Imputação de Dados Ausentes em Registros de Câncer do INCA: Uma Abordagem com Métodos de Aprendizado de Máquina

Nicolas dos Santos França[1,2], **Rodrigo Bonacin**[1], **Filipe Loyola Lopes**[1]

[1]Divisão de Metodologias da Computação – DIMEC
CTI Renato Archer – Campinas/SP

[2]Instituto de Computação - IC
Universidade Estadual de Campinas (Unicamp) – Campinas/SP

{nicolas.franca, rbonacin, flopes}@cti.gov.br

Abstract. *This paper addresses the challenge of missing data in cancer patient records provided by the Brazilian National Cancer Institute (INCA). The dataset, originally in .dbf format from the SisRHC system, was converted to .csv using TabWin software. Based on a provided data dictionary, missing values were identified and marked as empty cells. Imputation was performed using Scikit-learn methods: SimpleImputer (with mean, mode, and constant strategies), KNNImputer , and IterativeImputer. KNNImputer excels in capturing local data patterns by leveraging nearest neighbors, while IterativeImputer models relationships between features for robust imputation. Results demonstrate successful imputation, with post-processing to convert floating-point outputs to integers. Processing times were compared, revealing that SimpleImputer is significantly faster than KNNImputer and IterativeImputer, though the latter two offer improved accuracy for complex datasets. This approach enables the use of these data in machine learning models for cancer research and management.*

Resumo. *Este artigo aborda o desafio de dados ausentes em registros de pacientes com câncer fornecidos pelo Instituto Nacional de Câncer (INCA). O conjunto de dados, originalmente no formato .dbf do sistema SisRHC, foi convertido para .csv utilizando o software TabWin. Com base em um dicionário de dados fornecido, os valores ausentes foram identificados e marcados como células vazias. A imputação foi realizada com métodos da biblioteca Scikit-learn: SimpleImputer (com estratégias de média, moda e constante), KNNImputer e IterativeImputer. O KNNImputer destaca-se por capturar padrões locais nos dados ao utilizar vizinhos mais próximos, enquanto o IterativeImputer modela relações entre características, oferecendo imputação robusta. Os resultados demonstram imputação bem-sucedida, com pós-processamento para converter saídas de ponto flutuante em inteiros. Os tempos de processamento foram comparados, revelando que o SimpleImputer é significativamente mais rápido que o KNNImputer e o IterativeImputer, embora estes últimos proporcionem maior precisão em conjuntos de dados complexos. Essa abordagem permite o uso desses dados em modelos de aprendizado de máquina para pesquisa e gerenciamento do câncer.*

1. Introdução

O **Instituto Nacional de Câncer (INCA)** é o órgão brasileiro pertencente ao Ministério da Saúde, que atua no desenvolvimento e coordenação de ações integradas para a prevenção e controle do câncer no Brasil. Constitui-se como Centro de Referência de Assistência de Alta Complexidade em Oncologia do MS. O **INCA** direciona sua atuação multidisciplinar ao desenvolvimento de programas e ações, incluindo projetos, estudos, pesquisas e experiências eficazes de gestão com instituições governamentais e não governamentais, além de manter acordos internacionais de cooperação em várias frentes, formando redes de conhecimento técnico e científico e buscando reduzir o impacto regional e global da doença [INCA 2025].

O **INCA** trouxe uma demanda ao **Centro de Tecnologia da Informação Renato Archer (CTI Renato Archer)** relacionada com dados anonimizados de pacientes com câncer. Nesses dados existem variáveis que poderiam ser potenciais atributos preditivos para algorítmos de classificação de doenças por *Aprendizado de Máquina* (**AM**). No entanto, nesses dados tem valores ausentes, o que impossibilita usar esses dados em modelos de **AM** sem um pré-processamento [Faceli et al. 2025].

Dados Ausentes (*Missing Data*, **MD**) são um problema comum relatado na área médica, que pode impedir uma análise confiável dos dados [Lopes et al. 2025]. De acordo com a diretriz sobre *Princípios Estatísticos para Ensaios Clínicos* (ICH E9), os **MD** representam uma fonte potencial de viés. Portanto, todos os esforços devem ser feitos para minimizar a quantidade de **MD** e seus efeitos [Burzykowski et al. 2010].

