Explainability of Predictive Models in Healthcare: An Analysis of SHAP's Local and Global Explanation Plots

Matheus Henrique Pereira Borba[1], **Fabrício Jailson Barth**[1]

[1]INSPER
São Paulo - SP, Brasil

matheushpb@al.insper.edu.br, fabriciojb@insper.edu.br

Abstract. *This article analyzes how explanations generated by the* SHAP *tool are communicated to end users in supervised machine learning models applied to healthcare. A predictive model was developed using the CatBoost algorithm, and its local and global explanations were evaluated by 22 physicians according to the attributes of user orientation, context orientation, objectivity, comprehensibility, and informativeness. The results indicated a more positive perception of local explanations, while clarity and informativeness received relatively lower evaluations. As an additional contribution, a consistent questionnaire was developed and applied to assess explainability tools from the end-user perspective.*

Resumo. *Este artigo analisa como as explicações geradas pela ferramenta* SHAP *são comunicadas ao usuário final em modelos de aprendizado de máquina supervisionada aplicados à saúde. Para isso, foi desenvolvido um modelo preditivo com o algoritmo CatBoost, cujas explicações locais e globais foram avaliadas por 22 médicos segundo atributos de orientação ao usuário e ao contexto, objetividade, compreensibilidade e informatividade. Os resultados indicaram uma percepção mais positiva das explicações locais, enquanto clareza e informatividade receberam avaliações relativamente menores. Como contribuição adicional, foi desenvolvido e aplicado um questionário consistente para avaliar ferramentas de explicabilidade sob a perspectiva do usuário final.*

1. Introdução

O uso de modelos de *Machine Learning* (ML) em aplicações médicas tem se tornado cada vez mais comum, especialmente com o avanço de algoritmos capazes de lidar com uma extensa quantidade de dados tabulares e sua popularização. Entretanto, muitos desses modelos, apesar de terem uma elevada acurácia, são considerados "caixa-preta", o que dificulta a interpretação do porquê e do como as decisões são tomadas [Hassija 2023]. Em contextos médicos, esses modelos poderiam ser uma ferramenta de auxílio no processo de análise de resultados e de diagnósticos, por possibilitarem a extração das principais informações do modelo.

Por mais que existam essas vantagens e otimizações, as decisões clínicas podem ser críticas, pois envolvem a vida e a saúde de pacientes. Essa natureza sensível das decisões tomadas traz consigo um ceticismo por parte dos médicos, por mais que tenham um potencial de otimizar os processos de decisão. Visando essas vantagens tanto para o dia a dia do médico, quanto para o dos pacientes, se faz necessária a implementação

de técnicas que tornam esse fator de explicabilidade mais transparente, já que as decisões clínicas devem ser justificáveis e compreensíveis para ambos. Estudos recentes [Bharati et al. 2023, da Silva 2023, Markus et al. 2021] reforçam a importância da explicabilidade em modelos aplicados à saúde, não apenas pela sensibilidade das decisões envolvidas, mas também pela necessidade de confiança dos profissionais clínicos.

Assim, as técnicas de XAI, do inglês *eXplainable Artificial Intelligence*, surgem com a intenção de evidenciar os fatores que influenciam as previsões dos modelos, ou seja, fornecer uma ferramenta que transforme uma matemática complexa em algo inteligível à linguagem humana e menos técnica [Chaddad et al. 2023]. Atualmente, existem diversas soluções que se propõem a fazer isto, entre elas, a ferramenta SHAP (*SHapley Additive exPlanations*) [Lundberg and Lee 2017], LIME (*Local Interpretable Model-agnostic Explanations*) [Ribeiro et al. 2016] e Anchor [Ribeiro et al. 2018].

No entanto, para que uma ferramenta seja considerada adequadamente explicável, não basta apenas fornecer uma justificativa técnica para a decisão do modelo. É necessário que a explicação seja apresentada de forma que efetivamente favoreça a compreensão do usuário, tornando o processo de decisão do sistema mais transparente, acessível e útil para a interpretação dos resultados [Kyrimi 2024].

1.1. Atributos da explicabilidade

A explicabilidade é definida como a capacidade de compreender o mecanismo interno e o comportamento de um sistema, bem como de explicar por que uma ação específica foi realizada [Salih et al. 2024]. Com isso, evidencia-se que a explicabilidade não se limita à visualização de resultados, mas envolve a construção de mecanismos que tornem o comportamento do modelo inteligível para diferentes perfis de usuários. Diante disso, no contexto da área da saúde, considerando que o usuário médio terá uma menor compreensão técnica, a utilização de uma técnica com boa explicabilidade torna-se imprescindível.

Uma boa explicabilidade pode ser classificada com base em alguns atributos essenciais. Nesse contexto, um atributo pode ser compreendido como uma característica ou propriedade que define a qualidade e a adequação de um método explicativo em relação ao seu propósito. Trata-se de um critério utilizado para avaliar se a explicação fornecida por determinado método cumpre sua função de tornar o modelo mais compreensível e confiável.

