Explicando os "E se?" por Trás da Detecção de Violência Física Infantil: Uma Abordagem Contrafactual

Lívia Xavier[1], Marcelo Balbino[2], Hasheem Mannan[3], Cristiane Nobre[1*]

[1]Pontifícia Universidade Católica de Minas Gerais (PUC – MG)
[2]Centro Federal de Educação Tecnológica de Minas Gerais (CEFET – MG)
[3]School of Nursing, Midwifery and Health Systems, University College Dublin, IE

livia.camara@sga.pucminas.br, marcelobalbino@gmail.com
hasheem.mannan@ucd.ie, nobre@pucminas.br

Abstract. *Machine learning models are widely used, but limited interpretability compromises transparency, especially in sensitive cases. This study uses the CSSE method to produce counterfactual explanations for violent physical discipline using the UNICEF MICS dataset and an XGBoost model. The results show that beliefs about physical punishment and behavioral redirection are key factors, small shifts in these variables change the model's predictions. These findings show that counterfactual explanations clarify model behavior and help guide interventions to support child well-being.*

Resumo. *Os modelos de aprendizado de máquina são amplamente utilizados, mas a interpretabilidade limitada compromete a transparência, especialmente em casos sensíveis. Este estudo utiliza o método CSSE para produzir explicações contrafactuais para a disciplina física violenta, usando o conjunto de dados MICS da UNICEF e um modelo XGBoost. Os resultados mostram que as crenças sobre punição física e redirecionamento comportamental são fatores-chave, e pequenas mudanças nessas variáveis alteram as previsões do modelo. Essas descobertas demonstram que as explicações contrafactuais esclarecem o comportamento do modelo e ajudam a orientar intervenções para promover o bem-estar infantil.*

1. INTRODUÇÃO

O uso de técnicas de inteligência artificial (IA) tem se expandido significativamente em áreas como medicina, direito e ciências sociais [Guidotti et al. 2018; Arrieta et al. 2020]. Apesar dos avanços, muitos modelos ainda operam como caixas-pretas, fornecendo apenas o resultado final sem explicitar os fatores que motivaram suas decisões [Samek et al. 2023; Rudin 2019]. Em contrapartida, modelos intrinsecamente interpretáveis, como árvores de decisão e modelos lineares, oferecem maior transparência, embora nem sempre alcancem o mesmo desempenho preditivo que algoritmos mais complexos [Caruana et al. 2015; Rudin 2019].

*Os autores gostariam de agradecer ao Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq: 311573/2022-3), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES - Auxílio PROAP 88887.842889/2023-00 - PUC/MG - Informática e Código de Financiamento 001 e FIP 2025/32409), e à Fundação de Amparo à Pesquisa de Minas Gerais (APQ-03104-24 e APQ-05058-23).

Essa limitação compromete a confiança dos usuários e dificulta a adoção desses sistemas em contextos sensíveis, nos quais a interpretabilidade é essencial para evitar vieses e decisões discriminatórias [Reddy et al. 2020; Selbst et al. 2019]. Nesse cenário, as explicações contrafactuais surgem como uma abordagem promissora, permitindo compreender os fatores que influenciam as decisões e identificar mudanças que poderiam levar a resultados diferentes [Wachter et al. 2018; Karimi et al. 2022].

Este trabalho investiga dados da pesquisa MICS da UNICEF relacionados à disciplina violenta, entendida como práticas punitivas físicas ou psicológicas aplicadas por responsáveis [UNICEF 2019; Sinhorinho e de Moura 2022]. A agressão psicológica refere-se ao uso de estratégias verbais por parte de um adulto para controlar o comportamento da criança, incluindo gritos, insultos ou humilhações. Já a agressão física envolve o emprego de força com a intenção de causar dor como forma de disciplina, como bater na criança com a mão ou com objetos. Em contraste, a disciplina não agressiva baseia-se em abordagens construtivas, como explicar por que determinado comportamento é inadequado (raciocínio verbal), retirar privilégios de forma proporcional e redirecionar a atenção da criança por meio de alternativas ou atividades apropriadas. Neste trabalho, focaremos na análise da disciplina física violenta. Apesar de amplamente discutida em termos de impactos negativos, essa temática ainda é pouco explorada quanto às motivações que levam à sua ocorrência, sendo fundamental compreender tais fatores para orientar políticas públicas e estratégias de prevenção.

Para isso, utiliza-se o método CSSE (Counterfactual, Selected, and Social Explanations) [de Sousa Balbino et al. 2023], com o objetivo de identificar determinantes associados à disciplina física violenta e explorar cenários alternativos que poderiam evitá-la. Dessa forma, busca-se não apenas ampliar a compreensão sobre o problema, mas também evidenciar o potencial de abordagens mais transparentes e interpretáveis em inteligência artificial.

