Imputação de Valores Omissos em Análise Descritiva de Dados, em R

Luzizila Salambiaku[1,2], Paula Prata[1,2], Maria E. Ferrão[1,3]

[1]Universidade da Beira Interior, Portugal

[2]Instituto de Telecomunicações (IT), Portugal

[3]Centro de Matemática Aplicada e Economia (CEMAPRE), Portugal

luzizila2009@gmail.com, pprata@di.ubi.pt, meferrao@ubi.pt

Abstract. *Missing values represent a frequent problem in the data analysis process. In this paper, six different imputation methods, available in software R, were used and compared. Their performance was evaluated in datasets related to the education area, namely data from the national evaluation of school performance (Prova Brasil). A sample of 20408 students was studied to test the six algorithms in four subsets of data with different percentages of missing values, considering 5%, 10%, 15% and 20% in the variables of interest. Single imputation methods (Mean, Median and Mode), methods based on machine learning (kNN and bPCA) and a multiple imputation method (MICE) were explored. The performance of each method adopted in this work was evaluated by calculating the respective imputation errors using the metrics RMSE and MAE. The results obtained show that the method of imputation by Mode provided almost constantly lower values of error.*

Resumo. *Os valores omissos representam um problema frequente no processo de análise de dados. Neste artigo foram comparados seis métodos distintos de imputação, disponíveis no software R e avaliado o seu desempenho em conjuntos de dados relacionados com a área da educação. Foi estudada uma amostra de 20408 estudantes para testar os seis algoritmos em quatro conjuntos de dados gerados por simulação com diferentes percentagens de valores omissos, considerando 5%, 10%, 15% e 20% nas variáveis de interesse. Foram explorados métodos de imputação simples (Média, Mediana e Moda), métodos baseados em aprendizagem automática (kNN e bPCA) e um método de imputação múltipla (MICE). Foi avaliado o desempenho de cada método calculando os respetivos erros de imputação através as métricas RMSE e MAE. Os resultados obtidos mostram que* a *imputação pela Moda forneceu quase de forma constante menores valores de erro.*

1. Introdução

Com o crescimento explosivo de dados disponíveis provenientes de diversas fontes, a questão sobre a garantia da qualidade desses mesmos dados torna-se cada vez mais importante (Wu *et al*., 2004). Portanto, ao lidar com um grande volume de dados no nosso dia a dia, existe um problema comum nos *datasets* (conjunto de dados), a

presença de *missing* data (dados omissos) que ocorrem numa parte ou em todas as variáveis em estudo. Rubin (1976) classificou os dados omissos considerando três mecanismos, baseados na forma como essas omissões ocorrem. Assim, diz-se que os dados omissos são MCAR - *Missing completely at random*, se ocorrem de forma completamente aleatória, isto é, quando a probabilidade de um item ter respostas omissas não depender nem dos valores observados nem dos valores omissos; os dados omissos são MAR - *Missing at random*, se ocorrem de forma aleatória, isto é, quando a ausência está relacionada com valores observados noutras variáveis, mas a causa da omissão não está relacionada com os valores ausentes em si (Zainuri *et al*., 2015) e os dados omissos saõ NMAR - *Not missing at random*, quando são não aleatórios, isto é, quando a probabilidade de um registo com dado omisso em uma variável pode depender do valor do item (variável) (Little & Rubin, 2002). A omissão de dados pode causar perda do poder estatístico e consequentemente resultar na tomada de decisões erradas (Nunes *et al*., 2009).

Neste estudo foram utilizados e comparados seis métodos distintos de imputação, disponíveis no software R e avaliado o seu desempenho em conjuntos de dados relacionados com a área da educação, nomeadamente dados da avaliação nacional do rendimento escolar do ensino básico no Brasil. Foi considerada uma amostra de 20408 estudantes para testar os diferentes algoritmos em quatro subconjuntos de dados gerados por simulação com diferentes percentagens de valores omissos. Foram explorados métodos de imputação simples como a Média (Zainuri et al., 2015; Scrobote, 2017) a Mediana (Zainuri et al., 2015; Scrobote, 2017) e a Moda (Scrobote, 2017); métodos baseados em aprendizagem automática como o kNN – K-nearest neighbors (Driss et al., 2020) e o bPCA - Bayesian Principal Component Analysis (Qu et al., 2008) e o método de imputação múltipla, MICE - Multiple Imputations by Chained Equations (Nunes et al., 2010; Azur et al., 2007).

