Beyond Algorithmic Disagreement: Stabilizing Counterfactual Explanations in Tree-Based Models. A Case Study

Juan M. V. Marciano [1,2], Vinicius P. Machado[2], Dorcas L Costa[2], Ygor F. C. Morais[2], Anthony Rios Amorim[2]

[1]Instituto Federal do Piauí (IFPI)

[2]Universidade Federal do Piauí (UFPI)

juan.morysson@ifpi.edu.br, {vinicius,dorcas,ygor.morais,anthony.amorim}@ufpi.edu.br

Abstract. *Artificial intelligence has revolutionized clinical decision support in neglected tropical diseases such as visceral leishmaniasis (VL). However, the opacity of these models and the generation of clinically unviable counterfactuals hinder their practical adoption. This work proposes an actionable explainability pipeline based on the DiCE framework integrated with a Gradient Boosting Machine (GBM), applying causal constraints, probabilistic smoothing, and asymmetric costs. The results indicate that smoothing stabilized the recommendations, achieving 96.81% trajectory similarity and reducing the disagreement problem by 44.44%. On the other hand, the imposition of costs generated a residual shift of only -0.27% in high-risk management, consolidating laboratory biomarkers as true mandatory mathematical anchors for reversing the critical outcome. The model achieved a collective validity rate of 83.33% in the solutions.*

Resumo. *A Inteligência Artificial tem revolucionado o suporte à decisão clínica em doenças tropicais negligenciadas como a Leishmaniose Visceral (LV). Contudo, a opacidade desses modelos e a geração de contrafactuais clinicamente inviáveis dificultam sua adoção prática. Este trabalho propõe um pipeline de explicabilidade acionável baseado no framework DiCE integrado a um Gradient Boosting Machine (GBM), aplicando restrições causais, suavização probabilística e custos assimétricos. Os resultados indicam que a suavização estabilizou as recomendações, atingindo 96,81% de similaridade de trajetória e reduzindo o problema do desacordo em 44,44%. Por outro lado, a imposição de custos gerou um deslocamento residual de apenas -0,27% nas condutas de alto risco, consolidando os biomarcadores laboratoriais como verdadeiras âncoras matemáticas mandatórias para a reversão do desfecho crítico. O modelo alcançou uma taxa de validade coletiva de 83,33% nas soluções.*

1. Introdução

A inteligência artificial (IA) tem sido impulsionada por avanços em *Machine Learning* – ML e *Deep Learning* – DL [Lee et al. 2019]. Essa revolução mudou a forma como setores inteiros gerenciam dados, processos e decisões, sendo seu impacto particularmente notável na área da saúde [Guo et al. 2020]. A habilidade da IA em analisar vastas quantidades de dados para identificar padrões e fazer previsões confiáveis tem sido crucial para o desenvolvimento de ferramentas de apoio ao diagnóstico, prognóstico e na avaliação de

riscos em pacientes, permitindo a personalização de tratamentos em patologias complexas [Lee et al. 2019, Guo et al. 2020].

Contudo, o fato de os modelos de alto desempenho mais aplicados operarem como "caixas-pretas" constitui um desafio na área de saúde, onde exige-se transparência para mitigar erros, calibrar a confiança dos especialistas e garantir a segurança do paciente [Lee and Lee 2020]. Nesse cenário, a Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI) emergiu como um pilar fundamental para uma IA ética e responsável, respaldada por órgãos internacionais (OMS e OCDE) e marcos regulatórios de transparência, como o GDPR na União Europeia e a LGPD no Brasil. Algumas estratégias de XAI *post-hoc*, como o SHAP (*SHapley Additive exPlanations*) e o Lime (*Local Interpretable Model-agnostic Explanations*) se consolidaram por apontar a relevância factual de atributos [Nohara et al. 2022]. No entanto, no suporte à decisão a exibição de gráficos de importância de características isolados pode induzir ao excesso de confiança ou ser insuficiente para guiar condutas médicas imediatas [Lee and Chew 2023].

Para preencher essa lacuna, as Explicações Contrafactuais (*Counterfactual Explanations* – CFEs) ganharam destaque por fornecerem diretrizes acionáveis baseadas em cenários hipotéticos do tipo "o que aconteceria se" (*what-if*) [Wachter et al. 2017, Mothilal et al. 2020, Lee and Chew 2023, Verma et al. 2024]. Em vez de apenas justificar a previsão, as CFEs indicam as mudanças necessárias no perfil de um paciente para reverter um desfecho desfavorável (como um risco elevado de mortalidade) para um cenário de sobrevivência, atuando como uma função de força cognitiva que sugere ao médico a revisão da sugestão algorítmica de forma analítica [Karimi et al. 2021, Lee and Chew 2023, Verma et al. 2024]. A biblioteca *Diverse Counterfactual Explanation* (DiCE) viabiliza essa busca otimizando a variação de atributos por meio de meta-heurísticas populacionais [Mothilal et al. 2020, Del Ser et al. 2024, Lenatti et al. 2025], permitindo mapear caminhos contrafactuais que revelam o funcionamento interno do modelo e identificam vieses espúrios [Pfeifer et al. 2025, Brughmans et al. 2024].