Os **MD** podem surgir por vários motivos, como imprecisões durante a coleta de dados por erro humano ou mesmo erro em sistemas, sensores ou equipamentos [Pereira et al. 2024, Tlamelo et al. 2021]. Os **MD** são particularmente prevalente em dados médicos, industriais e de pesquisa [Fazakis et al. 2020] e podem ocasionar implicações significativas na extração de conhecimento e no desempenho de modelos de **AM** [Mangussi et al. 2025]. Portanto, é uma área que requer grande atenção.

A literatura mostra que três principais mecanismos produzem os **MD**: *completamente ao acaso* (**MCAR**), relacionados com uma ou mais variáveis *presentes no conjunto de dados* (**MAR**), ou relacionados com uma ou mais variáveis *não presentes no conjunto de dados* (**MNAR**) [Pereira et al. 2020]. De acordo com [García-Laencina et al. 2010], existem três formas principais de trabalhar com **MD**: (a) *Exclusão de dados incompletos*; (b) Procedimentos baseados em modelos, como *modelos máxima verossimilhança*; e (c) *Imputação*, que é a abordagem mais utilizada. A imputação consiste em ”substituir”os valores ausentes de acordo com critérios específicos, como métodos estatísticos ou imputação baseada em **AM**. Métodos de imputação estatística incluem técnicas como imputação média/moda, imputação múltipla e imputação de regressão, entre outras. Em contraste, métodos baseados em **AM** abrangem abordagens como *k-vizinhos mais próximos* (**kNN**) e imputação baseada em *autocodificadores* [Pereira et al. 2020].

O objetivo desta pesquisa é fazer a imputação de dados ausentes, utilizando diferentes técnicas de imputação consideradas como estado da arte, em dados obtidos do **INCA** relacionados com tratamento de câncer, e computar o tempo de processamento para cada algortimo.

O restante do artigo está organizado da seguinte forma: a Seção II apresenta o

referencial teórico, revisando trabalhos relacionados que abordam métodos de imputação de dados ausentes na área da saúde, com foco em algoritmos como **MICE**, **kNN** e outros métodos baseados em aprendizado de máquina; a Seção III detalha a metodologia utilizada, descrevendo o processo de conversão dos dados do **SisRHC** do **INCA**, a identificação de valores ausentes com base no dicionário de dados e a aplicação dos métodos de imputação **SimpleImputer**, **KNNImputer** e **IterativeImputer** da biblioteca **Scikit-learn**; a Seção IV apresenta os resultados obtidos, incluindo exemplos de imputação, a conversão de valores de ponto flutuante para inteiros e a comparação dos tempos de processamento de cada método; a Seção V discute as diferenças de desempenho entre os métodos de imputação, destacando a eficiência do **SimpleImputer** e a robustez do **KNNImputer** e **IterativeImputer** para dados complexos; a Seção VI conclui o trabalho, resumindo os resultados, destacando as lições aprendidas e sugerindo direções futuras para a pesquisa.

2. Referencial

A seguir, apresentamos alguns registros da literatura científica acerca do problema de imputação de **MD** na área da saúde. Os artigos demonstram alguns dos algoritmos mais utilizados na atualidade no que diz respeito a imputação, como o *Multiple Imputation by Chained Equations* (**MICE**), e *k-Nearest Neighbors* (**kNN**).

O algoritmo *Multiple Imputation by Chained Equations* (**MICE**) é um método para imputar dados faltantes, criando múltiplas imputações para refletir incertezas. Valores ausentes são inicialmente preenchidos com estimativas simples (ex., média). Iterativamente, cada variável com dados faltantes é modelada usando as demais como preditoras, atualizando valores em cadeia. O processo gera diversos conjuntos imputados, analisados separadamente e combinados por regras estatísticas [Van Buuren 2018]. O **MICE** é flexível, adaptando-se a diferentes tipos de variáveis, sendo amplamente usado em estatística.