2. REFERENCIAL TEÓRICO

2.1. Machine Learning e suas aplicações

Modelos baseados em inteligência artificial têm se tornado cada vez mais presente em diferentes áreas do conhecimento, como saúde, engenharia e direito [Musman et al. 2011; Khan et al. 2022; Medvedeva et al. 2022]. Embora ofereçam resultados expressivos, muitos desses sistemas permanecem caracterizados como verdadeiras caixas-pretas, já que o processo decisório interno raramente é transparente. Essa falta de clareza gera preocupações quanto à confiabilidade e à capacidade de auditar o comportamento desses modelos, o que intensifica a necessidade de ferramentas que tornem suas decisões mais compreensíveis para especialistas e usuários finais.

Alguns algoritmos, como as Árvores de Decisão, apresentam natureza intrinsecamente interpretável, permitindo a visualização de regras e caminhos utilizados em cada inferência [Silva 2023]. Contudo, esse nível de transparência não se estende aos modelos mais robustos e amplamente utilizados, como Redes Neurais, Random Forest e SVMs. Nesses casos, a lógica interna permanece inacessível, o que motiva o desenvolvimento e o uso de técnicas voltadas especificamente para explicar o funcionamento de modelos caixa-preta. Esse cenário tem impulsionado a área de interpretabilidade, pois

compreender, ainda que de forma aproximada, os critérios adotados por sistemas complexos é fundamental para assegurar sua aplicação responsável, especialmente em domínios sensíveis.

2.2. Interpretabilidade contrafactual

Entre as diversas abordagens propostas para aprimorar a interpretabilidade de modelos de Inteligência Artificial, as explicações contrafactuais destacam-se por sua clareza e proximidade com o raciocínio humano [Mothilal et al. 2020]. Essa abordagem consiste em realizar pequenas modificações na instância original com o objetivo de identificar quais alterações seriam suficientes para mudar o resultado da predição.

Esse tipo de explicação é especialmente útil por ser intuitiva e diretamente passível de aplicação, uma vez que pode ser compreendida por usuários finais e permite a adoção de medidas para alterar o resultado de uma decisão automatizada. Além disso, explicações contrafactuais são importantes para avaliar o funcionamento e a imparcialidade dos modelos de IA. Por exemplo, uma pessoa que solicita um empréstimo e é classificada como "alto risco" pode ter um contrafactual,uma versão alternativa mínima dessa mesma pessoa, que altera a decisão para "baixo risco". Na Tabela 1, a instância original (21 anos, baixa poupança e crédito elevado) recebe classificação negativa. Já os contrafactuais mostram pequenas mudanças, como aumentar a idade, melhorar a poupança ou reduzir o crédito já seriam suficientes para uma classificação positiva.

Tabela 1. Exemplo de explicação contrafactual

	Atributos modificados		
	Idade	**Poupança**	**Valor do crédito**
Instância original classe: 1 (Ruim)	21	1 (... <100 DM*)	15653
Contrafactuais classe: 0 (Bom)	30 –	2 (100 $\leq$ ... <500 DM) –	– 9157
Nenhum atributo estático			

*DM = Deutsche Mark = Marco Alemão

Diversos métodos têm sido propostos na literatura para a geração de explicações contrafactuais. O trabalho de [Wachter et al. 2018] foi um dos primeiros a formalizar o conceito de contrafactual e seu respectivo problema de otimização. Posteriormente, Guidotti et al. [2019] definiram contrafactuais como instâncias alternativas que diferem minimamente da observação original, mas que levam a uma predição desejada.

Nesse contexto, o método LORE [Guidotti et al. 2019] explora o uso de modelos substitutos locais e abordagens baseadas em regras, gerando regras factuais e contrafactuais por meio de um algoritmo genético que percorre a vizinhança local das instâncias. De forma semelhante, o método DiCE [Mothilal et al. 2020] propõe a geração simultânea de múltiplos contrafactuais diversos, otimizando diferentes instâncias para aumentar sua utilidade, ao mesmo tempo em que respeita restrições definidas pelo usuário relacionadas à viabilidade e acionabilidade das soluções.

Mais recentemente, o método CSSE [de Sousa Balbino et al. 2023] direciona-se explicitamente à perspectiva do usuário final, utilizando busca evolutiva para retornar um conjunto de contrafactuais que equilibram critérios como esparsidade e similaridade. Dessa forma, o método busca produzir explicações concisas, válidas e diversas, evitando redundâncias e garantindo maior interpretabilidade.

