INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

eXplainable AI XAI

Profa. Anita Fernandes

Contextualização

No cenário atual da inteligência artificial, modelos complexos, frequentemente referidos como "caixas pretas", dominam diversas aplicações, desde diagnósticos médicos até sistemas de recomendação financeira.

Embora esses modelos demonstrem alta performance em suas tarefas, a falta de transparência em seu processo decisório levanta preocupações significativas em termos de confiança, responsabilidade, justiça e conformidade regulatória.

É nesse contexto que surge a Inteligência Artificial Explicável (XAI - Explainable Artificial Intelligence), um campo emergente que busca tornar os sistemas de IA mais compreensíveis e transparentes para os seres humanos.

Contextualização

A IA tradicional muitas vezes operava como uma “caixa preta”. No mundo da IA, as “caixas pretas” representam sistemas em que o funcionamento interno permanece oculto aos olhos dos usuários. Basicamente, esses sistemas são alimentados por uma variedade de **dados**.

É nesse contexto que a **XAI** é uma evolução, ao abrir essa caixa preta para explicar as etapas e as decisões tomadas durante todo seu processo.

Recentemente um estudo da MarketsandMarkets, plataforma de inteligência competitiva e pesquisa de mercado, projetou que o tamanho global do mercado de XAI deve crescer de US$ 6,2 bilhões em 2023 para US$ 16,2 bilhões até 2028, a uma taxa composta de crescimento anual (CAGR) de 20,9%.

Contextualização
Empresas Pioneiras da XAI

A **Epic** firmou parceria estratégica com a **Microsoft**, com o objetivo de integrar a tecnologia generativa de IA ao domínio da saúde.

Esta colaboração ampliada aproveitou os recursos do serviço **Azure OpenAI** e do software de registro eletrônico de saúde (EHR) amplamente reconhecido da Epic, com o objetivo de oferecer as vantagens da IA ao setor de saúde.

Epic

Microsoft

Contextualização
Empresas Pioneiras da XAI

A **NVIDIA** anunciou uma colaboração com a Microsoft para acelerar a IA generativa pronta para empresas.

A colaboração envolve a integração do software NVIDIA AI Enterprise ao Azure Machine Learning da Microsoft para ajudar as empresas a criar modelos de IA.

Azure Machine Learning

Contextualização
Empresas Pioneiras da XAI

O **Instituto SAS** colaborou com o **Erasmus MC** para desenvolver um algoritmo de IA para prever se os pacientes deveriam permanecer no hospital após a cirurgia oncológica ou poderiam ser dispensados com segurança.

Erasmus MC

Conceito

XAI é um conjunto de processos e métodos que permite aos usuários humanos entenderem e confiarem nos resultados e produção criados por algoritmos de aprendizado de máquina.

A IA explicável é usada para descrever um modelo de IA, seu impacto esperado e potenciais vieses.

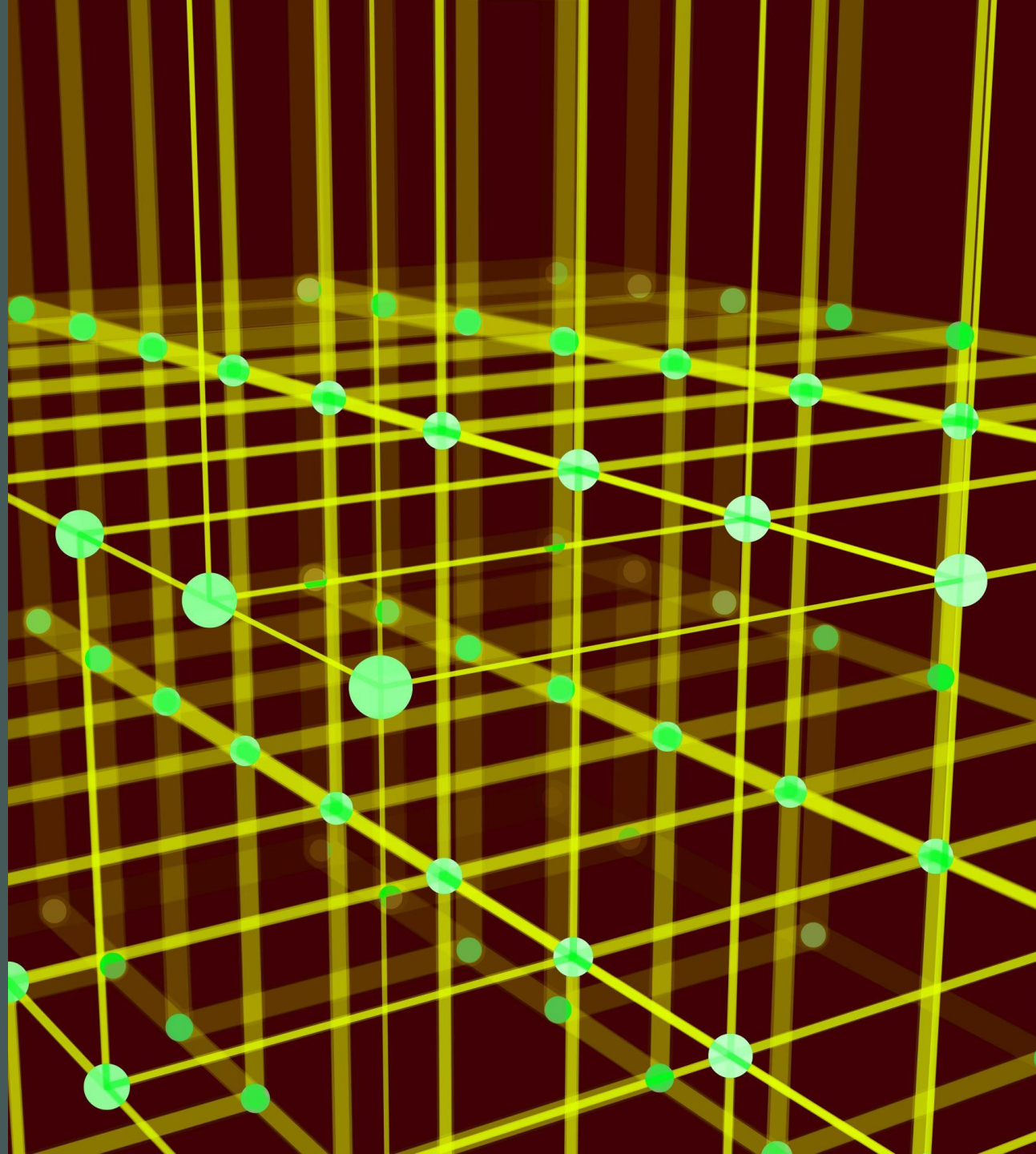
- Ajuda a caracterizar a precisão, justiça, transparência e resultados em tomadas de decisão alimentadas por IA.

A IA explicável é crucial para uma organização na construção de confiança e segurança ao colocar modelos de IA em produção.

A explicabilidade da IA também ajuda uma organização a adotar uma abordagem responsável para o desenvolvimento de IA.

Conceito

IA Explicável (XAI): conjunto de métodos e processos para tornar decisões de modelos de IA interpretáveis para humanos.



Interpretabilidade x Explicabilidade

Pimenta Chipotle dá dor de estômago?
Barulho alto acelera a perda auditiva?
Mulheres são menos agressivas que homens?

Se um modelo de aprendizado de máquina puder criar uma definição para essas relações, ele é **interpretável**.

Interpretabilidade x Explicabilidade

Todos os modelos devem começar com uma hipótese.
A curiosidade humana impulsiona um ser a intuir que uma coisa se relaciona com outra.
As pessoas criam modelos internos para interpretar seus arredores.
No campo do aprendizado de máquina, esses modelos podem ser testados e verificados como representações precisas ou imprecisas do mundo.

Interpretabilidade x Explicabilidade

Interpretabilidade significa que a causa e o efeito podem ser determinados.

