Capítulo 2

Interpretação de Modelos de Aprendizado de Máquina Aplicados à Saúde com SHAP

Letícia Martins Raposo, Vinicius Navega Stelet

Abstract

The growing adoption of complex machine learning models in healthcare has substantially enhanced the predictive capability of computational systems while intensifying challenges related to interpretability and transparency of automated decisions. In this context, explainable artificial intelligence methods have become essential to foster trust, enable clinical validation, and support the responsible use of such models in sensitive domains. This chapter presents the theoretical, methodological, and practical foundations of predictive model interpretation through SHAP (SHapley Additive exPlanations), a method grounded in Cooperative Game Theory. Core concepts of interpretability and black-box models are discussed, followed by the mathematical formalization of Shapley values, the axiomatic properties of SHAP, and its formulation as an additive explanation model. The chapter examines local and global interpretations, implementation in R and Python with a cardiovascular risk case study, visualization strategies, and result analysis, as well as limitations, methodological challenges, best practices, and ethical and regulatory implications for the responsible use of SHAP in health computing.

Resumo

A crescente adoção de modelos complexos de aprendizado de máquina em saúde amplia a capacidade preditiva dos sistemas computacionais, mas intensifica os desafios de interpretabilidade e transparência das decisões automatizadas. Nesse contexto, métodos de inteligência artificial explicável tornaram-se fundamentais para promover confiança, validação clínica e uso responsável desses modelos em domínios sensíveis. Este capítulo apresenta os fundamentos teóricos, metodológicos e práticos da interpretação de modelos preditivos por meio do SHAP (SHapley Additive exPlanations), método fundamentado na Teoria dos Jogos Cooperativos. São discutidos os conceitos centrais de interpretabilidade e modelos caixa-preta, a formalização matemática dos valores de Shapley, as propriedades

axiomáticas do SHAP e sua formulação como modelo de explicação aditivo. O capítulo examina interpretações locais e globais, implementação em R e Python com estudo de caso em risco cardiovascular, estratégias de visualização e análise de resultados, além de limitações, desafios metodológicos, boas práticas e implicações éticas e regulatórias para o uso responsável do método em computação aplicada à saúde.

2.1. Introdução à Interpretabilidade em Aprendizado de Máquina para Saúde

Nas últimas décadas, o aprendizado de máquina tem se consolidado como uma ferramenta central na área da saúde, impulsionado pela crescente disponibilidade de dados clínicos digitais, pelo avanço do poder computacional e pelo desenvolvimento de algoritmos capazes de lidar com grandes volumes de dados heterogêneos. Modelos preditivos baseados em *ensembles*, redes neurais profundas e técnicas avançadas de otimização têm demonstrado desempenho competitivo em tarefas específicas, embora com variação importante entre contextos, dados e critérios de validação [Miotto et al. 2018, Rajkomar et al. 2018, Topol 2019].

Esses avanços se manifestam em múltiplos domínios da informática em saúde. Em prontuários eletrônicos, modelos preditivos são empregados para identificar pacientes de alto risco, antecipar reinternações e otimizar a alocação de recursos hospitalares [Goldstein et al. 2017, Mahmoudi et al. 2020]. Em imagens médicas, redes neurais convolucionais atingem desempenho comparável ou superior ao de especialistas humanos na detecção de retinopatia diabética [Bajwa et al. 2023], na classificação de lesões dermatológicas [Esteva et al. 2017] e na análise de exames radiológicos [Chen et al. 2024]. Na análise de biossinais, algoritmos de aprendizado de máquina têm sido aplicados com sucesso à detecção precoce de arritmias [Hannun et al. 2019] e de crises epilépticas [Daoud and Bayoumi 2019].

Apesar desses resultados promissores, o aumento do desempenho preditivo é frequentemente acompanhado por uma redução da transparência dos modelos. Muitos dos algoritmos mais eficazes operam como *modelos caixa-preta*, cujos mecanismos internos são difíceis de compreender mesmo para especialistas em ciência de dados [Lipton 2018, Binzagr 2024]. No contexto da saúde, essa opacidade é particularmente problemática, uma vez que decisões automatizadas podem influenciar diretamente diagnósticos, condutas terapêuticas e desfechos clínicos.

Nesse cenário, a interpretabilidade em aprendizado de máquina emerge como um campo de pesquisa fundamental para aplicações em saúde. Neste capítulo, adotamos *interpretabilidade* como a capacidade de compreender, em termos humanamente significativos, como um modelo relaciona entradas e saídas, e *explicabilidade* como o conjunto de métodos empregados para tornar esse comportamento inteligível, especialmente em modelos complexos. De forma mais específica, a interpretabilidade pode ser entendida como o grau em que um ser humano — seja um profissional de saúde, pesquisador ou gestor — consegue compreender as razões pelas quais um modelo produz determinada saída a partir de suas entradas clínicas [Stiglic et al. 2020]. Embora não exista uma definição única e universalmente aceita, o conceito está intimamente relacionado à possibilidade de atribuir significado clínico às decisões do modelo e de estabelecer relações compreensíveis entre variáveis de entrada e predições [Molnar 2020].

A relevância da interpretabilidade em aplicações médicas vai além de considerações

estritamente técnicas. Em contextos clínicos sensíveis, como triagem em unidades de emergência, estratificação de risco para doenças crônicas, diagnóstico assistido por computador e sistemas de alerta precoce em unidades de terapia intensiva, compreender os fatores que sustentam uma predição é tão importante quanto a própria acurácia do modelo [Tonekaboni et al. 2019]. Explicações interpretáveis são essenciais para promover confiança, possibilitar auditoria clínica, apoiar a tomada de decisão compartilhada e facilitar a adoção segura de sistemas baseados em aprendizado de máquina.

Além disso, a interpretabilidade desempenha papel central em debates éticos e regulatórios. Modelos opacos podem perpetuar vieses e desigualdades sem que tais problemas sejam facilmente detectáveis, como demonstrado em estudos que identificaram vieses raciais em algoritmos amplamente utilizados em sistemas de saúde [Obermeyer et al. 2019]. Paralelamente, legislações recentes, como o Regulamento Geral sobre a Proteção de Dados (GDPR, em inglês) na União Europeia, a Lei Geral de Proteção de Dados Pessoais (LGPD) no Brasil e regulamentações específicas para *Software as a Medical Device* (SaMD) por agências como FDA e ANVISA, reforçam a necessidade de transparência, explicabilidade e justificativa das decisões automatizadas em saúde [Goodman and Flaxman 2017, U.S. Food and Drug Administration 2021].

Nesse contexto, métodos de explicabilidade *post hoc* têm ganhado destaque por permitirem interpretar modelos complexos sem comprometer seu desempenho preditivo. Entre esses métodos, o SHAP (*SHapley Additive exPlanations*) destaca-se por sua fundamentação teórica na Teoria dos Jogos Cooperativos e por fornecer explicações consistentes, locais e globais para uma ampla classe de modelos [Lundberg and Lee 2017]. Sua capacidade de atribuir contribuições individuais às variáveis de entrada o torna particularmente útil em dados tabulares estruturados, frequentes em prontuários eletrônicos, desde que a interpretação produzida seja compreendida como explicação do comportamento do modelo e não como evidência causal sobre o fenômeno clínico.

O argumento central deste capítulo é que o SHAP pode ampliar a transparência e a auditabilidade de modelos preditivos em saúde, sem eliminar limitações relacionadas à qualidade dos dados, ao viés algorítmico, à dependência das escolhas metodológicas e à diferença entre associação preditiva e interpretação causal. Com esse enquadramento, o capítulo parte da caracterização dos modelos caixa-preta e da motivação para métodos de explicabilidade *post hoc*, avança para os fundamentos matemáticos dos valores de Shapley e sua formulação no SHAP, e então apresenta sua implementação prática em R e Python em um estudo de caso sobre risco cardiovascular. As seções finais discutem limitações, boas práticas e implicações éticas e regulatórias para o uso responsável do método em contextos clínicos.

2.2. Modelos Caixa-Preta e Métodos de Explicação

O uso crescente de modelos de aprendizado de máquina em saúde levanta uma questão central: em que medida decisões automatizadas podem ser consideradas confiáveis quando seus mecanismos internos não são transparentes para os usuários humanos? Esta seção examina o problema da opacidade em modelos complexos, suas implicações no contexto clínico e as principais estratégias desenvolvidas no campo de *Explainable Artificial Intelligence* (XAI), culminando na justificativa do uso do método SHAP.

2.2.1. Modelos caixa-preta e opacidade estrutural

Modelos caixa-preta são sistemas cujos processos internos, embora formalmente especificados, não são acessíveis à interpretação humana em razão de sua complexidade estrutural, interações não lineares e alta dimensionalidade [Burrell 2016, Bodria et al. 2021, Hassija et al. 2023]. Em aplicações clínicas, essa opacidade manifesta-se de diferentes formas. Redes neurais profundas aplicadas à análise de imagens médicas envolvem milhões de parâmetros distribuídos em múltiplas camadas convolucionais, tornando inviável a interpretação direta de seus pesos [Mienye et al. 2025]. Modelos baseados em arquiteturas *transformer*, utilizados em processamento de linguagem natural sobre registros clínicos, apresentam complexidade ainda maior, sendo empregados em tarefas como codificação automatizada e detecção de eventos adversos [Yang et al. 2022, Yuan et al. 2025]. De modo análogo, métodos de *ensemble*, como florestas aleatórias e algoritmos de *gradient boosting*, combinam múltiplos modelos base, dificultando a análise interpretativa de seu comportamento agregado [Stiglic et al. 2020].

A opacidade de modelos complexos não implica ausência de rigor metodológico, mas sim uma limitação epistêmica na tradução de seus mecanismos internos em explicações compreensíveis. Essa limitação possui implicações diretas na prática clínica: evidências indicam que profissionais de saúde tendem a resistir ou reinterpretar recomendações provenientes de sistemas cujo funcionamento não compreendem, mesmo quando estes apresentam alto desempenho preditivo [Tonekaboni et al. 2019, Tun et al. 2024].

2.2.2. Interpretabilidade intrínseca e seus limites

Uma resposta tradicional a esse problema consiste no uso de modelos interpretáveis por construção, cuja estrutura permite a compreensão direta de suas decisões. Exemplos clássicos na medicina incluem escores de risco como o SCORE2, o CHA_2DS_2–VASc e o SOFA, nos quais os parâmetros possuem interpretações clínicas claras [working group and risk collaboration 2021, Lip et al. 2010, Vincent et al. 1996]. Contudo, a adoção desses modelos impõe um compromisso entre interpretabilidade e capacidade preditiva.

Em cenários caracterizados por alta dimensionalidade e relações não lineares — como prontuários eletrônicos e séries temporais clínicas — modelos simples frequentemente apresentam desempenho preditivo inferior [Patharkar et al. 2024]. Esse compromisso torna-se particularmente problemático em contextos nos quais tanto a acurácia quanto a interpretabilidade possuem valor crítico.

2.2.3. Explicabilidade em modelos preditivos e métodos pós-hoc

Como resposta a esse problema, métodos de explicação pós-hoc têm sido amplamente desenvolvidos com o objetivo de tornar modelos complexos mais interpretáveis sem modificar sua estrutura interna. Tais métodos produzem explicações a partir da relação entre entradas e saídas do modelo, tratando-o, em muitos casos, como uma função caixa-preta [Hassija et al. 2023, Markus et al. 2021].

Do ponto de vista regulatório, essas abordagens são particularmente relevantes para atender exigências de transparência e prestação de contas em sistemas automatizados [Goodman and Flaxman 2017, U.S. Food and Drug Administration 2021]. No

contexto clínico, permitem a integração de recomendações automatizadas ao raciocínio médico, além de possibilitar a identificação de vieses, inconsistências e potenciais riscos associados ao uso dessas tecnologias [Tonekaboni et al. 2019, Amann et al. 2020, Obermeyer et al. 2019].

Entretanto, a adoção de métodos pós-hoc introduz uma tensão fundamental: as explicações produzidas constituem aproximações do comportamento do modelo, e não descrições fiéis de seus mecanismos internos. Como consequência, explicações de baixa fidelidade podem induzir interpretações equivocadas, com implicações clínicas potencialmente graves [Rudin 2019]. Esse problema evidencia a necessidade de critérios analíticos rigorosos para classificar, avaliar e selecionar métodos de explicação.

Nesse contexto, a literatura em XAI converge para a identificação de duas dimensões analíticas centrais: o **escopo da explicação** e o **grau de dependência do modelo**. Essas dimensões estruturam diferentes formas de acesso ao comportamento de modelos complexos e permitem compreender as limitações e potencialidades das abordagens existentes.

No que se refere ao escopo, distinguem-se **explicações globais** e **locais**. Explicações globais buscam caracterizar o comportamento médio do modelo ao longo da distribuição de dados, permitindo identificar padrões gerais, importância média de variáveis e relações funcionais predominantes. Essas abordagens são úteis para avaliar plausibilidade clínica, detectar vieses sistemáticos e apoiar validação externa, mas podem obscurecer comportamentos raros ou não lineares relevantes em subpopulações específicas [Hakkoum et al. 2024, Greenwell 2017, Lundberg et al. 2020].

Por outro lado, **explicações locais** focalizam instâncias individuais, decompondo predições específicas em contribuições das variáveis de entrada ou aproximando o modelo em uma vizinhança restrita do espaço de dados. Essa perspectiva é particularmente relevante em decisões clínicas centradas no paciente, nas quais a justificativa individual é essencial [Hakkoum et al. 2024, Salih et al. 2023, Vimbi et al. 2024]. Contudo, tais explicações podem apresentar instabilidade e limitada representatividade global, exigindo cautela em sua interpretação [Zafar and Khan 2021].

Paralelamente, a distinção entre métodos **específicos do modelo** (*model-specific*) e **agnósticos ao modelo** (*model-agnostic*) constitui uma segunda dimensão analítica relevante. Métodos específicos exploram propriedades internas de determinadas arquiteturas (por exemplo, árvores de decisão ou redes neurais convolucionais), frequentemente alcançando maior fidelidade explicativa e eficiência computacional dentro de seu domínio [Roshinta and Gábor 2024, Belle and Papantonis 2020, Lundberg et al. 2020]. No entanto, sua aplicabilidade é limitada à arquitetura para a qual foram desenvolvidos.

Em contraste, métodos agnósticos operam exclusivamente sobre a relação entrada–saída, ampliando sua aplicabilidade a diferentes modelos e contextos [Elshawi et al. 2019, Ribeiro et al. 2016]. Essa flexibilidade, contudo, é frequentemente acompanhada por maior custo computacional, possíveis problemas de estabilidade e limitações na captura de propriedades estruturais profundas do modelo [Roshinta and Gábor 2024, Lundberg et al. 2020].

À luz dessas dimensões, as principais estratégias de explicação podem ser compreendidas como respostas parciais a um problema comum: tornar modelos complexos

inteligíveis sem comprometer sua capacidade preditiva. Métodos baseados em importância de variáveis, como importância por permutação ou medidas derivadas de árvores, oferecem interpretações globais intuitivas, mas são sensíveis à correlação entre preditores e não capturam heterogeneidade individual [Greenwell and Boehmke 2020, Hooker et al. 2019]. Abordagens baseadas em modelos substitutos locais, como o *Local Interpretable Model-agnostic Explanations* (LIME), procuram explicar predições individuais por meio de aproximações locais [Ribeiro et al. 2016], mas sua dependência de perturbações aleatórias e da definição de vizinhança pode comprometer a estabilidade e a fidelidade das explicações [Zafar and Khan 2021, Tan et al. 2023].

Por fim, técnicas de visualização, como gráficos de dependência parcial e mapas de saliência, desempenham papel relevante na comunicação dos resultados, mas não resolvem, isoladamente, os desafios associados à fidelidade e à consistência das explicações [Greenwell 2017, Selvaraju et al. 2020, Brenning 2021, Singh et al. 2025]. Em particular, métodos globais podem mascarar efeitos condicionais importantes, enquanto abordagens locais frequentemente carecem de robustez.

2.2.4. SHAP como abordagem teoricamente fundamentada

As limitações discutidas anteriormente — incluindo a tensão entre interpretabilidade e fidelidade, a instabilidade de explicações locais, a sensibilidade a correlações entre variáveis e a ausência de garantias teóricas consistentes — convergem para a necessidade de métodos que conciliem propriedades frequentemente conflitantes: (i) fidelidade ao modelo original, (ii) consistência teórica, (iii) aplicabilidade a diferentes arquiteturas e (iv) capacidade de produzir explicações coerentes em níveis local e global.

É nesse contexto que métodos baseados em valores de Shapley se destacam como uma abordagem formalmente fundamentada. Derivados da teoria dos jogos cooperativos, os valores de Shapley oferecem um esquema de atribuição de importância a variáveis que satisfaz um conjunto de propriedades axiomáticas desejáveis, incluindo eficiência (a soma das contribuições corresponde à predição do modelo), simetria (variáveis com contribuições equivalentes recebem o mesmo valor) e aditividade [Lundberg and Lee 2017, Shapley 1953, Bifarin 2022].