2.3. O método CSSE

CSSE é um dos métodos de explicação baseados em contrafactuais, ou seja, busca identificar quais atributos, se fossem modificados, alterariam a predição de uma instância. Esse modelo apresenta três características centrais: ser agnóstico ao modelo, múltiplo e acionável. Isso significa que o CSSE pode ser aplicado a qualquer algoritmo de classificação, independentemente de sua estrutura interna (modelo agnóstico), é capaz de gerar múltiplas explicações contrafactuais para uma mesma instância (múltiplo), e permite que o usuário defina atributos que não devem ser alterados (acionável).

Para gerar essas explicações, o CSSE considera quatro métricas principais. Segundo Kraus et al. [2020], a *esparsidade* avalia a quantidade de atributos que precisam ser modificados para gerar o contrafactual. A *similaridade* mede a distância entre a instância original e o contrafactual, sendo, no CSSE, calculada por meio da distância Euclidiana. A *validade* verifica se a classe do contrafactual é, de fato, diferente da classe original, garantindo que a explicação tenha efeito sobre a decisão do modelo. Por fim, a *prolixidade* identifica explicações redundantes, geradas a partir de aumentos ou reduções nos mesmos atributos sem acrescentar novas justificativas, penalizando esse tipo de repetição.

Além disso, o CSSE utiliza um algoritmo genético (uma técnica de otimização inspirada na seleção natural, que evolui soluções por meio de operações como seleção, cruzamento e mutação) para gerar múltiplos contrafactuais. No processo de geração de contrafactuais esse algoritmo usa três operações fundamentais: elitismo, crossover e mutação. O *elitismo* garante que os melhores indivíduos, ou seja, as soluções com maior qualidade de explicação segundo os critérios de avaliação, sejam preservados entre as gerações, assegurando que o conhecimento adquirido não seja perdido. O *crossover* (ou cruzamento) combina partes de dois indivíduos selecionados para gerar novos contrafactuais, promovendo diversidade e explorando novas regiões do espaço de busca. Já a *mutação* introduz pequenas alterações aleatórias em alguns atributos dos indivíduos, com o objetivo de evitar a convergência prematura em soluções locais e permitir a descoberta de contrafactuais potencialmente mais eficazes. Esses três mecanismos trabalham em conjunto para refinar progressivamente o conjunto de explicações geradas, mantendo um equilíbrio entre exploração e otimização.

A partir disso, o CSSE permite que o usuário defina quais características deseja priorizar, uma característica específica do modelo, que possibilita a seleção de atributos nos quais o usuário deseja dar maior ênfase, conforme a situação. Para isso, uma função de avaliação é empregada para medir a similaridade e a esparsidade das soluções geradas. Assim como nos algoritmos genéticos tradicionais, os melhores indivíduos são selecionados e armazenados a cada geração.

Ao final do processo, os melhores contrafactuais são ordenados com base no valor obtido na função de avaliação e na quantidade de instâncias definida pelo usuário. Dessa forma, o CSSE assegura que todas as explicações contrafactuais geradas correspondam a instâncias válidas, garantindo que as alterações nos atributos sejam suficientes para justificar a mudança de classe da instância original.

3. TRABALHOS RELACIONADOS

No campo da Inteligência Artificial Explicável (XAI), as explicações contrafactuais têm se consolidado como uma das abordagens mais relevantes, isso porque, fornecerem res-

postas orientadas à ação e interpretáveis por usuários não especialistas. Um exemplo é o método AdVice [Gomez et al. 2021] propõe explicações contrafactuais visuais agregadas, permitindo validar modelos por meio de alterações sintéticas em imagens, o que amplia sua aplicabilidade em tarefas supervisionadas de visão computacional.

No contexto educacional, Carvalho et al. [2024] avançam essa discussão ao adaptar o método LIME para gerar contrafactuais, incorporando validação estatística e métricas de equidade. O estudo demonstra como explicações podem revelar disparidades de desempenho entre grupos sensíveis, destacando a importância de métodos explicáveis para identificar possíveis vieses e apoiar decisões pedagógicas mais justas. Além disso, o estudo brasileiro [Marcolino et al. 2025] aplica contrafactuais para identificar intervenções que possam melhorar o desempenho estudantil, evidenciando o potencial da técnica para apoiar a tomada de decisão pedagógica. Juntos, esses estudos demonstram que abordagens explicáveis, especialmente as contrafactuais, são essenciais para aprimorar transparência, segurança e confiabilidade de modelos em domínios de alto impacto social, como saúde, educação, visão computacional e controle inteligente.