Foi avaliado o desempenho de cada método explorado neste estudo calculando os respetivos erros de imputação através as métricas RMSE (*Root Mean Square Error*) e MAE (*Mean Absolute Error*). Neste artigo, apresentamos na secção 2 os dados e os métodos usados; na secção 3 descrevem-se os resultados obtidos, e a análise em termos do erro obtido para cada método e subconjunto de dados estudado. Finalmente, a secção 4, apresenta as conclusões.

2. Material e Métodos

Nesta secção apresentamos o conjunto de dados em análise, os métodos de imputação de valores omissos aplicados, as métricas de erro usadas para avaliar a qualidade da imputação e um esquema geral das etapas do estudo realizado.

2.1. Descrição de dados

Para avaliação dos diferentes métodos de imputação, foi considerado um conjunto de dados completo relativo ao estudo de Avaliação Nacional do Rendimento Escolar, conhecido como Prova Brasil, disponível em http://portal.mec.gov.br/prova-brasil. O conjunto de dados original (Prova Brasil 2017) contém informações relativas a centenas de milhares de estudantes tendo uma grande ocorrência de valores omissos.

Para a simulação realizada neste trabalho, foi utilizada uma amostra composta com valores completos das variáveis utilizadas. Os mesmos dados utilizados por (Ferrão

& Prata, 2019) e Ferrão, Prata & Alves, 2020) e que foram obtidos por eliminação *listwise*. Partindo desse conjunto de dados com 20408 registos completos em três variáveis, foram gerados quatro conjuntos de dados com 5% (Miss5), 10% (Miss10), 15% (Miss15) e 20% (Miss20) de valores omissos. As variáveis consideradas foram: DL (desempenho do estudante na leitura), SSE (situação socioeconómica do estudante) e TSR (trajetória do estudante sem repetição). A geração dos quatro conjuntos de dados foi realizada através de extrações aleatórias nas variáveis DL e SSE, respetivamente no grupo de estudantes com desempenho mais baixo e no grupo de estudantes mais pobres. A variável "TSR" é uma variável dicotómica e sem valores omissos.

2.2. Métodos de imputação

Existem diversas técnicas e vários métodos de imputação de valores omissos na literatura (Vinha & Laros, 2018). Nesta pesquisa, foram usados e comparados seis métodos de imputação, disponíveis no programa R.

Resumidamente, foram usados três métodos de imputação simples (Média, Mediana e Moda). Estes métodos são classificados como métodos de imputação por constantes (Mcknight, Mcknight, & Sidani, 2007) e consistem em substituir todos os valores omissos de uma variável por um único valor. Foram também usados dois métodos baseados em aprendizagem automática (kNN e bPCA), estes métodos recorrem aos dados observados e aprendem com esses dados para posteriormente permitir dar como entrada os dados para as variáveis completas de uma dada observação e estimar os valores para as variáveis onde existem valores omissos (Costa, 2018); finalmente, foi aplicado um método de imputação múltipla (MICE). Com o intuito de tentar construir um método que reflectisse a incerteza sobre as previsões de valores omissos, Little & Rubin (1987) propuseram o método de imputação múltipla, o qual substitui cada valor omisso por um conjunto de valores plausíveis que representam essa incerteza. Foi usado o método PMM - *Predictive Mean Matching* (Schnitt *et al.*, 2015), disponível no package MICE, com m=5 o número de conjuntos de dados imputados, sendo realizadas 50 iterações. De notar que neste método o número de iterações depende da convergência da imputação que neste estudo não foi avaliada.

