# Avaliação de diferentes técnicas de agrupamento no contexto da Imputação em Cascata

**Tarsila Tavares**[1], **Kele Belloze**[1], **Ronaldo Goldschmidt**[2], **Jorge Soares**[1]

Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

[2]Instituto Militar de Engenharia - IME

tarsila.tavares@aluno.cefet-rj.br, kele.belloze@cefet-rj.br,

ronaldo.rgold@ime.eb.br, jorge.soares@cefet-rj.br

***Abstract.*** *It is common for databases to have missing values, which may require the application of imputation techniques. In this article, we propose a variation of the cascading imputation approach, suitable for handling missing values in multiple columns, where imputed values in one attribute are reintegrated into the database before imputation and used for the completion of the next attribute. The results reveal that the variation of clustering algorithms does not impact the quality of the imputed data. However, there are gains when compared to mean imputation.*

***Resumo.*** *É comum que bases de dados apresentem valores ausentes, o que pode demandar a aplicação de técnicas de imputação. Neste artigo, propomos uma variação da abordagem de imputação em cascata, adequada para tratar valores ausentes em múltiplas colunas, em que valores imputados em um atributo são reintegrados à base de dados antes da imputação, e utilizados para a complementação do próximo atributo. Os resultados revelam que a variação dos algoritmos de agrupamento não impacta na qualidade do dado imputado. Entretanto, há ganhos quando comparado à imputação por média.*

## 1. Introdução

A problemática da ausência de dados pode manifestar-se em diversas fases do ciclo de vida dos dados, desde a coleta inicial até a sua transformação e integração antes da análise final. Com o crescente volume de dados gerados, a lacuna de informações torna-se um desafio cada vez mais significativo, tanto para analistas quanto para pesquisadores, impondo obstáculos adicionais nos contextos acadêmico e empresarial [Coneglian and Segundo, 2017] .

A importância de explorar técnicas eficazes de complementação de dados ausentes em *datasets* - problema popularmente conhecido como *imputação de dados* - é sublinhada pela necessidade que muitos algoritmos de análise de dados têm de operar com conjuntos de dados completos, sem valores ausentes. A qualidade dos dados, além de sua completude, é crucial no apoio a sistemas de inteligência artificial, destacando-se recentemente o conceito de Inteligência Artificial Centrada em Dados (*Data-Centric AI* - DCAI) [Jarrahi et al., 2022] e, devido a isso, à necessidade de implementar e aplicar técnicas de imputação adequadas. Nesse aspecto, Rubin [1976] realizou contribuições substanciais

ao campo, propondo teorias sobre a análise de dados faltantes, fundamentadas em princípios de verossimilhança. Ao sistematizar essas abordagens, criou uma base teórica e experimental vital para futuras pesquisas em métodos de imputação de dados.

Nesse contexto, a *imputação em cascata*, proposta por Ferlin [2008], surge com o objetivo de melhorar a qualidade do proceso. Nela, os valores imputados em um atributo são reintegrados à base de dados antes da imputação do atributo subsequente, ao invés de serem adicionados somente ao final de todas as iterações, como é comum nesse tipo de procedimento. Esse processo é repetido até que a imputação seja concluída para todo o conjunto de dados.

Até onde foi possível observar, os experimentos envolvendo a imputação em cascata realizados em Ferlin [2008] se limitaram a testar um único algoritmo de agrupamento: a rede neural SOM (*Self-Organizing Maps*). Assim sendo, o presente trabalho levanta a seguinte questão de pesquisa: *O quanto a utilização de outros algoritmos de agrupamento na implementação pode afetar os resultados obtidos com a imputação em cascata proposta por Ferlin [2008]*? Tal questão mostra-se relevante, uma vez que o desempenho de algoritmos de agrupamento com diferentes estratégias de busca pelos grupos pode variar em função da distribuição dos dados a serem agrupados.

Diante da questão levantada, esta pesquisa tem como objetivo realizar experimentos em cinco conjuntos de dados com *outliers* e diferentes graus de correlação entre seus atributos, variando os algoritmos de agrupamento. Essa abordagem diferencia-se da proposta de Ferlin [2008] na escolha dos *datasets*. Os resultados obtidos nos experimentos até o momento não apontam diferenças significativas decorrentes da variação do algoritmo de agrupamento utilizado. Entretanto, pode-se perceber cenários em que a imputação em cascata mostra-se superior entre os métodos de imputação testados quando comparados à imputação por média.