O SHAP (*SHapley Additive exPlanations*) operacionaliza esses princípios no contexto de modelos de aprendizado de máquina, fornecendo uma decomposição aditiva da predição em contribuições atribuíveis a cada variável. Diferentemente de métodos como o LIME, que dependem de aproximações locais potencialmente instáveis, o SHAP estabelece uma conexão direta entre a explicação e propriedades formais do modelo, o que contribui para maior consistência e interpretabilidade [Lundberg et al. 2020, Ribeiro et al. 2016, Kelodjou et al. 2024]. Além disso, a estrutura aditiva dos valores de Shapley permite uma integração natural entre análises locais (explicações de instâncias individuais) e globais (agregação das contribuições ao longo da população).

Do ponto de vista computacional, avanços como o algoritmo TreeSHAP tornam viável a aplicação eficiente dessa abordagem em modelos amplamente utilizados na prática clínica, como árvores de decisão e métodos de *gradient boosting*, preservando exatidão polinomial na estimação dos valores [Lundberg et al. 2020, Nohara et al. 2021, Luo et al. 2024].

Apesar de suas vantagens, o SHAP não deve ser interpretado como uma solução definitiva. Embora ofereça garantias teóricas sob certas condições, sua interpretação depende de pressupostos sobre a distribuição e independência das variáveis, e diferentes escolhas metodológicas podem levar a explicações distintas. Ainda assim, o SHAP representa um avanço importante ao combinar fundamentos matemáticos bem definidos com aplicabilidade prática em problemas de interpretabilidade.

Dessa forma, o SHAP pode ser entendido como uma resposta teoricamente fundamentada às limitações das abordagens anteriores, oferecendo um quadro unificado para a análise explicativa de modelos complexos. As seções seguintes apresentam seus fundamentos matemáticos e suas aplicações no contexto da saúde.

2.3. Fundamentos da Teoria dos Jogos e Valores de Shapley

A interpretação de modelos preditivos pelo SHAP baseia-se na Teoria dos Jogos Cooperativos, que fornece um modelo formal para dividir, de forma justa, o resultado de uma cooperação entre diferentes participantes. No contexto de aprendizado de máquina, essa ideia é utilizada para decompor a predição de um modelo em contribuições individuais associadas a cada variável clínica, preservando propriedades matemáticas bem definidas [Shapley 1953, Nohara et al. 2021].

2.3.1. Jogos Cooperativos e Função Característica

Um jogo cooperativo é definido por um par (N, v), em que $N = \{1, 2, \ldots, n\}$ representa o conjunto de jogadores e $v : 2^N \to \mathbb{R}$ é a função característica, que associa um valor a cada subconjunto $S \subseteq N$. Esse subconjunto, chamado de *coalizão*, representa um grupo de jogadores considerados em conjunto. Em aprendizado de máquina aplicado à saúde, os "jogadores" correspondem às variáveis clínicas utilizadas no modelo [Nohara et al. 2021, Ghasemi et al. 2024, Vimbi et al. 2024, Ghosh and Khandoker 2024]. No contexto clínico, isso significa tratar cada variável observada como participante formal da predição produzida pelo modelo, e não como portadora, por si só, de significado causal ou terapêutico [Hettikankanamage et al. 2025].

Por exemplo, em um modelo de risco cardiovascular, os jogadores podem incluir idade, sexo, pressão arterial, colesterol, presença de diabetes e hábitos como tabagismo [working group and risk collaboration 2021]. A função característica $v(S)$ representa o desempenho do modelo quando apenas o conjunto de variáveis S está disponível. Assume-se que $v(\emptyset) = 0$, ou seja, na ausência de informação clínica, o modelo retorna apenas uma predição base, como a prevalência do desfecho na população [Ghosh and Khandoker 2024, Binzagr 2024, Covert et al. 2020, Shapley 1953].

Nesse contexto, o problema de explicabilidade pode ser formulado como um problema de alocação: quanto cada variável contribui para alterar a predição em relação a esse valor base?

Entretanto, calcular $v(S)$ na prática exige lidar com variáveis ausentes, o que implica realizar imputação ou marginalização. Essa escolha não é neutra: diferentes estratégias podem levar a explicações distintas. Em termos práticos, isso significa que a forma como as variáveis não observadas são tratadas pode alterar a explicação obtida. Além disso, ao

avaliar o modelo em combinações de variáveis que não ocorrem naturalmente nos dados, podem surgir cenários clinicamente implausíveis, o que dificulta a interpretação dos resultados [Hettikankanamage et al. 2025, Contreras et al. 2024, Hooshyar and Yang 2024].

2.3.2. O Problema da Alocação Justa de Contribuições

A questão central passa a ser: como distribuir o valor total $v(N)$ entre as variáveis de forma justa? Em termos práticos, isso corresponde a decompor uma predição em contribuições atribuíveis a cada variável [Nohara et al. 2021, Ghosh and Khandoker 2024, Salih et al. 2023].

Em aplicações clínicas, a alocação das contribuições possui implicações diretas para a interpretação e a tomada de decisão. Considere um modelo de predição de sepse em UTI que estima uma probabilidade de 78% para um paciente cuja probabilidade base é de 15%. Surge, então, a questão: como atribuir os 63 pontos percentuais de aumento a variáveis como lactato elevado, taquicardia e hipotensão? A decomposição dessa diferença em contribuições individuais constitui exatamente o tipo de explicação local necessário em contextos de alto risco clínico, nos quais a compreensão dos fatores que influenciam a predição é tão importante quanto o valor predito em si [Stenwig et al. 2022, Hu et al. 2022, Hu et al. 2024a].

No entanto, o conceito de "contribuição justa" é definido em termos matemáticos, não clínicos. Duas variáveis podem apresentar contribuições simétricas ao modelo preditivo e, ainda assim, possuir significados clínicos inteiramente distintos — uma associação espúria e um mecanismo fisiopatológico estabelecido podem, em princípio, receber contribuições equivalentes. Isso não invalida o método, mas delimita seu escopo: os valores atribuídos descrevem o comportamento do modelo, mas não necessariamente refletem relações causais no mundo real [Salih et al. 2023, Bifarin 2022, Hettikankanamage et al. 2025].

Os valores de Shapley se destacam porque são a única solução que satisfaz simultaneamente um conjunto de axiomas formais de equidade e consistência [Shapley 1953, Salih et al. 2023, Lundberg and Lee 2017], o que justifica sua adoção no SHAP.

2.3.3. Definição Formal dos Valores de Shapley

Os valores de Shapley, propostos por Shapley em 1953 [Shapley 1953], fornecem uma forma única de distribuir o valor total entre os jogadores. O valor $\phi_i(v)$ associado à variável i é dado por:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right] \tag{1}$$

Intuitivamente, esse valor ($\phi_i(v)$) corresponde à contribuição média da variável i ao longo de todas as possíveis combinações de variáveis. O termo $v(S \cup \{i\}) - v(S)$ mede quanto a inclusão de i melhora a predição, enquanto o fator combinatório pondera essa contribuição considerando todas as ordens possíveis de entrada das variáveis [Shapley 1953, Ghosh and Khandoker 2024].

Outra forma de interpretar esse valor ($\phi_i(v)$) é probabilística: ele representa a contribuição média de uma variável quando as variáveis são consideradas em ordem aleatória. No entanto, essa interpretação pressupõe independência entre variáveis, o que raramente ocorre em dados clínicos. Quando há correlação entre preditores, parte da contribuição pode refletir efeitos compartilhados entre variáveis [Hu et al. 2024a, Belle and Papantonis 2020].

2.3.4. Axiomas Caracterizadores

Os valores de Shapley são definidos por quatro axiomas principais:

Eficiência. $\sum_{i \in N} \phi_i(v) = v(N)$. A soma das contribuições reproduz exatamente a predição do modelo. No SHAP, essa propriedade corresponde à *local accuracy* [Lundberg and Lee 2017, Auzine et al. 2024]: a explicação é completa no sentido contábil. Contudo, completude contábil não implica completude causal; a eficiência garante que a predição seja integralmente decomposta, mas não que cada parcela corresponda a um mecanismo clínico independente [Hettikankanamage et al. 2025, Bifarin 2022].

Simetria. Se i e j contribuem igualmente para todas as coalizões, então $\phi_i(v) = \phi_j(v)$. Matematicamente, isso é uma propriedade de equidade formal [Shapley 1953, Bifarin 2022]. Clinicamente, porém, duas variáveis com efeito preditivo idêntico podem ter relevância terapêutica ou fisiopatológica inteiramente distinta. A simetria garante consistência na atribuição de contribuições, mas não substitui o julgamento clínico na interpretação de seu significado [Hettikankanamage et al. 2025].

Jogador nulo. Se a inclusão de i não altera nenhuma coalizão, então $\phi_i(v) = 0$. Variáveis sem poder preditivo não recebem contribuições espúrias — no SHAP, essa ideia corresponde à propriedade de *missingness* [Alabi et al. 2023, Auzine et al. 2024]. Note-se que esse axioma é definido em relação ao modelo, não ao fenômeno clínico: uma variável pode ser um fator causal estabelecido e, ainda assim, receber contribuição nula se sua informação for inteiramente redundante com outros preditores presentes no modelo.

Aditividade. Para dois jogos v e w, $\phi_i(v+w) = \phi_i(v) + \phi_i(w)$. Em termos simples, a contribuição de cada variável em um modelo combinado é igual à soma de suas contribuições em cada modelo separado. Isso permite analisar modelos complexos por partes, mantendo consistência nas explicações. No contexto clínico, em modelos *ensemble*, a explicação final pode ser entendida como a soma das explicações de cada modelo individual [Shapley 1953, Lundberg et al. 2019].

Esses axiomas garantem uma solução formalmente consistente para o problema de atribuição, mas não asseguram que as explicações reflitam relações causais.

2.3.5. Complexidade Computacional e Motivação para Aproximações

O cálculo exato dos valores de Shapley em (1) exige a avaliação de todas as possíveis combinações de variáveis, isto é, 2^n coalizões. Esse crescimento é exponencial: com 20 variáveis clínicas, são necessárias mais de um milhão de avaliações do modelo, enquanto

com 30 variáveis esse número ultrapassa um bilhão, tornando a abordagem inviável em aplicações reais [Nohara et al. 2021, Ghosh and Khandoker 2024, Salih et al. 2023]. Diante dessa limitação, foram desenvolvidas estratégias que tornam o SHAP aplicável na prática, destacando-se o KernelSHAP e o TreeSHAP.

O KernelSHAP é um método agnóstico ao modelo que estima os valores de Shapley por meio da amostragem de diferentes coalizões e ajuste de uma regressão linear ponderada [Lundberg et al. 2019, Olsen and Jullum 2025]. Sua principal vantagem é a generalidade, permitindo sua aplicação a uma ampla variedade de modelos, incluindo redes neurais, máquinas de vetor de suporte e *ensembles*. No entanto, essa flexibilidade tem custo: o método pode exigir um número elevado de avaliações do modelo para explicar uma única instância e, por depender de amostragem, introduz variabilidade nas estimativas. Em cenários com muitas variáveis ou forte dependência entre preditores, diferentes execuções podem gerar explicações distintas, o que exige cuidado na configuração dos parâmetros e na análise de robustez [Kelodjou et al. 2024, Olsen and Jullum 2025]. Extensões como DeepSHAP e GradientSHAP procuram reduzir o custo computacional em modelos mais complexos, mas mantêm o caráter aproximado da abordagem [Lundberg and Lee 2017].

O TreeSHAP, por sua vez, explora a estrutura de modelos baseados em árvores para calcular valores SHAP exatos de forma eficiente, com complexidade $O(TLD^2)$ por instância, onde T é o número de árvores, L o número médio de folhas e D a profundidade máxima [Lundberg et al. 2020]. Em modelos de *gradient boosting*, amplamente utilizados em dados clínicos, esse cálculo pode ser realizado em milissegundos por paciente, o que viabiliza sua aplicação em larga escala [Nohara et al. 2021, Luo et al. 2024, Hu et al. 2024a]. Ainda assim, a exatidão computacional não elimina limitações interpretativas: os valores obtidos refletem o comportamento do modelo treinado, incluindo possíveis vieses, padrões espúrios e dependências entre variáveis. Em particular, a presença de multicolinearidade pode levar à redistribuição das contribuições entre variáveis correlacionadas, dificultando a interpretação individual de seus efeitos [Salih et al. 2023, Hu et al. 2024a, Hettikankanamage et al. 2025].

Em síntese, enquanto o KernelSHAP oferece maior flexibilidade ao custo de variabilidade e maior demanda computacional, o TreeSHAP proporciona eficiência e determinismo, mas permanece dependente da estrutura do modelo e dos dados. Essa distinção é central para a interpretação dos resultados, uma vez que os valores SHAP devem ser analisados considerando tanto o método de estimação quanto o modelo subjacente. Em aplicações com dados tabulares de prontuários eletrônicos, nas quais modelos de *gradient boosting* são predominantes, o TreeSHAP tende a ser a escolha mais viável, permitindo a obtenção de explicações locais e globais em escala clínica.

2.4. SHAP: Formulação Teórica e Propriedades

O SHAP (*SHapley Additive Explanations*), proposto por Lundberg e Lee em 2017, utiliza os valores de Shapley para explicar modelos de aprendizado de máquina. A ideia central é decompor a predição de um modelo em contribuições individuais associadas a cada variável, permitindo interpretar de forma transparente como cada fator influencia o resultado [Lundberg and Lee 2017, Nohara et al. 2021, Ghosh and Khandoker 2024].

Essa abordagem fornece uma base matemática rigorosa para atribuir importância

às variáveis, sendo aplicável a diferentes tipos de modelos preditivos. Nesta seção, são apresentados os principais elementos da formulação do SHAP, bem como suas propriedades teóricas, incluindo um exemplo ilustrativo no contexto de risco cardiovascular.

Apesar de sua fundamentação formal, é importante destacar que a interpretação dos resultados não é automática. As explicações produzidas pelo SHAP dependem de suposições sobre a distribuição dos dados e sobre como as variáveis interagem entre si no modelo. Assim, mesmo sendo matematicamente consistentes, essas explicações podem variar conforme as escolhas metodológicas adotadas [Hettikankanamage et al. 2025, Belle and Papantonis 2020].

2.4.1. SHAP como Modelo de Explicação Aditivo

O SHAP pertence à classe dos modelos de explicação aditivos, nos quais a predição de um modelo é representada como a soma de contribuições individuais associadas a cada variável [Lundberg and Lee 2017, Binzagr 2024, Alabi et al. 2023]. Seja $f : \mathbb{R}^p \to \mathbb{R}$ um modelo preditivo com p variáveis de entrada. Para uma instância $x \in \mathbb{R}^p$, a explicação SHAP é dada por:

$$f(x) = \phi_0 + \sum_{i=1}^{p} \phi_i(x), \tag{2}$$

onde $\phi_0 = \mathbb{E}[f(X)]$ é o **valor base**, isto é, a predição média do modelo na população de referência, e $\phi_i(x)$ representa a **contribuição SHAP** da variável i, indicando quanto essa variável altera a predição em relação a esse valor base [Nohara et al. 2021, Alabi et al. 2023, Binzagr 2024]. Essa propriedade, conhecida como *local accuracy*, garante que a soma das contribuições individuais seja exatamente igual à predição do modelo, sem deixar parte não explicada [Lundberg and Lee 2017, Auzine et al. 2024, Alabi et al. 2023].

Na prática, isso significa que a predição é "quebrada" em partes atribuídas a cada variável. Essa forma de decomposição facilita a interpretação, pois permite entender como cada fator contribui para o resultado final. No entanto, essa representação também envolve simplificações. Ao distribuir a predição entre variáveis individuais, o SHAP assume que os efeitos podem ser separados dessa maneira, mesmo quando existem interações complexas entre elas [Lundberg et al. 2020, Lundberg et al. 2019].

Por exemplo, em um modelo de risco cardiovascular, a combinação de hipertensão e diabetes pode ter um efeito conjunto relevante. No SHAP, esse efeito é dividido entre as variáveis com base em contribuições médias. Embora essa divisão seja matematicamente bem definida, ela nem sempre corresponde a uma interpretação clínica direta.

Assim, o caráter aditivo do SHAP traz duas consequências importantes: por um lado, torna as explicações mais claras e consistentes; por outro, pode simplificar interações complexas, que na prática não são facilmente separáveis em contribuições independentes.

