

MIST-Saúde: Uma Ferramenta Flexível de Imputação Semântica de Dados Clínicos Tabulares

Arthur Henrique Quaresma¹, Caio Fabri¹, Davi Augusto Rodrigues¹
Izadora Ganem¹, Guilherme Dal Bianco², José Carlos Serufo Filho¹
Leonardo Rocha³, Milena Marcolino¹, Marcos André Gonçalves¹

¹ Universidade Federal de Minas Gerais,

² Universidade Federal da Fronteira Sul,

³ Universidade Federal de São João del-Rei

{caiofabri, daviamrodrigues, mgoncalv}@dcc.ufmg.br

{izaganem, milenamarc, serufo, tuquaresma}@ufmg.br

guilherme.dalbianco@uffs.edu.br, lcrocha@ufs.j.edu.br

Abstract. *Missing values in clinical datasets degrade model performance and compromise downstream decisions. This paper presents MIST-Saúde, a demonstrable tool for semantic imputation of tabular clinical data. The tool integrates statistical baselines and semantic approaches based on pre-trained foundational models (e.g., TabPFN), combined in an ensemble to improve robustness under temporal drifts. MIST-Saúde is delivered as a Python library with Docker packaging, a FastAPI REST API, and a web interface for configuration, execution, and export. We report the motivation, stakeholders, development process, requirements elicitation, key functional/non-functional requirements, and the module design. As the work is still evolving, the detailed architecture, final tech stack, and full test battery remain under consolidation; nevertheless, a prototype and an installation demo video are available for demonstration.*

Resumo. *Valores ausentes em dados clínicos degradam a performance de modelos e comprometem decisões. Este trabalho apresenta o MIST-Saúde, uma ferramenta demonstrável de imputação semântica de dados tabulares clínicos. A ferramenta integra métodos estatísticos e abordagens semânticas com modelos fundacionais pré-treinados (ex.: TabPFN), combinados em ensemble para robustez frente a drifts temporais. O MIST-Saúde é entregue como biblioteca Python com Docker, API REST em FastAPI e interface web para configuração, execução e exportação dos resultados. Relatamos motivação, stakeholders, processo de desenvolvimento, elicitação e requisitos, além do desenho do módulo. Como o trabalho segue em construção, a arquitetura detalhada, as tecnologias finais e a bateria completa de testes ainda estão sendo consolidadas; ainda assim, há um protótipo e um vídeo de instalação disponíveis para demonstração.*

Palavras-chave: Imputação de dados, Inteligência Artificial, Saúde, Modelos Fundacionais, Drifts Temporais, Ensemble.

1. Introdução

A imputação de valores ausentes em dados clínicos é uma etapa crítica para assegurar qualidade e confiabilidade em modelos preditivos. Em cenários reais, falhas na coleta

e integração de sistemas produzem lacunas que podem distorcer estimativas, reduzir acurácia e comprometer decisões clínicas.

Nesse contexto, apresentamos o MIST-Saúde, uma ferramenta voltada à imputação de dados tabulares clínicos, que combina métodos tradicionais e abordagens semânticas baseadas em modelos fundacionais, com foco em robustez a *drifts* temporais. O trabalho descreve a motivação, *stakeholders*, processo de desenvolvimento, elicitação e requisitos, desenho do módulo, tecnologias e avaliação, além de disponibilizar URL pública e vídeo de demonstração.

2. Visão Geral da Ferramenta

O MIST-Saúde implementa um fluxo completo de imputação para dados tabulares clínicos: (i) ingestão e validação do *dataset*; (ii) seleção e parametrização de métodos; (iii) execução dos algoritmos e *ensemble* semântico; e (iv) análise de métricas e exportação dos resultados. A ferramenta oferece interface *web* e *API REST* para integração a fluxos de pesquisa e operação.