O algoritmo *k-Nearest Neighbors* (**kNN**) é um método de aprendizado supervisionado usado para classificação e regressão, baseado na proximidade entre pontos de dados. Ele identifica os k vizinhos mais próximos de um ponto, usando uma métrica de distância (ex., euclidiana), e toma decisões com base na maioria ou média desses vizinhos. Para imputação de dados faltantes, o **kNN** é adaptado para estimar valores ausentes, encontrando os k pontos mais próximos com dados completos e usando suas informações (ex., média ou moda) para preencher lacunas. O método é não paramétrico, simples e eficaz, mas sensível à escolha de k e à escala dos dados [Hastie et al. 2009]. É amplamente usado em ciência de dados para tarefas preditivas e imputação.

Em [García-Laencina et al. 2015], é analisado o impacto da imputação de dados ausentes na predição de sobrevida de 5 anos para pacientes com câncer de mama, utilizando um conjunto de dados real do Instituto Português de Oncologia do Porto, com 399 casos, 16 variáveis majoritariamente categóricas e uma taxa média de 18% de dados ausentes. Quatro cenários são avaliados: predição sem imputação (usando **kNN** e Árvores de Classificação, que falham devido à alta incompletude) e com imputação por métodos de *Moda*, *Expectativa-Maximização* (**EM**) e **kNN** em dados limpos. Modelos de sobrevida são construídos com **kNN**, *Árvores de Classificação*, *Regressão Logística* e *Máquinas de Vetores de Suporte*, avaliados por validação cruzada aninhada de 10 do-

bras para *acurácia*, *sensibilidade*, *especificidade* e *Área sob a curva* (**AUC**). Os resultados mostram que a imputação **kNN** oferece o melhor desempenho, especialmente com classificação **kNN**, alcançando mais de 81% de acurácia e 0,78 de **AUC**, superando outros métodos e mantendo equilíbrio entre classes. Descartar variáveis com alta taxa de ausência melhora ligeiramente, mas não significativamente, os resultados, destacando a robustez do **kNN** para dados categóricos ausentes em tarefas de predição clínica.

O artigo de [Pereira et al. 2024] aborda a imputação de dados ausentes, introduzindo quatro novas estratégias de geração artificial de **MNAR** para simular cenários realistas em variáveis numéricas e categóricas, incluindo dependências em valores próprios e não observados. Essas estratégias superam limitações das abordagens existentes, permitindo experimentos mais precisos. Um estudo de benchmark compara seis métodos de imputação estado-da-arte, como **MICE**, **kNN**, *autoencoders* (**AE**) e variantes, em 10 conjuntos de dados médicos públicos com taxas de ausência variando de 10% a 80%. Os resultados, avaliados pelo *erro médio absoluto* (**MAE**), mostram que o **MICE** é superior na maioria das taxas e conjuntos de dados, enquanto *autoencoders* se destacam em cenários de alta ausência. A pesquisa oferece diretrizes para seleção de métodos de imputação em contextos **MNAR**, enfatizando a resiliência a mecanismos complexos de dados ausentes.

[Lopes et al. 2025] propõe uma abordagem inovadora para imputação de dados ausentes utilizando o algoritmo de propagação de rótulos, denominado *Label Propagation for Missing Data Imputation* (**LPMD**). O método modela conjuntos de dados como grafos de proximidade, propagando valores conhecidos para imputação de dados, combinando e ponderando informações de features observadas. Experimentos em dez conjuntos de dados de referência avaliam o desempenho sob mecanismos **MAR** e **MNAR** com taxas de ausência variadas (de 1% a 60%), comparando com métodos como *Mean*, **MICE**, **kNN**, *Partial Multiple Imputation with Variational Autoencode* (**PMIVAE**) e *Siamese Autoencoder-Based Approach for Missing Data Imputation* (**SAEI**). A variante **LPMD2** destaca-se pela velocidade de processamento, com acelerações de 0,7 a 23 vezes em relação aos baselines, e estabilidade no *erro médio absoluto* (**MAE**) para diferentes mecanismos e taxas. Os resultados em modelos de **AM** são comparáveis aos resultados linha de base, enfatizando a robustez do **LPMD** em cenários reais com mecanismos de ausência mistos e desconhecidos. A abordagem beneficia-se do paradigma semi-supervisionado, explorando dados rotulados e não rotulados para imputação.

