Um *Framework* Explicável para Predições Clínicas: Integração de Raciocínio Baseado em Casos, *Clustering* e Contrafactuais

Eduardo C. Radaelli[1], **Isabel C. Reinheimer**[1], **Joaquim V. C. Assunção**[1], **Carlos E. Poli-de-Figueiredo**[2], **Luís A. L. Silva**[1]

[1]Programa de Pós-Graduação em Ciência da Computação
Univ. Federal de Santa Maria (UFSM) – Santa Maria, RS – Brazil

[2]Programa de Pós-Graduação em Medicina e Ciências da Saúde
Pontifícia Univ. Católica do Rio Grande do Sul (PUCRS) – Porto Alegre, RS – Brazil

{ecradaelli, joaquim, luisalvaro}@inf.ufsm.br
cristinareinheimer@gmail.com, cepolif@pucrs.br

Abstract. *This work proposes an explainable artificial intelligence (XAI) framework for generating counterfactual explanations in clinical data, integrating Case-Based Reasoning (CBR), Clustering, and counterfactual methods. A Deep Learning (DL) model predicts the label of a query case, while similar cases of a different class are grouped into comparable clinical profiles. To enhance the diversity and interpretability of the generated counterfactuals, sampled cases from these groups guide minimal and plausible attribute modifications.*

Resumo. *Este trabalho propõe um framework de inteligência artificial explicável (XAI) para geração de explicações contrafactuais em dados clínicos, integrando Raciocínio Baseado em Casos (CBR), Clustering e métodos contrafactuais. O modelo de Deep Learning (DL) prediz o rótulo de um caso consulta, enquanto casos similares, porém de classe diferente, são agrupados em perfis clínicos semelhantes. Para melhorar a diversidade e interpretabilidade dos contrafactuais gerados, casos amostrados destes grupos orientam alterações mínimas e plausíveis nos atributos do caso.*

1. Introdução

Modelos de inteligência artificial (IA) têm sido aplicados com sucesso na análise de registros eletrônicos de saúde, apoiando decisões clínicas em diferentes contextos. Contudo, a falta de explicabilidade ainda representa um desafio, uma vez que justificativas compreensíveis e úteis são essenciais em ambientes clínicos [Minh et al. 2022]. A inteligência artificial explicável (XAI) busca promover transparência no raciocínio computacional e apoiar profissionais de saúde na compreensão das predições. Diversos métodos têm sido explorados para a geração de explicações contrafactuais, incluindo *Group-CFE* [Warren et al. 2023], que modifica variáveis de forma coletiva em grupos de instâncias similares; *DisCERN* [Wiratunga et al. 2021], que utiliza vizinhos mais próximos de classes diferentes (NUNs) combinados com medidas de relevância, como *SHAP*; e *NICE* [Brughmans et al. 2024], que adota um processo iterativo de substituição de atributos orientado por funções de recompensa para maximizar plausibilidade e esparsidade. Apesar destes avanços, essas abordagens ainda apresentam limitações quanto à diversidade das explicações geradas e à coerência clínica das alterações propostas.

Este trabalho propõe um *framework* que integra (i) Raciocínio Baseado em Casos (*CBR*), (ii) *Clustering* e (iii) explicações contrafactuais para ampliar a diversidade e a confiabilidade das decisões produzidas por algoritmos de *Deep Neural Network (DNN)*. A proposta utiliza *CBR* para justificar decisões com base em casos semelhantes, *Clustering* para organizar pacientes em grupos com perfis similares e métodos de explicação contrafactuais para ilustrar como pequenas alterações hipotéticas no perfil podem modificar o desfecho da condição do paciente. O método é aplicado a dados sobre Doença Renal Crônica (DRC) [Iimori et al. 2018], onde fatores de risco (diabetes, hipertensão e histórico cardiovascular) são usados como fatores clínicos (sub-rótulos) para avaliar grupos de casos construídos a partir de resultados de consultas *CBR*.

2. Geração de Explicações Contrafactuais baseadas em Grupos de Casos Mais Próximos de Classe Diferente

O *framework* de *XAI* proposto, descrito no Algoritmo 1, adota uma arquitetura *twin-system* que integra um modelo preditivo baseado em *DNN* a um módulo explicativo fundamentado em *CBR* e *Clustering*. O caso consulta é avaliado por ambos os modelos (*DNN* e *KNN* com decisão majoritária). Havendo divergência, o processo é interrompido; caso contrário, o rótulo da *DNN* é atribuído ao caso. Em seguida, os vizinhos de classe oposta (*NUNs*) são identificados e agrupados via *K-Means*, com o número de *clusters* definido pelo método do cotovelo. A qualidade desses grupos é avaliada por métricas como *Silhouette Score* e pureza clínica em relação a fatores relevantes para a análise de DRC (hipertensão, histórico cardiovascular, diabetes). Apenas os *clusters* válidos seguem para a etapa explicativa. Para cada grupo selecionado, escolhe-se um caso representativo (mais similar, medoide ou aleatório) e, a partir dele, geram-se explicações contrafactuais iterativas (Algoritmo 2). O resultado são explicações contrafactuais individualizadas e interpretáveis, ancoradas em grupos de pacientes clinicamente semelhantes, o que amplia a diversidade e a consistência das decisões do modelo.