2.4.2. Definição da Função Característica

Para aplicar os valores de Shapley em modelos preditivos, é necessário definir a chamada função característica. No contexto do SHAP, essa função é dada por:

$$v_x(S) = \mathbb{E}\left[f(X) \mid X_S = x_S\right], \tag{3}$$

onde $S \subseteq N$ representa um subconjunto de variáveis observadas, x_S são os valores dessas variáveis para a instância x, e a expectativa é calculada considerando a distribuição das variáveis não observadas $X_{\bar{S}}$ [Lundberg and Lee 2017, Covert et al. 2020]. Em termos simples, $v_x(S)$ responde à seguinte pergunta: *qual seria a predição esperada do modelo se conhecêssemos apenas as variáveis em S?*

Essa definição envolve uma escolha importante: como lidar com as variáveis ausentes ao calcular essa expectativa. Na prática, isso significa decidir como "preencher" os valores de $X_{\bar{S}}$. Uma abordagem comum utiliza a distribuição marginal dos dados, isto é, cada variável ausente é considerada de forma independente das demais. Embora simples, essa estratégia pode levar o modelo a avaliar combinações de características que não ocorrem na prática clínica. Por exemplo, ao fixar um valor elevado de pressão arterial sistólica e estimar a pressão diastólica de forma independente, o modelo pode gerar pares de valores que não são fisiologicamente plausíveis [Salih et al. 2023, Hooker et al. 2019, Hooshyar and Yang 2024].

Esse problema se torna ainda mais relevante em dados clínicos, nos quais muitas variáveis estão naturalmente correlacionadas, como pressão sistólica e diastólica, creatinina e ureia, ou índice de massa corporal e circunferência abdominal. Quando essas relações são ignoradas, o modelo passa a ser avaliado em regiões do espaço de dados que não foram observadas durante o treinamento, o que pode comprometer a interpretação das contribuições atribuídas [Salih et al. 2023, Hu et al. 2024a, Hooker et al. 2019].

Uma alternativa consiste em utilizar a distribuição condicional das variáveis ausentes, isto é, considerar $X_{\bar{S}}$ condicionado a $X_S = x_S$. Nesse caso, os valores ausentes são gerados de forma consistente com as variáveis observadas. Retomando o exemplo anterior, ao fixar uma pressão sistólica elevada, a pressão diastólica seria estimada dentro de uma faixa compatível com esse valor, preservando relações fisiológicas conhecidas. Essa abordagem, conhecida como SHAP condicional, tende a produzir cenários mais realistas [Olsen and Jullum 2025, Covert et al. 2020].

No entanto, essa escolha também altera o significado das explicações. Enquanto a abordagem marginal busca isolar o efeito de cada variável, a abordagem condicional passa a atribuir contribuições levando em conta o contexto das demais variáveis, o que pode redistribuir a importância entre preditores correlacionados. Além disso, a implementação é mais complexa e pode não preservar integralmente algumas propriedades teóricas do SHAP original [Lundberg and Lee 2017, Salih et al. 2023, Olsen and Jullum 2025, Hettikankanamage et al. 2025].

Assim, a escolha entre utilizar a distribuição marginal ou condicional não é apenas técnica: ela define o tipo de cenário hipotético considerado na explicação e influencia diretamente o significado dos valores SHAP obtidos.

2.4.3. Cálculo dos Valores SHAP

A contribuição de cada variável no SHAP é dada pelo valor de Shapley aplicado à função característica v_x. Para uma variável i e uma instância x, esse valor é definido por:

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(p-|S|-1)!}{p!} \left[v_x(S \cup \{i\}) - v_x(S) \right], \tag{4}$$

onde $\phi_i(x)$ representa a contribuição da variável i para a predição, calculada como a média de seu efeito marginal ao ser adicionada a diferentes combinações das demais variáveis [Lundberg and Lee 2017, Nohara et al. 2021, Binzagr 2024, Bifarin 2022]. O valor base da predição corresponde a $\phi_0 = v_x(\emptyset) = \mathbb{E}[f(X)]$.

Em termos intuitivos, esse cálculo considera todas as formas possíveis de incluir a variável i no modelo e mede quanto ela altera a predição em cada caso. O valor SHAP final é a média desses efeitos, ponderada de forma a garantir propriedades de equidade e consistência.

Quando há correlação entre variáveis (colinearidade), a interpretação se torna mais delicada. Nesse caso, o método distribui o efeito conjunto entre as variáveis correlacionadas, com base na média de suas contribuições marginais. Isso significa que, se duas variáveis carregam informações semelhantes, a contribuição atribuída a cada uma depende de como o modelo aprendeu essa relação durante o treinamento. Como resultado, a decomposição continua sendo única do ponto de vista matemático, mas sua interpretação clínica pode ser ambígua: parte do efeito atribuído a uma variável pode, na prática, refletir a influência de outra variável correlacionada [Nohara et al. 2021, Salih et al. 2023, Hu et al. 2024a].

Como o cálculo direto dessa expressão é inviável em modelos com muitas variáveis, métodos como TreeSHAP e KernelSHAP são utilizados para estimar os valores SHAP de forma eficiente em aplicações práticas [Nohara et al. 2021, Ghosh and Khandoker 2024, Luo et al. 2024, Lundberg and Lee 2017, Lundberg et al. 2020].

2.4.4. Exemplo Clínico: Estratificação de Risco Cardiovascular

Para ilustrar o uso dos valores SHAP, considere um modelo de *gradient boosting* treinado para prever a presença de doença cardiovascular a partir de variáveis demográficas, clínicas e comportamentais. Entre os preditores estão idade, sexo, pressão arterial sistólica e diastólica, índice de massa corporal (IMC), níveis de colesterol e glicemia, além de indicadores de tabagismo, consumo de álcool e atividade física. O valor base do modelo, $\phi_0 \approx -0{,}02$, corresponde à predição média no conjunto de treinamento, expressa em *log-odds*.

Considere agora um paciente do sexo masculino, com 52,4 anos, pressão arterial sistólica de 140 mmHg, diastólica de 90 mmHg, IMC de 30,1 kg/m^2, colesterol e glicemia normais (`cholesterol` = `gluc` = 1), não tabagista, não etilista e sedentário. Para esse paciente, o SHAP decompõe a predição da seguinte forma:

$$\begin{aligned} f(x) = -0{,}02 + \underbrace{1{,}32}_{\text{PAS=140}} + \underbrace{0{,}18}_{\text{PAD=90}} + \underbrace{0{,}14}_{\text{Sedentarismo}} \\ + \underbrace{(-0{,}06)}_{\text{Idade}} + \underbrace{(-0{,}04)}_{\text{Colesterol}} + \underbrace{0{,}01}_{\text{IMC}} + \underbrace{(-0{,}01)}_{\text{Demais}} \approx 1{,}54 \end{aligned}$$

Essa decomposição mostra como cada variável contribui para o valor final da predição. A pressão arterial sistólica elevada é o principal fator, respondendo pela maior parte do aumento do risco. A pressão diastólica e o sedentarismo também contribuem positivamente, mas com menor intensidade. Por outro lado, o fato de o paciente ter colesterol normal e idade intermediária reduz parcialmente o risco previsto [Ponce-Bobadilla et al. 2024, Stenwig et al. 2022].

No entanto, essa interpretação deve ser feita com cautela. A pressão sistólica e a diastólica são variáveis fisiologicamente relacionadas, e o SHAP distribui o efeito conjunto entre elas com base em como o modelo aprendeu essa relação. Assim, a maior contribuição atribuída à pressão sistólica não significa necessariamente que seu efeito seja independente ou mais "importante" do ponto de vista clínico. Mudanças na distribuição dos dados ou no próprio modelo podem alterar essa divisão das contribuições sem modificar a predição final [Salih et al. 2023].

Portanto, os valores SHAP devem ser entendidos como uma descrição de como o modelo utiliza as variáveis para gerar a predição, e não como uma representação direta dos mecanismos causais envolvidos no risco cardiovascular.

2.4.5. Considerações sobre Escala de Explicação

Em modelos de classificação binária, os valores SHAP podem ser expressos em duas escalas: probabilidade ou *log-odds* [Lundberg and Lee 2017, Ponce-Bobadilla et al. 2024]. Embora essa escolha não altere a ordem de importância das variáveis, ela influencia a forma como os resultados são interpretados e comunicados.

Na **escala de probabilidade**, as contribuições indicam diretamente quanto cada variável aumenta ou reduz a probabilidade do desfecho, $p(x) = \mathbb{P}(Y = 1 \mid x)$. Essa forma é mais intuitiva e, por isso, costuma ser mais adequada para comunicação clínica. No entanto, como a relação entre *log-odds* e probabilidade é não linear, a decomposição pode não ser estritamente aditiva. Isso significa que o impacto de uma variável pode parecer maior ou menor dependendo do nível de risco basal do paciente, mesmo que seu efeito real no modelo seja constante [Ponce-Bobadilla et al. 2024].

Na **escala de *log-odds***, a decomposição é exatamente aditiva:

$$\text{logit}(p(x)) = \phi_0 + \sum_{j=1}^{p} \phi_j^{(\text{logit})}(x) \tag{5}$$

Nessa escala, cada contribuição pode ser interpretada de forma multiplicativa em termos de *odds ratio* (e^{ϕ_j}), o que é particularmente útil em análises epidemiológicas e na comparação entre fatores de risco [Ponce-Bobadilla et al. 2024, Lundberg and Lee 2017, Nohara et al. 2021].

Na prática, uma abordagem comum é utilizar a escala de *log-odds* para análise técnica, especialmente ao estudar interações entre variáveis, e a escala de probabilidade para comunicação com profissionais de saúde. Em qualquer caso, é importante explicitar qual escala está sendo utilizada, pois isso afeta a interpretação dos resultados. As propriedades teóricas do SHAP permanecem válidas em ambas as escalas, desde que a função explicada

seja claramente definida [Lundberg and Lee 2017, Ponce-Bobadilla et al. 2024].

A seção seguinte discute como essas contribuições podem ser organizadas em análises locais e globais voltadas ao suporte à decisão clínica, retomando essas limitações no contexto de aplicações práticas.

2.5. Interpretação Local e Global com SHAP

O SHAP permite interpretar modelos preditivos em dois níveis complementares: o nível local e o nível global. No nível local, o objetivo é explicar uma predição específica, ou seja, entender como cada variável contribuiu para o resultado de um indivíduo. Já no nível global, busca-se compreender o comportamento geral do modelo, identificando padrões médios, variáveis mais influentes e possíveis relações entre os preditores [Bodria et al. 2021, Hakkoum et al. 2024].

No contexto da saúde, essa distinção é especialmente importante. Explicações locais são úteis para apoiar decisões clínicas individualizadas, permitindo que o profissional entenda por que um paciente foi classificado como de alto ou baixo risco. Por outro lado, análises globais ajudam a avaliar o modelo como um todo, identificar fatores de risco em nível populacional e atender a exigências regulatórias relacionadas à transparência e validação [Salih et al. 2023, Ponce-Bobadilla et al. 2024, Bachmann 2025, Joachim et al. 2026].

Uma característica importante do SHAP é que esses dois níveis estão diretamente conectados. As explicações globais são obtidas a partir da agregação das explicações locais, o que garante coerência entre a interpretação de casos individuais e o comportamento médio do modelo [Lundberg et al. 2020, Bodria et al. 2021]. Isso diferencia o SHAP de métodos como o LIME, que se concentram apenas em explicações locais e não oferecem, de forma direta, uma visão global consistente [Alabi et al. 2023, Vimbi et al. 2024].

2.5.1. Interpretação Local: Explicando Decisões Individuais

A interpretação local no SHAP tem como objetivo explicar a predição feita para um paciente específico. Isso é feito decompondo o valor previsto $f(x)$ em contribuições individuais de cada variável, de acordo com a expressão:

$$f(x) = \phi_0 + \sum_{i=1}^{p} \phi_i(x) \tag{6}$$

Nessa decomposição, ϕ_0 representa o valor base do modelo (a predição média), enquanto cada $\phi_i(x)$ indica quanto a variável i aumenta ou reduz a predição para aquele paciente.

Na prática, isso permite identificar, para cada paciente, quais fatores estão mais associados ao aumento ou à redução do risco e qual é a magnitude de cada efeito. Esse tipo de explicação é particularmente útil na clínica, pois conecta diretamente o resultado do modelo com características observáveis do paciente [Stenwig et al. 2022, Hu et al. 2024a, Luo et al. 2024].

Essa capacidade de rastrear a origem da predição é especialmente importante em sistemas de alerta precoce. Por exemplo, em uma unidade de terapia intensiva (UTI), um

alerta de sepse pode ser explicado por contribuições elevadas de variáveis como lactato e pressão arterial baixa. Nesse caso, o modelo não apenas sinaliza o risco, mas também indica quais fatores estão mais associados a esse alerta, oferecendo ao profissional de saúde uma justificativa clara e imediatamente interpretável [Hu et al. 2022].

Por serem calculados em relação a um mesmo valor base e obedecerem aos axiomas de Shapley, os valores SHAP são diretamente comparáveis entre indivíduos [Lundberg and Lee 2017, Binzagr 2024]. Em outras palavras, o efeito de uma variável tem o mesmo significado quantitativo em todos os indivíduos, o que permite analisar padrões de forma consistente.

Essa propriedade possibilita diferentes aplicações relevantes na prática clínica. Uma delas é a identificação de fenótipos clínicos: ao comparar os padrões de contribuição entre pacientes, é possível identificar subgrupos com perfis de risco distintos, como aqueles em que o risco é mais influenciado por fatores renais ou por fatores cardiovasculares [Luo et al. 2024, Bifarin 2022, Lundberg et al. 2020].

Outra aplicação é o monitoramento longitudinal. Ao acompanhar um mesmo paciente ao longo do tempo, pode-se observar como as contribuições das variáveis mudam, indicando alterações no perfil de risco [Ponce-Bobadilla et al. 2024, Patharkar et al. 2024]. Isso pode ajudar, por exemplo, a avaliar a evolução clínica ou a resposta a intervenções.

Além disso, a comparabilidade dos valores SHAP permite identificar casos atípicos. Quando um paciente apresenta um padrão de explicação muito diferente dos demais, isso pode indicar tanto um perfil clínico incomum quanto possíveis limitações do modelo, como falta de representatividade desse tipo de caso nos dados de treinamento [Salih et al. 2023, Auzine et al. 2024, Liu et al. 2022].

2.5.2. Interpretação Global por Agregação Estatística

As interpretações globais no SHAP são obtidas a partir da agregação dos valores individuais calculados para cada paciente em uma amostra representativa [Lundberg and Lee 2017, Ponce-Bobadilla et al. 2024]. Em vez de analisar um caso isolado, o objetivo aqui é entender como o modelo se comporta, em média, na população.

Duas medidas são particularmente importantes nesse contexto. A primeira é a **importância global média**, que indica o quanto, em média, cada variável contribui para a predição, independentemente da direção do efeito. Ela é definida como:

$$I_i = \mathbb{E}\left[|\phi_i(X)|\right] \approx \frac{1}{n}\sum_{j=1}^{n} |\phi_i(x_j)| \tag{7}$$

Essa métrica permite identificar quais variáveis têm maior impacto no modelo como um todo [Wang et al. 2024, Bifarin 2022].

A segunda medida avalia a **direcionalidade do efeito**. Em vez de usar a média aritmética das contribuições brutas — que, por definição matemática do SHAP, se aproxima de zero ao longo do conjunto de dados de referência [Nohara et al. 2021] —, a direção é frequentemente quantificada pela correlação (por exemplo, coeficiente de Spearman) entre os valores da variável e os seus respectivos valores SHAP [Joachim et al. 2026]. Essa

medida indica se o aumento de uma variável tende a aumentar ou reduzir o risco.

A análise conjunta da importância média global e da direcionalidade é útil para identificar padrões mais complexos. Quando o efeito de uma variável é consistente na população (por exemplo, sempre aumentando o risco), a correlação tende a ser forte. No entanto, quando há heterogeneidade, isto é, quando a variável aumenta o risco em alguns pacientes e reduz em outros, a importância global média pode permanecer alta, enquanto a medida de direcionalidade se aproxima de zero devido ao cancelamento entre efeitos positivos e negativos [Ponce-Bobadilla et al. 2024, Joachim et al. 2026]. Esse comportamento indica que o efeito da variável não é uniforme na população.

Além disso, essas medidas permitem comparar os resultados do modelo com o conhecimento epidemiológico existente e explorar padrões mais complexos, como relações não lineares e interações entre variáveis, especialmente quando combinadas com técnicas de visualização apropriadas [Ghasemi et al. 2024, Bifarin 2022]. Para além das métricas agregadas, a análise da **distribuição completa dos valores SHAP** fornece uma visão mais rica do comportamento do modelo. Uma alta dispersão nas contribuições de uma variável pode indicar diferentes fenômenos: relações não lineares (como a presença de limiares clínicos), interações com outras variáveis (por exemplo, quando o efeito de um fator depende da presença de outro) ou heterogeneidade real e presença de efeitos raros de alta magnitude entre pacientes [Ponce-Bobadilla et al. 2024, Cha et al. 2021, Lundberg et al. 2020].