O artigo de [Tiwaskar et al. 2025] investiga o impacto de técnicas de imputação baseadas em aprendizado de máquina em conjuntos de dados médicos, especialmente para diabetes, onde a incompletude é comum. As técnicas comparadas incluem **kNN**, **MICE** e *MissForest*, avaliadas em taxas de ausência de 10% a 50% no conjunto de dados UCI Diabetes. A análise abrange desempenho de modelos (*acurácia*, *precisão*, *recall*, *F1-score*, *Matthews Correlation Coefficient*) com classificadores como *Random Forest*, *Support Vector Machine* (**SVM**), *AdaBoost* e *XGBoost*; *erro de imputação* (**MAE**, *Raiz do Erro Quadrático Médio* (**RMSE**), *R-Squared* ($\mathbf{R^2}$)); *análise de correlação de Pearson*; e critérios de seleção (*O critério de informação de Akaike* (**AIC**), *Critério de Informação Bayesiano* (**BIC**)). Os resultados mostram que *MissForest* supera **kNN** e **MICE** na maioria dos casos, preservando associações entre variáveis semelhantes ao conjunto completo. A pesquisa conclui que *MissForest* é a técnica superior para lidar com dados ausentes em

estudos sobre diabetes, facilitando predições precisas.

Em [Benhamza et al. 2025], é apresentado uma revisão abrangente dos métodos de imputação de dados ausentes em análises de *big data* médico, destacando sua importância para avanços na pesquisa em saúde e melhoria do cuidado ao paciente. O estudo avalia técnicas tradicionais, abordagens estatísticas avançadas, modelos baseados em aprendizado de máquina e aprendizado profundo, considerando desafios como alta dimensionalidade e heterogeneidade dos dados médicos. Tecnologias de *big data* são exploradas para imputação eficiente em larga escala, e métricas de avaliação de qualidade são analisadas. Por fim, direções futuras são discutidas para aprimorar estratégias de imputação e apoiar decisões baseadas em dados na era do *big data* médico.

3. Metodologia

O **INCA** disponibilizou um dataset do Sistema para Informatização dos dados de Registros Hospitalares de Câncer (***SisRHC***) com 4GB de arquivos do tipo *.dbf* e *.cnv*. O ***SisRHC*** é um sistema de *software* desenvolvido pelo **INCA** para que os hospitais que tratam câncer possam coletar, armazenar e gerenciar informações detalhadas sobre os pacientes com diagnóstico de câncer. Ele funciona como uma ferramenta essencial para a vigilância do câncer no Brasil, pois as informações coletadas são usadas para:

- Entender o panorama do câncer no país: os dados alimentam uma base nacional que permite ao **INCA** e às secretarias de saúde estaduais e municipais analisar a incidência da doença, os tipos de tumores mais comuns e a sobrevida dos pacientes.
- Apoiar a pesquisa epidemiológica: as informações são fontes primárias para estudos sobre fatores determinantes do câncer.
- Melhorar o planejamento de políticas de saúde: com base nos dados, é possível planejar e otimizar ações de prevenção, diagnóstico e tratamento da doença em nível nacional e regional.

O funcionamento dos **Registros Hospitalares de Câncer (RHC)** e o envio de dados para o sistema **IntegradorRHC** do **INCA** é obrigatório para hospitais habilitados na Atenção Especializada em Oncologia do SUS e facultativo para os demais.

As extensões de arquivo provindas do SisRHC não são muito comum de se trabalhar, por conta disso usamos um programa chamado ***TabWin***[1] para converter esses datasets principalmente em *.dbf* para *.csv*. O TabWin funciona em complemento com o **TABNET**, uma aplicação de tabulação de dados do **DATASUS (Departamento de Informática do SUS)**, que permite criar tabelas, gráficos e mapas com base nas diversas bases de dados do **Sistema Único de Saúde (SUS)** e outras informações de saúde relevantes.