O Algoritmo 2 ilustra o processo de geração de explicações contrafactuais, apresentando o Caso Consulta e o Caso Comparado (*NUN*) com predições distintas pelo modelo *DNN* para progressão de DRC. As variáveis são ordenadas por importância (*SHAP*), guiando alterações graduais no Caso Consulta. Cada tentativa do algoritmo modifica uma variável até que a predição seja alterada, identificando o conjunto mínimo e mais relevante para gerar a contrafactual.

A Figura 1 ilustra o processo de geração de explicações contrafactuais descrito no Algoritmo 2. O exemplo apresenta o Caso Consulta e o Caso *NUN*, acompanhados pela lista de importância das variáveis, obtida pelo *SHAP*. As alterações são aplicadas de forma hierárquica e gradual. Na primeira iteração, o atributo mais importante do Caso Consulta substituído pelos valores do caso NUN. Por fim, o modelo *DNN* é executado para verificar se ocorreu a mudança na classe predita. Não ocorrida, a alteração é mantida e faz-se o mesmo para o próximo atributo, e assim por diante, até que a modificação do rótulo seja obtida. Por fim, é gerado um texto contrafactual que sintetiza o processo realizado, descrevendo de forma interpretável as modificações necessárias para que o Caso Consulta alterasse sua classificação. Esse método resultou em uma explicação contrafactual plausível e clinicamente coerente, evidenciando quais variáveis exerceram maior influência sobre a decisão do modelo (e.g., a presença de hipertensão e a ausência de histórico de doença cardiovascular).

Algorithm 1 Geração de Explicações Contrafactuais usando Grupos de Casos Mais Próximos de Classe Diferente (*NUNs*)

Require: $queryCase, dnnModel, caseBase, clusterCriteria, cfCriteria$
Ensure: $counterfactuals$
$labelDNN \leftarrow$ **execute**$(queryCase, dnnModel)$ ▷ Predição DNN
$labelCBR \leftarrow$ **execute**$(queryCase, KNNAlg, caseBase)$ ▷ Predição CBR
if $labelDNN \neq labelCBR$ **then**
  return $\emptyset$ ▷ Retornar se houver conflito entre predições DNN e CBR
else
  $queryCase.label \leftarrow labelDNN$ ▷ Atribuir rótulo da DNN ao caso de consulta
end if
$casesNUN \leftarrow$ **execute**$(queryCase, NUNAlg, caseBase)$ ▷ Recuperar casos NUNs
$clustersNUN \leftarrow$ **K-Means**$(casesNUN, "Elbow")$ ▷ Formar grupos de casos NUNs via K-Means
$counterfactuals \leftarrow [\,]$ ▷ Inicializar lista de contrafactuais
for all $cluster \in clustersNUN$ **do**
  $evalCluster \leftarrow$ **evalCluster**$(cluster, clusterCriteria)$ ▷ Avaliar *Sillhouette Score*, pureza, etc
  if isValidCluster$(evalCluster)$ **then**
    $clusterCase \leftarrow$ **selectCase**$(cluster, selectCriteria)$ ▷ Selecionar caso mais similar, medoid, aleatório
    $shapFeatures \leftarrow$ **execute**$(clusterCase, SHAPAlg, labelDNN)$ ▷ Relevância atributos do caso para a predição
    $candidateCF \leftarrow$ **generateCounterfactual**$($
      $queryCase, clusterCase, dnnModel, shapFeatures)$ ▷ Executar Algoritmo 2
    if $candidateCF \neq \emptyset$ **then**
      $evalCF \leftarrow$ **evalCounterfactual**$(candidateCF, cfCriteria)$ ▷ Avaliar plausibilidade, diversidade, etc
      if isValidCounterfactual$(evalCF)$ **then**
        $counterfactuals \leftarrow counterfactuals \cup \{candidateCF\}$ ▷ Contrafactuais válidos
      end if
    end if
  end if
end for
return $counterfactuals$ ▷ Retornar contrafactuais válidos

Variável	Caso Consulta	Caso Comparado (NUN)
TFGe (mL/min)	14.6	43.6
Pressão Sistólica (mmHg)	155	142
IMC	29.7	27.2
Glicose	150	120
Progressão de DRC (Rótulo)	**1**	**0**

Lista de Importância de Features *SHAP*	
Variável	**Peso**
TFGe	0.44
IMC	0.25
Glicose	0.17
Pressão Sistólica	0.14

Variável	1ª Tentativa (TFGe)	2ª Tentativa (IMC)	3ª Tentativa (Idade)
TFGe	**43.6**	**43.6 (mantém)**	**43.6 (mantém)**
Pressão Sistólica	155	155	155
IMC	29.7	**27.2**	**27.2 (mantém)**
Glicose	150	150	**120**
Predição	**1 (Não mudou)**	**1 (Não mudou)**	**0 (Mudou) → Contrafactual Gerada**

Figura 1. Processo de geração de contrafactuais, mostrando as tentativas de alteração de rótulo baseadas na importância das variáveis.

