Avaliando as Limitações e Potenciais do Algoritmo k-Vizinhos Mais Próximos (kNN) na Imputação de Dados Clínicos

Izadora Monken Ganem[1], Guilherme D. Bianco[3], José Carlos Serufo Filho[1], Luciano Lima[1], Leonardo Rocha[2], Marcos André Gonçalves[1]

[1] Universidade Federal de Minas Gerais (UFMG)
[2]Universidade Federal de São João del-Rei (UFSJ)
[3]Universidade Federal da Fronteira Sul (UFFS)
{izadoraganem@gmail.com, guilherme.dalbianco@uffs.edu.br, serufo@ufmg.br, luciano@dcc.ufmg.br, lcrocha@ufsj.edu.br, magoncalv@gmail.com}

Resumo. *A qualidade dos dados é crucial para a eficácia das soluções de Aprendizado de Máquina na saúde, sendo a ausência de valores um problema crítico e comum. Este estudo compara os métodos de imputação MissForest e MICE aplicados a dados clínicos de mais de 16.000 pacientes com COVID-19, destacando a precisão superior do MissForest, associada a alto custo computacional. Propomos um imputador baseado em KNN, otimizado para este contexto. Embora menos preciso, a eficiência computacional foi significativamente melhor. Uma análise mostrou que o desempenho do KNN é influenciado pela similaridade de vizinhança e entropia local. Em cenários homogêneos, o KNN se mostrou competitivo, sugerindo seu potencial para estratégias de imputação híbridas que combinem eficiência e robustez.*

Abstract. *Data quality is crucial for the effectiveness of Machine Learning solutions in healthcare, with missing values being a common and critical issue. This study compares the MissForest and MICE imputation methods applied to clinical data from over 16,000 COVID-19 patients, highlighting the superior accuracy of MissForest, associated with high computational cost. We propose a KNN-based imputer optimized for this context. Although less accurate, its computational efficiency was significantly better. An analysis showed that KNN performance is influenced by neighborhood similarity and local entropy. In homogeneous scenarios, KNN proved competitive, suggesting its potential for hybrid imputation strategies that combine efficiency and robustness.*

1. Introdução

Soluções para problemas na área de Saúde tem recaído cada vez mais em técnicas de Aprendizado de Máquina (AM) no intuito de se obter alta acurácia e escalabilidade [Paiva et al. 2023]. No entanto, a efetividade dessas soluções depende fortemente da qualidade dos dados utilizados em seu desenvolvimento, os quais são, frequentemente, inconsistentes ou incompletos. Uma das principais causas dessas inconsistências/incompletudes é a presença de valores ausentes, que podem ocorrer devido a uma variedade de fatores, como falhas na integração de múltiplas fontes de dados, erros nos instrumentos de coleta e/ou transmissão, além da omissão motivada pela não obrigatoriedade de preenchimento. Embora indesejáveis, tais lacunas são comuns em bases

de dados reais na área da saúde [Emmanuel et al. 2021] e, quando não tratadas adequadamente, podem introduzir vieses analíticos ou comprometer a acurácia de modelos preditivos voltados ao diagnóstico ou prognóstico clínico, com sérias implicações práticas.

O tratamento automático de valores ausentes tem sido amplamente estudado na literatura por meio de estratégias conhecidas como *imputação de dados*, que consistem no preenchimento dos dados ausentes com base em métodos estatísticos e/ou algoritmos de AM [Liu et al. 2023]. No contexto da saúde, essa tarefa é particularmente crítica, uma vez que valores imputados de forma imprecisa podem distorcer substancialmente a interpretação do estado clínico de um paciente e, mais grave ainda, influenciar equivocadamente decisões médicas.

