Evaluation of methods of counterfactual explanation - A qualitative and quantitative analysis

Omar F. de P. e Krauss[1], Marcelo de S. Balbino[1 2], Cristiane N. Nobre[1]

[1] Pontifícia Universidade Católica de Minas Gerais, Brasil
[2] Centro Federal de Educação Tecnológica de Minas Gerais, Brasil
virionkrauss@gmail.com, marcelobalbino@gmail.com, nobre@pucminas.br

Abstract.

There is currently a growing concern about the explainability of machine learning algorithms. Explainability refers to the ability to understand and interpret the decisions made by the models, that is, the process by which a model arrives at a given prediction or classification. The counterfactual explanation involves creating alternative examples where the model's prediction differs from the original. This work aims to raise and discuss essential features in the context of counterfactual explanation methods. For this, the CSSE and LORE methods will be evaluated and applied to twelve public databases, considering different characteristics regarding the number of attributes and data types. In this way, we can better understand their strengths and weaknesses using standardized metrics for different methods. This facilitates the selection and development of more effective strategies and helps to identify cases where one approach may outperform another regarding the quality of explanations. The survey measured the metrics validity, prolixity, sparsity, similarity, and hit rate. In general terms, the CSSE performed better in these metrics, except for sparsity.

CCS Concepts: • **Supervised learning by classification**; • **Applied computing** → **Law**;

Keywords: Machine Learning, Counterfactual Explanation, LORE, CSSE

1. INTRODUÇÃO

Algoritmos de Aprendizado de Máquina (AM) têm sido usados para solucionar problemas e realizar tomadas de decisões cada vez mais complexas e que exigem um alto nível de transparência. Algoritmos de AM que não são capazes de fornecer ao usuário esta transparência são chamados de modelos 'caixa-preta', ou também chamados de algoritmos não interpretáveis [El Shawi et al. 2019].

Ao utilizar uma grande quantidade de atributos disponibilizados para a predição, o modelo pode ser tendencioso ou mesmo preconceituoso, dependendo dos tipos de dados que são fornecidos a ele durante seu treinamento. De acordo com Guidotti et al. [2018], ao contarmos com modelos de aprendizagem com estruturas escaláveis e de alto desempenho com uma quantidade massiva de dados para treinamento, estamos nos arriscando a utilizar um sistema de decisões que não entendemos por completo, o que pode afetar tanto a parte ética, quanto questões de segurança e responsabilidade.

Entretanto, essas demandas, necessidades e crescimento massivo são inevitáveis [Zhang e Lu 2021]. Por outro lado, não se pode deixar de exigir uma total transparência das tomadas de decisão obtidas pela Inteligência Artificial (IA). Portanto, em situações como estas, normalmente aplicam-se méto-

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq, Código: 311573/2022-3), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, Código: APQ-03076-18), ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) e à Pontifícia Universidade Católica de Minas Gerais (Código: FIP-2023/29184-1S). O trabalho foi desenvolvido no laboratório de Inteligência Computacional Aplicada – LICAP/PUC Minas.

dos que extraem a interpretabilidade desses modelos. Esses métodos são aqueles que irão dizer de que maneira determinada IA alcançou tal resultado. Existem diversas famílias de métodos de interpretabilidade e explicabilidade, como os métodos factuais, extração de regras e análise de sensibilidade [Guidotti et al. 2018], mas, neste trabalho, analisaremos especialmente os métodos contrafactuais.

Dado uma instância x e um classificador c no qual $c(x) = y$, um contrafactual consiste em uma instância x' cuja decisão de c sob x' seja diferente de y, ou seja, $c(x') = y'$ e $y' \neq y$. Além disso, as diferenças entre a instância x e a instância contrafactual x' devem ser mínimas [Guidotti 2022].

Assim, este trabalho visa apresentar as principais métricas utilizadas para avaliação de métodos de explicações contrafactuais, além de exemplificar o funcionamento destas métricas de forma qualitativa e quantitativa. A aplicabilidade de cada métrica será realizada em dois métodos conhecidos na literatura: o LORE (*Local Rule-Based Explanation Method*) [Guidotti et al. 2019] e o CSSE (Social Explanations for Classification Models) [Balbino et al. 2023], utilizando doze bases de dados. Ao analisar diferentes métricas, é possível avaliar o quão bem as explicações fornecidas pelos métodos se alinham com as expectativas dos usuários e com a interpretabilidade desejada para um determinado contexto. Isso ajuda a determinar se os métodos estão produzindo explicações úteis e relevantes.