Interpretabilidade x Explicabilidade

Se um modelo puder receber as entradas e rotineiramente obter as mesmas saídas, o modelo é interpretável:

- Se você come macarrão demais na hora do jantar e sempre tem problemas para dormir, a situação é interpretável.
- Se todas as pesquisas de 2016 mostraram uma vitória democrata e o candidato republicano assumiu o cargo, todos esses modelos apresentaram baixa interpretabilidade.
- Se o objetivo dos pesquisadores é ter um bom modelo, o que a instituição jornalística é obrigada a fazer — relatar a verdade —, então o erro mostra que seus modelos precisam ser atualizados.

Interpretabilidade x Explicabilidade

A interpretabilidade não representa problema em cenários de baixo risco.

- Se um modelo estiver recomendando filmes para assistir, essa pode ser uma tarefa de baixo risco. (A menos que você seja um dos grandes provedores de conteúdo e todas as suas recomendações sejam ruins a ponto de as pessoas acharem que estão perdendo tempo, mas você entendeu.)
- Se um modelo estiver gerando qual será a sua cor favorita do dia ou gerando metas simples de ioga para você se concentrar ao longo do dia, eles jogam jogos de baixo risco e a interpretabilidade do modelo é desnecessária.

Interpretabilidade x Explicabilidade

Às vezes, a interpretabilidade precisa ser alta para justificar por que um modelo é melhor que outro.

Em Moneyball, os olheiros da velha guarda tinham um modelo interpretável que usavam para escolher bons jogadores para times de beisebol; não eram modelos de aprendizado de máquina, mas os olheiros haviam desenvolvido seus métodos (basicamente um algoritmo) para selecionar qual jogador teria um bom desempenho em uma temporada em comparação com a outra. Mas o técnico queria mudar esse método.

Para que os métodos de Billy Beane funcionassem e para que a metodologia se consolidasse, seu modelo precisava ser altamente interpretável quando contrariasse tudo o que a indústria acreditava ser verdade. A alta interpretabilidade do modelo vence discussões.

Interpretabilidade x Explicabilidade

Um modelo com alta interpretabilidade é desejável em um cenário de alto risco.

Modelos com alta interpretabilidade equivalem à capacidade de responsabilizar outra parte.

E quando os modelos preveem se uma pessoa tem câncer, as pessoas precisam ser responsabilizadas pela decisão tomada.

Modelos com alta interpretabilidade e a manutenção de alta interpretabilidade como padrão de projeto podem ajudar a construir confiança entre desenvolvedores e usuários.

Interpretabilidade x Explicabilidade

A alta interpretabilidade permite que as pessoas sigam as regras do sistema.

- Se o professor distribuir uma explicação que mostre como está avaliando a prova, tudo o que o aluno precisa fazer é seguir as regras.
- Se o professor for fanático por "Wayne's World", o aluno sabe que deve contar anedotas sobre o assunto.
- Ou, se o professor realmente quiser ter certeza de que o aluno entende o processo de como as bactérias decompõem as proteínas no estômago, o aluno não deve descrever os tipos de proteínas e bactérias existentes.
    - Em vez disso, deve ir direto ao que as bactérias estão fazendo.

Interpretabilidade x Explicabilidade

Interpretabilidade: refere-se à capacidade de compreender o processo de tomada de decisão de um modelo de IA.

- Um modelo interpretável é transparente em sua operação e fornece informações sobre as relações entre entradas e saídas.
- Um algoritmo interpretável pode ser explicado de forma clara e compreensível por um ser humano.
- A interpretabilidade é, portanto, importante para garantir que os usuários possam compreender e confiar nos modelos de IA.

Interpretabilidade x Explicabilidade

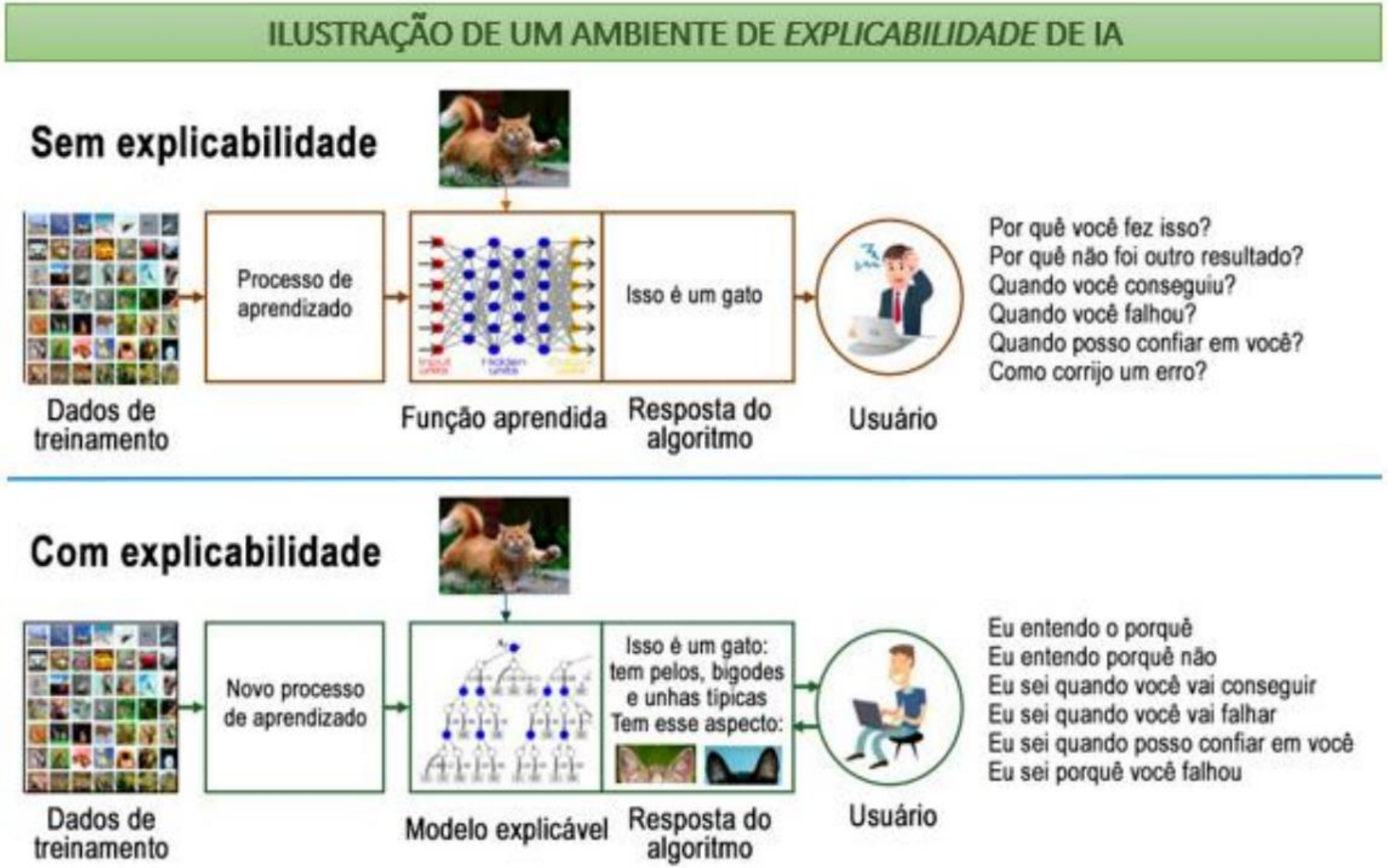
Explicabilidade envolve a capacidade de descrever em termos compreensíveis como o sistema de IA chegou a uma decisão ou resultado específico.

Explicabilidade está mais relacionada à lógica ou ao raciocínio por trás das decisões individuais da IA, tornando os processos da IA acessíveis e relacionáveis aos usuários finais.

Um exemplo de explicabilidade em um algoritmo de IA pode ser um modelo de aprendizado de máquina usado em pontuação de crédito, onde a IA avalia a capacidade creditícia de um indivíduo com base em vários fatores, como renda, histórico de crédito, situação profissional e níveis de dívida.

O aspecto explicável seria a IA fornecer razões para sua decisão, como declarar que um pedido de empréstimo foi negado devido a uma pontuação de crédito baixa e uma alta relação dívida/renda.

