CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

IMPUTAÇÃO *HOT-DECK*: UMA REVISÃO SISTEMÁTICA DA LITERATURA

Leandro Maia Gonçalves

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador(a): Jorge de Abreu Soares

Rio de Janeiro,
Janeiro de 2021.

IMPUTAÇÃO *HOT-DECK*: UMA REVISÃO SISTEMÁTICA DA LITERATURA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Leandro Maia Gonçalves

Banca Examinadora:

_____

Presidente, Dr. Professor Jorge de Abreu Soares (CEFET/RJ)

_____

Professor Dr. Eduardo Soares Ogasawara (CEFET/RJ)

_____

Professor Dr. José Maria da Silva Monteiro Filho (UFC)

Rio de Janeiro,
Janeiro de 2021.

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

G635 Gonçalves, Leandro Maia
Imputação Hot-Deck: uma revisão sistemática da literatura / Leandro Maia Gonçalves — 2021.
131f : il. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2021.
Bibliografia : f. 112-131
Orientador: Jorge de Abreu Soares

1. Mineração de dados (Computação). 2. Bases de dados. 3. Aprendizado de máquina. I. Soares, Jorge de Abreu (Orient.). II. Título.

CDD 005.74

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

RESUMO

Imputação *Hot-deck*: Uma revisão Sistemática da Literatura

As organizações têm percebido que investir na transformação de dados em informação com o objetivo de auxiliar o processo de tomada de decisões pode trazer vantagens competitivas. À vista disso, no cenário atual em que os dados crescem em volume, velocidade e variedade, nota-se que tal expansão é acompanhada do aumento de dados ausentes, que podem trazer problemas de interpretação para analistas e pesquisadores. A exclusão destes casos não pode necessariamente ser considerada uma solução, independente do volume dos dados, devido aos seus riscos de geração de vieses ou tendências. Logo, a imputação de dados revela-se uma tarefa fundamental no pré-processamento de dados, capaz de melhorar a sua análise. A imputação *hot-deck* é uma abordagem que se destaca neste contexto devido à sua capacidade de estimar com melhor precisão e preservar as diferenças individuais entre os sujeitos no processo de imputação. Neste estudo, é apresentada uma revisão sistemática sobre técnicas de imputação *hot-deck* realizada na base *Scopus*, com o objetivo de avaliar como ocorre a evolução dos estudos sobre este tema ao longo dos anos. Este trabalho também propõe uma taxonomia que busca classificar, ordenar e estabelecer hierarquias para as técnicas de imputação. Como resultado deste trabalho, verificou-se 63% dos artigos investigados não identificaram adequadamente os mecanismos de ausência em seus experimentos, 72% dos algoritmos de agrupamento utilizados na abordagem *hot-deck* estão contidos na categoria *Partitioning Based*, sendo 75% desta representada pelos algoritmos *Random hot-deck, K-Nearest-Neighbor e K-means.* Com relação à reprodutibilidade dos experimentos, 30% dos artigos apresentaram pseudocódigos dos algoritmos utilizados, 42% utilizaram conjuntos de dados públicos, 45% compararam os resultados da imputação com o conjunto de dados original. Destaca-se que apenas 1% dos artigos apresentou código fonte em repositório aberto, deixando uma importante lacuna no que tange à reprodutibilidade de experimentos nesta área.

Palavras-chave: Imputação de dados *hot-deck*. Revisão Sistemática. Dados Ausentes.

ABSTRACT

Hot-deck imputation: A systematic review of literature

Organizations have realized that investing in transforming data into information to assist the decision-making process can bring competitive advantages. Thus, in the current scenario in which data grows in volume, speed, and variety, such expansion is accompanied by an increase in missing data, bringing interpretation problems for analysts and researchers. The exclusion of these cases cannot necessarily be considered a solution, regardless of data volume, due to its risks of generating bias or trends. Therefore, data imputation proves to be a fundamental task in data pre-processing, capable of improving its analysis. Hot-deck imputation is an approach that stands out in this context due to its ability to estimate more accurately and preserve individual differences between subjects in the imputation process. In this study, a systematic review of hot-deck imputation techniques performed on the Scopus database evaluates how the evolution of studies on this topic has occurred over the years. This work also proposes a taxonomy that aims to classify, order, and establish hierarchies for imputation techniques. As a result, 63% of the investigated articles did not adequately identify the missing mechanisms in their experiments; the hot-deck approach used 72% of the clustering algorithms in the Partitioning Based category; and 75% represented by the algorithms random hot-deck, K-Nearest-Neighbor, and K-means. Regarding the experiment's reproducibility, 30% of the articles presented pseudocodes for the algorithms used, 42% used public data sets, and 45% compared to the original data set's imputation results. It is noteworthy that only 1% of the articles presented source code in an open repository, leaving an essential lack regarding the reproducibility of experiments in this area.

