# IA Aplicada à Saúde

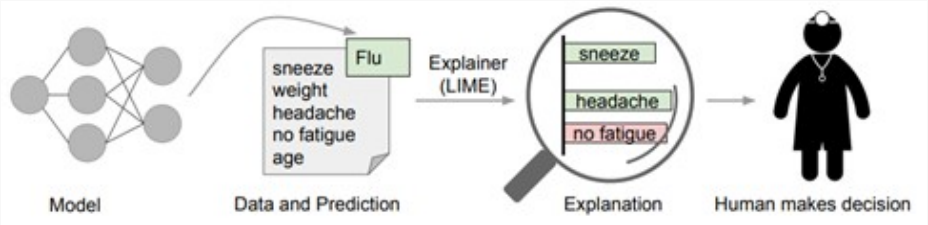
UNIDADE 05

Explainable IA na Saúde

## Explainable IA na Saúde

Nas semanas anteriores você entrou em contato com diversas técnicas de IA para solução de problemas de saúde, entre elas, os algoritmos de *Machine Learning* (ML), que tem enorme potencial na melhoria de processos clínicos e de pesquisa.

Todavia, uma das **barreiras para sua adoção**, em sistemas de apoio a decisão clínica por exemplo, é uma **falta de transparência quanto às predições dos modelos**, ou seja, a carência de indicação acerca de quais características levaram o sistema a tomar a decisão. Este aspecto é especialmente importante na área da saúde, onde a ferramenta de IA deve trabalhar em conjunto com um corpo médico, que por sua vez emite uma decisão final sobre a condição e tratamento do paciente (conforme exemplo da Figura 1).




**Fonte:** RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 2016. p. 1135–1144. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939778

Um modelo prevê que um paciente está gripado e uma ferramenta de explicação de ML, chamada LIME, destaca os sintomas na história do paciente que levaram à previsão. Espirro e dor de cabeça são retratados como contribuintes para a previsão da gripe (em verde), enquanto "sem fadiga" é uma evidência contra isso (em vermelho). Com isso, um médico pode tomar uma decisão informado sobre se deve confiar na previsão do modelo.

Uma série de algoritmos de ML já podem ser considerados "interpretáveis" por já fornecerem, por padrão, evidências de suas predições como as árvores de decisão e regressão linear. A árvore de decisão pode plotar visualmente uma árvore que indica de maneira clara como o modelo chegou em sua predição. Já a regressão linear pode explicar suas decisões ao plotar a linha de regressão utilizada para efetuar sua saída, conforme este exemplo, em que há uma correlação entre peso e altura, que nos permite predizer o peso de um paciente.

O grande problema reside nos modelos que chamamos de "**caixas preta**", como as **redes neurais**, em que a falta de atributos explícitos e a complexidade dos pesos distribuídos pela rede, dificultam a interpretação dos resultados. Nestes casos, técnicas adicionais devem ser executadas na tentativa de explicar as predições (falaremos um pouco mais sobre isso depois).

## Devemos desenvolver apenas modelos interpretáveis/explicáveis?

No editorial "*Should Health Care Demand Interpretable Artificial Intelligence or Accept Black Box Medicine?*", publicado pelo "*American College of Physicians*", algumas ponderações são feitas acerca do assunto, conforme trechos traduzidos a seguir.

> *Muitos médicos se sentem desconfortáveis modelos caixa preta, mesmo que ele atinja altos graus de precisão diagnóstica ou prognóstica, e tem havido pedidos para mais pesquisas sobre como esses modelos funcionam. Embora os médicos provavelmente prefiram modelos intrinsecamente interpretáveis acreditamos que os modelos de caixa preta desempenharão um papel importante na medicina e, em muitos casos, não são tão diferentes de outras áreas nas quais não temos conhecimento biológico ou clínico. Os médicos nem sempre podem explicar por que chegaram a um determinado diagnóstico. Muitos medicamentos eficazes - aspirina, paracetamol e penicilina - foram amplamente utilizados por décadas antes de seu mecanismo de ação ser compreendido. Ainda não está claro como a terapia eletroconvulsiva ou os inibidores seletivos da recaptação da serotonina funcionam. Devemos então, manter a IA em um padrão explicativo mais elevado do que medicamentos e médicos?*

Parece sensato exigir níveis de explicabilidade distintos de diferentes tipos de IA. Pode haver expectativas de explicabilidade e requisitos regulatórios mais baixos para modelos que tratam de tarefas discretas e conhecidas, em que há probabilidade relativamente baixa de incorporar preconceitos sociais invisíveis e em que a saída representa apenas uma parte de uma avaliação clínica maior onde os médicos humanos podem intervir prontamente. Portanto, podemos estar relativamente confortáveis com o uso de modelos de caixa preta para análise de imagens, testes de laboratório ou processamento de linguagem natural de notas clínicas. Por outro lado, podemos ficar menos confortáveis com o uso de modelos de caixa preta para problemas abstratos ou inexplorados, nos quais a probabilidade de viés é alta, ou quando tais modelos são os principais motivadores das decisões de diagnóstico e tratamento - como pode ser o caso de alguns tipos de apoio à decisão clínica, estratificação de risco e distribuição de recursos escassos.

