UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Rio Claro

Instituto de Geociências e Ciências Exatas

Gabriel Covello Furlanetto

CausalBioCF: Contrafactuais Causais para Interpretabilidade de Modelos de Aprendizado de Máquina

Rio Claro-SP

2026

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Gabriel Covello Furlanetto

CausalBioCF: Contrafactuais Causais para Interpretabilidade de Modelos de Aprendizado de Máquina

Tese de Doutorado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Doutor em Ciência da Computação.

Orientador: Prof. Dr. Alexandro José Baldassin

Rio Claro – SP
2026

F985c

Furlanetto, Gabriel Covello
CausalBioCF: contrafactuais causais para interpretabilidade de modelos de aprendizado de máquina / Gabriel Covello Furlanetto. -- Rio Claro, 2026
88 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Alexandro José Baldassin

1. Contrafactuais. 2. Causalidade. 3. Aprendizado do máquina. 4. Conhecimento de domínio. 5. Acionabilidade. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

Esta pesquisa é uma contribuição na área de interpretabilidade de algoritmos de aprendizado de máquina. Por meio da integração entre análise de causalidade, geradores de contrafactuais e a proposição de uma metodologia para verificação e mensuração da acionabilidade de contrafactuais, o trabalho possibilita maior transparência e justiça em decisões automatizadas. Além disso, a pesquisa é aplicável a áreas multidisciplinares, como finanças, principalmente na análise de crédito e inadimplência, pesquisas relacionadas à saúde, marketing, compondo a recomendação de produtos e serviços públicos, com potencial de parcerias com órgãos e empresas.

POTENTIAL IMPACT OF THIS RESEARCH

This research contributes to the field of interpretability of machine learning algorithms. By integrating causality analysis, counterfactual generators and proposing a methodology to verify and measure the actionability of counterfactuals, the work enables greater transparency and fairness in automated decisions. Furthermore, research is applicable to multidisciplinary fields, such as finance, particularly in credit and default analysis, healthcare-related research, marketing, and recommendations for public products and services, with potential for partnerships with agencies and companies.

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Gabriel Covello Furlanetto

CausalBioCF: Contrafactuais Causais para Interpretabilidade de Modelos de Aprendizado de Máquina

Tese de Doutorado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Doutor em Ciência da Computação.

Comissão examinadora

Prof. Dr. Alexandro José Baldassin
IGCE/UNESP/Rio Claro (SP)
Orientador

Prof. Dr. Fabio Gagliardi Cozman
Escola Politécnica da Universidade de São Paulo/USP/São Paulo (SP)

Prof. Dr. Marcos Medeiros Raimundo
Instituto de Computação da Universidade de Campinas/UNICAMP/Campinas (SP)

Prof. Dr. João Paulo Papa
FC/UNESP/ Bauru (SP)

Prof. Dr. Arnaldo Cândido Júnior
IBILCE/UNESP/São José do Rio Preto (SP)

Conceito: Aprovado.

Rio Claro-SP, 13 de janeiro de 2026.

Dedico este trabalho aos meus pais Ana e Rogerio e a minha esposa Jéssica

Agradecimentos

Neste momento, gostaria de agradecer a Deus e a todas as pessoas que sempre me deram e me dão forças para a continuidade de meu trabalho.

Agradeço aos meus pais, Ana e Rogerio, que sempre incentivaram meus estudos, principalmente à minha mãe, que sugeriu o início do doutorado em um momento difícil, de pandemia, que passávamos. À minha esposa, Jéssica, que enfrentou comigo esta jornada e ao nosso primeiro filho Miguel, que estará em breve conosco, mas já inspira este esforço.

Agradeço aos meus orientadores, Alexandro, que desde a conclusão do mestrado, já citava a possibilidade de continuar com os estudos e topou fazer parte desta nova fase e Aleardo, que já faz parte de minha vida acadêmica/pessoal há mais de 15 anos. Não posso deixar de citar a professora Renata, que apesar de não estar tão próxima ao longo do doutorado, foi uma grande incentivadora e uma inspiração para que eu iniciasse o trabalho como pesquisador.

Agradeço à Daniela e ao Denys, que não mediram esforços para que eu pudesse conciliar a jornada de estudos com a de trabalho, ingressando no programa. A Beatriz, Ana e Thiago que sempre apoiaram a participação em atividades de pesquisa como forma de aprendizado e em especial ao Rafhael, que contribuiu em muitas discussões de conceitos estatísticos.

Agradeço aos amigos do IBILCE que me ajudaram ao longo desta jornada, Vitória, Diego e Fernanda.

Agradeço aos professores Fabrício e Verônica, da Unesp de Rio Claro, que tive a oportunidade de conhecer e trabalhar ao longo das disciplinas.

Por fim, mas não menos importante, agradeço aos demais familiares que sempre estiveram me apoiando.

"Tenha sempre muita força, muita determinação e sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá."

- Ayrton Senna.

Resumo

À medida que a inteligência artificial se torna comum aos processos de tomada de decisão, explicar o comportamento do modelo é essencial para promover a confiança, detectar vieses e garantir a conformidade com as regulamentações de proteção de dados, como a LGPD do Brasil e o GDPR da União Europeia. Entre as técnicas de interpretabilidade, as explicações contrafactuais oferecem esclarecimentos intuitivos sobre as decisões do modelo, ilustrando as mudanças mínimas necessárias para alterar um resultado. No entanto, as abordagens existentes frequentemente falham em gerar contrafactuais acionáveis que respeitem as restrições causais, lógicas e físicas, e carecem de meios confiáveis para avaliar a qualidade das explicações geradas. Este trabalho contribui para a explicabilidade contrafactual ao introduzir uma abordagem orientada por diretrizes de causalidade que restringem e direcionam o processo de geração, resultando em explicações mais realistas e viáveis. Além disso, propõe um método de avaliação capaz de medir sistematicamente a acionabilidade e a utilidade prática dessas explicações, uma dimensão frequentemente negligenciada na literatura. Essas contribuições são materializadas no CausalBioCF, um método que integra abordagens de avaliação causal a um processo de otimização bioinspirado baseado em Algoritmo Genético, no qual o conhecimento de domínio é incorporado na forma de restrições, permitindo a geração de contrafactuais válidos e efetivamente acionáveis. Experimentos realizados sobre múltiplos conjuntos de dados tabulares, abrangendo domínios como crédito, marketing bancário, justiça criminal e análise de comportamento do usuário, mostram que o CausalBioCF produz contrafactuais que são causalmente fundamentados e praticamente acionáveis, alcançando desempenho competitivo ou superior em comparação com métodos estabelecidos como DiCE, NICE e CFNOW em vários critérios, incluindo acionabilidade, esparsidade, proximidade, validade e tempo de execução. O trabalho também evidencia que adicionar informações causais a outros algoritmos, como o DiCE, pode melhorar significativamente a qualidade e a utilidade dos contrafactuais que eles produzem. Observa-se um aumento de 71,4% na proporção de contrafactuais acionáveis produzidos pelo CausalBioCF em comparação ao DiCE, representando o maior ganho estatisticamente significativo entre os métodos avaliados. Em relação ao NICE, o sistema apresenta uma melhoria de 15,2%. Além disso, a incorporação de conhecimento causal ao DiCE resulta em um incremento de 39,1% na geração de contrafactuais acionáveis quando comparado à versão original do algoritmo.

Palavras-chaves: contrafactuais; causalidade; aprendizado de máquina; conhecimento de domínio; acionabilidade.

Abstract

As artificial intelligence becomes increasingly embedded in decision-making processes, explaining model behavior is essential to foster trust, detect bias, and ensure compliance with data protection regulations such as Brazil's LGPD and EU's GDPR. Among interpretability techniques, counterfactual explanations offer intuitive insights into model decisions by illustrating the minimal changes needed to alter an outcome. However, existing approaches often fail to generate actionable counterfactuals that respect causal, logical, and physical constraints, and lack reliable means to evaluate the quality of the generated explanations. This work contributes to counterfactual explainability by introducing an approach guided by principles of causality that constrains and directs the generation process, resulting in more realistic and viable explanations. In addition, it proposes an evaluation method capable of systematically measuring the actionability and practical utility of these explanations, a dimension that has been frequently overlooked in the literature. These contributions are materialized in CausalBioCF, a method that integrates causal evaluation approaches into a bioinspired optimization process based on a Genetic Algorithm, in which domain knowledge is incorporated in the form of constraints, enabling the generation of valid and effectively actionable counterfactuals. Experiments conducted on multiple tabular datasets, covering domains such as credit scoring, bank marketing, criminal justice, and user behavior analysis, show that CausalBioCF produces counterfactuals that are both causally grounded and practically actionable, achieving competitive or superior performance compared to established methods such as DiCE, NICE, and CFNOW across multiple criteria, including actionability, sparsity, proximity, validity, and execution time. The work also shows that adding causal information to systems like DiCE can significantly enhance the quality and usefulness of the counterfactuals they produce. We observe a 71.4% increase in the proportion of actionable counterfactuals produced by CausalBioCF compared to DiCE, representing the largest statistically significant improvement among the evaluated methods. Compared to NICE, the system achieves an improvement of 15.2%. Furthermore, incorporating causal knowledge into DiCE results in a 39.1% increase in the generation of actionable counterfactuals compared to the original algorithm.

Keywords: counterfactuals; causality; machine learning; knowledge domain; actionability.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AED	Análise Exploratória de Dados
ATE	Average Treatment Effect
AG	Algoritmo Genético
CF	Contrafactuais
DAG	Directed Acyclic Graph
IA	Inteligência Artificial
IPW	Inverse Propensity Weighting
PSM	Propensity Score Matching
RF	Random Forest
XGBoost	Extreme Gradient Boosting
KNN	K-Nearest Neighbors
MLP	Multilayer Perceptron
DiCE	***Di**verse **C**ounterfactual **E**xplanations*
NICE	***N**earest **I**nstance **C**ounterfactual **E**xplanations*
CFNOW	*Counterfactuals Now*
FDR	*False Discovery Rate*
XAI	*e**X**planaible **A**rtificial **I**ntelligence*

Lista de símbolos

α	Letra Grega Alfa
ε	Letra Grega Epsilon

Sumário

1 INTRODUÇÃO

Algoritmos de aprendizado de máquina são cada vez mais utilizados em diversos contextos de tomadas de decisões (JIN, 2020). Suas aplicações encontram-se tanto em cenários cujo risco dessas decisões para os usuários são pequenos, como na indicação de conteúdo de mídia (GOMEZ-URIBE; HUNT, 2015) ou de produtos para completar o carrinho de compras (SHANI; GUNAWARDANA, 2011), como também em cenários de elevado risco, como na área médica, para o diagnóstico de doenças (FATIMA; PASHA et al., 2017), e na área financeira, auxiliando na detecção de fraudes (SULAIMAN; SCHETININ; SANT, 2022) e na liberação de crédito financeiro (AUQUI et al., 2022).

Deste modo, é importante que as escolhas embasadas pelos dados gerados por esses sistemas sejam confiáveis. Assim, torna-se essencial a compreensão de como o sistema está fazendo suas previsões, requerendo-se transparência nos algoritmos utilizados. Esta situação é ainda mais crítica devido ao surgimento de políticas regulatórias que exigem o fornecimento de ao menos uma noção do motivo da tomada de decisão, a fim de criar confiança na tecnologia (GUIDOTTI, 2024).

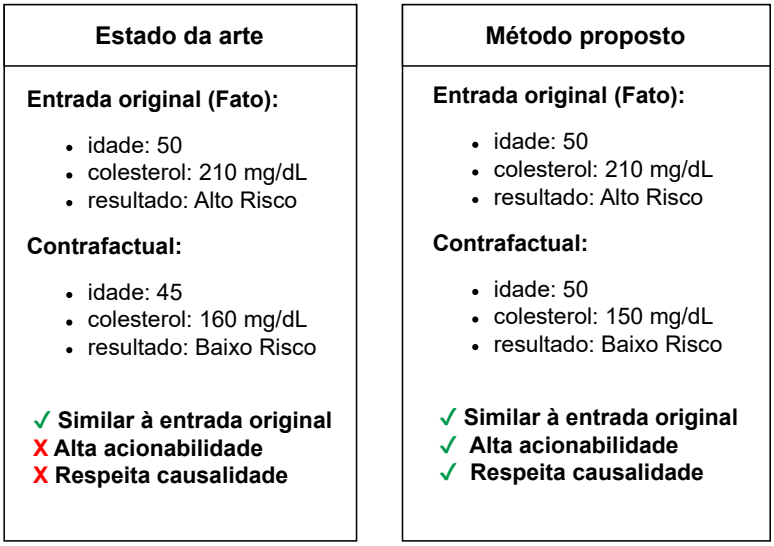
Neste contexto, surge a importância do conceito de explicabilidade contrafatual. Segundo Wachter, Mittelstadt e Russell (2017), explicabilidade contrafactual é um tipo de explicação utilizada em inteligência artificial e aprendizado de máquina que objetiva fornecer percepções a respeito da forma com que uma decisão automatizada foi tomada. Este tipo de metodologia trabalha sobre um cenário hipotético, no qual a escolha seria diferente caso certas variáveis de entrada para o modelo tivessem valores diferentes. Assim, ao comparar a decisão real com a hipotética, os contrafactuais podem auxiliar os usuários a entender por que um resultado específico foi alcançado e quais fatores foram mais importantes para esta conclusão.

Para fins ilustrativos, considere um modelo que classifica pacientes quanto ao risco cardiovascular com base nos níveis de colesterol. Em um sistema de apoio à decisão para diagnóstico médico, explicações contrafactuais permitem identificar quais atributos foram decisivos para a predição realizada. Nesse contexto, uma explicação pode assumir a forma: “caso o paciente não apresentasse o sintoma X, a predição do modelo seria negativa (contrafactual), em vez de positiva (fato)” (LOOVEREN; KLAISE, 2021).

1.1 Motivação

Após o trabalho de Wachter, Mittelstadt e Russell (2017), diversos métodos foram propostos para a geração de contrafactuais e estão sendo gradativamente utilizados para auxiliar decisões em domínios sensíveis, como concessão de crédito, avaliação de saúde e de riscos

Figura 1 – Comparação entre métodos tradicionais de geração de contrafactuais e a abordagem proposta.



Fonte: Elaborada pelo autor.

financeiros. Entretanto, as abordagens existentes frequentemente geram contrafactuais pouco realistas ou inviáveis, na prática (YANG et al., 2021). Para solucionar este problema, a literatura evidencia que muitos métodos dependem de conhecimento de domínio inserido manualmente, não capturam relações causais entre variáveis e podem gerar contrafactuais inconsistentes com o mundo real (MOTHILAL; SHARMA; TAN, 2020; VERMA et al., 2024).

A Figura 1 ilustra, de maneira comparativa, as limitações dos métodos tradicionais de geração de contrafactuais e as vantagens da abordagem proposta. No exemplo apresentado, os métodos do estado da arte modificam simultaneamente atributos que possuem dependências causais (idade e colesterol), produzindo uma explicação contrafactual não factível. A redução da idade de um paciente, de 50 para 45 anos, não é uma ação realizável, caracterizando baixa acionabilidade. Além disso, alterar idade e colesterol ao mesmo tempo, ignora a relação causal existente entre esses atributos, levando o modelo a propor cenários artificialmente possíveis, obtendo em troca ganhos em similaridade entre factual e contrafactual. Já a abordagem proposta gera alternativas plausíveis e coerentes com as relações entre variáveis. O contrafactual gerado permanece próximo à instância original, é causalmente plausível e representa uma ação executável pelo usuário (redução do colesterol sem alteração de idade), resultando em um contrafactual acionável.

Nesse contexto, este trabalho busca reduzir a dependência da intervenção manual ao extrair automaticamente restrições a partir dos dados e incorporá-las em métodos de geração já estabelecidos. Além disso, propõe-se uma métrica para avaliar a qualidade e a utilidade prática dos contrafactuais obtidos, permitindo mensurar sua aderência às restrições do mundo real, ou seja, a sua acionabilidade. Essa métrica possibilita mensurar em que proporção os contrafactuais

representam mudanças coerentes com as restrições reais, preenchendo uma lacuna importante nas abordagens atuais de avaliação, que não possuem uma metodologia clara.

1.2 Objetivos

O objetivo central deste trabalho é aprimorar a obtenção de contrafactuais por meio da integração implícita de dependências causais e conhecimento de domínio, de modo a produzir soluções mais acionáveis, plausíveis e coerentes com o contexto da aplicação. Desta forma, este trabalho propõe o CausalBioCF, um método que articula, de maneira inédita, técnicas existentes na literatura para derivar restrições de conhecimento de domínio na geração de contrafactuais por meio de um Algoritmo Genético (AG) (GOLBERG, 1989; HOLLAND, 1992). Adaptada para dados tabulares, essa abordagem garante que os contrafactuais resultantes sejam relevantes e aplicáveis em conjuntos de dados estruturados. O CasualBioCF emprega análise estatística para identificar distribuições de dados, intervalos válidos e potenciais relações causais entre variáveis.

Um algoritmo bioinspirado é adotado devido à sua capacidade de produzir múltiplos contrafactuais por execução e à redução da sobrecarga computacional em comparação com abordagens comumente utilizadas na geração de contrafactuais, como *autoencoders* (NEMIROVSKY et al., 2022; ZHANG; BARR; PAISLEY, 2022). Além disso, a fim de mensurar a qualidade dos contrafactuais gerados e compará-los com aqueles gerados por outros algoritmos, o trabalho propõe uma metodologia de cálculo padronizada para a métrica de acionabilidade. Esta metodologia incorpora indicadores qualitativos, aspecto negligenciado em trabalhos correlatos.

Em resumo, as principais contribuições deste trabalho são:

- Uma abordagem guiada pela causalidade para gerar explicações contrafactuais usando um algoritmo genético, que incorpora conhecimento do domínio derivado de análises estatísticas e causais, melhorando a viabilidade e reduzindo a necessidade de intervenção humana;
- Uma métrica padronizada para avaliar a aplicabilidade de contrafactuais, permitindo avaliações mais rigorosas, comparáveis e quantificáveis da qualidade dos contrafactuais entre diferentes métodos;
- Uma avaliação experimental detalhada mostrando que o algoritmo proposto tem um desempenho particularmente bom em cenários que envolvem conjuntos de dados com um grande número de características e/ou atributos predominantemente categóricos, em que a integração da inferência causal desempenha um papel central na orientação do processo de busca. Ao aproveitar a causalidade para definir quais características podem levar a um contrafactual melhor, o método é capaz de gerar contrafactuais que são significativamente acionáveis.

1.3 Publicações

Este texto organiza e consolida os resultados publicados nas seguintes conferências:

- Furlanetto, G. C., Baldassin, A., & Manacero, A. (2024, Maio). CausalBioCF: Causalidade e otimização bioinspirada para geração de contrafactuais factíveis em tempo real. In Escola Regional de Alto Desempenho de São Paulo (ERAD-SP) (pp. 77-80). SBC. (Trabalho que recebeu menção honrosa por seu destaque no evento)
- Furlanetto, G. C., Baldassin, A., & Manacero, A. (2025, June). CausalBioCF: Causal Counterfactuals for Machine Learning Interpretability. In International Conference on Computational Science and Its Applications (pp. 200-217). Cham: Springer Nature Switzerland.

Também está em avaliação uma submissão realizada em um periódico, como segue:

- Furlanetto, G. C., Baldassin, A., & Manacero, A. (2025). Causality-Guided Generation of Actionable Counterfactual Explanations for Machine Learning Models. Manuscrito em revisão na revista *Data Mining and Knowledge Discovery* (Springer).

1.4 Organização do Texto

A estrutura deste trabalho foi organizada para conduzir o leitor gradualmente desde os conceitos fundamentais até as contribuições e resultados obtidos. Ela está organizada em seis capítulos, sendo este o primeiro deles.

O Capítulo 2 apresenta a fundamentação teórica necessária para o desenvolvimento da pesquisa, incluindo conceitos relacionados a interpretabilidade, explicação contrafactual e técnicas de inferência causal.

O Capítulo 3 sintetiza o estado da arte sobre a geração de contrafactuais, apresentando uma revisão abrangente da literatura e destacando as principais abordagens existentes.

O Capítulo 4 detalha as principais contribuições do trabalho: a geração de contrafactuais guiada por relações causais, e a técnica desenvolvida para computação da acionabilidade.

O Capítulo 5 apresenta os testes realizados e as validações conduzidas, com o objetivo de demonstrar e avaliar as contribuições deste estudo.

Por fim, o Capítulo 6 reúne as conclusões alcançadas, discutindo as implicações dos resultados obtidos e sugerindo possíveis direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos fundamentais que sustentam o desenvolvimento deste trabalho. Cada conjunto de conceitos é introduzido de forma a motivar seu papel no contexto da explicabilidade contrafactual e na construção do método proposto.

Inicialmente, são discutidos conceitos de interpretabilidade e explicabilidade aplicados a modelos de aprendizado de máquina, com ênfase em métodos de explicações contrafactuais. Essa fundamentação é necessária para situar o leitor quanto às características, limitações e objetivos das explicações produzidas por abordagens contrafactuais, que constituem o foco central deste estudo.

Em seguida, são apresentadas noções estatísticas relevantes para a análise exploratória dos dados. Essa etapa é fundamental para compreender a estrutura das variáveis, identificar padrões e extrair conhecimento preliminar do domínio, orientando tanto a definição de restrições quanto a avaliação da plausibilidade das explicações produzidas. Juntamente aos conceitos de análise exploratória de dados, são introduzidos os principais elementos da inferência causal. A discussão sobre causalidade é essencial, pois grande parte das limitações dos métodos tradicionais de contrafactuais decorre da incapacidade de distinguir relações estruturais entre variáveis de meras associações observadas. Assim, a incorporação de conhecimento causal fundamenta as contribuições do método proposto neste trabalho.

2.1 Explicabilidade x Interpretabilidade de Modelos

Explicar ao usuário o que levou um modelo de aprendizado de máquina a uma tomada de decisão é cada vez mais importante no cenário atual. Com a popularização do termo “Inteligência Artificial” entre leigos, este tipo de interpretação de resultados pode garantir transparência e justiça sobre as decisões tomadas. No entanto, trazer este tipo de interpretação de resultados aos usuários comuns nem sempre é fácil. Muitas vezes, eles não possuem conhecimento suficiente para interpretar informações técnicas como a importância de atributos para um modelo, por exemplo.

Desta forma, ao tratar-se de compreensão de modelos de aprendizado de máquina e, mais amplamente, de inteligência artificial, sempre são apresentadas duas vertentes: uma com modelos inerentemente interpretáveis (modelos lineares, conjuntos de regras, árvores de decisão, entre outros) e outra de explicação de modelos complexos (*random forest*, redes neurais, entre outros) (RUDIN, 2019). Nestas vertentes, observam-se autores que consideram explicabilidade e interpretabilidade como conceitos diferentes (MILLER, 2019) e outros que os consideram como semelhantes (ADADI; BERRADA, 2018).

O primeiro grupo define interpretabilidade como o nível de compreensão que um ser humano pode ter do motivo de uma decisão, estando ela ligada à explicação por trás das saídas de um modelo, mas não de seu funcionamento. Estes mesmos autores definem a explicabilidade como sendo algo relacionado à compreensão da mecânica interna de um sistema de aprendizado de máquina. Quanto mais explicável é um modelo, mais profundo é a compreensão humana do que está ocorrendo ao longo da execução do algoritmo, enquanto o modelo de aprendizado está sendo treinado e executado (MILLER, 2019).

Neste trabalho utiliza-se ambos os conceitos de forma similar, pois se acredita que tanto o funcionamento do algoritmo quanto a compreensão da decisão apresentada por ele são fundamentais para o entendimento do resultado apresentado pelo modelo de aprendizado de máquina.

2.2 Interpretabilidade Global e Local

Ainda no contexto de compreensão de algoritmos de aprendizado de máquina, outros conceitos importantes referem-se à interpretabilidade global e local. A interpretabilidade global relaciona-se à compreensão da forma como o modelo toma decisões a fim de chegar em seus resultados. Neste caso, isso é feito a partir de uma visão completa dos dados e de cada um dos componentes aprendidos pelo modelo, dentre os quais estão, por exemplo, os pesos de variáveis e os parâmetros de entrada. A fim de simplificar a definição, pode-se dizer ainda que, como interpretabilidade global do modelo, entende-se a maneira como o modelo treinado faz previsões dentro do todo, ou a forma com que a lógica dele o leva a todos os diferentes resultados possíveis (DU; LIU; HU, 2019). Obter explicações globais é uma tarefa difícil de ser conseguida (ADADI; BERRADA, 2018).

Por outro lado, a interpretabilidade local relaciona-se à explicação de uma predição de maneira isolada e individual (DU; LIU; HU, 2019). Embora este tipo de interpretabilidade não forneça uma solução ótima, ela consiste em uma aproximação boa para obter-se explicação dos resultados obtidos. Isto ocorre pois, localmente, a previsão pode depender apenas linearmente das variáveis de entrada (ou atributos), ao invés de ter uma dependência complexa deles. Muitas vezes explicações locais podem ser mais precisas do que explicações globais (ADADI; BERRADA, 2018).

Este trabalho adota explicações locais (contrafactuais), em detrimento de explicações globais. O foco em explicações contrafactuais permite analisar decisões individuais do modelo, fornecendo informações ao usuário final diretamente relacionadas a instâncias específicas e às mudanças necessárias para a obtenção de resultados alternativos.