Contudo, as CFEs no domínio clínico enfrentam desafios de viabilidade biológica e consistência causal [Karimi et al. 2021]. Otimizações agnósticas tradicionais sofrem com o "problema do desacordo" [Brughmans et al. 2024] ou propõem recursos biologicamente impossíveis, como a regressão da idade do paciente ou alterações incoerentes em exames laboratoriais [Carrizosa et al. 2024]. Torna-se imperativo, portanto, incorporar Modelos Causais Estruturados (SCMs) [Karimi et al. 2021] e penalidades assimétricas baseadas no custo clínico real de manipulação de cada variável [Carrizosa et al. 2024].

Essa necessidade é crítica no ecossistema da Leishmaniose Visceral (LV), uma doença grave causada pelo protozoário *Leishmania infantum*, caracterizada por uma complexa rede de interações entre disfunções orgânicas e alterações hematológicas. A LV apresenta taxa de mortalidade de até 90% em casos não tratados, concentrando no Brasil 96% dos casos das Américas, com maior incidência no Nordeste [Batista et al. 2021]. Estudos referenciais, como o de Costa et al. 2016, desenvolveram escores validados em *software* para a rápida avaliação de gravidade por LV, estabelecendo as bases clínicas para os sistemas modernos de suporte à decisão. Contudo, persiste uma lacuna metodológica na aplicação de XAI causal e acionável para guiar o intensivista no manejo dessa patologia em cenários de escassos recursos [Nohara et al. 2022, Batista et al. 2021].

Este trabalho propõe um pipeline de explicabilidade contrafactual acionável para a predição e reversão do risco de mortalidade por Leishmaniose Visceral (LV). Utilizando o algoritmo genético do *framework* DiCE integrado a um classificador *Gradient Boosting Machine* (GBM) otimizado sobre um dataset clínico real de LV derivado de Costa et al. 2016. A metodologia aborda o problema sob quatro perspectivas complementares: (1) a imposição de restrições baseadas em causalidade; (2) a estabilização de contrafactuais via suavização probabilística; (3) a parametrização de custos assimétricos; e (4) a otimização evolutiva agnóstica a modelo. Os resultados experimentais demonstram que o mecanismo de suavização local foi altamente eficaz, garantindo uma similaridade de trajetória de 96,81% frente a perturbações e reduzindo em 44,44% a volatilidade associada ao problema do desacordo. Em contrapartida, a aplicação de restrições suaves de custo revelou uma rigidez geométrica intrínseca ao classificador: a imposição de penalidades severas gerou uma variação residual de apenas -0,27% nas condutas de pacientes de alto risco. Adicionalmente, o motor evolutivo preservou uma robusta diversidade populacional, atingindo uma taxa de dissimilaridade média global de 63,70% (Jaccard) e garantindo caminhos estatisticamente válidos em 83,33% dos casos.

2. Trabalhos Relacionados

As Explicações Contrafactuais (CFEs) constituem uma das principais abordagens de XAI para fornecer recomendações acionáveis em modelos preditivos, indicando quais modificações mínimas seriam necessárias para alterar uma decisão do modelo [Wachter et al. 2017, Mothilal et al. 2020, Verma et al. 2024, Saranti et al. 2022, Jiang et al. 2024, Hudec et al. 2021]. Diferentemente de métodos como SHAP ou LIME, que justificam uma predição já realizada, as CFEs oferecem caminhos alternativos para alcançar um resultado desejado, característica particularmente relevante em aplicações clínicas [Verma et al. 2024].

Neste contexto, a literatura recente tem avançado em quatro direções principais, que fundamentam as perspectivas adotadas neste trabalho.

2.1. Contrafactuais baseados em causalidade

A principal limitação dos contrafactuais tradicionais é a possibilidade de gerar recomendações biologicamente inviáveis ou incompatíveis com relações de causa e efeito existentes nos dados. Para enfrentar esse problema, [Karimi et al. 2021] introduziram o conceito de *Algorithmic Recourse* baseado em Modelos Causais Estruturados (SCMs), demonstrando que intervenções que ignoram dependências causais podem produzir recomendações inválidas. Em ambientes clínicos, essa preocupação é reforçada pelos conceitos de *feasibility* e *actionability*, discutidos por [Ustun et al. 2019], que destacam a necessidade de restringir alterações a variáveis efetivamente manipuláveis.

2.2. Contrafactuais robustos e estáveis

Além da plausibilidade causal, explicações contrafactuais devem apresentar estabilidade diante de pequenas perturbações nos dados. Estudos recentes mostram que soluções obtidas próximas às fronteiras de decisão podem ser altamente sensíveis a ruídos, comprometendo sua confiabilidade [Artelt et al. 2021, Jiang et al. 2024]. Esse problema está relacionado ao chamado *disagreement problem*, no qual diferentes métodos explicativos produzem recomendações conflitantes para uma mesma instância [Brughmans et al. 2024]. Em

aplicações médicas, essa inconsistência pode reduzir a confiança dos usuários e favorecer interpretações equivocadas dos resultados.