O ***TabWin*** permite as seguintes funcionalidades:

- Importar as tabulações efetuadas na Internet (geradas pelo aplicativo **TABNET**).
- Realizar operações aritméticas e estatísticas nos dados da tabela gerada ou importada pelo ***TabWin***.
- Elaborar gráficos de vários tipos, inclusive mapas, a partir dos dados dessa tabela;

[1] http://siab.datasus.gov.br/DATASUS/index.php?area=060805item=1

- Efetuar outras operações na tabela, ajustando-a às suas necessidades.

Já o programa **TABNET** pode facilitar:

- A construção e aplicação de índices e indicadores de produção de serviços, de características epidemiológicas (incidência de doenças, agravos e mortalidade) e dos aspectos demográficos de interesse (educação, saneamento, renda e etc) - por estado e município;
- O planejamento e programação de serviços; a avaliação e tomada de decisões relativas à alocação e distribuição de recursos; a avaliação do impacto de intervenções nas condições de saúde.

O **INCA** disponibilizou um arquivo pdf que é um dicionário das variáveis da base de dados do ***SisRHC*** que, pra cada tabela dos datasets, descrevia o que cada código numérico significava para aquela coluna específica como pode ser visto na **Tabela 1**.

Depois de termos os arquivos *.csv* pra cada dataset proveniente de arquivos *.dbf*, baseado na descrição da coluna *domínio* do dicionário de dados, analisamos quais domínios significavam "Sem informação", ou seja, considerando a coluna em si, pra quais variáveis aquela coluna significa não ter informações. Na maioria dos casos, o valor numérico 9 foi colocado pra representar isso. Uma exceção é a coluna "BASDIAGSP", em que o valor numérico que representa "Sem informação"é o número 4.

Montamos um *script* em Python que verificava no arquivo *.csv* quando temos uma determinada coluna, setamos o valor numérico que significa que aquela celula é "Sem informação"e esse *script* percorria todas as colunas substituindo o valor numérico que significa "Sem informação"por celulas vazias.

A partir do pré-processamento para identificar os dados ausentes foi realizado o processo de imputação a partir de diferentes algoritmos, notadamente média ou identificação de conjuntos de dados mais frequentes, **kNN** e o **MICE**. O pipeline de imputação foi construído na linguagem Python, ulizando a biblioteca ***Scikit-learn*** que é uma biblioteca de código aberto em Python para **AM**, construída sobre ***NumPy***, ***SciPy*** e ***Matplotlib***. Ela oferece ferramentas simples e eficientes para mineração de dados, análise de dados e tarefas de aprendizado de máquina. O ***Scikit-learn*** tem um módulo de imputação de dados em **sklearn.impute** que fornece ferramentas para lidar com valores ausentes em conjuntos de dados. Entre as classes implementadas pelo ***Scikit-learn*** para imputação de dados, estão:

- **SimpleImputer**: Uma classe para imputação básica, que substitui valores ausentes por estratégias como:
  - Média (**strategy="mean"**) ou mediana (**strategy="median"**) para variáveis numéricas.
  - Moda (**strategy="most-frequent"**) para variáveis categóricas.
  - Um valor constante (**strategy="constant"**, com **fill-value** definido)
- **KNNImputer** : Imputa valores ausentes com base nos ***k*** vizinhos mais próximos, usando a média ponderada ou não ponderada dos valores vizinhos. Ideal para dados com padrões locais.
- **IterativeImputer**: Realiza imputação iterativa, modelando cada característica com valores ausentes como uma função das demais. usando métodos como regressão (experimental, baseado em modelos como BayesianRidge).