2.3 Contrafactuais

Utilizados no passado em áreas não tão próximas à tecnologia, como, por exemplo, na psicologia e na filosofia (BYRNE, 2019; KAHNEMAN; MILLER, 1986), explicações contrafactuais possuem uma longa história. Segundo Lipton (1990), o objetivo de tais explicações é responder à pergunta "Por que P ao invés de Q?", em que P e Q seriam o fato (factual) e o contra fato (contrafactual), respectivamente. Além disso, estas explicações também possuem utilidade por serem geradas sem revelar informações confidenciais sobre como o sistema de decisões funciona. Wachter, Mittelstadt e Russell (2017) argumentam ainda que existem três finalidades principais para explicar/interpretar modelos de aprendizado de máquina:

- Informar e ajudar a pessoa que está utilizando o modelo, ou que está sendo afetada por ele, a entender o motivo de uma predição;
- Possibilitar que a decisão seja contestada caso resulte em uma consequência indesejável para uma das partes (beneficiado ou prejudicado pelo modelo);
- Compreender o que poderia ser alterado a fim de obter um resultado diferente do apontado pelo modelo.

Neste contexto, Wachter, Mittelstadt e Russell (2017) propõem o uso de explicabilidade contrafactual juntamente com modelos de aprendizado de máquina, trazendo uma perspectiva de que esse tipo de explicabilidade poderia tornar as decisões dos algoritmos mais compreensíveis aos humanos. A partir da definição destes autores, quando é apresentado um conjunto de dados de entrada, representado como uma tupla ou um ponto e um modelo de tomada de decisões, uma explicação contrafactual poderia ser definida como a menor perturbação ao conjunto de variáveis de entrada que gere um resultado diferente (contrafactual) do que foi gerado pelo ponto original (factual). Outros trabalhos subsequentes, como o de Sharma, Henderson e Ghosh (2020), expandem essa caracterização, reforçando o papel da proximidade e da mudança mínima como elementos centrais desse tipo de explicação.

Posteriormente, pesquisas ampliaram essa perspectiva ao identificar que a geração de contrafactuais ideal também deve utilizar instrumentos de *algorithmic recourse*, isto é, restrições que permitam alterar o resultado de um modelo de forma prática e acionável, gerando contrafactuais factíveis (KARIMI et al., 2020). Atualmente, os termos contrafactual e *algorithmic recourse* são utilizados para designar a mesma ideia na literatura, como sinônimos, uma vez que para manter factibilidade no mundo real, contrafactuais devem obedecer restrições de domínio (O'BRIEN; KIM; WEBER, 2023).

Trazendo esta definição para um contexto matemático, ainda segundo Wachter, Mittelstadt e Russell (2017), dado um modelo de decisão $f : X \rightarrow Y$ e uma instância de entrada

$x \in X$, com previsão $f(x) = y$, um contrafactual é uma nova instância x' que altera o resultado do modelo para uma classe alvo y', mantendo-se próxima da instância original:

$$f(x') = y' \neq f(x), \tag{2.1}$$

de modo que x' seja a solução do seguinte problema de otimização:

$$x' = \arg\min_{z \in X} D(x, z), \tag{2.2}$$

em que $D(\cdot, \cdot)$ representa uma medida de distância entre amostras. Restrições adicionais, frutos de abordagens de *algorithmic recourse*, podem ser aplicadas para garantir a viabilidade das soluções, como monotonicidade, causalidade ou restrições semânticas (SHARMA; HENDERSON; GHOSH, 2020; KARIMI et al., 2022).

O exemplo clássico deste cenário pode ser observado em um modelo de crédito. Supondo que um usuário de 40 anos, cujo grau de escolaridade é ensino médio completo e a renda mensal é de R$2.000,00, tenha sua solicitação de crédito negada por um modelo de aprendizado de máquina, uma explicação contrafactual poderia ser: "Se a renda mensal em vez de R$2.000,00 fosse R$4.000,00, o empréstimo seria aprovado". Desse modo, entregar contrafactuais para usuários de modelos de aprendizado de máquina pode ser uma forma de ajudá-lo a compreender quais fatores, ou quais atributos, foram determinantes para chegar àquela decisão. Além disso, é possível mostrar a eles quais valores estes atributos deveriam assumir para atingir um resultado diferente.

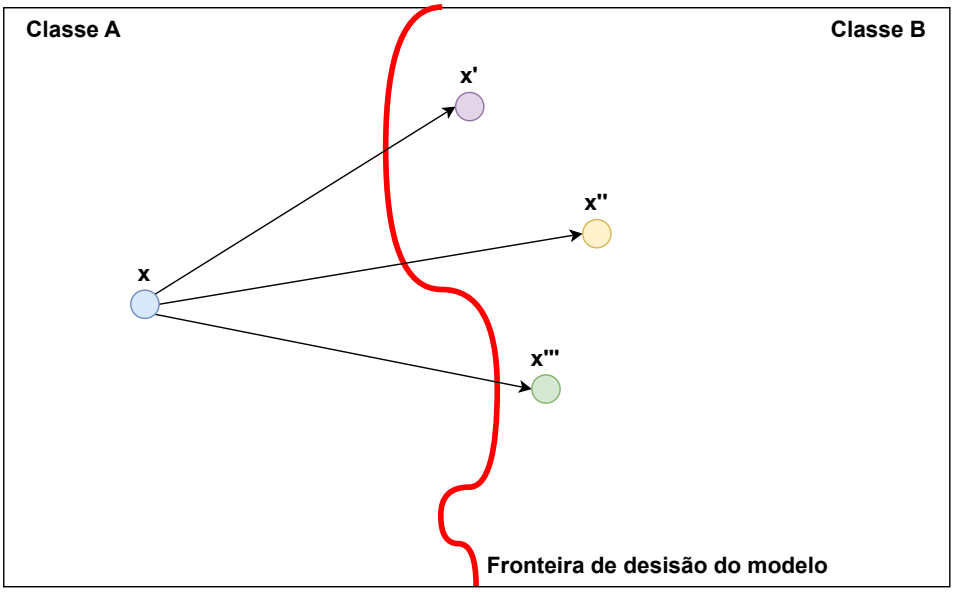
A Figura 2 ilustra o conceito geometricamente: dado um ponto original x, seus contrafactuais x', x'' e x''' são os pontos mais próximo possíveis, mas localizados do outro lado da fronteira de decisão do modelo. Juntamente com a proposição de Wachter *et al.*, diferentes métricas passaram a ser utilizadas para avaliar a qualidade dos contrafactuais gerados (WACHTER; MITTELSTADT; RUSSELL, 2017; KARIMI et al., 2022). Tais métricas são tipicamente divididas entre quantitativas e qualitativas, sendo detalhadas, respectivamente, nas Seções 2.3.1 e 2.3.2.

2.3.1 Métricas Quantitativas

Em geral, métricas quantitativas, como validade, esparsidade, tempo de execução e proximidade, são claramente definidas e amplamente padronizadas na literatura, permitindo comparações consistentes entre diferentes métodos. Os trabalhos na literatura definem normalmente as métricas da seguinte forma:

- **Distância ou Proximidade:** mede o quão próximos os contrafactuais estão dos pontos factuais. Para esse propósito, as distâncias Euclidiana (L2) (DANIELSSON, 1980) e

Figura 2 – Representação gráfica de uma explicação contrafacual genérica.



Fonte: Adaptado do trabalho de Mothilal, Sharma e Tan (2020).

Manhattan (L1) (BLACK, 1998) são geralmente empregadas (MOTHILAL; SHARMA; TAN, 2020; KURATOMI et al., 2022);

- **Validade:** avalia se um ponto gerado (contrafactual) pertence a uma classe diferente em comparação com o ponto original (factual) (GUIDOTTI, 2024; MOTHILAL; SHARMA; TAN, 2020; OLIVEIRA; MARTENS, 2021);
- **Tempo de execução:** considera o tempo necessário para gerar o contrafactual desejado (GUIDOTTI, 2024; KURATOMI et al., 2022);
- **Esparsidade:** quantifica o número de variáveis alteradas para gerar um conjunto de contrafactuais. Contrafactuais melhores têm menos alterações de variáveis (MOTHILAL; SHARMA; TAN, 2020; KURATOMI et al., 2022; BRUGHMANS; LEYMAN; MARTENS, 2024; OLIVEIRA; MARTENS, 2021).

2.3.2 Métricas Qualitativas

As métricas qualitativas, particularmente a acionabilidade, não possuem uma definição clara ou padronizada. Termos como acionabilidade, viabilidade e plausibilidade são frequentemente usados, às vezes com significados semelhantes, mas cada um com foco em ideias diferentes.

Plausibilidade refere-se ao quão realista é uma instância contrafactual em relação à distribuição dos dados observados. Contrafactuais plausíveis respeitam não apenas os intervalos

de valores individuais de cada atributo, mas também a estrutura conjunta da distribuição, de modo que a combinação de características seja consistente com exemplos reais presentes no conjunto de dados (GUIDOTTI, 2024).

Factibilidade ou viabilidade avalia se as mudanças sugeridas no contrafactual são lógica e fisicamente possíveis. Para esse conceito, alguns atributos são considerados imutáveis, enquanto outros podem mudar apenas em uma direção (por exemplo, a idade pode aumentar, mas não diminuir). Para ser viável, um contrafactual deve alterar apenas atributos mutáveis, e quaisquer alterações devem seguir restrições direcionais realistas (VERMA et al., 2024).

A acionabilidade concentra-se em saber se as alterações sugeridas pelo contrafactual podem ser aplicadas. Um contrafactual é acionável somente se todas as variáveis alteradas estiverem no conjunto de variáveis que o usuário pode realisticamente controlar. Por exemplo, alguém pode melhorar seu nível de escolaridade ou renda, mas não pode alterar sua data de aniversário ou seu tipo sanguíneo (GUIDOTTI, 2024). Um dos poucos trabalhos na literatura que tentam quantificar acionabilidade dos contrafactuais de maneira sistemática é o trabalho de Pascual-Triana et al. (2025), que faz isso por meio de marcação da realização de alterações ou não em atributos imutáveis. Porém, ainda assim não há uma explicação clara de como a metodologia é implementada.

Dessa forma, contrafactuais permitem explicar a tomada de decisão de modelos de aprendizado de máquina ao indicar quais atributos e valores precisariam ser modificados para um resultado alternativo ser obtido, fornecendo ao usuário final ou especialista de domínio uma orientação clara e direcionada sobre possíveis ações. Para avaliar a qualidade dessas explicações, diferentes métricas quantitativas e qualitativas têm sido propostas, devendo ser interpretadas de maneira complementar. Assim, os contrafactuais constituem uma ferramenta central na interpretabilidade de modelos de aprendizado de máquina, especialmente no apoio à compreensão e à tomada de decisão por parte de usuários humanos que detêm conhecimento do domínio dos dados, mas não necessariamente familiaridade com os mecanismos internos dos modelos, bem como na avaliação da capacidade desses modelos de fornecer explicações significativas e efetivamente aplicáveis.

2.4 Análise Estatística e Correlação

A compreensão das relações entre variáveis é um componente fundamental para o desenvolvimento deste trabalho. Antes de discutir causalidade e sua aplicação na geração de contrafactuais, é necessário reconhecer o papel da análise estatística descritiva e exploratória na construção do conhecimento de domínio. Em áreas nas quais o pesquisador não possui conhecimento prévio aprofundado, a Análise Exploratória de Dados (AED) (TUKEY et al., 1977) constitui uma etapa essencial para organizar, sintetizar e interpretar informações, permitindo identificar padrões, tendências e possíveis estruturas subjacentes aos dados.

A AED fornece subsídios iniciais para a formulação de hipóteses e preparação de análises posteriores, como a inferência estatística (MORETTIN; BUSSAB, 2017). Embora métricas descritivas, como medidas de tendência central e dispersão, não sejam objeto de aprofundamento neste trabalho, elas desempenharam papel instrumental na caracterização preliminar dos dados apresentados no Capítulo 4. Essa etapa, no entanto, possui limitações importantes, como a incapacidade de distinguir entre correlação e causalidade.

Essa distinção é fundamental. Relações observadas entre variáveis não representam, por si só, efeitos causais. Eventos podem ocorrer em conjunto por razões que não envolvem influência direta. Assim, correlação não implica causalidade, e confundir esses conceitos pode levar a conclusões equivocadas e decisões inadequadas (PEARL, 2009; HOLLAND, 1986). A dificuldade em diferenciar ambos manifesta-se tanto no senso comum quanto em análises técnicas. É frequente atribuir causalidade a coincidências. Exemplos cotidianos como "*Se eu tomar este medicamento, meu resfriado será curado?*" ou "*Está chovendo, então levarei uma blusa porque irá fazer frio*" ilustram como associações observadas podem ser interpretadas causalmente mesmo quando tal relação não existe.

De maneira informal, eventos A e B que ocorrem em conjunto não estabelecem, por si só, uma relação causal, mas sim uma correlação. Apenas quando a ocorrência de A altera a probabilidade de ocorrência de B é que se pode iniciar uma investigação sistemática de causalidade (PEARL, 2009; SILVA, 2021). A limitação existente nas análises puramente associativas motiva uma transição para o estudo da causalidade já que, para gerar contrafactuais acionáveis, não basta identificar padrões nos dados, é também necessário compreender como as variáveis de fato influenciam umas às outras. A seção seguinte aprofunda esse tema, apresentando os fundamentos conceituais e metodológicos da inferência causal.

2.5 Causalidade e Inferência Causal

A noção de causalidade desempenha papel fundamental na análise de relações entre variáveis, especialmente quando se busca compreender como mudanças em um fator podem produzir efeitos em outro. Segundo Nogueira et al. (2022), a causalidade pode ser definida como a influência pela qual um evento contribui para a produção de outros eventos. Assim, a causalidade busca explicar relações entre eventos distintos, dos quais fazem parte diferentes variáveis, indo além da simples observação de padrões ou associações.

Pearce e Lawlor (2016) também buscam definir causalidade de uma maneira simplificada. Segundo eles, em um caso em que uma variável X é uma causa de uma variável Y, Y depende de X para determinar seu valor, ou seja, a variável Y utiliza a informação fornecida por X para obter o valor que irá assumir. Essa definição sugere que a variável X desempenha um papel causal na determinação da variável Y que uma alteração em X pode resultar em uma mudança correspondente em Y, indicando uma relação de causa e efeito entre as variáveis.

Para este trabalho, o conceito de causalidade é particularmente relevante. Tanto modelos de aprendizado de máquina quanto métodos tradicionais de geração de contrafactuais baseiam-se predominantemente em padrões associativos presentes nos dados, não possuindo mecanismos internos para diferenciar correlação de dependência causal. Isso pode resultar em contrafactuais inviáveis, sugerindo alterações impossíveis, incoerentes ou incompatíveis com o domínio. A incorporação de conhecimento causal, seja proveniente de especialistas de domínio ou estimado a partir dos dados, é fundamental para orientar a geração de explicações que respeitem relações estruturais e representem possibilidades reais de ação (PEARL, 2009).

A investigação de causalidade organiza-se tipicamente em dois eixos (NOGUEIRA et al., 2022; PEARL, 2009; SILVA, 2021): a *descoberta causal*, que busca identificar relações estruturais diretamente de dados observacionais, e a *inferência causal*, que estima o impacto de intervenções específicas em variáveis de interesse. Esses conceitos são formalizados por meio de Modelos Causais Estruturais (*Structural Causal Models* – SCM) (PEARL, 2009) e Grafos Direcionados Acíclicos (*Direct Acyclic Graphs* – DAGs) (TENNANT et al., 2021), ferramentas que permitem distinguir dependências espúrias de relações de causa e efeito.

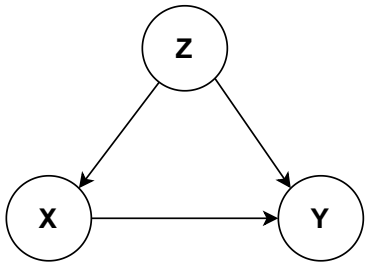
Por fim, embora definições conceituais de causalidade sejam importantes, sua utilidade prática depende de métodos capazes de avaliar relações causais em conjuntos de dados reais. A literatura em inferência causal oferece tais métodos, que serão apresentados a seguir, juntamente com suas características, limitações e a justificativa para a escolha da abordagem adotada neste trabalho.

2.5.1 Métodos de inferência causal

Métodos de inferência causal podem ser organizados, de forma geral, em dois grandes paradigmas. O primeiro, associado aos Modelos Causais Estruturais propostos por Pearl (PEARL, 2009), baseia-se na definição explícita de grafos causais e relações estruturais entre variáveis:

- **Redes Bayesianas Causais:** Podem ser vistas como uma instância particular dos Modelos Causais Estruturais, nas quais as relações causais são representadas exclusivamente por distribuições de probabilidade condicionais. São compostas por modelos gráficos de probabilidade a fim de apresentar dependência entre um conjunto de variáveis. Deste modo, elas são modeladas a partir de grafos acíclicos direcionados (*Directed Acyclic Graph* - DAG), em que os nós representam variáveis aleatórias e as arestas representam a relação entre elas. Cada um destes nós possui uma distribuição de probabilidade condicional associada. Elas são usadas com o objetivo de inferência causal em diversas áreas, como ciência da computação, estatística, biologia, medicina e economia (SPIRTES, 2010). Isso ocorre devido à sua capacidade de estruturar e representar informações objetivamente, permitindo uma representação gráfica de dados não estruturados e imprecisos a partir do conhecimento de especialistas. Característica que também acaba expondo algumas de suas

Figura 3 – Relação causal direta entre três variáveis. X exerce influência causal direta sobre Y, ou seja X pode alterar o valor de Y. Z representa um efeito confundidor na análise de inferência causal.



Fonte: Adaptado do trabalho de Pearl (2009).

desvantagens, como a construção de elementos gráficos de forma manual e subjetiva, já que para seu desenvolvimento é necessário conhecimento de um humano, referente ao campo de pesquisa que está sendo abordado (WIEGERINCK; KAPPEN; BURGERS, 2010);

- **Modelos Causais Estruturais:** são modelos que utilizam grafos, equações estruturais, lógica contrafactual e interventiva para mensurar a causalidade entre variáveis (PEARL, 2009). Embora os Modelos Causais Estruturais façam uso formal de contrafactuais, no sentido definido pela estrutura matemática de intervenções e mundos possíveis proposta por Pearl (2009), essa utilização não está relacionada à explicabilidade contrafactual aplicada a modelos de aprendizado de máquina. Neste contexto, o termo refere-se exclusivamente ao tema teórico da causalidade, que emprega contrafactuais para raciocinar sobre efeitos de intervenções e dependências causais entre variáveis. Trata-se, portanto, de um uso conceitual da mesma base matemática que fundamenta os métodos de explicabilidade, mas com finalidade distinta: analisar relações causais entre atributos, e não explicar decisões de modelos preditivos. Os Modelos Causais Estruturais são considerados um conjunto de variáveis endógenas e exógenas conectadas por um conjunto de funções que determinam os valores das variáveis endógenas com base nos valores das variáveis exógenas (PEARL, 2019). Cada modelo está associado a um grafo (Figura 3) no qual as funções correspondem a suposições causais, sendo que, se uma variável Y é filha de uma variável X, então dizemos que Y é causado por X, ou que X é a causa direta de Y; se a variável Y é descendente de uma variável X, então dizemos que Y é potencialmente causado por X, ou que X é a causa potencial de Y (PEARL, 2009).

O segundo paradigma, conhecido como estrutura de resultados potenciais (*potential outcomes*) (RUBIN, 1974), adota uma perspectiva estatística, na qual efeitos causais são estimados a partir da comparação entre grupos de tratamento e controle:

- **Ponderação de Probabilidade Inversa:** Do inglês *Inverse Probability Weighting* (IPW),

é um método que propõe balancear grupos de indivíduos, atribuindo a cada um deles um peso diferente, de forma que a distribuição ponderada das características em cada grupo seja semelhante (ROSENBAUM; RUBIN, 1983). Esta técnica utiliza uma pontuação de propensão que indica a probabilidade de um indivíduo pertencer a um grupo de tratamento A, dadas suas características representadas por X ($P(A \mid X)$). Essa probabilidade pode ser estimada por meio de um modelo de aprendizado de máquina. No entanto, para obter o efeito médio de tratamento, ATE (*Average Treatment Effect*), é necessária a definição de quais fatores confundidores podem afetar o problema. Tal definição é algo bastante subjetivo (SEAMAN; WHITE, 2013);

- **Pareamento por pontuação de propensão:** O *Propensity Score Matching* (PSM) é um método usado para reduzir o viés de seleção e estimar o efeito causal de um tratamento (ROSENBAUM; RUBIN, 1983; CALIENDO; KOPEINIG, 2008). Quatro etapas principais são seguidas para avaliar a causalidade. Primeiro, a pontuação de propensão é estimada para cada indivíduo usando um modelo de regressão logística. Os indivíduos são então pareados em grupos de tratamento e controle com base na similaridade de pontuações, usando técnicas como *Nearest-Neighbor Matching* (COVER; HART, 1967) ou *Caliper Matching* (ROSENBAUM; RUBIN, 1985). A avaliação das pontuações garante equilíbrio que ambos os grupos sejam comparáveis em termos de características observáveis. Finalmente, o efeito médio do tratamento é estimado comparando os resultados entre os grupos pareados (CALIENDO; KOPEINIG, 2008; ROSENBAUM; RUBIN, 1983). Entre as vantagens do PSM estão a redução do viés de seleção e sua flexibilidade, permitindo seu uso em diferentes contextos de pesquisa.

Este trabalho adota o segundo paradigma. Métodos baseados em grafos causais, como Redes Bayesianas Causais e Modelos Causais Estruturais também foram discutidos apenas informativamente e não são utilizados, principalmente devido à necessidade de especificações manuais de estruturas causais, contrariando o objetivo de automatização da abordagem proposta.

Dentre os métodos de resultados potenciais, o PSM foi escolhido devido à sua simplicidade de implementação, à possibilidade de automatizar a análise causal e ao desempenho correspondente em tempo de execução, uma vez que busca gerar grupos de tratamento e controle aproximadamente equivalentes, reduzindo o viés associado a fatores de confusão observáveis para calcular o ATE e, consequentemente, a causalidade entre as variáveis. Embora seja reconhecido que o modelo possui limitações, como controlar apenas variáveis observáveis e influenciar a estimativa da pontuação de propensão, entende-se que, no cenário de aplicação pretendido, essas limitações não são tão impactantes em comparação com o benefício de se obter uma análise causal independente do conhecimento fornecido por humanos.

2.6 Considerações Finais

Neste capítulo foram apresentados os conceitos teóricos fundamentais para o desenvolvimento deste trabalho. Inicialmente, foram discutidos métodos de interpretabilidade aplicados à compreensão de algoritmos de aprendizado de máquina e em especial explicabilidade baseada em contrafactuais. Em seguida, foram introduzidas definições estatísticas relacionadas à análise exploratória de dados, cuja aplicação ao longo deste trabalho pretende gerar restrições que viabilizem a criação de contrafactuais.

Também foram abordados conceitos referentes ao relacionamento entre variáveis em conjuntos de dados, com ênfase na investigação de relações causais e nas técnicas estatísticas de inferência que possibilitam essa análise. Com a base teórica estabelecida a fim de compreender as escolhas metodológicas adotadas ao longo da pesquisa, o Capítulo 3 apresenta a revisão de literatura conduzida, para mapear o estado da arte em pesquisas relacionadas à explicabilidade contrafactual e contextualizar as contribuições deste trabalho nesse cenário.

3 TRABALHOS RELACIONADOS

Neste capítulo é apresentado o estado da arte relacionado à geração de contrafactuais. Além das técnicas de geração, são examinadas as métricas utilizadas para avaliar contrafactuais e a relação complementar de indicadores quantitativos e qualitativos, porém evidenciando-se a ausência de uma metodologia qualitativa consolidada para mensurar realismo e viabilidade prática das explicações geradas. Essa análise permite identificar as limitações e lacunas presentes na literatura e estabelece as bases conceituais que motivam e justificam as contribuições desta pesquisa.

3.1 Fundamentação e Estado da Arte

Nos últimos anos, progressos significativos foram feitos no desenvolvimento de métodos para interpretar previsões de aprendizado de máquina, com explicações contrafactuais emergindo como uma abordagem particularmente intuitiva e centrada no ser humano (WACHTER; MITTELSTADT; RUSSELL, 2017). A pesquisa nessa área evoluiu em duas direções complementares que são diretamente relevantes para este trabalho: a geração e a avaliação de contrafactuais. A primeira direção é a geração de explicações contrafactuais, em que várias abordagens foram propostas, incluindo métodos baseados em otimização, em perturbação de pontos, em abordagens generativas e, por fim, em análise de causalidade. Cada uma delas emprega técnicas diferentes para produzir contrafactuais com propriedades desejáveis, como proximidade, esparsidade, diversidade e factibilidade (GUIDOTTI, 2024; FATIMA; PASHA et al., 2017; SULAIMAN; SCHETININ; SANT, 2022). Entre elas, as abordagens baseadas em causalidade são particularmente notáveis por buscarem garantir que os contrafactuais gerados respeitem a estrutura causal subjacente dos dados, aumentando assim sua viabilidade e aplicabilidade no mundo real (SHAO et al., 2023).