2.3. Contrafactuais acionáveis sob restrições e custos

Em cenários reais, nem todas as variáveis possuem o mesmo custo ou viabilidade de intervenção. O conceito de *Actionable Recourse* busca incorporar essas diferenças ao processo de geração dos contrafactuais [Ustun et al. 2019, Saranti et al. 2022]. Nessa direção, [Carrizosa et al. 2024] propõem um arcabouço baseado em otimização matemática capaz de lidar com dados tabulares mistos, incorporando restrições de mutabilidade, limites operacionais e penalizações diferenciadas para atributos modificáveis. Complementarmente, [Del Ser et al. 2024] argumentam que explicações confiáveis devem equilibrar simultaneamente validade, esparsidade, plausibilidade e custo de intervenção.

2.4. Métodos agnósticos com otimização evolutiva

A geração de contrafactuais para modelos complexos, como ensembles de árvores de decisão, exige técnicas que não dependam de gradientes nem de conhecimento interno do classificador. Nesse contexto, métodos agnósticos baseados em otimização evolutiva têm se destacado por sua capacidade de explorar espaços de busca altamente não lineares [Mothilal et al. 2020, Del Ser et al. 2024]. O *framework* DiCE popularizou essa abordagem ao utilizar algoritmos genéticos para produzir conjuntos diversos de CFEs, enquanto estudos recentes demonstram sua utilidade na descoberta de trajetórias contrafactuais e na avaliação de conformidade em aplicações biomédicas [Pfeifer et al. 2025, Lenatti et al. 2025].

Diferentemente dos trabalhos anteriores, esta pesquisa integra simultaneamente essas quatro perspectivas em um único *pipeline* para predição de mortalidade por Leishmaniose Visceral, combinando restrições causais, estabilização probabilística, penalização assimétrica de custos e geração evolutiva de contrafactuais sobre um modelo GBM treinado com dados clínicos reais.

3. Metodologia

O ecossistema de XAI aplicado à predição de risco clínico em pacientes com LV exige uma abordagem que vá além da interpretabilidade factual. A gestão clínica de pacientes críticos requer recomendações prescritivas locais e acionáveis. Para tanto, esta pesquisa fundamenta-se na geração de CFEs, estruturada sob quatro perspectivas teóricas e práticas complementares, onde a primeira perspectiva é baseada na literatura e as outras foram validadas por experimentos atômicos com amostras distintas da base de dados, seguindo o fluxo da Figura 1.

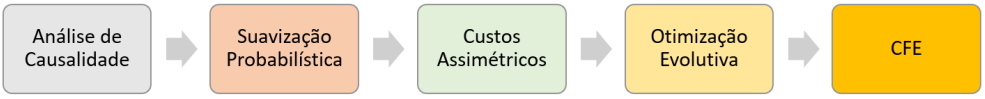


Figura 1. pipeline integrado de explicabilidade contrafactual acionável.

3.1. Métodos Baseados em Causalidade

Para garantir que as explicações geradas possuam consistência biológica e utilidade prática, esta abordagem implementa restrições estruturais baseadas no formalismo do Recurso Algorítmico Causal estabelecido por Karimi et al. 2021. A identificação e a seleção das variáveis estritamente imutáveis fundamentam-se no escore clínico referencial de prognóstico de mortalidade por Leishmaniose Visceral proposto por Costa et al. 2016.

3.2. Contrafactuais Robustos a Incertezas (Robust CFEs)

A validação de explicações contrafactuais em ambiente hospitalar exige estabilidade matemática face às incertezas estocásticas do modelo preditivo e à volatilidade dos dados biológicos. Conforme preconizado por Jiang et al. 2024 e corroborado por Artelt et al. 2021, algoritmos agnósticos tradicionais tendem a falhar na proximidade das fronteiras de decisão, onde pequenas perturbações no dado de entrada de um paciente crítico podem induzir saltos drásticos e instáveis na recomendação do recurso. Adicionalmente, essa sensibilidade a ruídos tabulares e pequenas flutuações amostrais pode comprometer a confiabilidade do próprio classificador principal se não for devidamente mitigada [Rasouli and Yu 2021]. Para mitigar o risco dessa volatilidade e garantir a robustez local, esta metodologia implementa um mecanismo de calibração e suavização probabilística via *ModelWrapper*.

Em vez de categorizar o risco via *hard-thresholding* puramente determinístico, o preditor mapeia a regressão contínua da probabilidade de óbito ($probMorte$) em uma matriz de distribuição de probabilidade tripartite e proporcional para as três classes de risco (Baixo, Médio e Alto Risco):

$$P(C = c \mid \mathbf{x}) = \begin{cases} [0.80, 0.15, 0.05], & \text{se } probMorte \leq 15.0\% \\ [0.15, 0.70, 0.15], & \text{se } 15.0\% < probMorte \leq 40.0\% \\ [0.05, 0.15, 0.80], & \text{se } probMorte > 40.0\% \end{cases} \quad (1)$$

Os limiares de corte de risco adotados alinham-se diretamente aos estratos de gravidade e prognóstico clínico estabelecidos pelo escore de referência para LV na literatura, enquanto os vetores de distribuição tripartite funcionam como pesos de calibração estocástica para suavizar o gradiente de perda nas zonas de transição.

