidente: instituição, profissionais, pesquisadores e desenvolvedores devem ter deveres documentados. O estudo de consenso encontrou maior convergência sobre deveres institucionais e profissionais do que sobre a atribuição ao desenvolvedor, mostrando que ausência de consenso não autoriza deixar a governança indefinida [5].

Para LLMs, RAG pode ancorar respostas em literatura ou diretrizes, mas a qualidade da explicação depende da recuperação, da fidelidade e da compreensão pelo usuário. Uma explicação plausível ou longa não substitui validação prospectiva, integração ao fluxo e avaliação de desfechos [6].

LIMITAÇÃO DOCUMENTAL

A cobertura jurídica detalhada de LGPD e GDPR permanece limitada; a revisão atualiza governança de IA generativa, mas não substitui interpretação legal das normas.

Síntese do encontro

Interpretabilidade e explicabilidade: modelos complexos mapeiam preditores para saídas por relações difíceis de representar em raciocínio humano.

Equidade e fontes de viés: viés pode ser introduzido na formulação, seleção de dados, construção e implantação, além da interpretação humana.
Privacidade, confidencialidade e segurança: privacidade diz respeito ao controle e ao uso apropriado de informações pessoais; confidencialidade envolve obrigações de quem recebe os dados; segurança reúne controles que protegem confidencialidade, integridade e disponibilidade [3].
Colaboração distribuída e vigilância: aprendizado federado mantém dados localmente e coordena atualização de um modelo comum.
Governança de IA generativa: ia generativa acrescenta riscos de alucinação, perda de contexto e comportamento mais autônomo.

Referências

[1] Li, R. C. et al. Explainability in Medical AI. In: Intelligent Systems in Medicine and Health. Springer, 2022.

[2] Korngiebel, D. M.; Solomonides, A.; Goodman, K. W. Ethical and Policy Issues. In: Intelligent Systems in Medicine and Health. Springer, 2022.

[3] DeMuro, P.; Norwood, H. Privacy and Security. In: Health Informatics. 2021.

[4] Cohen, T. A.; Patel, V. L.; Shortliffe, E. H. Reflections and Projections. In: Intelligent Systems in Medicine and Health. Springer, 2022.

[5] Cha, H. et al. Development of research ethics guidelines for healthcare generative artificial intelligence: deriving expert consensus through a Delphi study. BMC Medical Ethics, 2026.

[6] Zamanifar, A.; Taherkordi, A.; Farhadi, A. (eds.). Explainable Large Language Models in Healthcare Applications. Springer, 2026.