A segunda linha de pesquisa, que aparece até mesmo como foco secundário em muitos trabalhos, diz respeito à avaliação de explicações contrafactuais, particularmente à definição de métricas capazes de mensurar sua qualidade. Dentre elas, embora exista um corpo crescente de estudos voltados à avaliação quantitativa dessas explicações, com métricas amplamente adotadas, como proximidade, esparsidade, tempo de execução e validade (VERMA et al., 2024; GUIDOTTI, 2024), a dimensão qualitativa permanece substancialmente menos desenvolvida. Especificamente, aspectos relacionados à utilidade prática das explicações, como sua interpretabilidade humana e, sobretudo, sua viabilidade ou acionabilidade, ainda carecem de definições consolidadas e métodos de mensuração sistemáticos, fator este evidenciado pela falta de trabalhos publicados na literatura.

Essa lacuna motiva uma das contribuições deste trabalho: a definição de uma metodologia

para calcular se uma explicação contrafactual é acionável ou não, e uma métrica escalar para definir o quão acionável ela é. Tal abordagem busca complementar as métricas quantitativas tradicionais, incorporando um componente qualitativo essencial para a adoção prática de explicações contrafactuais. No restante desta seção, são discutidos esses dois aspectos com mais detalhes, apresentando os trabalhos relacionados mais relevantes.

3.1.1 Métodos Para Geração de Contrafactuais

Diversos métodos têm sido propostos na literatura para a geração de explicações contrafactuais, podendo estes ser organizados em algumas categorias principais. Dentre elas estão:

- Abordagens baseadas em otimização, que realizam a geração do contrafactual ao resolver um problema de minimização (MOTHILAL; SHARMA; TAN, 2020; de Oliveira; SöRENSEN; MARTENS, 2023);
- Métodos baseados em modelos generativos, que realizam seu aprendizado a partir de distribuições dos dados com a finalidade de produzir contrafactuais realistas (PAWELCZYK; BROELEMANN; KASNECI, 2020b; NEMIROVSKY et al., 2022);
- Heurísticas e metaheurísticas, como algoritmos genéticos e outras técnicas inspiradas em processos evolutivos, utilizadas para explorar o espaço de soluções eficientemente (MOTHILAL; SHARMA; TAN, 2020);
- Métodos baseados em perturbação, que geram pequenas variações nos atributos de entrada para explorar o espaço em torno de uma instância dada e identificar contrafactuais plausíveis (BRUGHMANS; LEYMAN; MARTENS, 2024; POYIADZI et al., 2020).

A maioria dos métodos de geração de contrafactuais propostos na literatura utiliza abordagens baseadas em otimização. Nessa abordagem, restrições e penalidades são adicionadas à função de perda para impor propriedades desejadas aos contrafactuais gerados (GUIDOTTI, 2024). Heurísticas e metaheurísticas, como algoritmos genéticos, também estão incluídas nessa família de algoritmos. Exemplos dessa classe de algoritmos são o **Di**verse ***C**ounterfactual **E**xplanations* (DiCE) (MOTHILAL; SHARMA; TAN, 2020), o CFNOW (de Oliveira; SöRENSEN; MARTENS, 2023) e o ***C**ounterfactuals **E**xplanations for **R**obustness, **T**ransparency, **I**nterpretability, and **Fai**rness* (CERTIFAI) (SHARMA; HENDERSON; GHOSH, 2020).

Nos métodos baseados em perturbação, tais perturbações são tipicamente introduzidas por meio de estratégias baseadas em gradiente ou, em alguns casos, amostragem aleatória. Outras abordagens ampliam essa ideia substituindo uma ou mais variáveis de instâncias reais no conjunto de dados para gerar contrafactuais plausíveis e alinhados com a distribuição de dados subjacente. Como exemplo dessa abordagem é possível citar ***N**earest **I**nstance **C**ountefactual **E**xplanations* (NICE) (BRUGHMANS; LEYMAN; MARTENS, 2024), ***F**easible*

*and **A**ctionable **C**ounterfactual **E**xplanations* (FACE) (POYIADZI et al., 2020) and, ***G**enerating **RA**ndom **C**ontrastive sampl**E**s* (GRACE) (LE; WANG; LEE, 2020).

Os métodos generativos utilizam redes neurais ou modelos probabilísticos para sintetizar exemplos contrafactuais realistas. Exemplos notáveis incluem o CounterGAN (NEMIROVSKY et al., 2022) e o Counterfactual Conditional Heterogeneous Variational Autoencoder (CCHVAE) (PAWELCZYK; BROELEMANN; KASNECI, 2020a). Nota-se ainda uma subdivisão clara entre autores que empregam algoritmos intitulados *Generative Adversarial Networks* (GAN) (GOODFELLOW et al., 2020), como Nemirovsky et al. (2022), Shao et al. (2023) e Yang et al. (2021), e os autores que utilizam *Variational Autoencoder* (VAE) (KINGMA; WELLING et al., 2019), como Zhang, Barr e Paisley (2022) e Pawelczyk, Broelemann e Kasneci (2020a). Embora essas abordagens visem gerar contrafactuais que se assemelhem a dados reais, um desafio fundamental reside em garantir sua viabilidade e alinhamento com as distribuições de dados reais, um problema frequentemente observado em modelos como GAN e VAE (VERMA et al., 2024). Além disso, os substanciais recursos computacionais necessários para o treinamento e a complexidade inerente dos modelos generativos apresentam limitações adicionais à implementação prática.

Embora ainda relativamente pouco explorado, o grupo de algoritmos que geram contrafactuais com base em relações causais tem se mostrado altamente relevante, dados os resultados promissores que demonstraram. A análise causal difere das associações puramente estatísticas por se concentrar nas relações de causa e efeito, em vez de meras correlações, que podem levar a conclusões enganosas. De acordo com Pearl (2009), as relações associativas refletem distribuições que ocorrem simultaneamente entre variáveis (por exemplo, correlação ou dependência condicional), enquanto as relações causais envolvem influência e não podem ser definidas puramente por meio de distribuições. Essa distinção é crucial em disciplinas como estatística, epidemiologia e ciência da computação. Nogueira et al. (2022) definem causalidade como a influência de um evento sobre outro, enquanto Pearce e Lawlor (2016) a descrevem por meio da dependência de variáveis, onde uma mudança na variável `X` resulta em uma mudança correspondente em `Y`, caracterizando uma relação causal.

A causalidade tem sido cada vez mais aplicada em aprendizado de máquina para melhorar a confiabilidade e a interpretabilidade dos modelos. Os Modelos Causais Estruturais (*Structural Causal Models* - SCMs) (PEARL, 2009) são amplamente utilizados para modelar essas relações, e classificadores causais construídos com base neles têm demonstrado desempenho superior (O'BRIEN; KIM; WEBER, 2023). Os contrafactuais também desempenham um papel central na inferência causal, particularmente na estimativa de efeitos de tratamento variáveis no tempo em áreas como a epidemiologia (MELNYCHUK; FRAUEN; FEUERRIEGEL, 2022). Pesquisas recentes, como o trabalho de Shao et al. (2023), demonstram o uso de SCMs para gerar contrafactuais de alta qualidade para a explicabilidade. Apesar de seu potencial, essas abordagens frequentemente dependem muito do conhecimento especializado do domínio de conhecimento

da aplicação para construir grafos causais precisos.

Dada a diversidade de abordagens para gerar explicações contrafactuais, selecionamos algoritmos representativos de diferentes famílias metodológicas para comparação. Abaixo, são detalhados os três métodos de referência. Esses métodos foram escolhidos por sua relevância na literatura e características complementares.

- **DiCE:** ***Di****verse* ***C****ounterfactual* ***E****xplanations* é um método amplamente utilizado para geração de contrafactuais, criado por Mothilal, Sharma e Tan (2020). Disponibilizado como uma biblioteca de explicações contrafactuais desenvolvida pela Microsoft Research, seu principal objetivo é produzir múltiplos contrafactuais válidos e diversos, minimizando as alterações na instância original. O DiCE suporta diferentes estratégias de geração, incluindo amostragem aleatória, KD-Tree (para instâncias nos dados de treinamento) e algoritmos genéticos. Apesar de sua popularidade, o método não incorpora relações causais, o que pode levar à geração de contrafactuais irreais ou não plausíveis;

- **NICE:** ***N****earest* ***I****nstance* ***C****ounterfactual* ***E****xplanations* é um método de geração de contrafactuais (BRUGHMANS; LEYMAN; MARTENS, 2024) projetado para aumentar a plausibilidade e reduzir o tempo de execução, particularmente em contextos de dados tabulares. Ao contrário das abordagens baseadas em otimização, o NICE opera substituindo diretamente os valores das características usando exemplos de um conjunto de dados fornecido pelo usuário, garantindo que todos os contrafactuais gerados sejam fundamentados em instâncias reais. O algoritmo visa modificar o número mínimo de características necessárias para alcançar o resultado desejado. No entanto, sua eficácia está intimamente ligada à qualidade e diversidade do conjunto de dados de entrada, o que pode limitar o desempenho quando a amostra de dados não é representativa;

- **CFNOW:** *Counterfactuals Now* (OLIVEIRA; SÖRENSEN; MARTENS, 2024) é um algoritmo versátil, agnóstico e compatível com diversos tipos de dados, incluindo tabulares, imagens e texto. Ele utiliza a meta-heurística denominada Busca Tabu (GLOVER, 1989) para explorar eficientemente o espaço de busca por contrafactuais, visando evitar a convergência para mínimos locais e a geração de instâncias não plausíveis ou fora da distribuição.

Deve-se destacar ainda que, apesar de seus resultados promissores e de ser um dos poucos trabalhos na literatura que incorpora relações causais na geração de contrafactuais, o método de Shao et al. (2023) não foi incluído nas comparações deste estudo. Isso se deve à sua forte dependência de conhecimento especializado fornecido por humanos para viabilizar a integração da causalidade por meio de grafos, o que contrasta com o foco adotado neste trabalho. Tanto no desenvolvimento quanto na validação da proposta aqui apresentada, buscou-se priorizar métodos agnósticos de modelo e independentes de intervenção humana.

Assim, enquanto Shao et al. (2023) utilizam causalidade a partir de conhecimento explícito fornecido por especialistas, o método proposto neste trabalho fundamenta-se em técnicas de inferência causal extraídas diretamente dos dados, dispensando a necessidade de conhecimento externo. A abordagem adotada também enfatiza o uso de estratégias de otimização bioinspiradas, que oferecem um equilíbrio adequado entre custo computacional, flexibilidade e capacidade de gerar contrafactuais diversos e acionáveis.

3.1.2 Avaliação de Contrafactuais

De acordo com Guidotti (2024), formalizar um conjunto de métricas é essencial para avaliar a qualidade das explicações contrafactuais. As métricas para avaliar explicações contrafactuais podem ser categorizadas em quantitativas e qualitativas. Embora as métricas quantitativas capturem as propriedades estatísticas dos contrafactuais, as métricas qualitativas avaliam sua viabilidade e utilidade no mundo real.

As métricas quantitativas são mais amplamente adotadas devido à sua facilidade de cálculo e ao consenso em torno de suas definições. As métricas qualitativas são mais difíceis de avaliar. Sua subjetividade inerente e a falta de metodologias de cálculo bem estabelecidas tornam sua avaliação menos direta. No entanto, esses dois tipos de métricas não devem ser vistos como categorias mutuamente exclusivas, mas sim como dimensões complementares da avaliação de contrafactuais. Um contrafactual ideal não deve apenas satisfazer critérios quantitativos, garantindo consistência técnica e estatística, mas também atender a requisitos qualitativos, garantindo relevância prática e utilidade no mundo real.

Neste trabalho, métricas quantitativas são calculadas para fornecer uma avaliação abrangente dos contrafactuais gerados. Uma das contribuições consiste no desenvolvimento de uma metodologia para mensurar a métrica qualitativa conhecida como acionabilidade, preenchendo uma lacuna importante nas práticas vigentes de avaliação de explicações contrafactuais.

Em geral, métricas quantitativas, como validade, esparsidade, tempo de execução e proximidade, são claramente definidas e amplamente padronizadas na literatura, permitindo comparações consistentes entre diferentes métodos. Enquanto isso, as métricas qualitativas, particularmente a acionabilidade, não possuem uma definição clara ou padronizada. Termos como acionabilidade, viabilidade e plausibilidade são frequentemente usados, às vezes com significados semelhantes, mas cada um com foco em ideias diferentes (GUIDOTTI, 2024; VERMA et al., 2024).

Assim, com base nas definições de plausibilidade, aplicabilidade e viabilidade (Capítulo 2), que avaliam contrafactuais a partir de três perspectivas distintas, porém complementares, este trabalho propõe não apenas aprimorar a forma de mensurar qualitativamente a métrica de acionabilidade, mas também fazê-lo de maneira sistemática e reproduzível, visando estabelecer uma metodologia que possa ser reaplicada e comparada em estudos futuros, aprimorando a

Tabela 1 – Métricas utilizadas em trabalhos relacionados.

Métrica	**Referências**	**Qtd**
Distância ou Proximidade	(RASOULI; YU, 2021; RASOULI; YU, 2022; NEMIROVSKY et al., 2022; SCHLEICH et al., 2021; KURATOMI et al., 2022; PAWELCZYK; BROELEMANN; KASNECI, 2020a; CHEN et al., 2022; MOTHILAL; SHARMA; TAN, 2020; OLIVEIRA; SÖRENSEN; MARTENS, 2024; DUONG; LI; XU, 2023; DANDL et al., 2024; PIAGGESI et al., 2024; DOMNICH; VICENTE, 2024; BRUGHMANS; LEYMAN; MARTENS, 2024; PANAGIOTOU et al., 2024; JI et al., 2023; PASCUAL-TRIANA et al., 2025)	17
Tempo de Execução	(ZHANG; BARR; PAISLEY, 2022; RASOULI; YU, 2021; RASOULI; YU, 2022; ALBINI et al., 2022; NEMIROVSKY et al., 2022; KANAMORI et al., 2020; SCHLEICH et al., 2021; KURATOMI et al., 2022; CHEN et al., 2022; OLIVEIRA; SÖRENSEN; MARTENS, 2024; DANDL et al., 2024; REDELMEIER et al., 2024)	12
Esparsidade	(RASOULI; YU, 2022; NEMIROVSKY et al., 2022; KURATOMI et al., 2022; CHEN et al., 2022; WANG et al., 2021; MOTHILAL; SHARMA; TAN, 2020; OLIVEIRA; SÖRENSEN; MARTENS, 2024; DANDL et al., 2024; PIAGGESI et al., 2024; REDELMEIER et al., 2024; BRUGHMANS; LEYMAN; MARTENS, 2024; PANAGIOTOU et al., 2024; PASCUAL-TRIANA et al., 2025)	13
Viabilidade ou Plausibilidade	(RASOULI; YU, 2022; ALBINI et al., 2022; NEMIROVSKY et al., 2022; SCHLEICH et al., 2021; CHEN et al., 2022; MOTHILAL; SHARMA; TAN, 2020; DANDL et al., 2024; PIAGGESI et al., 2024)	8
Acionabilidade	(PASCUAL-TRIANA et al., 2025)	1

Fonte: Elaborada pelo autor.

abordagem de Pascual-Triana et al. (2025), cuja medida atual considera apenas mudanças em atributos imutáveis.

A fim de quantificar a utilização das métricas na literatura, a Tabela 1 possibilita identificar os trabalhos nos quais as métricas mencionadas foram utilizadas. Ela sintetiza as métricas mais utilizadas nos trabalhos analisados, permitindo observar padrões recorrentes na avaliação de contrafactuais. Nota-se um foco predominante (métricas mais empregadas nos trabalhos revisados) da literatura em métricas quantitativas, capazes de mensurar propriedades estruturais dos contrafactuais, como o quão próximos estão da instância original, quantas variáveis são modificadas e o tempo de geração.

A métrica de viabilidade/plausibilidade, utilizada em oito trabalhos, indica uma preocupação crescente, porém ainda limitada, com a coerência dos contrafactuais em relação à distribuição dos dados reais. Sua baixa utilização evidencia que, sem avaliar esta métrica, os métodos podem gerar contrafactuais inviáveis, sugerindo ações impossíveis ou inconsistentes com o domínio.

Por fim, ainda se tratando de métricas qualitativas, observa-se uma lacuna significativa em que apenas um dos trabalhos revisados utiliza uma métrica explícita de acionabilidade. A baixa adoção desse tipo de métrica evidencia a dificuldade da área em avaliar se as mudanças sugeridas pelo contrafactual são realmente factíveis, compreensíveis e aplicáveis no mundo real. Isso ressalta uma das principais contribuições desta pesquisa: a de definir uma metodologia capaz de quantificar a acionabilidade dos contrafactuais. Assim, a Tabela 1 não apenas mapeia o uso das métricas existentes, mas também evidencia um desequilíbrio entre o uso evidente de métricas quantitativas e a quase ausência de métricas qualitativas.

3.2 Lacunas de Pesquisa

A partir da revisão da literatura realizada foi possível identificar um conjunto de lacunas recorrentes tanto nos métodos de geração de contrafactuais quanto em suas estratégias de avaliação. Essas lacunas não são necessariamente explicitadas pelos autores em seus respectivos trabalhos, mas emergiram da análise comparativa entre as propostas existentes, considerando suas limitações, pressupostos e resultados.

Para tornar esse mapeamento transparente ao leitor, as lacunas são listadas a seguir, acompanhadas de referências a trabalhos nos quais (i) a limitação é discutida explicitamente, ou (ii) a limitação se manifesta na forma de hipóteses adotadas ou trabalhos futuros, ainda que sem ser nomeada como "lacuna" pelos autores.

1. Definição das restrições que devem ser respeitadas pelo método de geração de contrafactuais (*human-in-the-loop* x automatização). Lacuna encontrada em (RASOULI; YU, 2022), (ALBINI et al., 2022), (DANDL et al., 2024), (SCHLEICH et al., 2021), (MOTHILAL; SHARMA; TAN, 2020) e (PANAGIOTOU et al., 2024);

2. Investigação do benefício do conhecimento de domínio para criar recursos acionáveis mais realistas e amigáveis. Lacuna encontrada em (RASOULI; YU, 2022), (DANDL et al., 2024), (MOTHILAL; SHARMA; TAN, 2020) e (ALBINI et al., 2022);

3. Seleção de contrafactuais considerando relações de causalidade. Lacuna encontrada em (ALBINI et al., 2022), (WANG et al., 2021), (SHAO et al., 2023) e (MOTHILAL; SHARMA; TAN, 2020);

4. Falta de definição de uma metodologia padrão para calcular acionabilidade, plausabilidade e/ou viabilidade. Lacuna encontrada em (REDELMEIER et al., 2024) e (PASCUAL-TRIANA et al., 2025).

A Tabela 2 sintetiza essas lacunas e destaca como elas se manifestam neste trabalho e nos que foram analisados, permitindo visualizar comparativamente em que medida a solução aqui

Tabela 2 – Lacunas preenchidas pelos métodos de geração de contrafactuais em comparação ao CausalBioCF.

Método	Restrições de domínio	Conhecimento de domínio	Causalidade	Acionabilidade	Tempo
CausalBioCF	Sim	Sim	Sim	Sim	Parcial
NICE	Parcial	Parcial	Não	Não	Sim
DiCE	Parcial	Parcial	Não	Não	Não
CFNOW	Não	Não	Não	Não	Sim

Fonte: Elaborada pelo autor.

desenvolvida avança em relação aos métodos selecionados do estado da arte. A tabela categoriza os trabalhos segundo cinco dimensões que representam lacunas recorrentes na área: (i) o respeito a restrições de domínio; (ii) a geração de conhecimento de domínio de maneira automatizada, sem interferência humana; (iii) a incorporação explícita de relações causais; (iv) a capacidade de avaliar a acionabilidade dos contrafactuais gerados; (v) e o tempo de execução para gerar um contrafactual.

Observa-se que os métodos selecionados não integram informações causais de forma sistemática, o que já era esperado, visto que até o momento, apenas o trabalho de Shao et al. (2023) utiliza essa abordagem, ainda que de maneira manual. Além disso, tais métodos não avaliam o quão acionáveis são os contrafactuais produzidos. Em muitos casos, a utilização de conhecimento de domínio para gerar restrições de domínio é parcial, incluída de maneira manual, como no DiCE ou por meio de uma base amostral, conforme feito no NICE. Em contraste, o CausalBioCF incorpora causalidade derivada diretamente dos dados, respeita restrições estruturais do domínio e introduz uma métrica específica para avaliação da acionabilidade, preenchendo lacunas que permanecem abertas nos métodos atuais.

Por fim, enquanto o tempo de geração de contrafactuais é uma limitação evidente no DiCE, esse aspecto não se mostra problemático nos demais métodos analisados. No CausalBioCF, embora o desempenho temporal permaneça adequado, a otimização dessa métrica é deliberadamente secundarizada em favor da obtenção de contrafactuais mais acionáveis. Essa comparação evidencia de maneira sintética como a metodologia proposta avança em relação ao estado da arte, tanto no processo de geração quanto na avaliação das explicações contrafactuais.

3.3 Considerações Finais

A investigação sobre geração de contrafactuais e sua aplicabilidade no campo de explicabilidade de algoritmos de aprendizado de máquina revelou que, embora apresente avanços relevantes, especialmente no desenvolvimento de algoritmos, ainda existem desafios expressivos no que diz respeito à avaliação e à qualidade das explicações geradas. Em particular, observou-se que a maioria dos estudos concentra-se em avaliar contrafactuais por meio de métricas quantitativas, enquanto abordagens que consideram aspectos qualitativos, como realismo, factibilidade e

acionabilidade das recomendações, permanecem pouco estruturadas na literatura.

Além disso, verificou-se que o uso de fatores e relações causais na geração de contrafactuais ainda é pouco explorada e, quando o fazem, dependem geralmente de intervenção humana. Essa escassez reforça a necessidade de métodos que integrem causalidade de forma sistemática e automática, promovendo explicações mais coerentes com o domínio e realmente aplicáveis. Esses achados forneceram a base conceitual e motivacional para o desenvolvimento da proposta apresentada no Capítulo 4, em que é detalhado a abordagem concebida para avançar nesse cenário, tanto no âmbito algorítmico quanto no metodológico.

4 CAUSALBIOCF: GERAÇÃO E ACIONABILIDADE DE CONTRAFACTUAIS

Neste capítulo é apresentada a proposta que fundamenta este trabalho, detalhando a abordagem adotada para alcançar os objetivos propostos. São descritas de forma estruturada as duas principais contribuições deste trabalho. Primeiro, o método CausalBioCF (Seção 4.1) é apresentado visando incorporar conhecimento de domínio obtido por meio de avaliações causais na geração de contrafactuais, para obtenção de explicações mais acionáveis. Posteriormente, é descrito um novo método para avaliação qualitativa de contrafactuais (Seção 4.2), baseada em acionabilidade.

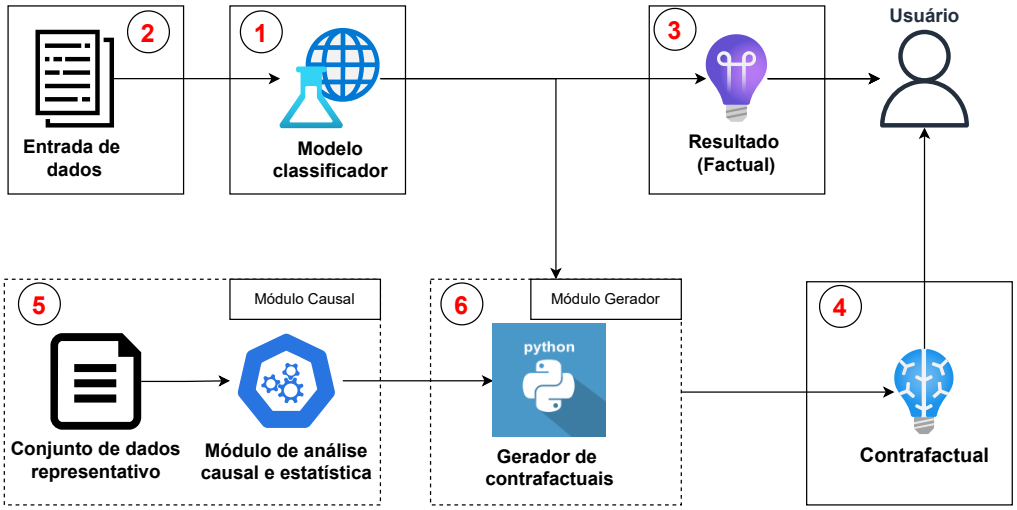
4.1 CausalBioCF

Denominado CausalBioCF, o método proposto incorpora técnicas da literatura para extrair conhecimento de domínio e empregá-lo na geração de contrafactuais por meio de um método de otimização bioinspirado. Essa abordagem é especificamente projetada para dados tabulares, garantindo que os contrafactuais gerados sejam significativos e aplicáveis em conjuntos de dados estruturados. O algoritmo de geração de conhecimento de domínio utiliza análise estatística para obter intervalos/distribuições de dados e possíveis relações causais entre eles (KLINE; LUO, 2022), utilizando-os como restrições e melhorando a qualidade dos contrafactuais gerados.