Um estudo recente [Kyrimi 2024] reuniu profissionais da saúde, desenvolvedores de inteligência artificial (IA) e pesquisadores da área de XAI com o objetivo de identificar os principais atributos considerados essenciais em um método explicável. A análise classificou os atributos das explicações em dois grupos: importantes e críticos, e desejáveis, mas não essenciais. Entre os importantes e críticos estão orientação ao propósito, clareza, caráter informativo, orientação ao usuário, orientação ao domínio, comunicação de incerteza, orientação ao contexto, fidelidade ao funcionamento do sistema e objetividade, garantindo que a explicação seja clara, adequada ao perfil do usuário, transparente quanto ao nível de confiança e fiel ao modelo. Já os atributos desejáveis incluem seletividade, robustez, interatividade, caráter causal, transparência, contraste, avaliativo, consciência temporal e capacidade contrafactual, que enriquecem a explicação e apoiam a tomada de decisão, embora não sejam indispensáveis.

Ainda de acordo com [Kyrimi 2024], a **orientação ao usuário** refere-se à capacidade da explicação de ser adaptada às necessidades, capacidades e nível de conhecimento do usuário que a recebe. Já a **orientação ao contexto** está relacionada à adequação da explicação ao contexto da decisão, de modo que as informações fornecidas sejam relevantes para a ação ou julgamento que será realizado. A **objetividade** consiste na apresentação das informações de forma imparcial, minimizando interpretações subjetivas e ambiguidades. Além disso, uma explicação deve ser **compreensível**, ou seja, clara e facilmente entendida pelos usuários, e **informativa**, fornecendo informações necessárias e suficientes para reduzir a lacuna de conhecimento do usuário e justificar adequadamente a decisão produzida pelo sistema.

Este trabalho tem como objetivo principal analisar como as explicações produzidas pela ferramenta SHAP são comunicadas a médicos por meio de suas representações visuais locais e globais. Especificamente, busca-se investigar em que medida essas representações auxiliam a interpretação dos resultados gerados pelo modelo e identificar as potencialidades e limitações das explicações sob a perspectiva desses profissionais.

Para essa análise, foram selecionados os atributos de orientação ao usuário, orientação ao contexto, objetividade, compreensibilidade e informatividade. A escolha desses atributos fundamenta-se no estudo de [Kyrimi 2024], que os classifica como importantes e críticos para a efetividade de explicações em sistemas de IA aplicados à saúde. Além disso, esses atributos estão diretamente relacionados à forma como as explicações são percebidas, compreendidas e utilizadas pelos usuários finais, alinhando-se ao propósito deste trabalho de avaliar a comunicação e a utilidade prática das explicações, em vez de aspectos associados à implementação técnica dos métodos de explicabilidade.

2. Ferramenta SHAP

O SHAP [Lundberg and Lee 2017] baseia-se na teoria dos jogos, ou seja, interpreta cada *feature* como um jogador que coopera para um prêmio e analisa a sua importância para a equipe. Nesse caso, a importância para a equipe seria o peso da *feature* e o prêmio seria a saída. Dessa forma, esse método mede quantitativamente quanto cada entrada contribuiu para a saída. Para encontrar esses pesos, o método consiste em pegar uma entrada local e fornecer uma *feature* por vez (isso significa atribuir um valor de referência/baseline às demais features e apenas colocar o valor real das *feature* já fornecidas) e, em cada uma delas, é registrado o quanto a predição aumenta ou diminui. Dessa forma, obtém-se o valor em que cada *feature* contribuiu. Entretanto, como a ordem em que a revelação ocorre interfere diretamente no valor previsto, é calculada a média de contribuição de cada uma das n *feature* das $n!$ possíveis combinações.

Esse método funciona definindo um valor base, o qual é a média das saídas do conjunto de dados e, a partir dele, soma os pesos de cada entrada, encontrando o valor de saída do método, a qual deve ser muito próxima, embora não necessariamente igual, à saída do modelo original. Como parte da ferramenta de explicabilidade SHAP, existem diversas formas de visualização para interpretar as decisões de um modelo, permitindo tanto a explicação local, ao analisar a contribuição das variáveis para uma única instância, quanto a explicação global, ao avaliar o comportamento do modelo em todo o conjunto de dados.

Neste trabalho, o estudo será concentrado em dois dos gráficos mais utilizados

para inferências locais e globais: *Waterfall* e *BeeSwarm*. As Figuras 1 e 2 representam uma tentativa de explicabilidade para um modelo de classificação treinado com um *dataset* sobre a COVID-19 [Nizri 2026]. A figura 1 apresenta uma explicação usando o método *BeeSwarm* e a figura 2 apresenta uma explicação usando o método *Waterfall*.