Neste contexto, este trabalho tem dois objetivos principais. O primeiro é realizar uma comparação de métodos de imputação aplicados a uma grande base de dados com informações de pacientes diagnosticados com COVID-19. Essa base engloba dados de mais de 16.000 pacientes coletados em 25 hospitais, incluindo sinais vitais, comorbidades e exames laboratoriais, refletindo um cenário clínico desafiador, com alta heterogeneidade e variabilidade dos casos. Nesta análise, comparamos 5 métodos de imputação: (i) média de valores das demais instâncias; (ii) uma estratégia baseada nos valores dos vizinhos mais próximos (KNN); (iii) uma estratégia baseada em redes neurais adversárias (GAIN); (iv) o MissForest [Shadbahr et al. 2023] e o MICE [Jarrett et al. 2022], sendo os dois últimos considerados o estado-da-arte em imputação. De fato, nossos experimentos mostraram que o MissForest e o MICE geraram imputações mais próximas dos valores reais na base COVID-19, mas com alto custo computacional.

O segundo objetivo deste trabalho foi investigar a seguinte hipótese: *"Imputar dados ausentes com base em pacientes que apresentam quadros clínicos semelhantes ao caso-alvo tende a produzir imputações precisas, devido à correspondência entre condições de saúde"*. A partir dessa premissa, implementamos e avaliamos um novo imputador também baseado no KNN (denominamos de KNN++), com parâmetros (número de vizinhos, atributos utilizados no cálculo de similaridade, etc.) ajustados para a base de dados da COVID-19 utilizada em nossos experimentos. Em uma análise experimental comparativa, o desempenho do KNN++ em termos de efetividade foi inferior ao dos métodos MissForest e MICE. Por outro lado, apresentou eficiência computacional significativamente superior, o que motivou uma terceira contribuição deste trabalho que consiste em uma investigação aprofundada sobre os *"quando (funciona)" e "porquês"* da performance do KNN++ em imputação nesta base.

Identificamos que o KNN++ é fortemente influenciado por dois fatores principais: a similaridade média entre os vizinhos (i.e., o quão semelhantes são ao paciente-alvo) e a entropia dentro da vizinhança (i.e., a dispersão dos valores da variável a ser imputada). Concluímos que, em cenários nos quais a vizinhança é mais homogênea — com alta similaridade média e baixa entropia — a efetividade do KNN++ pode ser comparável ao dos métodos estado-da-arte, com custo computacional significativamente menor (aproximadamente 70 e 2.850 vezes menor que tempo do MICE e MissForest, respectivamente). Esses resultados abrem caminho para o desenvolvimento de abordagens híbridas, que combinem a eficiência do KNN++ com a robustez de métodos avançados de imputação.

2. Fundamentação Teórica

Imputação de dados ausentes em bases clínicas é tema recorrente na literatura [Chen et al. 2023], dada a complexidade e a sensibilidade envolvidas na manipulação de dados no contexto da saúde. Diversos métodos têm sido investigados para aprimorar a a precisão de modelos preditivos. O MissForest [Shadbahr et al. 2023], considerado estado da arte, é um método não paramétrico que utiliza florestas aleatórias para imputar valores faltantes. Iterativamente, treina modelos com variáveis completas para estimar os dados ausentes, atualizando a matriz até a convergência ou um limite de iterações. É robusto a relações não lineares e *outliers*, sendo eficaz em cenários complexos de imputação.

Outros dois algoritmos com bons resultados reportados na literatura são o *Multivariate Imputation by Chained Equations* (MICE) [Jarrett et al. 2022] e o GAIN (*Generative Adversarial Imputation Nets*) [Yoon et al. 2018]. O MICE realiza imputação iterativa com base em modelos de regressão condicionais. Cada variável é modelada como função das demais, formando uma cadeia de imputações. Inicialmente, utiliza-se uma estimativa simples (média ou moda), seguida por imputações sucessivas com modelos ajustados, refinadas a cada ciclo até a convergência. O GAIN é um método baseado em redes adversariais generativas (GANs), composto por um gerador, que imputa valores ausentes, e um discriminador, que tenta distinguir valores reais dos imputados. O gerador utiliza variáveis observadas e uma máscara de ausência para produzir estimativas, enquanto o discriminador avalia os dados completos. GAIN destaca-se por capturar relações não lineares e estruturas complexas, sendo eficaz mesmo com alta proporção de dados ausentes.