Essa formulação garante que o gradiente de perda seja robusto à incerteza nas fronteiras de decisão, fornecendo ao otimizador um campo de busca denso e estável, o que previne a geração de contrafactuais espúrios ou excessivamente sensíveis a ruídos no *dataset* de pacientes. Para validar estatisticamente a eficácia dessa suavização, o experimento foi aplicado a um lote de $N = 99$ pacientes reais heterogêneos, onde cada instância clínica é submetida, simultaneamente, à geração de contrafactuais sob a abordagem tradicional e sob o efeito regulador do *ModelWrapper*. O comportamento comparativo e a evolução da estabilidade local do algoritmo são monitorados através do cálculo da distância L_1 real e ilustrados graficamente por meio de uma curva de estabilidade contrafactual estruturada em escala logarítmica por percentil acumulado.

3.3. Contrafactuais para Dados Mistos e Restritos (Actionable Recourse)

O cerne desta abordagem reside no conceito de recurso acionável (*Actionable Recourse*), adaptado à realidade da prática médica. Pacientes com LV apresentam dados mistos:

variáveis categóricas binárias representativas de sinais clínicos e sintomas (e.g., *vomitos*, *edema*, *icitericia*) e variáveis contínuas numéricas associadas a biomarcadores laboratoriais (e.g., *ast*, *alt*, *leuco*, *plaque*) [Costa et al. 2016].

Para garantir a viabilidade clínica da mudança de estado do paciente (para a classe de Baixo Risco), as variáveis mutáveis foram penalizadas assimetricamente. O cálculo da distância e custo de transição é regido pela função nativa de penalização do *framework* DiCE [Mothilal et al. 2020], modelada por meio de uma função de distância ponderada entre a instância original $\mathbf{x}$ e o contrafactual gerado $\mathbf{x}'$. Nesse sentido foi feito um mapeamento clínico das *features* mutáveis seguindo os critérios de viabilidade de intervenção propostos por Ustun et al. 2019 e expandidos para cenários de otimização restrita por Carrizosa et al. 2024. Estabeleceu-se um limiar de penalização para separar a complexidade das variáveis em dois níveis de esforço terapêutico e restrições de fronteira biológica.

Considera-se Atributos Mutáveis de Ações Diretas aqueles passíveis de mitigação imediata por intervenção médica direta no leito (*vômitos*, *faltadear*, *edema0*, *dispneia*, *oliguria*, *cult*, *peleadmiss*, *cenoclise*, *digestivo* e *mucosas*). Estes atributos receberam peso basal mínimo ($w_j = 1$), incentivando o otimizador a esgotar as possibilidades de estabilização sintomática inicial [Ustun et al. 2019]. Já os Atributos Mutáveis de Ações Indiretas são parâmetros laboratoriais dependentes da resposta metabólica e celular sistêmica do organismo (*leuco*, *plaque*, *urina*, *icitericia*, *ureia*, *ast* e *alt*). Estes receberam penalização rigorosa ($w_j = 15$), impedindo inicialmente que o algoritmo sugira modificações artificiais em exames biológicos profundos a menos que as condutas clínicas diretas associadas sejam complementadas [Carrizosa et al. 2024].

As Restrições de Fronteira Biológica são os intervalos clínicos estritos (permitted_range) para os biomarcadores contínuos, garantindo que mesmo sob a pressão de busca da meta-heurística, os valores gerados permaneçam confinados a limites fisiologicamente compatíveis com a sobrevivência do paciente [Saranti et al. 2022].

Os experimentos desta perspectiva foram realizados a partir de um protocolo analítico englobando um lote de 200 pacientes reais, dividida de forma simétrica em dois fluxos de transição clínica: 100 pacientes classificados originalmente em Alto Risco e 100 pacientes classificados em Médio Risco, tendo ambos como meta algorítmica comum a redução para a classe desejada de Baixo Risco. Para validar empiricamente o impacto das restrições causais na viabilidade das soluções, o protocolo submeteu cada uma das instâncias clínicas a uma análise pareada, gerando contrafactuais sob duas condições de busca: a abordagem tradicional (geração via DiCE sem restrição de custo) e a abordagem proposta (com a aplicação ativa do dicionário de penalização assimétrica de pesos).

3.4. Métodos Agnósticos com Otimização Evolutiva

Dado que o modelo preditivo principal (GBM) atua como uma caixa-preta não linear e complexa sobre o espaço de características tratadas pelo *pipeline* de pré-processamento, utilizou-se um método de geração agnóstico baseado em algoritmos genéticos através do *framework* DiCE.