4. MATERIAIS E MÉTODOS

4.1. Descrição da base de dados

A base de dados utilizada neste estudo corresponde ao Multiple Indicator Cluster Surveys (MICS), disponibilizado pela UNICEF. As informações do MICS são coletadas por meio de questionários aplicados às famílias, contemplando dados sobre a criança ou adolescente, bem como características dos pais ou responsáveis. O conjunto de dados também inclui variáveis relacionadas às condições de saúde, moradia e outros aspectos contextuais relevantes.

A partir do conjunto de dados MICS, a variável-alvo '*violent_discipline*' foi definida com base em seis atributos diretamente relacionados à agressão física: 'FCD2G', 'FCD2I', 'FCD2K', 'FCD2F', 'FCD2J' e 'FCD2C'. Esses atributos correspondem, respectivamente, a: *bater na criança na bunda ou em outra parte do corpo com cinto, escova ou vara; bater na criança no rosto, cabeça ou orelhas; espancar a criança o mais forte possível; dar palmadas, bater ou dar tapas na criança na bunda com a mão; bater na criança na mão, braço ou perna; e sacudir a criança.* Assim, a classe recebeu valor 1 (indicando que a criança sofreu disciplina violenta) quando pelo menos um desses atributos apresentava resposta positiva, e valor 0 quando todos eram negativos. Dessa forma, a base de dados resultante contem 94 atributos.

Inicialmente, realizou-se um levantamento com o objetivo de identificar a região com maior representatividade de respostas ao questionário (veja Tabela 2). Observa-se que Oceania e África apresentam níveis de representatividade semelhantes. Assim, optou-se por analisar os dados da África, por possuir um número ligeiramente maior de instâncias.

4.2. Pré-processamento da base de dados

Primeiramente, foram selecionados apenas os atributos que apresentavam pelo menos 30% de valores observados (isto é, valores não ausentes) em relação ao total de instâncias, assim atributos com alta proporção de dados faltantes (superior a 70%) foram descartados. Em seguida, os valores ausentes foram imputados utilizando o método KNN Imputer. Os

Tabela 2. Distribuição de classes por região

Região	Classe 0	Classe 1	Total	Classe 1 (%)	Países (%)
Africa	77.258	100.273	177.531	56,49	35,19
America	34.548	18.483	53.031	34,85	28,57
Asia	106.412	105.735	212.147	49,84	24,49
Europe	12.232	3.889	16.121	24,12	13,64
Oceania	3.443	5.917	9.360	63,22	35,71
Total	233.893	234.297	468.190	50,04	26,94

atributos categóricos e numéricos foram então pré-processados separadamente, incluindo a recodificação de faixas etárias e o agrupamento de categorias. Posteriormente, *outliers* foram identificados e removidos, e os dados foram normalizados por meio do StandardScaler [Pedregosa et al. 2011].

A divisão entre conjuntos de treino e teste foi realizada de forma a evitar vazamento de dados. Considerando o desbalanceamento do conjunto, aplicou-se a técnica de *undersampling*, utilizando o Random UnderSampling apenas no conjunto de treino e exclusivamente sobre a classe majoritária, preservando assim a integridade da validação [Lemaître et al. 2017]. Por fim, após a remoção de registros duplicados e inconsistentes, foi realizada a otimização de hiperparâmetros. Os modelos Decision Tree, Random Forest, XGBoost e SVM foram ajustados utilizando o método BayesSearch [Shahriari et al. 2016], enquanto as Redes Neurais tiveram seus parâmetros otimizados por meio do RandomSearch [Bergstra e Bengio 2012]. Entre os modelos avaliados, o que obteve o melhor desempenho geral foi selecionado para rodar o algoritmo CSSE e obter as explicações contrafactuais.

4.3. Métricas de avaliação

Para a avaliação dos modelos de aprendizado de máquina, foram utilizadas três métricas: *recall*[1], precisão[2], *F1score*[3], que corresponde à média harmônica entre precisão e *recall*. Além disso, foi empregada a técnica de validação cruzada com 10 dobras, considerando-se 30% dos dados reservados para teste.

Também foi implementada uma modificação no código do CSSE para permitir o registro, ao longo das iterações, dos atributos modificados em cada contrafactual gerado. Além disso, passou a ser possível selecionar especificamente quais atributos se deseja armazenar, por exemplo, apenas aqueles associados a contrafactuais que resultaram na mudança da classe de 1 para 0 ou no sentido inverso.