Apesar da robustez, métodos avançados de imputação apresentam alto custo computacional, o que limita sua aplicabilidade. Nesse contexto, abordagens mais simples, como o K-Nearest Neighbors (KNN), tornam-se atrativas por sua eficiência, facilidade de uso e interpretabilidade. O KNN imputa valores ausentes com base na média ou moda dos k vizinhos mais próximos, explorando padrões locais nos dados. Embora sensível à escolha de k, à escala das variáveis e à variabilidade dos dados, seu baixo custo computacional e fácil interpretação dos resultados justifica sua exploração na tarefa de imputação.

3. Metodologia de Avaliação

O primeiro objetivo deste trabalho é comparar 5 métodos de imputação (Média, GAIN, MICE, MissForest e KNN). Utilizamos uma coorte multi-hospitalar descrita em [Marcolino et al. 2021], com dados de cerca de 16 mil pacientes hospitalizados com COVID-19 no Brasil, provenientes de 25 hospitais. A coorte inclui 62 variáveis clínicas desde a admissão até os desfechos, com aproximadamente 16% de dados ausentes, o que dificulta a construção de modelos preditivos. Para um estudo mais controlado, foi selecionado um recorte da primeira onda da COVID-19 [Marcolino et al. 2021].

Realizamos a avaliação dos métodos na imputação de valores para a variável *Sat/Fio2* (saturação por fração inspirada de oxigênio), apontada em trabalhos anteriores como a de maior relevância na predição do desfecho óbito [Lana et al. 2025]. Removemos todas as tuplas que continham valores faltantes para essa variável, uma vez que não seria possível avaliar a eficácia dos métodos nesses casos. Para permitir essa avaliação, foram inseridos valores ausentes de forma artificial na variável em 50% das tuplas restantes utilizando a técnica *Missing Completely at Random* (MCAR) [Little and Rubin 2019]. A escolha dessa porcentagem de dados teve o objetivo de preservar o restante da base para servir como referência durante a imputação no aprendizado dos modelos.

Para garantir que a comparação entre os métodos pudesse ser validada estatisticamente, propomos um experimento de validação cruzada de 10 partições. Dividimos em 10 conjuntos distintos, os 50% da base que tiveram os valores da variável *Sat/Fio2* removidos, no qual nove eram usados para treinar o modelo (os valores reais de *Sat/Fio2* eram considerados) e a décima partição era utilizada para teste (os valores eram imputados pelos métodos). Esse processo foi repetido 10 vezes para cada método, garantindo que cada instância fosse utilizada como teste uma única vez. Para avaliar a efetividade dos modelos, os valores imputados pelos mesmos foram comparados com os valores reais utilizando duas métricas tradicionais: Erro Absoluto Médio (MAE), que mede a diferença média entre os valores reais e os imputados; e o Erro Quadrático Médio (MSE), que calcula a média dos quadrados dessas diferenças [Emmanuel et al. 2021].

Para investigar a hipótese principal desse trabalho (*"Imputar dados ausentes com base em pacientes que apresentam quadros clínicos semelhantes ao caso-alvo tende a produzir imputações precisas, devido à correspondência entre condições de saúde"*), realizamos ajustes nos parâmetros do KNN com base nos fatores que mais influenciam no seu desempenho, tais como: métricas de distância, tamanho da vizinhança, e quais variáveis considerar para melhor representar os pacientes. Nesse caso, realizamos uma avaliação preliminar, considerando apenas as porções de treino da coleção, para identificar quais as variáveis mais relevantes para representar os pacientes. Utilizando uma estratégia de *permutation importance* [Breiman 2001], avaliamos as variáveis consideradas para representar os pacientes no processo de imputação do KNN.