Desta forma, ao ter métricas bem definidas, é possível comunicar de forma mais clara os méritos e as limitações das explicações geradas. Isso é especialmente importante em contextos sensíveis, como em sistemas de tomada de decisão automatizados, onde é crucial entender como as explicações estão sendo produzidas e se estão sujeitas a vieses ou interpretações equivocadas.

Ressalta-se que em [Balbino et al. 2023], os autores também realizaram uma comparação entre o CSSE e LORE, mas no presente trabalho a proposta é discutir características que não foram abordadas, como prolixidade e validade, além de realizar uma análise mais completa com diferentes bases de dados.

2. REFERENCIAL TEÓRICO

2.1 Métodos de Explicações contrafactuais

De acordo com Wachter et al. [2017], explicações contrafactuais são capazes de fornecer as informações mais importantes para qualquer usuário, por ser de fácil entendimento e úteis na prática. Em outras palavras, caso a decisão y' sob a instância contrafactual x' der um resultado diferente de y, essa diferença se deu por conta da característica alterada x'_0 e, portanto, apresentar somente essa característica alterada facilita o entendimento do usuário e auxilia em tomadas de decisões na realidade.

O método contrafactual CSSE [Balbino et al. 2023], que faz uso de um algoritmo genético ad-hoc para geração dos exemplos contrafactuais, gera múltiplos contrafactuais com diversidade (baseado em diferentes conjuntos de atributos). Além disso, o CSSE é agnóstico, ou seja, é capaz de gerar explicações para qualquer modelo de classificação independente do algoritmo de aprendizado.

Ademais, outro recurso do CSSE é permitir ao usuário indicar quais atributos de um conjunto de dados são acionáveis. Por definição, acionabilidade é a capacidade do método de permitir ao usuário selecionar quaisquer características que ele não gostaria que fossem alteradas para gerar os contrafactuais, tornando-as imutáveis. Por último, o modelo CSSE possui a característica de prover ao usuário que o mesmo selecione quantas explicações ele gostaria de visualizar.

Guidotti et al. [2019] propuseram o LORE, um método agnóstico e que faz uso de regras lógicas. Ao receber uma saída de um algoritmo de aprendizado e uma instância a ser explicada, o LORE constrói uma árvore de classificação local e a treina com uma grande quantidade de vizinhos da instância analisada que foram gerados por um algoritmo genético. A partir dessa árvore são derivadas uma explicação factual formada por uma regra que explica a decisão e um conjunto de regras contrafactuais que dizem quais trocas nos valores dos atributos conseguiriam reverter esta mesma decisão.

Além destes métodos, existem outros, tais como DICE [Mothilal et al. 2020], SEDCT [Vermeire

et al. 2022] e Polaris [Zhang et al. 2018], que apresentam características diferentes quanto a forma como eles recuperam as explicações contrafactuais, ao tipo de explicação (agnóstico ou específico), ao tipo de dados aceitável (tabulares, imagem, etc), além de outras categorizações. Em [Guidotti 2022], o autor apresenta outros métodos com suas respectivas características.

2.2 Métricas de comparação

Em [Guidotti 2022], o autor apresenta algumas métricas para avaliação de métodos contrafactuais. Dentre elas, selecionamos as seguintes para avaliação neste trabalho: *esparsidade*, *similaridade*, taxa de *hit*, *validade* e *prolixidade*. Ressalta-se que *esparsidade* e a *similaridade* referem-se a duas formas diferentes de se observar as mudanças mínimas dos contrafactuais para a instância original.