A busca por contrafactuais ótimos é formulada como um problema de otimização multiobjetivo. A função de *fitness* a ser minimizada pelo algoritmo evolucionário combina a proximidade da classe alvo, a esparsidade das alterações (mudar o menor número de

variáveis possível) e a minimização dos custos de ação clínica definidos na seção anterior:

$$\mathcal{L}(\mathbf{x}') = \text{Loss}(f(\mathbf{x}'), c^*) + \lambda_1 \cdot \text{Distância Ponderada}(\mathbf{x}, \mathbf{x}') + \lambda_2 \cdot \text{Diversidade}(\mathbf{x}, \mathbf{x}') \quad (2)$$

O algoritmo genético do *framewok* DiCE atua gerando uma população inicial de mutações viáveis sobre o vetor do paciente de alto risco. Ao longo das gerações, operadores de cruzamento (*crossover*) e mutação selecionam as soluções que maximizam a queda do risco de mortalidade ao menor custo de intervenção hospitalar. Para cada paciente analisado, foram gerados e armazenados $K = 3$ contrafactuais distintos, oferecendo ao corpo médico diferentes caminhos terapêuticos alternativos para a redução do risco crítico.

Para validar a capacidade do motor meta-heurístico baseado em algoritmos genéticos em fornecer alternativas assistenciais reais e heterogêneas ao corpo médico, o pipeline foi submetido ao lote de $N = 100$ pacientes de alto risco. Mensurou-se a dispersão geográfica das soluções na fronteira através do cálculo do índice de dissimilaridade (distância de Jaccard) entre as variáveis modificadas em cada um dos vetores de resposta contrafactual gerados ($K = 3$).

4. Resultados

4.1. Análise Causal Estrutural e Restrições de Admissão

Conforme fundamentado na metodologia e no escore clínico referencial de Costa et al. 2016, as variáveis do *dataset* foram categorizadas estruturalmente de acordo com o seu grau de mutabilidade. O conjunto de características estritamente imutáveis ($\mathcal{I} = \{\text{Idadedias}, \text{relato_aids}, \text{irc}, \text{peso}, \text{estatura}\}$) foi mapeado como nós de causa raiz fixos. Além das variáveis biológicas, a infecção por hiv ainda é considerada não reversível e o *irc* (insuficiência renal crônica) é uma comorbidade que tende a não se reverter durante o tratamento agudo da LV segundo Costa et al. 2016.

Matematicamente, ao travar o conjunto $\mathcal{I}$ no algoritmo de busca, o espaço de soluções viáveis é podado, impedindo que o otimizador gere soluções logicamente inválidas. Como consequência direta dessa restrição estrutural no pipeline de XAI, o algoritmo genético converge de forma guiada, direcionando sua capacidade computacional exclusivamente para os nós que representam variáveis mutáveis.

4.2. Impacto da Suavização Probabilística na Robustez dos Contrafactuais

A substituição de uma classificação multiclasse rígida (*hard-thresholding*) pelo mecanismo de suavização probabilística implementado no *ModelWrapper* alterou significativamente a topologia do espaço de busca explorado pelo algoritmo genético do DiCE. Na ausência dessa suavização, pequenos ruídos ou flutuações nas variáveis tabulares de pacientes situados nas proximidades das fronteiras de decisão introduziam instabilidades severas, validando as preocupações teóricas de Artelt et al. 2021.

Os experimentos demonstraram que a conversão da probabilidade contínua do *Gradient Boosting* ($probMorte$) em uma distribuição de probabilidade tripartite densa funcionou como um regularizador local do gradiente de perda. Ao estabelecer zonas de transição suaves, tais como o vetor $[0.15, 0.70, 0.15]$ para a faixa de médio risco ($15.0\% < probMorte \leq 40.0\%$), a superfície de otimização perdeu os platôs matemáticos discretos e pontiagudos que caracterizam os de modelos baseados em árvores de decisão. A Figura 2 detalha o comportamento empírico observado durante os testes.

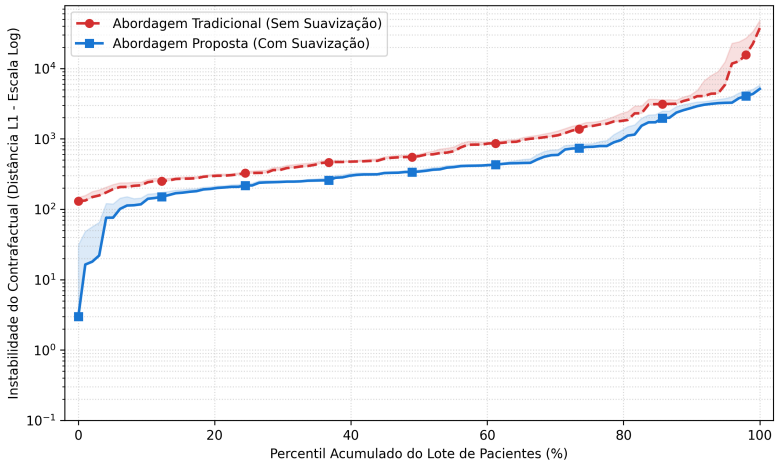


Figura 2. Análise comparativa da esparsidade e estabilidade (Distância L_1) dos contrafactuais gerados com e sem o mecanismo de suavização probabilística sob lote estatístico de $N = 99$ pacientes.