Keywords: Hot-deck imputation. Systematic review. Missing Data.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

3DMICE	*3-dimensional Multiple Imputation with Chained Equations*
ANN	*Artificial Neural Networks*
BHM	*Bayesian Hierarquical Model*
BIRCH	*Balanced Iterative Reducing and Clustering Using Hierarchies*
BN	*Bayesian Network*
CLARA	*Clustering Large Applications*
CLARANS	*Clustering Large Applications based upon RANdomize*
CLIQUE	*Clustering In QUEst*
CURE	*Clustering Using Representatives*
DBCLASD	*Distribution Based Clustering of LArge Spatial Databases*
DBSCAN	*Density-Based Spatial Clustering on Applications with Noise*
DENCLUE	*DENsitbased CLUstering*
Echidna	*Efficient Clustering of Hierarchical Data for Network Traffic Analysis*
ECM	*Evolving Clustering Method*
EM	*Expectation - Maximization*
ESBE	Engenharia de Software Baseada em Evidências
FCM	*Fuzzy C-Means*
FEFI	*Fully Efficient Fractional Imputation*
FHDI	*Fractional Hot Deck Imputation*
FIML	*Full Information Maximum Likelihood*
FKM	*Fuzzy K-Means*
GMM	*Gaussian Mixture Model*
HAC	*Hierarquical Agglomerative Clustering*
HDBB	*Hot-Deck Bayesian Bootstrap*
HDMM	*Hot-Deck Multiple Imputation*
IM	*Ignorable Missing*
KNN	*K-Nearest Neighbors*
MAR	*Missing at Random*
MCAR	*Missing Completely at Random*
MCMC	*Markov Chain Monte Carlo*
MICE	*Multiple Inputation by Chained Equation*
MLE	*Maximum Likelihood Estimation*
MM	*Multinominal Model*
MST	*Minimum Spanning Tree*
NMAR	*Not Missing at Random*
NNHD	*Nearest Neighbors Hot-Deck*
OPTICS	*Ordering Points To Identify the Clustering Structure*
OptiGrid	*Optimal Grid-Partitioning*
PAM	*Partitioning Around Medoids*
PBE	Prática Baseada em Evidências
PCA	*Principal Component Analysis*
PMM	*Predictive Mean Matching*
PSO	*Particle Swarm Optimization*

RHD	*Random Hot-Deck*
ROCK	*Clustering using linKs*
SHD	*Sequential Hot-Deck*
SOM	*Self-Organizing Map*
STING	*STatistical INformation Grid*
TDIDT	*Top-Down Induction of Decision Tree*
WaveCluster	*Clustering based on Wavelet Transforms*

SUMÁRIO

1- Introdução

A quantidade de dados em nosso mundo está crescendo a uma taxa impressionante, fazendo com que pesquisadores em diferentes áreas do conhecimento admitam que a humanidade entrou na “Era do Big Data” (WANG et al., 2014). Este aumento no volume de dados, na velocidade de sua produção, e na variedade de fontes em que os dados se encontram desafia os profissionais e pesquisadores da informação a trabalhar com maior eficiência e eficácia (CONEGLIAN; GONÇALEZ; SANTARÉM SEGUNDO, 2017).

Estas características de volume, velocidade e variedade dos dados com constantes mudanças quantitativas e qualitativas têm feito com que as organizações repensem seus negócios e processos, uma vez que uma coleta e análise de dados de forma consistente pode proporcionar vantagens competitivas para uma organização (FONTES; DA SILVA; DE ALMEIDA, 2016).