2.3. Métricas de avaliação

Para comparar os resultados e a qualidade dos valores obtidos na simulação dos seis métodos de imputação utilizados nesta pesquisa, foram adotadas duas métricas estatísticas:

RMSE (*Root Mean Square Error*) calcula a raiz quadrada do erro quadrático médio entre os valores observados e os valores imputados ver equação (1);

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_i^{Obs} - V_i^{Imp})^2} \qquad (1)$$

MAE (*Mean Absolute Error*) calcula o erro absoluto médio dos erros entre os valores observados e os valores imputados ver equação (2).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|V_i^{Obs} - V_i^{Imp}\right| \quad (2)$$

Finalmente, foram avaliados e comparados os tempos de execução para cada método de imputação.

2.4. Estudo de Simulação

A Figura 1 ilustra as diferentes etapas do estudo de simulação para o tratamento de valores omissos. Partindo do conjunto de dados original (sem valores omissos), foram introduzidas diferentes percentagens de valores omissos (5%, 10%, 15% e 20%) gerados como descrito na Secção 2.1. De seguida, estes valores foram imputados usando seis métodos e três critérios de avaliação, cálculo da raiz quadrada do erro quadrático médio (RMSE) cálculo do erro absoluto médio (MAE) e por fim foi medido o tempo de execução.

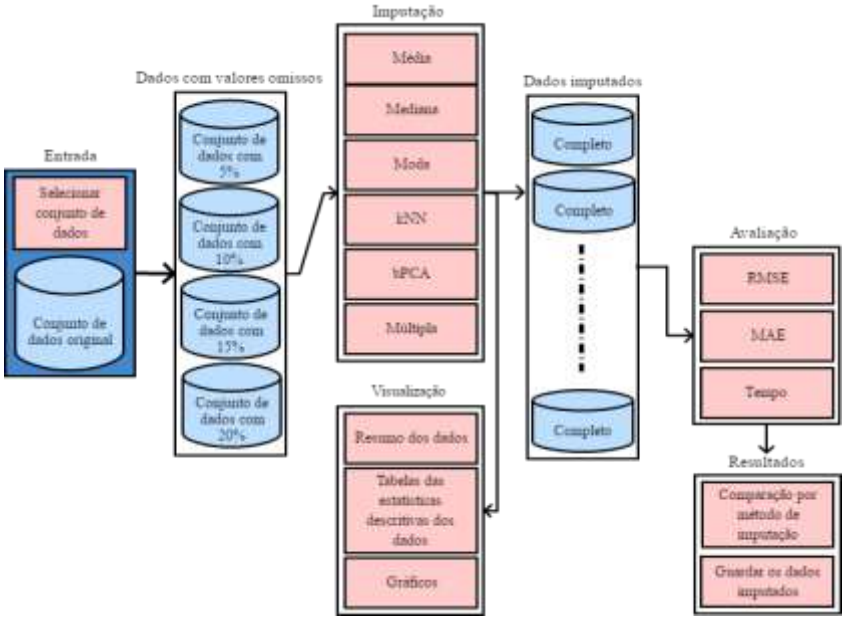

Figura 1. Etapas do estudo de simulação.

3. Resultados

São apresentados nesta secção os resultados obtidos pela imputação por cada método, nas diferentes percentagens de valores omissos, e a avaliação do erro correspondente. Começamos por analisar os padrões de valores omissos existentes em cada subconjunto de dados.

Na Figura 2, são mostrados os padrões de valores omissos em cada subconjunto de dados. As áreas com a cor azul representam os dados observados e a rosa indicam a localização dos valores não observados (omissos). No gráfico A (conjunto Miss5) observamos os seguintes padrões de valores omissos: 4,5% de valores omissos apenas na variável DL, 4,4% de valores omissos apenas na variável SSE e 0,3% de valores omissos nas duas variáveis. No gráfico B (conjunto Miss10) observamos 8,3% de omissos em SSE, 8,3% em DL e 1,4% de omissos nas duas variáveis. No gráfico C (Miss15) 12,3% de omissos em SSE, 12,3% em DL e 3,0% em ambas as variáveis. Finalmente, no gráfico D (Miss20) 15,3% em DL, 15,0% em SSE e 4,6% em ambas as variáveis.