Tabela 1. Exemplo de dicionário com algumas variáveis da base SisRHC

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
44	ALCOOLIS	Histórico de consumo de bebida alcoólica	1.Nunca; 2.Ex-consumidor; 3.Sim; 4.Não avaliado; 8.Não se aplica; 9.Sem informação
22	ANOPRIDI	Ano do diagnóstico	aaaa
42	ANTRI	Ano da triagem	dd/mm/aaaa
base de dados SP	BASDIAGSP	Base mais importante para o diagnóstico do tumor	1.Exame clínico 2.Recursos auxiliares não microscópicos 3.Confirmação microscópica 4.Sem informação
24	BASMAIMP	Base mais importante para o diagnóstico do tumor	1.Clínica; 2.Pesquisa clínica; 3.Exame por imagem; 4.Marcadores tumorais; 5.Citologia; 6.Histologia da metástase; 7.Histologia do tumor primário; 9. Sem informação
47	CLIATEN	Clínicas do primeiro atendimento - entrada do paciente	Codificação segundo Tabela de Clínicas do SisRHC
31	CLITRAT	Clínica de início do tratamento	Codificação segundo Tabela de Clínicas do SisRHC
SIS	CNES	Número do CNES do Hospital	Codificação segundo tabela do Cadastro Nacional de Estab. de Saúde
32	DATAINITRT	Data do início do primeiro tratamento específico para o tumor, no hospital	dd/mm/aaaa
36	DATAOBITO	Data do óbito	dd/mm/aaaa
21	DATAPRICON	Data da 1ª consulta	dd/mm/aaaa
23	DIAGANT	Diagnóstico e tratamento anteriores	1.Sem diag./Sem trat.; 2.Com diag./Sem trat.; 3.Com diag./Com trat.; 4.Outros; 9. Sem informação
22	DTDIAGNO	Data do primeiro diagnóstico	dd/mm/aaaa
32	DTINITRT	Ano do início do primeiro tratamento específico para o tumor, no hospital	aaaa
21	DTPRICON	Ano da 1ª consulta	aaaa
42	DTTRIAGE	Data da triagem	dd/mm/aaaa
28a	ESTADIAG	Estadiamento clínico do tumor (TNM) - Grupo	Codificação do grupamento do estádio clínico segundo classificação TNM
28a	ESTADIAM	Estadiamento clínico do tumor (TNM)	Codificação do grupamento do estádio clínico segundo classificação TNM
17	ESTADRES	UF de procedência (residência)	Sigla da UF de procêdencia
41	ESTCONJ	Estado conjugal atual	1.Solteiro; 2.Casado; 3.Viúvo; 4.Separado judicialmente; 5.União consensual; 9.Sem informação
35	ESTDFIMT	Estado da doença ao final do primeiro tratamento no hospital	1.Sem evidência da doença (remissão completa); 2.Remissão parcial; 3.Doença estável; 4.Doença em progressão; 5.Suporte terapêutico oncológico; 6. Óbito; 8. Não se aplica; 9. Sem informação
48	EXDIAG	Exames relevantes para o diagnóstico e planejamento da terapêutica do tumor	1.Exame clínico e patologia clínica; 2.Exames por imagem; 3.Endoscopia e cirurgia exploradora; 4.Anatomia patológica; 5.Marcadores tumorais; 8.Não se aplica; 9. Sem informação
43	HISTFAMC	Histórico familiar de câncer	1.Sim; 2.Não; 9.Sem informação

O *dataset* imputado possui **17432 linhas** e **47 colunas** de dados. Temos **7,36% de missing values** para serem tratados.

O processo de imputação, com cada algoritmo foi realizado em uma máquina com processador Intel Core i7-4770 CPU @ 3.40GHz, com 8 núcleos, 32 GB de memória RAM, GPU Mesa Intel HD Graphics 4600 (1,5GB). Os tempos de processamento foram computados. Os resultados obtidos estão apresentados na seção subsequente.

Os métodos do Scikit-learn (como KNNImputer e IterativeImputer) são tipicamente executados na CPU, sem menção a aceleração por GPU. A diferença de tempo parece ocorrer principalmente pelo processador (CPU), dado que os algoritmos envolvem operações sequenciais e iterativas que se beneficiam de múltiplos núcleos, mas não de processamento paralelo massivo como em GPUs.

4. Resultados

Um exemplo do resultado de imputação obtido pelo algortimo **kNN** pode ser observado na **Tabela 2**.