Atualmente vivemos em uma sociedade em rede, como resultado da adoção e apropriação das tecnologias de informação e comunicação, a qual criou um valor social e econômico para a informação que, há alguns anos, não era possível imaginar (MORENO, 2015). No entanto, neste mundo onde a demanda pela gestão de grande volume de dados é uma realidade, o problema da falta de dados (*missing data*) é generalizado, uma vez que é raro encontrar um banco de dados que não contenha valores ausentes (LAROSE; LAROSE, 2015). O modo como o analista lida com os dados ausentes pode alterar o resultado de sua análise. Por isso, é importante aprender metodologias para lidar com dados ausentes de forma que os resultados não sejam influenciados (LAROSE; LAROSE, 2015).

Isto posto, com o aumento do volume de dados, o problema da ausência de dados cresce cada vez mais e, por conseguinte, podemos cair em problemas estatísticos, tal como vieses, que podem fazer com que as organizações tomem decisões equivocadas.

A ausência de informação e a presença de dados que se afastam do que é tido como normal são um mal endêmico nas ciências sociais e nas análises quantitativas. Apesar de os usuários, de modo geral, serem informados da presença de registros sem informação (contendo valores ausentes), há várias implicações estatísticas que envolvem trabalhar neste contexto e que geram uma deficiente interpertação dos dados ou modelos preditivos imprecisos.

As rotinas dos modelos computacionais genericamente assumem que o usuário esteja trabalhando com dados completos e, desta forma, incorporam-se ações, nem sempre as mais adequadas, para imputar observações sem que o usuário esteja ciente disso. Logo, a substituição inadequada de informações introduz tendências e reduz o poder exploratório de métodos estatísticos, reduzindo a eficiência da fase de inferência e podendo até invalidar as conclusões de uma investigação.

Neste contexto, técnicas de imputação desenvolvidas por pesquisadores têm apresentado consideráveis avanços na qualidade do dado imputado, criando um grande universo de algoritmos e workflows científicos. A ênfase escolhida dentro desse assunto tão amplo foi a imputação *hot-deck*, já que esta não é recente e possui uma contribuição histórica para os novos métodos de imputação que surgiram posteriormente. Além dela ser muito utilizada até os dias de hoje, muitos outros algoritmos são a evolução da imputação *hot-deck*, como por exemplo a imputação múltipla. Denomina-se *hot-deck* o agrupamento que precede a imputação. Isto traz uma importante característica para o processo de preenchimento de dados ausentes e para a preservação da variedade das amostras, ou seja, das diferenças individuais, evitando tendências ou vieses, fundamental para os modelos preditivos, em especial os estatísticos.

Assim, o objetivo desta pesquisa é avaliar como ocorre a evolução dos estudos sobre imputação *hot-deck* ao longo dos anos, verificar a qualidade e a reprodutibilidade dos mesmos e identificar novos caminhos de pesquisa para trabalhos futuros. Propõe-se, neste estudo, uma taxonomia de imputação de dados, inspirada no trabalho de Soares (2007), haja vista que o estudo taxonômico busca fazer uma classificação correta, ordenada e hierarquizada.

A metodologia utilizada para o presente estudo é a revisão sistemática, já que tal método possibilita identificar as melhores evidências e sintetizá-las em prol de fundamentar propostas. A base de dados escolhida para a execução desta pesquisa foi a Scopus[1], do grupo Elsevier, dada sua capacidade de indexação de conteúdo científico – o que potencializa sobremaneira sua visibilidade.

Para as questões norteadoras do trabalho:

1. Quais são os principais periódicos que publicam artigos sobre imputação *hot-deck*?
2. Quantos estudos comparativos com a utilização de *hot-deck* foram realizados?

[1] https://www.scopus.com/

3. Os padrões de ausências são adequadamente identificados nos estudos experimentais?
4. Os mecanismos de ausências são adequadamente identificados nos estudos experimentais?
5. Ao longo dos anos está ocorrendo uma maior utilização de *hot-deck* em estudos de imputação híbrida?
6. Quantos estudos de imputação híbrida utilizaram a técnica *hot-deck*? Destes estudos, quantos realizaram a Imputação Múltipla?
7. Quantos estudos de imputação híbrida utilizaram métodos estatísticos?
8. Qual categoria de algoritmos de agrupamento é a mais utilizada na imputação *hot-deck*?
9. Quais são os principais algoritmos utilizados na etapa de agrupamento que precede a imputação?
10. Quantos artigos realizaram a remoção completa de casos? O estudo ser descritivo, preditivo ou prescritivo impacta na escolha pela remoção completa de casos?
11. Os estudos realizados podem ser reproduzidos? Estes estudos apresentaram pseudocódigo, código em repositório aberto ou conjunto de dados público? Os experimentos realizaram a comparação dos resultados com o conjunto de dados original?
12. Quais os principais conjuntos de dados utilizados nos estudos experimentais de imputação *hot-deck*?