3. Motivação

A ausência de valores em bases de dados clínicos é um problema recorrente na área da saúde, decorrente de falhas na coleta, integração de sistemas ou omissões intencionais [Emmanuel et al. 2021]. Essa lacuna compromete a integridade dos dados, introduzindo vieses e reduzindo significativamente a acurácia de modelos de aprendizado de máquina (AM) utilizados para previsões clínicas, como estimativas de desfecho de óbito em pacientes com COVID-19 [Paiva et al. 2023].

Como exemplo, utilizamos uma coorte multicêntrica [Marcolino et al. 2021, Paiva et al. 2022, Marcolino et al. 2025], composta por aproximadamente 16.000 pacientes hospitalizados com COVID-19 no Brasil, provenientes de 25 hospitais. A base contém 62 variáveis clínicas coletadas desde a admissão até os desfechos hospitalares e apresenta cerca de 16% de dados ausentes, o que impõe desafios relevantes para a construção de modelos preditivos confiáveis.

A motivação para o desenvolvimento do MIST-Saúde reside na necessidade de soluções robustas, eficientes e contextualmente conscientes para imputação de dados, equilibrando acurácia preditiva e custo computacional. A ferramenta foi construída a partir de um estudo comparativo extenso e incorpora mecanismos para lidar com *drifts* temporais (mudanças nos padrões de distribuição dos dados ao longo do tempo) [Lana et al. 2025], utilizando compreensão semântica para gerar imputações mais coerentes com o contexto clínico do paciente. Dessa forma, reduz riscos associados a decisões médicas baseadas em dados incompletos e contribui para o avanço de aplicações confiáveis e escaláveis de inteligência artificial na saúde [Ganem et al. 2025].

4. Stakeholders e Perfis de Usuário

- **Usuários finais:** pesquisadores e analistas de dados em saúde, que utilizam a ferramenta para preparar conjuntos de dados para treinamento e validação de modelos de aprendizado de máquina.
- **Pacientes:** beneficiários indiretos, uma vez que imputações mais precisas resultam em previsões clínicas mais confiáveis, favorecendo tratamentos personalizados e melhores desfechos.

- **Médicos e especialistas clínicos:** profissionais que dependem de dados íntegros para diagnóstico, prognóstico e planejamento terapêutico.
- **Organizações e instituições:** hospitais, universidades e órgãos de saúde pública, que podem adotar a solução para gerenciamento de grandes bases clínicas e resposta a emergências sanitárias.
- **Desenvolvedores e mantenedores:** equipe de IA responsável pela evolução contínua da ferramenta e por sua integração com sistemas hospitalares existentes.

Esses grupos subsidiaram a definição de requisitos por meio de entrevistas e discussões iniciais; a validação com usuários finais e demais *stakeholders* ainda está em andamento.

5. Processo de Desenvolvimento

O desenvolvimento do MIST-Saúde seguiu o *framework Scrum* com iterações curtas (*sprints* de 1 semana), o que permitiu adaptação rápida a mudanças de requisitos e ciclos de revisão [Schwaber and Sutherland 2020]. As revisões de *sprint* foram usadas para avaliar protótipos incrementais e registrar pontos de melhoria, e definir novas tarefas.

6. Elicitação e Especificação de Requisitos

A elicitação envolveu entrevistas semiestruturadas com especialistas em saúde e IA, observações em ambiente hospitalar e revisão de literatura. A especificação foi feita por histórias de usuário (ex.: “Como pesquisador, quero selecionar diferentes métodos de imputação para processar datasets clínicos”) e casos de uso com fluxos (*upload*, escolha de algoritmo, visualização e exportação). A priorização dos requisitos seguiu a técnica MoSCoW (Must-have, Should-have, Could-have, Won’t-have) [Cohn 2005].