Para cálculo das métricas de esparsidade, similaridade e taxa de *hit*, é necessário obter o número total de contrafactuais a serem considerados. Para tal, considere um conjunto de N instâncias x_i para as quais se deseja gerar k contrafactuais para cada uma. Idealmente, teríamos um total de $N \times k$ contrafactuais. No entanto, de acordo com Balbino et al. [2023], é possível que um método contrafactual não consiga gerar os k contrafactuais requisitados para uma determinada instância. Assim, se para cada instância original x_i, NC_i contrafactuais são gerados (onde $NC_i \leq k$), obtém-se o total de contrafactuais TC, conforme calculado pela Equação 1:

$$TC = \sum_{i=1}^{N} NC_i \tag{1}$$

A *esparsidade* é a característica que faz com que o método contrafactual produza explicações cujo número de trocas de atributos (NT) entre a instância original x e o contrafactual x' seja mínimo. Assim, calcular a esparsidade (ES) de um método é obter a quantidade média de trocas (NT_mean) necessária para gerar TC contrafactuais (Equação 2) e seu desvio padrão SD (Equação 3). A Equação 4 apresenta o cálculo de ES:

$$NT\_mean = \frac{\sum_{i=1}^{TC} NT_i}{TC} \tag{2}$$

$$SD = \sqrt{\frac{\sum_{i=1}^{TC}(NT_i - NT\_mean)^2}{TC - 1}} \tag{3}$$

$$ES = \begin{cases} NT\_mean \\ SD \end{cases} \tag{4}$$

A *similaridade* está associada a distância entre as instâncias x e x', obtida por meio de uma função de distância, tais como Euclidiana, Manhattan ou Cosseno. Para este trabalho, escolheu-se a distância Euclidiana. Considerando que cada contrafactual está a uma distância D da instância original, é possível calcular a distância média D_mean dos TC contrafactuais encontrados usando a Equação 5.

$$D\_mean = \frac{\sum_{i=1}^{TC} D_i}{TC} \tag{5}$$

A taxa de $hit(c - hit)$ é a métrica relacionada a *validade* dos exemplos contrafactuais. Segundo Guidotti [2022], dado uma instância x pertencente a classe y, seu contrafactual x' é considerado válido se as decisões para x e x' forem diferentes. Em outras palavras, caso $c(x') = y'$ e $y' \neq y$, considera-se que $hit = 1$, caso contrário, $hit = 0$. Assim, avaliando todos os contrafactuais, obtém-se a porcentagem de contrafactuais que de fato são válidos, como mostrado na Equação 6.

$$c\text{-}hit = \frac{\sum_{i=1}^{TC} hit_i}{TC} \tag{6}$$

A *prolixidade* é uma característica indesejável nas explicações contrafactuais, pois dizem respeito a exemplos gerados a partir de incrementos ou decrementos nos mesmos atributos sem resultar em novas explicações para os usuários [Balbino et al. 2023]. Por exemplo, considere um contrafactual x_1, tal que o atributo modificado é $idade = 21$. Caso outro contrafactual também possua somente idade como atributo modificado, esses contrafactuais são prolixos.

3. MATERIAIS E MÉTODOS

3.1 Seleção das métricas

Por meio de levantamento bibliográfico, foram identificadas as métricas a serem utilizadas na avaliação dos métodos contrafactuais. Para a análise qualitativa foram consideradas as métricas de *prolixidade* e *validade*. No quesito quantitativo, foram consideradas a *esparsidade*, *similaridade* e *taxa de hit*. Estas métricas foram formalmente apresentadas na Seção 2.2.

3.2 Seleção dos métodos contrafactuais

Para a seleção dos métodos utilizados para comparação, consideramos a análise feita em [Balbino et al. 2023], que avalia algumas abordagens presentes na literatura e compara suas características a fim de verificar quais atendem determinados critérios. Os critérios foram: 1) geração de múltiplos contrafactuais; 2) ser ou não agnóstico [Guidotti et al. 2018]; 3) permitir ou não tratar a acionabilidade dos atributos [Guidotti 2022]. Dentre os métodos analisados, somente o LORE e o CSSE possuem as três características e portanto foram os selecionados para comparação nesse artigo.

3.2.1 *Protocolo de comparação.* Foi desenvolvido um protocolo para tornar justa a comparação entre o LORE e o CSSE. Para isso, após a obtenção dos contrafactuais pelo LORE, é utilizado um filtro para retirar todos os casos prolixos. Em seguida, para cada não prolixo, verifica-se se a instância contrafactual gerada é válida. Neste caso, o contrafactual é considerado justo para comparação. Ao obter todos os contrafactuais gerados pelo LORE que atendem à esses requisitos, os mesmos são contabilizados para que o CSSE gere um número igual de contrafactuais para cada instância original.