Na Seção 2, é apresentada uma visão geral sobre imputação e a imputação em cascata. A Seção 3 aborda como foi realizada a busca sistemática por trabalhos relacionados. A Seção 4 apresenta a abordagem metodológica deste trabalho. Na Seção 5, estão descritos os experimentos comparativos preliminares realizados e as análises dos resultados obtidos. E, por fim, a Seção 6 reporta os principais resultados da pesquisa até o momento e indica as iniciativas de trabalhos futuros.

## 2. Fundamentação Teórica

### 2.1. Imputação

A imputação de dados é formalmente reconhecida como um campo de pesquisa que busca preencher lacunas de dados faltantes por meio de métodos estatísticos e técnicas de inteligência artificial [Gelman and Hill, 2006]. Este processo envolve substituir dados ausentes ou inválidos por estimativas que, apesar de não serem originais, são consideradas adequadas para análises subsequentes.

Os métodos de *imputação simples*, ou *imputação univariada*, consistem na geração de um único valor para cada valor ausente em um atributo do conjunto de dados. Existem diversos métodos de imputação simples, muitos deles baseados no paradigma estatístico como média, moda, regressão simples e regressão logística [Rubin, 1976; Little and Rubin, 2019]. Embora consigam preencher as lacunas de informação ausente nos

conjuntos de dados, esses métodos podem estimar valores com variância reduzida, o que potencialmente resulta em análises enviesadas [Cartwright et al., 2003]. A *imputação múltipla* emerge como uma alternativa à imputação simples. Segundo ela, cada valor ausente é substituído por um conjunto de valores estimados pela técnica [Rubin, 1988]. Os métodos de imputação múltipla proporcionam uma abordagem mais robusta ao tratamento de dados ausentes, produzindo estimativas mais precisas [Wayman, 2003].

Existem motivadores que causam as ausências em conjuntos de dados. Chamamos de *mecanismo de ausência* a razão geradora do dado inexistente. Existem três tipos de mecanismos: (i) MCAR (*Missing Completely At Random*) – a ausência é totalmente aleatória e sem qualquer correlação entre os atributos; (ii) MAR (*Missing At Random*) - a probabilidade de dados ausentes está condicionada apenas aos dados observados, não aos dados ausentes e (iii) NMAR (*Not Missing At Random*) - a probabilidade de um dado estar ausente está diretamente relacionada aos valores ausentes.

## 2.2. Imputação em Cascata

O método de imputação em cascata proposto por Ferlin [2008], inicia-se com a separação do *dataset* numérico original em duas partes: uma que contém somente tuplas completas (casos completos); e outra, nas quais as linhas apresentam valores ausentes em algum(ns) atributo(s) (casos incompletos), como ilustrado na Figura 1.

Na sequência, para os casos incompletos, procede-se a separação das tuplas em grupos, em que os registros incompletos são categorizados em grupos baseados no arranjo espacial dos dados ausentes nas tuplas. Essa fase foi chamada de identificação da morfologia da ausência, e tem como objetivo a identificação de conjuntos de casos com padrões similares de ausência. Durante esta fase, ocorre a binarização dos dados nos registros incompletos, transformando-os em uma representação binária onde os dados faltantes são marcados com 1 e os presentes com 0. Este processo de binarização permite um agrupamento eficiente, pois habilita a comparação entre registros com base na similaridade de seus padrões de ausência, independentemente dos valores específicos dos dados.

Em seguida, a imputação é realizada sequencialmente de grupo em grupo. Diferentes critérios podem ser utilizados na ordenação dos grupos a serem submetidos à imputação - por exemplo, na ordem crescente do percentual de dados ausentes em relação ao total de dados no grupo. Seguindo este critério, a imputação dá preferência a preencher primeiro os grupos com maior quantidade de dados originais e, portanto, mais confiáveis. Os dados imputados em cada etapa são utilizados para facilitar a imputação nas etapas subsequentes, gerando um efeito cascata.