Este trabalho serve como base para o uso de relações causais para impor restrições que aprimoram a qualidade dos contrafactuais gerados. Os resultados obtidos indicam que algoritmos de otimização bioinspirados, mais especificamente o algoritmo genético (GOLBERG, 1989; HOLLAND, 1992), podem ser adaptados para produzir contrafactuais mais viáveis e contextualmente apropriados na estrutura do modelo aplicado. O CausalBioCF aborda os desafios mencionados anteriormente e, particularmente, a difícil tarefa de encontrar relações causais sem a necessidade de intervenção de especialistas.

A Figura 4 ilustra a arquitetura geral da abordagem proposta. O sistema é construído em torno de um modelo de classificação ①, que pode ser um algoritmo de aprendizado de máquina caixa preta ou uma função estatística. Este modelo processa dados estruturados de entrada ②, como tuplas compostas de atributos numéricos, categóricos ou de outros tipos, para determinar uma classe prevista ou a probabilidade de uma instância pertencer a uma determinada classe. O CausalBioCF usa este classificador para avaliar a instância factual ③ e para verificar se os contrafactuais gerados por algoritmos bioinspirados pertencem à classe oposta, formando assim uma população de contrafactuais válidos ④.

Figura 4 – Arquitetura do CausalBioCF.



Fonte: Elaborada pelo autor.

Os dois módulos principais do CausalBioCF são destacados na Figura 4 por meio de caixas tracejadas. O Módulo Causal ⑤ é responsável pela extração de conhecimento de domínio e fornece as informações necessárias para orientar o Módulo Gerador ⑥, o qual efetivamente constrói contrafactuais que são acionáveis, plausíveis, diversos e válidos. O Módulo Gerador desempenha papel centralizador no fluxo do método, recebendo informações de outros módulos e transformando as restrições derivadas da análise causal em contrafactuais que respeitam relações estruturais entre variáveis, evitando alterações inviáveis ou inconsistentes, problema recorrente nos métodos tradicionais de geração. O Módulo Causal, por sua vez, opera a partir de um conjunto de dados representativo do domínio em análise, o qual descreve o comportamento observacional das instâncias. Esse conjunto de dados é utilizado para extrair conhecimento estrutural, incluindo relações causais entre atributos, intervalos válidos para variáveis contínuas, categorias permitidas para atributos categóricos e demais restrições contextuais necessárias para garantir a viabilidade dos contrafactuais.

O conjunto de dados representativo utilizado pelo Módulo Causal não é produzido pelo método proposto, mas sim fornecido pelo usuário do sistema, seja ele o desenvolvedor, analista ou pesquisador responsável pelo modelo de aprendizado de máquina. Em geral, trata-se dos mesmos dados utilizados para treinar o modelo, ou de uma amostra representativa deles, contendo as variáveis de entrada e a classe alvo devidamente identificadas. Assim, o CausalBioCF utiliza diretamente os dados observados no problema real para obter automaticamente as restrições que guiarão a geração dos contrafactuais. A Tabela 3 mostra um exemplo de um conjunto de dados representativo relativo à base `Adult` (BECKER; KOHAVI, 1996), comumente utilizado em tarefas de classificação salarial. Nele, a Instância 1 representa dados de exemplo para o público cuja renda é $\leq$ 50 mil e a Instância 2 representa dados do público da classe complementar (Renda

Tabela 3 – Exemplo de instâncias do conjunto de dados `Adult` (BECKER; KOHAVI, 1996) utilizadas como entrada para o Módulo Causal.

Atributo	**Instância 1**	**Instância 2**
Idade	39	52
Classe de Trabalho	Governo estadual	Autônomo (não incorporado)
Escolaridade	Bacharelado	Ensino Médio
Anos de Estudo (*education-num*)	13	9
Estado Civil	Nunca casado	Casado – cônjuge civil
Ocupação	Administrativo	Executivo-gerencial
Relação Familiar	Não pertence à família	Marido
Gênero	Masculino	Masculino
Ganho de Capital	2174	0
Perda de Capital	0	0
Horas Trabalhadas por Semana	40	45
País de Origem	Estados Unidos	Estados Unidos
Variável Alvo (Renda)	$\leq$ 50 mil	$>$ 50 mil

Fonte: Adaptado do conjunto de dados Adult. (BECKER; KOHAVI, 1996).

> 50 mil). O arquivo deve estar em um formato específico de banco de dados, com colunas nomeadas conforme os atributos utilizados no modelo e os valores das instâncias representados linha a linha.

As seções a seguir explicam a metodologia de inferência causal aplicada para extrair conhecimento de domínio ⑤ e detalham o algoritmo bioinspirado responsável por gerar contrafactuais ⑥ na estrutura do CausalBioCF.

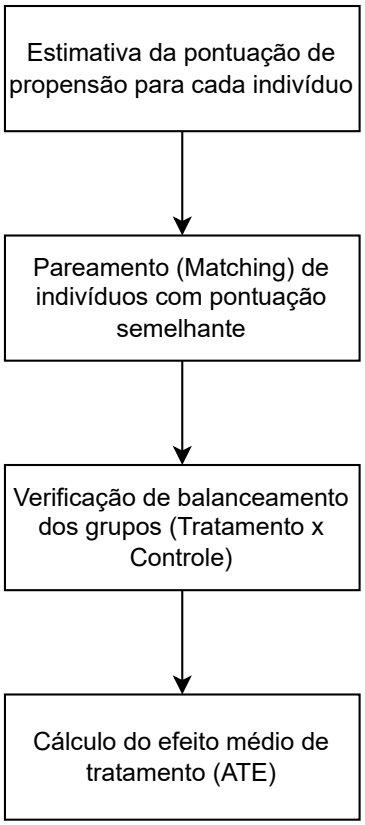
4.1.1 Inferência Causal

A fim de identificar uma boa técnica para selecionar as melhores relações causais entre preditores e variáveis alvo, sem a necessidade de definições subjetivas por especialistas, avaliaram-se alguns métodos encontrados na literatura. A partir dessa avaliação, descartou-se o uso de grafos como o Modelo Causal Estrutural, *Structural Causal Model* – SCM (PEARL, 2009), devido ao seu alto nível de subjetividade, e a Ponderação por Propensão Inversa, *Inverse Propensity Weight* – IPW (SEAMAN; WHITE, 2013), devido à dificuldade em definir variáveis de confusão que possam afetar o problema.

Deste modo, a escolha foi adaptar um método chamado Pareamento por Pontuação de Propensão, *Propensity Score Matching* – PSM (CALIENDO; KOPEINIG, 2008; ROSENBAUM; RUBIN, 1983), para guiar a geração de contrafactuais no CausalBioCF. O PSM é um método utilizado para reduzir o viés de seleção e estimar o efeito causal de um tratamento. Quatro etapas principais são seguidas para avaliar a causalidade (Figura 5):

1. Estimativa da pontuação de propensão para cada instância usando regressão logística;

Figura 5 – Fluxograma das quatro etapas que compõem o método de Propensity Score Matching (PSM).



Fonte: Elaborada pelo autor.

2. Emparelhamento de indivíduos entre os grupos de tratamento e controle com base na similaridade de pontuações (usando técnicas como Nearest-Neighbor (COVER; HART, 1967) ou Caliper Matching (ROSENBAUM; RUBIN, 1985));

3. Avaliação do equilíbrio entre os grupos de tratamento e controle, verificando se a distribuição das variáveis observáveis é semelhante após o pareamento, de modo a assegurar que eventuais diferenças no desfecho possam ser atribuídas ao efeito do tratamento;

4. Estimativa do Efeito Médio do Tratamento (*Average Treatement Effect* - ATE) comparando os resultados entre os grupos emparelhados (KANE et al., 2020; CALIENDO; KOPEINIG, 2008).

Entre as vantagens do PSM estão sua capacidade de mitigar o viés de seleção, seu bom desempenho computacional e sua flexibilidade, que permite seu uso em diversos domínios de pesquisa. O PSM também foi escolhido por sua simplicidade de implementação, sua possibilidade de adequação para análise causal automatizada e seu desempenho em tempo de execução. O método busca criar grupos de tratamento e controle equivalentes, minimizando o viés e os efeitos das variáveis de confusão, de modo que o efeito médio do tratamento (ATE) possa ser

calculado, quantificando assim as relações causais entre as variáveis. Embora se reconheça que o modelo apresenta limitações, como o controle apenas de variáveis observáveis e a influência na estimação da pontuação de propensão, entende-se que, no cenário de aplicação pretendido, essas limitações não são tão impactantes em comparação com o benefício de alcançar uma análise causal independente do conhecimento fornecido por humanos.

Do ponto de vista metodológico, é importante destacar que foram consideradas alternativas para o tratamento de causalidade, incluindo o desenvolvimento de um novo método específico para o contexto de contrafactuais. Assim, a opção por adaptar o PSM foi motivada tanto por sua robustez e maturidade teórica quanto pela possibilidade de integrá-lo eficientemente ao fluxo de geração do CausalBioCF. Como resultado, foi preservado o foco da tese nas contribuições diretamente relacionadas à obtenção de conhecimento de domínio a fim de gerar restrições para os métodos existentes que garantam a acionabilidade dos contrafactuais gerados.

Para explorar o PSM durante a geração de contrafactuais, a abordagem proposta por Kline e Luo (2022) foi adaptada para torná-la adequada ao contexto específico de explicações contrafactuais. Em seu formato original, o PSM é aplicado para estimar o efeito causal de um tratamento definido pela variável de interesse. Neste trabalho, tal procedimento foi generalizado para permitir a avaliação sistemática do efeito causal entre cada variável explicativa do conjunto de dados e a variável alvo. Dessa forma, o PSM pôde ser reutilizado para mapear relações causais potenciais em um cenário em que não existe um tratamento natural previamente definido, exigindo a reinterpretação de cada atributo como possível interveniente no valor previsto pelo modelo.

Essa adaptação implica em uma flexibilização das premissas clássicas associadas ao PSM. Assim, a hipótese de ignorabilidade condicional e a ausência de confundidores não observados, por exemplo, não são explicitamente verificadas neste trabalho. Essa escolha é deliberada e visa possibilitar a automatização completa do processo, eliminando a necessidade de intervenção de especialistas humanos na definição de restrições de domínio.

Nesse contexto, o PSM é empregado não como um método de inferência causal estrita, mas como um mecanismo heurístico orientado por causalidade para identificar relações potencialmente relevantes e estruturar o espaço de busca do gerador de contrafactuais. A incorporação de estratégias para lidar explicitamente com confundidores não observados é deixada como diretriz para trabalhos futuros.

Após essa adaptação, o ATE passou a ser calculado individualmente para cada variável, permitindo quantificar sua influência causal sobre o desfecho. Para distinguir variáveis relevantes daquelas cuja contribuição causal era desprezível, adotou-se uma classificação baseada na significância do efeito, seguindo a convenção de *effect size* de Cohen, também conhecida como *Cohen's d* (COHEN, 2013). Essa medida padroniza o ATE, dividindo-o pelo desvio padrão combinado, o que permite comparar efeitos entre variáveis que operam em escalas distintas. Com base nessa classificação, variáveis cujo tamanho de efeito apresentava significância pequena

($\|d\| \leq 0{,}2$) foram consideradas irrelevantes do ponto de vista causal e tratadas como variáveis de confusão ou ruído, sendo removidas do conjunto de atributos candidatos à otimização. Em contraste, variáveis classificadas com efeito médio ($0{,}2 < \|d\| \leq 0{,}5$) ou alto ($\|d\| > 0{,}5$) foram preservadas, por apresentarem influência causal significativa sobre a variável-alvo.

Esse procedimento reduziu expressivamente o espaço de busca do algoritmo genético, diminuindo o custo computacional e mitigando o risco de produzir contrafactuais inviáveis ou inconsistentes com o domínio, ao mesmo tempo, em que preserva as relações causais mais relevantes para o modelo. Para que o PSM pudesse funcionar como um mecanismo automatizado de extração de restrições, foi necessário reinterpretar o pareamento como uma ferramenta para identificar efeitos estruturais entre variáveis, e não como um procedimento de comparação entre grupos de indivíduos, como ocorre em sua aplicação tradicional. Além disso, os critérios de balanceamento foram ajustados para lidar com conjuntos de dados que contêm simultaneamente atributos contínuos e categóricos em escalas distintas, de modo a garantir que o pareamento considere adequadamente a similaridade entre instâncias, sem que variáveis com maior variância ou cardinalidade dominem a estimação dos efeitos causais.

Por fim, as saídas do método precisaram ser transformadas em insumos adequados para o gerador de contrafactuais, deixando de operar exclusivamente como estimativas de efeito causal direcionadas à inferência científica. Essa transformação envolveu derivar, a partir dos resultados do PSM, informações diretamente utilizáveis como restrições: quais variáveis são mutáveis ou imutáveis, quais limites de valor são plausíveis para cada atributo, obtidos a partir da amostra de dados utilizada como entrada, e quais relações de dependência devem ser mantidas. Todo esse processo foi estruturado de forma totalmente automatizada. O método retorna uma lista contendo apenas as variáveis cujo efeito ultrapassa o limiar definido (efeito causal médio ou alto), indicando quais atributos devem participar da etapa de otimização durante a geração de contrafactuais.

Com o método proposto não é necessária a presença de um especialista humano (*man-in-the-loop*) para selecionar variáveis ou definir restrições, superando justamente uma das limitações observadas em métodos anteriores que incorporam causalidade manualmente, como em (SHAO et al., 2023). Dessa forma, o método passou a atuar não apenas como ferramenta de análise causal, mas como componente ativo na definição do espaço de busca para a geração de contrafactuais acionáveis.

Juntamente com essa etapa de inferência causal, análises estatísticas básicas também foram incorporadas ao processo de aquisição de conhecimento de domínio. Enquanto o PSM fornece quais variáveis podem ser modificadas, as análises estatísticas determinam como elas podem ser modificadas. Intervalos válidos para atributos numéricos foram estimados a partir de medidas robustas, evitando que o gerador produza valores fora do domínio plausível. Para atributos categóricos, o conjunto de categorias observadas no dado foi utilizado como domínio permitido, evitando combinações inexistentes ou sem significado real. Essa combinação de

inferência causal e análise estatística é essencial para a proposta deste trabalho: diferentemente da aplicação direta dos métodos existentes, que tratam o PSM como ferramenta para avaliação causal tradicional, aqui ele foi reinterpretado e adaptado como mecanismo de restrição e estruturação do espaço de busca para geração de contrafactuais acionáveis, superando limitações típicas observadas nas abordagens puramente associativas.

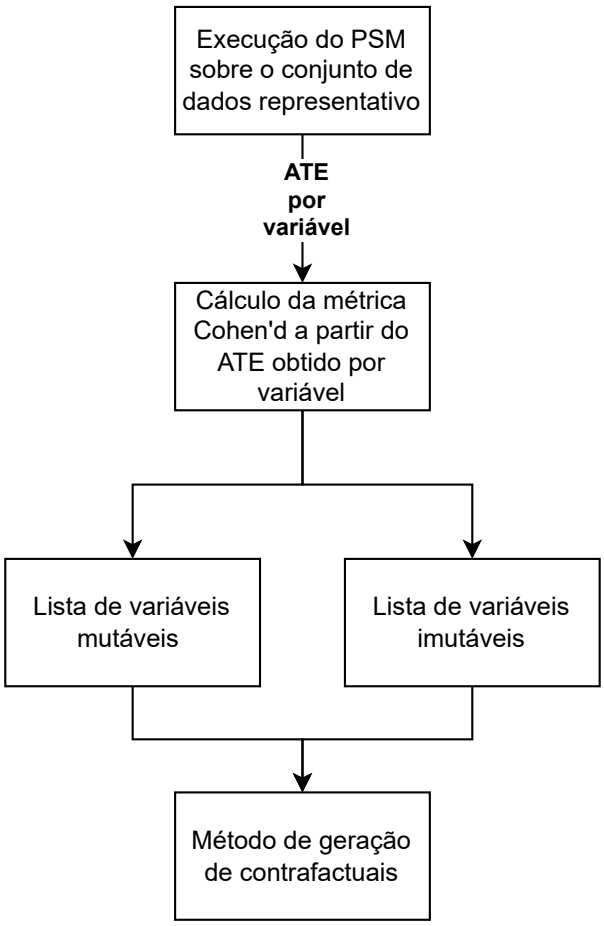
A Figura 6 ilustra o processo adotado para transformar os resultados do PSM em restrições estruturadas para orientar a geração de contrafactuais. Cada etapa do fluxograma contribui para identificar, quantificar e selecionar as variáveis que podem ser modificadas pelo gerador, garantindo que as intervenções propostas permaneçam causais e plausíveis no domínio do problema. É importante ressaltar que a etapa baseada em PSM é dedicada exclusivamente a avaliar relacionamentos causais entre os atributos e a variável alvo, gerando como resultado uma lista de variáveis consideradas causalmente relevantes e, portanto, candidatas a serem tratadas como mutáveis durante a geração de contrafactuais. Assim, ela não determina por si só todas as restrições necessárias. Para completar essa estrutura, análises estatísticas adicionais são aplicadas aos dados originais a fim de fornecer limites plausíveis ao nível marginal, isto é, variável a variável. Esses limites são utilizados para evitar valores numéricos fora da distribuição observada e categorias inexistentes para atributos categóricos.

4.1.2 Algoritmo de Geração de Contrafactuais

Para utilizar todo o conhecimento de domínio obtido por meio da análise causal e superar o desafio de usar a causalidade na geração de contrafactuais, decidiu-se trabalhar com otimização por algoritmos bioinspirados. Este tipo de algoritmo oferece um equilíbrio favorável entre simplicidade de implementação e eficiência computacional em comparação com métodos tradicionais de otimização matemática e abordagens generativas. Especificamente, utilizou-se o algoritmo genético (HOLLAND, 1992), conhecido por suas robustas capacidades de busca em espaços de alta dimensão. Juntamente com o conhecimento do domínio, este método permite que o CausalBioCF gere contrafactuais coerentes e acionáveis.

Considerando um algoritmo genético com etapas como inicialização da população, avaliação de aptidão, seleção, cruzamento e mutação (HOLLAND, 1992; GOLBERG, 1989), algumas modificações foram necessárias para gerar contrafactuais causais. Um dos desafios mais críticos na adaptação de um algoritmo bioinspirado para geração de contrafactuais é definir uma estratégia eficaz para gerar a população inicial. Essa etapa desempenha um papel fundamental na orientação do processo de busca e na garantia de que a função objetivo tenha convergência para uma solução ótima em um tempo de execução razoável. Para gerar a população inicial do CausalBioCF, duas estratégias são empregadas. A primeira aproveita o conjunto de dados representativo usado no módulo de inferência causal para preencher os indivíduos iniciais, total ou parcialmente, dependendo de um parâmetro percentual definido pelo usuário. Se não houver indivíduos válidos disponíveis, ou se o número for insuficiente para atingir o tamanho

Figura 6 – Fluxo de extração automática de restrições causais a partir do PSM.



Fonte: Elaborada pelo autor.

populacional desejado, os indivíduos restantes são gerados aleatoriamente. Essa estratégia visa reduzir o número de iterações necessárias para alcançar uma solução ótima, partindo de uma população que seja, pelo menos parcialmente, fundamentada no conhecimento do domínio.

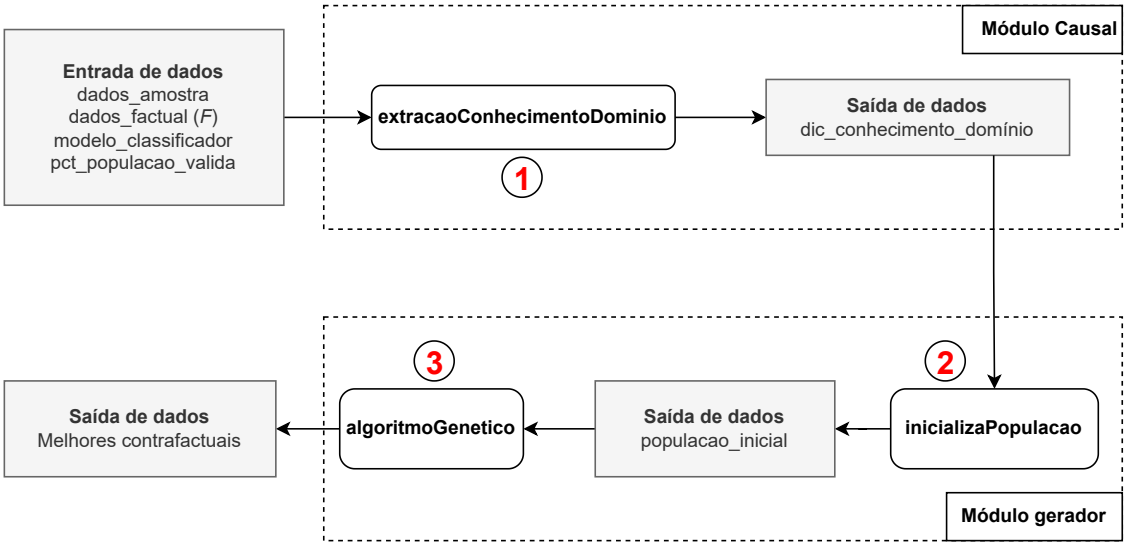
Além disso, outra modificação fundamental envolveu a função objetivo do algoritmo. Essa função desempenha um papel central no método e é diretamente influenciada pelas definições de esparsidade e distância, conforme definidas nas Equações 4.1 e 4.2, respectivamente.

$$esparsidade = \sum_{i=1}^{n} 1(x_i \neq x_i'), \tag{4.1}$$

em que n é o número de variáveis ou atributos no conjunto de dados; x_i é a *i-ésima* variável da instância factual; x_i' é o *i-ésimo* atributo da instância contrafactual; e a expressão $x_i \neq x_i'$ avalia se ambos os valores de variáveis diferem, retornando 0 quando são diferentes e 1 quando são iguais.

Para a definição de distância, utilizamos a distância euclidiana (DANIELSSON, 1980), conforme descrito pela Equação 4.2.

Figura 7 – Fluxo de execução do algoritmo CausalBioCF.



Fonte: Elaborada pelo autor.

$$distancia = \sqrt{\sum_{i=1}^{n}(x_i - x_i')^2}, \tag{4.2}$$

em que x_i é a *i-ésima* variável da instância factual e x_i' é a *i-ésima* variável da instância contrafactual.

Finalmente, a função objetivo é dada pela Equação 4.3.

$$\text{função\_objetivo} = -\big((w_1 \times \text{esparsidade}) + (w_2 \times \text{distância})\big), \tag{4.3}$$

em que w_1 e w_2 são ponderações ou pesos utilizados para atribuir maior importância à esparsidade ou à distância. Esses pesos são hiperparâmetros passados para o algoritmo bioinspirado sendo definidos, por padrão, com valores iguais, para garantir que nenhuma preferência seja dada a qualquer métrica no início da otimização. Essa inicialização reflete o objetivo do algoritmo de otimizar simultaneamente múltiplas métricas de validação. No entanto, é possível ajustar esses pesos para atribuir maior importância a uma métrica em detrimento da outra, dependendo da aplicação específica ou do resultado desejado.

Apesar de o problema envolver múltiplos critérios, optou-se por uma função objetivo escalar em vez de um método multiobjetivo, como o NSGA-II (DEB et al., 2002), a fim de simplificar a otimização e permitir uma ordenação direta dos contrafactuais gerados. Essa escolha facilita a seleção e comparação das soluções finais, mantendo flexibilidade por meio da ponderação explícita dos critérios.

O fluxo geral de execução do CausalBioCF é apresentado na Figura 7. O processo inicia quando o procedimento `extracaoConhecimentoDominio`, ainda do Módulo Causal, é chamado, recebendo todos os dados de entrada, incluindo a amostra representativa (`dados_amostra`), o factual (`dados_factual(F)`), o classificador (`modelo_classificador`) e a porcentagem da população válida que deve ser utilizada (`pct_populacao_valida`). Este procedimento então gera como saída o `dic_conhecimento_dominio`, que consiste nas restrições entregues como entradas para o método de geração de contrafactuais, ou seja, os resultados obtidos por meio da inferência causal e do PSM. Na sequência, a população é inicializada ② e, por fim, é realizada a execução do algoritmo genético para a geração de contrafactuais ③.

Olhando mais especificamente para o Módulo gerador, a geração dos contrafactuais é iniciada pelo procedimento `inicializaPopulacao` ②, responsável por criar a população inicial que servirá de entrada para o algoritmo genético. Como já mencionado, essa população é construída, sempre que possível, utilizando o mesmo conjunto de dados usado para a extração de conhecimento do domínio, garantindo que os indivíduos iniciais sejam realistas e consistentes com a distribuição dos dados. Se o tamanho da população necessário ou o número mínimo de indivíduos válidos (definido pelo parâmetro de entrada `pct_populacao_valida`) não puder ser alcançado, indivíduos adicionais são gerados aleatoriamente. Na ausência de exemplos válidos, toda a população inicial é gerada aleatoriamente. No entanto, presume-se que, se existir uma base de conhecimento válida para inferência causal, o mesmo conjunto de dados estará disponível para fornecer exemplos representativos para a inicialização da população.