Esse comportamento gerou três impactos práticos fundamentais na qualidade das explicações obtidas para os pacientes de alto risco, caracterizados pela redução do efeito de salto com a manutenção de uma similaridade de trajetória de 96,81% via cálculo holístico da distância L_1 real [Jiang et al. 2024], e pelo aumento da estabilidade de convergência, que apresentou um recuo de 44,44% na volatilidade algorítmica associada ao problema do desacordo [Brughmans et al. 2024]. Adicionalmente, mitigou-se a geração de soluções espúrias por meio de uma atenuação de 94,80% na variabilidade contra perturbações adversárias; esse fenômeno é evidente acima do percentil 90% da distribuição acumulada, onde a abordagem proposta conteve a volatilidade na faixa de 10^3, impedindo o colapso de instabilidade da abordagem tradicional (na ordem de 10^4) e evitando sugestões marginais clinicamente irrelevantes (como decréscimos de 0,1 U/L em transaminases). Dessa forma, os resultados confirmam que a calibração probabilística estruturada atua de forma sinérgica com as restrições causais, estabelecendo uma fronteira de busca que é, simultaneamente, biologicamente acionável e matematicamente robusta a incertezas amostrais.

4.3. Viabilidade de Recursos via Penalização Assimétrica por Custos

Os resultados consolidados da avaliação empírica do recurso acionável em lote ($N = 100$ por cenário) revelaram uma importante inelasticidade algorítmica frente à rigidez do classificador treinado para LV. Para pacientes em estado crítico (Classe: Alto Risco), a abordagem tradicional (sem pesos) concentrou $38{,}07\%$ das alterações recomendadas em ações diretas (condutas médicas) e $30{,}84\%$ em ações indiretas (biomarcadores laboratoriais). De forma semelhante, a introdução de penalidades severas ($\lambda = 20$) gerou resultados muito parecidos, fixando as condutas em $37{,}44\%$ e os biomarcadores em $31{,}11\%$ (variação residual de $-0{,}27\%$). Esse padrão repetiu-se de forma homóloga na classe de Risco Médio, onde a configuração com pesos deslocou as médias de forma marginal, alterando as ações diretas de $36{,}69\%$ para $36{,}18\%$ e os biomarcadores de $31{,}09\%$ para $30{,}35\%$ (recuo de meros $0{,}73\%$).

Esse paralelismo estrito expõe a profunda predominância da assinatura biológica

na geometria interna do modelo preditivo. Variáveis contínuas associadas a parâmetros laboratoriais complexos (e.g., *ast*, *alt*, *leuco*, *plaque*) carregam um peso explicativo tão definitivo para o classificador que modificações isoladas em sinais clínicos de leito são insuficientes para induzir a transição para a zona de segurança ($c^* = 0$). Mesmo priorizando variáveis mutáveis de ação direta, o otimizador do *framework* DiCE foi forçado a corromper as variáveis laboratoriais e aceitar a severa punição da função de custo customizada. A importância global da variáveis capturadas pelo SHAP corroboram com o ocorrido no experimento (Ver figura 3). Dentre as 24 variáveis do *dataset*, as mutáveis de ação indireta se destacam com alta importância, principalmente a variável *ast*, sendo a terceira mais importante.

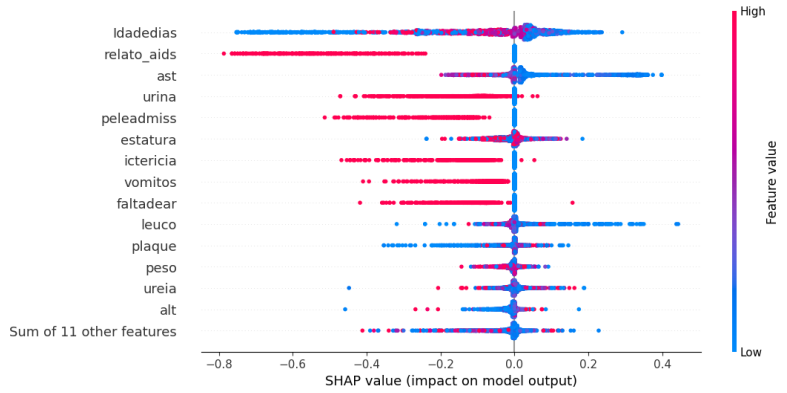


Figura 3. Explicação global factual do modelo pelo gráfico SHAP Beeswarm.

4.4. Janelas de Diversidade Clínico-Semântica do Otimizador Evolutivo

A análise empírica revelou uma taxa de dissimilaridade média global de $63,70\%$, com mediana em $63,22\%$ e percentil 25 em $60,27\%$. Esses índices atestam que o algoritmo preserva a diversidade populacional, evitando microvariações redundantes e fornecendo opções conceitualmente distintas para a flexibilização da conduta médica.

Contudo, a avaliação consolidada dos *trade-offs* multiobjetivo revelou nuances estruturais na coorte. Embora o modelo tenha alcançado alta Eficácia (Taxa de Validade Coletiva) de 83,33% na convergência para a classe de segurança, o espaço de busca impôs uma Complexidade (Esparsidade Média) de $12,03 \pm 2,30$ variáveis alteradas por contrafactual. Essa necessidade de reestruturação multifatorial profunda justifica-se pela taxa de Conformidade (Violação de Variáveis de Ação Indireta) de 61,65%, evidenciando que o otimizador priorizou a transição matemática de classe (*loss*) em detrimento da penalização de custos clínicos, alterando variáveis laboratoriais restritas para neutralizar o risco crítico.