Esses padrões de heterogeneidade e não linearidade podem ser explorados por meio de visualizações, que serão apresentadas mais adiante. Os *beeswarm plots* permitem observar a distribuição das contribuições de cada variável em toda a amostra, enquanto os *dependence plots* mostram como essas contribuições variam em função do valor da variável e de possíveis interações [Ponce-Bobadilla et al. 2024, Lundberg et al. 2020]. Essas visualizações tornam visíveis relações não lineares e efeitos condicionais que não são capturados por métricas baseadas em médias.

Por fim, a análise global pode ser aprofundada por meio da **estratificação por subgrupos clínicos relevantes**. Comparar, por exemplo, pacientes cirúrgicos e clínicos, diferentes faixas etárias ou a presença de comorbidades permite avaliar se o modelo se comporta de forma consistente entre populações distintas. Essa etapa é fundamental para detectar vieses, limitações de generalização e possíveis diferenças na relevância das variáveis entre grupos [Hakkoum et al. 2024, Lundberg et al. 2020].

2.5.3. Implicações para Validação e Auditoria Regulatória

O uso do SHAP tem se tornado cada vez mais relevante em processos de validação clínica e auditoria regulatória. Embora ainda não exista uma padronização formal consolidada, diferentes aplicações já demonstram como essas análises podem apoiar a avaliação e o monitoramento de modelos preditivos.

Uma primeira aplicação importante é a **validação de plausibilidade clínica**. Por meio de análises globais, é possível verificar se o modelo utiliza as variáveis de forma consistente com o conhecimento médico estabelecido. Em estudos multicêntricos, por exemplo, o SHAP tem sido utilizado para confirmar que os principais preditores identificados pelo modelo correspondem a fatores de risco reconhecidos na literatura. Quando surgem incon-

sistências relevantes, isso pode indicar problemas nos dados, presença de confundimento ou falhas no próprio modelo [Prendin et al. 2023, Zeng et al. 2025, Kheder et al. 2025].

Outra aplicação central é a **detecção de vieses algorítmicos**. Ao analisar as contribuições das variáveis em diferentes subgrupos — como sexo, raça ou condição socioeconômica — é possível identificar padrões discriminatórios. Esse tipo de análise também ajuda a distinguir entre efeitos biológicos reais e variáveis que funcionam como proxies de desigualdade no acesso ao sistema de saúde [Momenzadeh et al. 2022, Hettikankanamage et al. 2025, Obermeyer et al. 2019, Chen et al. 2021].

O SHAP também tem sido utilizado na **documentação regulatória**, especialmente em sistemas classificados como SaMD. Nesses casos, análises locais e globais são incluídas na documentação técnica para permitir a verificação externa do comportamento do modelo e facilitar a reprodutibilidade dos resultados [Singh et al. 2025, Kiseleva et al. 2022, U.S. Food and Drug Administration 2021].

Além disso, o SHAP pode apoiar o **monitoramento pós-implantação**. A análise periódica das contribuições permite identificar possíveis mudanças no comportamento do modelo ao longo do tempo, fenômeno conhecido como *drift*, que pode ocorrer devido a alterações na população atendida ou nas práticas clínicas. Esse tipo de monitoramento contribui para auditorias contínuas de desempenho e equidade (*fairness*), embora ainda seja mais comum como prática conceitual do que como exigência regulatória formal [Lundberg et al. 2020, Stogiannos et al. 2023, Singh et al. 2025].

Em conjunto, essas aplicações mostram que o SHAP não se limita à explicação de casos individuais. Ele também pode ser utilizado como ferramenta de validação, governança e monitoramento de modelos ao longo de todo o seu ciclo de vida.

2.6. Implementação Computacional do SHAP em R e Python

Esta seção apresenta, de forma paralela, a implementação do SHAP em R e em Python para um mesmo problema clínico: a classificação binária de risco cardiovascular. Em ambas as linguagens, utiliza-se o mesmo conjunto de variáveis e um modelo XGBoost equivalente. O objetivo é permitir que o leitor acompanhe as etapas nas duas abordagens, compare suas estruturas e compreenda boas práticas de reprodutibilidade.

Os experimentos são realizados com uma base de dados pública, anonimizada e amplamente utilizada em estudos metodológicos de aprendizado de máquina em saúde, composta por variáveis demográficas, clínicas e comportamentais associadas ao risco cardiovascular[1]. O fluxo de trabalho adotado é o mesmo em R e Python e inclui cinco etapas principais: (1) preparação dos dados e particionamento estratificado em treino e teste; (2) ajuste de um modelo XGBoost; (3) avaliação do desempenho por meio da AUC-ROC e da calibração; (4) cálculo de valores SHAP em nível local e global; e (5) geração de visualizações voltadas à interpretação clínica.

Os códigos apresentados a seguir foram organizados para serem autocontidos e executáveis no Google Colab[2], sem necessidade de configuração adicional do ambiente.

[1]Disponível em: `https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset`

[2]`https://colab.research.google.com`

No caso do Python, basta utilizar o kernel padrão da plataforma. Para a implementação em R, é necessário selecionar a opção *Ambiente de execução* > *Alterar o tipo de ambiente de execução* > *R* antes de iniciar a sessão. Em ambos os casos, o arquivo `cardio_train.csv` deve ser carregado previamente, seja pelo painel *Files* do Colab, seja por código de upload executado no próprio notebook. A instalação das dependências é feita na primeira célula de cada implementação e, uma vez concluída, pode ser dispensada nas execuções seguintes dentro da mesma sessão.

Cabe destacar que o objetivo desta seção não é apresentar um tutorial completo de construção de modelos preditivos, mas sim ilustrar, de forma prática, como aplicar o SHAP em um conjunto de dados clínicos. O foco está na interpretação do modelo já ajustado, mostrando como extrair e analisar as contribuições das variáveis em nível local e global. Dessa forma, as etapas de modelagem são apresentadas de maneira objetiva, apenas na medida necessária para viabilizar o uso e a compreensão das explicações geradas pelo SHAP.

2.6.1. Estrutura dos Dados e Pré-processamento

Considerando o conjunto de dados descrito anteriormente, as variáveis preditoras abrangem diferentes dimensões clínicas relevantes para o risco cardiovascular. Entre as variáveis **demográficas**, incluem-se idade (convertida de dias para anos) e sexo. As variáveis **antropométricas** compreendem o IMC, calculado a partir de altura e peso. Os **sinais vitais** incluem as pressões arteriais sistólica e diastólica, medidas em mmHg. O conjunto também contempla **marcadores metabólicos categóricos**, como níveis de colesterol e glicemia codificados em categorias ordinais, além de **fatores comportamentais**, como tabagismo, consumo de álcool e prática de atividade física, representados como variáveis binárias.

A variável resposta é binária, $Y \in \{0, 1\}$, em que $Y = 1$ indica a presença de doença cardiovascular.

Antes do treinamento do modelo, algumas etapas de pré-processamento são particularmente importantes em dados clínicos. Em primeiro lugar, **valores ausentes** são frequentes em bases de prontuário eletrônico. Modelos como XGBoost e LightGBM lidam diretamente com dados faltantes, o que reduz a necessidade de intervenção. Ainda assim, pode-se optar por métodos de imputação desde que essa escolha seja explicitamente documentada.

Em relação às **variáveis categóricas**, é necessário convertê-las para formato numérico. Uma prática comum é utilizar *label encoding* para variáveis ordinais (como colesterol e glicemia) e *one-hot encoding* para variáveis nominais, o que é compatível com o uso posterior do TreeSHAP.

Quanto à **escala das variáveis**, modelos baseados em árvores são invariantes a transformações monotônicas, o que dispensa normalização. No entanto, manter as unidades clínicas originais facilita a interpretação dos valores SHAP.

Por fim, a presença de **outliers** deve ser avaliada com cuidado. Valores clinicamente implausíveis — como níveis extremos de pressão arterial — podem distorcer tanto o treinamento do modelo quanto as explicações geradas, devendo ser investigados e tratados

previamente.

Cabe ressaltar que essas decisões de pré-processamento não afetam apenas o desempenho do modelo, mas também influenciam diretamente a interpretação dos valores SHAP, uma vez que alteram a forma como as variáveis são representadas e utilizadas na predição.

2.6.2. Implementação em Python com `shap` e `xgboost`

Em Python, a implementação do SHAP é facilitada pela biblioteca `shap`, que oferece suporte nativo ao algoritmo TreeSHAP para modelos baseados em árvores, como o XGBoost. Isso permite calcular explicações locais e globais de forma eficiente, sem necessidade de implementar manualmente os valores de Shapley [Lundberg and Lee 2017]. Nesta seção, apresentamos um exemplo completo, desde a preparação dos dados até a geração de gráficos interpretativos.

2.6.2.1. Preparação dos Dados, Limpeza e Treinamento do Modelo

Antes de calcular os valores SHAP, é necessário preparar a base, realizar uma limpeza mínima dos dados e treinar o modelo preditivo. No Google Colab, a instalação das bibliotecas pode ser feita com o comando abaixo:

Instalação (executar uma vez no Colab):

```
!pip install shap xgboost scikit-learn pandas numpy
matplotlib -quiet
```

O código a seguir carrega a base de dados, remove valores clinicamente implausíveis, cria variáveis derivadas e ajusta um modelo XGBoost para classificação binária.

```
import numpy as np
import pandas as pd

from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score, brier_score_loss

from xgboost import XGBClassifier
import shap

df = pd.read_csv("cardio_train.csv")

# Limpeza: pressão arterial
df = df[
(df["ap_hi"].between(70, 250)) &
(df["ap_lo"].between(40, 150)) &
(df["ap_hi"] > df["ap_lo"])
].copy()
```



```
# Altura e peso
df = df[
(df["height"].between(120, 220)) &
(df["weight"].between(40, 200))
].copy()

df["age_years"] = df["age"] / 365.25
df["imc"] = df["weight"] / ((df["height"] / 100) ** 2)
df = df[df["imc"].between(15, 60)].copy()
print("N após limpeza:", df.shape[0])

features = [
"age_years", "gender", "imc",
"ap_hi", "ap_lo",
"cholesterol", "gluc",
"smoke", "alco", "active"
]
target = "cardio"
X = df[features]
y = df[target].astype(int)

# Particionamento estratificado (25% teste)
X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size=0.25, random_state=42, stratify=y
)

model = XGBClassifier(
n_estimators=300,
max_depth=4,
learning_rate=0.05,
subsample=0.8,
colsample_bytree=0.8,
reg_lambda=1.0,
reg_alpha=0.1,
random_state=42,
eval_metric="logloss",
use_label_encoder=False,
n_jobs=-1
)

model.fit(X_train, y_train, eval_set=[(X_test, y_test)],
 verbose=False)

# Avaliação
y_prob = model.predict_proba(X_test)[:, 1]
```



```
auc = roc_auc_score(y_test, y_prob)
brier = brier_score_loss(y_test, y_prob)

print(f"AUC-ROC: {auc:.3f}")
print(f"Brier Score: {brier:.4f}")
```

Esse bloco executa cinco tarefas principais. Primeiro, importa as bibliotecas necessárias para manipulação de dados, treinamento do modelo, avaliação e cálculo de valores SHAP. Em seguida, lê o arquivo `cardio_train.csv`. Depois, aplica uma limpeza básica para excluir valores incompatíveis com a prática clínica, como pressões arteriais inconsistentes, alturas ou pesos extremos, e valores de IMC fora de uma faixa plausível.

Na sequência, duas variáveis derivadas são criadas: idade em anos (`age_years`) e índice de massa corporal (`imc`). Essas transformações tornam os dados mais adequados tanto para o treinamento quanto para a interpretação clínica posterior. Depois disso, define-se o conjunto de variáveis preditoras e a variável resposta.

O particionamento entre treino e teste é feito de forma estratificada, preservando a proporção de casos com e sem doença cardiovascular em ambos os subconjuntos. Em seguida, ajusta-se um modelo `XGBClassifier` com hiperparâmetros moderados, suficientes para fins ilustrativos. Por fim, o modelo é avaliado por duas métricas: a AUC-ROC, que mede sua capacidade discriminativa, e o *Brier Score*, que avalia a calibração das probabilidades previstas.

2.6.2.2. Cálculo dos Valores SHAP com `TreeExplainer`

Depois do treinamento, o próximo passo é calcular os valores SHAP. Como o modelo utilizado é baseado em árvores, a forma mais eficiente de fazer isso é por meio do `TreeExplainer`, que implementa o TreeSHAP.

```
# TreeSHAP (log-odds)
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test)
base_value = explainer.expected_value
print(f"Valor base (phi_0): {base_value:.4f}")

# Verificação da decomposição aditiva
i = 0
pred_logodds = base_value + shap_values[i].sum()
pred_prob_check = 1 / (1 + np.exp(-pred_logodds))

print(f"Predição reconstruída: {pred_prob_check:.4f}")
print(f"Predição do modelo: {y_prob[i]:.4f}")
```

Nesse trecho, o objeto `explainer` aprende como decompor as predições do modelo em contribuições individuais das variáveis. O resultado, armazenado em

`shap_values`, contém um valor SHAP para cada variável e para cada paciente do conjunto de teste. Já `base_value` corresponde ao valor base da predição, isto é, à saída média do modelo antes de considerar as características específicas de cada indivíduo.

O código também mostra uma verificação importante: a soma dos valores SHAP de um paciente, adicionada ao valor base, deve reconstruir a predição do modelo na escala de *log-odds*. Em seguida, essa quantidade é transformada em probabilidade pela função logística. Essa etapa ajuda o leitor a visualizar, na prática, a propriedade de aditividade local do SHAP.

2.6.2.3. Visualizações para Interpretação

Uma das principais vantagens do SHAP é a possibilidade de gerar visualizações que ajudam a interpretar o modelo tanto em nível global quanto local. O código abaixo apresenta alguns dos gráficos mais utilizados.

```
import matplotlib.pyplot as plt

# Summary plots (beeswarm e bar)
shap.summary_plot(shap_values, X_test, plot_type="dot")
shap.summary_plot(shap_values, X_test, plot_type="bar")

# Dependence plot: ap_hi x colesterol
shap.dependence_plot(
"ap_hi",
shap_values,
X_test,
interaction_index="cholesterol"
)

# Waterfall plot (explicação local)
idx_paciente = 0
# Garantir escalar para base_value
if isinstance(base_value, (list, tuple, np.ndarray)):
bv = float(np.array(base_value).reshape(-1)[0])
else:
bv = float(base_value)

exp = shap.Explanation(
values=np.array(shap_values)[idx_paciente],
base_values=bv,
data=X_test.iloc[idx_paciente].to_numpy(),
feature_names=features
)

shap.plots.waterfall(exp)
```



```
# Force plot
shap.force_plot(bv, shap_values[idx_paciente],
X_test.iloc[idx_paciente])
```

Os primeiros comandos geram gráficos de resumo, que permitem uma visão global do comportamento do modelo. O *beeswarm plot* mostra, para cada variável, a distribuição dos valores SHAP em toda a amostra, permitindo observar simultaneamente a importância, a direção do efeito (aumento ou redução do risco) e a variabilidade entre pacientes. Já o gráfico de barras apresenta a importância média global das variáveis, calculada a partir da magnitude média de suas contribuições.

Em seguida, o *dependence plot* explora como a contribuição de uma variável específica varia ao longo de seus valores. No exemplo apresentado, analisa-se a pressão arterial sistólica (`ap_hi`) e sua possível interação com os níveis de colesterol. Esse tipo de visualização é útil para identificar relações não lineares e efeitos condicionais entre variáveis.

Por fim, os gráficos *waterfall* e *force plot* apresentam explicações locais, isto é, detalham a predição de um paciente específico. Nesses gráficos, a predição individual é construída a partir do valor base e das contribuições positivas e negativas de cada variável, permitindo compreender de forma direta quais fatores levaram o modelo a classificar aquele paciente como de maior ou menor risco.

Cabe destacar que cada um desses gráficos será discutido em maior detalhe na seção seguinte, com foco em sua interpretação clínica e estatística. Para fins de padronização visual, as figuras apresentadas no texto serão geradas em R; no entanto, todas essas visualizações podem ser produzidas de forma equivalente em Python, como ilustrado neste exemplo.

De modo geral, esse fluxo em Python demonstra como treinar um modelo compatível com o TreeSHAP, calcular explicações locais e globais e gerar visualizações interpretáveis com poucas linhas de código. O objetivo não é otimizar exaustivamente o modelo, mas mostrar, de forma prática, como o SHAP pode ser incorporado à análise de um problema clínico real.