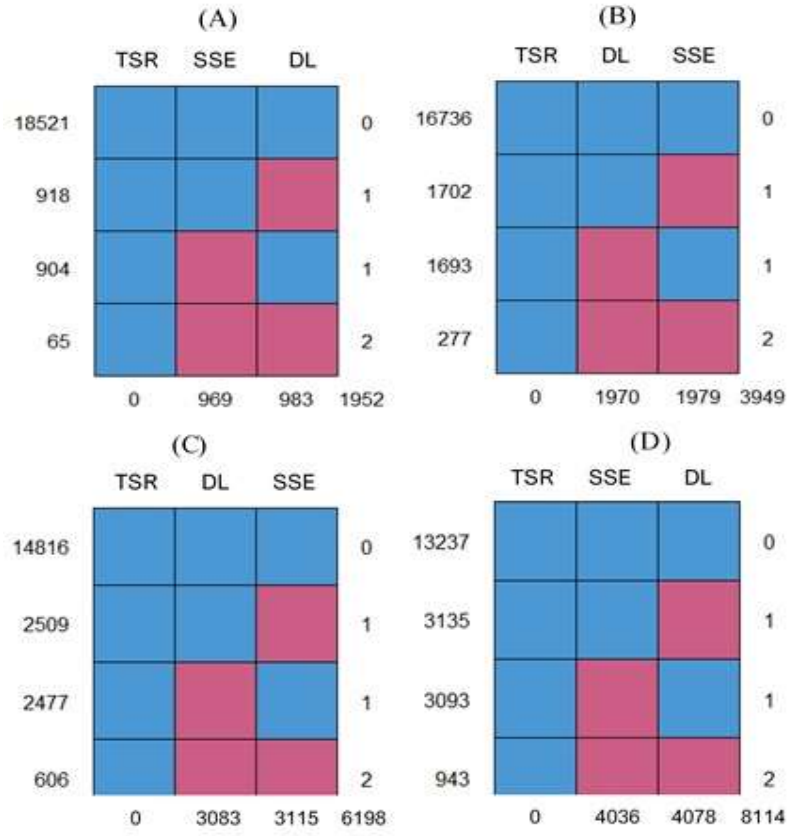

Figura 2. Representação gráfica de padrões de valores omissos por variável nos diferentes subconjuntos

Na Tabela 1, são apresentadas as estatísticas descritivas do conjunto de dados original (completo) à esquerda e as estatísticas descritivas dos quatro subconjuntos de dados com valores omissos nas variáveis de interesse. A última linha da tabela contém o número de valores omissos para cada caso.

Tabela 1. Estatísticas descritivas dos conjuntos de dados estudados (antes da imputação)

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.500	4.400	-0.400	4.500	-0.300	4.600	-0.200	4.700
Mediana	0.040	4.960	0.100	5.000	0.200	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.151	5.166	0.208	5.229	0.261	5.289	0.297	5.327
3º Quartil	0.690	5.710	0.700	5.800	0.800	5.800	0.800	5.900	0.900	5.900
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.864**	**1.011**	**0.850**	**0.998**	**0.848**	**1.001**	**0.859**	**1.014**
NA's	0	0	983	969	1970	1979	3083	3115	4078	4036

Na Tabela 2 são apresentados os resultados das estatísticas descritivas para os conjuntos obtidos por imputação pela média nas diferentes percentagens. Os valores com a cor azul são os que são diferentes dos apresentados na TabelaTabela **1**, depois da imputação.

Tabela 2. Estatísticas descritivas dos conjuntos com dados imputados pela Média

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.400	4.500	-0.300	4.600	-0.200	4.700	-0.100	4.800
Mediana	0.040	4.960	0.151	5.100	0.208	5.200	0.261	5.289	0.297	5.327
Média	0.073	5.082	0.151	5.166	0.208	5.227	0.261	5.289	0.297	5.327
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.843**	**0.987**	**0.808**	**0.948**	**0.781**	**0.922**	**0.769**	**0.908**

De seguida apresentam-se tabelas equivalentes à Tabela 2, para cada um dos outros métodos de imputação: estatísticas descritivas após imputação pela Mediana (Tabela 3), após imputação pela Moda (Tabela 4), após imputação com k-vizinhos mais próximos, kNN, com $K = 6$ (Tabela 5), após imputação por análise de componentes principais, bPCA, (Tabela 6) e finalmente, após imputação múltipla com MICE (Tabela 7).