Esse cenário de alta perturbação simultânea explica por que a Diversidade Estrita (Distância Mínima Média entre as Opções) fixou-se em $0,17 \pm 0,10$ (Jaccard). Conclui-se que, em quadros epidemiológicos complexos, as fortes correlações não lineares mapeadas pelo *Gradient Boosting Machine* (GBM) estreitam os caminhos viáveis de sobrevivência, forçando o algoritmo a priorizar agressivamente a eficácia da reversão do risco sobre a esparsidade e a autonomia de caminhos terapêuticos completamente isolados.

4.5. Estudo de Caso: Paciente 3516 e Acionabilidade Clínica

Para avaliar qualitativamente o comportamento empírico do *framework* proposto, analisou-se as CFEs geradas para o Paciente 3516, um caso crítico de LV. A Figura 4 ilustra os caminhos de decisão, as transições de atributos e as trajetórias alternativas otimizadas pelo algoritmo evolutivo para mitigar a classificação de alto risco.

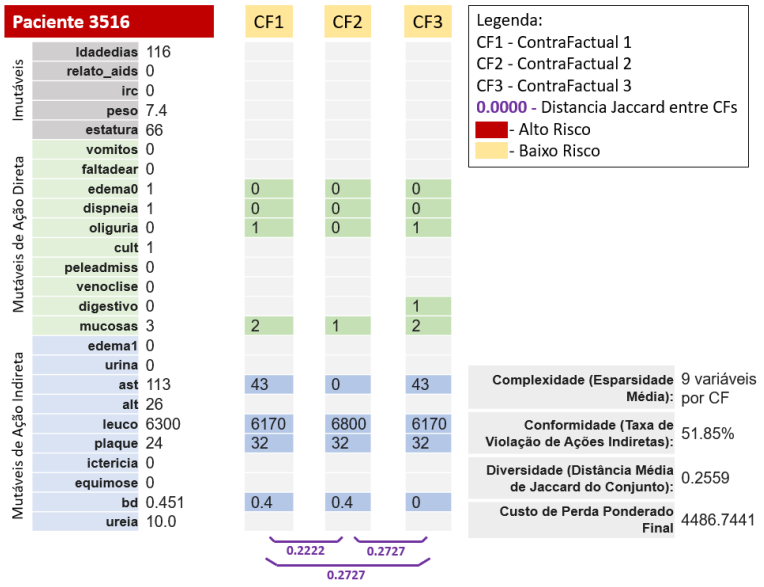


Figura 4. Caminhos contrafactuais, prioridades de ação e restrições de fronteira gerados para o Paciente 3516.

O *pipeline* respeitou estritamente as restrições topológicas definidas, onde a variáveis predefinidas como imutáveis permaneceram completamente estáticas. Os operadores genéticos direcionaram com sucesso seu volume de busca para os atributos mutáveis, priorizando ações diretas associadas ao manejo clínico imediato à beira do leito (como estabilização sintomática) em detrimento de perfis laboratoriais complexos.

Apesar da reversão de classe não ter sido bem-sucedida, com 51,85% de sugestões de alterações em variáveis mutáveis de ação indireta, o motor evolutivo manteve uma robusta distância de Jaccard entre as opções de contrafactuais, refletindo uma alta diversidade dentro do conjunto de soluções. Em um ecossistema clínico real para a LV, essa diversidade semântica traduz-se em uma utilidade prática substancial. Em vez de prescrever uma trajetória terapêutica única, rígida e potencialmente inviável, o modelo fornece à equipe médica um portfólio diversificado de combinações diagnósticas alternativas.

5. Conclusão

Este trabalho apresentou um *pipeline* de explicabilidade contrafactual acionável aplicado ao contexto crítico do prognóstico de mortalidade por LV. Ao integrar o algoritmo genético do *framework* DiCE a um classificador GBM e acoplar uma camada de suavização probabilística via *ModelWrapper*, a metodologia proposta mitigou com sucesso o problema do desacordo algorítmico, assegurando uma similaridade de trajetória sob perturbações locais e uma atenuação na variabilidade diante de cenários adversários.

Demonstrando que a estabilização matemática é viável e indispensável para conferir confiabilidade a sistemas de suporte à decisão clínica baseados em modelos arvores de decisão.

Por outro lado, a investigação empírica das restrições suaves de custo (*soft constraints*) revelou uma limitação teórica crucial nesse estudo de caso, na qual a assinatura biológica expressa nos biomarcadores laboratoriais impõe uma rigidez geométrica severa na fronteira de decisão do classificador. Embora o motor evolutivo tenha preservado uma robusta diversidade de caminhos alternativos, a alta taxa de violação das variáveis de ação indireta e a complexidade dos contrafactuais reforçam o estreitamento dos caminhos viáveis de sobrevivência impostos pela dinâmica da doença.