2.6.3. Implementação em R com `xgboost` e `shapviz`

No ecossistema R, a combinação entre `xgboost` para modelagem e `shapviz` para interpretação permite reproduzir, de forma equivalente, o fluxo apresentado em Python. A principal diferença está na sintaxe e nas convenções da linguagem, mas as etapas centrais permanecem as mesmas: preparação dos dados, ajuste do modelo, cálculo dos valores SHAP e geração de gráficos interpretativos.

2.6.3.1. Preparação dos Dados, Limpeza e Treinamento do Modelo

No Google Colab com kernel R, as bibliotecas necessárias podem ser instaladas com o comando abaixo:

Instalação (executar uma vez no Colab com kernel R):

```
install.packages(c("data.table", "xgboost", "shapviz",
"pROC"))
```

O código a seguir realiza a leitura da base, aplica critérios básicos de limpeza, constrói variáveis derivadas, separa os dados em treino e teste e ajusta um modelo XGBoost para classificação binária.

```
library(data.table)
library(xgboost)
library(shapviz)
library(pROC)

df <- fread("cardio_train.csv")

# Limpeza: pressão arterial
df <- df[
ap_hi >= 70 & ap_hi <= 250 &
ap_lo >= 40 & ap_lo <= 150 &
ap_hi > ap_lo
]

# Altura e peso
df <- df[
height >= 120 & height <= 220 &
weight >= 40  & weight <= 200
]

df[, age_years := age / 365.25]
df[, imc := weight / (height / 100)^2]
df <- df[imc >= 15 & imc <= 60]
cat("N após limpeza:", nrow(df), "\n")

features <- c(
"age_years", "gender", "imc",
"ap_hi", "ap_lo",
"cholesterol", "gluc",
"smoke", "alco", "active"
)
target <- "cardio"
stopifnot(all(c(features, target) %in% names(df)))
X <- as.matrix(df[, ..features])
y <- as.numeric(df[[target]])

# Particionamento estratificado (25% teste)
```



```
set.seed(42)
idx1 <- which(y == 1)
idx0 <- which(y == 0)

idx_test <- c(
sample(idx1, size = floor(0.25 * length(idx1))),
sample(idx0, size = floor(0.25 * length(idx0)))
)

X_train <- X[-idx_test, , drop = FALSE]
y_train <- y[-idx_test]
X_test  <- X[idx_test,  , drop = FALSE]
y_test  <- y[idx_test]

dtrain <- xgb.DMatrix(data = X_train, label = y_train)
dtest  <- xgb.DMatrix(data = X_test,  label = y_test)

params <- list(
objective = "binary:logistic",
eval_metric = "logloss",
max_depth = 4,
eta = 0.05,
subsample = 0.8,
colsample_bytree = 0.8,
lambda = 1.0,
alpha = 0.1
)

model <- xgb.train(
params = params,
data = dtrain,
nrounds = 300,
watchlist = list(test = dtest),
verbose = 0
)

# Avaliação
pred_prob <- predict(model, dtest)
auc <- pROC::auc(y_test, pred_prob)
cat("AUC-ROC:", round(as.numeric(auc), 3), "\n")
```

Esse bloco segue a mesma lógica da implementação em Python. Primeiro, carrega as bibliotecas responsáveis pela leitura dos dados, treinamento do modelo, cálculo das explicações e avaliação de desempenho. Em seguida, lê o arquivo `cardio_train.csv` e aplica uma limpeza básica para remover valores incompatíveis com a prática clínica, como pressões arteriais inconsistentes, alturas e pesos fora de faixas plausíveis e valores

extremos de IMC.

Depois disso, são construídas duas variáveis derivadas: idade em anos (`age_years`) e índice de massa corporal (`imc`). Essas transformações tornam os dados mais adequados à interpretação clínica. Em seguida, selecionam-se as variáveis preditoras e a variável resposta, e os dados são convertidos para o formato matricial exigido pelo `xgboost`.

O particionamento em treino e teste é feito de forma estratificada, preservando a proporção entre indivíduos com e sem doença cardiovascular. Após isso, os objetos `xgb.DMatrix` são criados para armazenar os dados em um formato otimizado para o treinamento. O modelo é então ajustado com parâmetros moderados, suficientes para fins ilustrativos, e avaliado por meio da AUC-ROC, que mede sua capacidade de discriminar entre pacientes com e sem o desfecho.

2.6.3.2. Cálculo de Valores SHAP

Depois de treinado o modelo, o próximo passo é obter os valores SHAP. No `xgboost` em R, isso pode ser feito diretamente com o argumento `predcontrib = TRUE`, que retorna a contribuição de cada variável para cada predição individual.

```
# SHAP (log-odds; última coluna = BIAS)
shap_contrib <- predict(model, dtest, predcontrib = TRUE)
shap_matrix <- shap_contrib[, 1:(ncol(shap_contrib) - 1),
 drop = FALSE]
base_value  <- shap_contrib[1, ncol(shap_contrib)]
colnames(shap_matrix) <- features
sv <- shapviz(model, X_pred = dtest, X = X_test)
```

Nesse trecho, `shap_contrib` armazena uma matriz em que cada linha corresponde a um paciente e cada coluna representa a contribuição de uma variável. A última coluna contém o valor base da predição, também chamado de *BIAS*, que corresponde ao ponto de partida a partir do qual as contribuições individuais são somadas. A matriz `shap_matrix` contém apenas os valores SHAP das variáveis, já com os nomes das colunas ajustados para facilitar a leitura.

Em seguida, a função `shapviz()` organiza essas informações em um objeto próprio para visualização. Esse objeto reúne o modelo, os dados utilizados nas predições e os valores SHAP, permitindo gerar gráficos de forma direta e consistente.

2.6.3.3. Visualizações com `shapviz`

Uma vez calculados os valores SHAP, o pacote `shapviz` oferece funções específicas para gerar visualizações globais e locais de forma simples.

```
# Bar plot
```



```
sv_importance(sv, kind = "bar")

# Summary plot
sv_importance(sv, kind = "beeswarm")

# Dependence
sv_dependence(sv, v = "ap_hi", color_var = "cholesterol")

# Waterfall local
sv_waterfall(sv, row_id = 1)

# Force plot
sv_force(sv, row_id = 1)
```

Os dois primeiros comandos geram gráficos de resumo do modelo. O gráfico de barras apresenta a importância global média das variáveis, enquanto o *beeswarm plot* mostra a distribuição completa das contribuições em toda a amostra, permitindo observar importância, direção do efeito e heterogeneidade entre pacientes.

O *dependence plot* examina como a contribuição de uma variável específica varia de acordo com seus valores. No exemplo acima, analisa-se a pressão arterial sistólica (`ap_hi`) e sua possível interação com o colesterol. Esse tipo de gráfico é útil para investigar relações não lineares e efeitos condicionais.

Já os gráficos *waterfall* e *force plot* apresentam explicações locais, isto é, mostram como a predição de um paciente específico é construída a partir do valor base e das contribuições positivas e negativas de cada variável. Essas visualizações ajudam a entender, de forma direta, por que o modelo classificou aquele indivíduo como de maior ou menor risco.

Cabe destacar que cada um desses gráficos será discutido em maior detalhe na seção seguinte, com foco em sua interpretação clínica e estatística. Para fins de padronização visual, as figuras apresentadas no texto serão geradas em R. Ainda assim, como mostrado anteriormente, todas essas visualizações também podem ser produzidas de forma equivalente em Python.

De modo geral, esse fluxo em R demonstra como treinar um modelo compatível com o TreeSHAP, calcular explicações locais e globais e gerar visualizações interpretáveis de forma relativamente simples. Assim como na implementação em Python, o objetivo não é ensinar exaustivamente a modelagem preditiva, mas mostrar, de maneira prática, como o SHAP pode ser aplicado à interpretação de um banco de dados clínico.

2.6.4. Considerações de Desempenho para Grandes Bases Clínicas

Em aplicações com grandes volumes de dados, como bases de prontuários eletrônicos com milhões de registros, o cálculo de valores SHAP para todos os pacientes pode se tornar computacionalmente custoso. Para viabilizar o uso do método em larga escala, é comum combinar estratégias de amostragem, otimização algorítmica e técnicas de engenharia.

Uma abordagem prática é a **amostragem estratificada**. Para análises globais, não

é necessário utilizar toda a base de dados: amostras representativas, tipicamente entre 1.000 e 10.000 instâncias, costumam ser suficientes para estimar importâncias médias com boa precisão. Estudos mostram que, quando a amostra preserva a estrutura da população, especialmente a distribuição do desfecho e de subgrupos clínicos relevantes, as estimativas permanecem estáveis [Bachmann 2025].

Outra estratégia importante é o **cálculo sob demanda**. Em sistemas de produção, nem sempre é necessário gerar explicações para todos os pacientes. Em vez disso, pode-se calcular valores SHAP apenas para casos de interesse, como pacientes que acionam alertas clínicos ou situações em que o usuário solicita uma explicação. Essa abordagem reduz significativamente o custo computacional e está alinhada com o uso prático do modelo no contexto clínico [Bachmann 2025, Hu et al. 2024a].

Do ponto de vista computacional, a **paralelização** é uma das principais formas de ganho de desempenho. O algoritmo TreeSHAP permite calcular explicações de forma independente para cada instância, o que facilita sua execução em múltiplos núcleos de CPU ou em GPUs. Em Python, isso pode ser explorado por meio de parâmetros como `n_jobs` ou bibliotecas como `joblib`; em R, pacotes como `parallel`, `future` e implementações otimizadas oferecem funcionalidades semelhantes para acelerar a computação [Greenwell and Boehmke 2020].

Além disso, o uso de **cache de resultados** é particularmente relevante em ambientes clínicos e regulatórios. Armazenar os valores SHAP já calculados permite reutilizar explicações em auditorias futuras, garantir rastreabilidade das decisões e evitar recálculos desnecessários. Essa prática é especialmente importante quando se deseja reconstruir, de forma retrospectiva, o comportamento do modelo em situações específicas para fins de conformidade [Kiseleva et al. 2022, U.S. Food and Drug Administration 2021].

Por fim, a literatura recente tem proposto métodos que reduzem ainda mais o custo computacional do SHAP. Entre eles estão algoritmos acelerados, como o Fast TreeSHAP, e abordagens aproximadas, como o FastSHAP, que utilizam estratégias de aprendizado para estimar contribuições de forma mais eficiente. Há também métodos baseados em decomposição ou aproximações mais refinadas, que buscam manter a qualidade das explicações com menor custo [Olsen and Jullum 2025, Gevaert and Saeys 2022].

Em conjunto, essas estratégias tornam o uso do SHAP viável mesmo em cenários clínicos de grande escala, permitindo equilibrar custo computacional e qualidade interpretativa.

2.7. Visualização e Análise de Resultados SHAP

A visualização dos valores SHAP tem como objetivo transformar resultados matemáticos em representações que possam ser interpretadas por diferentes públicos, como clínicos, pesquisadores e gestores. Enquanto métricas tradicionais, como AUC-ROC, indicam o desempenho geral do modelo, as visualizações SHAP ajudam a entender *como* e *por que* o modelo chega a uma determinada predição. Em outras palavras, elas tornam explícita a contribuição de cada variável, tanto no nível individual quanto no nível populacional.

Nesta seção, são apresentadas as principais estratégias de visualização associadas ao SHAP, utilizando como referência o conjunto de dados de risco cardiovascular discutido

ao longo do capítulo. O foco está na interpretação dos gráficos e em sua aplicação prática no contexto clínico.

2.7.1. Princípios de Visualização para Aplicações Clínicas

O uso de visualizações em saúde exige cuidado especial, pois os resultados precisam ser compreendidos rapidamente e, ao mesmo tempo, manter rigor técnico. Quatro princípios orientam o uso de gráficos SHAP nesse contexto.

O primeiro é a **clareza e acessibilidade**. As informações devem ser apresentadas em unidades que façam sentido para o clínico (por exemplo, mmHg ou probabilidade de risco), com uso consistente de elementos visuais como cores e posições. Além disso, é importante permitir diferentes níveis de detalhamento, de modo que usuários com mais ou menos experiência possam interpretar os resultados de forma adequada [Patel et al. 2024].

O segundo princípio é a **fidelidade ao modelo**. As visualizações devem refletir de forma fiel o comportamento do modelo, sem simplificações que distorçam sua interpretação. Isso inclui, sempre que possível, indicar incertezas, limitações dos dados e possíveis fontes de erro, aspectos fundamentais para a confiança no uso clínico dessas ferramentas [Harrigan et al. 2022, Markus et al. 2021].

O terceiro princípio é a **relevância clínica**. Os padrões observados nos gráficos precisam ser interpretados à luz do conhecimento médico existente e conectados a tarefas reais, como diagnóstico, triagem ou acompanhamento de pacientes. Sem essa contextualização, a visualização pode ser tecnicamente correta, mas pouco útil na prática [Patel et al. 2024, Tonekaboni et al. 2019].

Por fim, destaca-se a **acionabilidade**. Sempre que possível, as visualizações devem apontar fatores que podem ser modificados, indicar limiares clínicos importantes e destacar situações de risco que demandem intervenção. O objetivo não é apenas explicar o modelo, mas apoiar decisões concretas no cuidado ao paciente [Harrigan et al. 2022, Tonekaboni et al. 2019].

2.7.2. Summary Plot: Visão Global do Comportamento do Modelo

O *summary plot*, também conhecido como *beeswarm plot*, é uma das visualizações mais completas para análise global com SHAP. Em um único gráfico, ele combina três informações importantes: a importância das variáveis no modelo, a distribuição de seus efeitos e a relação entre o valor da variável e sua contribuição para a predição [Lundberg et al. 2019, Lundberg and Lee 2017, Ponce-Bobadilla et al. 2024].

Nesse gráfico, as variáveis são organizadas no eixo vertical de acordo com sua importância global, medida pela média do valor absoluto das contribuições ($\mathbb{E}[|\phi_i|]$). Cada ponto representa um paciente da amostra. A posição horizontal do ponto indica o valor SHAP, ou seja, o quanto aquela variável aumentou ou reduziu a predição para aquele indivíduo. A cor do ponto representa o valor original da variável (por exemplo, valores mais altos ou mais baixos), o que permite visualizar, ao mesmo tempo, a direção do efeito, possíveis relações monotônicas, padrões não lineares e diferenças entre pacientes [Lundberg et al. 2020, Liu et al. 2022, Ponce-Bobadilla et al. 2024].

Na Figura 2.1, observa-se que a pressão arterial sistólica (`ap_hi`) e a idade

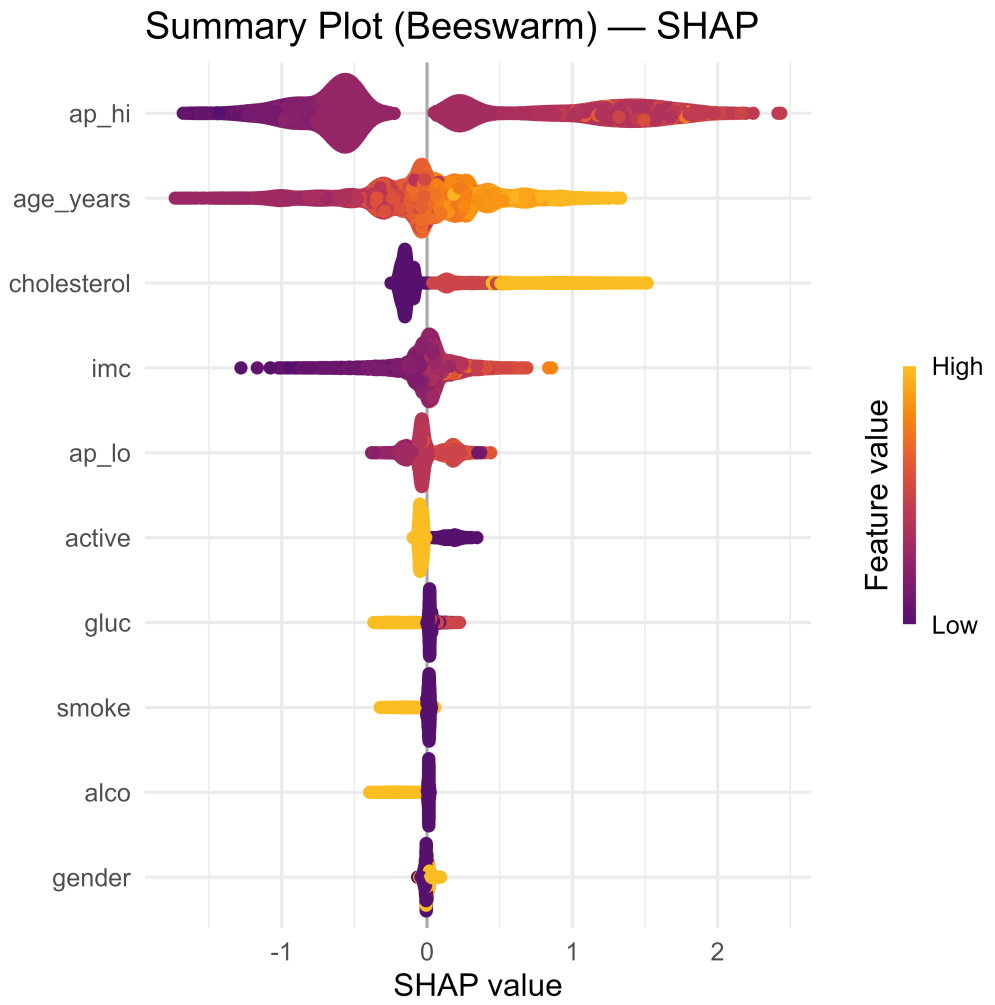

Figura 2.1. Summary plot (beeswarm) dos valores SHAP para o modelo XGBoost treinado no conjunto de dados cardiovascular. Variáveis ordenadas por importância global; coloração indica o valor original do preditor.