O gráfico *BeeSwarm* (Figura 1) é uma técnica de explicação global que apresenta a distribuição dos valores SHAP para todas as instâncias do conjunto de dados. Nele, cada ponto representa o impacto de uma variável em uma observação específica, e as variáveis são organizadas de acordo com sua importância média. O gráfico evidencia tanto a magnitude quanto a direção do efeito de cada variável, sendo que a cor indica o valor da própria variável (vermelho para valores altos e azul para valores baixos), enquanto a posição no eixo horizontal corresponde ao valor SHAP, que reflete a contribuição daquela variável para a predição do modelo.

O gráfico *Waterfall* (Figura 2) é uma técnica de explicação local que ilustra como cada variável contribui, de forma positiva ou negativa, para a predição de uma instância específica. Ele é representado por um gráfico de barras horizontais, uma para cada variável, no qual as barras em vermelho indicam contribuições que aumentam o valor da variável de saída, enquanto as barras em azul representam contribuições que o reduzem.

3. Método

Este trabalho foi conduzido com o objetivo de avaliar se as explicações geradas pela ferramenta SHAP atendiam a atributos essenciais de explicabilidade em modelos de aprendizado de máquina sob a perspectiva de profissionais da área da saúde. Para isso, o método foi composto por quatro etapas: (i) desenvolvimento de um modelo preditivo utilizando um dataset sobre COVID-19 [Nizri 2026]; (ii) geração de explicações por meio da ferramenta SHAP; (iii) coleta de dados sobre a percepção de profissionais da saúde em relação às explicações apresentadas em um questionário; e (iv) análise dos dados obtidos.

Inicialmente, foi desenvolvido um modelo preditivo baseado no algoritmo *CatBoost* utilizando o dataset COVID-19 disponibilizado por [Nizri 2026]. Este dataset possui 1.048.575 exemplos e é caracterizado com 21 atributos. O objetivo do modelo foi predizer o desfecho de pacientes atendidos com suspeita de COVID-19, classificando-os entre sobreviventes e não sobreviventes. A variável target deste dataset é desbalanceada (sobreviventes = 92,66%; não sobreviventes = 7,34%). Os hiperparâmetros do modelo, o pré-processamento utilizado no dataset e outras escolhas podem ser acessadas no repositório do projeto[1]. O modelo com melhor capacidade preditiva teve F1-Score igual a 0,62 e ROC-AUC igual a 0,95. Ressalta-se que o foco deste trabalho não esteve na maximização do desempenho preditivo do modelo, mas na avaliação das explicações geradas pela ferramenta SHAP. Após o treinamento do modelo, foram produzidas explicações globais e locais utilizando essa ferramenta.

Este questionário foi validado por uma pesquisadora na área da saúde e disponibilizado a médicos convidados a participar da pesquisa por meio da rede de contatos do grupo de pesquisa. Antes de responder ao questionário, todos os participantes tiveram acesso a um *termo de consentimento livre e esclarecido* e manifestaram sua concordância em participar da pesquisa anonimamente. Os participantes avaliaram diretamente as explicações apresentadas, sem a intermediação de profissionais especializados

[1] `https://github.com/mthperera/XAI-Research`

em ciência de dados ou aprendizado de máquina, buscando reproduzir uma situação na qual o próprio médico interpreta as informações fornecidas pelo sistema.

O questionário foi estruturado em três partes. Na primeira, foram coletadas informações sociodemográficas e profissionais com o objetivo de caracterizar os participantes: data de nascimento, sexo, estado de atuação, cidade de atuação, área(s) de atuação/especialidades (era possível escolher entre 55 especialidades reconhecidas pelo Conselho Federal de Medicina), tempo de atuação em anos, principais locais de trabalho (múltipla seleção entre hospital privado, hospital público, hospital-escola, UBS/PSF ou consultório particular) e categoria atual (seleção única entre residente ou médico especialista). Na segunda parte, foram realizadas perguntas sobre conhecimento prévio em inteligência artificial e técnicas de explicabilidade (XAI), como apresentado na Tabela 1. Na terceira parte, os participantes avaliaram explicações globais e locais geradas pela ferramenta SHAP aplicada ao modelo treinado com o dataset COVID-19 [Nizri 2026]. As questões de avaliação utilizadas nessa etapa são apresentadas na Tabela 2.

Embora a literatura empregue o termo compreensibilidade, a operacionalização desse atributo no instrumento de coleta foi realizada por meio do conceito de clareza, entendido como a facilidade de entendimento da explicação pelo usuário. Dessa forma, os termos são tratados como equivalentes neste estudo.