O método final, `algoritmoGenetico` ③, executa o próprio AG usando `dados_factual` e `modelo_classificador` como entradas, guiado pelas restrições estabelecidas nas etapas anteriores. Durante o processo de otimização, essas restrições garantem que os contrafactuais gerados respeitem as dependências causais, resultando em soluções que sejam factíveis e acionáveis. O pseudocódigo que descreve a implementação do AG utilizado pode ser observado no Algoritmo 1. Nele, observa-se que cada candidato a solução é representado como um *indivíduo* (*ind*), isto é, um vetor de atributos que representa um possível contrafactual para a instância factual. Nesse sentido, o indivíduo corresponde ao *cromossomo*, e cada atributo representa um *gene*. O conjunto de indivíduos em uma iteração constitui a *população* (*pop*), inicializada pelo método INICIAPOP (linha 1), de modo a conter um número mínimo de candidatos válidos segundo o classificador.

A cada geração, os indivíduos são avaliados por uma função de aptidão (OBJFUNCTION) (linha 3) que combina distância e esparsidade em relação ao factual, com pesos *w1* e *w2* respectivamente, de forma a privilegiar pequenas modificações (linhas 20-22). Em seguida, aplica-se elitismo para preservar uma fração dos melhores indivíduos e utiliza-se seleção por roleta (SELECIONAPAIS) como mecanismo simples de pressão seletiva proporcional à aptidão (linhas 4-7). Novos indivíduos são produzidos por *crossover* (CROSSOVER) de um ponto e mutação com taxa (*tx_mutacao*), introduzindo diversidade e evitando uma convergência prematura do método (linhas 7-10). Por fim, candidatos que não atingem a classe desejada são removidos

Algoritmo 1 Algoritmo Genético adaptado para geração de contrafactuais

Entrada de dados: dados_factual, classificador, tam_pop, tx_mutacao, num_geracoes, w_1, w_2, min_ind_validos, dados_amostra, num_contrafactuais
Saída de dados: contrafactuais

1: $pop \leftarrow$ INICIAPOP($min\_ind\_validos, dados\_amostra, dados\_factual, classificador$)
2: **for** $g = 1$ **to** num_geracoes **do**
3: $val\_aptidao \leftarrow$ OBJFUNCTION($dados\_factual, ind, w_1, w_2$) **for all** $ind \in pop$ ▷ Função de minimizacao
4: $nova\_pop \leftarrow$ ELITISMO($pop, val\_aptidao, 0.3$) ▷ Mantém 30% melhores indivíduos
5: $tam\_nova\_pop \leftarrow len(nova\_pop)$
6: **while** $tam\_nova\_pop < tam\_pop$ **do**
7: $(p_1, p_2) \leftarrow$ SELECIONAPAIS($pop, val\_aptidao$) ▷ Seleção por roleta
8: $(u_1, u_2) \leftarrow$ CROSSOVER(p_1, p_2) ▷ Crossover realizado por corte
9: $u_1 \leftarrow$ MUTACAO($u_1, tx\_mutacao$) ▷ Mutação realizada em um atributo aleatório
10: $u_2 \leftarrow$ MUTACAO($u_2, tx\_mutacao$) ▷ Mutação realizada em um atributo aleatório
11: adiciona u_1, u_2 em $nova\_pop$
12: $tam\_nova\_pop \leftarrow len(nova\_pop)$
13: **end while**
14: $pop \leftarrow$ FILTRAVALIDOS($nova\_pop, dados\_factual, classificador$) ▷ mantém somente classe desejada
15: $pop \leftarrow$ COMPLETAPOP($pop, tam\_pop, classificador$) ▷ gera indivíduos se necessário
16: **end for**
17: $val\_aptidao \leftarrow$ OBJFUNCTION($dados\_factual, ind, w_1, w_2$) **for all** $ind \in pop$ ▷ Função de minimizacao
18: $contrafactuais \leftarrow$ TOPK($pop, val\_aptidao, num\_contrafactuais$) ▷ retorna os K melhores contrafactuais
19: **return** $contrafactuais$
20: **function** OBJFUNCTION($dados\_factual, ind, w_1, w_2$)
21: **return** $-((w_1)\cdot$ESPARSIDADE(dados_factual,ind) $+ (w_2)\cdot$DISTANCIA(dados_factual,ind))
22: **end function**

(FILTRAVALIDOS) e, quando necessário, a população é complementada por novos indivíduos (COMPLETAPOP) (linhas 14-15), garantindo o tamanho (*tam_pop*). Ao término, retornam-se os K melhores indivíduos como conjunto de contrafactuais gerados (linhas 17-19).

4.1.3 Incorporação de Causalidade na Geração de Contrafactuais

É importante ressaltar que, embora o foco desta tese seja uma abordagem baseada em algoritmos evolutivos, a estratégia proposta para extrair e incorporar conhecimento causal automaticamente a partir de uma amostra de dados é, em princípio, independente do método de geração adotado. Tal integração é possível desde que o algoritmo de geração de contrafactuais permita a inserção de restrições de conhecimento de domínio, como limites válidos para atributos, relações estruturais entre variáveis e a definição explícita de quais características podem ou não ser modificadas durante o processo de geração.

Entre os métodos utilizados para validação (NICE, DiCE e CFNOW), apenas o DiCE oferece suporte nativo para a incorporação desse tipo de restrição. Sua estrutura permite que um especialista de domínio especifique manualmente quais variáveis podem ser otimizadas e quais valores são admissíveis. No contexto deste trabalho, essa funcionalidade possibilitou substituir essa etapa manual pela utilização direta e automatizada da saída do módulo causal apresentado na Figura 4, aprimorando os contrafactuais gerados sem necessidade de intervenção humana.

Por outro lado, NICE e CFNOW não oferecem mecanismos internos para a inclusão dessas restrições. A adaptação de seus códigos-fonte seria necessária para explorar plenamente as informações extraídas pelo módulo causal, mas tais modificações não foram realizadas por não constituírem o escopo deste trabalho.

4.2 Metodologia para Avaliação da Acionabilidade

A fim de solucionar a ausência de uma metodologia padronizada para a avaliação qualitativa de contrafactuais, bem como a falta de consenso nas definições de viabilidade, acionabilidade e plausibilidade, conforme discutido no Capítulo 3, este trabalho propõe uma metodologia para o cálculo de uma variável denominada acionabilidade.

O trabalho de Pascual-Triana et al. (2025) define uma métrica de acionabilidade projetada para avaliar se as características modificadas em um contrafactual são verdadeiramente mutáveis. Sua métrica fornece uma avaliação booleana, na qual um contrafactual é acionável se todas as características modificadas possuírem a propriedade de mutabilidade e não acionável caso contrário. Essa abordagem representa um passo importante para distinguir explicações viáveis daquelas que são teoricamente impossíveis ou impraticáveis. No entanto, essa avaliação booleana apresenta uma limitação significativa ao comparar contrafactuais que são parcialmente acionáveis. A definição de Pascual-Triana *et al.* apenas determina se um contrafactual é acionável ou não, sem fornecer uma maneira para os usuários avaliarem o quão acionável ele é ou compararem a acionabilidade relativa entre diferentes contrafactuais.

4.2.1 Quantificando a Acionabilidade

Para abordar a limitação encontrada na literatura e aprimorar a metodologia de Pascual-Triana et al. (2025), o presente trabalho estende o conceito de acionabilidade de uma classificação booleana para uma métrica quantitativa, porém considerando fatores qualitativos no seu cálculo. A metodologia proposta atribui um número real entre 0 e $+\infty$, representando graus de acionabilidade, em vez de um simples resultado sim/não. Valores menores indicam maior acionabilidade. Assim, a abordagem permite comparações refinadas entre cenários contrafactuais, priorizando aqueles que são mais realistas, lógica e fisicamente viáveis e práticos.

Esta proposta é particularmente relevante porque a avaliação da acionabilidade dos contrafactuais, juntamente com as métricas tradicionais, pode melhorar significativamente a explicabilidade do modelo. Por exemplo, considere um cenário contrafactual que modifica apenas uma variável, alterando a idade de um indivíduo de 45 para 44 anos. Nesse caso, a distância entre as instâncias factuais e contrafactuais seria mínima e a esparsidade envolveria apenas uma única variável. No entanto, a acionabilidade ficaria totalmente comprometida, já que a variável modificada representa um atributo que não pode ser logicamente reduzido.

Algoritmo 2 Pseudocódigo para cálculo da métrica de Acionabilidade

Entrada de dados: bd_conhecimento, dados_factual, dados_cf, ε, α

dic_acionabilidade $\leftarrow$ {}
penalidade $\leftarrow 0$
for all atributo F **in** bd_conhecimento **do**
    if $F \notin$ dados_factual **or** $F \notin$ dados_cf **then**
        dic_acionabilidade[F] $\leftarrow$ "Desconhecido"
        continue
    end if
    tipo_dados $\leftarrow$ bd_conhecimento[F]["tipo"] ▷ "categórico" or "numérico"
    direcao $\leftarrow$ bd_conhecimento[F]["direcao"] ▷ "inc.", "dec.", or None
    mutacao $\leftarrow$ bd_conhecimento[F]["mutacao"] ▷ Verdadeiro/Falso
    VF $\leftarrow$ dados_factual[F] ▷ valor factual
    VC $\leftarrow$ dados_cf[F] ▷ valor contrafactual
    if (mutacao = False) **and** (VF $\neq$ VC) **then**
        dic_actionabilidade[F] $\leftarrow$ "Nao Acionavel"
        continue
    end if
    if tipo_dados = "categorico" **then**
        if VF = VC **then**
            dic_acionabilidade[F] $\leftarrow 0$
        else
            dic_actionabilidade[F] $\leftarrow 1$
        end if
    else ▷ numérico
        diferenca $\leftarrow$ VC $-$ VF
        if direcao contradiz diferenca **then** ▷ alteração viola a direção esperada
            dic_acionabilidade[F] $\leftarrow$ "Nao Acionavel"
        else
            dic_acionabilidade[F] $\leftarrow$ diferenca
        end if
    end if
end for
for all v **in** dic_acionabilidade **do**
    if $v =$ "Nao acionavel" **then**
        penalidade $\leftarrow$ penalidade $+ \varepsilon$
    end if
end for
metrica_acionabilidade $\leftarrow (\alpha \cdot distancia\_norm(dados\_factual, dados\_cf, dic\_acionabilidade) + (1-\alpha) \cdot esparsidade\_norm(dados\_factual, dados\_cf, dic\_acionabilidade)) + penalidade$
return metrica_acionabilidade

Portanto, ao refinar a metodologia para calcular a métrica de acionabilidade, de modo a também considerar a distância e a esparsidade, garantimos a geração de contrafactuais que sejam simultaneamente próximos, concisos e verdadeiramente factíveis, abordando assim três dimensões fundamentais da qualidade dos contrafactuais. O procedimento completo para o cálculo da métrica de acionabilidade é descrito a seguir. Seu fluxo de execução é descrito no Algoritmo 2. A metodologia proposta requer quatro entradas principais:

- Uma base de dados de conhecimento de domínio (`bd_conhecimento`), fornecida por um especialista no domínio ou inferida por meio de dados. Essa base deve especificar a condição de mutabilidade dos atributos (mutável ou imutável) e as direções permitidas de variação para cada variável (aumento, diminuição ou ambos);

- O par factual/contrafactual (`dados_factual` e `dados_cf`), representando a instância de dados original e seu correspondente contrafactual gerado;
- ε representa uma penalidade adicionada aos contrafactuais não acionáveis para que eles sempre recebam uma pontuação de acionabilidade pior (maior) do que qualquer contrafactual acionável, preservando uma classificação consistente;
- α representa um parâmetro de ponderação que equilibra a contribuição da distância e da esparsidade normalizadas para o cálculo da métrica de acionabilidade.

Usando essas informações, o algoritmo inicializa um dicionário de dados vazio chamado `dic_acionabilidade` (linha 1) e itera por cada variável F da base de conhecimento, `bd_conhecimento` (linhas 2–30). Primeiro, é verificado se a variável existe nos conjuntos de dados factual e contrafactual (linhas 3–6). Se `F` estiver ausente em qualquer um dos conjuntos de dados, seu valor no dicionário é definido como "Desconhecido", indicando que nenhuma avaliação pode ser feita.

Se `F` estiver presente, o algoritmo recupera de `bd_conhecimento` seu tipo, sua condição de mutabilidade e as direções permitidas de modificação (linhas 7–9). Caso uma variável seja marcada como imutável e seu valor tenha sido alterado entre o registro factual e o contrafactual (linhas 10–15), ela é imediatamente classificada como "Não Acionável".

Para variáveis categóricas (linhas 16–21), se não houver mudança entre os valores factuais e contrafactuais, o algoritmo atribui uma pontuação de valor 0; caso contrário, atribui 1, indicando que ocorreu uma mudança. Caso a variável não seja categórica (linhas 22–29), o algoritmo calcula a diferença entre os valores contrafactuais e factuais. Em seguida, verifica se a direção da mudança viola a regra predefinida (por exemplo, um "aumento" quando apenas uma "diminuição" é permitida). Nesses casos, a variável também é rotulada como "Não Acionável". Caso contrário, a diferença numérica é armazenada em um atributo do dicionário de acionabilidade.

Ao final desta etapa, o algoritmo produz um dicionário de acionabilidade completo, contendo as pontuações de acionabilidade ao nível de variável/atributo. Esses valores são usados na segunda etapa (linhas 31–36), em que o método agrega os resultados presentes no `dic_acionabilidade` para produzir a pontuação final da métrica de acionabilidade para a instância contrafactual. Nesta etapa, os contrafactuais que violam as restrições de acionabilidade recebem uma penalidade, um custo fixo elevado (ε), que garante que eles sejam classificados pior do que qualquer contrafactual válido. Isso impede que modificações inviáveis ou irreais pareçam competitivas ao longo da avaliação. Finalmente, a métrica combina a distância euclidiana e a esparsidade normalizadas, calculadas por seus respectivos métodos (linha 36), e ponderadas por um parâmetro que equilibra ambos os componentes, com a pontuação de penalidade, para gerar o valor final de acionabilidade.

A metodologia proposta apresenta diversos aspectos positivos, entre eles o estabelecimento de uma abordagem sistemática para identificar contrafactuais acionáveis e calcular um índice associado. Inclui verificações importantes, como a detecção de modificações indevidas em variáveis imutáveis e a avaliação de se as mudanças em atributos numéricos seguem direções logicamente consistentes. Entretanto, alguns pontos merecem atenção em trabalhos futuros. A metodologia avalia os atributos de forma independente, sem considerar interações entre múltiplas variáveis (por exemplo, comparações em pares ou em grupos), o que pode limitar sua aplicabilidade em cenários mais complexos. Além disso, as variáveis categóricas não são avaliadas quanto a possíveis ordens ou estruturas inerentes, o que pode gerar incerteza sobre se as mudanças categóricas propostas são realmente acionáveis ou realistas.

Por fim, observa-se que a métrica de acionabilidade depende de um elemento que introduz certo grau de subjetividade, o `bd_conhecimento`, responsável por registrar restrições e características operacionais das variáveis envolvidas no problema.

4.2.2 Base de Conhecimento para Calcular Acionabilidade

Representado como uma das entradas do algoritmo, esse componente corresponde a uma base de conhecimento de domínio estruturada, que agrega informações essenciais para orientar a avaliação dos contrafactuais. Entre essas informações inclui-se a mutabilidade de cada variável, indicada por um valor binário, que especifica se ela pode ou não ser modificada, e as direções permitidas de alteração, classificadas como *Aumento* (quando a variável só pode assumir valores maiores do que os originais), *Redução* (quando apenas valores menores são viáveis) ou *Ambos* (quando incrementos e decrementos são considerados plausíveis para aquele atributo).

Esse tipo de definição para variáveis de mutabilidade e direção de alteração, empregado no cálculo da métrica de acionabilidade, inevitavelmente envolve juízo humano. Tal subjetividade pode impactar diretamente o valor da métrica caso as propriedades de mutabilidade ou direção de alteração sejam especificadas inadequadamente. Em situações assim, um contrafactual que seria, na prática, acionável, pode ser penalizado indevidamente, ou, inversamente, um contrafactual inviável pode deixar de receber a penalização apropriada.

Para ilustrar a utilização da base de dados de conhecimento `bd_conhecimento`, é apresentado a seguir um exemplo simplificado envolvendo três contrafactuais gerados a partir de uma mesma instância factual. Esta base de conhecimento é geralmente fornecida pelo usuário do sistema e deve relacionar as variáveis avaliadas, incluindo sua mutabilidade e as direções permitidas de alteração. A Tabela 4 apresenta o formato utilizado no exemplo aqui apresentado. Neste exemplo são consideradas cinco variáveis: (i) *Idade*; (ii) *Glicemia*; (iii) *Índice de Massa Corporal (IMC)*; (iv) *Nível de Atividade Física*; e (v) *Histórico Familiar de Diabetes*. Note que as três primeiras são todas variáveis numéricas, enquanto as duas últimas são variáveis categóricas. Note ainda que a idade pode ser modificada, mas apenas aumentando seu valor. A glicemia, o IMC e o nível de atividade física podem variar com aumento ou redução. Já o histórico familiar

Tabela 4 – Exemplo da base de conhecimento de domínio necessária para o cálculo da métrica de acionabilidade.

Variável	**Mutável**	**Direção de Alteração**
Idade	Sim	Aumento
Glicemia (mg/dL)	Sim	Ambos
IMC (kg/m^2)	Sim	Ambos
Nível de Atividade Física	Sim	Ambos
Histórico Familiar de Diabetes	Não	–

Fonte: Elaborada pelo autor.

Tabela 5 – Factual e contrafactuais avaliados.

Registro	**Idade**	**Glicemia**	**IMC**	**Atividade Física**	**Histórico Diabetes**	**Classe**
Factual	33	140	32	Baixo	Sim	Diabético
CF1	33	95	29	Moderado	Sim	Normal
CF2	31	99	32	Baixo	Não	Normal
CF3	33	99	32	Baixo	Não	Normal

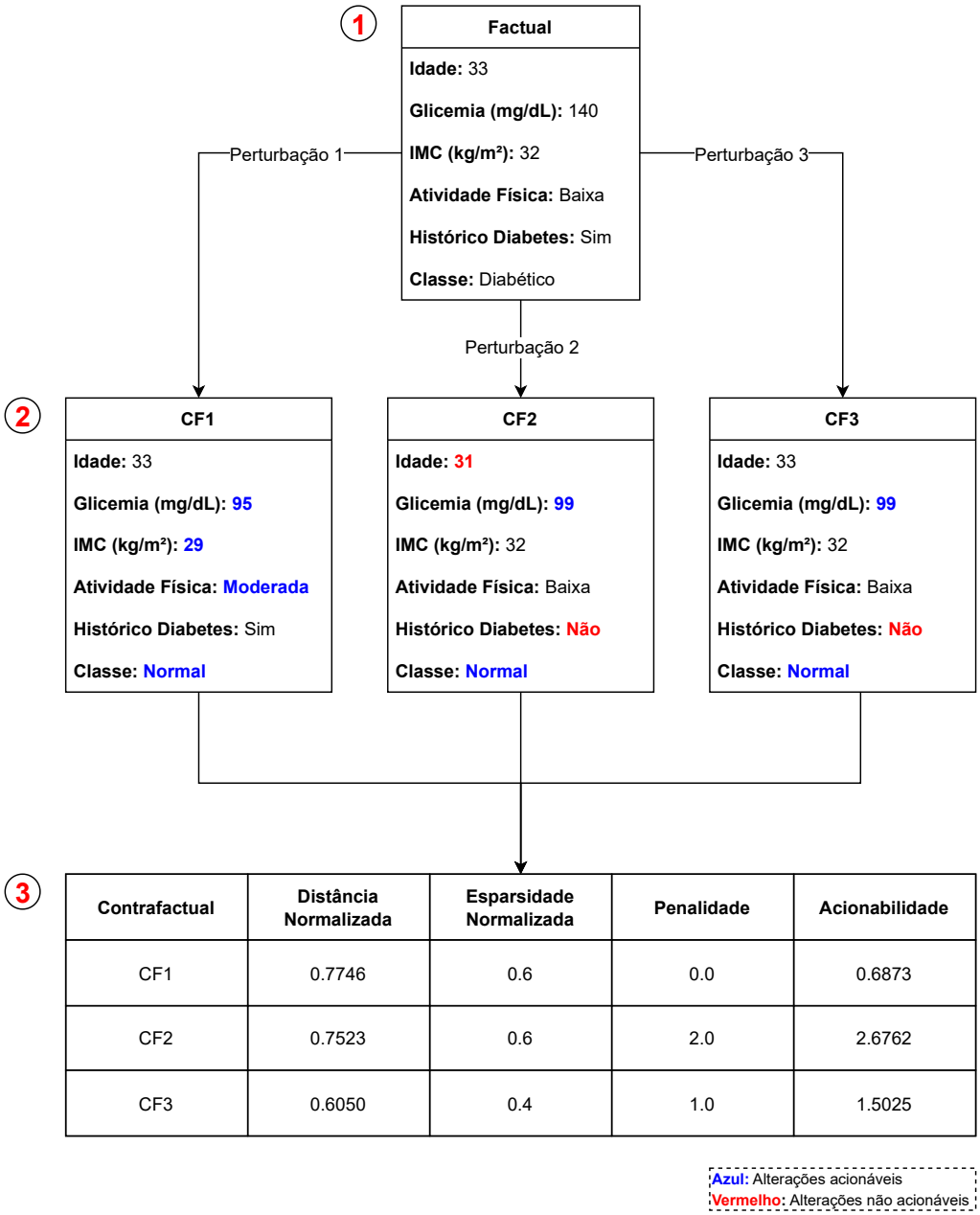
Fonte: Elaborada pelo autor.

de diabetes não pode ser modificado.

A Tabela 5 apresenta o factual e os três contrafactuais utilizados no exemplo. A partir desses valores, inicia-se o processo de verificação de acionabilidade. Em vez de simplesmente listar as diferenças entre factual e contrafactuais, o objetivo desta etapa é identificar, de forma estruturada, quais atributos sofreram alteração e em que magnitude, distinguindo mudanças em variáveis numéricas e categóricas. Este processamento é ilustrado na Figura 8, que evidencia onde ocorreram as modificações que alimentarão as etapas subsequentes da análise. Nela, observa-se que o factual é demarcado por ①. Em seguida, a primeira etapa de cálculo consiste em identificar, para cada contrafactual, quais atributos foram modificados em relação ao factual ②. Para tornar essa etapa mais clara, as alterações são destacadas por cores, sendo que em azul são marcadas as mudanças permitidas segundo a base de conhecimento de domínio (por exemplo, aumento de idade, redução de glicemia, entre outros) e, em vermelho, são destacadas as mudanças não acionáveis, isto é, que envolvam violações diretas das restrições definidas (como tentar reduzir a idade ou modificar um atributo imutável, como histórico familiar). Além disso, a fim de quantificar as alterações que ocorreram a partir das perturbações inseridas no factual, para variáveis numéricas, utiliza-se a diferença direta entre os valores. Enquanto isso, para variáveis categóricas, atribui-se 0 quando não há alteração e 1 quando os valores diferem.

Na sequência, calcula-se a distância euclidiana entre factual e contrafactual, representando o grau de alteração necessário. A distância é normalizada pelo maior valor observado no conjunto avaliado. A esparsidade é calculada como a proporção de variáveis alteradas, também normalizada. A métrica final incorpora ambos os componentes, distância normalizada e espar-

Figura 8 – Exemplo ilustrativo do processamento da métrica de acionabilidade, mostrando as diferenças entre o factual e cada contrafactual.



Contrafactual	Distância Normalizada	Esparsidade Normalizada	Penalidade	Acionabilidade
CF1	0.7746	0.6	0.0	0.6873
CF2	0.7523	0.6	2.0	2.6762
CF3	0.6050	0.4	1.0	1.5025

Fonte: Elaborada pelo autor.

sidade normalizada, sendo estes ponderados por um parâmetro $\alpha = 0{,}5$. Além disso, aplica-se uma penalidade quando um contrafactual viola restrições de conhecimento de domínio contidas em `bd_conhecimento`, como a impossibilidade de reduzir idade ou alterar o histórico familiar de diabetes. No exemplo, o CF1 respeita todas as restrições, o CF2 reduz idade, o que não é permitido, e altera o histórico de diabetes. Por fim, o CF3 também altera o histórico de diabetes, considerado imutável. Assim, CF2 recebe penalidade igual a 2 e CF3 recebe penalidade igual a 1. Por fim, em ③ apresenta-se os valores que compõem a métrica de acionabilidade calculados

para cada contrafactual.

A análise evidencia que contrafactuais com menor magnitude de alterações (menor distância) e com menor número de atributos modificados (menor esparsidade), caso respeitem o conhecimento de domínio, possuem maior acionabilidade. No exemplo, o CF1 é o contrafactual mais acionável, enquanto CF2 e CF3 apresentam valores mais elevados devido às penalidades associadas a violações de restrições de domínio, sendo CF3 mais acionável que CF2. Esse exemplo demonstra como a métrica proposta sintetiza, em um único indicador, aspectos de plausibilidade, esforço e coerência estrutural.

4.2.3 O Papel da Subjetividade

Conforme citado anteriormente, a construção da base de conhecimento de domínio (`bd_conhecimento`) para avaliação da acionabilidade, embora fundamentada em critérios racionais e alinhada ao contexto do tema central do conjunto de dados, introduz uma camada de subjetividade ao processo. Essa subjetividade é provinda principalmente de dois fatores:

- A definição de quais variáveis são mutáveis ou não (Exemplo: histórico de diabetes familiar é uma variável imutável);
- A atribuição de direções permitidas de alteração (Exemplo: Idade de um indivíduo pode apenas aumentar).