O principal interesse concentrou-se nos casos em que a classificação da instância passou de 1 (sofreu disciplina violenta) para 0 (não sofreu disciplina violenta). Os atributos modificados foram registrados ao longo de toda a execução, possibilitando que, ao final da execução do CSSE, fossem identificados quais atributos apareceram com maior frequência como responsáveis pela mudança de classe.

Adicionalmente, tornou-se possível analisar a frequência do número de atributos modificados em cada iteração. Dessa forma, ao término da execução, o método apresenta

[1]Recall = $\frac{TP}{TP+FN}$
[2]Precision = $\frac{TP}{TP+FP}$
[3]F1Score = $2 \times \frac{Precision \times Recall}{Precision+Recall}$

quantos atributos precisaram ser alterados para que a classificação da instância mudasse. A distribuição resultante evidencia, por exemplo, quantos contrafactuais exigiram apenas uma modificação e quantos demandaram duas, três ou mais alterações, permitindo identificar se mudanças pequenas ou mais extensas foram predominantes no processo.

5. RESULTADOS

Após a avaliação do desempenho dos modelos, observou-se que o XGBoost apresentou o melhor resultado médio de *F1-score*[4], conforme apresentado na Tabela 3. Para avaliar a significância estatística das diferenças observadas, foram aplicados os testes não paramétricos de Friedman, seguido do pós-teste de Nemenyi, conforme recomendado por Demšar [2006], sendo adequados para comparação de múltiplos modelos em cenários com validação cruzada. O teste de Friedman indicou diferenças estatisticamente significativas entre os algoritmos ($p < 0{,}05$).

A análise dos *rankings* médios obtidos no teste de Nemenyi mostrou que o XGBoost apresentou o melhor desempenho ($\approx 1{,}7$), seguido por Random Forest ($\approx 2{,}1$) e SVM ($\approx 2{,}9$), enquanto Rede Neural ($\approx 4{,}0$) e Decision Tree ($\approx 4{,}9$) apresentaram resultados inferiores.

Tabela 3. Comparação de desempenho dos modelos

Modelo	Classe	Precisão	Recall	F1-Score
Decision Tree	0	0.46	0.66	0.55
	1	0.81	0.65	0.72
Random Forest	0	0.57	0.75	0.65
	1	0.87	0.74	0.80
XGBoost	0	0.57	0.75	0.65
	1	0.87	0.75	0.80
SVM	0	0.57	0.74	0.64
	1	0.86	0.74	0.80
Rede Neural	0	0.50	0.71	0.59
	1	0.84	0.68	0.75

Assim, a partir do modelo treinado pelo algoritmo XGBoost, aplicou-se o método de explicação contrafacual CSSE. A Figura 1 apresenta, em (a), os atributos mais relevantes para a transição da classe "sofreu disciplina violenta" para "não sofreu" e, em (b), o número de atributos necessários para essa mudança.

Esses resultados foram obtidos a partir da análise de até três contrafactuais (número previamente definido) gerados para 1.000 instâncias selecionadas aleatoriamente. Ao todo, foram produzidos 1.668 contrafactuais, uma vez que não foi possível gerar três contrafactuais para todas as instâncias.

O atributo mais influente foi '*redirecionou a criança para outra atividade*' ("*redirect_child*" - FCD2E), que representou 42% das modificações, indicando que práticas de redirecionamento comportamental, ações nas quais cuidadores guiam a criança à outra atividade diferente da que está fazendo, têm um efeito substancial na mudança de classe. Em seguida está '*acredita na necessidade de punição física*' ("*believe_physical_punishment*" -

[4]Os hiperparâmetros finais deste modelo foram: *colsample_bytree* = 0.8, *gamma* = 0.5, *learning_rate* = 0.05, *max_depth* = 13, *n_estimators* = 200, *reg_alpha* = 0.1, *reg_lambda* = 1 e *subsample* = 0.8