A estrutura dos demais capítulos desta dissertação é a que segue: no Capítulo 2 são analisados os fundamentos teóricos de um amplo conjunto de métodos de pesquisa sobre imputação de dados. A primeira parte descreve os vieses sobre ausência de dados, a teoria em que se baseia e a forma como se aplica a imputação de dados, levando em consideração seus benefícios e limitações. Neste sentido, conceitua-se o agrupamento de dados e suas classificações e os principais algoritmos de cada conjunto. Na sequência, o Capítulo 3 descreve a metodologia de revisão sistemática empregada e os procedimentos empregados para a seleção e avaliação dos achados. O Capítulo 4 apresenta os resultados obtidos e sua discussão. Por fim, as considerações finais e as direções para trabalhos futuros figuram no Capítulo 5.

2- Referencial Teórico

Neste capítulo, serão apresentados os conceitos referentes à metodologia de pesquisa utilizada nesta dissertação e à teoria de imputação de dados, tema deste estudo. A primeira seção expõe os fundamentos de revisão sistemática, sua importância e contribuição para a comunidade acadêmica. Em seguida, a segunda seção mostra os principais conceitos sobre ausência de dados, descrevendo suas possíveis causas, tipos, padrões e mecanismos de ausência.

Na terceira parte deste capítulo, as técnicas de imputação de dados são abordadas: imputação global, imputação *hot-deck*, imputação múltipla e imputação composta. Por fim, a última seção deste capítulo apresenta o conceito de agrupamento de dados e suas classificações de acordo com a literatura, seguido da explicação sobre cada categoria de classificação e os principais algoritmos de cada categoria.

2.1 – Revisão Sistemática

A pesquisa e a prática baseadas em evidências foram desenvolvidas inicialmente na área de medicina, uma vez que pesquisas indicaram que a opinião de especialistas com base puramente na opinião médica não era tão confiável quanto o conselho baseado na acumulação de resultados de experimentos científicos (KITCHENHAM et al., 2009; WOHLIN, 2014). Desde então, muitos domínios de conhecimento adotaram esta abordagem, por exemplo, criminologia, política social, economia, enfermagem, dentre outros.

A partir da medicina baseada em evidências, o objetivo da Engenharia de Software Baseada em Evidências (ESBE) é fornecer os meios pelos quais as melhores evidências atuais de pesquisa podem ser integradas com a experiência prática e valores humanos no processo de tomada de decisão em relação ao desenvolvimento e manutenção do software (DYBA; KITCHENHAM; JORGENSEN, 2005). Neste contexto, a evidência é definida como uma síntese da melhor qualidade de estudos científicos sobre um tópico específico ou questão de pesquisa.

A revisão de literatura refere-se à fundamentação teórica que será adotada para tratar o tema e o problema de pesquisa. Por meio da análise da literatura publicada será

possível traçar um quadro teórico e fará a estruturação conceitual que dará sustentação ao desenvolvimento da pesquisa. Para elaborar uma revisão de literatura é recomendável a adoção da metodologia de pesquisa bibliográfica, a qual possui como base a análise da literatura já publicada em forma de livros, artigos e relatórios técnicos (ZOBEL, 2004).

De acordo com Cardoso *et al.* (2019), os métodos de revisão de pesquisa existentes podem ser classificados em: (i) **revisão sistemática**, a qual é planejada e utiliza métodos explícitos e sistemáticos para identificar, selecionar e avaliar criticamente estudos primários (pesquisas relacionados a um problema específico); (ii) **revisão integrada**, que se diferencia da revisão sistemática por avaliar não só estudos primários (pesquisas), como revisões teóricas, relatos, e outros tipos de estudos, possuindo assim uma questão de pesquisa mais ampla do que aquela que gera uma revisão sistemática; (iii) **meta-análise**, a qual utiliza técnicas estatísticas desenvolvidas para integrar os resultados de dois ou mais estudos independentes, sobre uma mesma questão de pesquisa, consolidando os resultados de tais estudos e (iv) **pesquisa híbrida**, a qual combina métodos quantitativos e qualitativos.. Neste rol, verifica-se que a revisão sistemática é focada principalmente na pesquisa quantitativa e é utilizada como um resumo para produzir uma meta-análise de coleta das melhores evidências possíveis para desenvolver a prática baseada em evidência, ou por sua utilidade de conhecimento de pesquisa em estudos primários, sejam quantitativos ou qualitativos (CARDOSO et al., 2019).