7. Requisitos Funcionais e Não Funcionais

7.1. Requisitos Funcionais

- Suporte a múltiplos algoritmos de imputação: média, KNN, MissForest, MICE, GAIN e ensemble com TabPFN [Hollmann et al. 2025].
- Processamento eficiente de *datasets* tabulares clínicos de alta dimensionalidade (ex.: 62 variáveis).
- Adaptação a *drifts* temporais por meio de detecção e mecanismos de reajuste semântico.
- Exposição de *API RESTful* via *FastAPI*, permitindo integração de interfaces, inferência remota, *upload* de dados, seleção de algoritmo e retorno de resultados imputados em formato *JSON* ou arquivo.
- Interface para *upload* de dados, visualização de como o dado está estruturado, configuração de parâmetros e exportação dos dados imputados (via *web* ou *API*).

7.2. Requisitos Não Funcionais

- **Segurança e Privacidade:** conformidade com a LGPD (Lei geral de proteção de dados), não armazenamento dos dados e criptografia em trânsito (*HTTPS*).
- **Desempenho:** tempo de execução otimizado e escalável para conjuntos com mais de 16.000 instâncias.
- **Usabilidade:** interface *web* intuitiva e documentação completa em português.
- **Confiabilidade:** garantia de efetividade mínima em cenários homogêneos, avaliada por métricas como *MAE* e *MSE*.
- **Portabilidade:** containerização via *Docker* para implantação local ou em nuvem.

8. Protótipo e Arquitetura

A arquitetura adota o padrão limpo/hexagonal, garantindo separação clara de responsabilidades e alta modularidade. Os principais componentes são:

- **Core:** camada em Python contendo os algoritmos de imputação (estatísticos, baseados em AM e semânticos), extensível por injeção de dependências.
- **API:** desenvolvida com FastAPI, expondo endpoints REST para integração e inferência em sistemas externos.
- **Frontend:** interface web para upload, configuração e visualização de resultados e métricas.
- **Containerização:** Docker para garantir portabilidade e facilidade de deployment.

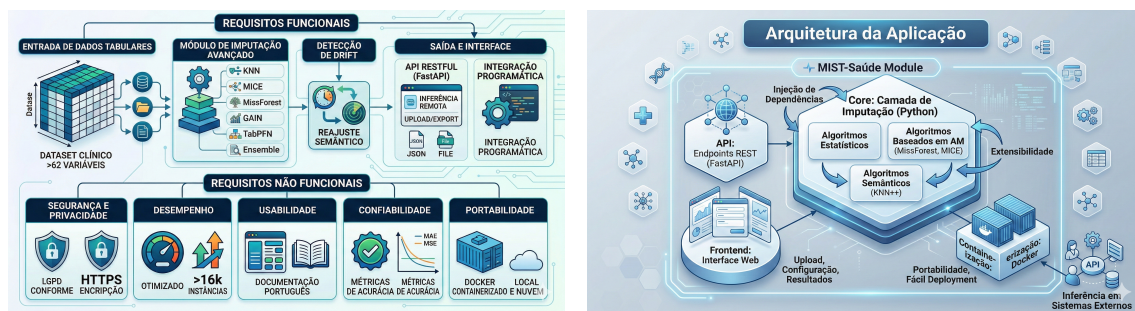


Figura 1. Resumo dos requisitos e arquitetura da aplicação.

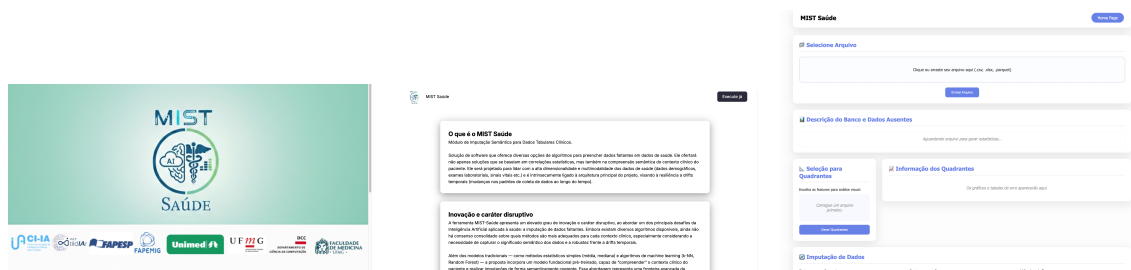


Figura 2. Tela inicial, ajuda e tela principal (com seleção da base de dados).