Esse protocolo é necessário pois o LORE não garante que seus contrafactuais não são prolixos e nem válidos, bem como não possui a característica de fornecer ao usuário a capacidade de escolher a quantidade de contrafactuais que ele deseja como saída. Tendo em vista que o CSSE possui essas três características, os valores obtidos para comparação são todos provindos de uma mesma quantidade de contrafactuais não prolixos e válidos.

3.3 Seleção das bases de dados

Para comparação entre os métodos, foram selecionadas doze bases de dados públicas. Em relação a quantidade de atributos, as bases foram classificadas em pequena, média e grande, considerando o cálculo de percentil entre as bases. Assim, as bases com até 13 atributos foram classificadas como pequenas, de 14 a 27 atributos como médias e as demais como grandes. A quantidade de entradas[1] varia de 5 até 57. Considerando o tipo do atributo predominante, as mesmas foram classificadas como numérica, categórica ou mista. As escolhas foram realizadas de forma a permitir avaliar os métodos em bases de diferentes características. As bases de dados estão listadas na Tabela I.

Em especial, utilizou-se a base *Compas* para exemplificar as explicações geradas pelos métodos. A base Compas inclui dados demográficos, histórico criminal, tempo de prisão e três pontuações *Compas* (risco de reincidência de crimes, risco de violência e o risco de não comparecer em ocasiões agendades

[1] Quantidade de entradas equivale aos atributos após a sua codificação para numérico. Neste trabalho, atributos nominais ordinais foram codificados usando a codificação ordinal encoding, e os atributos nominais não ordinais foram binarizados, utilizando a codificação one hot encoding.

Tabela I: Bases de dados utilizadas nos experimentos.

Base	Qtd. entradas	Tamanho	Tipo
Adult	24	Média	Mista
Australian	14	Média	Mista
Compas	11	Pequena	Mista
Diabetes	8	Pequena	Numérica
Ionosphere	33	Grande	Numérica
Mammographic masses	5	Pequena	Categórica
Phishing Websites	30	Grande	Categórica
Spambase	57	Grande	Numérica
Student Performance - MAT	43	Grande	Mista
Titanic	26	Média	Categórica
Votes	16	Média	Categórica
Wine	11	Pequena	Numérica

por conta de sua situação legal) de 7214 réus do condado de Broward, Flórida. Cada réu recebe uma pontuação Compas que varia de 1 a 10. Nos experimentos, considerou-se as transformações realizadas por Guidotti et al. [2019] que resultaram em 12 atributos, incluindo a classe. Para o atributo classe, os autores rotularam as pontuações de 1 a 6 como "Risco Médio-Baixo" e de 7 a 10 como "Alto Risco". A base possui 5.219 réus classificados como "Risco Médio-Baixo" e 1.995 como "Alto Risco".

3.4 Experimentos e análise dos resultados

Foram realizados experimentos para avaliação do CSSE e LORE utilizando as métricas selecionadas. Para cada base de dados foi gerado um modelo baseado em *Random Forest* (RF) e outro com Redes Neurais (RN). Ambos modelos utilizaram todos os hiper-parâmetros em seu valor padrão[2].

Os experimentos foram realizados utilizando *trinta* instâncias do conjunto de teste de cada base de dados. A análise dos resultados foi dividida em duas partes: qualitativa e quantitativa. Na análise qualitativa foi observado como a saída gerada pelos métodos trata os critérios de *validade* e *prolixidade*. Na análise quantitativa foram calculadas as métricas de *esparsidade*, *similaridade* e *taxa de hit* de cada método, para cada base/classificador.

4. RESULTADOS E DISCUSSÕES

Para investigar as métricas nos métodos CSSE e LORE foram realizados experimentos com 12 bases de dados de diferentes tipos e tamanhos (quantidade de atributos). Na Seção 4.1 é realizada uma análise qualitativa apresentando exemplos das saídas geradas pelos métodos e alguns recursos e limitações. Na Seção 4.2 avalia-se a capacidade dos métodos para gerar contrafactuais válidos e com mudanças mínimas em relação à instância original. Para tal, compara-se o desempenho dos mesmos utilizando as métricas de *esparsidade*, *similaridade* e *taxa de hit*.