Nesses experimentos preliminares, variamos também a métrica de distância (distância de Manhattan, distância euclidiana e similaridade de cossenos) e o número de vizinhos a ser considerado, e os melhores resultados (3, 5 e 10), buscando otimizar o KNN. Além disso, para reduzir a dimensionalidade do conjunto de dados e melhorar a acurácia do método, foi realizada uma seleção de variáveis preditoras mais correlacionadas com o alvo de imputação. Isso foi feito por meio de uma permutação de *features* considerando tanto a relevância estatística quanto a relevância clínica para o problema, resultando em um conjunto de 10 variáveis utilizadas na predição:*vm* (ventilação mecânica), *po2/fio2* (pressão parcial de oxigênio/fração inspirada de oxigênio), *fr* (frequência respiratória), *óbito, HospitalGDP, plaquetas, fc* (frequência cardíaca),*bicarbonato, ph* (ph sanguíneo), *pcr* (proteína c reativa). Essa versão ajustada do KNN, a qual denominamos de KNN++, foi então avaliada no procedimento de validação cruzada descrito anteriormente e comparada aos demais métodos.

Aprofundamos a análise do KNN++ avaliando seu desempenho sob diferentes perspectivas. Ao final das 10 execuções de validação cruzada, todos os pacientes tiveram a variável *Sat/Fio2* imputada. Cada paciente foi representado em um espaço bidimensional, com o eixo X indicando a distância média aos k vizinhos mais próximos e o eixo Y, a entropia da variável *Sat/Fio2* entre esses vizinhos, calculada pela entropia diferencial [Beirlant et al. 1997]. Esse espaço foi dividido em quatro quadrantes: 1) baixa distância/baixa entropia; 2) baixa distância/alta entropia; 3) alta distância/baixa entropia; 4) alta distância/alta entropia. O objetivo é investigar a relação entre erro de imputação e estrutura das vizinhanças, comparando os métodos KNN++, MissForest e MICE em cada quadrante, com validação estatística via teste de Wilcoxon com correção de Bonferroni [Lana et al. 2025].

4. Resultados

Nessa seção apresentamos os resultados obtidos a partir da aplicação da metodologia descrita na seção anterior. Dividimos nossas analises em duas linhas: (*i*) avaliação comparativa entre 5 métodos de imputação; e (*ii*) investigação das limitações e potenciais do KNN.

4.1. Comparação entre métodos de imputação

Foram comparados os cinco métodos de imputação descritos na Seção 2, três baseados em aprendizado de máquina (KNN, MICE, MissForest) e um em aprendizagem profunda (GAIN). Consideramos também um método bem simples que imputa o valor baseado na média do mesmo nas demais instâncias da coleção (Média). A variável imputada (*Sat/Fio2*) possui média de 369.83, o que nos permite ter uma referência para avaliar os erros introduzidos por cada método. A Tabela 1 apresenta os valores médios de MSE e MAE para cada estratégia, bem como seus respectivos tempos médio de execução.

Os resultados mostram maior efetividade dos métodos MICE e MissForest em relação aos demais, enquanto Média e GAIN apresentaram desempenho limitado. Embora o baixo desempenho da Média fosse esperado, a baixa performance do GAIN requer investigação. Neste estudo, utilizou-se a configuração padrão do método, o que pode ter limitado seu desempenho, já que ele é sensível à calibração de parâmetros que afetam diretamente sua convergência e aprendizado. Assim, é importante investigar se ajustes mais refinados trazem melhorias significativas, ainda que aumentem o tempo de execução. Por fim, o KNN teve desempenho modesto, ligeiramente superior ao da Média, mas com tempo de execução muito menor. Devido ao baixo custo computacional, simplicidade e interpretabilidade, o KNN foi escolhido para uma análise mais aprofundada.

Métricas	Média	KNN	KNN++	MissForest	MICE	GAIN
MAE	95.74	94.47	87.32	52.17	64.88	94.51
MSE	19061	25731	20101	11281	15742	18013
Tempo (seg.)	0.014	0.16	0.082	228.95	5.15	9.39

Tabela 1. Efetividade e Custo Computacional (tempo) dos Métodos de Imputação.