Tabela 1. Perguntas sobre conhecimento prévio

Pergunta	**Formato da Resposta**
Com que frequência você utiliza sistemas baseados em IA ou algoritmos de apoio à decisão clínica?	Escala Likert: Nunca; Raramente; Às vezes; Frequentemente; Muito frequentemente
Antes desta pesquisa, qual era seu nível de conhecimento sobre explicabilidade em modelos de IA (XAI)?	Escala Likert: Nenhum conhecimento; Conhecimento muito básico; Conhecimento intermediário; Conhecimento avançado; Conhecimento muito avançado
Qual é o seu nível de experiência com ferramentas que explicam as decisões de modelos de IA?	Escala Likert: Nenhuma experiência; Experiência muito limitada; Experiência moderada; Experiência significativa; Experiência extensa

Antes da etapa de avaliação, os participantes visualizaram os gráficos de explicação global e local gerados pela ferramenta SHAP acompanhados de uma breve descrição interpretativa e de uma tabela contendo as variáveis clínicas utilizadas pelo modelo e seus respectivos significados. A explicação global é apresentada na Figura 1 e o texto utilizada para descrevê-la é apresentado a seguir:

> O gráfico apresenta uma análise abrangendo todos os pacientes do conjunto de dados. À esquerda, encontram-se as variáveis consideradas pelo modelo, como, por exemplo, a presença ou ausência de diabetes, onde estão dispostas por relevância, ou seja, as variáveis do topo são consideradas pela ferramenta como mais importantes para a tomada de decisão do modelo. À direita, observa-se uma barra de cores (heatmap), que representa a intensidade dos valores associados a cada variável. Na parte inferior do gráfico, está indicado o SHAP value, que expressa a contribuição de cada variável para a decisão do modelo. As variáveis dispostas no lado esquerdo do gráfico estão ordenadas de acordo com seus respectivos valores de SHAP, de modo que aquelas posicionadas no topo exercem maior influência na previsão do modelo. Por exemplo, as variáveis idade e pneumonia apresentam maior contribuição para a decisão do modelo quando comparadas às variáveis asma e doença cardiovascular. Além disso, a coloração dos pontos fornece informação adicional sobre os valores assumidos por cada variável. No caso da variável pneumonia, os pontos em vermelho representam pacientes que apresentaram pneumonia, sendo que tais valores contribuem de forma mais significativa para que o modelo preveja a ocorrência de óbito, já que apresentam maior SHAP value.

A explicação local é apresentada na Figura 2 e o texto utilizada para descrevê-la é apresentado a seguir:

Tabela 2. Perguntas avaliativas sobre os atributos

Atributo	Pergunta	Descrição
Orientado ao usuário	A linguagem e o nível de complexidade da explicação são adequados à minha prática profissional?	Esta pergunta avalia se a explicação está adaptada ao perfil do usuário, considerando sua formação, experiência profissional e necessidades específicas.
Orientado ao contexto	A explicação apresenta informações relevantes que apoiam minha tomada de decisão clínica?	Esta pergunta verifica se a explicação é pertinente ao contexto clínico e contribui de forma prática para a tomada de decisão. Avalia se as informações fornecidas têm utilidade real na avaliação do caso e na definição de condutas.
Objetividade	A explicação é objetiva e minimiza minhas possíveis interpretações subjetivas?	Esta pergunta avalia se a explicação é apresentada de maneira direta, baseada em critérios claros e dados explícitos. O objetivo é reduzir ambiguidades e interpretações pessoais, promovendo maior precisão e confiança na compreensão da decisão do sistema.
Compreensível (clareza)	A explicação é clara e fácil de entender?	Esta pergunta avalia se a explicação é apresentada de forma clara e organizada. Verifica se a estrutura, os termos utilizados e a forma de apresentação facilitam o entendimento, permitindo que o usuário compreenda a mensagem principal sem dificuldade.
Informativo	A explicação fornece informações suficientes para que eu compreenda como o sistema chegou à sua decisão?	Esta pergunta avalia se a explicação apresenta os elementos essenciais que conectam os dados analisados à decisão final. O foco está na suficiência das informações, garantindo que o usuário consiga entender o raciocínio do sistema sem lacunas importantes.