4.1 Análise qualitativa

Para exemplificar a aplicação do CSSE e LORE e analisar as explicações geradas pelos mesmos foi utilizado o modelo de classificação para a base de dados *Compas* baseado em *Random Forest*. Para tal, entre as instância do conjunto de teste, foi escolhida uma pertencente a classe *Alto Risco*.

As Tabelas II e III apresentam os contrafactuais gerados pelo LORE e CSSE, respectivamente, para a instância escolhida. Apenas os atributos alterados para reverter a classe estão presentes na tabela. Além disso, para o LORE, foram avaliados se os mesmos são prolixos e válidos. Ressalta-se que o CSSE garante que os contrafactuais não são prolixos e são válidos. Seguindo o protocolo apresentado na Seção 3, como o LORE gerou dois contrafactuais válidos e não prolixos[3], o CSSE foi executado para geração do mesmo número de exemplos. Na execução dos métodos, não se restringiu a utilização de quaisquer atributos nas explicações.

[2] Para o algoritmo Random Forest: n_estimators = 100, criterion = gini e max_features = sqrt. Para a Rede neural: activation = relu, solver = adam, learning_rate_init = 0.001 e hidden_layers_sizes = 100. [3] Observe que o LORE apresentou dois contrafactuais prolixos: o segundo exemplo faz apenas um decremento na idade e o quarto um decremento no atributo antecedente.

Tabela II: Resultado de aplicação do LORE

	Atributos modificados			
	Idade	**Antecedentes**	**Prolixo**	**Válido**
Instância original: **classe**: Alto	26	7	–	–
Contrafactuais **classe**: Médio-Baixo	38.6	–	Não	Sim
	29.5	–	Sim	Não
	–	3.0	Não	Sim
	–	0.5	Sim	Sim

Tabela III: Resultado de aplicação do CSSE

	Atributos modificados	
	Idade	**Antecedentes**
Instância original: **classe**: Alto	26	7
Contrafactuais **classe**: Médio-baixo	34	–
	–	1

A instância original escolhida consiste, dentre outros atributos, de um indivíduo de 26 anos de idade, com 7 antecedentes criminais e que pertence a classe de *Alto Risco* de reincidência criminal. Considerando os contrafactuais não prolixos e válidos encontrados pelo LORE, esse indivíduo teria sido classificado como *Risco Médio-Baixo* se tivesse 38.6 anos ou 3 antecedentes criminais. Para o CSSE, o mesmo indivíduo seria da classe *Risco Médio-Baixo* se sua idade fosse 34 anos ou tivesse 1 antecedente criminal.

No exemplo apresentado, ambos os métodos encontraram contrafactuais pela alteração de um único atributos (idade ou número de antecedentes). Em relação a distância para instância original, o CSSE obteve o contrafacual mais próximos a partir de uma mudança menor no atributo idade e o LORE obteve o contrafactual mais similar em se tratando da mudança no número de antecedentes criminais.

Entende-se que o CSSE garantir que os contrafactuais gerados não são prolixos e são válidos constitui uma vantagem significativa do método em relação ao LORE. Avaliando outros recursos que não foram utilizados neste exemplo, percebe-se que outra vantagem do CSSE é permitir ao usuário escolher a quantidade de contrafactuais que deseja gerar. No LORE, o número de contrafactuais é definido pelas regras geradas pela árvore de decisão. Além disso, o CSSE permite ajustar o peso a ser atribuído a esparsidade e similaridade de forma que o usuário pode priorizar contrafactuais menos esparsos ou mais similares, o que pode ser necessário em alguns contextos.

4.2 Análise quantitativa

A partir dos experimentos realizados com o CSSE e o LORE, calculou-se os valores médios do *números de trocas* (esparsidade), a *distância média* (similaridade) e *taxa de hit* médio de cada método para os dois classificadores em cada base de dados selecionada. A Figura 1 apresenta os resultados obtidos. Nas comparações entre os resultados alcançados para *esparsidade* e *similaridade* foi utilizado o Teste-t com intervalo de confiança de 95% ($p\text{-value} > 0.05$) e *Valor crítico* = 2.001717484. No caso da *taxa de hit*, a comparação foi realizada pelo valor absoluto.