Apesar da forma de síntese da literatura ter maior expressão atualmente, esta não é uma concepção nova. Para Sousa e Firmino (2018 p.46): “James Lind, em 1753, realizou o primeiro ensaio clínico aleatório, reconheceu o valor dos métodos sistemáticos para identificar, extrair e avaliar as informações de estudos de modo a evitar interpretações tendenciosas da investigação”. À vista disso, ao longo do percurso histórico, vê-se marcos que seguiram valorizando a acumulação de resultados para construir um balanço da produção de conhecimento até um dado momento (MANCINI, 2017).

Para Mancini (2017, p.83):

> As revisões sistemáticas são desenhadas para ser metódicas, explícitas e passíveis de reprodução. Esse tipo de estudo serve para nortear o desenvolvimento de projetos, indicando novos rumos para futuras investigações e identificando quais métodos de pesquisa foram utilizados em uma área.

Neste ínterim, vê-se que a revisão sistemática é uma proposta adequada para obtenção de provas e a síntese do conhecimento sobre o tema, uma vez que há o incentivo, a convergência e a inclusão de todos os estudos relevantes, proporcionando uma visão ampla do escopo em foco (SOUSA; FIRMINO, 2018)

A revisão sistemática é um processo desenvolvido para identificar o núcleo de uma revisão da literatura de interesse para a prática, realizando pesquisa e extração mais relevante de acordo com os critérios que foram avaliados e respeitados por outros. Logo, compreende-se que é uma investigação em si, com métodos planejados com antecedência e com uma montagem dos estudos originais considerados como seus assuntos (MUNN et al., 2018).

Já outros autores entendem que a revisão sistemática sintetiza os resultados de múltiplas investigações usando estratégias para reduzir o preconceito e erros do acaso. Essas estratégias incluem a pesquisa exaustiva de todos os artigos potencialmente relevantes e reproduzíveis nas seleções de artigos para revisão (HOLLY; SALMOND; SAIMBERT, 2017). Kitchenham et al. (2009) complementam que a revisão sistemática não apenas agrega todas as evidências existentes em uma pesquisa em questão, mas também se destina a apoiar o desenvolvimento de diretrizes com base em evidências para os profissionais.

Desta forma, as revisões sistemáticas são diferentes das revisões tradicionais de literatura porque visam identificar todos os estudos que abordam uma questão específica, e sua metodologia tem como objetivo minimizar o efeito da seleção, publicação e viés de extração de dados, fornecendo assim um resumo equilibrado e imparcial da literatura (NIGHTINGALE, 2009).

Wazlawick (2009) também evidencia a importância de proceder à pesquisa bibliográfica de maneira sistemática e sugere passos a sua realização, informando que o pesquisador pode adaptar esses passos de acordo com suas necessidades ou disponibilidade.

Desta forma, apropriando do conceito de Holly; Salmond e Saimbert (2017), Nightingale (2009) e Wazlawick (2009), compreende-se que a revisão sistemática é um processo que agrega estratégias científicas que limitam os vieses da montagem sistemática, avaliação crítica e síntese de todos os estudos relevantes sobre um tópico específico, como uma ferramenta científica que pode ser utilizada, de modo geral, para resumir, extrair e comunicar os resultados e implicações de uma série de investigações que não poderiam ser administradas de outra forma. Em realização ao processo supracitado, o rigor científico é evidente, tal que a revisão sistemática é considerada

uma investigação por seus próprios méritos.