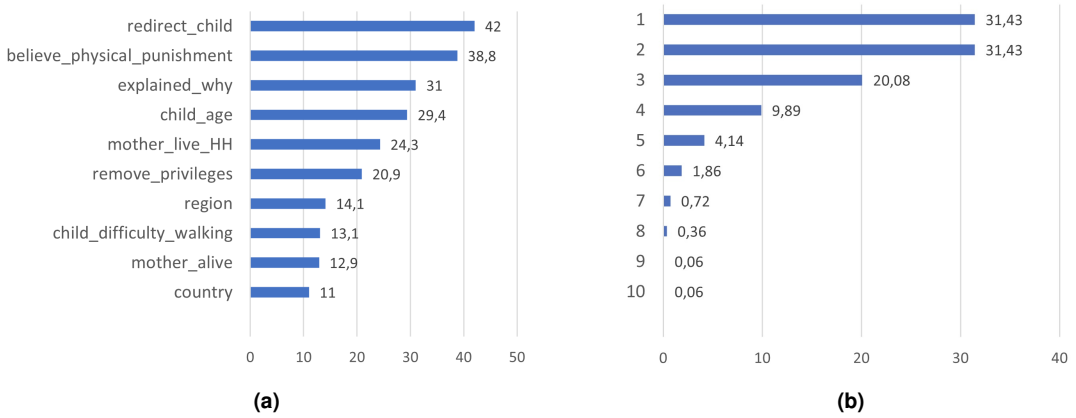

Figura 1. Avaliação das explicações utilizando o método CSSE. (a) Percentual de importância dos principais atributos; (b) Percentual de atributos alterados por contrafactual gerado.

FCD5), com 38.8%, reforçando que a crença na necessidade da punição física permanece como um dos principais fatores associados ao uso de práticas disciplinares violentas. O atributo '*explicou por que o comportamento é inadequado*' ("*explained_why*" - FCD2B) aparece em terceiro, com 31%, sugerindo que explicar à criança por que seu comportamento é inadequado é essencial. Em seguida, '*idade de ingresso da criança na escola*' ("*child_age*" - schage), com 29.4%, indicando que características relacionadas à idade em que a criança entrou na escola também influenciam o risco de exposição à disciplina violenta. Finalmente, '*a mãe reside no mesmo domicílio*' ("*mother_live_HH*" - HL13), responsável por 24.3% das modificações, mostra que a ausência da mãe no domicílio é um aspecto essencial do ambiente familiar que pode alterar a probabilidade de práticas disciplinares violentas. De forma geral, o CSSE permitiu identificar, de maneira interpretável, os fatores mais relevantes para a alteração das classificações, evidenciando aqueles que exercem maior influência no risco de ocorrência de disciplina violenta.

A Figura 1b também mostra que, na maioria dos casos, apenas um ou dois atributos modificados (ambos com 31.43%) foram suficientes para alterar a classe de uma instância, indicando que pequenos ajustes já podem produzir um impacto significativo.

Além da avaliação global, também foi realizada uma análise local em nível de instância. O CSSE examinou detalhadamente instâncias do conjunto de dados, identificando possíveis contrafactuais capazes de alterar sua classificação de "sofreu disciplina violenta" para "não sofreu disciplina violenta".

A instância original possuía as seguintes características: ela sofria disciplina violenta, era do país Malawi e seus responsáveis costumavam redirecioná-la para outras atividades. Os resultados apresentados na Tabela 4 mostram que, caso a criança fosse de outro país, como o Zimbabwe (*country*), ou se os responsáveis não tivessem o hábito de redirecioná-la (FCD2E - "*redirect_child*"), sua classificação mudaria para não sofreu disciplina violenta. Ressalta-se que, de todos os países disponíveis do continente africano, Zimbabwe possui a menor taxa de disciplina física violenta, 38,5% contra 59,5% em Malawi.

Quanto ao atributo FCD2E, embora o redirecionamento da criança seja classifi-

cado como uma prática não agressiva, o modelo indicou que sua ausência está associada à redução da probabilidade de disciplina violenta, o que pode parecer contraintuitivo. Esse comportamento pode ser parcialmente explicado pelos dados do Malawi: entre as crianças que sofreram disciplina violenta, a proporção de responsáveis que relataram redirecionamento foi de 60,5%, enquanto entre aquelas que não sofreram, essa proporção foi de 38,7%. Entre aqueles que não sofreram disciplina violenta, o comportamento é o oposto (61,2% relataram não direcionamento, contra 38,7% que direcionavam).

Esse padrão sugere que o redirecionamento não ocorre de forma isolada, podendo coexistir com práticas agressivas no mesmo ambiente familiar. Assim, o atributo FCD2E pode refletir perfis comportamentais nos quais múltiplas estratégias disciplinares são utilizadas simultaneamente, além de possíveis interações com fatores contextuais, como o país. Dessa forma, a explicação contrafactual deve ser interpretada com cautela, pois reflete associações estatísticas aprendidas pelo modelo, e não relações causais diretas.