Tabela 3. Estatísticas descritivas dos conjuntos com dados imputados pela Mediana

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.400	4.500	-0.300	4.600	-0.200	4.700	-0.100	4.800
Mediana	0.040	4.960	0.100	5.000	0.200	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.148	5.158	0.207	5.217	0.252	5.275	0.298	5.302
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.843**	**0.988**	**0.808**	**0.949**	**0.782**	**0.922**	**0.769**	**0.910**

Tabela 4. Estatísticas descritivas dos conjuntos com dados imputados pela Moda

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.400	4.500	-0.300	4.600	-0.200	4.600	-0.100	4.800
Mediana	0.040	4.960	0.000	5.000	0.000	5.000	0.000	5.000	0.000	5.000
Média	0.073	5.082	0.139	5.139	0.178	5.168	0.206	5.184	0.218	5.263
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.845**	**0.994**	**0.813**	**0.966**	**0.792**	**0.955**	**0.785**	**0.917**

Tabela 5. Estatísticas descritivas dos conjuntos com dados imputados com kNN

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.500	4.450	-0.400	4.600	-0.300	4.700	-0.100	4.750
Mediana	0.040	4.960	0.100	5.000	0.150	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.144	5.152	0.205	5.199	0.244	5.265	0.306	5.281
3º Quartil	0.690	5.710	0.700	5.700	0.800	5.700	0.800	5.800	0.800	5.800
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.853**	**0.994**	**0.828**	**0.966**	**0.849**	**0.937**	**0.800**	**0.939**

Tabela 6. Estatísticas descritivas dos conjuntos com dados imputados com bPCA

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.489	4.500	-0.400	4.600	-0.293	4.700	-0.200	4.800
Mediana	0.040	4.960	0.100	5.000	0.200	5.100	0.200	5.200	0.300	5.292
Média	0.073	5.082	0.141	5.163	0.193	5.225	0.240	5.283	0.274	5.320
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.848**	**0.988**	**0.816**	**0.949**	**0.794**	**0.924**	**0.786**	**0.911**

Tabela 7. Estatísticas descritivas dos conjuntos com dados imputados com MICE

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.500	4.400	-0.400	4.500	-0.300	4.600	-0.200	4.700
Mediana	0.040	4.960	0.100	5.000	0.100	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.141	5.162	0.192	5.222	0.231	5.282	0.263	5.309
3º Quartil	0.690	5.710	0.700	5.700	0.800	5.800	0.800	5.900	0.800	5.900
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.865**	**1.013**	**0.851**	**1.001**	**0.852**	**1.005**	**0.866**	**1.009**

Nas Figura 3 e Figura 4 apresenta-se o comportamento do erro RMSE para os vários métodos de imputação nos diferentes conjuntos de dados imputados, respectivamente para as variáveis SSE e DL.

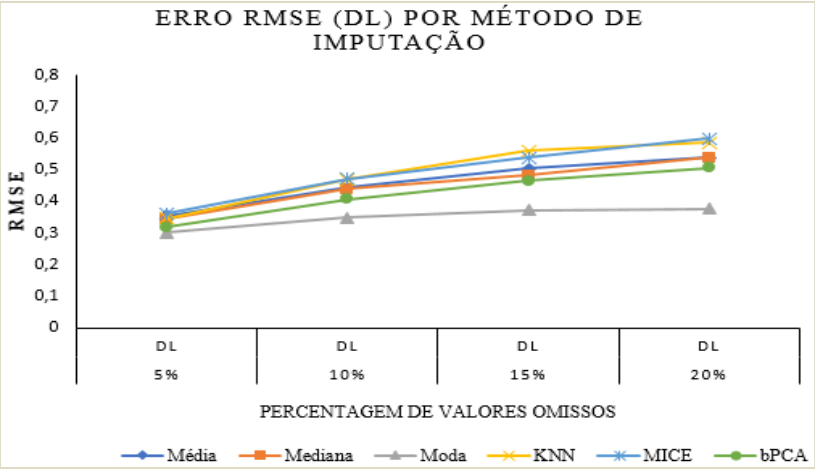

Figura 3. Erro RMSE para a situação socioeconómica do estudante, SSE.