## Quais as vantagens de modelos interpretáveis/explicáveis?

Os benefícios de modelos de ML explicáveis são vários, e sempre devemos ter em mente que a tendência é contarmos cada vez com mais soluções que forneçam mecanismos de interpretação para aumento da confiança no uso da tecnologia. A seguir uma série de aspectos importantes relacionados a utilização de modelos explicáveis.

- Entendimento do comportamento geral dos modelos, e principalmente, de suas predições
- Garantia de utilização de atributos e variáveis úteis para predições
- Detecção de viés, falhas, preconceito e vulnerabilidades no modelo
- Fornecimento de insights que podem alimentar novas soluções
- Aumento na confiança das aplicações
- Atendimento a questões regulatórias

## Qual a diferença entre Interpretabilidade e Explicabilidade?

Quando pesquisamos sobre o tema abordado nesta unidade, nos deparamos com vários termos similares como “interpretável”, “explicável”, “inteligível”, “transparente”, “entendível”, etc. Alguns autores fazem um distinção entre os termos, conforme apresentado a seguir.

A **Interpretabilidade** ocorre quando humanos podem entender a causa e efeito, a entrada e saída, de um modelo de ML. Dizer que um modelo tem um alto nível de interpretabilidade significa que você pode descrever de uma forma interpretável por humanos sua inferência. Em outras palavras, por que uma entrada para um modelo produz uma saída específica? Quais são os requisitos e restrições dos dados de entrada? Quais são os limites de confiança das previsões? Ou, por que uma variável tem um efeito mais substancial do que outra? Para fins de interpretabilidade, detalhar como um modelo funciona só é relevante na medida em que pode explicar suas previsões e justificar que é o modelo certo para o caso de uso.

No exemplo citado anteriormente, você poderia explicar que existe uma relação linear entre a altura e o peso humanos, portanto, o uso de regressão linear em vez de um modelo não linear faz sentido. Você pode provar isso estatisticamente porque as variáveis envolvidas não violam os pressupostos da regressão linear.

A **Explicabilidade** abrange tudo o que a interpretabilidade é. A diferença é que ele vai mais fundo no requisito de transparência do que na interpretabilidade, porque exige explicações amigáveis para o funcionamento interno de um modelo e o seu processo de treinamento, e não apenas a sua inferência. Dependendo da aplicação, esse requisito pode se estender a vários graus de modelo (i.e., explicar passo a passo como funciona o treinamento), design (i.e., explicar as escolhas de arquitetura e hiperparâmetros) e outros.

Portanto, alcançarmos a completa explicabilidade dos modelos é muito difícil, pois os algoritmos caixa preta geralmente “perdem” uma série de atributos e parâmetros durante o treinamento. Além disso, muitos destes algoritmos (i.e., redes neurais) iniciam seus pesos aleatoriamente, não possibilitando a reprodução de experimentos.

## Quais são as abordagens existentes?

As abordagens para explicação dos modelos se diferenciam em função da técnica e objetivo, conforme as categorias apresentadas a seguir.