Em se tratando da *média de trocas* necessária para gerar os contrafactuais com o modelo baseado em *Random Forest* (Figura 1a), os cálculos com o Teste-t indicaram que o CSSE e o LORE apresentam comportamento equivalentes em todas as bases analisadas. Nessa mesma métrica, no modelo com Redes Neurais (Figura 1b), o LORE alcançou desempenho superior ao CSSE nas bases *Adult*[4], *Ionosphere*[5] e *Votes*[6]. Nas demais bases, os resultados dos métodos foram equivalentes.

A Figura 1c apresenta os resultados obtidos pelo CSSE e LORE em relação a *distância média* dos contrafactuais para as respectivas instâncias originais. No modelo com *Random Forest*, o CSSE foi superior ao LORE nas bases *Diabetes*[7], *Ionosphere*[8], *Student Performance – MAT*[9] e *Wine*[10]. O LORE teve desempenho superior na base *Votes*[11]. Nas demais bases, os métodos foram equivalentes.

A Figura 1d apresenta os valores de *distância média* para o modelo com Redes Neurais. O CSSE ap-

[4] $t = 3.5679$, $p\text{-value} = 0.0007$ [5] $t = 2.4494$, $p\text{-value} = 0.0173$ [6] $t = 5.0853$, $p\text{-value} = 4.12175E-06$
[7] $t = 3.5804$, $p\text{-value} = 0.0007$ [8] $t = 3.9865$, $p\text{-value} = 0.0002$ [9] $t = 3.8203$, $p\text{-value} = 0.0003$ [10] $t = 7.6762$, $p\text{-value} = 2.1404E-10$ [11] $t = 2.6724$, $p\text{-value} = 0.0098$

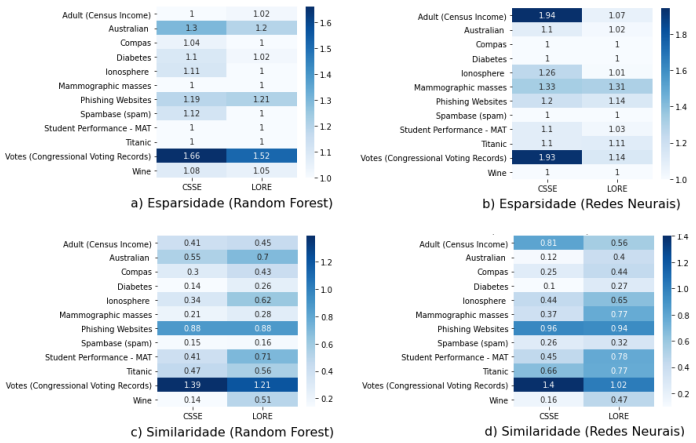


Fig. 1: Resultados de esparsidade e similaridade.

resentou desempenho superior ao LORE nas bases *Australian*[12], *Compas*[13], *Diabetes*[14], *Ionosphere*[15], *Mammographic masses*[16], *Student Performance – MAT*[17] e *Wine*[18]. Novamente, o LORE atingiu melhor desempenho apenas na base *Votes*[19]. Nas outras bases, os resultados foram equivalentes.

A Tabela IV apresenta a *taxa de hit* média obtida pelo LORE para cada classificador em cada base de dados. Destacadamente, o CSSE é favorecido neste critério, uma vez que retorna apenas contrafactuais válidos, ou seja, a taxa de *hit* = 100% em todas as bases de dados.

Tabela IV: Taxas de *hit*, em porcentagem, obtidas com LORE.

Base	Hit RF	Hit RN
Adult	46.77	68.85
Australian	44.00	36.94
Compas	50.00	58.23
Diabetes	45.28	59.62
Ionosphere	17.97	50.99
Mammographic masses	73.20	79.45
Phishing Websites	57.73	77.31
Spambase	19.42	37.44
Student Performance - MAT	40.54	53.78
Titanic	64.34	48.72
Votes	71.64	64.56
Wine	44.52	81.40

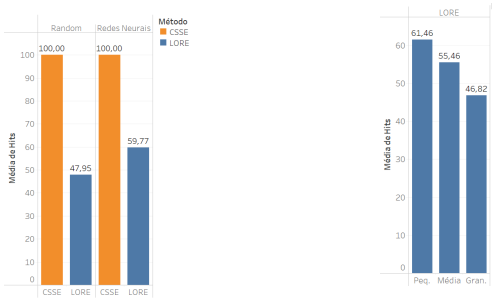
Ainda em relação a taxa de *hit*, calculou-se valores médios, considerando todas as bases e, no caso do LORE, separadas por tamanho. Os valores médios apresentados na Figura 2a confirmam as observações feitas para as bases individualmente, ou seja, o CSSE apresenta desempenho significamente superior ao LORE neste critério. Sobre a taxa de *hit* do LORE em bases de diferentes tamanhos, a Figura 2b indica que o método apresenta maior dificuldade na geração de contrafactuais válidos nas bases com maior quantidade de atributos.