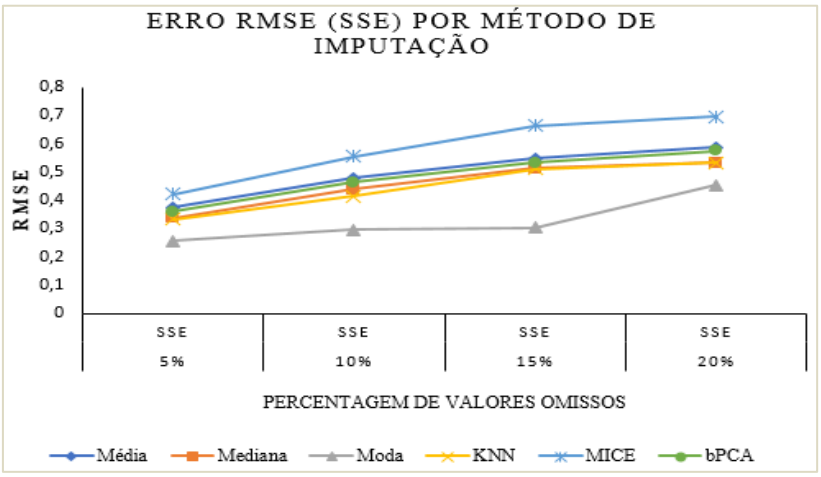

Figura 4. Erro RMSE para o desempenho do estudante na leitura, DL.

As Figuras 5 e 6 mostram o comportamento do erro MAE para os seis métodos de imputação nos diferentes conjuntos de dados imputados, respectivamente para as variáveis SSE e DL.

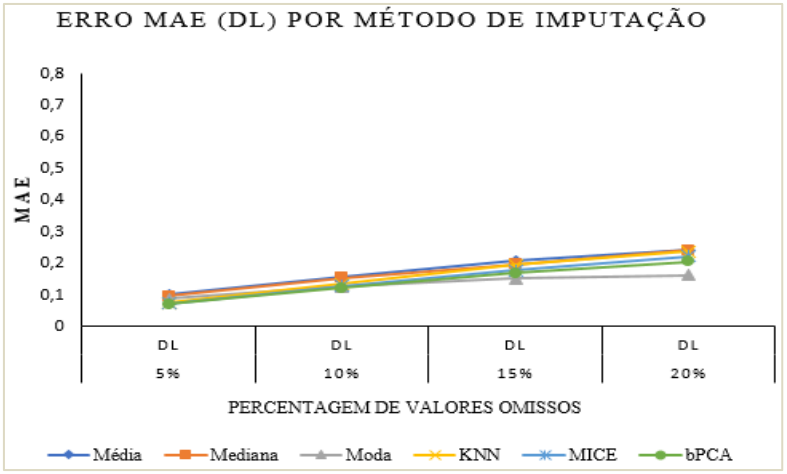

Figura 5. Erro MAE para a situação socioeconómica do estudante, SSE.