Algorithm 2 Geração Iterativa de Contrafactual usando *SHAP*.

Require: $queryCase, clusterCase, model, shapFeatures$
Ensure: $counterfactual$ (ou $\emptyset$ se não válido)
$caseCF \leftarrow queryCase$; $targetLabel \leftarrow$ labelDNN; ▷ Copia o caso e o rótulo da consulta
for $feature \in shapFeatures$ **do**
  Substituir valor de $feature$ em $caseCF$ pelo valor de $clusterCase$
  $cfLabel \leftarrow$ **execute**$(caseCF, dnnModel)$ ▷ Predição DNN para o contrafactual
  if $cfLabel = targetLabel$ **then**
    break ▷ Parar se o rótulo já for o desejado
  end if
end for
if $cfLabel = targetLabel$ **then**
  return $counterfactual$ ▷ Retornar contrafactual válido
else
  return $\emptyset$ ▷ Retornar vazio se não convergiu
end if

***Contrafactual**: em um cenário onde fosse possível alterar os resultados de TFGe para 43.6, IMC para 27.2 e Glicose para 120, as características clínicas do paciente passariam a se assemelhar àquelas observadas em pacientes organizados em um grupo caracterizado pela presença de hipertensão e ausência de histórico de doença cardiovascular".*

Resultados preliminares com 275 casos consulta são apresentados na Tabela 1, onde: i) Diversidade mede o quão distante a contrafactual está do caso consulta, isto é, quanto mais baixa, mais próxima é a contrafactual, facilitando a(s) mudança(s); ii) Plausibilidade reflete o quão plausível é o contrafactual dentro do contexto clínico; e iii) Suavidade avalia se as alterações nas variáveis ocorrem de forma gradual, evitando mudanças artificiais. Os resultados indicaram que os métodos *NiCE* e *Group-CFE Cluster* 3 alcançaram as maiores médias de plausibilidade (1,000), sugerindo que as explicações geradas são clinicamente coerentes. O método *NiCE* também apresentou a melhor suavidade (0,836), evidenciando transições graduais nas variáveis modificadas. Em contrapartida, o *Group-CFE Cluster* 3 obteve maior diversidade (0,225), refletindo maior variação entre os contrafactuais. Esses achados demonstram um equilíbrio desejável entre plausibilidade, diversidade e suavidade, reforçando a robustez interpretativa do *framework* proposto.

Tabela 1. Resultados médios entre melhores explicações contrafactuais.

Método	Diversidade	Plausibilidade	Suavidade
DisCERN	0,190	0,857	0,794
NICE	0,179	**1,000**	**0,836**
Group-CFE Cluster 3	0,225	**1,000**	0,750
Casos Medoides dos Clusters	**0,142**	0,906	0,825

3. Conclusão

Este trabalho apresenta um *framework* de XAI que integra *CBR*, *Clustering* e métodos de geração de contrafactuais para aprimorar a interpretabilidade e a utilidade clínica de explicações de modelos de DL aplicados em dados clínicos. A abordagem permite gerar explicações diversificadas e clinicamente coerentes, reusando grupos de casos com perfis semelhantes para orientar alterações mínimas nos atributos dos pacientes.

Referências

Brughmans, D., Leyman, P., and Martens, D. (2024). Nice: an algorithm for nearest instance counterfactual explanations. *Data mining and knowledge discovery*, 38(5):2665–2703.

Iimori, S., Naito, S., Noda, Y., Sato, H., Nomura, N., Sohara, E., Okado, T., Sasaki, S., Uchida, S., and Rai, T. (2018). Prognosis of chronic kidney disease with normal-range proteinuria: The ckd-route study. *PLOS ONE*, 13(1):e0190493.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. In *Artificial Intelligence Review*, volume 55, pages 3503–3568.

Warren, G., Guéret, C., Keane, M. T., and Delaney, E. (2023). Explaining groups of instances counterfactually for xai: A use case, algorithm and user study for group-counterfactuals. In *Proc. of the ACM Conf. on Explainable AI*, pages 123–134.

Wiratunga, N., Wijekoon, A., Nkisi-Orji, I., Martin, K., Palihawadana, C., and Corsar, D. (2021). Discern: Discovering counterfactual explanations using relevance features from neighbourhoods. In *Proc. of the Int. Conf. on Machine Learning*, pages 1–10.