Tabela 4. Contrafactuais gerados pelo CSSE (Exemplo 1)

Instância	FCD2E	country	Classe
Instância original	**1**	**12**	**1**
Contrafactual 1	2	-	0
Contrafactual 2	-	18	0

Além disso, o CSSE permite examinar a direção oposta, identificando quais mudanças fariam uma instância não violenta ser classificada como violenta, possibilitando compreender tanto fatores de risco quanto de proteção. Neste sentido, a Tabela 5 apresenta um exemplo em que a criança original não sofreu disciplina violenta (Classe =0) e que apenas a mudança do atributo "*explained_why*" (FCD2B), que corresponde à prática de explicar à criança por que seu comportamento é inadequado, passando de explicar para não explicar, seria suficiente para ela vir a sofrer disciplina violenta. Esse resultado sugere que estratégias educativas, baseadas no diálogo e na explicação, pode aumentar a probabilidade da haver práticas de disciplinares não violentas, evidenciando a relevância de abordagens comunicativas no contexto da educação infantil.

Tabela 5. Contrafactuais gerados pelo CSSE (Exemplo 3)

Instância	FCD2B	Classe
Instância original	**1**	**0**
Contrafactual 1	2	1

6. DISCUSSÕES

Os resultados da trabalho evidenciam a necessidade de compreender mais profundamente os fatores culturais, sociais e familiares que sustentam a prática da disciplina física violenta no conjunto de países selecionados. Observa-se que uma parcela significativa das crianças presentes no conjunto de dados foi exposta a violência física, o que revela a persistência da crença de que medidas agressivas fazem parte do processo educativo. Essa percepção costuma estar enraizada em tradições culturais, em hábitos transmitidos ao longo das gerações e em contextos socioeconômicos que dificultam o acesso a práticas alternativas de cuidado e educação baseadas no diálogo e na resolução não violenta de conflitos.

A análise contrafactual obtida com o algoritmo CSSE permitiu identificar atributos específicos que, quando modificados, poderiam alterar a classificação de uma criança como vítima de disciplina física violenta. Entre os fatores mais relevantes estão comportamentos como gritos, insultos ou xingamentos dirigidos à criança, além de crenças que justificam o uso do castigo físico como forma de disciplinar. Esses resultados indicam que a redução da violência requer não apenas a eliminação de agressões físicas diretas, mas também uma transformação nos padrões de comunicação familiar e nos valores culturais que naturalizam tais práticas, oferecendo assim *insights* importantes para intervenções sociais mais direcionadas. Esses achados estão alinhados aos resultados de Ward et al. [2022], que evidenciam a associação entre normas culturais, crenças parentais e o uso de práticas disciplinares físicas violentas.

Nesse cenário, torna-se fundamental pensar em ações concretas capazes de apoiar pais, responsáveis e comunidades na adoção de práticas de disciplina não violentas. Programas de orientação parental, por exemplo, podem oferecer ferramentas práticas para lidar com comportamentos desafiadores sem recorrer à agressão, ensinando técnicas como estabelecimento de rotinas, comunicação assertiva, negociação de limites e formas de corrigir a criança preservando sua integridade emocional.

Além disso, campanhas de conscientização pública, veiculadas em escolas, rádios locais, redes sociais e outros meios de comunicação de grande alcance, podem contribuir para desconstruir a ideia, ainda muito difundida, de que gritar, humilhar ou aplicar castigos físicos é uma forma aceitável de educar.

7. CONCLUSÃO

Os resultados obtidos mostram que a disciplina física violenta permanece como uma prática amplamente difundida em diversos países do continente escolhido, sustentada por fatores culturais, sociais e familiares que reforçam a crença de que a agressão faz parte do processo educativo. A análise contrafactual realizada com o algoritmo CSSE evidenciou que mudanças específicas em comportamentos e crenças parentais podem alterar significativamente o risco de exposição das crianças à violência, destacando a importância de intervenções focadas na comunicação familiar, na orientação parental e na transformação de valores profundamente enraizados. Assim, o estudo reforça a necessidade de políticas públicas integradas que promovam práticas educativas não violentas, ampliem o acesso a informações qualificadas e envolvam comunidades na construção de ambientes mais seguros, respeitosos e acolhedores para o desenvolvimento infantil.

7.1. Limitações e Trabalhos Futuros

Embora este estudo ofereça contribuições relevantes, algumas limitações precisam ser reconhecidas. A análise baseia-se em informações autorreferidas da base MICS, o que pode introduzir vieses de resposta e diferenças culturais na forma como a disciplina física violenta é percebida e relatada. Além disso, embora as explicações contrafactuais forneçam indicações úteis e interpretáveis sobre possíveis mudanças de comportamento, elas não permitem estabelecer relações de causalidade, devendo ser entendidas apenas como associações estatísticas. Como perspectivas futuras, recomenda-se ampliar o escopo da pesquisa para outras regiões do mundo, possibilitando comparações interculturais, bem como integrar técnicas de inferência causal, como modelos causais estruturais,

para fortalecer a interpretação dos contrafactuais. A inclusão de dados de múltiplas rodadas de pesquisa também pode aprofundar a compreensão sobre a evolução das práticas disciplinares ao longo do tempo, aumentando a robustez das evidências para subsidiar políticas públicas mais eficazes.