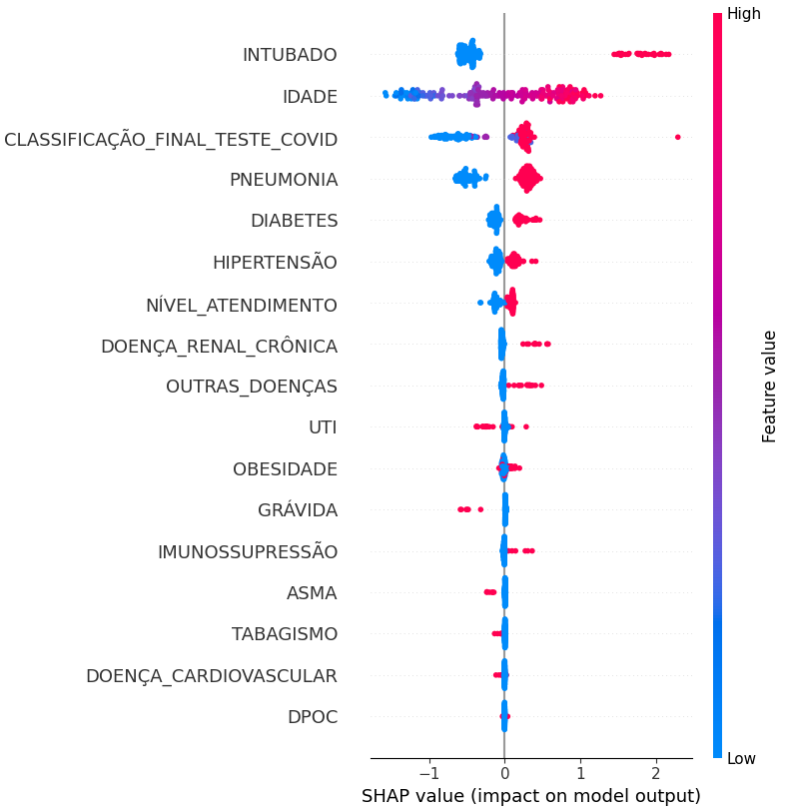


Figura 1. Explicação Global do SHAP usada no questionário

O gráfico apresentado corresponde a um SHAP waterfall referente a um paciente específico. Trata-se do caso de um paciente com diagnóstico de pneumonia, 68 anos de idade e histórico de diabetes, que evoluiu para óbito por COVID-19, apresentando um SHAP value final do modelo de f(x) = 1,698. Esse valor é transformado por uma fórmula matemática própria do modelo para ser convertido em probabilidade. Após essa transformação, o valor corresponde a uma probabilidade estimada de aproximadamente 84,5% de óbito. O gráfico tem início no valor base do modelo, que corresponde à média das predições considerando todos os pacientes do conjunto de dados. A partir desse valor, observa-se como cada variável contribui individualmente para deslocar a predição em direção ao desfecho final. As barras em vermelho indicam variáveis que aumentam a probabilidade estimada de óbito, enquanto as barras em azul representam variáveis que reduzem essa probabilidade. As contribuições são acumuladas sequencialmente até que se atinja o valor final do SHAP (f(x)). Nesse caso específico, as variáveis intubado e idade foram as que mais contribuíram para elevar o risco estimado, ao passo que outras variáveis exerceram efeito contrário, reduzindo a probabilidade final predita pelo modelo.

Após essa contextualização, os participantes responderam às questões apresentadas na Tabela 2. As respostas às questões de avaliação seguiram uma escala Likert de cinco pontos, variando de (1) Discordo Totalmente a (5) Concordo Totalmente. Os itens avaliados contemplaram atributos relacionados à explicabilidade, incluindo compreensibilidade, objetividade, informatividade, adequação ao contexto e orientação ao usuário.

4. Resultados

O questionário foi respondido por médicos, dos quais 12 (54,5%) eram mulheres e 10 (45,5%) homens. Entre os participantes, 21 eram especialistas e apenas 1 era residente. A idade variou entre 25 e 73 anos, com média de 46,2 anos. O tempo de experiência profissional variou de 1 a 45 anos, com média de 18,7 anos. Em relação às classificações de atuação profissional, foram analisadas tanto as áreas de trabalho quanto as especialidades médicas dos participantes. Observou-se predominância de atuação em clínicas privadas (14 profissionais), seguidas por hospitais públicos (9 profissionais), hospitais privados (8