Explicabilidade



Interpretabilidade x Explicabilidade

Explicabilidade: refere-se à capacidade de explicar o processo de tomada de decisão de um modelo de IA em termos compreensíveis para o usuário final.

Um modelo explicável fornece uma explicação clara e intuitiva das decisões tomadas, permitindo que os usuários entendam por que o modelo produziu um determinado resultado.

Em outras palavras, a explicabilidade se concentra em por que um algoritmo tomou uma decisão específica e como essa decisão pode ser justificada.

Interpretabilidade x Explicabilidade

Embora interpretabilidade e explicabilidade sejam importantes para a compreensão de modelos de IA, existem algumas diferenças importantes entre os dois conceitos:

Nível de detalhe:

A interpretabilidade se concentra na compreensão do funcionamento interno dos modelos, enquanto a explicabilidade se concentra em explicar as decisões tomadas.

Consequentemente, a interpretabilidade requer um nível de detalhe maior do que a explicabilidade.

Interpretabilidade x Explicabilidade

Complexidade do modelo:

Modelos de IA mais complexos, como redes neurais profundas, podem ser difíceis de interpretar devido à sua estrutura intrincada e às interações entre diferentes partes do modelo.

Nesses casos, a explicabilidade pode ser mais viável, pois se concentra em explicar decisões em vez de compreender o modelo em si.

Interpretabilidade x Explicabilidade

Comunicação:

A interpretabilidade diz respeito à compreensão do modelo por especialistas e pesquisadores de IA, enquanto a explicabilidade se concentra mais em comunicar as decisões do modelo aos usuários finais.

Como resultado, a explicabilidade requer uma apresentação de informações mais simples e intuitiva.

Motivação

Crescimento do uso de modelos “caixa-preta” (deep learning, ensemble complexos).

Necessidade de confiança, auditoria e conformidade com regulamentos (ex.: GDPR – direito à explicação, AI Act da União Europeia). de métodos e processos para tornar decisões de modelos de IA interpretáveis para humanos.

A importância de uma IA explicável – Exemplo na Visão Computacional

A interpretabilidade dos modelos de visão por computador ajuda os usuários a compreenderem melhor como foi feita uma previsão e a lógica que lhe está subjacente.

A transparência contribui para este objetivo, tornando o funcionamento do modelo claro para todos, definindo claramente as limitações do modelo e garantindo que os dados são utilizados de forma ética.

A importância de uma IA explicável – Exemplo na Visão Computacional

Por exemplo, a visão computacional pode ajudar os radiologistas a identificar de maneira eficaz, complicações de saúde em imagens de raios X.

Radiografias do tórax analisadas com o Vision AI, mostrando as classificações de tuberculose saudável, doente, ativa e latente.

A importância de uma IA explicável – Exemplo na Visão Computacional

Um sistema de visão que seja apenas exato não é suficiente.

- O sistema também precisa de ser capaz de explicar as suas decisões.
  - Imaginemos que o sistema podia mostrar quais as partes da imagem que levaram às suas conclusões - então, quaisquer resultados seriam mais claros.
  - Este nível de transparência ajudaria os profissionais de saúde a verificar as suas conclusões e a garantir que os cuidados prestados aos pacientes cumprem as normas médicas.

Como funciona a XAI

Técnicas de XAI

A configuração das técnicas da XAI consiste em três métodos principais.

A precisão e a rastreabilidade da previsão lidam com os requisitos tecnológicos, enquanto a compreensão da decisão atende às necessidades humanas.

Como funciona a XAI

Precisão da Previsão

A precisão é um componente-chave do sucesso do uso da IA nas operações cotidianas.

Ao executar simulações e comparar a saída da XAI com os resultados no conjunto de dados de treinamento, a precisão da previsão pode ser determinada.

A técnica mais popular usada para isso é de **Explicações Locais Interpretáveis Independentes de Modelo** (LIME), que explica a previsão dos classificadores pelo algoritmo de ML.

Como funciona a XAI

Precisão da Previsão

Em essência, ele se refere a quão confiável é a explicação de uma previsão feita por um modelo de IA.

Em outras palavras, a explicação que a XAI te dá realmente se alinha com o que o modelo está "pensando"?

Imagine que você trabalha em um banco e precisa usar um modelo de IA para decidir se um cliente deve receber um empréstimo ou não.

Como funciona a XAI

Precisão da Previsão

Passo 1: O modelo faz a previsão

O cliente João entra com um pedido de empréstimo.

O modelo de IA analisa os dados dele, como idade (35 anos), renda (R$ 8.000), histórico de crédito (bom) e dívidas atuais (baixas).

O modelo de IA faz a sua previsão: **aprovar o empréstimo**.

Como funciona a XAI

Precisão da Previsão

Passo 2: A XAI entra em cena e gera a explicação

Como você não quer apenas uma resposta de "sim" ou "não", você usa uma ferramenta de XAI (como **SHAP** ou **LIME**) para entender o porquê da decisão.

A XAI analisa a previsão do modelo e te dá uma explicação.

A explicação da XAI é: "O empréstimo foi aprovado principalmente porque a **renda do cliente é alta** e o **histórico de crédito é bom**."

Como funciona a XAI

Precisão da Previsão

Passo 3: A validação da precisão da explicação (a Precisão da Previsão)

A pergunta crucial: essa explicação da XAI é realmente o que o modelo usou para tomar a decisão?

Para validar a **precisão da previsão**, você pode fazer um "teste de fidelidade".

Imagine que você "perturba" ou modifica os dados do cliente para ver se a previsão do modelo muda de acordo com a explicação.

Como funciona a XAI

Precisão da Previsão

Teste 1: Reduzir a renda de João de R$ 8.000 para R$ 1.500, mantendo o histórico de crédito bom.

Resultado esperado (com base na explicação da XAI): O modelo deve mudar a previsão para "negar o empréstimo", já que a renda era um fator-chave.

O que acontece: O modelo muda a previsão para "negar o empréstimo". Isso mostra que a explicação da XAI era **precisa**.

Como funciona a XAI

Precisão da Previsão

Teste 2: Piorar o histórico de crédito de João (de "bom" para "ruim"), mantendo a renda em R$ 8.000.

Resultado esperado (com base na explicação da XAI): O modelo deve mudar a previsão para "negar o empréstimo", já que o histórico de crédito era outro fator-chave.

O que acontece: O modelo muda a previsão para "negar o empréstimo". Novamente, a explicação da XAI foi **precisa**.

Como funciona a XAI

Precisão da Previsão

Teste 3 (um exemplo de baixa precisão): Agora imagine que você muda a idade de João (de 35 para 20 anos), mas a explicação da XAI nem sequer mencionou a idade.

Resultado esperado (com base na explicação da XAI): O modelo não deve mudar a previsão.

O que acontece: O modelo muda a previsão para "negar o empréstimo", mesmo que a explicação da XAI não tenha falado nada sobre a idade.

Conclusão: Neste caso, a **precisão da previsão** da XAI é **baixa**, porque ela não capturou a importância da idade na decisão do modelo. A explicação não era uma representação fiel do que o modelo estava fazendo.

Como funciona a XAI

Precisão da Previsão

Em resumo, a **precisão da previsão** na XAI é uma medida de quão bem a explicação gerada pela ferramenta de XAI reflete o processo de decisão real do modelo.

Um alto grau de precisão significa que você pode confiar na explicação e entender genuinamente por que a IA tomou uma decisão, o que é fundamental para a ética, transparência e confiança em sistemas de IA.

Como funciona a XAI

Rastreabilidade

É um conceito fundamental que permite entender e acompanhar todo o ciclo de vida de uma previsão feita por um modelo de IA.

É como ter um "rastro de migalhas" digital que nos leva desde o dado de entrada até a decisão final, com todas as explicações no meio.

O exemplo prático, novamente é no contexto de um banco que usa IA para decidir sobre empréstimos.

Como funciona a XAI

Rastreabilidade

Passo 1: O pedido do cliente e a coleta de dados

Um cliente, a Sra. Ana, solicita um empréstimo. O sistema de IA do banco coleta os dados dela.