Cada uma dessas escolhas, ainda que justificáveis, depende do conhecimento e de julgamentos humanos e pode, portanto, variar entre especialistas, instituições ou cenários de aplicação. Assim, torna-se inevitável a avaliação de qual proporção a métrica de acionabilidade permanece estável quando essa representação do conhecimento de domínio sofre variações, ou então, sobre o quão robusta ela é.

A fim de manter a transparência, todas as bases de conhecimento utilizadas nos experimentos conduzidos neste trabalho encontram-se documentadas no Apêndice A, permitindo que leitores e futuros pesquisadores verifiquem, reproduzam e ajustem eventualmente os critérios adotados. Finalmente, é importante distinguir a participação do fator humano na definição das restrições causais, usadas na geração dos contrafactuais, da sua participação na avaliação da acionabilidade: enquanto a subjetividade na geração dos contrafactuais foi eliminada por meio da avaliação causal e estatística automatizada, a subjetividade remanescente na avaliação é inerente ao conceito de acionabilidade e, enquanto não puder ser realizada de maneira independente, deve ser tratada como um componente explícito e auditável do processo.

4.3 Considerações Finais

Este capítulo apresentou a abordagem proposta, na forma do CausalBioCF, para geração de contrafactuais que considera relações causais, além do método para cálculo da acionabilidade. Desta forma, foi possível conhecer a composição do algoritmo de análise causal entre atributos e a variável alvo, o algoritmo de geração de contrafactuais utilizando o conhecimento de domínio obtido e, por fim, a metodologia de avaliação de acionabilidade dos contrafactuais.

5 AVALIAÇÃO EXPERIMENTAL

Este capítulo apresenta os resultados obtidos a partir da implementação e experimentação do método proposto neste trabalho. O objetivo é avaliar o desempenho do CausalBioCF, tanto em relação à geração de contrafactuais, quanto à aplicação da metodologia de cálculo da métrica de acionabilidade proposta. Mais especificamente, os experimentos visam responder às seguintes questões de pesquisa (QP):

QP1 Quão diferentes são os contrafactuais gerados de suas instâncias factuais correspondentes?

QP2 Quão realistas, viáveis e significativos são os contrafactuais dentro de seu contexto de domínio, e de que forma o método proposto para avaliar acionabilidade aprimora e torna mais robusta essa avaliação?

QP3 Qual é o impacto da inferência causal nas métricas quantitativas e na qualidade geral dos contrafactuais?

Para responder a essas questões de pesquisa, combinamos análises quantitativas e qualitativas. A avaliação quantitativa concentra-se em métricas orientadas ao desempenho, como validade, proximidade, esparsidade e tempo de execução, que, em conjunto, indicam a eficiência e a eficácia com que cada algoritmo gera contrafactuais. A avaliação qualitativa, por sua vez, baseia-se na métrica de acionabilidade proposta, que avalia o quão realistas, viáveis e interpretáveis são os contrafactuais dentro de suas restrições causais e de domínio. Juntas, as perspectivas quantitativa e qualitativa fornecem uma visão abrangente dos pontos fortes e limitações de cada método.

O restante deste capítulo está organizado da seguinte forma. A Seção 5.1 descreve o ambiente experimental, conjuntos de dados e métodos utilizados nos experimentos. A Seção 5.2 apresenta os resultados dos experimentos realizados, enquanto a Seção 5.3 aborda diretamente as questões de pesquisa levantadas aqui em face dos resultados experimentais.

5.1 Ambiente Experimental

Todos os experimentos foram realizados em uma máquina com sistema operacional Windows 11 de 64 bits, equipada com 16 GB de RAM e um processador Intel(R) Core(TM) Ultra 7 155H de 3.80 GHz. Os modelos classificadores utilizados nas avaliações foram: Random Forest (RF) (BREIMAN, 2001), eXtreme Gradient Boosting (XGBoost ou XGB) (CHEN; GUESTRIN, 2016), KNN Classifier (KNN) (COVER; HART, 1967) e Multilayer Perceptron (MLP)

Tabela 6 – Acurácia de cada conjunto de dados por classificador.

	KNN		**MLP**		**RF**		**XGB**	
Conjunto de dados	Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste
Adult	85.03	77.84	87.60	79.26	79.06	73.91	86.19	82.97
Bank Marketing	81.22	76.82	92.97	82.24	82.07	80.89	91.24	85.18
Churn	77.13	67.66	**99.65**	**70.95**	90.16	89.22	99.88	92.81
Compas Scores	72.13	61.00	68.12	71.35	68.41	67.15	71.46	67.47
Credit Card Default	77.22	67.89	72.30	77.33	72.32	76.50	80.15	73.83
German Credit	77.59	69.00	**100.00**	**68.00**	82.02	69.00	**100.00**	**69.00**
Online Shoppers	83.82	76.48	90.31	84.34	85.47	86.13	96.30	84.59

Fonte: Elaborada pelo autor.

(ROSENBLATT, 1958). É importante ressaltar que os modelos foram treinados e avaliados especificamente para este estudo, sem o uso de modelos pré-treinados.

Os conjuntos de dados foram divididos em 90% para treinamento e 10% para teste. O MLP foi configurado com regularização L2 ($\alpha = 0,0009$) e limite máximo de 1000 iterações. O RF utilizou profundidade máxima igual a 5. O XGBoost foi treinado com 200 estimadores, profundidade máxima 5 e taxa de aprendizado de 0,1. O classificador KNN foi configurado com k=5. Todo o código-fonte foi implementado em Python 3.9[1].

Na Tabela 6 é possível verificar a acurácia de cada um dos modelos treinados por conjunto de dados balanceados. A partir da análise dos dados, é possível observar que todos os modelos apresentaram acurácia superior a 60%. É importante notar que, para os conjuntos de dados Churn e German Credit, os modelos Multilayer Perceptron e XGBoost apresentaram desempenhos que aparentemente demonstram uma condição de *overffiting* (destacados em negrito na tabela). No entanto, para confirmar o comportamento agnóstico ao modelo que os contrafactuais possuem, esses modelos foram mantidos, permitindo que os leitores observem se há diferenças nas métricas avaliadas devido a essa condição.

Nos experimentos foram gerados 10 contrafactuais para cada uma das 10 instâncias factuais selecionadas aleatoriamente do conjunto de dados de teste, com uma repetição de 20 execuções. No total, foram gerados 200 contrafactuais por conjunto de dados por modelo, correspondendo a 2000 contrafactuais no total.

Para obter os parâmetros ótimos para o algoritmo genético, testes exaustivos de otimização de hiperparâmetros (GridSearch) foram realizados (BERGSTRA; BENGIO, 2012). Nesses experimentos, foram analisados parâmetros-chave que afetam significativamente o desempenho e a convergência do algoritmo genético. A Tabela 7 apresenta os intervalos avaliados e os valores ótimos utilizados nos experimentos finais. Os intervalos considerados foram definidos a partir de testes exploratórios preliminares, nos quais faixas mais amplas de valores foram inicialmente

[1] <https://github.com/gcovfur/CausalBioCF>

Tabela 7 – Hiperparâmetros avaliados e valores ótimos selecionados para o CausalBioCF.

Hiperparâmetro	**Intervalo Avaliado**	**Valor Ótimo**
Percentual mínimo de contrafactuais válidos	70% – 100%	70%
Número de gerações / iterações	10 – 15	10
Tamanho mínimo da população	100 – 1000	100
Taxa de mutação	10% – 30%	10%

Fonte: Elaborada pelo autor.

avaliadas. Esses parâmetros otimizados foram considerados a melhor configuração para a geração de CausalBioCF contrafactual, enquanto para os demais algoritmos (DiCE, NICE, CFNOW) foram utilizadas as configurações de parâmetros padrão fornecidas por seus autores.

Além desses parâmetros, para a avaliação da métrica de acionabilidade adotou-se α = 0.5, atribuindo pesos iguais aos componentes de distância e esparsidade na composição da métrica. Essa escolha visa evitar viés em favor de um único critério quantitativo, permitindo uma avaliação equilibrada das alterações propostas pelos contrafactuais. Por fim, o valor de ε, correspondente à penalidade aplicada em casos de violação das restrições de conhecimento de domínio, foi estimado individualmente para cada conjunto de dados, de modo a refletir suas características específicas. O valor de ε é obtido adaptativamente a partir da razão entre o valor médio da métrica de acionabilidade e a proporção de contrafactuais acionáveis gerados em cada execução, com a adição de uma constante pequena ($1x10^{-9}$) para evitar divisão por zero. Em situações nas quais não há contrafactuais acionáveis, aplica-se uma penalidade robusta baseada na escala dos custos observados em execuções válidas. Essa definição garante que contrafactuais que ferem mais fortemente o conceito de acionabilidade recebam valores de custo mais elevados, penalizando tanto a intensidade das modificações quanto a ausência de soluções efetivamente acionáveis.

5.1.1 Conjuntos de Dados

Os conjuntos de dados selecionados para este estudo são amplamente reconhecidos e frequentemente adotados em pesquisas de explicação contrafactual (VERMA et al., 2024). Isso inclui conjuntos como *Credit Card Default* (YEH, 2009), *German Credit* (HOFMANN, 1994), *Adult* (BECKER; KOHAVI, 1996), *Churn* (KAGGLE, 2020), *Online Shoppers Purchasing Intention* (SAKAR; KASTRO, 2018), *Bank Marketing* (MORO; RITA; CORTEZ, 2014) e *COMPAS Scores* (PROPUBLICA, 2016), abrangendo diversos domínios de conhecimento como finanças, varejo, demografia e justiça criminal. Sua inclusão destaca a adaptabilidade do método proposto em diferentes áreas de aplicação.

Além de sua relevância na literatura, esses conjuntos apresentam características complementares, tanto em número de atributos quanto em quantidade de instâncias. Essa diversidade estrutural permite avaliar o método proposto em cenários com diferentes níveis de dimensio-

Tabela 8 – Visão geral dos conjuntos de dados utilizados para validar o método proposto.

Conjunto de dados	**Descrição**	**Atributos**	**Linhas**
Credit Card Default	Informações sobre inadimplência de clientes em Taiwan.	23	30.000
German Credit	Dados de avaliação de risco de crédito classificando indivíduos como bons ou maus pagadores.	20	1.000
Adult	Informações do censo usadas para prever se a renda anual ultrapassa US$ 50 mil/ano.	14	48.842
Churn	Dados comportamentais do cliente usados para prever a probabilidade de saída ou abandono de clientes.	20	3.333
Online Shoppers	Dados sobre o comportamento de compra dos clientes em um ambiente de comércio eletrônico online.	17	12.330
Compas	Dados sobre o risco de reincidência criminal nos Estados Unidos.	52	7.214
Bank Marketing	Dados referentes a campanhas de marketing, por meio de ligações telefônicas, realizadas por uma instituição bancária portuguesa.	16	45.211

Fonte: Elaborada pelo autor.

nalidade, reforçando a generalidade e adaptabilidade da abordagem. Mais detalhes sobre os conjuntos de dados podem ser encontrados na Tabela 8.

Outros dois conjuntos de dados comuns, *Diabetes* (KAHN, 1994) e *Titanic* (KAGGLE, 2012), foram excluídos dos experimentos devido a limitações tanto no número de registros quanto no número de atributos, o que poderia comprometer a confiabilidade dos resultados.

5.1.2 Métodos de Referência

Dada a diversidade de abordagens existentes para gerar explicações contrafactuais, foram selecionados algoritmos representativos de diferentes famílias metodológicas para compor a comparação experimental. Estes métodos foram escolhidos por sua relevância na literatura e características complementares. Esta seção detalha os três métodos de referência escolhidos e destaca como cada um incorpora uma forma distinta de geração de contrafactuais.

- O DiCE (MOTHILAL; SHARMA; TAN, 2020), foi selecionado por ser um dos métodos mais difundidos para geração de múltiplos contrafactuais diversos. Foi selecionado por ser comparável ao CausalBioCF e por suportar abordagens baseadas em heurísticas. Embora eficiente e flexível, não incorpora relações causais, podendo produzir contrafactuais pouco acionáveis;
- O NICE (BRUGHMANS; LEYMAN; MARTENS, 2024) foi selecionado por representar

Tabela 9 – Métodos de geração de contrafactuais utilizados na avaliação experimental.

Algoritmo	**Abreviação**	**Descrição**
CausalBioCF	`CBio-GA`	Algoritmo bioinspirado proposto neste trabalho.
DiCE	`DiCE`	Algoritmo para geração de contrafactuais usando otimização por algoritmo genético.
Causal DiCE	`C-DiCE`	Variante do DiCE que incorpora restrições causais a partir do CausalBioCF.
NICE	`NICE-all`	Algoritmo NICE baseado em alterações nos dados do factual para gerar contrafactuais.
CFNOW	`CFNOW`	Algoritmo de geração de contrafactuais baseado em um modelo de vizinhanças.

Fonte: Elaborada pelo autor.

algoritmos baseados na substituição direta de atributos por valores observados, tendendo a aumentar a acionabilidade das soluções. O algoritmo possui diferentes variantes, cada uma otimizada para um critério específico: uma voltada à minimização da proximidade (buscando contrafactuais mais próximos do factual) e outra orientada à esparsidade (privilegiando alterações em poucos atributos). Neste trabalho, emprega-se a melhor entre suas variantes direcionadas à proximidade ou à esparsidade. Esta versão ótima será referida como `NICE-all`;

- O CFNOW (de Oliveira; SöRENSEN; MARTENS, 2023) foi selecionado, pois seu objetivo é evitar soluções inviáveis ou fora da distribuição de dados, oferecendo uma alternativa heurística capaz de lidar com diferentes domínios. Além disso, trata-se de um método recentemente proposto, juntamente com o `NICE-all`, reforçando sua relevância como representantes das abordagens mais atuais para geração de contrafactuais.

Dentre esses algoritmos, o DiCE também foi adaptado para incorporar as restrições de inferência causal produzidas pelo CausalBioCF. A partir delas, o modelo incorporou definições de quais atributos deveriam ser mutáveis e imutáveis, bem como intervalos numéricos válidos (é importante ressaltar que, em sua versão original, era o único algoritmo capaz de aceitar tais restrições como entrada). Essa versão modificada é denominada Causal DiCE (`C-DiCE`). No total, cinco variações dos quatro algoritmos são usadas para realizar os testes, conforme apresentado na Tabela 9.

5.2 Avaliação

Nesta seção comparam-se os algoritmos de geração de contrafactuais utilizando métricas quantitativas tradicionais, como distância Euclidiana/Manhattan, validade, esparsidade e tempo de execução, além do método proposto para o cálculo da acionabilidade. A avaliação evidencia como a métrica de acionabilidade pode revelar limitações de realismo e viabilidade que não

são capturadas pelas métricas quantitativas clássicas, aprimorando assim a avaliação global da qualidade das explicações produzidas.

Para esta avaliação, utiliza-se uma análise estatística não paramétrica para identificar diferenças significativas entre os algoritmos de geração de contrafactuais. Os testes não paramétricos foram selecionados porque os valores de desempenho em diferentes conjuntos de dados e modelos raramente seguem uma distribuição normal, e seu uso é fortemente recomendado na literatura de avaliação de aprendizado de máquina (DEMŠAR, 2006; GARCÍA et al., 2010). Isso torna a escolha robusta e apropriada para o contexto comparativo deste trabalho. Além disso, trabalhos recentes na literatura também seguem uma abordagem bastante similar, dentre eles `NICE-all` e `CFNOW`.

As diferenças gerais entre os algoritmos foram inicialmente avaliadas com o teste de Friedman (FRIEDMAN, 1940). Quando o teste de Friedman é significativo, realizam-se comparações *post-hoc* aos pares utilizando o teste *Wilcoxon Signed-Rank* (WILCOXON, 1945). Para lidar com múltiplas comparações, os valores de *p* (probabilidade de observar um resultado tão extremo quanto o obtido caso a hipótese nula seja verdadeira) foram ajustados usando o procedimento *False Discovery Rate* (FDR) de Benjamini–Hochberg (BENJAMINI; HOCHBERG, 1995), que oferece maior poder estatístico do que correções baseadas em erro, além de reduzir o número de falsos positivos nos resultados. Essa metodologia segue as melhores práticas estabelecidas na avaliação de inteligência artificial e explicabilidade de IA (*eXplanaible Artificial Intelligence* - XAI) (DEMŠAR, 2006; GARCÍA et al., 2010) e está alinhada com estudos recentes sobre geração de contrafactuais (BRUGHMANS; LEYMAN; MARTENS, 2024; PASCUAL-TRIANA et al., 2025). A única divergência na metodologia utilizada é o teste *post-hoc* de Wilcoxon em vez do teste de Nemenyi, devido ao menor tamanho amostral em comparação com a configuração de Brughmans, além da adoção do FDR para reduzir falsos positivos. Para maior clareza, todos os resultados experimentais são apresentados em forma de tabelas.

5.2.1 Análise Centrada em Dados Quantitativos

Para avaliar quantitativamente os contrafactuais gerados, foram utilizadas as seguintes métricas: (i) o número de contrafactuais válidos entre os produzidos; (ii) a distância entre o factual e o contrafactual (euclidiana e de Manhattan); (iii) a esparsidade; e (iv) o tempo de execução. A Tabela 10 apresenta as comparações estatísticas entre todos os algoritmos em diferentes conjuntos de dados, modelos e métricas. A tabela também inclui os valores calculados pelo método proposto para acionabilidade (última coluna), permitindo um contraste direto com as medidas tradicionais, mais centradas em aspectos quantitativos.

Na tabela, uma vitória significativa é registrada quando um algoritmo obtém uma classificação média inferior à de um concorrente (quanto menor for o ranking, maior a classificação), utilizando o método de Friedman para avaliação, e essa diferença é estatisticamente significativa após o ajuste FDR ($p < 0.05$). Assim, vitórias significativas indicam não apenas uma melhor

Tabela 10 – Taxa de sucesso significativa (%) dos algoritmos em todas as métricas.

Algoritmo	Dist. Euclidiana	Dist. Manhattan	Esparsidade	Tempo	Acionabilidade
`NICE-all`	**80.4**	**77.7**	**76.8**	**88.4**	57.1
`CBio-GA`	67.0	67.0	64.3	31.2	**72.3**
`CFNOW`	18.8	25.0	31.2	59.8	19.6
`C-DiCE`	28.6	30.4	28.6	10.7	40.2
`DiCE`	6.2	5.4	0.9	7.1	0.9

Obs. A taxa de sucesso significativa representa a porcentagem de casos (combinações de conjunto de dados e modelo) em que um algoritmo obteve um resultado estatisticamente significativo melhor do que os demais, com base no teste de Wilcoxon ($p < 0.05$). O número total de casos (combinações de conjunto de dados e modelo) é sempre 112.
Fonte: Elaborada pelo autor.

classificação média, mas também que a melhoria observada é robusta em diferentes conjuntos de dados e classificadores.

A métrica de validade (ou cobertura) mede quantos contrafactuais gerados conseguem inverter a classe da instância factual correspondente. Não apresentamos comparações estatísticas para essa métrica na Tabela 10 porque a maioria dos algoritmos teve desempenho semelhante, sem diferenças significativas. A única exceção foi o `CFNOW`, que consistentemente apresentou validade inferior a todos os outros métodos em todos os cenários. Também notamos que, embora `DiCE` e `C-DiCE` produzissem todos os contrafactuais válidos, nem sempre geravam o número total de contrafactuais solicitados.

As métricas de distância quantificam o quanto as variáveis de um contrafactual precisam mudar em relação à sua instância original para que ocorra uma inversão de classe. Elas capturam a magnitude das modificações. Portanto, distâncias menores indicam contrafactuais que requerem menos modificações, ajudando a responder à QP1. Os resultados dos testes não paramétricos e post-hoc para as distâncias euclidiana e de Manhattan (Tabela 10) mostram que o `NICE-all` alcança o melhor desempenho geral, gerando contrafactuais que permanecem mais próximos de suas instâncias factuais correspondentes. O `CBio-GA` também apresenta um desempenho sólido, obtendo distâncias significativamente menores do que o `CFNOW`, o `C-DiCE` e o `DiCE`.

A esparsidade reflete quantas variáveis precisam ser alteradas para obter um contrafactual e, portanto, serve como um indicador de simplicidade e interpretabilidade: contrafactuais que modificam menos atributos são geralmente mais fáceis de entender e usar para as ações dos usuários. Essa avaliação também ajuda a responder à QP1. Os resultados de esparsidade repetem o padrão observado nas métricas baseadas em distância. O algoritmo `NICE-all` alcança o maior número de acertos significativos em todos os conjuntos de dados e modelos, seguido pelo `CBio-GA`. Os algoritmos `CFNOW` e `C-DiCE` também produzem contrafactuais relativamente esparsos, embora com desempenho ligeiramente inferior.

Ao analisar os tempos de execução (5ª coluna) para geração de contrafactuais, observa-

mos uma grande vantagem para o NICE-all, seguido pelo CFNOW, um algoritmo projetado para execução rápida. Embora não seja o de melhor desempenho, o CBio-GA ainda supera o DiCE e o C-DiCE, que também utilizam algoritmos genéticos para gerar contrafactuais. É importante notar que, embora o CBio-GA não figure entre os mais rápidos, seu tempo de execução permanece viável. Em média, ele é apenas cerca de 0,75 segundos mais lento que o NICE-all (aproximadamente 1 segundo de execução para a geração de contrafactuais pelo CBio-GA em relação à 0.25 segundos para o NICE-all), o algoritmo mais rápido.

Em conjunto, os resultados quantitativos evidenciam um padrão consistente: NICE-all alcança o melhor desempenho geral, seguido por CBio-GA e CFNOW, que apresentam resultados competitivos na maioria das métricas. Em contraste, DiCE e C-DiCE geralmente têm um desempenho pior. No entanto, deve-se notar que a incorporação de restrições causais leva a uma melhoria substancial em C-DiCE em comparação com o DiCE original. Isso destaca que a adição de causalidade pode aprimorar significativamente a capacidade do DiCE de gerar contrafactuais mais coerentes e eficazes, embora ainda não alcance os níveis de desempenho dos melhores métodos. Notavelmente, ao considerar o novo método para calcular a acionabilidade (última coluna da Tabela 10), o C-DiCE supera até mesmo o CFNOW. Mais importante ainda é a grande melhoria em termos de acionabilidade obtida pelo CBio-GA. Os motivos por trás desse comportamento serão examinados nas próximas seções.

5.2.2 A Necessidade de Uma Métrica Mais Centrada na Qualidade

Embora as métricas quantitativas forneçam informações úteis sobre as propriedades numéricas dos contrafactuais, elas não são suficientes para avaliar sua qualidade geral. Essas métricas não podem determinar se os contrafactuais gerados são realistas ou viáveis, na prática. Consequentemente, um bom desempenho apenas em termos de distância ou esparsidade não indica necessariamente que um método produza explicações significativas ou acionáveis.

Essa limitação é ilustrada pelos exemplos na Tabela 11. Cada exemplo mostra, na primeira linha, um subconjunto de características da instância factual usada por NICE-all e CBio-GA, seguido por fragmentos dos contrafactuais correspondentes produzidos por cada método. A última coluna mostra o valor da métrica usada para avaliar cada contrafactual. No Exemplo 1, NICE-all atinge um valor de esparsidade menor, mas faz isso diminuindo a idade, um atributo imutável, tornando o contrafactual não factível. CBio-GA, em contraste, gera um contrafactual que permanece consistente com as restrições do domínio, exigindo alterações apenas em mais uma variável do que o método anterior. Este caso ilustra a necessidade de avaliar tanto o desempenho da métrica quantitativa de esparsidade quanto sua acionabilidade.

Um padrão semelhante ocorre no Exemplo 2. Aqui, o método NICE-all obtém uma distância menor até a instância factual, sugerindo uma perturbação menor na instância original. No entanto, ele diminui novamente a idade, violando as restrições de imutabilidade e comprometendo a utilidade prática da explicação. O método CBio-GA produz um contrafactual com

Tabela 11 – Alguns exemplos mostrando contrafactuais produzidos pelo `CBio-GA` e pelo `NICE-all`. As variáveis modificadas estão destacadas. Valores em vermelho indicam que não são acionáveis.

Exemplo 1					
Algoritmo	**Limite_Crédito**	**Idade**	**Pag_0**	**Pag_2**	**Esparsidade**
Factual	500000	51	0	0	-
`NICE-all`	**380000**	**40**	0	0	2
`CBio-GA`	**430879**	51	**4**	**2**	3
Exemplo 2					
Algoritmo	**Limite_Crédito**	**Idade**	**Pag_0**	**Pag_2**	**Distância**
Factual	500000	51	0	0	-
`NICE-all`	500000	**50**	0	0	1
`CBio-GA`	500000	71	0	0	21

Fonte: Elaborada pelo autor.

uma distância maior, mas que permanece acionável e respeita tanto as direções causais quanto a imutabilidade. Esses exemplos mostram que métodos otimizados unicamente para proximidade ou esparsidade podem gerar contrafactuais que parecem numericamente atraentes, mas não atendem aos requisitos fundamentais de factibilidade, problemas que a acionabilidade expõe, mas não são capturados por métricas quantitativas tradicionais.