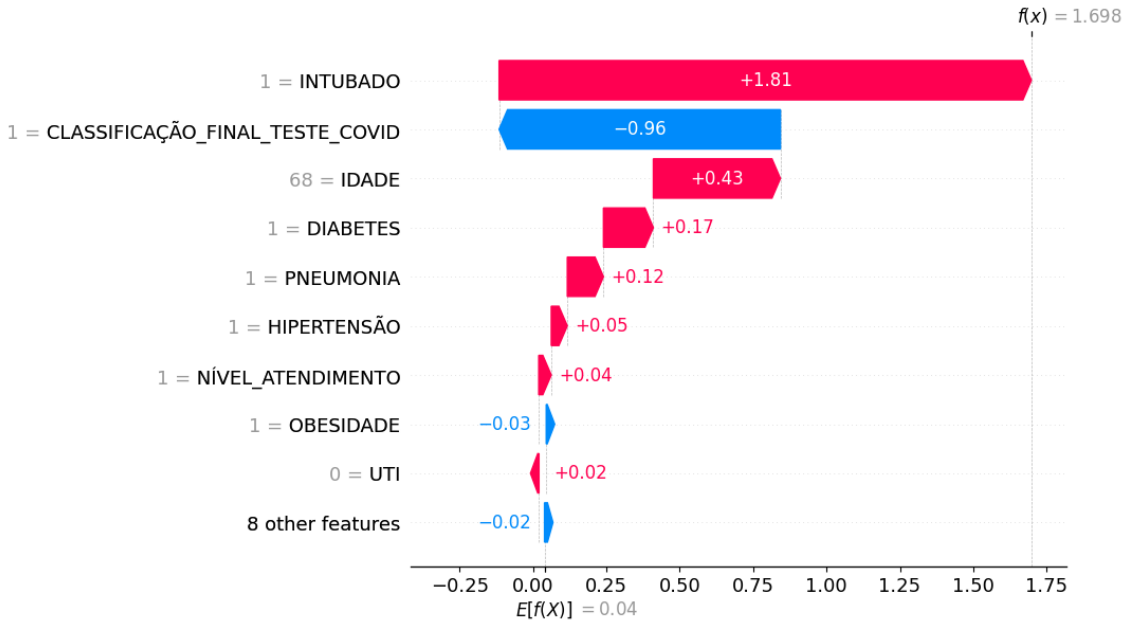


Figura 2. Explicação Local do SHAP usada no questionário

profissionais) e hospitais universitários (8 profissionais). O local de atuação destes profissionais é majoritariamente o estado de São Paulo (16 - 73%), com participação de Minas Gerais (4), Espirito Santo (1) e Ceará (1).

Quanto às especialidades, houve predominância de Psiquiatria (6 profissionais). Também foram identificadas as especialidades de Ginecologia e Obstetrícia (2 profissionais), Dermatologia (2 profissionais) e Radiologia e Diagnóstico por Imagem (2 profissionais). As demais especialidades apresentaram frequência individual de 1 profissional, incluindo Infectologia, Geriatria, Pediatria, Reumatologia, Coloproctologia, Neurologia, Endocrinologia e Metabologia, Anestesiologia, Medicina do Trabalho, Ortopedia e Traumatologia e Clínica Médica.

A frequência de utilização de sistemas baseados em IA ou algoritmos de apoio à decisão clínica apresentou média aproximada de 2,5 pontos na escala Likert de 1 a 5, situando-se entre os níveis “Raramente” e “Às vezes”. Esse resultado indica baixa frequência de uso dessas ferramentas entre os participantes. O nível de conhecimento prévio sobre explicabilidade em modelos de IA (XAI) apresentou média aproximada de 2,0 ponto, correspondente a um nível entre “Conhecimento muito básico”. Observou-se concentração das respostas nos níveis inferiores da escala. Em relação à experiência com ferramentas capazes de explicar decisões de modelos de IA, como justificativas do modelo, destaque de fatores relevantes e visualizações de risco, a média observada foi de aproximadamente 1,9 pontos, situando-se entre “Experiência muito limitada” e “Experiência moderada”. As distribuições completas das respostas para as três variáveis estão apresentadas na Tabela 3.

A frequência de utilização de sistemas baseados em IA não apresentou correlação com o nível de conhecimento prévio sobre XAI (ρ de Spearman = $0,03$). Em contrapartida, o conhecimento prévio sobre XAI e a experiência com esse tipo de ferramenta apresentaram correlação positiva moderada ($\rho = 0,58$). O conjunto das três questões re-

Tabela 3. Descrição das respostas sobre conhecimento prévio dos participantes

Pergunta	Média	Moda	Min	Max
Uso de IA em apoio à decisão clínica	2.50	2	1	5
Conhecimento prévio sobre XAI	2.00	2	1	3
Experiência com ferramentas de explicabilidade	1.86	2	1	3

lacionadas ao conhecimento prévio apresentou coeficiente alfa de Cronbach igual a 0,52, indicando baixa consistência interna. Dessa forma, as questões foram analisadas separadamente, pois não houve evidência suficiente para tratá-las como uma única medida composta. Idade e anos de experiência não tem correlação com o conhecimento prévio dos participantes sobre IA ou XAI.

4.1. Avaliação de atributos

Na Tabela 4 são apresentados valores estatísticos das respostas referentes às avaliações dos atributos do SHAP. O alfa de Cronbach calculado para as questões relacionadas às explicações globais foi de 0,951, enquanto o valor obtido para as questões referentes às explicações locais foi de 0,919, indicando elevada consistência interna em ambos os conjuntos de perguntas. Dessa forma, os itens de cada conjunto foram agregados por meio da média das respostas, gerando uma medida única para a avaliação das explicações globais e locais. Esta medida única é chamada de qualidade média na Tabela 4.