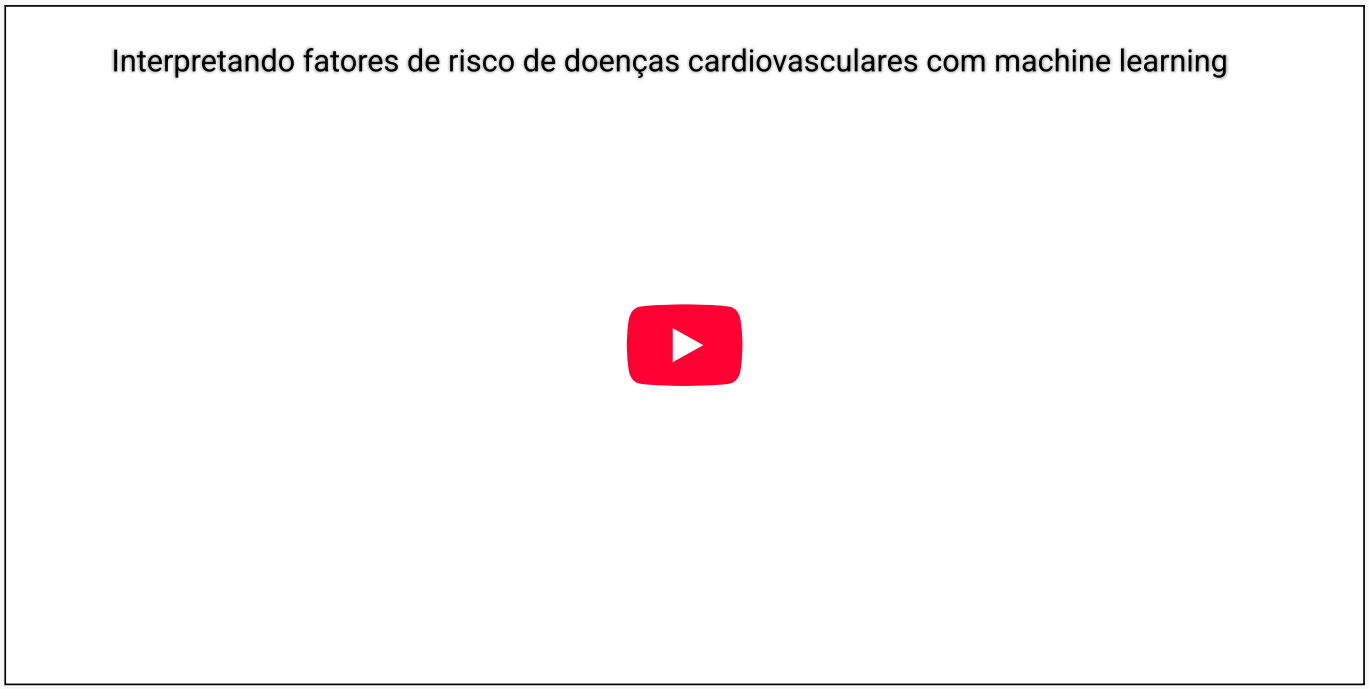
**Global vs. Local**: esta classificação diz respeito ao nível de interpretação, sendo **global** quando explicamos o comportamento geral do modelo, qual sua lógica e principais aspectos de funcionamento. Já o **local** é caracterizado quando explicamos o comportamento do modelo para uma instância do *dataset*, ou seja, uma predição específica (como no exemplo mostrado na Figura 1).

**Transparente vs. Post-hoc**: aqui definimos quando a explicação do modelo é gerada. O método **transparente**, que pode também ser chamado de intrínseco ou caixa branca, fornece interpretações na própria estrutura da arquitetura escolhida, como as árvores de decisão ou modelos lineares, por exemplo. Quando falamos de **post-hoc** estamos abordando técnicas aplicadas no modelo após seu treinamento, para assim, gerar as explicações.

**Específico vs. *Model-agnostic***: esta classificação diz quais modelos a técnica de explicação suportará. Em técnicas **específicas**, a explicação é desenvolvida para um modelo ou família de algoritmos específica (e.g., somente para redes neurais). Isto ocorre, pois, algumas técnicas utilizam-se de estruturas internas de um tipo de arquitetura para extrair informações. Enquanto o ***model-agnostic*** permite a aplicação em qualquer modelo preditivo.

## Interpretando fatores de risco de doenças cardiovasculares com Machine Learning?

Nesta videoaula nós aplicaremos conceitos de Explainable AI em um modelo que manipula fatores de risco de doenças cardiovasculares.




## Referências Bibliográficas

COLICCHIO, T. K. Introdução à informática em saúde: fundamentos, aplicações e lições aprendidas com a informatização do sistema de saúde americano. Porto Alegre: Artmed, 2020. [Minha Biblioteca].

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: KDD ’16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, San Francisco California USA. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 2016. p. 1135–1144. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939778. Acesso em: 01 jan. 2021.

MASÍS, Serg. Interpretable machine learning with Python: learn to build interpretable high-performance models with hands-on real-world examples. Birmingham: Packt, 2021.

WANG, Fei; KAUSHAL, Rainu; KHULLAR, Dhruv. Should Health Care Demand Interpretable Artificial Intelligence or Accept “Black Box” Medicine? Annals of Internal Medicine, [s. l.], v. 172, n. 1, p. 59, 2020.

PAYROVNAZIRI, Seyedeh Neelufar et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association, [s. l.], v. 27, n. 7, p. 1173–1185, 2020.

CRAIG, Erin; ARIAS, Carlos; GILLMAN, David. Predicting readmission risk from doctors’ notes. arXiv:1711.10663 [stat], [s. l.], 2017. Disponível em: http://arxiv.org/abs/1711.10663. Acesso em: 01 jan. 2021.