Para superar essa limitação, foi introduzido um novo método para calcular a acionabilidade, focado especificamente no realismo e na viabilidade de cenários contrafactuais (Seção 4.2). Embora as avaliações qualitativas frequentemente envolvam julgamento subjetivo, o método proposto para cálculo da métrica de acionabilidade formaliza essas considerações de maneira transparente e comparável, permitindo uma avaliação consistente entre diferentes métodos. A última coluna da Tabela 10 apresenta os valores obtidos com essa nova forma de avaliar a acionabilidade. O `CBio-GA` surge como o método com melhor desempenho, seguido por `NICE-all`. Um fator fundamental por trás do desempenho superior do `CBio-GA` é o uso explícito de restrições causais. Essa influência é reforçada pela melhoria substancial observada ao comparar o `C-DiCE` com sua versão original, o `DiCE`.

Além dos experimentos apresentados anteriormente, foram conduzidas análises adicionais com o objetivo de reforçar a suspeita de que métodos tradicionais, como o `NICE-all`, podem gerar contrafactuais com boas métricas quantitativas, porém sem acionabilidade. Ao analisar os resultados por meio de comparações semelhantes às da Tabela 11, foi possível identificar diferentes situações que evidenciam limitações das métricas tradicionais como, por exemplo, contrafactuais gerados pelo `NICE-all` ou pelo `CBio-GA` com baixa distância, mas ainda assim não acionáveis. Outras situações mostram casos com baixa esparsidade que, mesmo assim, violam restrições do domínio e até mesmo contrafactuais que simultaneamente apresentam baixa distância e baixa esparsidade, mas que não atendem aos critérios de acionabilidade. A

consolidação dessas análises resultou nos valores apresentados na Tabela 12, na qual são comparados apenas os algoritmos `CBio-GA` e `NICE-all`, uma vez que são os algoritmos com melhor desempenho significativo.

Essa tabela organiza os resultados da seguinte forma:

- Cada linha corresponde a um conjunto de dados, analisado individualmente;
- As colunas representam três cenários distintos, sempre contabilizando apenas o número de contrafactuais significativos não acionáveis, isto é, aqueles que violam pelo menos uma restrição de viabilidade;
- Em cada cenário, considera-se que o algoritmo produziu contrafactuais com determinada métrica quantitativa em nível *"baixo"*, mas ainda assim com baixa acionabilidade:
  - Baixa distância, porém baixa acionabilidade (segunda coluna)
  - Baixa esparsidade, porém baixa acionabilidade (terceira coluna)
  - Baixa distância + baixa esparsidade, porém baixa acionabilidade (quarta coluna)

Os valores aparecem no formato X/Y, em que `X` é a quantidade de contrafactuais não acionáveis produzidos pelo `NICE-all` e `Y` é a quantidade produzida pelo `CBio-GA`. O algoritmo destacado em negrito é aquele que produziu menos contrafactuais não acionáveis naquele cenário, indicando desempenho superior em termos de acionabilidade. Desta forma, notamos que para o conjunto de dados `Credit Card`, na coluna Distância, o `NICE-all` gerou 67 contrafactuais que, apesar de próximos ao factual, não são acionáveis. Enquanto isso, o `CBio-GA` não gerou nenhum contrafactual nessas condições.

Esse padrão se repete na maioria dos conjuntos de dados, ou seja, mesmo quando o `NICE-all` obtém métricas quantitativas mais favoráveis, como menor distância ou maior esparsidade, o `CBio-GA` produz contrafactuais mais acionáveis e realistas. Tal comportamento é consistente tanto por conjunto de dados quanto na análise agregada entre conjuntos de dados (Total Geral). Há exceções apenas na métrica de distância no conjunto de dados `Adult` e na métrica de esparsidade no conjunto de dados `Bank Marketing`. Uma análise mais detalhada desses resultados de acionabilidade é apresentada na seção a seguir.

5.2.3 Avaliação da Acionabilidade

Para refinar a avaliação da nova metodologia para calcular a acionabilidade, esta seção fornece uma análise detalhada, dividida por conjunto de dados e modelos de classificação, para oferecer uma compreensão mais precisa de como cada algoritmo de geração de contrafactuais se comporta em diferentes condições. Para simplificar a análise, apenas o `CBio-GA` e o `NICE-all` são incluídos nesta comparação, por serem os algoritmos que obtiveram mais

Tabela 12 – Comparação `NICE-all` vs `CBio-GA` por conjunto de dados, considerando distância, esparsidade e acionabilidade.

Conjunto de dados	**Distância**	**Esparsidade**	**Dist. e Espar.**
Credit Card	67/**0**	7/**0**	7/**0**
Churn	182/**0**	166/**0**	136/**0**
Compas	164/**24**	8/**2**	4/**0**
Adult	**47**/76	32/**7**	18/**2**
German Credit	89/**32**	51/**2**	21/**2**
Online Shoppers	68/**20**	57/**1**	32/**1**
Bank Marketing	35/**21**	**0**/7	**0/0**
Total Geral	652/**173**	321/**19**	218/**5**

Obs. Comparação entre algoritmos quanto à qualidade dos contrafactuais. Olhamos para menor distância, menor esparsidade e menor combinação distância e esparsidade (quando o algoritmo possui menor resultado em ambos os critérios), considerando apenas os casos em que o contrafactual gerado não é acionável, ou seja, há baixa acionabilidade. Os valores estão no formato X/Y, em que X = `NICE-all` e Y = `CBio-GA`. O algoritmo superior, ou seja, o que gera menor número de contrafactuais não acionáveis, é marcado em negrito.
Fonte: Elaborada pelo autor.

vitórias estatisticamente significativas do que derrotas, indicando desempenho competitivo no contexto da acionabilidade. Assim, os demais métodos são omitidos por não atingirem um nível de significância comparável e, portanto, não serem competitivos.

Detalhes da avaliação por conjunto de dados

A Tabela 13 apresenta uma comparação por conjunto de dados entre `CBio-GA` e `NICE-all`. Analisar o desempenho ao nível de conjunto de dados é importante porque a acionabilidade é fortemente influenciada por propriedades estruturais dos dados, como dimensionalidade e a proporção de atributos numéricos e categóricos, que podem favorecer um método em detrimento de outro. A tabela apresenta, para cada conjunto de dados, o número de instâncias ou linhas (2ª coluna), o total de atributos ou variáveis (3ª coluna) e a porcentagem de atributos numéricos e categóricos (4ª e 5ª colunas, respectivamente). `CBio-GA` e `NICE-all` respondem de maneiras distintas a conjuntos de dados com muitos atributos, forte presença de variáveis categóricas ou combinações heterogêneas de tipos de atributos. Os resultados podem ser visualizados nas três últimas colunas da tabela. As colunas `CBio-GA` e `NICE-all` indicam o número de vitórias obtidas pelos respectivos algoritmos considerando os quatro classificadores. A última coluna (`Empates`) representa o número de casos em que a diferença entre os dois métodos não é significativa. A soma de vitórias e empates corresponde ao número total de classificadores considerados.

Os resultados apresentados na Tabela 13 permitem responder a algumas questões de pesquisa, particularmente aquelas referentes ao contraste entre instâncias factuais e contrafactuais (QP2) e ao realismo e viabilidade das explicações geradas (QP3). O método `NICE-all`

Tabela 13 – Resultados da avaliação de acionabilidade por conjunto de dados. Colunas 2–5 mostram as principais características de cada conjunto de dados, seguido pelo número de vezes que o `CBio-GA` (6ª coluna) ou o `NICE-all` (7ª coluna) vencem, considerando os 4 classificadores, ou se há um empate entre eles (última coluna).

Conjunto de dados	**Linhas**	**Cols.**	**(%) Var. numéricas**	**(%) Var categóricas**	**Vitórias** `CBio-GA`	**Vitórias** `NICE-all`	**Empates**
Compas	7.214	53	38	62	2	0	2
Credit Card	30.000	24	100	0	2	1	1
Churn	3.333	21	81	19	2	1	1
German Credit	1.000	21	38	62	3	0	1
Online Shoppers	12.330	18	89	11	1	1	2
Bank Marketing	45.211	17	41	59	0	3	1
Adult	48.842	15	40	60	0	2	2

Fonte: Elaborada pelo autor.

tende a apresentar melhor desempenho em conjuntos de dados de baixa dimensionalidade e predominantemente numéricos, enquanto o `CBio-GA` demonstra resultados mais robustos em contextos onde as dependências estruturais entre as características parecem mais relevantes e onde a estrutura causal se mostra mais influente. Isso sugere que a escolha entre os métodos deve depender das características do conjunto de dados.

Detalhes da análise por classificador

Analisar o desempenho no nível do classificador é importante porque os contrafactuais podem ser influenciados pelas estruturas de fronteiras de decisão dos modelos preditivos subjacentes. Por exemplo, o KNN produz decisões baseadas em vizinhanças locais, geralmente gerando contrafactuais que seguem mais de perto a distribuição dos dados. As MLPs aprendem fronteiras não lineares e contínuas, permitindo que os algoritmos explorem mudanças graduais no espaço de atributos. O XGBoost cria regiões de decisão compactas e focadas em atributos específicos, que muitas vezes se alinham bem com dependências causais. Em contraste, Random Forests geram fronteiras mais fragmentadas devido às árvores treinadas de forma independente, o que faz com que pequenas alterações isoladas sejam suficientes para mudar a classe de previsão. Isso tende a beneficiar o `NICE-all`, que otimiza esparsidade e proximidade, mas pode ser limitante do ponto de vista da acionabilidade. Considerar essas características proporciona uma compreensão mais profunda de como a natureza do classificador interage com a geração de contrafactuais para cada método.

A Tabela 14 oferece uma visão complementar em relação à análise com base nos conjuntos de dados, ajudando a responder às questões QP2 e QP3. Ela mostra, para cada classificador (1ª coluna), quantas vezes o `CBio-GA` (2ª coluna) ou o `NICE-all` (3ª coluna) vence considerando os sete conjuntos de dados avaliados. Os nomes exatos dos conjuntos de dados e o método que obteve melhor desempenho são apresentados na última coluna. Observe que a soma

Tabela 14 – Resultados de acionabilidade por classificador. Colunas 2–4 mostram quantas vezes o `CBio-GA` ou o `NICE-all` venceram, considerando os 7 conjuntos de dados. A última coluna indica, por conjunto de dados, em qual algoritmo o desempenho foi superior.

Classificador	**Vitórias** `CBio-GA`	**Vitórias** `NICE-all`	**Empates**	**Onde cada algoritmo é melhor?**
KNN	3	0	4	`CBio-GA`: Churn, Credit Card e Online Shoppers
MLP	3	1	3	`CBio-GA`: Churn, Compas e German Credit; `NICE-all`: Adult
RF	1	3	3	`CBio-GA`: German Credit; `NICE-all`: Bank Marketing, Churn e Online Shoppers
XGB	3	2	2	`CBio-GA`: Compas, Credit Card e German Credit; `NICE-all`: Adult e Bank Marketing

Fonte: Elaborada pelo autor.

de vitórias e empates corresponde ao total de conjuntos de dados considerados na comparação (sete).

Uma possível explicação para as diferenças de desempenho observadas entre os classificadores é que o `CBio-GA` tende a se beneficiar mais de modelos cujo comportamento decisório reflete a estrutura dos dados de forma mais coerente. O KNN, por exemplo, baseia-se diretamente em vizinhanças locais do conjunto de dados, o que naturalmente leva à geração de contrafactuais que permanecem próximos de regiões densas e realistas. As MLPs aprendem relações contínuas e não lineares entre os atributos, de modo que os contrafactuais gerados sob esses modelos frequentemente seguem trajetórias consistentes no espaço de dados. O XGBoost, embora também baseado em árvores, difere do Random Forest ao utilizar boosting e regularização, produzindo padrões de decisão mais refinados e estáveis. Como resultado, os contrafactuais gerados pelo XGBoost tendem a respeitar melhor as dependências entre atributos quando comparados àqueles produzidos por árvores independentes.

Por outro lado, o `NICE-all` tende a favorecer modelos baseados em Random Forest, pois as métricas utilizadas em sua otimização (esparsidade e proximidade) recompensam contrafactuais que exigem poucas alterações nas variáveis. Como o Random Forest é composto por árvores com divisões simples e independentes, ele cria fronteiras de decisão em que pequenas mudanças em um ou dois atributos já são suficientes para alterar a classe, resultando em contrafactuais com baixa esparsidade. O XGBoost também se beneficia desse critério, pois, ao combinar boosting e regularização, concentra as decisões em atributos-chave, permitindo saltos curtos e direcionados para a mudança de classe. Entretanto, mesmo com essa vantagem, tal característica não é

suficiente para superar o `CBio-GA` na maioria dos conjuntos de dados.

5.3 Discussão

Os resultados apresentados nas seções anteriores permitem responder de forma sistemática às três questões de pesquisa definidas neste capítulo (QP1–QP3). A seguir, são discutidos os principais achados experimentais à luz de cada questão, integrando evidências quantitativas e qualitativas.

5.3.1 QP1 – Quão diferentes são os contrafactuais gerados de suas instâncias factuais correspondentes?

A QP1 investiga o grau de modificação necessário para transformar uma instância factual em um contrafactual válido, sendo operacionalizada principalmente por métricas quantitativas de distância (Euclidiana e Manhattan) e esparsidade. Os resultados evidenciam que o `NICE-all` alcança, consistentemente, os melhores valores nessas métricas, produzindo contrafactuais muito próximos do factual e com poucas alterações nos atributos. O `CBio-GA` apresenta desempenho competitivo, superando `CFNOW`, `DiCE` e `C-DiCE`, embora não atinja os mesmos níveis de proximidade e esparsidade do `NICE-all`.

Esses resultados indicam que, do ponto de vista puramente quantitativo, métodos orientados à substituição direta de atributos e à otimização local tendem a gerar contrafactuais menos distantes do factual. No entanto, como discutido nas seções subsequentes, essa proximidade nem sempre reflete modificações plausíveis ou realizáveis no mundo real. Assim, embora a QP1 seja respondida positivamente em favor do `NICE-all` no plano numérico, esses achados precisam ser interpretados em conjunto com critérios qualitativos.

5.3.2 QP2 – Quão realistas, viáveis e significativos são os contrafactuais dentro de seu contexto de domínio, e de que forma o método proposto para avaliar acionabilidade aprimora e torna mais robusta essa avaliação?

A QP2 desloca o foco da quantidade de mudança para a qualidade das mudanças, analisando se os contrafactuais respeitam restrições de imutabilidade, direcionalidade causal e viabilidade de domínio. Os exemplos qualitativos (Tabela 11) e a análise agregada apresentada na Tabela 12 evidenciam que contrafactuais com baixa distância e/ou esparsidade podem, ainda assim, ser não acionáveis.

Nesse contexto, a métrica de acionabilidade proposta revela diferenças substanciais entre os métodos. O `CBio-GA` gera significativamente menos contrafactuais não acionáveis, mesmo em cenários nos quais o `NICE-all` apresenta métricas quantitativas superiores. Esse padrão se mantém tanto na análise por conjunto de dados quanto na análise agregada, indicando que a

incorporação explícita de restrições causais direciona o processo de busca para soluções mais realistas.

Esses resultados evidenciam que métricas tradicionais são insuficientes para capturar a significância prática das explicações. A QP2 é, portanto, respondida ao demonstrar que o `CBio-GA` produz contrafactuais mais viáveis e acionáveis, enquanto métodos otimizados apenas para fatores quantitativos (proximidade e/ou esparsidade) tendem a gerar soluções que violam restrições fundamentais do domínio.

5.3.3 QP3 – Qual é o impacto da inferência causal nas métricas quantitativas e na qualidade geral dos contrafactuais?

A QP3 avalia diretamente o efeito da incorporação de inferência causal na geração e avaliação de contrafactuais. Os resultados evidenciam que métodos de geração de contrafactuais que não incorporam nenhuma noção de causalidade, como o `NICE-all`, apresentam melhores resultados em métricas puramente quantitativas, como distância e esparsidade. No entanto, quando a qualidade geral das explicações é avaliada por meio da acionabilidade, o efeito da causalidade torna-se claramente positivo.

Isso é evidenciado pela comparação entre `DiCE` e `C-DiCE`, na qual a adição de restrições causais resulta em um aumento expressivo na proporção de contrafactuais acionáveis. O mesmo padrão é observado ao comparar `CBio-GA` com métodos não causais: o uso explícito de conhecimento causal melhora substancialmente o realismo e a viabilidade das soluções geradas.

O `CBio-GA` demonstra uma clara vantagem na geração de contrafactuais acionáveis, particularmente em conjuntos de dados com muitos atributos ou dominados por variáveis categóricas. Nesses cenários, a estrutura causal desempenha um papel crucial na orientação do processo de busca, levando a soluções que refletem melhor as mudanças realistas. O `NICE-all`, em contraste, apresenta bom desempenho em conjuntos de dados de baixa dimensionalidade ou predominantemente numéricos, em que a obtenção de mudanças de classe requer menos ajustes estruturais.

Além disso, a análise por classificador indica que o impacto da causalidade depende da geometria das fronteiras de decisão do modelo preditivo. Classificadores como KNN, MLP e XGBoost tendem a se beneficiar mais de abordagens causais, enquanto modelos baseados em Random Forest favorecem métodos orientados à esparsidade local. Esses achados reforçam que o valor da inferência causal está fortemente ligado ao contexto estrutural dos dados e do classificador utilizado.

5.3.4 Síntese Geral

Em conjunto, os resultados evidenciam haver um *trade-off* fundamental entre proximidade numérica e acionabilidade. Métodos não causais, como o `NICE-all`, destacam-se em

métricas quantitativas clássicas, enquanto métodos que incorporam causalidade, como o `CBio-GA`, produzem contrafactuais mais alinhados com restrições reais. A métrica de acionabilidade proposta desempenha um papel central ao revelar limitações ocultas das métricas tradicionais e ao permitir uma avaliação mais fiel da utilidade prática das explicações. Um resumo dessas conclusões pode ser encontrado na Tabela 15.

Tabela 15 – Síntese dos resultados experimentais em relação às questões de pesquisa.

QP	Foco	Métricas / Evidências	Principais Resultados	Conclusão
QP1	Métricas quantitativas e grau de modificações para inverter a classe	Distâncias (Euclidiana e Manhattan), Esparsidade, Tempo (Tab. 10)	`NICE-all` mostra menores distâncias e esparsidade; `CBio-GA` é competitivo e supera `CFNOW`, `DiCE` e `C-DiCE`	Em termos de métricas quantitativas, é possível obter-se bons resultados sem utilizar causalidade
QP2	Viabilidade e realismo prático	Acionabilidade, exemplos qualitativos (Tab. 11), contagem de inviáveis (Tab. 12)	`NICE-all` viola condições de mutabilidade e direcionalidade; `CBio-GA` gera mais contrafactuais acionáveis	Proximidade e esparsidade não garantem acionabilidade, portanto as métricas quantitativas e qualitativas devem ser avaliadas em conjunto
QP3	Impacto da inferência causal	Comparações causais e não causais (Tabs. 13 e 14)	Causalidade melhora substancialmente a acionabilidade, inclusive ao acrescentá-la em algoritmos da literatura como o `DiCE`; impacto varia por conjunto de dados e classificador	Causalidade é decisiva para a qualidade prática dos contrafactuais

Fonte: Elaborada pelo autor.

Em termos quantitativos, observa-se um aumento de 71,4% na proporção de contrafactuais acionáveis produzidos pelo `CBio-GA` em comparação ao `DiCE`, representando o maior ganho estatisticamente significativo entre os métodos avaliados. Em relação ao `NICE-all`, o sistema apresenta uma melhoria de 15,2%. Além disso, a incorporação de conhecimento causal ao `DiCE` resulta em um incremento de 39,1% na geração de contrafactuais acionáveis quando comparado à versão original do algoritmo.

Em geral, os resultados evidenciam que os métodos que consideram a causalidade oferecem benefícios substanciais em cenários em que as restrições de domínio são importantes, e que a acionabilidade é uma dimensão crítica para avaliar explicações contrafactuais. Essas constatações reforçam a necessidade de métodos de avaliação que considerem não apenas o quanto um contrafactual altera a entrada, mas também se essas alterações fazem sentido no mundo real.

6 CONCLUSÃO

Este trabalho introduziu o CausalBioCF, um método de geração de contrafactuais guiado por causalidade que integra obtenção de conhecimento do domínio e otimização bioinspirada para aprimorar o realismo e a viabilidade das explicações contrafactuais. Além disso, foi proposta uma forma sistemática de calcular a métrica de acionabilidade que avalia os contrafactuais para além das medidas quantitativas tradicionais, incorporando explicitamente as avaliações de restrições de mutabilidade e a direção da mudança. Juntas, essas contribuições abordam duas lacunas significativas na pesquisa existente sobre explicabilidade contrafactual: a falta de fundamentação causal nos métodos de geração e a ausência de uma maneira fundamentada de medir a acionabilidade dos contrafactuais gerados.

A avaliação experimental do trabalho foi realizada utilizando sete conjuntos de dados de referência e quatro modelos de classificação mostrando que o CausalBioCF é competitivo em métricas quantitativas, enquanto supera claramente as abordagens existentes em termos de acionabilidade. Ao incorporar conhecimento causal extraído automaticamente dos dados, o método evita mudanças não realísticas, como modificar atributos imutáveis ou propor alterações incompatíveis com o domínio, produzindo contrafactuais que permanecem válidos e acionáveis na prática.

Além disso, o método proposto revela limitações importantes dos algoritmos atuais. Abordagens que se destacam em métricas de proximidade ou esparsidade podem gerar explicações inviáveis em cenários reais justamente por ignorarem dependências causais entre variáveis. A inclusão explícita da causalidade no processo decisório do algoritmo genético demonstrou reduzir substancialmente esse problema. Ao avaliar também a acionabilidade, a métrica proposta fornece uma análise mais completa e confiável da qualidade contrafactual, capturando dimensões da explicação que as métricas tradicionais não conseguem medir.

Assim, este trabalho destaca o valor da combinação do raciocínio causal com a geração de contrafactuais baseada em otimização e impulsiona o desenvolvimento de explicações acionáveis, significativas e confiáveis para modelos de aprendizado de máquina.

6.1 Trabalhos Futuros

Apesar das contribuições citadas, ainda existem diversas possibilidades para trabalhos futuros. Dentre elas, primeiro, pode-se estender o CausalBioCF para classificação multiclasse, requerendo a adaptação do processo de inferência causal para lidar com múltiplos grupos de tratamento, representando um passo importante e complexo para uma aplicabilidade mais ampla. Segundo, a validação dos contrafactuais gerados com especialistas da área proporcionaria uma

avaliação mais completa de sua utilidade e garantiria o alinhamento com os processos de tomada de decisão do mundo real. Pesquisas futuras poderiam explorar a integração de técnicas de descoberta causal para inferir automaticamente a estrutura causal quando o conhecimento da área for escasso, bem como aprofundar a análise sobre como diferentes geometrias de classificadores influenciam a viabilidade dos contrafactuais. Além disso, uma comparação entre o CausalBioCF e métodos que utilizam causalidade por métodos dependentes de conhecimento de especialistas de domínio poderia ser proposta a fim de verificar se há benefícios nesse tipo de abordagem menos automatizada.

Para a métrica de acionabilidade proposta, poderia haver uma ampliação da metodologia de avaliação para considerar interações entre atributos, uma vez que, na forma atual, cada variável é analisada de maneira independente. Essa ampliação permitiria incorporar relações entre múltiplas variáveis ou grupos de variáveis, aspecto especialmente relevante em domínios mais complexos. Além disso, a abordagem aplicada às variáveis categóricas pode ser aprimorada, pois não explora possíveis estruturas internas, como ordenações ou hierarquias. Investigar formas mais robustas de avaliar mudanças em atributos categóricos poderia tornar o processo de geração e validação de contrafactuais ainda mais realista e aplicável a cenários de maior diversidade.

Por fim, abre-se espaço para expandir o estudo acerca da robustez da métrica de acionabilidade proposta diante de sua dependência de uma base de conhecimento construída por especialistas de domínio. Deste modo, há a possibilidade de conduzir experimentos adicionais em ambiente controlado, avaliando de forma sistemática como diferentes tipos de perturbações na base (por exemplo, erros na definição de mutabilidade e direções de alteração) afetam o valor final da métrica. Tal investigação permitiria caracterizar a sensibilidade da métrica a imprecisões humanas e, consequentemente, reforçar sua confiabilidade e aplicabilidade em cenários reais, podendo ser realizada com a inserção de mais conjuntos de dados nas avaliações.

Referências

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, IEEE, v. 6, p. 52138–52160, 2018. Citado 2 vezes nas páginas 20 e 21.

ALBINI, E. et al. Counterfactual shapley additive explanations. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. [S.l.: s.n.], 2022. p. 1054–1070. Citado 2 vezes nas páginas 36 e 37.

AUQUI, J. A. O. et al. Machine learning for personal credit evaluation: A systematic review. *WSEAS Transactions on Computer Research*, World Scientific and Engineering Academy and Society, 2022. Citado na página 16.

BECKER, B.; KOHAVI, R. *Adult*. 1996. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20. Citado 3 vezes nas páginas 41, 42 e 61.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 57, n. 1, p. 289–300, 1995. Citado na página 64.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *The journal of machine learning research*, JMLR. org, v. 13, n. 1, p. 281–305, 2012. Citado na página 60.