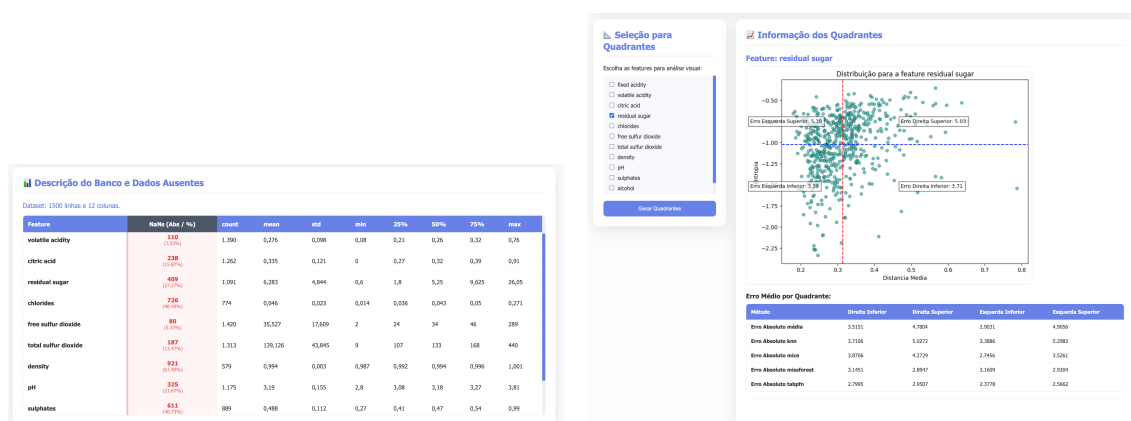


Figura 3. Tela de análise da base de dados.

Imputação de Dados

Features para **Imputar**:

- ☐ fixed acidity
- ☐ citric acid
- ☐ chlorides
- ☐ total sulfur dioxide
- ☐ pH
- ☐ alcohol
- ☐ volatile acidity
- ☐ residual sugar
- ☐ free sulfur dioxide
- ☐ density
- ☐ sulphates
- ☐ quality

Features para **Ignorar**:

- ☐ fixed acidity
- ☐ citric acid
- ☐ chlorides
- ☐ total sulfur dioxide
- ☐ pH
- ☐ alcohol
- ☐ volatile acidity
- ☐ residual sugar
- ☐ free sulfur dioxide
- ☐ density
- ☐ sulphates
- ☐ quality

Método e Ação: TABPFN

Aplicar Imputação e Baixar

Figura 4. Escolha de *features* e métodos de imputação.

9. Tecnologias

Python 3.10+, bibliotecas Pandas, NumPy, scikit-learn, implementações customizadas dos algoritmos de imputação, integração com modelos fundacionais (TabPFN). API desenvolvida com FastAPI + Pydantic, utilizando Uvicorn como servidor ASGI e Nginx como proxy reverso, com frontend em HTML, CSS e JavaScript (Vanilla JS), containerização via Docker e Docker Compose. Controle de versão com Git e prototipagem/avaliação em Jupyter Notebooks. Essas tecnologias foram escolhidas pela maturidade no ecossistema científico, ampla adoção na comunidade de dados em saúde e boa integração entre processamento, serviços e interface. O uso de Python e bibliotecas consolidadas acelera a experimentação, enquanto FastAPI, Uvicorn e Nginx garantem APIs performáticas. Docker e Docker Compose asseguram reprodutibilidade e portabilidade, facilitando implantação em diferentes ambientes.