(`age_years`) são as variáveis com maior influência no modelo. Valores mais altos dessas variáveis (indicados por cores mais claras) estão consistentemente associados a maior risco predito, ao passo que valores baixos contribuem negativamente para a predição. O colesterol, representado como variável ordinal, apresenta um padrão monotônico claro, com valores elevados associados a SHAP positivos. O IMC (`imc`) exibe comportamento assimétrico: valores baixos concentram-se em SHAP negativos, enquanto valores altos produzem contribuições ligeiramente positivas.

A pressão diastólica (`ap_lo`) também contribui para o risco, porém com menor magnitude e dispersão. A variável atividade física (`active`) apresenta SHAP predominantemente próximos de zero, com leve tendência protetora para indivíduos ativos. Por outro lado, variáveis como glicemia (`gluc`), tabagismo (`smoke`), consumo de álcool (`alco`) e sexo (`gender`) têm impacto relativamente menor quando consideradas em conjunto com as demais variáveis do modelo.

Uma forma mais simples de visualizar a importância das variáveis é o *bar plot*, que representa cada variável pela média do valor absoluto de suas contribuições ($\mathbb{E}[|\phi_i|]$) [Lundberg et al. 2020, Ponce-Bobadilla et al. 2024]. Esse gráfico oferece uma leitura direta do peso relativo de cada preditor no modelo.

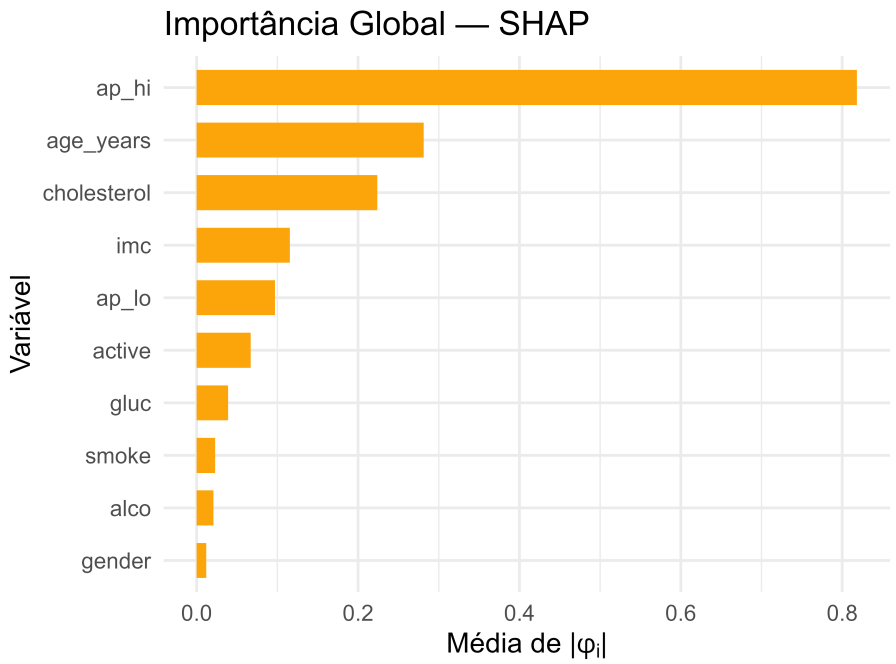

Figura 2.2. Bar plot de importância global baseado em $\mathbb{E}[|\phi_i|]$. As barras indicam a contribuição relativa de cada variável no modelo XGBoost treinado com os dados cardiovasculares.

Na Figura 2.2, observa-se novamente a predominância da pressão arterial sistólica, seguida pela idade e pelo colesterol. Variáveis comportamentais e sexo aparecem com menor contribuição relativa. Esse tipo de gráfico é útil para uma visão rápida e resumida do modelo, especialmente em contextos de comunicação. No entanto, ele não mostra a direção dos efeitos nem a variabilidade entre pacientes — informações que estão presentes no *beeswarm plot* [Lundberg et al. 2020, Ponce-Bobadilla et al. 2024].

2.7.3. Dependence Plot: Relações Não Lineares e Limiares Clínicos

O *dependence plot* é uma visualização que permite analisar como o efeito de uma variável muda ao longo de seus valores. Nesse gráfico, o eixo x representa o valor observado da variável, enquanto o eixo y mostra sua contribuição SHAP para a predição. Cada ponto corresponde a um paciente da base, o que permite observar não apenas tendências médias, mas também a variabilidade entre indivíduos [Lundberg et al. 2020, Ponce-Bobadilla et al. 2024].

Diferentemente de métodos tradicionais, como gráficos de dependência parcial, o *dependence plot* preserva a heterogeneidade individual. Ou seja, ele não mostra apenas o comportamento médio do modelo, mas revela como o efeito da variável varia entre pacientes com perfis diferentes [Lundberg et al. 2020, Cha et al. 2021].

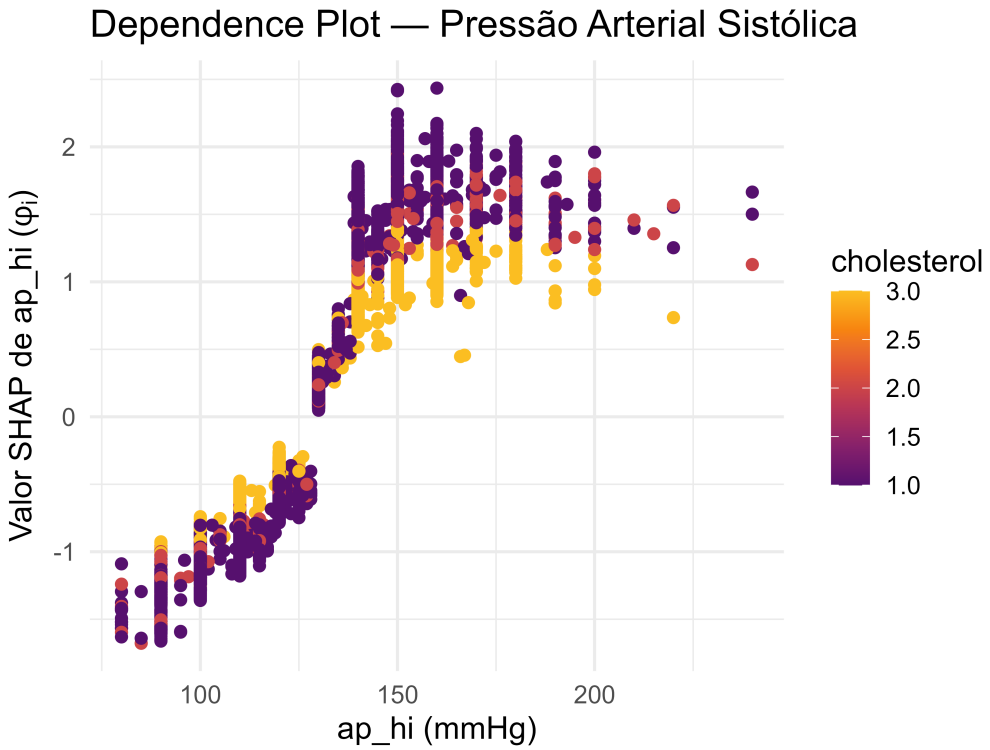

Figura 2.3. Dependence plot dos valores SHAP para `ap_hi`, com coloração pela variável ordinal de colesterol. Cada ponto representa uma observação individual.

Na Figura 2.3, observa-se uma relação não linear entre a pressão arterial sistólica (`ap_hi`) e o risco predito pelo modelo. Para valores abaixo de aproximadamente 130–135 mmHg, as contribuições SHAP são predominantemente negativas, indicando que pressões mais baixas reduzem o risco predito. A partir de cerca de 135–140 mmHg, há uma transição abrupta para contribuições positivas e crescentes. Esse ponto de inflexão é consistente com os limiares utilizados nas diretrizes clínicas para definição de hipertensão. Acima desse intervalo, o efeito permanece positivo e crescente, com dispersão vertical considerável dos pontos, refletindo a variação interindividual no impacto dessa variável.

A cor dos pontos, representando o nível de colesterol (`cholesterol`), permite identificar uma possível interação entre as duas variáveis. Na faixa de pressão baixa a

moderada (aproximadamente 80–130 mmHg), pacientes com colesterol mais elevado tendem a apresentar contribuições SHAP ligeiramente maiores do que aqueles com colesterol baixo para um mesmo valor de pressão. De modo geral, a dispersão vertical dos pontos indica que o efeito da pressão arterial sistólica não é isolado, mas modulado pelo conjunto das demais características clínicas do paciente.

2.7.4. Visualizações Locais: Explicando Decisões Individuais

Para compreender como o modelo toma decisões em nível individual, duas visualizações são amplamente utilizadas: o *waterfall plot* e o *force plot*. Ambas mostram como a predição de um paciente é construída a partir das contribuições das variáveis, mas o fazem de maneiras complementares [Bifarin 2022, Lin et al. 2025].

Waterfall plot. O *waterfall plot* apresenta a decomposição da predição de forma sequencial. O gráfico começa no valor base ϕ_0, que corresponde à predição média do modelo sobre o conjunto de referência utilizado pelo SHAP, e adiciona (ou subtrai) as contribuições SHAP de cada variável até alcançar a predição final [Ponce-Bobadilla et al. 2024, Lin et al. 2025, Bifarin 2022]. Contribuições positivas, que aumentam o risco, são representadas à direita; contribuições negativas, que reduzem o risco, aparecem à esquerda. As contribuições são expressas na escala de *log-odds*, e não em probabilidade. A conversão para probabilidade deve ser realizada apenas após a soma de todas as contribuições, por meio da função logística [Lundberg and Lee 2017, Ponce-Bobadilla et al. 2024].

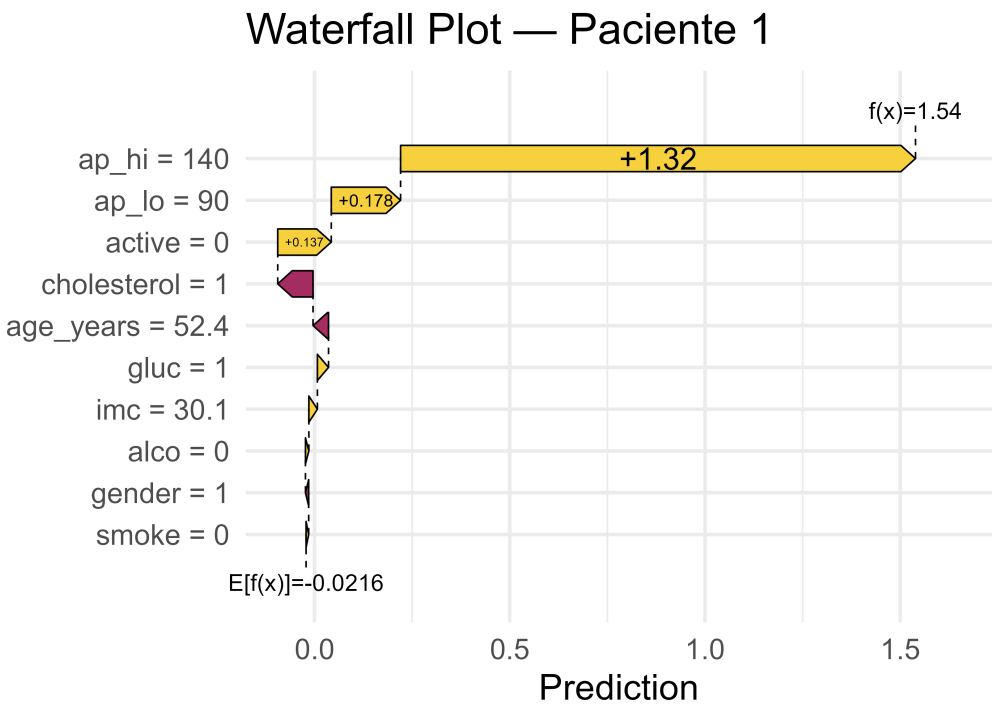

Figura 2.4. Waterfall plot dos valores SHAP para o paciente exemplo, ilustrando a decomposição da predição a partir de ϕ_0 até a predição final $f(x) \approx 1{,}54$.

Na Figura 2.4, observa-se que a pressão arterial sistólica (`ap_hi` = 140 mmHg)

é, de longe, o principal fator que eleva o risco predito ($\varphi = +1{,}32$), seguida pela pressão diastólica (`ap_lo` = 90 mmHg; $\varphi = +0{,}18$) e pelo sedentarismo (`active` = 0; $\varphi = +0{,}14$). Por outro lado, o colesterol normal (`cholesterol` = 1) e a idade de 52,4 anos contribuem modestamente para reduzir o risco predito, enquanto as demais variáveis — glicemia, IMC, álcool, sexo e tabagismo — apresentam contribuições próximas de zero. A predição final é $f(x) = 1{,}54$ em escala de *log-odds*, partindo de um valor base de $E[f(x)] = -0{,}02$.

Force plot. O *force plot* apresenta as mesmas informações, mas em um formato horizontal mais compacto. Nesse gráfico, as contribuições positivas (fatores de risco) são representadas por tons quentes, enquanto as contribuições negativas (fatores protetores) aparecem em tons frios. Diferentemente do *waterfall plot*, que enfatiza a sequência de acumulação, o *force plot* destaca o balanço entre forças que aumentam e reduzem o risco [Stenwig et al. 2022, Hu et al. 2024a].

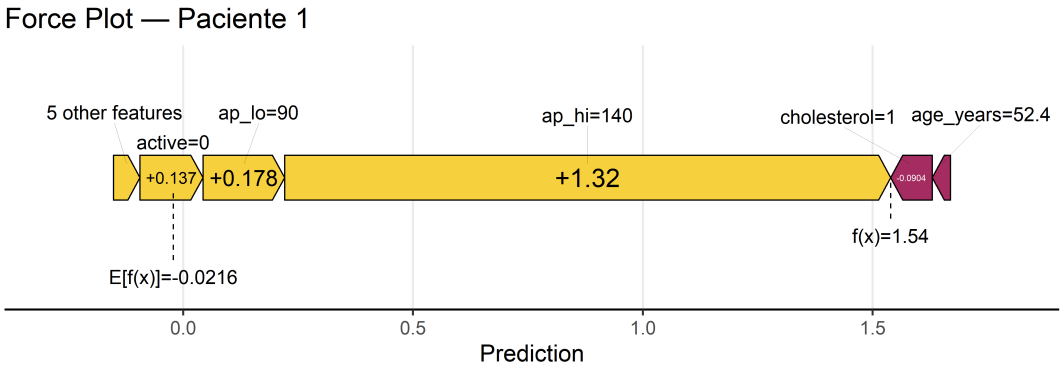

Figura 2.5. Force plot dos valores SHAP para o paciente exemplo. Contribuições positivas (tons quentes) e negativas (tons frios) são apresentadas simultaneamente, ilustrando o deslocamento da predição de ϕ_0 até $f(x)$.

Esse formato é particularmente útil em interfaces clínicas, onde é necessário visualizar rapidamente os fatores mais relevantes para a decisão. No entanto, quando o número de variáveis é elevado, o gráfico pode se tornar mais difícil de interpretar, pois várias contribuições são exibidas simultaneamente [Ponce-Bobadilla et al. 2024, Contreras et al. 2024].

2.7.5. Armadilhas e Limitações na Interpretação Visual

Apesar de sua utilidade, as visualizações SHAP devem ser interpretadas com cautela. Elas facilitam a compreensão do comportamento do modelo, mas também podem induzir conclusões equivocadas se analisadas de forma acrítica.

Correlação não implica causalidade. Uma contribuição elevada de uma variável não significa que ela seja causa direta do desfecho. Por exemplo, o colesterol pode aparecer como um fator importante não apenas por seu papel fisiológico, mas também por refletir aspectos indiretos, como hábitos alimentares ou acesso ao sistema de saúde. Nesse sentido, intervenções focadas apenas na variável observada podem não produzir o efeito esperado

se os determinantes subjacentes não forem considerados [Ponce-Bobadilla et al. 2024, Salih et al. 2023, Nohara et al. 2021, Bifarin 2022, Hettikankanamage et al. 2025].