BLACK, P. E. *Dictionary of algorithms and data structures*. [S.l.]: Paul E. Black, 1998. Citado na página 24.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 59.

BRUGHMANS, D.; LEYMAN, P.; MARTENS, D. Nice: an algorithm for nearest instance counterfactual explanations. *Data mining and knowledge discovery*, Springer, v. 38, n. 5, p. 2665–2703, 2024. Citado 6 vezes nas páginas 24, 32, 34, 36, 62 e 64.

BYRNE, R. M. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: *IJCAI*. [S.l.: s.n.], 2019. p. 6276–6282. Citado na página 22.

CALIENDO, M.; KOPEINIG, S. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, Wiley Online Library, v. 22, n. 1, p. 31–72, 2008. Citado 3 vezes nas páginas 29, 42 e 43.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794. Citado na página 59.

CHEN, Z. et al. Relax: Reinforcement learning agent explainer for arbitrary predictive models. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. [S.l.: s.n.], 2022. p. 252–261. Citado na página 36.

COHEN, J. *Statistical power analysis for the behavioral sciences*. [S.l.]: routledge, 2013. Citado na página 44.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado 3 vezes nas páginas 29, 43 e 59.

DANDL, S. et al. Countarfactuals–generating plausible model-agnostic counterfactual explanations with adversarial random forests. In: SPRINGER. *World Conference on Explainable Artificial Intelligence*. [S.l.], 2024. p. 85–107. Citado 2 vezes nas páginas 36 e 37.

DANIELSSON, P.-E. Euclidean distance mapping. *Computer Graphics and image processing*, Elsevier, v. 14, n. 3, p. 227–248, 1980. Citado 2 vezes nas páginas 23 e 47.

de Oliveira, R. M. B.; SöRENSEN, K.; MARTENS, D. A model-agnostic and data-independent tabu search algorithm to generate counterfactuals for tabular, image, and text data. *European Journal of Operational Research*, 2023. ISSN 0377-2217. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377221723006598>. Citado 2 vezes nas páginas 32 e 63.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, IEEE, v. 6, n. 2, p. 182–197, 2002. Citado na página 48.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, v. 7, n. Jan, p. 1–30, 2006. Citado na página 64.

DOMNICH, M.; VICENTE, R. Enhancing counterfactual explanation search with diffusion distance and directional coherence. In: SPRINGER. *World Conference on Explainable Artificial Intelligence*. [S.l.], 2024. p. 60–84. Citado na página 36.

DU, M.; LIU, N.; HU, X. Techniques for interpretable machine learning. *Communications of the ACM*, ACM New York, NY, USA, v. 63, n. 1, p. 68–77, 2019. Citado na página 21.

DUONG, T. D.; LI, Q.; XU, G. Ceflow: A robust and efficient counterfactual explanation framework for tabular data using normalizing flows. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2023. p. 133–144. Citado na página 36.

FATIMA, M.; PASHA, M. et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, Scientific Research Publishing, v. 9, n. 01, p. 1, 2017. Citado 2 vezes nas páginas 16 e 31.

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, v. 11, n. 1, p. 86–92, 1940. Citado na página 64.

GARCÍA, S. et al. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information sciences*, Elsevier, v. 180, n. 10, p. 2044–2064, 2010. Citado na página 64.

GLOVER, F. Tabu search—part i. *ORSA Journal on computing*, Informs, v. 1, n. 3, p. 190–206, 1989. Citado na página 34.

GOLBERG, D. E. Genetic algorithms in search, optimization, and machine learning. *Addion wesley*, v. 1989, n. 102, p. 36, 1989. Citado 3 vezes nas páginas 18, 40 e 46.

GOMEZ-URIBE, C. A.; HUNT, N. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, ACM New York, NY, USA, v. 6, n. 4, p. 1–19, 2015. Citado na página 16.

GOODFELLOW, I. et al. Generative adversarial networks. *Communications of the ACM*, ACM New York, NY, USA, v. 63, n. 11, p. 139–144, 2020. Citado na página 33.

GUIDOTTI, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, Springer, v. 38, n. 5, p. 2770–2824, 2024. Citado 6 vezes nas páginas 16, 24, 25, 31, 32 e 35.

HOFMANN, H. *Statlog (German Credit Data)*. 1994. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77. Citado na página 61.

HOLLAND, J. H. Genetic algorithms. *Scientific american*, JSTOR, v. 267, n. 1, p. 66–73, 1992. Citado 3 vezes nas páginas 18, 40 e 46.

HOLLAND, P. W. Statistics and causal inference. *Journal of the American statistical Association*, Taylor & Francis, v. 81, n. 396, p. 945–960, 1986. Citado na página 26.

JI, J. et al. Unified counterfactual explanation framework for black-box models. In: SPRINGER. *Pacific Rim International Conference on Artificial Intelligence*. [S.l.], 2023. p. 422–433. Citado na página 36.

JIN, W. Research on machine learning and its algorithms and development. *Journal of Physics: Conference Series*, IOP Publishing, v. 1544, n. 1, p. 012003, may 2020. Disponível em: <https://dx.doi.org/10.1088/1742-6596/1544/1/012003>. Citado na página 16.

KAGGLE. *Titanic - Machine Learning from Disaster*. 2012. <https://www.kaggle.com/competitions/titanic>. Accessed: 2024-10-04. Citado na página 62.

KAGGLE. *Telco Customer Churn*. 2020. <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>. Accessed: 2024-10-04. Citado na página 61.

KAHN, M. *Diabetes*. 1994. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G. Citado na página 62.

KAHNEMAN, D.; MILLER, D. T. Norm theory: Comparing reality to its alternatives. *Psychological review*, American Psychological Association, v. 93, n. 2, p. 136, 1986. Citado na página 22.

KANAMORI, K. et al. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. In: *IJCAI*. [S.l.: s.n.], 2020. p. 2855–2862. Citado na página 36.

KANE, L. T. et al. Propensity score matching: a statistical method. *Clinical spine surgery*, LWW, v. 33, n. 3, p. 120–122, 2020. Citado na página 43.

KARIMI, A.-H. et al. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 5, p. 1–29, 2022. Citado na página 23.

KARIMI, A.-H. et al. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, v. 33, p. 265–277, 2020. Citado na página 22.

KINGMA, D. P.; WELLING, M. et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, Now Publishers, Inc., v. 12, n. 4, p. 307–392, 2019. Citado na página 33.

KLINE, A.; LUO, Y. PsmPy: a package for retrospective cohort matching in Python. In: IEEE. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. [S.l.], 2022. p. 1354–1357. Citado 2 vezes nas páginas 40 e 44.

KURATOMI, A. et al. JUICE: Justified counterfactual explanations. In: SPRINGER. *Discovery Science: 25th International Conference, DS 2022, Montpellier, France, October 10–12, 2022, Proceedings*. [S.l.], 2022. p. 493–508. Citado 2 vezes nas páginas 24 e 36.

LE, T.; WANG, S.; LEE, D. GRACE: Generating concise and informative contrastive sample to explain neural network model's prediction. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. [S.l.: s.n.], 2020. p. 238–248. Citado na página 33.

LIPTON, P. Contrastive explanation. *Royal Institute of Philosophy Supplements*, Cambridge University Press, v. 27, p. 247–266, 1990. Citado na página 22.

LOOVEREN, A. V.; KLAISE, J. Interpretable counterfactual explanations guided by prototypes. In: SPRINGER. *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*. [S.l.], 2021. p. 650–665. Citado na página 16.

MELNYCHUK, V.; FRAUEN, D.; FEUERRIEGEL, S. Causal transformer for estimating counterfactual outcomes. In: PMLR. *International conference on machine learning*. [S.l.], 2022. p. 15293–15329. Citado na página 33.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, Elsevier, v. 267, p. 1–38, 2019. Citado 2 vezes nas páginas 20 e 21.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. [S.l.]: Saraiva Educação SA, 2017. Citado na página 26.

MORO, S.; RITA, P.; CORTEZ, P. *Bank Marketing [Dataset]*. 2014. UCI Machine Learning Repository. Disponível em: <https://archive.ics.uci.edu/dataset/222/bank+marketing>. Citado na página 61.

MOTHILAL, R. K.; SHARMA, A.; TAN, C. Explaining machine learning classifiers through diverse counterfactual explanations. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. [S.l.: s.n.], 2020. p. 607–617. Citado 7 vezes nas páginas 17, 24, 32, 34, 36, 37 e 62.

NEMIROVSKY, D. et al. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In: PMLR. *Uncertainty in Artificial Intelligence*. [S.l.], 2022. p. 1488–1497. Citado 4 vezes nas páginas 18, 32, 33 e 36.

NOGUEIRA, A. R. et al. Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, Wiley Online Library, v. 12, n. 2, p. e1449, 2022. Citado 3 vezes nas páginas 26, 27 e 33.

OLIVEIRA, R. M. B. de; MARTENS, D. A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences*, MDPI, v. 11, n. 16, p. 7274, 2021. Citado na página 24.

OLIVEIRA, R. M. B. de; SÖRENSEN, K.; MARTENS, D. A model-agnostic and data-independent tabu search algorithm to generate counterfactuals for tabular, image, and text data. *European Journal of Operational Research*, Elsevier, v. 317, n. 2, p. 286–302, 2024. Citado 2 vezes nas páginas 34 e 36.

O'BRIEN, A.; KIM, E.; WEBER, R. Investigating causally augmented sparse learning as a tool for meaningful classification. In: IEEE. *2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. [S.l.], 2023. p. 33–37. Citado 2 vezes nas páginas 22 e 33.

PANAGIOTOU, E. et al. Tabcf: Counterfactual explanations for tabular data using a transformer-based vae. In: *Proceedings of the 5th ACM International Conference on AI in Finance*. [S.l.: s.n.], 2024. p. 274–282. Citado 2 vezes nas páginas 36 e 37.

PASCUAL-TRIANA, J. D. et al. Overlap number of balls model-agnostic counterfactuals (ONB-MACF): A data-morphology-based counterfactual generation method for trustworthy artificial intelligence. *Information Sciences*, Elsevier, v. 701, p. 121844, 2025. Citado 5 vezes nas páginas 25, 36, 37, 51 e 64.

PAWELCZYK, M.; BROELEMANN, K.; KASNECI, G. Learning model-agnostic counterfactual explanations for tabular data. In: *Proceedings of The Web Conference 2020*. [S.l.: s.n.], 2020. p. 3126–3132. Citado 2 vezes nas páginas 33 e 36.

PAWELCZYK, M.; BROELEMANN, K.; KASNECI, G. On counterfactual explanations under predictive multiplicity. In: PMLR. *Conference on Uncertainty in Artificial Intelligence*. [S.l.], 2020. p. 809–818. Citado na página 32.

PEARCE, N.; LAWLOR, D. A. Causal inference—so much more than statistics. *International journal of epidemiology*, Oxford University Press, v. 45, n. 6, p. 1895–1903, 2016. Citado 2 vezes nas páginas 26 e 33.

PEARL, J. *Causality*. [S.l.]: Cambridge university press, 2009. Citado 5 vezes nas páginas 26, 27, 28, 33 e 42.

PEARL, J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, ACM New York, NY, USA, v. 62, n. 3, p. 54–60, 2019. Citado na página 28.

PIAGGESI, S. et al. Counterfactual and prototypical explanations for tabular data via interpretable latent space. *IEEE Access*, IEEE, 2024. Citado na página 36.

POYIADZI, R. et al. FACE: feasible and actionable counterfactual explanations. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. [S.l.: s.n.], 2020. p. 344–350. Citado 2 vezes nas páginas 32 e 33.

PROPUBLICA. *COMPAS Recidivism Risk Score Data and Analysis*. 2016. <https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis>. Accessed: 2024-10-04. Citado na página 61.

RASOULI, P.; YU, I. C. Analyzing and improving the robustness of tabular classifiers using counterfactual explanations. In: IEEE. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.], 2021. p. 1286–1293. Citado na página 36.

RASOULI, P.; YU, I. C. CARE: Coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, Springer, p. 1–26, 2022. Citado 2 vezes nas páginas 36 e 37.

REDELMEIER, A. et al. MCCE: Monte carlo sampling of valid and realistic counterfactual explanations for tabular data. *Data Mining and Knowledge Discovery*, Springer, v. 38, n. 4, p. 1830–1861, 2024. Citado 2 vezes nas páginas 36 e 37.

ROSENBAUM, P. R.; RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, Oxford University Press, v. 70, n. 1, p. 41–55, 1983. Citado 2 vezes nas páginas 29 e 42.

ROSENBAUM, P. R.; RUBIN, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, v. 39, n. 1, p. 33–38, 1985. Citado 2 vezes nas páginas 29 e 43.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958. Citado na página 60.

RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, American Psychological Association, v. 66, n. 5, p. 688, 1974. Citado na página 28.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, Nature Publishing Group UK London, v. 1, n. 5, p. 206–215, 2019. Citado na página 20.

SAKAR, C.; KASTRO, Y. *Online Shoppers Purchasing Intention Dataset*. 2018. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5F88Q. Citado na página 61.

SCHLEICH, M. et al. Geco: Quality counterfactual explanations in real time. *arXiv preprint arXiv:2101.01292*, 2021. Citado 2 vezes nas páginas 36 e 37.

SEAMAN, S. R.; WHITE, I. R. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 22, n. 3, p. 278–295, 2013. Citado 2 vezes nas páginas 29 e 42.

SHANI, G.; GUNAWARDANA, A. Evaluating recommendation systems. *Recommender systems handbook*, Springer, p. 257–297, 2011. Citado na página 16.

SHAO, X. et al. Cube: Causal intervention-based counterfactual explanation for prediction models. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 36, n. 6, p. 2416–2429, 2023. Citado 7 vezes nas páginas 31, 33, 34, 35, 37, 38 e 45.

SHARMA, S.; HENDERSON, J.; GHOSH, J. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. [S.l.: s.n.], 2020. p. 166–172. Citado 3 vezes nas páginas 22, 23 e 32.

SILVA, A. A. M. d. *Introdução à inferência causal em epidemiologia: uma abordagem gráfica e contrafatual*. Rio de Janeiro: Editora Fiocruz, 2021. Citado 2 vezes nas páginas 26 e 27.

SPIRTES, P. Introduction to causal inference. *Journal of Machine Learning Research*, v. 11, n. 5, 2010. Citado na página 27.

SULAIMAN, R. B.; SCHETININ, V.; SANT, P. Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, Springer, v. 2, n. 1-2, p. 55–68, 2022. Citado 2 vezes nas páginas 16 e 31.

TENNANT, P. W. et al. Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, Oxford University Press, v. 50, n. 2, p. 620–632, 2021. Citado na página 27.

TUKEY, J. W. et al. *Exploratory data analysis*. [S.l.]: Reading, MA, 1977. v. 2. Citado na página 25.

VERMA, S. et al. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, ACM New York, NY, v. 56, n. 12, p. 1–42, 2024. Citado 6 vezes nas páginas 17, 25, 31, 33, 35 e 61.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, HeinOnline, v. 31, p. 841, 2017. Citado 4 vezes nas páginas 16, 22, 23 e 31.

WANG, Y. et al. The skyline of counterfactual explanations for machine learning decision models. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. [S.l.: s.n.], 2021. p. 2030–2039. Citado 2 vezes nas páginas 36 e 37.

WIEGERINCK, W.; KAPPEN, B.; BURGERS, W. Bayesian networks for expert systems: Theory and practical applications. *Interactive collaborative information systems*, Springer, p. 547–578, 2010. Citado na página 28.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80–83, 1945. Citado na página 64.

YANG, F. et al. Model-based counterfactual synthesizer for interpretation. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. [S.l.: s.n.], 2021. p. 1964–1974. Citado 2 vezes nas páginas 17 e 33.

YEH, I.-C. *Default of Credit Card Clients*. 2009. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H. Citado na página 61.

ZHANG, W.; BARR, B.; PAISLEY, J. An interpretable deep classifier for counterfactual generation. In: *Proceedings of the Third ACM International Conference on AI in Finance*. [S.l.: s.n.], 2022. p. 36–43. Citado 3 vezes nas páginas 18, 33 e 36.

APÊNDICE A – Bases de conhecimento utilizadas nos testes

Este apêndice apresenta as bases de conhecimento utilizadas nos experimentos, de modo a garantir a reprodutibilidade completa das validações realizadas e a transparência dos procedimentos adotados no cálculo da métrica de acionabilidade.

Tabela 16 – Base de conhecimento de domínio utilizada para o conjunto de dados *Credit Card Default*.

Variável	**Mutável**	**Direção de Alteração**
LIMIT_BAL	Sim	Ambos
SEX	Não	–
EDUCATION	Sim	Ambos
MARRIAGE	Sim	Ambos
AGE	Sim	Aumento
PAY_0	Sim	Ambos
PAY_2	Sim	Ambos
PAY_3	Sim	Ambos
PAY_4	Sim	Ambos
PAY_5	Sim	Ambos
PAY_6	Sim	Ambos
BILL_AMT1	Sim	Ambos
BILL_AMT2	Sim	Ambos
BILL_AMT3	Sim	Ambos
BILL_AMT4	Sim	Ambos
BILL_AMT5	Sim	Ambos
BILL_AMT6	Sim	Ambos
PAY_AMT1	Sim	Ambos
PAY_AMT2	Sim	Ambos
PAY_AMT3	Sim	Ambos
PAY_AMT4	Sim	Ambos
PAY_AMT5	Sim	Ambos
PAY_AMT6	Sim	Ambos

Fonte: Elaborada pelo autor.

Tabela 17 – Base de conhecimento de domínio utilizada para o conjunto de dados *Adult*.

Variável	**Mutável**	**Direção de Alteração**
age	Sim	Aumento
workclass	Sim	Ambos
fnlwgt	Não	–
education	Sim	Ambos
education-num	Sim	Aumento
marital-status	Sim	Ambos
occupation	Sim	Ambos
relationship	Sim	Ambos
race	Não	–
sex	Não	–
capital-gain	Sim	Ambos
capital-loss	Sim	Ambos
hours-per-week	Sim	Ambos
native-country	Não	–

Fonte: Elaborada pelo autor.

Tabela 18 – Base de conhecimento de domínio utilizada para o conjunto de dados *Churn*.

Variável	**Mutável**	**Direção de Alteração**
state	Não	–
account length	Sim	Aumento
area code	Não	–
phone number	Não	–
international plan	Sim	Ambos
voice mail plan	Sim	Ambos
number vmail messages	Sim	Ambos
total day minutes	Sim	Ambos
total day calls	Sim	Ambos
total day charge	Sim	Ambos
total eve minutes	Sim	Ambos
total eve calls	Sim	Ambos
total eve charge	Sim	Ambos
total night minutes	Sim	Ambos
total night calls	Sim	Ambos
total night charge	Sim	Ambos
total intl minutes	Sim	Ambos
total intl calls	Sim	Ambos
total intl charge	Sim	Ambos
customer service calls	Sim	Ambos

Fonte: Elaborada pelo autor.

Tabela 19 – Base de conhecimento de domínio utilizada para o conjunto de dados *Compas Scores*.

Variável	**Mutável**	**Direção**
id	Não	–
name	Não	–
first	Não	–
last	Não	–
compas-screening-date	Não	–
sex	Não	–
dob	Não	–
age	Sim	Aumento
age-cat	Não	–
race	Não	–
juv-fel-count	Sim	Aumento
decile-score	Sim	Ambos
juv-misd-count	Sim	Aumento
juv-other-count	Sim	Aumento
priors-count	Sim	Aumento
days-b-screening-arrest	Sim	Ambos
c-jail-in	Não	–
c-jail-out	Não	–
c-case-number	Não	–
c-offense-date	Não	–
c-arrest-date	Não	–
c-days-from-compas	Não	–
c-charge-degree	Sim	Ambos
c-charge-desc	Sim	Ambos
is-recid	Não	–
r-case-number	Não	–
r-charge-degree	Sim	Ambos
r-days-from-arrest	Sim	Ambos
r-offense-date	Não	–
r-charge-desc	Sim	Ambos
r-jail-in	Não	–
r-jail-out	Não	–
violent-recid	Não	–
is-violent-recid	Não	–
vr-case-number	Não	–
vr-charge-degree	Sim	Ambos
vr-offense-date	Não	–
vr-charge-desc	Sim	Ambos
type-of-assessment	Não	–
score-text	Sim	Ambos
screening-date	Não	–
v-type-of-assessment	Não	–
v-decile-score	Sim	Ambos
v-score-text	Sim	Ambos
v-screening-date	Não	–
in-custody	Não	–
out-custody	Não	–
priors-count	Sim	Aumento
start	Não	–
end	Não	–
event	Não	–

Fonte: Elaborada pelo autor.

Tabela 20 – Base de conhecimento de domínio utilizada para o conjunto de dados *Online Shoppers Purchasing Intention*.

Variável	**Mutável**	**Direção de Alteração**
Administrative	Sim	Ambos
Administrative-Duration	Sim	Ambos
Informational	Sim	Ambos
Informational-Duration	Sim	Ambos
ProductRelated	Sim	Ambos
ProductRelated-Duration	Sim	Ambos
BounceRates	Sim	Ambos
ExitRates	Sim	Ambos
PageValues	Sim	Ambos
SpecialDay	Sim	Ambos
Month	Não	–
OperatingSystems	Não	–
Browser	Não	–
Region	Não	–
TrafficType	Não	–
VisitorType	Não	–
Weekend	Não	–

Fonte: Elaborada pelo autor.

Tabela 21 – Base de conhecimento de domínio utilizada para o conjunto de dados *Bank Marketing*.

Variável	**Mutável**	**Direção de Alteração**
age	Sim	Aumento
job	Sim	Ambos
marital	Sim	Ambos
education	Sim	Aumento
default	Sim	Ambos
balance	Sim	Ambos
housing	Sim	Ambos
loan	Sim	Ambos
contact	Não	–
day	Não	–
month	Não	–
duration	Sim	Ambos
campaign	Não	–
pdays	Não	–
previous	Não	–
poutcome	Não	–

Fonte: Elaborada pelo autor.

Tabela 22 – Base de conhecimento de domínio utilizada para o conjunto de dados *German Credit*.

Variável	**Mutável**	**Direção de Alteração**
account-check-status	Sim	Ambos
duration-in-month	Sim	Aumento
credit-history	Sim	Ambos
purpose	Não	–
credit-amount	Sim	Ambos
savings	Sim	Aumento
present-emp-since	Sim	Aumento
installment-as-income-perc	Sim	Ambos
personal-status-sex	Não	–
other-debtors	Não	–
present-res-since	Sim	Aumento
property	Sim	Ambos
age	Sim	Aumento
other-installment-plans	Sim	Ambos
housing	Sim	Ambos
credits-this-bank	Não	–
job	Não	–
people-under-maintenance	Não	–
telephone	Sim	Ambos
foreign-worker	Não	–

Fonte: Elaborada pelo autor.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	Gabriel Covello Furlanetto Data nasc. 24/04/1992
Nacionalidade	Brasileiro
Nome em citações bibliográficas	Furlanetto, Gabriel Furlanetto, GC.
Currículo Lattes	http://lattes.cnpq.br/9697669047428418
ORCID	https://orcid.org/0000-0003-2917-9182
FORMAÇÃO ACADÊMICA	
2010/2013	Bacharel em Ciência da Computação UNESP/IBILCE
2014/2016	Mestrado em Ciência da Computação UNESP/IBILCE
2017/2018	Especialização em Tecnologias e Inovações para Web Centro Universitário Senac
2021/2026	Doutorado em Ciência da Computação UNESP/IGCE
PRODUÇÃO BIBLIOGRÁFICA	
FURLANETTO, Gabriel Covello; BALDASSIN, Alexandro; MANACERO, Aleardo. CausalBioCF: Causal Counterfactuals for Machine Learning Interpretability. In: International Conference on Computational Science and Its Applications. Cham: Springer Nature Switzerland, 2025. p. 200-217.	
FURLANETTO, Gabriel Covello; BALDASSIN, Alexandro; MANACERO, Aleardo. CausalBioCF: Causalidade e otimização bioinspirada para geração de contrafactuais factíveis em tempo real. In: Escola Regional de Alto Desempenho de São Paulo (ERAD-SP). SBC, 2024. p. 77-80.	
FURLANETTO, Gabriel Covello; GOMES, Vitoria Zanon; BREVE, Fabricio Aparecido. Artificial Bee Colony Algorithm for Feature Selection in Fraud Detection Process. In: International Conference on Computational Science and Its Applications. Cham: Springer Nature Switzerland, 2023. p. 535-549.	
FURLANETTO, Gabriel C. et al. A tool for model conversion between simulators of grid computing. In: SpringSim (ANSS). 2015. p. 9-16.	
PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS	
ANNUAL SIMULATION SYMPOSIUM (ANSS), 15., 2015, (Washington, Estados Unidos). Título do trabalho apresentado: A tool for model conversion between simulators of grid computing. 2015. (Seminário).	
ESCOLA REGIONAL DE ALTO DESEMPENHO, 15., 2024, (Rio Claro). Título do trabalho apresentado: Causalidade e otimização bioinspirada para geração de contrafactuais factíveis em tempo real. 2024. (Seminário).	
INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND ITS APPLICATION, 25., 2025, (Istambul, Turquia). Título do trabalho apresentado: Causal Counterfactuals for Machine Learning Interpretability. 2025. (Seminário).	