10. Avaliação

A avaliação foi multifacetada e incluiu:

- Testes unitários (pytest)
- Testes de integração (core + API)
- Testes com datasets reais (métricas MAE, MSE, tempo de execução)

11. Demonstração

O protótipo está disponível em URL pública

(https://github.com/CacabaGV/SBCAS_MIST), acompanhado de um vídeo de demonstração que apresenta: (i) instalação/execução; (ii) configuração de métodos e parâmetros; (iii) execução da imputação; e (iv) interpretação das métricas e exportação dos resultados. (<https://www.youtube.com/watch?v=HRVoXGgMWRE>)

12. Lições Aprendidas

No planejamento, destacou-se a necessidade de mapear drifts temporais desde as fases iniciais. Durante o desenvolvimento, a heterogeneidade dos dados desafiou a parametrização de algoritmos como KNN++ (desenvolvido no artigo em que esse produto foi baseado [Ganem et al. 2025]), exigindo várias iterações. Nos testes, verificou-se que cenários homogêneos favorecem métodos simples, enquanto regiões heterogêneas demandam ensembles. A eficiência computacional foi identificada como diferencial competitivo a partir de discussões preliminares. Conclui-se que abordagens híbridas oferecem melhor equilíbrio entre robustez e custo, e que a validação contínua com usuários reais é indispensável para garantir adoção.

13. Limitações

Apesar do protótipo demonstrável e dos resultados apresentados, algumas limitações permanecem: (i) necessidade de ampliar testes com mais domínios clínicos e diferentes distribuições de ausência; (ii) validação adicional em cenários com maior heterogeneidade e *drifts* mais severos; e (iii) análise aprofundada de custo computacional em ambiente de produção com cargas variáveis.

14. Trabalhos Futuros

Planeja-se expandir o MIST-Saúde para cenários multimodais, validar em novos domínios (oncologia, cardiologia), implementar auto-seleção de métodos via meta-aprendizado, conduzir uma avaliação de usabilidade mais formal com stakeholders e avançar para TRL 7–9 com parcerias hospitalares.

15. Agradecimentos

INCT-TILDIAR, FINEP, CAPES, CNPq, CIIA-Saúde, FAPEMIG, FAPESP e UNIMED-BH

Referências

- Cohn, M. (2005). *Agile Estimating and Planning*. Prentice Hall.
- Emmanuel, T. et al. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8:1–37.
- Ganem, I. M., Bianco, G. D., Serufo Filho, J. C., Lima, L., Rocha, L., and Gonçalves, M. A. (2025). Avaliando as limitações e potenciais do algoritmo k-vizinhos mais próximos (knn) na imputação de dados clínicos. In *Simpósio Brasileiro de Banco de Dados (SBB D)*, pages 809–815. SBC.
- Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326.
- Lana, F. C. B. et al. (2025). Unraveling relevant cross-waves pattern drifts in patient-hospital risk factors among hospitalized covid-19 patients. *BMC Infectious Diseases*, 25(1):537.
- Marcolino, M. S. et al. (2021). Clinical characteristics and outcomes of patients hospitalized with covid-19 in brazil. *International Journal of Infectious Diseases*, 107:300–310.
- Marcolino, M. S., Schettini, I., do Nascimento, G., da Rocha, L., Lana, F., and Gonçalves, M. A. (2025). Explainable artificial intelligence for predicting cardiovascular events in hospitalised covid-19 patients. *BMC Infectious Diseases*, 25(1):1569.
- Paiva, B. B. M. et al. (2023). Potential and limitations of machine meta-learning methods for predicting covid-19 mortality. *Scientific Reports*, 13(1):3463.
- Paiva, B. B. M., Pereira, P. D., Gomes, V. M. R., Silva, M. V. R. S., Valiense, C., Marcolino, M. S., and Gonçalves, M. A. (2022). Characterizing and understanding temporal effects in covid-19 data. In *Proceedings of Machine Learning Research*, volume 184 of *Proc. of Mach. Learn. Research*, pages 33–40.
- Schwaber, K. and Sutherland, J. (2020). The scrum guide. Scrum.org. The Definitive Guide to Scrum: The Rules of the Game.