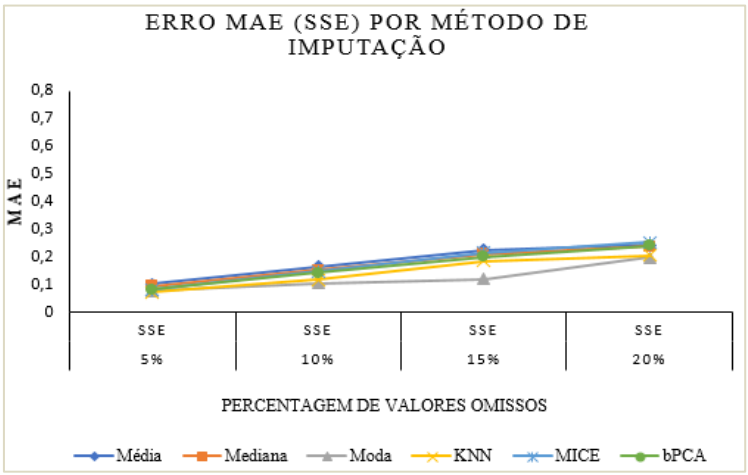

Figura 6. Erro MAE para o desempenho do estudante na leitura DL

A Figura 7 apresenta graficamente os resultados dos tempos de execução de cada método nas diferentes percentagens de valores omissos. Para medir os tempos foram utilizadas duas funções *tic()* e *toc()* disponíveis no *package tictoc*.

Figura 7. Evolução dos tempos médios da execução nos diferentes conjuntos de dados por método de imputação.

Podemos observar facilmente que o maior erro RMSE foi introduzido pelo MICE no caso da variável SSE, e para a variável DL o MICE e o kNN partilham os piores resultados. Para o erro MAE, as diferenças são bastante diluídas. A Moda é o método que na generalidade dos casos apresenta menor erro, mas a diferença para os outros métodos só é clara para os conjuntos com mais valores omissos.

Enquanto com o RMSE o MICE foi o pior método com valores de erro mais acentuados a seguir da média, a moda foi consistentemente o melhor método entre as diferentes percentagens de valores omissos e métricas de avaliação de desempenho, superando todos os outros métodos quando aplicado ao conjunto de dados estudados com base nos critérios de RMSE e MAE.

Quanto ao desempenho, podemos observar que os métodos de imputação pela Média, Mediana e Moda foram todos muito mais rápidos, com uma duração de algumas décimas de segundos. Os métodos kNN e bPCA, para estes conjuntos de dados, tem tempos de execução na ordem das dezenas de segundos, enquanto o MICE tem um tempo de execução na ordem das várias centenas de segundos.

4. Conclusão

A existência de valores omissos em conjuntos de dados é um problema frequente na análise de dados, pelo que encontrar as melhores maneiras de lidar com estes valores é uma área de estudo importante. Neste estudo foram realizadas algumas análises da aplicação de métodos de imputação de valores omissos no conjunto de dados do estudo de Avaliação Nacional do Rendimento Escolar (Prova Brasil 2017). Foram testados seis métodos de imputação de valores omissos sendo três de imputação por valores constantes (Média, Mediana e Moda), dois baseados em aprendizagem automática (kNN

e bPCA) e um baseado na imputação múltipla com MICE com o objetivo de comparar e determinar melhores métodos e técnicas de alcançar valores dos resultados mais próximos aos originais.

Os seis algoritmos de imputação foram estudados em quatros subconjuntos de dados dos estudantes com distintas percentagens de valores omissos, tais como 5%, 10%, 15% e 20% de valores omissos gerados aleatoriamente. A seguir cada subconjunto de valores imputados passou por avaliação de resultados com duas métricas de erro RMSE e MAE. Calcularam-se estatísticas descritivas simples para o conjunto de dados original e para cada conjunto de valores imputados com os diferentes métodos. Perante os resultados obtidos, pode-se concluir o seguinte:

Em relação aos valores originais, observa-se em todas as percentagens de valores omissos (5%, 10%, 15% e 20%) imputados pela Moda, valores menores de erro na maioria das posições.

Em 15 e 20% de valores omissos, o método de imputação pela Moda forneceu menores valores das estimativas de erros nas duas variáveis de interesse. Isto é valores mais próximos dos originais.

Na medida que cresce a percentagem de valores omissos, o valor da média também aumenta e o desvio padrão entre as imputações decresce na maior parte dos casos.

O método que apresentou um erro mais levado foi na generalidade dos casos a imputação por MICE. No entanto, neste estudo não se avaliou a convergência dos valores imputados, tendo sido realizadas 50 iterações em todos os casos.

Dos métodos de imputação simples, a imputação pela média foi a que apresentou maiores valores de erro.

Por se tratar de resultados obtidos em situações particulares, utilizando conjuntos de dados educacionais, estes não podem ser generalizados.