Em uma análise geral, por voto majoritário, o CSSE e LORE mostraram-se equivalentes em relação à esparsidade no modelo com *Random Forest* e o LORE apresentou melhor desempenho com as Redes Neurais. No que se refere à similaridade, o CSSE foi superior ao LORE em ambos os classificadores. Por fim, em relação a taxa de *hit*, o CSSE apresentou desempenho superior em todas as bases.

5. CONCLUSÃO

No campo da interpretabilidade, os métodos contrafactuais têm ganhado evidência em função da sua capacidade de entendimento e comunicação com os usuários. Tal destaque traz consigo a necessidade de

[12] $t = 3.973063667$, $p\text{-}value = 0.000198621$ [13] $t = 2.3013$, $p\text{-}value = 0.0249$ [14] $t = 4.2064$, $p\text{-}value = 9.1080E-05$
[15] $t = 3.8046$, $p\text{-}value = 0.0003$ [16] $t = 6.5819$, $p\text{-}value = 1.4713E-08$ [17] $t = 3.96317$, $p\text{-}value = 0.0002$
[18] $t = 5.7108$, $p\text{-}value = 4.0650E-07$ [19] $t = 5.9589$, $p\text{-}value = 1.5931E-07$



(a) Taxa de *hit* CSSE x LORE

(b) *Hit* LORE por tamanho

Fig. 2: Taxa de *hit* médias com todas as bases.

avaliar os méritos e limitações de cada método contrafactual. Neste artigo, realizamos o levantamento de métricas de avaliação de métodos contrafactuais que permitam verificar o quanto suas explicações estão alinhadas com as necessidades dos usuários. Destacamos ainda o CSSE e o LORE, dois relevantes métodos contrafactuais, os quais foram avaliados com essas mesmas métricas.

Experimentos foram realizados com dois classificadores em doze bases de dados, e as análises foram divididas em qualitativa e quantitativa. No âmbito qualitativo, entendemos que o CSSE apresenta algumas vantagens sobre o LORE, tendo em vista que o primeiro permite que o usuário escolha a quantidade desejada de contrafactuais e todos os exemplos obtidos são garantidamente não prolixos e válidos. Na parte quantitativa, em relação a *esparsidade*, os métodos foram equivalentes com o modelo em *Random Forest* e o LORE foi superior com as Redes Neurais. Para *similaridade* e *taxa de hit*, o CSSE foi superior em ambos classificadores.

Para trabalhos futuros, é fundamental realizar experimentos com mais bases de dados variando em quantidade e tipos de atributos para analisar o comportamento dos modelos. Além disso, testes adicionais devem ser conduzidos com outros métodos de explicação contrafactual, visando uma maior diversidade em relação às diferentes abordagens.

REFERÊNCIAS

Balbino, M. d. S., Zárate, L. E. G., and Nobre, C. N. Csse - an agnostic method of counterfactual, selected, and social explanations for classification models. *Expert Systems with Applications*, 2023.

El Shawi, R., Sherif, Y., Al-Mallah, M., and Sakr, S. Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, pp. 275–280, 2019.

Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 2022.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34 (6): 14–23, 2019.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51 (5), aug, 2018.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. pp. 607–617, 2020.

Vermeire, T., Brughmans, D., Goethals, S., de Oliveira, R. M. B., and Martens, D. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications* 25 (2): 315–335, May, 2022.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.* vol. 31, pp. 841, 2017.

Zhang, C. and Lu, Y. Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration* vol. 23, pp. 100224, 2021.

Zhang, X., Solar-Lezama, A., and Singh, R. Interpreting neural network judgments via minimal, stable, and symbolic corrections. *Advances in neural information processing systems* vol. 31, 2018.