Dados de Entrada: Idade (45 anos), Renda Mensal (R$ 12.000), Histórico de Crédito (excelente), Dívidas Atuais (baixas), Tipo de Moradia (própria).

Rastreabilidade aqui: Cada um desses dados deve ser rastreável. Você deve saber a fonte exata de onde cada dado veio (ex: "renda obtida do extrato bancário do dia X").

Como funciona a XAI

Rastreabilidade

Passo 2: O modelo de IA faz a previsão

O modelo de IA processa os dados da Sra. Ana e faz sua previsão.

Previsão do Modelo: Aprovar o empréstimo.

Rastreabilidade aqui: Você deve registrar exatamente qual versão do modelo de IA foi usada para fazer essa previsão (ex: "Modelo de aprovação de crédito v3.1").

Se o modelo for atualizado na semana seguinte, a decisão sobre a Sra. Ana ainda estará vinculada à versão original.

Como funciona a XAI

Rastreabilidade

Passo 3: A ferramenta de XAI gera a explicação

Uma ferramenta de XAI, como o SHAP, é usada para explicar a decisão.

Explicação da XAI: O empréstimo foi aprovado principalmente devido ao **histórico de crédito excelente** e à **alta renda mensal**. A idade e o tipo de moradia tiveram uma influência neutra.

Rastreabilidade aqui: A explicação da XAI também precisa ser rastreada. Você deve registrar que a explicação foi gerada pelo "módulo de XAI SHAP v1.5", atrelado àquela previsão específica.

Como funciona a XAI

Rastreabilidade

Passo 4: O registro da decisão e a validação humana

Com a previsão e a explicação em mãos, um analista de crédito revisa o caso da Sra. Ana. Ele vê a previsão "aprovar" e a explicação do modelo.

Ação Final: O analista concorda com a previsão e aprova o empréstimo.

Rastreabilidade aqui: O sistema deve registrar que a decisão final foi tomada pelo "analista João da Silva, em 17/08/2025 às 10:30", com base na previsão e na explicação do modelo.

Como funciona a XAI

Rastreabilidade

O valor da rastreabilidade

Seis meses depois, uma auditoria interna questiona por que o empréstimo da Sra. Ana foi aprovado. Graças à rastreabilidade, é possível reconstruir o processo completo passo a passo:

Dados de Entrada: Os dados de Ana foram coletados do sistema X e Y na data Z.

Modelo Usado: A decisão foi tomada pela versão v3.1 do modelo, que estava em produção naquele momento.

Explicação: A ferramenta de XAI mostrou que os fatores mais importantes foram "histórico de crédito" e "renda".

Decisão Humana: O analista João da Silva revisou o caso e aprovou a decisão.

Como funciona a XAI

Rastreabilidade

Em resumo, a **rastreabilidade na XAI** não se limita a explicar uma decisão; ela cria um registro auditável e completo de como essa decisão foi tomada, desde a fonte dos dados até a intervenção humana.

Isso é vital para garantir a transparência, a responsabilidade e o cumprimento de regulamentações, especialmente em indústrias críticas como finanças e saúde.

Como funciona a XAI

Compreensão da decisão

Esse é o fator humano. Muitas pessoas têm desconfiança em relação à IA, mas para trabalhar com ela de forma eficiente, precisam aprender a confiar nela.

Isso se consegue educando a equipe que trabalha com a IA para que possam entender como e por que a IA toma decisões.

Como funciona a XAI

Compreensão da decisão

A **compreensão** na XAI é o objetivo final de todo o processo de explicação.

Não basta apenas que um modelo seja rastreável ou que sua explicação seja precisa; a explicação precisa ser **entendível** por um humano.

Como funciona a XAI

Compreensão da decisão

Podemos usar como analogia a diferença entre dar a alguém uma lista de ingredientes (os dados) e uma receita completa e clara (a explicação).

O objetivo é que a pessoa consiga entender como o bolo (a decisão) foi feito.

Como funciona a XAI

Compreensão da decisão

Passo 1: O modelo de IA faz a previsão

Um médico carrega a imagem do raio-X do pulmão de um paciente no sistema de IA.

O modelo de aprendizado de máquina analisa a imagem.

Previsão do Modelo: O modelo de IA classifica a imagem como **"pneumonia"** com 85% de confiança.

Como funciona a XAI

Compreensão da decisão

Passo 2: A ferramenta de XAI gera a explicação

O sistema de IA não apenas diz "pneumonia", mas também usa uma ferramenta de XAI (como LIME ou Grad-CAM) para gerar uma explicação visual. Essa explicação não é um texto, mas uma **sobreposição visual** na própria imagem do raio-X.

Explicação da XAI: A ferramenta de XAI realça as áreas na imagem do raio-X que mais contribuíram para a previsão. Ela destaca uma mancha esbranquiçada e opaca no lóbulo inferior do pulmão direito. A área realçada é marcada em vermelho, indicando que é um "fator positivo" para a previsão de pneumonia.

Como funciona a XAI

Compreensão da decisão

Passo 3: A verificação da compreensão humana

Agora vem a parte mais importante: a **compreensão**. O médico precisa ser capaz de interpretar essa explicação e usá-la.

Interpretação do Médico: Ao ver a sobreposição, o médico imediatamente reconhece a mancha realçada em vermelho. Ele sabe que **essa opacidade é um sinal clínico clássico de pneumonia** em raios-X. A explicação visual do modelo se alinha perfeitamente com o seu conhecimento médico.

Validação do Médico: O médico usa a explicação para confirmar sua própria suspeita clínica. Ele não precisa simplesmente confiar na previsão de "pneumonia" do modelo; ele entende o **porquê**. Isso aumenta sua confiança na ferramenta de IA e o ajuda a tomar uma decisão informada sobre o tratamento do paciente.

Como funciona a XAI

Compreensão da decisão

E se a compreensão for baixa?

Vamos supor que a explicação da XAI não seja tão boa.

Explicação ruim: A ferramenta de XAI realça as bordas da imagem, a identificação do paciente, ou uma área do coração.

Interpretação do Médico: O médico olha para a explicação e fica confuso. "Por que o modelo está focando nas bordas da imagem? Isso não tem relação com pneumonia."

Resultado: A **compreensão é baixa**. O médico perde a confiança na ferramenta de IA porque a explicação não faz sentido para ele, e ele provavelmente ignorará a previsão do modelo.

Como funciona a XAI

Compreensão da decisão

Em resumo, a **compreensão na XAI** é a medida de quão bem a explicação de um modelo de IA pode ser interpretada e utilizada por um ser humano.

No final das contas, o objetivo da XAI não é apenas gerar explicações, mas torná-las **intuitivas e úteis**, permitindo que os usuários confiem e colaborem efetivamente com a IA.

Tipos de Modelo

- **Intrinsecamente interpretáveis** (ex.: regressão linear, árvore de decisão).

- **Caixa-preta** (ex.: redes neurais profundas, boosting).

Tipos de Modelo

Existem modelos algorítmicos que são "auto"-interpretáveis e existem aqueles que necessitam ser explicados por meio de técnicas externas de explicabilidade.

Um modelo interpretável (ou transparente) de machine learning é aquele que não requer técnicas adicionais para que o humano possa compreendê-lo.

Entre estes sistemas encontram-se: regressão linear/logística, árvores de decisão, k-vizinhos mais próximos, RBML (Rule-based machine learning), GAM (generalized additive model) e os modelos bayesianos.

Tipos de Modelo

O fato de um modelo algorítmico ser interpretável, não significa, todavia, que ele dispense a explicabilidade.

Pelo contrário, sua transparência tornará a explicabilidade ainda mais viável.

Por exemplo a árvore de decisão.

Os modelos baseados em árvore dividem os dados várias vezes de acordo com determinados valores de corte nos recursos.

Por meio da divisão, diferentes subconjuntos do conjunto de dados são criados, com cada instância pertencendo a um subconjunto.