Instabilidade com variáveis correlacionadas. Quando há forte correlação entre variáveis, como entre creatinina e taxa de filtração glomerular, as contribuições SHAP podem ser distribuídas de forma instável entre elas. Essa instabilidade pode variar entre execuções, especialmente em métodos que assumem independência entre variáveis ou utilizam estratégias de amostragem [Ponce-Bobadilla et al. 2024, Salih et al. 2023, Hu et al. 2024a].

Dependência da amostra de referência. Os valores SHAP dependem da distribuição dos dados utilizada como referência. O valor base ϕ_0 e as contribuições individuais são calculados com base nessa distribuição. Assim, se a amostra não for representativa da população de interesse, as conclusões sobre a importância das variáveis podem ser distorcidas [Liu et al. 2022, Ponce-Bobadilla et al. 2024].

Extrapolação visual arriscada. Os padrões observados em gráficos, como os *dependence plots*, são válidos apenas dentro da faixa de valores efetivamente presente nos dados. Inferências feitas fora dessas regiões, especialmente em áreas com pouca ou nenhuma observação, podem ser enganosas, sobretudo quando o modelo apresenta comportamento não linear [Hooker et al. 2019, Ponce-Bobadilla et al. 2024].

Ilusão de compreensão completa. Visualizações bem construídas podem transmitir a impressão de que o modelo foi totalmente compreendido. No entanto, os valores SHAP descrevem apenas como o modelo se comporta diante dos dados disponíveis. Eles não revelam, por si só, os mecanismos causais do fenômeno clínico. Trabalhos recentes destacam que interpretações sem validação externa podem ser equivocadas quando tratadas como evidência objetiva [Huang and ao Marques-Silva 2024, Hooshyar and Yang 2024, Singh et al. 2025].

Em conjunto, essas limitações mostram que as visualizações SHAP devem ser entendidas como ferramentas de apoio à interpretação, e não como evidência definitiva. Sua utilização deve ser sempre complementada por conhecimento clínico, análise crítica e validação empírica. A seção seguinte aprofunda essas questões ao discutir limitações metodológicas e boas práticas para o uso responsável do SHAP em ambientes clínicos.

2.8. Limitações, Desafios e Boas Práticas

O uso do SHAP em aplicações clínicas traz benefícios importantes em termos de interpretabilidade, mas também envolve limitações conceituais, desafios técnicos e riscos de interpretação inadequada. Esses aspectos precisam ser reconhecidos e gerenciados de forma ativa. Esta seção discute essas questões e apresenta boas práticas para o uso responsável do método em saúde.

2.8.1. Limitações Conceituais Fundamentais

2.8.1.1. SHAP Explica o Modelo, Não o Fenômeno Clínico

A limitação mais importante do SHAP é que ele explica o comportamento do modelo, e não o fenômeno clínico em si. Os valores SHAP refletem as relações que o modelo aprendeu a partir dos dados, incluindo sua estrutura, suas hipóteses implícitas e a distribuição da amostra. No entanto, eles não estabelecem relações causais nem explicam mecanismos fisiopatológicos [Lundberg and Lee 2017, Ponce-Bobadilla et al. 2024, Salih et al. 2023, Bifarin 2022].

Na prática, isso significa que uma variável pode aparecer como altamente relevante no modelo sem ser, de fato, a causa do desfecho. Por exemplo, o número de exames realizados pode estar associado a maior mortalidade hospitalar não porque cause o desfecho, mas porque pacientes mais graves tendem a receber mais exames. Interpretar esse resultado de forma causal levaria a decisões equivocadas ditadas por vieses de presença no sistema de saúde [Momenzadeh et al. 2022].

As explicações SHAP devem ser interpretadas como indicações de como o modelo utiliza as variáveis, e não como evidência causal. Para responder a perguntas causais, são necessários métodos específicos, como modelos estruturais, variáveis instrumentais ou ensaios clínicos [Chen et al. 2021]. Nesse sentido, o SHAP é mais adequado como ferramenta de geração de hipóteses, que deve ser integrada ao conhecimento clínico e a outras fontes de evidência empírica [Hettikankanamage et al. 2025, Viswan et al. 2023].

2.8.1.2. Dependência da Qualidade e Representatividade dos Dados

A qualidade das explicações depende diretamente da qualidade dos dados. Se o modelo é treinado com dados enviesados ou incompletos, as explicações refletirão essas limitações.

Vieses de seleção podem comprometer a generalização do modelo. Erros de mensuração podem distorcer as associações aprendidas. Vieses históricos, como subtratamento de determinados grupos, podem ser reproduzidos e até evidenciados pelas explicações SHAP [Obermeyer et al. 2019, Chen et al. 2021]. Além disso, a presença de confundidores não controlados pode levar o modelo a aprender associações que não representam relações reais [Ponce-Bobadilla et al. 2024, Chen et al. 2021].

Outro ponto importante é a escolha do conjunto de referência (*background*) utilizado para calcular os valores SHAP. Amostras pequenas, desbalanceadas ou não representativas podem introduzir variabilidade excessiva na estimativa de importância das variáveis [Liu et al. 2022] e até permitir manipulação das explicações por meio de seleção enviesada de dados (ataques adversariais), ocultando comportamentos discriminatórios do modelo [Wood et al. 2023, Xin et al. 2025].

2.8.2. Desafios Técnicos e Metodológicos

2.8.2.1. Multicolinearidade e Ambiguidade de Atribuição

Em dados clínicos, é comum que variáveis estejam correlacionadas. Nesses casos, o SHAP distribui a contribuição entre elas de acordo com os axiomas de Shapley. Embora essa decomposição seja matematicamente consistente, ela pode ser difícil de interpretar do ponto de vista clínico.

Por exemplo, variáveis como pressão arterial sistólica e diastólica ou diferentes marcadores renais podem compartilhar informações semelhantes. O resultado é que a importância pode ser diluída entre essas variáveis, dificultando a identificação de qual delas é mais relevante na prática [Salih et al. 2023, Hu et al. 2024b]. Em métodos aproximados, esse problema pode ser agravado pela geração de combinações de dados pouco realistas durante o cálculo das contribuições [Hu et al. 2024b].

Para lidar com esse problema, é recomendável analisar variáveis correlacionadas em conjunto, explorar valores de interação SHAP e documentar explicitamente as relações conhecidas entre variáveis [Contreras et al. 2024, Lundberg et al. 2020].

2.8.2.2. Complexidade Computacional em Escala

O cálculo exato dos valores de Shapley é computacionalmente complexo e, em muitos casos, inviável. Embora o TreeSHAP torne esse cálculo eficiente para modelos baseados em árvores [Lundberg et al. 2020], o custo ainda pode ser elevado em bases muito grandes.

Na prática, estratégias como amostragem estratificada, cálculo sob demanda, paralelização e armazenamento de resultados são utilizadas para tornar o método viável em larga escala. Abordagens adicionais, como o uso de agrupamentos (clustering) combinado com SHAP, também podem reduzir o custo mantendo boa fidelidade das explicações [Bachmann 2025]. Sempre que métodos aproximados forem utilizados, seu impacto deve ser avaliado e documentado.

2.8.2.3. Estabilidade e Reprodutibilidade

Os valores SHAP podem variar entre diferentes execuções do modelo. Mudanças na divisão dos dados, na inicialização do algoritmo ou nos hiperparâmetros podem alterar o ranking de importância das variáveis. Esse efeito é particularmente relevante em métodos baseados em amostragem, como o KernelSHAP [Kelodjou et al. 2024, Bachmann 2025].

Para garantir maior robustez, recomenda-se avaliar a estabilidade das explicações. Isso pode ser feito treinando múltiplos modelos, utilizando reamostragem dos dados e verificando se os padrões observados se mantêm consistentes. Em aplicações críticas, deve-se dar maior peso a variáveis cujas contribuições são estáveis em diferentes configurações.

2.8.3. Riscos de Uso Indevido em Contextos Clínicos

As visualizações SHAP, por serem intuitivas e visualmente claras, podem gerar uma falsa sensação de compreensão completa do modelo. Isso pode levar a excesso de confiança nas predições ou a interpretações equivocadas [Singh et al. 2025, Salih et al. 2023].

Em ambientes clínicos com alta carga cognitiva, como unidades de terapia intensiva, explicações muito detalhadas podem contribuir para fadiga de alertas. Nesses casos, é preferível apresentar informações de forma concisa, destacando apenas os fatores mais relevantes [Ponce-Bobadilla et al. 2024, Rasheed et al. 2021, Patel et al. 2024].

Além disso, explicações locais não garantem que a predição esteja correta. Elas apenas mostram como o modelo chegou àquele resultado. Portanto, não devem ser utilizadas como única base para decisões clínicas de alto risco [Singh et al. 2025]. A responsabilidade final permanece com o profissional de saúde [Kheder et al. 2025].

2.8.4. Boas Práticas para Aplicações em Saúde

O uso responsável do SHAP requer atenção ao longo de todo o ciclo de vida do modelo [Ponce-Bobadilla et al. 2024, Hu et al. 2022].

Antes da análise, é fundamental garantir que o modelo tenha desempenho adequado, incluindo boa calibração, validação externa e avaliação por subgrupos. Também é importante documentar todo o pipeline e definir claramente os objetivos da interpretação [Kheder et al. 2025, Chen et al. 2021].

Durante a análise, recomenda-se combinar explicações globais e locais, avaliar a plausibilidade clínica dos resultados, investigar discrepâncias com especialistas e considerar diferentes subgrupos populacionais [Hakkoum et al. 2024, Prendin et al. 2023]. A escala utilizada (probabilidade ou *log-odds*) deve ser explicitada [Ponce-Bobadilla et al. 2024].

Na comunicação dos resultados, a linguagem deve ser adaptada ao público, evitando interpretações causais indevidas [Arrieta et al. 2020, Bifarin 2022]. As limitações do método devem ser explicitadas, e sempre que possível as interpretações devem ser validadas com profissionais de saúde [Ghasemi et al. 2024].

Em produção, é importante monitorar continuamente o comportamento do modelo e de suas explicações, armazenar resultados para auditoria e revisar o sistema após atualizações ou mudanças na população atendida [Ponce-Bobadilla et al. 2024, Stogiannos et al. 2023, Singh et al. 2025].

Em síntese, o SHAP é uma ferramenta poderosa para interpretação de modelos, mas seu uso seguro exige análise crítica, integração com conhecimento clínico e atenção constante às suas limitações.

A seção seguinte discute as implicações éticas, regulatórias e de governança associadas ao uso de modelos explicáveis em saúde.

2.9. Considerações Éticas, Regulatórias e Governança de Modelos Explicáveis em Saúde

O uso de modelos explicáveis em saúde não envolve apenas questões técnicas, mas também implicações éticas, legais e organizacionais. A adoção responsável dessas ferramentas exige atenção a regulamentações, definição clara de responsabilidades, avaliação de equidade e implementação de mecanismos de governança ao longo de todo o ciclo de vida do modelo.

Leis e diretrizes recentes, como a LGPD no Brasil, o GDPR na União Europeia e o *EU AI Act*, estabelecem requisitos explícitos relacionados à transparência e à prestação de contas em decisões automatizadas [Goodman and Flaxman 2017, Rasheed et al. 2021, Khan et al. 2024, Aldhafeeri 2025]. Nesse contexto, métodos como o SHAP podem contribuir para tornar essas exigências operacionais, ao fornecer explicações sobre o funcionamento dos modelos.

No entanto, é importante reconhecer que a simples disponibilização de explicações técnicas não é suficiente. Essas explicações precisam ser traduzidas para formatos compreensíveis e relevantes para o contexto clínico. Além disso, a explicabilidade não substitui outras obrigações legais, como o consentimento informado, a proteção de dados pessoais e a minimização do uso de informações sensíveis [Amann et al. 2020].

Embora o SHAP contribua para tornar decisões mais transparentes e rastreáveis, ele não transfere a responsabilidade para o sistema automatizado. A decisão final continua sendo do profissional de saúde, que deve interpretar as informações fornecidas pelo modelo à luz do contexto clínico [Kiseleva et al. 2022, Kheder et al. 2025]. Assim, a *accountability* permanece inerente ao agente humano, mesmo quando há apoio de sistemas baseados em inteligência artificial.

As explicações SHAP podem ser utilizadas para identificar possíveis vieses no modelo, especialmente quando as contribuições são analisadas em diferentes subgrupos populacionais. No entanto, essa análise tem limitações. Variáveis que funcionam como *proxies* de atributos protegidos, como raça ou condição socioeconômica, podem ocultar mecanismos de discriminação estrutural [Chen et al. 2021, Obermeyer et al. 2019]. Por isso, a avaliação de equidade não deve se basear apenas em explicações, mas também em métricas específicas de *fairness* e análises complementares [Amann et al. 2020, Tung et al. 2025, Chen et al. 2021].

Em sistemas classificados como *Software as a Medical Device* (SaMD), órgãos regulatórios como FDA, ANVISA e entidades europeias exigem documentação detalhada sobre o propósito clínico, o desempenho do modelo e seus mecanismos de explicação [Rasheed et al. 2021, Aldhafeeri 2025, U.S. Food and Drug Administration 2021, Kiseleva et al. 2022].

Para atender a essas exigências, é necessário estabelecer uma estrutura de governança que acompanhe todo o ciclo de vida do modelo. Isso inclui a participação de equipes multidisciplinares, validação antes da implantação, monitoramento contínuo após o uso em produção e manutenção de documentação versionada e auditável [Reddy et al. 2020, Upadhyay et al. 2023, Stogiannos et al. 2023, Singh et al. 2025].

Em síntese, a explicabilidade é um componente importante, mas não suficiente para o uso responsável de modelos em saúde. Ela deve ser integrada a um conjunto mais amplo de práticas que envolvem governança, responsabilidade e avaliação contínua. As considerações finais retomam esses pontos à luz do estado atual e das perspectivas futuras do campo.

2.10. Considerações Finais e Perspectivas

O SHAP consolidou-se, na última década, como uma das principais abordagens para explicabilidade de modelos de aprendizado de máquina em saúde. Esse destaque não é acidental. Sua base nos valores de Shapley fornece garantias matemáticas que muitos métodos heurísticos não possuem: a decomposição da predição é exata, as contribuições individuais somam ao valor final do modelo e variáveis irrelevantes não recebem importância [Lundberg and Lee 2017, Bifarin 2022]. Além disso, a existência de implementações eficientes, como o TreeSHAP, e a possibilidade de análise tanto local quanto global tornam o método particularmente adequado ao contexto clínico, que exige compreender decisões individuais e, ao mesmo tempo, avaliar padrões populacionais.

Ao longo deste capítulo, foi possível observar, no entanto, que essas propriedades não eliminam limitações importantes. Os valores SHAP descrevem o comportamento do modelo, não o fenômeno clínico em si. Eles capturam associações presentes nos dados, mas não estabelecem relações causais [Bifarin 2022, Hettikankanamage et al. 2025]. Além disso, suas interpretações dependem da qualidade dos dados, da presença de variáveis correlacionadas e das escolhas feitas na implementação [Hu et al. 2024a, Liu et al. 2022].

Esses limites não invalidam o uso do SHAP, mas indicam a necessidade de uma aplicação cuidadosa. As explicações devem ser interpretadas em conjunto com o conhecimento clínico, avaliadas quanto à plausibilidade e analisadas em diferentes subgrupos [Prendin et al. 2023, Hakkoum et al. 2024]. Também é fundamental explicitar as limitações do método, especialmente em relação ao que ele não permite afirmar.

Um ponto central é que explicabilidade não deve ser tratada como um objetivo isolado. Uma explicação pode ser matematicamente correta e ainda assim pouco útil na prática clínica, se não for compreensível, relevante ou acionável [Tonekaboni et al. 2019]. O valor do SHAP depende de sua integração em um contexto mais amplo, que inclui validação clínica rigorosa, análise de equidade, consideração de aspectos causais e estruturas de governança bem definidas [Stogiannos et al. 2023].

Diversas questões permanecem em aberto e representam oportunidades de avanço. A integração entre métodos de explicabilidade e abordagens causais formais pode ampliar o valor interpretativo das análises, aproximando-as de decisões clínicas interventivas. A adaptação do SHAP a dados longitudinais e multimodais, cada vez mais comuns em sistemas de saúde, ainda demanda desenvolvimento metodológico [Hettikankanamage et al. 2025, Aldhafeeri 2025]. Além disso, há uma lacuna importante na avaliação do impacto das explicações sobre desfechos clínicos reais, indo além de medidas de usabilidade ou aceitação por parte dos usuários [Hettikankanamage et al. 2025, Singh et al. 2025]. Por fim, a consolidação de padrões regulatórios específicos para explicabilidade em sistemas classificados como *Software as a Medical Device* permanece um desafio relevante [U.S. Food and Drug Administration 2021, Singh et al. 2025].