4.2. Limitações e Potenciais do KNN++

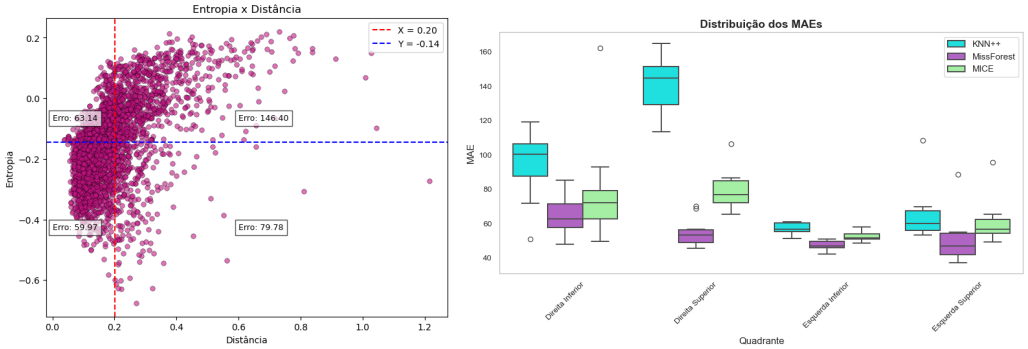
Realizamos ajustes finos nos parâmetros (número de vizinhos, atributos utilizados no cálculo de similaridade, etc.) do imputador baseado no KNN (denominamos de KNN++) para a base de dados da COVID-19. Avaliando os resultados apresentados na Tabela 2, observamos uma melhora em relação à sua versão não otimizada, mas ainda aquém dos resultados obtidos pelos métodos MissForest e MICE, mas novamente com um custo computacional significativamente menor em relação ao tempo. Mais especificamente, o desempenho do KNN++ (MAE = 86.32) permaneceu inferior ao dos métodos mais competitivos - MissForest (MAE = 52.17) e MICE (MAE = 64.88), mas com um custo computacional 70 e 2.850 vezes menor que o MICE e o MissForest, respectivamente.

Quadrante	KNN++	MissForest	MICE	Porcentagem da Coorte
Direita Inferior	79.78 (↓,*)	65.61	79.03	9%
Direita Superior	146.40 (↓,↓)	54.94	79.32	28%
Esquerda Inferior	59.97 (↓,↓)	47.25	52.59	42%
Esquerda Superior	63.14 (↓,*)	50.44	60.62	21%

Tabela 2. Média dos Erros por Quadrante com Indicação de Significância Estatística entre os erros do KNN em relação aos demais. Os símbolos ↓ e * reportam perda e empate estatístico do KNN++ em relação aos métodos.

Os pacientes foram distribuídos em quadrantes com base nas médias de dois parâmetros: a distância média aos k vizinhos mais próximos (eixo X), como medida de similaridade, e a entropia das vizinhanças (eixo Y), que reflete a heterogeneidade estrutural. O objetivo é avaliar a hipótese de que regiões com estrutura local estável (alta entropia e alta similaridade) favorecem a imputação com o KNN++. A Figura 1(a) ilustra o desempenho do KNN++ por quadrante. A Tabela 2 complementa essa análise, apresentando os erros de imputação por quadrante, e evidenciando que o KNN++ é competitivo em regiões com alta similaridade e baixa entropia, como nos quadrantes Inferior Esquerdo e Superior Esquerdo, onde seus erros não diferem estatisticamente dos obtidos com o método MICE.

Em contrapartida, em regiões mais heterogêneas, como o quadrante Superior Direito, o KNN++ apresenta desempenho inferior. A Figura 1(b) evidencia a distribuição e variabilidade dos erros por método e quadrante, em concordância com os testes estatísticos da Tabela 2, que confirmam a equivalência entre KNN++ e MICE no quadrante Superior Esquerdo. Destaca-se a robustez do MissForest, com as menores medianas em todos os quadrantes, e a capacidade do KNN++ de alcançar desempenho competitivo em cenários específicos, ressaltando a importância da homogeneidade local na imputação. O quadrante Inferior Esquerdo, onde o KNN++ apresenta o menor erro médio, concentra 42% dos pacientes, o que reforça sua aplicabilidade prática nesse contexto.