Tipos de Modelo

As previsões individuais de uma árvore de decisão podem ser explicadas decompondo-se o caminho de decisão em um componente por recurso.

Pode-se rastrear uma decisão por meio da árvore e explicar uma previsão pelas contribuições adicionadas em cada nó de decisão.

Tipos de Modelo

A árvore de decisão simples é um exemplo de '**modelo transparente**'.

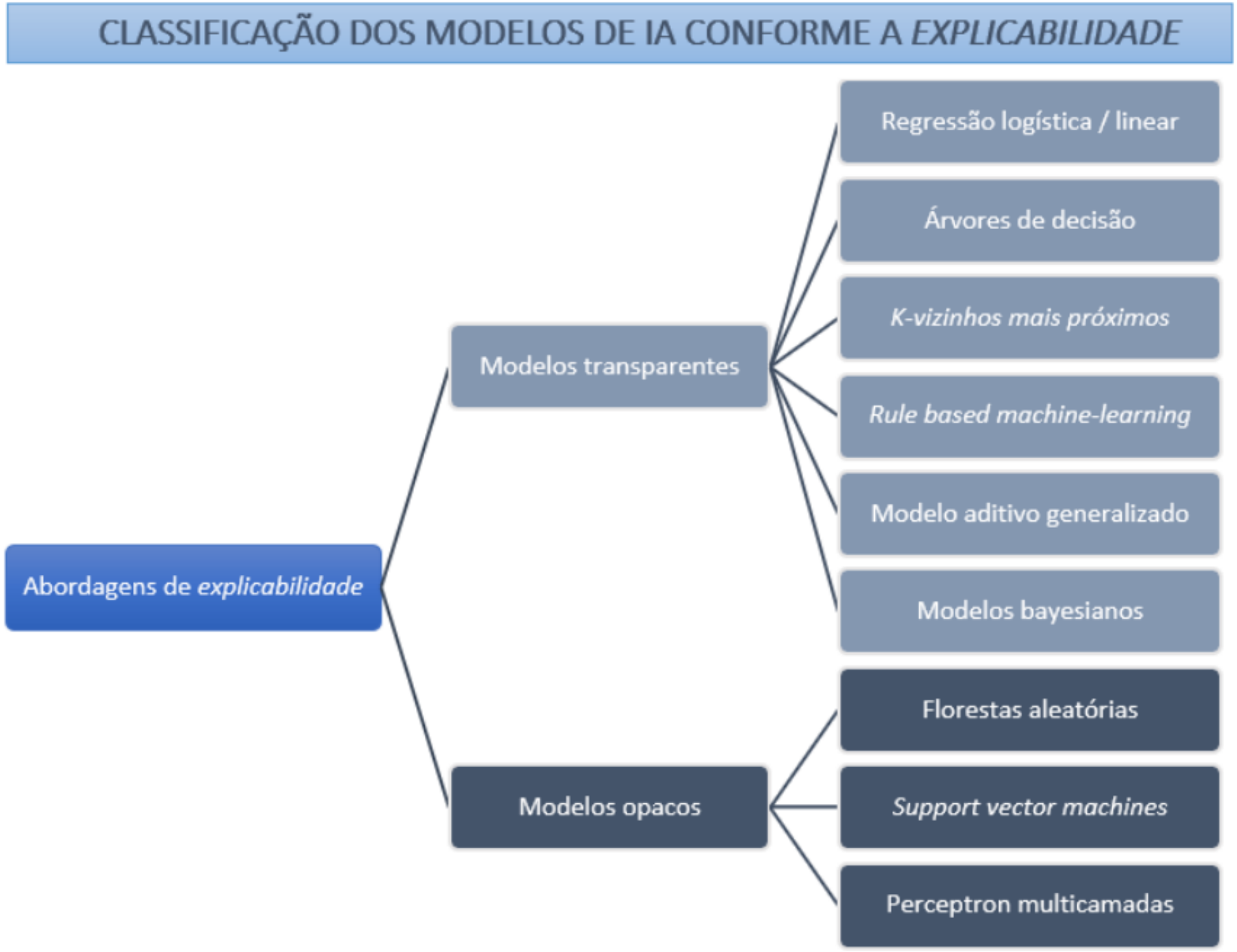
A cada pergunta, ela classifica os elementos analisados.

Tipos de Modelo

Em oposição aos "modelos transparentes", existem os "modelos opacos", cuja compreensão irá requerer um processo de explicação adicional, chamado de "explicabilidade post-hoc".

Entre esses modelos opacos, é possível apontar as Support Vector Machines (SVM), o Perceptron multicamadas e as florestas de decisão aleatórias (random forests).

Tipos de Modelo



Dimensões da Explicabilidade

- Local vs. Global.
- Pós-modelo (post-hoc) x Durante o treinamento (ante-hoc).
- Focada no modelo x focada no usuário.

Dimensões da Explicabilidade

Explicações Locais

As **explicações locais** se concentram no "porquê" de uma **única previsão**.

Elas fornecem um nível de detalhe cirúrgico, explicando a decisão para um ponto de dado específico.

O que respondem? "Por que o modelo aprovou o empréstimo do cliente João?" ou "Por que o modelo diagnosticou pneumonia neste raio-X específico?"

Dimensões da Explicabilidade

Explicações Locais

Quando usar? São cruciais para a **tomada de decisão em tempo real** e para a **responsabilidade**.

Em áreas como medicina, finanças ou direito, é fundamental entender a razão por trás de cada caso individual.

Se um cliente é negado, ele tem o direito de saber o porquê.

Se um médico usa a IA para um diagnóstico, ele precisa entender os fatores específicos daquele paciente.

Dimensões da Explicabilidade

Explicações Locais

Exemplos de métodos:

LIME (Local Interpretable Model-agnostic Explanations) e

SHAP (SHapley Additive exPlanations) são os métodos mais comuns.

Eles funcionam perturbando os dados de entrada ou avaliando a contribuição de cada recurso para aquela previsão específica.

Dimensões da Explicabilidade

Explicações Globais

As **explicações globais** se concentram no "como" do modelo **como um todo**. Elas fornecem uma visão de alto nível, explicando os padrões e o comportamento geral do modelo em todo o conjunto de dados.

O que respondem?

"Quais são os fatores mais importantes para a aprovação de empréstimos em geral?" ou

"Quais características de imagem são mais relevantes para o modelo detectar pneumonia em todos os casos?"

Dimensões da Explicabilidade

Explicações Globais

Quando usar?

São essenciais para **auditoria, depuração de modelos e conformidade regulatória**.

Uma explicação global ajuda a identificar vieses (o modelo está dando menos peso para renda e mais para gênero?) ou a entender se o modelo aprendeu o que deveria (ele realmente se baseia em sinais clínicos ou apenas em artefatos da imagem?).

Dimensões da Explicabilidade

Explicações Globais

Exemplos de métodos:

Gráficos de **Importância de Recurso** (Feature Importance),

Gráficos de Dependência Parcial (Partial Dependence Plots) e explicações agregadas do **SHAP**.

Esses métodos avaliam o impacto médio ou a tendência de um recurso em todas as previsões do modelo.

Métodos e Algoritmos de XAI

Métodos Intrínsecos (*ante-hoc*)

São uma classe de modelos de inteligência artificial que são **explicáveis por design**. Isso significa que a explicabilidade não é adicionada depois que o modelo é criado, mas sim construída na sua própria estrutura, desde o início.

A expressão latina "ante-hoc" significa "antes do fato", "antes da ocorrência". Nesse contexto, refere-se ao momento em que a explicabilidade é incorporada: **antes do treinamento** do modelo.

Métodos e Algoritmos de XAI

Métodos Intrínsecos (*ante-hoc*)

Como eles funcionam?

Diferente dos modelos de "caixa-preta" (como redes neurais profundas), que são opacos e exigem ferramentas de XAI "post-hoc" (depois do fato) para serem compreendidos, os modelos intrínsecos são "caixas de vidro". Você pode ver e entender como eles tomam suas decisões simplesmente inspecionando sua estrutura e seus parâmetros.