Referências

A. B. Arrieta et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 2020.

J. Bergstra e Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

R. Caruana et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD*, 2015.

C. S. Carvalho, J. C. Mattos, e M. S. Aguiar. Interpretabilidade e justiça algorítmica: Avançando na transparência de modelos preditivos de evasão escolar. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1658–1673. SBC, 2024.

M. de Sousa Balbino, L. E. Z. Gálvez, e C. N. Nobre. CSSE - an agnostic method of Counterfactual, Selected, and Social Explanations for classification models. *Expert Systems with Applications*, 228:120373, 2023. ISSN 0957-4174. URL `https://www.sciencedirect.com/science/article/pii/S0957417423008758`.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

O. Gomez, S. Holter, J. Yuan, e E. Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)*, pages 31–35. IEEE, 2021.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, e D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5): 93, 2018. doi: 10.1145/3236009.

R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, e F. Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, PP:1–1, 12 2019. doi: 10.1109/MIS.2019.2957223.

A.-H. Karimi, G. Barthe, B. Schölkopf, e I. Valera. A survey of algorithmic recourse: contrastive explanations and counterfactuals. *ACM Computing Surveys (CSUR)*, 55(5): 1–29, 2022.

K. Khan, W. Ahmad, M. N. Amin, e A. Ahmad. A systematic review of the research development on the application of machine learning for concrete. *Materials*, 15(13): 4512, 2022. doi: 10.3390/ma15134512. URL `https://doi.org/10.3390/ma15134512`.

M. Kraus, A. Baumann, e S. Feuerriegel. Algorithmic counterfactual explanations for robustness, transparency and accountability in predictive decision-making. *European Journal of Operational Research*, 282(1):92–105, 2020. doi: 10.1016/j.ejor.2019.09.040. URL `https://arxiv.org/abs/1909.12481`.

G. Lemaître, F. Nogueira, e C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

M. R. Marcolino et al. Análise preditiva do desempenho acadêmico em ambientes virtuais de aprendizagem: uma abordagem com aprendizado de máquina otimizado. 2025.

M. Medvedeva, M. Wieling, e M. Vols. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212, 2022. doi: 10.1007/s10506-021-09306-3. URL https://doi.org/10.1007/s10506-021-09306-3.

R. K. Mothilal, A. Sharma, e C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 607–617. ACM, 2020. doi: 10.1145/3351095.3372850. URL https://arxiv.org/abs/1905.07697.

S. Musman, V. M. d. A. Passos, I. B. R. Silva, e S. M. Barreto. Avaliação de um modelo de predição para apneia do sono em pacientes submetidos a polissonografia. *Jornal Brasileiro de Pneumologia*, 37:75–84, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, e E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

D. J. M. Reddy, S. Regella, e S. R. Seelam. Recruitment prediction using machine learning. In *2020 5th international conference on computing, communication and security (ICCCS)*, pages 1–4. IEEE, 2020.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, e K.-R. Müller. Explainable artificial intelligence: Interpreting, explaining and visualizing deep learning. *Cognitive Computation*, 2023. doi: 10.1007/s12559-023-10179-8.

A. Selbst et al. Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, e N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016.

J. J. d. D. d. Silva. Interpretabilidade de modelos de aprendizado de máquina: uma abordagem baseada em árvores de decisão. 2023.

S. M. Sinhorinho e A. T. M. S. de Moura. Uso de disciplina violenta na infância: percepções e práticas na estratégia saúde da família. *Revista Brasileira de Medicina de Família e Comunidade*, 17(44):2835–2835, 2022.

UNICEF. Violent discipline. Technical report, UNICEF MENA Regional Office, 2019. URL https://www.unicef.org/mena/sites/unicef.org.mena/files/2019-02/ViolentDiscipline-28Jan19_0.pdf.

S. Wachter, B. Mittelstadt, e C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard journal of law technology*, 31: 841–887, 04 2018. doi: 10.2139/ssrn.3063289.

K. P. Ward, S. J. Lee, A. C. Grogan-Kaylor, J. Ma, e G. T. Pace. Patterns of caregiver aggressive and nonaggressive discipline toward young children in low- and middle-income countries: A latent class approach. *Child Abuse Neglect*, 128:105606, 2022. URL https://www.sciencedirect.com/science/article/pii/S0145213422001260.