Para Munn et al. (2018), os objetivos da revisão sistemática estão no benefício do resumo das evidências encontrado em um tema, uma vez que este fornece um mínimo erro e preconceito, que às vezes pode interferir com uma revisão ou seleção de literatura apropriada. Logo, ao se ter um sistema claramente definido, com um conjunto padronizado, com critérios de elegibilidade predefinido para estudos face a uma metodologia explícita e reproduzível; a busca sistemática identificará todos os estudos que se enquadrariam no critério de eleição, com uma avaliação da validade dos resultados dos estudos incluídos e uma apresentação sistemática com a síntese das características e resultados dos estudos incluídos.

Alguns autores argumentam que também existe a metodologia denominada **mapeamento sistemático**, a qual possui o objetivo de fornecer uma visão geral de uma área de pesquisa por meio de classificação e quantificação de contribuições em relação às categorias dessa classificação. Esta metodologia envolve pesquisar a literatura para saber quais tópicos foram abordados e onde a literatura foi publicada. O mapeamento sistemático e a revisão sistemática da literatura compartilham alguns pontos em comum (por exemplo, no que diz respeito à organização sistemática da pesquisa e da seleção de estudos). No entanto estes estudos diferem em termos de objetivos e, portanto, abordagens para a análise de dados, uma vez que as revisões sistemáticas visam sintetizar evidências, também considerando a força da evidência, enquanto os mapas sistemáticos estão preocupados em estruturar uma área de pesquisa (PETERSEN; VAKKALANKA; KUZNIARZ, 2015).

Kitchenham, Budgen e Brereton (2010) avaliaram as diferentes características entre as revisões sistemáticas da literatura e estudos de mapeamento sistemático. Existem diferenças no que diz respeito às questões de pesquisa, processos de pesquisa, requisitos de estratégia de pesquisa, avaliação de qualidade e resultados.

As questões de pesquisa em estudos de mapeamento sistemático são gerais, pois visam descobrir tendências de pesquisa (por exemplo, tendências de publicação ao longo do tempo, tópicos abordados na literatura). Por outro lado, as revisões sistemáticas visam agregar evidências e, portanto, um objetivo mais específico de pesquisa precisa ser formulado (KITCHENHAM; BUDGEN; BRERETON, 2010).

Dada essa diferença principal, o processo de pesquisa é impactado, pois a busca por estudos no mapeamento sistemático possui como base uma área de pesquisa, enquanto as revisões sistemáticas da literatura são orientadas por questões de pesquisa específicas (KITCHENHAM; BUDGEN; BRERETON, 2010).

Kitchenham, Budgen e Brereton (2010) também argumentam que os requisitos de pesquisa são menos rigorosos para estudos de mapeamento, uma vez que o seu interesse é limitado às tendências de pesquisa, enquanto todos os estudos devem ser encontrados em revisões sistemáticas. Porém, alguns autores discutem que encontrar todos os estudos muitas vezes não é realista, nem para revisões sistemáticas, nem para estudos de mapeamento; logo, obter uma boa amostra no que diz respeito às características das publicações seria suficiente (WOHLIN et al., 2013).

Com relação à avaliação da qualidade, ela é essencial em revisões sistemáticas para determinar o rigor e a relevância dos estudos primários. Já no mapeamento sistemático nenhuma avaliação de qualidade precisa ser realizada. Artigos que não possuem evidência empírica não devem ser incluídos em uma revisão sistemática, embora nos mapeamentos sistemáticos eles sejam importantes para detectar tendências na área que está sendo investigada (PETERSEN; VAKKALANKA; KUZNIARZ, 2015).

Tratando-se de resultados, a revisão sistemática apresenta uma síntese de evidências encontradas, enquanto o resultado de um estudo de mapeamento é um inventário de documentos sobre área de pesquisa, mapeados a partir de uma classificação. Portanto, um mapeamento sistemático fornece uma visão de uma área de pesquisa, permitindo descobrir lacunas e tendências de pesquisa (PETERSEN; VAKKALANKA; KUZNIARZ, 2015).

2.2 – Ausência de Dados

O problema da ausência de dados pode se fazer presente em diversas etapas, desde a origem do dado, passando pela sua transformação até a integração dos dados anterior à sua análise. Neste contexto, a ocorrência de dados faltantes cresce proporcionalmente ao volume de dados gerados, tornando-se um problema cada vez mais relevante para analistas e pesquisadores, assim como um desafio adicional ao processo de análise de dados para os meios acadêmico e empresarial.