A principal característica desses modelos é que a sua arquitetura é, por natureza, interpretável. A lógica de decisão é inerente, clara e, muitas vezes, pode ser representada de forma simples.

Métodos e Algoritmos de XAI

Métodos Intrínsecos (*ante-hoc*)

Vantagens

Transparência e Confiança: A lógica de decisão é clara e auditável, o que aumenta a confiança dos usuários e dos *stakeholders*.

Fidelidade à Explicação: Não há risco de a explicação ser imprecisa ou enganosa, pois a explicação é o próprio funcionamento do modelo.

Simplicidade: Modelos intrínsecos são geralmente mais fáceis de entender e depurar.

Métodos e Algoritmos de XAI

Métodos Intrínsecos (*ante-hoc*)

Desvantagens

Precisão Limitada: Geralmente, modelos intrínsecos não conseguem alcançar a mesma alta precisão preditiva que os modelos de "caixa-preta" mais complexos. Para problemas muito complexos (como reconhecimento de imagens), essa limitação é um grande problema.

Menor Flexibilidade: A simplicidade que os torna explicáveis também limita sua capacidade de capturar padrões e relacionamentos complexos nos dados.

Métodos e Algoritmos de XAI

Métodos Intrínsecos (*ante-hoc*) – Exemplos

- **Modelos de Árvore de Decisão:** São talvez o exemplo mais clássico.
  - O processo de decisão é uma sequência de perguntas "se/então" (if/then) que pode ser visualizada como um fluxograma.
  - É muito fácil seguir o caminho que o modelo percorreu para chegar a uma decisão específica.

Métodos e Algoritmos de XAI

Métodos Intrínsecos (*ante-hoc*) – Exemplos

•**Regressão Linear e Logística:** As equações desses modelos são totalmente transparentes.

- O peso (coeficiente) de cada variável de entrada mostra a sua influência na previsão.
- Se uma variável tem um peso de 0.5, por exemplo, um aumento de uma unidade nessa variável leva a um aumento de 0.5 na previsão, mantendo as outras variáveis constantes.

Métodos e Algoritmos de XAI

Métodos Intrínsecos (*ante-hoc*) – Exemplos

- **Modelos de Regras de Associação:** Usados para encontrar relações entre variáveis, como "clientes que compram produto X também compram produto Y em 80% das vezes".
  - A explicação é a própria regra gerada pelo modelo.

Métodos e Algoritmos de XAI

Métodos Post Hoc

- **Globais**
  - Feature Importance (permutação, ganho de informação).
  - Partial Dependence Plots (PDP).
  - Surrogate Models (ex.: árvore de decisão para explicar rede neural).
- **Locais**
  - LIME (Local Interpretable Model-agnostic Explanations).
  - SHAP (SHapley Additive exPlanations).
  - Anchors.

Métodos e Algoritmos de XAI

Métodos Post Hoc

- **Métodos Visuais**
  - Mapas de calor em CNNs (Grad-CAM, Layer-wise Relevance Propagation).
  - Visualização de embeddings (t-SNE, UMAP).

Métodos e Algoritmos de XAI

Modelos Agnósticos

É uma abordagem que, de propósito, não leva em conta a estrutura ou o funcionamento interno de um modelo de machine learning.

O termo "agnóstico" vem da ideia de "não saber", significando que o método de explicação **ignora a arquitetura interna do modelo**.

Métodos e Algoritmos de XAI

Modelos Agnósticos

Em vez de tentar entender os detalhes complexos de uma rede neural profunda ou de um algoritmo de boosting, um método agnóstico ao modelo trata o modelo de IA como uma "caixa-preta".

Ele se concentra exclusivamente na relação entre as entradas (dados) e as saídas (previsões) do modelo.

Métodos e Algoritmos de XAI

Modelos Agnósticos

Como funciona?

Esses métodos se baseiam em um princípio simples: **perturbar a entrada para observar o que acontece com a saída.** A ideia é testar o modelo de fora, sem abrir a "caixa-preta".

Escolha um modelo de "caixa-preta": Pode ser qualquer modelo, de uma rede neural complexa a um Random Forest, desde que você possa passar dados de entrada para ele e obter uma previsão.

Selecione uma instância para explicar: Pegue um único ponto de dado (por exemplo, os dados de um cliente específico ou uma imagem de raio-X).

Métodos e Algoritmos de XAI

Modelos Agnósticos

Como funciona?

Perturbe os dados: Crie diversas variações do dado de entrada original. Por exemplo, mude um valor, remova uma parte da imagem ou embaralhe um conjunto de dados.

Observe as previsões: Passe todas essas variações pelo modelo de "caixa-preta" e colete as novas previsões.

Crie uma explicação: Analise como as previsões mudaram em resposta às perturbações. Se uma pequena mudança em uma variável causou uma grande alteração na previsão, essa variável é considerada importante.

Métodos e Algoritmos de XAI

Modelos Agnósticos

Vantagens dos Modelos Agnósticos

Flexibilidade: A maior vantagem é que eles funcionam com **qualquer tipo de modelo** de machine learning. Isso significa que você pode usar o mesmo método de XAI para explicar uma regressão logística simples e um modelo de deep learning gigantesco.

Foco no resultado: Eles permitem que você interprete o comportamento de um modelo sem ter que entender a complexidade matemática e algorítmica por trás dele. Isso é ótimo para usuários que não são especialistas em machine learning, como médicos, gerentes ou advogados.

Métodos e Algoritmos de XAI

Modelos Agnósticos

Exemplos de Métodos Agnósticos em XAI

SHAP (SHapley Additive exPlanations): É uma das ferramentas mais populares. Ela calcula a importância de cada variável para uma previsão, baseando-se em um conceito da teoria dos jogos para garantir que a contribuição de cada variável seja justa e consistente.

LIME (Local Interpretable Model-agnostic Explanations): Este método cria um modelo substituto, mais simples e interpretável (como uma regressão linear), que "imita" o comportamento do modelo complexo apenas em uma região próxima ao ponto de dado que você quer explicar.

Métodos e Algoritmos de XAI

Método	Tipo	Nível	Modelo-agnóstico?	Vantagens	Limitações
LIME	Pós-hoc	Local	Sim	Fácil de aplicar	Instável em certos casos
SHAP	Pós-hoc	Local+Global	Sim	Base teórica sólida	Custo computacional alto
PDP	Pós-hoc	Global	Sim	Boa visão geral	Perde interações complexas
Grad-CAM	Pós-hoc	Local	Não	Visualização intuitiva	Aplicável só a CNNs