Tabela 4. Média, mediana e moda das avaliações das explicações globais e locais

	Figura 1 (Global)			Figura 2 (Local)		
Atributo	Média	Mediana	Moda	Média	Mediana	Moda
Orientação ao usuário	3,64	4	4	3,95	4	4
Orientação ao contexto	3,59	4	4	3,91	4	4
Objetividade	3,45	4	4	3,77	4	4
Compreensibilidade	3,36	3	3	3,55	4	4
Informatividade	3,32	4	4	3,55	4	4
Qualidade Média	**3,47**	3,8	-	**3,75**	4,0	-

Para os gráficos referentes às explicações globais, observou-se média de $3{,}64 \pm 1{,}09$ para orientação ao usuário, $3{,}59 \pm 1{,}10$ para orientação ao contexto, $3{,}45 \pm 1{,}06$ para objetividade, $3{,}36 \pm 1{,}00$ para clareza e $3{,}32 \pm 1{,}17$ para capacidade informativa. A qualidade média da explicação global apresentou uma média de $3{,}47 \pm 0{,}99$ pontos. Para os gráficos referentes às explicações locais, foram observadas médias de $3{,}95 \pm 0{,}72$ para orientação ao usuário, $3{,}91 \pm 0{,}81$ para orientação ao contexto, $3{,}77 \pm 0{,}87$ para objetividade, $3{,}55 \pm 1{,}14$ para clareza e $3{,}55 \pm 1{,}01$ para capacidade informativa. A qualidade média da explicação local apresentou uma média de $3{,}75 \pm 0{,}8$ pontos. Nenhuma correlação forte foi identificada entre as qualidades médias, de ambas as explicações, com os atributos que medem o conhecimento prévio dos participantes. Entre os atributos sociodemográficos, identificou-se uma correlação negativa e estatisticamente significante: *médicos com idade maior tendem a dar notas menores para as explicações globais* (correlação de pearson = -0.56, $p = 0,007$). O mesmo não foi identificado com

explicações locais (correlação de pearson = -0.29, $p = 0,190$). Esse mesmo comportamento foi observado para os anos de experiência, indicando que profissionais mais experientes tendem a atribuir notas menores às explicações globais.

5. Discussão

Os resultados indicam que as explicações locais do SHAP foram avaliadas de maneira consistentemente mais positiva do que as explicações globais em todos os atributos analisados. As médias observadas para as explicações locais foram superiores nas dimensões de orientação ao usuário, orientação ao contexto, objetividade, clareza e capacidade informativa. Além disso, o indicador geral (qualidade média) também apresentou valor médio maior nas explicações locais ($3,75 \pm 0,8$) em comparação às globais ($3,47 \pm 0,99$).

Esse resultado sugere que os participantes perceberam as explicações locais como mais adequadas ao contexto clínico e mais úteis para compreender a decisão do modelo. Uma possível justificativa está relacionada à própria natureza das explicações locais do SHAP, especialmente o gráfico *Waterfall*, que apresenta de forma direta como cada variável contribui para uma predição individual. Por outro lado, as explicações globais apresentaram avaliações inferiores em todos os atributos. O gráfico *BeeSwarm* exige interpretação simultânea de múltiplas distribuições, cores e direções de impacto, o que pode aumentar a complexidade cognitiva. Embora as explicações locais tenham apresentado avaliações superiores em todos os atributos, o teste de Wilcoxon para amostras pareadas não identificou diferenças estatisticamente significativas após a correção de Holm ($p > 0,05$). Isso indica uma tendência favorável às explicações locais, mas sem evidência estatística suficiente para afirmar superioridade definitiva entre as abordagens na amostra analisada.

Entre os atributos avaliados nas explicações globais, a maior média foi observada para orientação ao usuário ($3,64 \pm 1,09$), seguida de orientação ao contexto ($3,59 \pm 1,10$) e objetividade ($3,45 \pm 1,06$). Já os menores valores ocorreram para clareza/compreensibilidade ($3,36 \pm 1,00$) e capacidade informativa ($3,32 \pm 1,17$). Nas explicações locais, o padrão foi semelhante. Os maiores valores foram observados para orientação ao usuário ($3,95 \pm 0,72$), orientação ao contexto ($3,91 \pm 0,81$) e objetividade ($3,77 \pm 0,87$), enquanto clareza ($3,55 \pm 1,14$) e capacidade informativa ($3,55 \pm 1,01$) apresentaram médias menores.

Esse comportamento sugere que os participantes perceberam as explicações do SHAP como relativamente alinhadas ao contexto clínico e ao perfil do usuário, mas ainda insuficientes em relação à clareza e à quantidade de informação necessária para compreender completamente o raciocínio do sistema. Em outras palavras, mesmo quando as explicações pareciam relevantes para o contexto médico, os participantes ainda encontraram dificuldades para interpretar plenamente o significado das informações apresentadas.

Os resultados do teste de Friedman indicaram que não houve diferença estatisticamente significativa entre os atributos avaliados nas explicações globais ($p = 0,123$). Já para as explicações locais, foi identificada uma diferença global significativa entre os atributos ($p = 0,012$), sugerindo que alguns aspectos foram percebidos de maneira mais positiva do que outros. As maiores médias nas explicações locais foram observadas em orientação ao usuário e orientação ao contexto, enquanto clareza e capacidade informativa

apresentaram valores relativamente menores. No entanto, após a realização dos pós-testes de Wilcoxon com correção de Holm, nenhuma comparação par a par permaneceu estatisticamente significativa ($p > 0{,}05$), indicando que as diferenças observadas entre atributos específicos não foram suficientemente fortes após o controle para múltiplas comparações.