	TPCASO.1	SEXO	IDADE	LOCALNAS	RACACOR	INSTRUC	CLIATEN
0	2.0	1.0	59.0	0.0	1.0	2.0	24.0
1	2.0	1.0	72.0	0.0	2.0	2.2	24.0
2	1.0	2.0	20.0	0.0	1.0	3.0	0.0
3	1.0	1.0	54.0	0.0	1.0	2.0	0.0
4	1.0	2.0	43.0	0.0	1.0	6.0	0.0
5	1.0	1.0	53.0	0.0	1.0	4.0	0.0
6	1.0	2.0	62.0	0.0	1.0	2.0	0.0
7	1.0	2.0	60.0	0.0	1.0	3.0	0.0
8	1.0	1.0	55.0	0.0	1.0	3.4	0.0
9	1.0	1.0	65.0	0.0	2.0	4.0	0.0
10	1.0	1.0	66.0	0.0	1.0	4.0	0.0
11	1.0	2.0	81.0	0.0	1.0	3.0	0.0
12	1.0	2.0	61.0	0.0	1.0	4.0	0.0
13	1.0	2.0	50.0	0.0	1.0	4.0	0.0

Tabela 2. Exemplo com parte dos resultads da imputação através do algoritmo de imputação kNN.

Conforme pode ser observado na **Tabela 2**, o resultado da imputação retornou com vírgulas, porém, os valores originais do dataset eram números naturais. Para resolver isso, montamos um script que usa o ***Pandas*** para fazer manipulações em arquivos *.csv* e convertemos esses números com virgula para números naturais. A conversão foi feita primeiramente substituindo a virgula por ponto, e em seguida convertendo essa string para float. Logo após essa conversão para ponto flutuante, convertemos esse valor para inteiro diretamente. Ou seja, o valor 3,4 retornaria o valor 3, e o valor 5,9 retornaria 5, por exemplo. Na **Tabela 3** pode ser observado o resultado após o tratamento dado.

Enquanto faziamos nossas análises, criamos mais um *script* que calculava o tempo em segundos para o processamento de cada método de imputação de dados (**Tabela 4**).

	TPCASO.1	SEXO	IDADE	LOCALNAS	RACACOR	INSTRUC	CLIATEN
0	2	1	59	0	1	2	24
1	2	1	72	0	2	2	24
2	1	2	20	0	1	3	0
3	1	1	54	0	1	2	0
4	1	2	43	0	1	6	0
5	1	1	53	0	1	4	0
6	1	2	62	0	1	2	0
7	1	2	60	0	1	3	0
8	1	1	55	0	1	3	0
9	1	1	65	0	2	4	0
10	1	1	66	0	1	4	0
11	1	2	81	0	1	3	0
12	1	2	61	0	1	4	0
13	1	2	50	0	1	4	0

Tabela 3. Resultado da imputação usando KNNImputer logo após o script de conversão citado.

Método	Tempo (s)
SimpleImputer (Média)	0.021005
SimpleImputer (Moda)	0.036131
SimpleImputer (Constante)	0.012014
KNNImputer (Neighbors = 5)	48.786462
IterativeImputer (MICE)	32.343818

Tabela 4. Tempo em segundos para processar o dataset citado.

5. Discussão

Alguns pontos que gostariamos de destacar fazendo comparação entre os métodos de imputação de dados é que, especialmente o **KNNImputer** e o **IterativeImputer** em comparação com o **SimpleImputer** demorou muito mais que o esperado. E quanto mais dados tem no ***.csv***, mais demorado vai ser o termino da imputação em métodos como o **KNNImputer** e o **IterativeImputer** em comparação com o **SimpleImputer**.

6. Conclusão

O objetivo desta pesquisa, que era realizar a imputação de dados ausentes em um conjunto de dados de pacientes com câncer fornecido pelo **INCA**, foi plenamente alcançado. Utilizando métodos da biblioteca **Scikit-learn**, como **SimpleImputer**, **KNNImputer** e **IterativeImputer**, foi possível tratar valores ausentes no dataset convertido de .dbf para .csv, tornando-o adequado para uso em modelos de aprendizado de máquina. O pós-processamento para conversão de números de ponto flutuante para inteiros garantiu a compatibilidade com os formatos esperados, conforme descrito no dicionário de dados do SisRHC.