Algoritmos

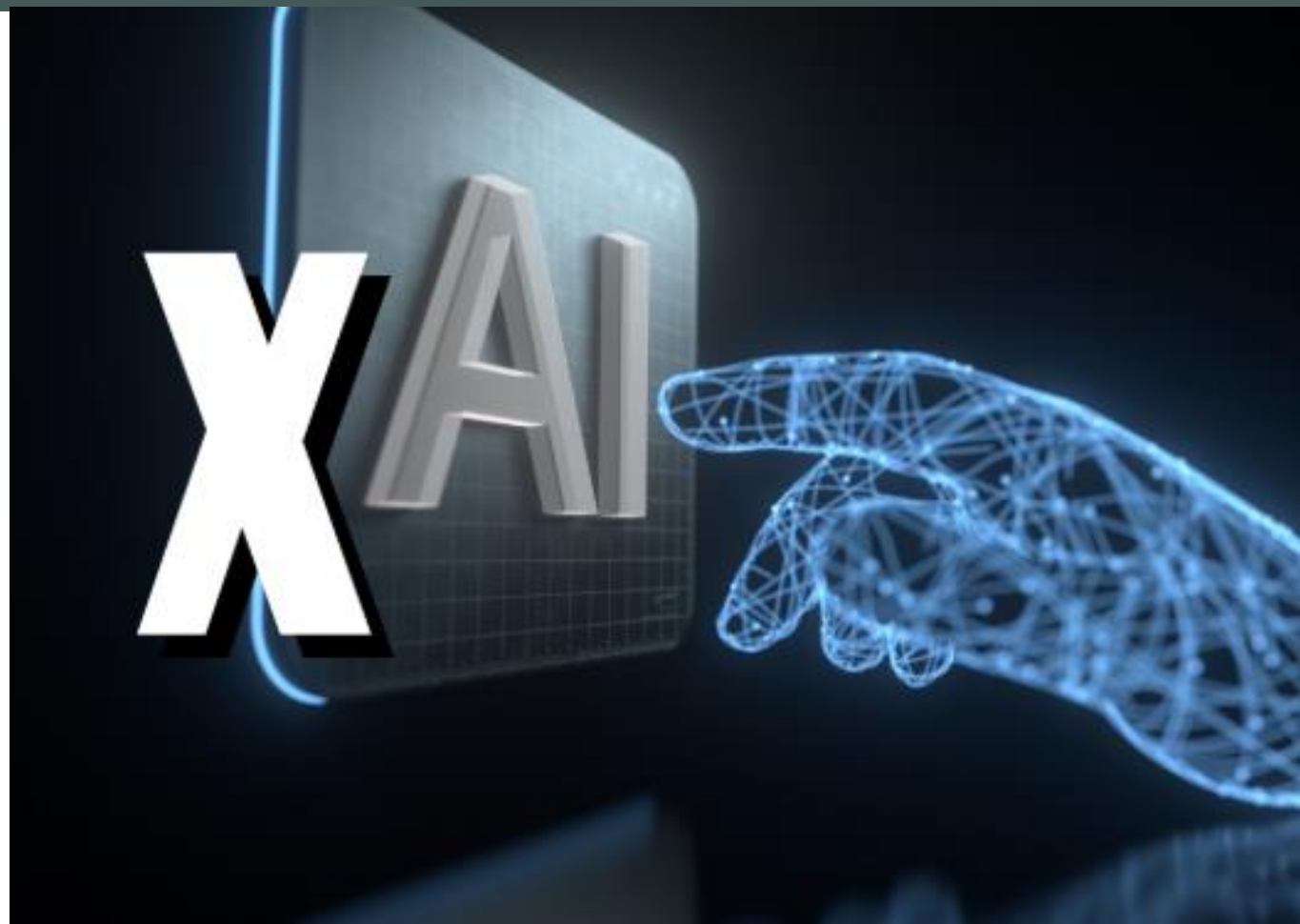
- **Modelos Substitutos (Surrogate Models)**: Treinar um modelo interpretável (substituto) para aproximar as previsões de um modelo complexo (caixa-preta). A interpretabilidade do modelo substituto é então usada para entender o modelo original.
- **Mapas de Saliência (Saliency Maps)**: Usados principalmente em visão computacional, esses mapas destacam as regiões de uma imagem que mais contribuíram para a previsão do modelo.
- **Explicações Contra-factuais (Counterfactual Explanations)**: Descrevem a menor mudança nas características de entrada que alteraria a previsão do modelo para um resultado desejado. Isso ajuda a entender o que precisaria ser diferente para obter um resultado diferente.

Algoritmos

- **Decomposição de Relevância em Camadas (Layer-wise Relevance Propagation - LRP):** Uma técnica para decompor a previsão de uma rede neural em contribuições de neurônios de entrada, mostrando a relevância de cada entrada para a saída final.
- **Modelos Aditivos Generalizados (Generalized Additive Models - GAMs):** Modelos que permitem que a contribuição de cada característica seja modelada de forma flexível, mantendo a interpretabilidade ao mostrar o efeito de cada característica individualmente.

Ferramentas

- LIME (pacote lime)
- SHAP (pacote shap)
- InterpretML
- Captum (PyTorch)
- Alibi (framework open-source)



Estudo de Caso 1: Concessão de Crédito

Cenário

Um banco utiliza um modelo de aprendizado de máquina para decidir se concede ou não um empréstimo a um cliente.

O modelo é complexo (por exemplo, uma rede neural profunda) e, para um determinado cliente, o modelo nega o empréstimo.

O cliente, insatisfeito, solicita uma explicação para a decisão.

Estudo de Caso 1: Concessão de Crédito

Aplicação da XAI (usando SHAP):

Neste cenário, poderíamos aplicar a técnica SHAP para explicar a decisão do modelo.

O SHAP atribuiria um "valor SHAP" a cada característica do cliente (idade, renda, histórico de crédito, dívidas, etc.) para a previsão específica de negação do empréstimo.

Esses valores indicariam a contribuição de cada característica para a decisão final do modelo.

Estudo de Caso 1: Concessão de Crédito

Exemplo de Explicação SHAP:

Renda Mensal: -0.8 (contribuição negativa, ou seja, renda mais baixa contribuiu para a negação)

Histórico de Crédito (Pontuação): -0.5 (contribuição negativa, pontuação de crédito baixa contribuiu para a negação)

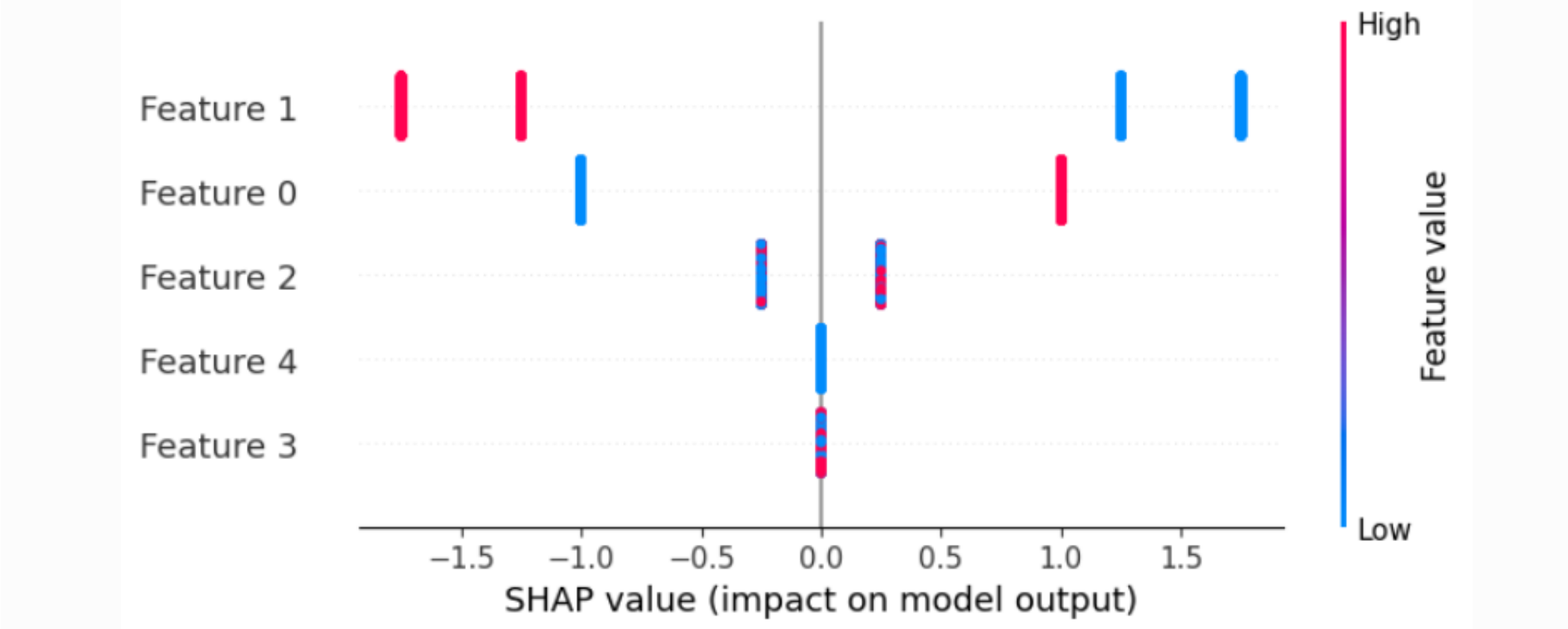
Idade: +0.1 (contribuição positiva, idade não foi um fator significativo para a negação)

Dívidas Atuais: -0.6 (contribuição negativa, alto nível de dívidas contribuiu para a negação)

Estudo de Caso 1: Concessão de Crédito

Exemplo de Explicação SHAP

```
[17]: shap.plots.beeswarm(explanation)
```



Estudo de Caso 1: Concessão de Crédito

Conclusões:

Com base nesses valores SHAP, o banco poderia explicar ao cliente que a negação do empréstimo foi principalmente devido à sua baixa renda mensal e ao seu histórico de crédito desfavorável, além do alto nível de dívidas.