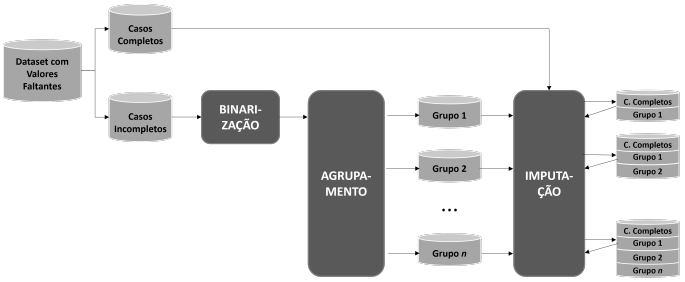


**Figura 1. Metodologia da imputação em cascata proposta por Ferlin [2008]**

## 3. Trabalhos Relacionados

A fim de identificar trabalhos com abordagem similar a esta proposta, executou-se, na base Scopus, a string de busca *TITLE-ABS-KEY ( ( "cascade imputation" OR "cascade missing data" OR "recursive imputation" OR "recursive missing data" ) )*. Obteve-se como resultado apenas dois artigos, sendo um deles relacionado a dados censurados, que não são objeto de nosso estudo. Assim, restou o trabalho de Montiel et al. [2018], que introduz o método *Cascade Imputation* (CIM) - uma técnica incremental baseada em aprendizado de máquina supervisionado. Este método ordena o *dataset* em ordem crescente, separando e aplicando modelos de regressão ou de classificação, dependendo da variável-alvo. Na rodada subsequente do processo, o atributo previamente imputado é incluído como uma variável preditora nos modelos de regressão ou classificação. Este ciclo de imputação e reintegração de variáveis continua até não restarem atributos com valores ausentes no conjunto de dados.

Nosso estudo se diferencia da abordagem de Montiel et al. [2018] por implementar a imputação *hot-deck* (na qual a imputação considera apenas as linhas de dados com maior similaridade no processo de imputação, ao invés da imputação clássica, que leva em conta todo o conjunto de dados nesse processo) com algoritmos de agrupamento de quatro categorias distintas (particionamento, hierárquico, modelo e densidade), além de realizar o processo de imputação com um algoritmo (k-NN) que não demanda a divisão do *dataset* em conjunto de treino e teste.

## 4. Método

Com o objetivo de responder à pergunta de pesquisa, os algoritmos de agrupamento e respectivas estratégias avaliados neste trabalho concentram-se na etapa de agrupamento da Figura 1. O trabalho original de Ferlin [2008] utiliza, na etapa de agrupamento, apenas o algoritmo SOM, algoritmo de agrupamento baseado em modelos. Este trabalho expande a abordagem variando os algoritmos de agrupamento. O critério principal é que eles pertençam a outras categorias, com o propósito de diversificar as escolhas e, com isso, verificar se há ganhos na qualidade do dado imputado. Para isso, utilizamos DBSCAN (baseado em densidade), K-Modes (baseado em particionamento) e Agglomerative Clustering (baseado em hierarquia sucessiva além do próprio SOM). Os valores faltantes foram introduzidos sinteticamente nos *datasets* por meio de diferentes mecanismos de ausência de dados (MCAR, MAR e MNAR), aplicando três percentuais distintos de ausência (10%, 20% e 30%).

## 5. Avaliação experimental

Nesta pesquisa, selecionamos cinco conjuntos de dados, sem ocorrências de dados ausentes, provenientes do repositório da Universidade da Califórnia, Irvine (UCI)[1]: *Wine*, *Pima Indians Diabetes*, *Abalone*, *Glass Identification* e *Yeast*. Tais conjuntos são frequentemente utilizados em estudos de imputação, conforme destacado pela pesquisa de Gonçalves [2021]. Os cinco conjuntos de dados possuem um esquema da forma $S(A_1, A_2, ..., A_n, C)$ em que $C$ é considerado o alvo na construção de modelos de classificação e os atributos $A_i, i = 1, 2, ..., n$, são atributos quantitativos.