Durante o desenvolvimento do trabalho, aprendemos que a escolha do método de imputação impacta tanto a precisão dos resultados quanto o tempo de processamento. O

SimpleImputer destacou-se pela sua rapidez, sendo ideal para datasets grandes quando o tempo é um fator crítico. Por outro lado, o **KNNImputer** e o **IterativeImputer**, embora mais demorados, demonstraram maior capacidade de capturar padrões locais e relações complexas entre variáveis, o que pode ser vantajoso em datasets com alta dimensionalidade ou padrões não lineares. Além disso, a conversão inicial dos dados e a identificação precisa de valores ”Sem informação”com base no dicionário de dados foram etapas para garantir a qualidade da imputação.

Embora tenha sido interessante o estudo do problema de *Missing Data* (**MD**), percebemos que o dado que foi imputado pelos imputadores foi o valor 0. Esse valor não significa nada considerando o *dóminio de dados* das colunas com dados ausentes. Considerando esse fator, e os nossos estudos da literatura acerca do assunto, no lugar dos imputadores disponíveis no ***Scikit-learn***, talvez seria melhor para a finalidade da pesquisa usar métodos de *autoencoder*, considerando o número de linhas do dataset com dados ausentes. Também seria importante dedicar mais tempo analisando os datasets entregues e pré-processando eles para reduzir a quantidade de *Missing Data* para um estudo posterior.

Referências

Benhamza, K., Benselim, R., Naidja, H., and Seridi, H. (2025). A comprehensive survey of imputation methods in medical missing data analysis. *Applied Intelligence*, 55(11):778.

Burzykowski, T., Carpenter, J., Coens, C., Evans, D., France, L., Kenward, M., Lane, P., Matcham, J., Morgan, D., Phillips, A., Roger, J., Sullivan, B., White, I., and Yu, L.-M. (2010). Missing data: Discussion points from the PSI missing data expert group. *Pharmaceutical statistics : the journal of the pharmaceutical industry*, 9(4):288–297.

Faceli, K., Lorena, A. C., Gama, João, A. T. A. d., and de Carvalho, A. C. P. d. L. F. (2025). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, Rio de Janeiro, 3 edition.

Fazakis, N., Kostopoulos, G., Kotsiantis, S., and Mporas, I. (2020). Iterative robust semi-supervised missing data imputation. *IEEE Access*, 8(1):90555–90569.

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., and Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in Biology and Medicine*, 59:125–133.

García-Laencina, P. J., Sancho-Gómez, J.-L., and Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.

INCA (2025). Instituto nacional de câncer (inca). `https://www.gov.br/inca/pt-br`. Acesso em 02 de setembro de 2025.

Lopes, F. L., Mangussi, A. D., Pereira, R. C., Santos, M. S., Abreu, P. H., and Lorena, A. C. (2025). A label propagation approach for missing data imputation. *IEEE Access*, 13:65925–65938.

Mangussi, A. D., Santos, M. S., Lopes, F. L., Pereira, R. C., Lorena, A. C., and Abreu, P. H. (2025). mdatagen: A python library for the artificial generation of missing data. *Neurocomputing*, 625:129478.

Pereira, R. C., Abreu, P. H., Rodrigues, P. P., and Figueiredo, M. A. (2024). Imputation of data missing not at random: Artificial generation and benchmark analysis. *Expert Systems with Applications*, 249:123654.

Pereira, R. C., Santos, M. S., Rodrigues, P. P., and Abreu, P. H. (2020). Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. *Journal of Artificial Intelligence Research*, 69:1255–1285.

Tiwaskar, S., Rashid, M., and Gokhale, P. (2025). Impact of machine learning-based imputation techniques on medical datasets- a comparative analysis. *Multimedia Tools and Applications*, 84(9):5905–5925.

Tlamelo, E., Maupong Thabiso, Dimane, M., Semong Thabo, Banyatsang, M., and Oteng, T. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1).

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, FL.