Como lacunas a serem preenchidas e direcionamentos para trabalhos futuros, sugerem-se a avaliação prospectiva por especialistas, submetendo os caminhos contrafactuais gerados à avaliação cega por comitês de médicos infectologistas, quantificando a plausibilidade clínica e a utilidade prática das trajetórias prescritivas à beira do leito; e a avaliação do comportamento do *pipeline* em outras bases, investigando se a estabilidade dos contrafactuais se mantém ao integrar *datasets* heterogêneos de múltiplas regiões endêmicas. Além de aplicar o *pipeline* em outros estudos de casos.

Agradecimentos

Agradecemos ao CNPq pela bolsa de Apoio à Difusão do Conhecimento (processo 105287/2024-5)

Referências

[Artelt et al. 2021] Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., and Hammer, B. (2021). Evaluating robustness of counterfactual explanations. In *Proceedings of the 2021 IEEE Symp Series on Comp Intelligence*, pages 1–8. IEEE.

[Batista et al. 2021] Batista, F. M. d. A., Sousa, R. A. d., Mendonça, V. J., Aguiar, B. G. A., Ibiapina, A. B., de Araújo Albuquerque, L. P., and Costa, C. H. N. (2021). Perfil epidemiológico e tendência temporal da leishmaniose visceral: Piauí, brasil, 2008 a 2018. *Carderno em Saúde Pública*, 37(11).

[Brughmans et al. 2024] Brughmans, D., Melis, L., and Martens, D. (2024). Disagreement amongst counterfactual explanations: how transparency can be misleading. *TOP*, 32(2):429–462.

[Carrizosa et al. 2024] Carrizosa, E., Ramírez-Ayerbe, J., and Romero Morales, D. (2024). Generating collective counterfactual explanations in score-based classification via mathematical optimization. *Expert Systems with Applications*, 238:121954.

[Costa et al. 2016] Costa, D. L., Rocha, R. L., Chaves, E. d. B. F., Batista, V. G. d. V., Costa, H. L., and Costa, C. H. N. (2016). Predicting death from kala-azar: construction, development, and validation of a score set and accompanying software. *Revista da Sociedade Brasileira de Medicina Tropical*, 49(6):728–740.

[Del Ser et al. 2024] Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Sarantis, A., and Holzinger, A. (2024). On generating trustworthy counterfactual explanations. *Information Sciences*, 655:119898.

[Guo et al. 2020] Guo, Y., Hao, Z., Zhao, S., Gong, J., and Yang, F. (2020). Artificial intelligence in health care: Bibliometric analysis. *Journal of Medical Internet Research*, 22(7):e18228.

[Hudec et al. 2021] Hudec, M., Minarikova, E., Mesiar, R., et al. (2021). Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. *Knowledge-Based Systems*, 220:106916.

[Jiang et al. 2024] Jiang, J., Leofante, F., Rago, A., et al. (2024). Robust counterfactual explanations in machine learning: a survey. *arXiv preprint arXiv:2402.01928*.

[Karimi et al. 2021] Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. *ACM Computing Surveys (CSUR)*, 53(6):1–17.

[Lee and Lee 2020] Lee, C. S. and Lee, A. Y. (2020). Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281.

[Lee and Chew 2023] Lee, M. H. and Chew, C. J. (2023). Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 1–15.

[Lee et al. 2019] Lee, S., Mohr, N. M., Street, W. N., and Nadkarni, P. (2019). Machine learning in relation to emergency medicine clinical and operational scenarios: An overview. *Western Journal of Emergency Medicine*, 20(2):219–227.

[Lenatti et al. 2025] Lenatti, M., Carlevaro, A., Guergachi, A., Keshavjee, K., Mongelli, M., and Paglialonga, A. (2025). Estimation and conformity evaluation of multi-class counterfactual explanations for chronic disease prevention. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6132–6142.

[Mothilal et al. 2020] Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conf. on Fairness, Accountability, and Transparency*, pages 607–617.

[Nohara et al. 2022] Nohara, Y., Matsumoto, K., Soejima, H., and Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214:106584.

[Pfeifer et al. 2025] Pfeifer, B., Krzyzinski, M., Baniecki, H., Holzinger, A., and Biecek, P. (2025). Explaining and visualizing black-box models through counterfactual paths. *Pattern Analysis and Applications*, 28(154):1–19.

[Rasouli and Yu 2021] Rasouli, P. and Yu, I. C. (2021). Analyzing and improving the robustness of tabular classifiers using counterfactual explanations. In *Proceedings of the 2021 20th IEEE International Conf on ML and Applications*, pages 1–8. IEEE.

[Saranti et al. 2022] Saranti, A., Hudec, M., Minarikova, E., et al. (2022). Actionable explainable AI (AXAI): A practical example with aggregation functions for adaptive classification and textual explanations for interpretable machine learning. *Machine Learning and Knowledge Extraction*, 4(4):924–953.

[Ustun et al. 2019] Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 10–19. ACM.

[Verma et al. 2024] Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., and Shah, C. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(312):1–42.

[Wachter et al. 2017] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841.