Essa explicação é transparente e acionável, permitindo que o cliente entenda os motivos e, potencialmente, tome medidas para melhorar sua situação financeira no futuro.

Estudo de Caso 2:
Diagnóstico Médico por Imagem

Cenário:

Um sistema de IA é treinado para detectar a presença de uma doença em imagens de raios-X.

Para um paciente específico, o modelo prevê a presença da doença.

O médico precisa entender por que o modelo chegou a essa conclusão para confirmar o diagnóstico e planejar o tratamento.

Estudo de Caso 2:
Diagnóstico Médico por Imagem

Aplicação da XAI (usando Mapas de Saliência):

Mapas de saliência são particularmente úteis em aplicações de visão computacional.

Eles destacam as regiões da imagem que mais influenciaram a decisão do modelo.

Estudo de Caso 2:
Diagnóstico Médico por Imagem

Aplicação da XAI (usando Mapas de Saliência):

Visão lado a lado da imagem e seu respectivo mapa de saliência construído a partir do modelo

Estudo de Caso 2:
Diagnóstico Médico por Imagem

Exemplo de Explicação com Mapa de Saliência:

O mapa de saliência gerado para a imagem de raios-X do paciente mostraria as áreas da imagem (por exemplo, regiões do pulmão) que o modelo considerou mais relevantes para prever a doença.

As áreas mais "quentes" no mapa indicariam os pixels ou regiões da imagem que tiveram a maior contribuição para a previsão positiva da doença.

Estudo de Caso 2:
Diagnóstico Médico por Imagem

Conclusão:

Essa visualização permite que o médico veja exatamente onde o modelo está "olhando" na imagem para fazer seu diagnóstico.

Se o modelo estiver focando em uma área clinicamente relevante, isso aumenta a confiança do médico na previsão.

Se estiver focando em artefatos ou regiões irrelevantes, isso pode indicar um problema no modelo ou nos dados de treinamento.

Estudo de Caso 3:
Recomendação de Produtos em E-commerce

Cenário:

Uma plataforma de e-commerce utiliza um sistema de recomendação de IA para sugerir produtos aos usuários.

Para um usuário específico, o sistema recomenda um conjunto de produtos.

A empresa deseja entender por que esses produtos foram recomendados para otimizar suas estratégias de marketing e melhorar a experiência do usuário.

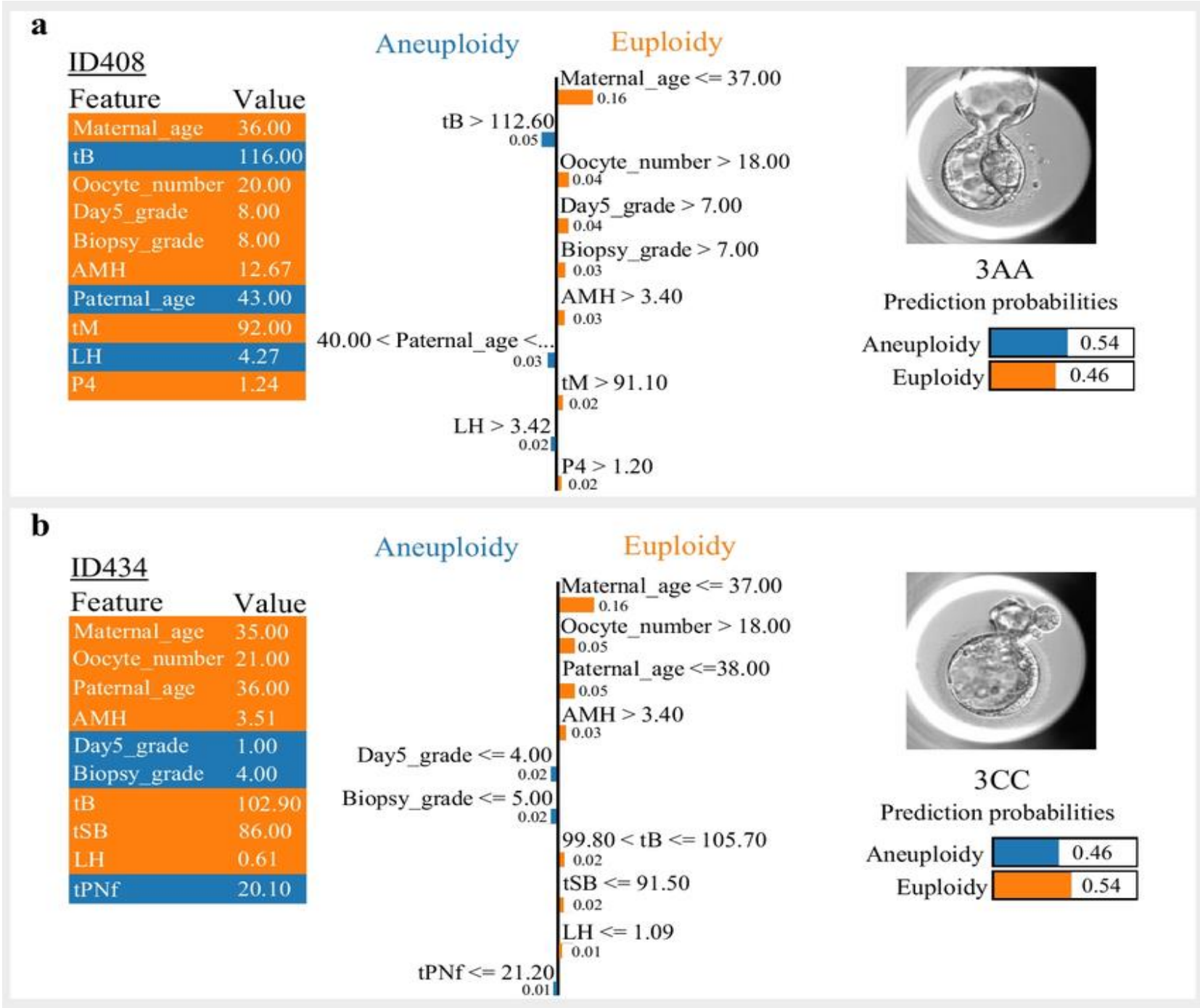
Estudo de Caso 3:
Recomendação de Produtos em E-commerce

Aplicação da XAI (usando LIME):

LIME pode ser usado para explicar as recomendações para um usuário individual, criando um modelo local interpretável que aproxima o comportamento do modelo complexo.

Estudo de Caso 3: Recomendação de Produtos em E-commerce

Aplicação da XAI (usando LIME)



Estudo de Caso 3: Recomendação de Produtos em E-commerce

Exemplo de Explicação LIME:

Para a recomendação de um "fone de ouvido sem fio" para um usuário, o LIME poderia gerar uma explicação como:

- Histórico de Compras: O usuário comprou "acessórios de áudio" recentemente (contribuição positiva forte).
- Visualizações de Produtos: O usuário visualizou "fones de ouvido com fio" e "caixas de som portáteis" (contribuição positiva).
- Interesses Declarados: O usuário indicou interesse em "tecnologia" (contribuição positiva moderada).
- Demografia: Idade e localização não foram fatores significativos (contribuição neutra).

Estudo de Caso 3:
Recomendação de Produtos em E-commerce

Conclusões:

Essa explicação permite que a equipe de marketing entenda que a recomendação foi impulsionada principalmente pelo histórico de compras e visualizações do usuário, confirmando que o sistema está alinhado com o comportamento do cliente.

Isso também pode revelar oportunidades para refinar as recomendações ou personalizar ainda mais a experiência do usuário.