Referências

Azur, M. J., Stuart, E. A., and C. F. & Leaf, P. J. (2007) “A dimensional approach to developmental psychopathology”, Int. J. Methods Psychiatr. Res., vol. 16, no. S1, pp. S16–S23, http://doi.wiley.com/10.1002/mpr.329.

Driss, K., Boulila, W., Batool, A. and Ahmad, J. (2020) “A Novel approach for classifying diabetes’ patients based on imputation and machine learning”, 2020 Int. Conf. UK-China Emerg. Technol. UCET 2020, pp. 14–17. DOI: 10.1109/UCET51115.2020.9205378.

da Costa, C. F. G. (2018) “Exploração de dados em falta: Uma abordagem visual”, Dissertação de mestrado, http://hdl.handle.net/10316/86107.

Ferrão, M. E., & Prata, P. (2019). Computing topics on multiple imputation in Big Identifiable Data using R: An application to educational research. In S. M. et Al. (Ed.), Lecture Notes in Computer Science (Vol. 11621, pp. 12–24). Springer Cham. https://doi.org/10.1007/978-3-030-24302-9_2.

Ferrão, M. E., Prata, P., & Alves, M. T. G. (2020). Multiple imputation in big identifiable data for educational research: An example from the Brazilian education assessment system. Ensaio: Avaliação e Políticas Públicas Em Educação, 28(108), 599–621. https://doi.org/10.1590/s0104-4036202000280234.

Little, R.J.A., Rubin, D.B. (2002) "Statistical Analysis with Missing Data", (2nd ed.).

Little, R. J. A. and Rubin, D. B. (1987) "Statistical analysis with missing data", John Wiley & Sons, USA.

Mcknight, B. P. E., Mcknight, K. M. and Sidani, S. (2007) "A Gentle Introduction to Missing Data", Guilford Press, 251 pages.

Nunes, L., Klück, M. and Fachel, J. (2009) "Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos Multiple imputations for missing data: a simulation with epidemiological data", *Cad. Saude Publica*, vol. 25, no. 2, pp. 268–278. http://dx.doi.org/10.1590/S0102-311X2009000200005.

Nunes, L. N., Klück, M. M. and Fachel, J. M. G. (2010) "Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica", Rev. Bras. Epidemiol., vol. 13, no. 4, pp. 596–606. http://dx.doi.org/10.1590/S1415-790X2010000400005.

Qu, L., Zhang, Y., Hu, J., Jia, L. and Li, L. (2008) "A BPCA based missing value imputing method for traffic flow volume data", IEEE Intell. Veh. Symp. Proc., vol. D, no. 10, pp. 985–990. DOI: 10.1109/IVS.2008.4621153.

Rubin, D.B. (1976) "Inference and missing data", Biometrika 63: 581-592.

Scrobote, A. (2017) "Uma análise da aplicação de algoritmos de imputação de valores faltantes em bases de dados multirrótulo", http://repositorio.utfpr.edu.br/jspui/handle/1/15935.

Schmitt, P., Mandel, J. and Guedj, M. (2015) "A Comparison of Six Methods for Missing Data Imputation", J. Biom. Biostat., vol. 06, no. 01, pp. 1–6. DOI: 10.4172/2155-6180.1000224.

Vinha, L. G. do A. and Laros, J. A. (2018) "Dados ausentes em avaliações educacionais: comparação de métodos de tratamento", *Estud. em Avaliação Educ.*, no. x, p. 1. DOI: 10.18222/eae.v0ix.3916.

Wu, C., Wun, C., Chou, H. (2004) "Using association rules for completing missing data", Proceedings of the 4th International Conference on Hybrid Intelligent Systems, Taiwan, pp. 236-241. DOI: 10.1109/ICHIS.2004.91.

Zainuri, N. A., Jemain, A. A. and Muda, N. (2015) "A Comparison of Various Imputation Methods for Missing Values in Air Quality Data (Perbandingan Pelbagai Kaedah Imputasi bagi Data Lenyap untuk Data Kualiti Udara)", Sains Malaysiana, vol. 44, no. 3, pp. 449–456. DOI: 10.17576/jsm-2015-4403-17.