[1]https://archive.ics.uci.edu/

Durante a seleção dos conjuntos de dados, foram analisadas as correlações entre os atributos e a presença de *outliers*, ambos critérios importantes para permitir uma análise mais clara dos resultados de imputação. A avaliação da correlação envolveu: calcular a correlação entre cada par de atributos $A_i$ e $A_j$, $i \neq j$ e a média da correlação entre os atributos. Os *outliers* foram identificados por análise dos interquartis, quantificados por coluna, e calculou-se a proporção de tuplas com *outliers* em relação à quantidade de atributos. Os resultados dessas classificações encontram-se resumidos na Tabela 1, que mostra os conjuntos escolhidos e suas características.

**Tabela 1. Principais características dos conjuntos de dados selecionados**

| *Datasets* | % Correlação | % *Outliers* | Atributos | Tuplas |
|---|---|---|---|---|
| Wine | 50% | 0,8% | 13 | 178 |
| Pima Indians Diabetes | 22% | 2,1% | 8 | 768 |
| Abalone | 93% | 1,7% | 7 | 4.177 |
| Glass Identification | 50% | 16,4% | 8 | 214 |
| Yeast | 25% | 28,8% | 7 | 1.484 |

Para simular a ausência de dados, aplicou-se taxas de ausências de 10%, 20% e 30%, utilizando os mecanismos de ausências MCAR, MNAR e MAR, nos *datasets* analisados. Esses percentuais de ausências foram aplicados na base como um todo e mantiveram a proporcionalidade dentro de cada atributo - ou seja, se um *dataset* apresenta *x*% de dados ausentes, cada coluna conterá *x*% de ausência, mantendo a uniformidade. Como exemplo, uma taxa de ausência de 10% determinará esse percentual de ausência em cada atributo do *dataset*.

A implementação dos algoritmos de agrupamento foi realizada na linguagem Python com apoio da Biblioteca Scikit-learn. Na bateria inicial de experimentos, foram adotados os valores *default* dos hiperparâmetros das implementações dos algoritmos de agrupamento.

Uma vez imputados todos os dados ausentes do conjunto de dados, a avaliação do método de imputação foi medida pelo desempenho do algoritmo dos k-vizinhos mais próximos (*k-Nearest Neighbors* ou k-NN) na tarefa de classificação aplicado ao conjunto de dados. Assim, dado um registro de dados $r$ a ser classificado, o k-NN identifica os $k$ registros do conjunto de dados mais próximos de $r$. A distância adotada pelo k-NN em nossos experimentos foi a Euclidiana. Para o presente estudo, foram adotados valores de $k$ iguais a 1, 3, 5 e 10, conforme utilizado por Ferlin [2008] na implementação do método de imputação em cascata.

Para fins de avaliação, foram comparados os valores imputados com os dados originais, calculando-se o percentual do erro. Adicionalmente, o erro da imputação foi calculado pela diferença entre os valores médios imputados e os respectivos valores reais existentes nos conjuntos de dados. Tal metodologia é comumente utilizada como referência básica para avaliação do desempenho de métodos de imputação.

A Figura 2 apresenta uma comparação entre os percentuais dos erros médios da imputação em cascata, e a imputação por média aritmética. Observa-se que não há variação significativa entre os algoritmos de agrupamento. De forma geral, as ausências MAR

10%, MAR 20%, MCAR 10% e MNAR 10% tiveram um desempenho melhor, ou seja, menor percentual de erro, na imputação em cascata frente à imputação por média. Os resultados produzidos no contexto deste trabalho sugerem que a imputação em cascata desempenha de forma mais satisfatória quando o percentual de ausência é baixo - ressaltando, portanto, o efeito da similaridade promovida pela imputação *hot-deck*. Tal diferencial, entretanto, não é sensível ao tipo de algoritmo de agrupamento utilizado. Ressalte-se também que o mecanismo de ausência dos dados pouco influenciou o comportamento global da técnica.