Dessa forma, o SHAP é uma ferramenta poderosa para tornar modelos mais transparentes e auditáveis, mas seu valor depende de como é utilizado. Quando aplicado de forma criteriosa e integrado ao conhecimento clínico e às exigências regulatórias, ele contribui para o desenvolvimento de sistemas de inteligência artificial mais confiáveis e responsáveis [Amann et al. 2020]. Esse é o objetivo final da explicabilidade: não apenas tornar modelos compreensíveis, mas garantir que possam ser utilizados de forma segura, crítica e justificada na prática clínica.

Referências

[Alabi et al. 2023] Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A., and Mäkitie, A. (2023). Machine learning explainability in nasopharyngeal cancer survival using lime and shap. *Scientific Reports*, 13.

[Aldhafeeri 2025] Aldhafeeri, F. (2025). Governing artificial intelligence in radiology: A systematic review of ethical, legal, and regulatory frameworks. *Diagnostics*, 15.

[Amann et al. 2020] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., and Consortium, P. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):310.

[Arrieta et al. 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

[Auzine et al. 2024] Auzine, M. M., Khan, M. H.-M., Baichoo, S., Sahib, N. G., Bissoonauth-Daiboo, P., Gao, X., and Heetun, Z. (2024). Development of an ensemble cnn model with explainable ai for the classification of gastrointestinal cancer. *PLOS ONE*, 19.

[Bachmann 2025] Bachmann, S. (2025). Efficient xai: A low-cost data reduction approach to shap interpretability. *J. Artif. Intell. Res.*, 83.

[Bajwa et al. 2023] Bajwa, A., Nosheen, N., Talpur, K., and Akram, S. (2023). A prospective study on diabetic retinopathy detection based on modify convolutional neural network using fundus images at sindh institute of ophthalmology and visual sciences. *Diagnostics*, 13.

[Belle and Papantonis 2020] Belle, V. and Papantonis, I. (2020). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4.

[Bifarin 2022] Bifarin, O. O. (2022). Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *PLOS ONE*, 18.

[Binzagr 2024] Binzagr, F. (2024). Explainable ai-driven model for gastrointestinal cancer classification. *Frontiers in medicine*, 11:1349373.

[Bodria et al. 2021] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., and Rinzivillo, S. (2021). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37:1719–1778.

[Brenning 2021] Brenning, A. (2021). Interpreting machine-learning models in transformed feature space with an application to remote-sensing classification. *Machine Learning*, 112:3455–3471.

[Burrell 2016] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, 3.

[Cha et al. 2021] Cha, Y., Shin, J., Go, B., Lee, D.-S., Kim, Y., Kim, T., and Park, Y.-S. (2021). An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates. *Journal of environmental management*, 291:112719.

[Chen et al. 2024] Chen, C., Isa, N. A. M., and Liu, X. (2024). A review of convolutional neural network based methods for medical image classification. *Computers in biology and medicine*, 185:109507.

[Chen et al. 2021] Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4(1):123–144.

[Contreras et al. 2024] Contreras, J., Winterfeld, A., Popp, J., and Bocklitz, T. (2024). Spectral zones-based shap/lime: Enhancing interpretability in spectral deep learning models through grouped feature analysis. *Analytical Chemistry*, 96:15588 – 15597.

[Covert et al. 2020] Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in neural information processing systems*, 33:17212–17223.

[Daoud and Bayoumi 2019] Daoud, H. G. and Bayoumi, M. (2019). Efficient epileptic seizure prediction based on deep learning. *IEEE Transactions on Biomedical Circuits and Systems*, 13:804–813.

[Elshawi et al. 2019] Elshawi, R., Al-Mallah, M., and Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, 19.

[Esteva et al. 2017] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

[Gevaert and Saeys 2022] Gevaert, A. and Saeys, Y. (2022). Pdd-shap: Fast approximations for shapley values using functional decomposition.

[Ghasemi et al. 2024] Ghasemi, A., Hashtarkhani, S., Schwartz, D. L., and Shaban-Nejad, A. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. *Cancer Innovation*, 3.

[Ghosh and Khandoker 2024] Ghosh, S. K. and Khandoker, A. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, 14.

[Goldstein et al. 2017] Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208.

[Goodman and Flaxman 2017] Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

[Greenwell 2017] Greenwell, B. M. (2017). pdp: An r package for constructing partial dependence plots. *R J.*, 9:421.

[Greenwell and Boehmke 2020] Greenwell, B. M. and Boehmke, B. C. (2020). Variable importance plots - an introduction to the vip package. *R J.*, 12:343.

[Hakkoum et al. 2024] Hakkoum, H., Idri, A., and Abnane, I. (2024). Global and local interpretability techniques of supervised machine learning black box models for numerical medical data. *Eng. Appl. Artif. Intell.*, 131:107829.

[Hannun et al. 2019] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69.

[Harrigan et al. 2022] Harrigan, C. F., Morgenshtern, G., Goldenberg, A., and Chevalier, F. (2022). Considerations for visualizing uncertainty in clinical machine learning models.

[Hassija et al. 2023] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. (2023). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16:45–74.

[Hettikankanamage et al. 2025] Hettikankanamage, N. D., Shafiabady, N., Chatteur, F., Wu, R. M. X., Din, F. U., and Zhou, J. (2025). explainable artificial intelligence (xai): A systematic review for unveiling the black box models and their relevance to biomedical imaging and sensing. *Sensors (Basel, Switzerland)*, 25.

[Hooker et al. 2019] Hooker, G., Mentch, L., and Zhou, S. (2019). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31.

[Hooshyar and Yang 2024] Hooshyar, D. and Yang, Y. (2024). Problems with shap and lime in interpretable ai for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*, 12:137472–137490.

[Hu et al. 2024a] Hu, C., Gao, C., Li, T., Liu, C., and Peng, Z. (2024a). Explainable artificial intelligence model for mortality risk prediction in the intensive care unit: a derivation and validation study. *Postgraduate medical journal*, 100(1182):219–227.

[Hu et al. 2022] Hu, C., Li, L., ping Huang, W., Wu, T., Xu, Q., Liu, J., and Hu, B. (2022). Interpretable machine learning for early prediction of prognosis in sepsis: A discovery and validation study. *Infectious Diseases and Therapy*, 11:1117 – 1132.

[Hu et al. 2024b] Hu, X., Zhu, M., Feng, Z., and Stanković, L. (2024b). Manifold-based shapley explanations for high dimensional correlated features. *Neural networks : the official journal of the International Neural Network Society*, 180:106634.

[Huang and ao Marques-Silva 2024] Huang, X. and ao Marques-Silva, J. (2024). On the failings of shapley values for explainability. *Int. J. Approx. Reason.*, 171:109112.

[Joachim et al. 2026] Joachim, K., Sparks, O., Perrotta, A., Lin, A., Gettleman, B., Hamad, C., Jeong, S., Dingle, E., Stavrakis, A., and Christ, A. B. (2026). Evaluating the methodological suitability of partial dependence plots and shapley additive explanations for population-level interpretation of machine learning models in total joint arthroplasty. *Arthroplasty*, 8.

[Kelodjou et al. 2024] Kelodjou, G., Rozé, L., Masson, V., Galárraga, L., Gaudel, R., Tchuente, M., and Termier, A. (2024). Shaping up shap: enhancing stability through layer-wise neighbor selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13094–13103.

[Khan et al. 2024] Khan, M. M., Shah, N., Shaikh, N., Thabet, A., Alrabayah, T., and belkhair, S. (2024). Towards secure and trusted ai in healthcare: A systematic review of emerging innovations and ethical challenges. *International journal of medical informatics*, 195:105780.

[Kheder et al. 2025] Kheder, W., Leblouba, M., Rego, R., and Hamdoon, Z. (2025). Multicentre validation and clinical interpretation of an explainable gradient-boosting model for dental-implant survival/failure prediction. *Journal of dentistry*, page 106166.

[Kiseleva et al. 2022] Kiseleva, A., Kotzinos, D., and Hert, P. (2022). Transparency of ai in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations. *Frontiers in Artificial Intelligence*, 5.

[Lin et al. 2025] Lin, Q., Zhao, W., Zhang, H., Chen, W., Lian, S., Ruan, Q., Qu, Z., Lin, Y., Chai, D., and Lin, X. (2025). Predicting the risk of heart failure after acute myocardial infarction using an interpretable machine learning model. *Frontiers in Cardiovascular Medicine*, 12.

[Lip et al. 2010] Lip, G. Y., Nieuwlaat, R., Pisters, R., Lane, D. A., and Crijns, H. J. (2010). Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272.

[Lipton 2018] Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57.

[Liu et al. 2022] Liu, M., Ning, Y., Yuan, H., Ong, M. E. H., and Liu, N. (2022). Balanced background and explanation data are needed in explaining deep learning models with shap: An empirical study on clinical decision making.

[Lundberg and Lee 2017] Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

[Lundberg et al. 2020] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

[Lundberg et al. 2019] Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles.

[Luo et al. 2024] Luo, H., Xiang, C., Zeng, L., Li, S., Mei, X., Xiong, L., Liu, Y., Wen, C., Cui, Y., Du, L., Zhou, Y., Wang, K., Li, L., Liu, Z., Wu, Q., Pu, J., and Yue, R. (2024). Shap based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation. *Scientific Reports*, 14.

[Mahmoudi et al. 2020] Mahmoudi, E., Kamdar, N., Kim, N., Gonzales, G., Singh, K., and Waljee, A. (2020). Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *The BMJ*, 369.

[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.

[Mienye et al. 2025] Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., and Ilono, P. (2025). Deep convolutional neural networks in medical image analysis: A review. *Inf.*, 16:195.

[Miotto et al. 2018] Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246.

[Molnar 2020] Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 3 edition.

[Momenzadeh et al. 2022] Momenzadeh, A., Shamsa, A., and Meyer, J. G. (2022). Bias or biology? importance of model interpretation in machine learning studies from electronic health records. *JAMIA Open*, 5.

[Nohara et al. 2021] Nohara, Y., Matsumoto, K., Soejima, H., and Nakashima, N. (2021). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer methods and programs in biomedicine*, 214:106584.

[Obermeyer et al. 2019] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

[Olsen and Jullum 2025] Olsen, L. H. B. and Jullum, M. (2025). Improving the weighting strategy in kernelshap. In *World Conference on Explainable Artificial Intelligence*, pages 194–218. Springer.

[Patel et al. 2024] Patel, A. M., Baxter, W., and Porat, T. (2024). Toward guidelines for designing holistic integrated information visualizations for time-critical contexts: Systematic review. *Journal of Medical Internet Research*, 26.

[Patharkar et al. 2024] Patharkar, A., Cai, F., Al-Hindawi, F., and Wu, T. (2024). Predictive modeling of biomedical temporal data in healthcare applications: review and future directions. *Frontiers in Physiology*, 15.

[Ponce-Bobadilla et al. 2024] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C., Mensing, S., and Stodtmann, S. (2024). Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, 17.

[Prendin et al. 2023] Prendin, F., Pavan, J., Cappon, G., Favero, S. D., Sparacino, G., and Facchinetti, A. (2023). The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using shap. *Scientific Reports*, 13.

[Rajkomar et al. 2018] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18.

[Rasheed et al. 2021] Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A. I., Razi, A., and Qadir, J. (2021). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in biology and medicine*, 149:106043.

[Reddy et al. 2020] Reddy, S., Allan, S., Coghlan, S., and Cooper, P. (2020). A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27(3):491–497.

[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.

[Roshinta and Gábor 2024] Roshinta, T. A. and Gábor, S. (2024). A comparative study of lime and shap for enhancing trustworthiness and efficiency in explainable ai systems. *2024 IEEE International Conference on Computing (ICOCO)*, pages 134–139.

[Rudin 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

[Salih et al. 2023] Salih, A. M. A., Raisi-Estabragh, Z., Galazzo, I., Radeva, P., Petersen, S. E., Lekadir, K., and Menegaz, G. (2023). A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, 7.

[Selvaraju et al. 2020] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128(2):336–359.

[Shapley 1953] Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317. Reprinted in: Kuhn, H.W. (ed.) Classics in Game Theory, Princeton University Press, 1997.

[Singh et al. 2025] Singh, Y., Hathaway, Q. A., Keishing, V., Salehi, S., Wei, Y., Horvat, N., Vera-Garcia, D. V., Choudhary, A., Kh, A. M., Quaia, E., and Andersen, J. (2025). Beyond post hoc explanations: A comprehensive framework for accountable ai in medical imaging through transparency, interpretability, and explainability. *Bioengineering*, 12.

[Stenwig et al. 2022] Stenwig, E., Salvi, G., Rossi, P. S., and Skjærvold, N.-K. (2022). Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Medical Research Methodology*, 22.

[Stiglic et al. 2020] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., and Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10.

[Stogiannos et al. 2023] Stogiannos, N., Malik, R., Kumar, A., Barnes, A., Pogose, M., Harvey, H., McEntee, M., and Malamateniou, C. (2023). Black box no more: a scoping review of ai governance frameworks to guide procurement and adoption of ai in medical imaging and radiotherapy in the uk. *The British Journal of Radiology*, 96.

[Tan et al. 2023] Tan, Z., Tian, Y., and Li, J. (2023). Glime: General, stable and local lime explanation.

[Tonekaboni et al. 2019] Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. *Proceedings of Machine Learning Research*, 106:359–380. Machine Learning for Healthcare Conference.

[Topol 2019] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

[Tun et al. 2024] Tun, H., Rahman, H., Naing, L., and Malik, O. A. (2024). Trust in artificial intelligence–based clinical decision support systems among health care workers: Systematic review. *Journal of Medical Internet Research*, 27.

[Tung et al. 2025] Tung, T., Hasnaeen, S. M. N., and Zhao, X. (2025). Ethical and practical challenges of generative ai in healthcare and proposed solutions: a survey. *Frontiers in Digital Health*, 7.

[Upadhyay et al. 2023] Upadhyay, U., Gradisek, A., Iqbal, U., Dhar, E., Li, Y., and Syed-Abdul, S. (2023). Call for the responsible artificial intelligence in the healthcare. *BMJ Health & Care Informatics*, 30.

[U.S. Food and Drug Administration 2021] U.S. Food and Drug Administration (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Accessed: 2026.

[Vimbi et al. 2024] Vimbi, V., Shaffi, N., and Mahmud, M. (2024). Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection. *Brain Informatics*, 11.

[Vincent et al. 1996] Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C. K., Suter, P., and Thijs, L. G. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix). *Intensive care medicine*, 22(7):707–710.

[Viswan et al. 2023] Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., and Hajamohideen, F. (2023). Explainable artificial intelligence in alzheimer's disease classification: A systematic review. *Cognitive Computation*, 16:1–44.

[Wang et al. 2024] Wang, H., Liang, Q., Hancock, J. T., and Khoshgoftaar, T. (2024). Feature selection strategies: a comparative analysis of shap-value and importance-based methods. *Journal of Big Data*, 11:1–16.

[Wood et al. 2023] Wood, D., Papamarkou, T., Benatan, M., and Allmendinger, R. (2023). Model-agnostic variable importance for predictive uncertainty: an entropy-based approach. *Data Mining and Knowledge Discovery*, 38:4184 – 4216.

[working group and risk collaboration 2021] working group, S. and risk collaboration, E. C. (2021). Score2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in europe. *European Heart Journal*, 42(25):2439–2454.

[Xin et al. 2025] Xin, X., Hooker, G., and Huang, F. (2025). Pitfalls in machine learning interpretability: Manipulating partial dependence plots to hide discrimination. *Insurance: Mathematics and Economics*, page 103135.

[Yang et al. 2022] Yang, X., Chen, A., Pournejatian, N. M., Shin, H.-C., Smith, K. E., Parisien, C., Compas, C. B., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C., Lipori, G. P., Mitchell, D. A., Hogan, W., Shenkman, E., Bian, J., and Wu, Y. (2022). A large language model for electronic health records. *NPJ Digital Medicine*, 5.

[Yuan et al. 2025] Yuan, K., Yoon, C. H., Gu, Q., Munby, H., Walker, A., Zhu, T., and Eyre, D. W. (2025). Transformers and large language models are efficient feature extractors for electronic health record studies. *Communications Medicine*, 5.

[Zafar and Khan 2021] Zafar, M. R. and Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Mach. Learn. Knowl. Extr.*, 3:525–541.

[Zeng et al. 2025] Zeng, X., Chen, J., Zeng, X., Tang, X., and Peng, J. (2025). Integrating multiparametric mri radiomics and clinical models to assess sensitivity to neoadjuvant chemotherapy in breast cancer: A multicenter study. *Journal of Applied Clinical Medical Physics*, 26.