Além disso, o fato de que apenas as explicações globais apresentaram correlação negativa moderada com a idade e com os anos de experiência, em contraste com a correlação fraca e não significativa observada nas explicações locais, sugere que o *BeeSwarm* é mais dependente do perfil e da idade do usuário para ser bem recebido, enquanto o *Waterfall* mostrou-se mais uniforme entre os participantes.

6. Considerações Finais

Os resultados obtidos reforçam a importância de considerar o perfil do usuário final no desenvolvimento de técnicas de explicabilidade para aplicações médicas. Embora o SHAP seja amplamente utilizado na literatura científica como ferramenta de interpretação de modelos preditivos, os resultados deste estudo indicam que sua forma de apresentação ainda pode representar dificuldades para médicos sem formação técnica em inteligência artificial.

A preferência relativa pelas explicações locais sugere que abordagens centradas em casos individuais podem ser mais adequadas para cenários clínicos reais, nos quais o profissional frequentemente precisa compreender a justificativa de uma decisão específica relacionada a um paciente. Por outro lado, explicações globais parecem exigir maior familiaridade com conceitos estatísticos e computacionais, reduzindo sua acessibilidade para usuários não especialistas.

Dessa forma, os achados deste trabalho indicam que a explicabilidade em saúde não depende apenas da fidelidade matemática da técnica utilizada, mas também da capacidade de comunicação da explicação. Isso evidencia a necessidade de desenvolver interfaces e estratégias de visualização que reduzam a complexidade interpretativa e aproximem os resultados do raciocínio clínico humano.

Por outro lado, considerando o baixo nível de conhecimento prévio dos participantes sobre conceitos de XAI (Tabela 3), as avaliações obtidas provavelmente não refletem apenas a percepção sobre as visualizações produzidas pela ferramenta SHAP. Em vez disso, elas representam a experiência conjunta proporcionada pelos gráficos e pelas descrições interpretativas que os acompanharam. Dessa forma, não é possível dissociar completamente o efeito da representação visual do efeito da contextualização textual fornecida aos participantes, o que deve ser considerado na interpretação e generalização dos resultados.

Por fim, além da análise da ferramenta SHAP, este trabalho contribui com a proposição e aplicação de um protocolo de avaliação de explicabilidade centrado em usuários finais da área da saúde. O instrumento, fundamentado em atributos reconhecidos na literatura de XAI, apresentou elevada consistência interna nas avaliações das explicações locais e globais, evidenciada pelos altos valores do coeficiente Alfa de Cronbach, o que reforça seu potencial como instrumento de pesquisa para a comparação de diferentes métodos de explicabilidade. Estudos futuros poderão utilizar esse protocolo em outros contextos clínicos, com diferentes perfis de usuários e técnicas de XAI.

Código e dados

Todo o código e os dados utilizados nesse estudo estão publicamente disponíveis em um repositório para auxiliar a reprodutibilidade durante o processo de revisão. O repositório pode ser acessado em: `https://github.com/mthperera/XAI-Research`

Declaração sobre uso de Inteligência Artificial

Os autores deste artigo utilizaram ferramentas de Inteligência Artificial com o objetivo de auxiliar na revisão gramatical e no aprimoramento da fluidez textual. Além disso, foi utilizado o Claude Sonnet 4.6 para gerar o código da análise dos dados.

Referências

[Bharati et al. 2023] Bharati, S., Mondal, M. R. H., and Podder, P. (2023). A review on explainable artificial intelligence for healthcare: why, how, and when? *arXiv*, abs/2304.04780.

[Chaddad et al. 2023] Chaddad, A., Peng, J., Xu, J., and Bouridane, A. (2023). Survey of explainable ai techniques in healthcare. *Sensors*, 23(2):634.

[da Silva 2023] da Silva, M. V. S. (2023). Explainable artificial intelligence on medical images: a survey. *arXiv*, abs/2305.07511.

[Hassija 2023] Hassija, V. (2023). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74.

[Kyrimi 2024] Kyrimi, E. (2024). Explainable ai: definition and attributes of a good explanation for health ai. *arXiv*, abs/2409.15338.

[Lundberg and Lee 2017] Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Markus et al. 2021] Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655.

[Nizri 2026] Nizri, M. (2026). Covid-19 dataset. `https://www.kaggle.com/datasets/meirnizri/covid19-dataset`. Kaggle dataset. Accessed: 2026-05-25.

[Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, NY. Association for Computing Machinery.

[Ribeiro et al. 2018] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: high-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[Salih et al. 2024] Salih, A. M., Galazzo, I. B., Gkontra, P., Rauseo, E., Lee, A. M., Lekadir, K., Radeva, P., Petersen, S. E., and Menegaz, G. (2024). A review of evaluation approaches for explainable ai with applications in cardiology.