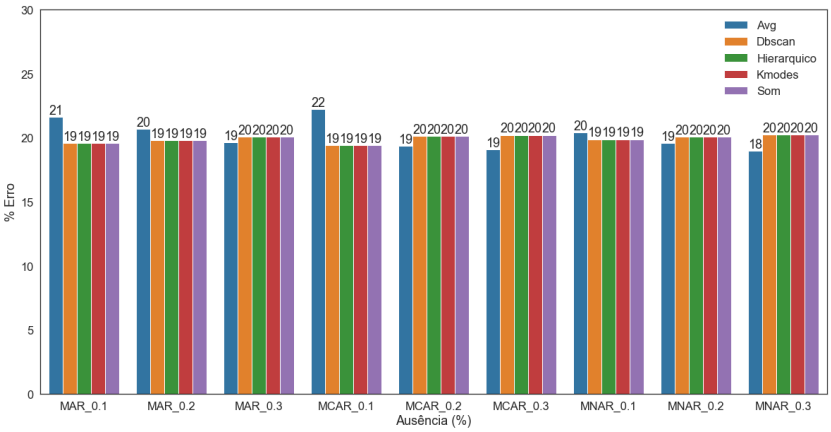


**Figura 2. Comparativo geral dos métodos testados por tipos de ausências**

## 6. Considerações Finais

Este artigo aborda a imputação em cascata, um tema pouco explorado que envolve a reutilização de valores imputados para a imputação de atributos subsequentes. Realizado em continuidade ao trabalho de Ferlin [2008], o presente estudo empregou quatro algoritmos de agrupamento — SOM, DBSCAN, K-modes e Clustering Hierárquico Aglomerativo — e utilizou o algoritmo k-NN para a imputação. Os resultados foram comparados com aqueles obtidos por meio da imputação por média, calculando-se o percentual de erro. Foram induzidas ausências em todos os atributos dos conjuntos de dados selecionados, nas proporções de 10%, 20% e 30%, sob os mecanismos MCAR, MNAR e MAR.

Referindo-se à pergunta de pesquisa, observou-se que não houve diferença significativa entre os algoritmos de agrupamento. Todavia, há ganhos na sua utilização quando comparado à imputação por média. Em bases de dados com baixa correlação entre os atributos, a imputação por média superou os outros métodos. Em contraste, nos cenários MAR 10%, MAR 20%, MCAR 10% e MNAR 10%, a imputação em cascata apresentou menores erros percentuais comparada à imputação por média, indicando uma tendência a ser mais efetiva em conjuntos de dados com menor índice de valores ausentes, independentemente do mecanismo causador das ausências ou do tipo de algoritmo de agrupamento utilizado na imputação *hot-deck*. Como trabalhos futuros, espera-se realizar novos experimentos com os mesmos conjuntos de dados, porém variando a configuração dos hiperparâmetros dos algoritmos de agrupamento. Variações na estrutura do método originalmente proposto também serão alvo de análise e experimentação. Planeja-se, outrossim, realizar novos experimentos com conjuntos de dados maiores e com outros algoritmos de agrupamento.

## Referências

Cartwright, M., Shepperd, M., and Song, Q. (2003). Dealing with missing software project data. *Software Metrics, IEEE International Symposium on*, 0:154.

Coneglian, G. and Segundo (2017). Missing data: Our view of the state of the art. *Encontros Bibli*, 22.

Ferlin, C. (2008). *Imputação em cascata: uma abordagem para imputação multivariada de dados*. PhD thesis, Tese (Doutorado em Engenharia de Sistemas e Computação) - Universidade Federal do Rio de Janeiro - Rio de Janeiro.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression And Multilevel/Hierarchical Models*, volume 3.

Gonçalves, L. M. (2021). Imputação hot-deck: Uma revisão sistemática da literatura. Master's thesis, Dissertação (Mestrado em Ciência da Computação) - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ).

Jarrahi, M. H., Memariani, A., and Guha, S. (2022). The principles of data-centric ai (dcai). *Communications of the ACM*.

Little, R. and Rubin, D. (2019). *Statistical Analysis with Missing Data*. Wiley.

Montiel, J., Read, J., Bifet, A., and Abdessalem, T. (2018). Scalable model-based cascaded imputation of missing data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10939 LNAI:64 – 76.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1988). An overview of multiple imputation. *Journal of the American Statistical Association*, page 79–84.

Wayman, J. (2003). Multiple imputation for missing data: What is it and how can i use it. *Annual Meeting of the American Educational Research Associantion*.