((a)) Entropia x Distância média. **((b)) Distribuição de erros por quadrante.**

Figura 1. Avaliação do Impacto da Estrutura Local dos Dados na Imputação.

5. Conclusões e Trabalhos Futuros

Este estudo avaliou cinco métodos de imputação (Média, KNN, MICE, MissForest e GAIN) em uma coorte clínica relacionada à COVID-19. MICE e MissForest apresentaram os menores erros médios, embora com alto custo computacional. O KNN, apesar da maior eficiência, teve desempenho inferior, motivando uma análise aprofundada. Verificou-se que sua performance é sensível à similaridade entre vizinhos, entropia local e número de variáveis. Com base nessas observações, foi proposta uma versão aprimorada (KNN++), que apresentou desempenho competitivo com menor custo computacional. Os resultados indicam o potencial de abordagens híbridas que conciliem alta acurácia (e.g., MissForest) e eficiência (e.g., KNN++), promovendo equilíbrio entre qualidade e desempenho na imputação de dados médicos. Destaca-se, ainda, a importância de aprimorar a representação clínica dos pacientes, aspecto fundamental para a eficácia do KNN++ em contextos mais complexos. Seguiremos nesse caminho em trabalhos futuros.

Agradecimentos

Este trabalho foi apoiado por CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde e INCT-TILDIAR (408490/2024-1).

Referências

Beirlant, J., Dudewicz, E. J., Györfi, L., and van der Meulen, E. (1997). Estimating differential entropy with kernel methods. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, Z., Tan, S., Chajewska, U., Rudin, C., and Caruana, R. (2023). Missing values and imputation in healthcare data: Can interpretable machine learning help? In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, volume 209, pages 88–108. PMLR.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, 8:1–37.

Jarrett, D., Cebere, B., Liu, T., Curth, A., and van der Schaar, M. (2022). Hyperimpute: Generalized iterative imputation with automatic model selection.

Lana, F. C. B., Marinho, C. C., de Paiva, B. B. M., Valle, L. R., do Nascimento, G. F., da Rocha, L. C. D., Carneiro, M., Batista, J. d. L., Anschau, F., Paraiso, P. G., Bartolazzi, F., Cimini, C. C. R., Schwarzbold, A. V., Rios, D. R. A., Gonçalves, M. A., and Marcolino, M. S. (2025). Unraveling relevant cross-waves pattern drifts in patient-hospital risk factors among hospitalized covid-19 patients using explainable machine learning methods. *BMC Infectious Diseases*, 25(1):537.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, M., Li, S., Yuan, H., Ong, M. E. H., Ning, Y., Xie, F., Saffari, S. E., Shang, Y., Volovici, V., Chakraborty, B., et al. (2023). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial intelligence in medicine*, 142:102587.

Marcolino, M. S., Ziegelmann, P. K., Souza-Silva, M. V. R., Nascimento, I. J. B., Oliveira, L. M., and et al. (2021). Clinical characteristics and outcomes of patients hospitalized with covid-19 in brazil: Results from the brazilian covid-19 registry. *International Journal of Infectious Diseases*, 107:300–310.

Paiva, B. B. M. et al. (2023). Potential and limitations of machine meta-learning (ensemble) methods for predicting covid-19 mortality in a large in-hospital brazilian dataset. *Scientific Reports*, 13(1):3463.

Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Torné, R. V., Sala, E., Lió, P., Patel, M., Preller, J., Rudd, J. H. F., Mirtti, T., Rannikko, A. S., Aston, J. A. D., Tang, J., and Schönlieb, C.-B. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5689–5698. PMLR.