Um motivo que justifica a necessidade de investigar as técnicas de imputação de dados é o fato de a maioria dos algoritmos utilizados para as análises adotar as matrizes de dados completas, ou seja, sem a ocorrência de valores ausentes. Sendo

assim, a execução dos algoritmos de análise exige uma base de dados completa. Desta forma, teorias para a análise de dados ausentes com suporte em verossimilhança foram apresentadas e esses métodos sistematizados, fornecendo uma base teórica e experimental fundamental para futuras investigações sobre técnicas de imputação (RUBIN, 1976).

A ausência de respostas constitui uma das principais limitações de qualquer estudo (ALLISON, 2001). Do ponto de vista conceitual, a ausência de dados abarca dois aspectos distintos: (i) a não participação de um sujeito no estudo por não responder ao questionário; e (ii) os valores faltantes (dados faltantes) das pessoas que responderam o questionário de forma incompleta por não responder a uma ou mais variáveis.

Para Rubin (1976) os dados ausentes podem levar a uma perda considerável do tamanho da amostra ao analisar com técnicas estatísticas, como a multivariada, uma vez que mesmo quando um sujeito tem apenas um dado ausente em uma das variáveis, sua exclusão da análise elimina todos os outros valores a ele associados.

2.2.1 – Possíveis causas de ausência de dados

Compreende-se que são várias as causas que provam a existência de dados desconhecidos em um conjunto de dados, sendo que algumas mais habituais implicam na falta de tempo ou no engajamento e envolvimento dos entrevistados, carga dos dados com custo elevado, ausência de um treinamento apropriado na coleta dos dados, dentre outros (CARTWRIGHT; SHEPPERD; SONG, 2004). Podem constar ausência de alguns dados simplesmente por não fazerem nenhum sentido. Nesta perspectiva, Brown e Kros (2003) elencam outras causas para esta situação, conforme são expostas, a seguir: (i) Fatores operacionais; (ii) Recusa na resposta em pesquisas; e (iii) Respostas não aplicáveis.

Para os fatores operacionais, erros na entrada de dados, bem como estimativas erradas, são as causas mais frequentes. Já para a questão da recusa na resposta em pesquisas acontece quando um entrevistado deixa uma ou mais questões sem resposta, seja por motivos de ordem emocional, ideológica, religiosa, entre outros. Para as respostas não aplicáveis, estas são questões demonstradas que não encontram de alguma forma aderência aos entrevistados.

Verifica-se que os analistas de dados, de modo geral, não se familiarizam com

as particularidades da carga de dados, tais como se vê em pesquisas. Logo, criam-se vantagens frente a um tratamento de dados com mais cuidado para os valores ausentes (CHIU; SEDRANSK, 1986). Isto posto, Ford (1983) indica algumas das causas que justificam o motivo de a responsabilidade do ajuste de valores ausentes ser de quem coleta os dados.

Portanto, em primeiro lugar, deve-se possuir compreensão sobre o universo dos dados; em segundo, quem coleta os dados deve, de modo geral, fazer estimativas sobre o conjunto de dados; em terceiro, grande parte dos analistas de dados não querem a responsabilidade do ajuste dos valores ausentes; e, em quarto, considera-se que em um grupo de dados sem valores ausentes há a possibilidade de que todas as análises futuras tenham um ponto em comum sem que cada analista de dados coloque o seu próprio valor inicial (FORD, 1983).

Há duas razões para existirem dados ausentes, conforme relatam alguns autores, sendo que a primeira é denominada de dado ausente unitário (*unit nonresponse*) (RAGHUNATHAN, 2004; SCHAFER; GRAHAM, 2002; TWALA; CARTWRIGHT; SHEPPERD, 2005). Neste caso, alguns dos eventos completos dos dados podem não ter sido coletados da amostra original, bem como alguns entrevistados podem ter se recusado a responder a todos os questionamentos. A segunda razão para existirem os dados ausentes seria a dos itens de dados ausentes (*item nonresponse*), neste enquadramento, alguns dos dados coletados estão preenchidos e outros não.

Schafer e Graham (2002) e Twala; Cartwright e Shepperd (2005) adicionam uma terceira razão, na qual ocorrem levas de dados ausentes (*"wave" nonresponse*). Neste cenário, não se encontram respostas para um assunto de uma seção em uma ou mais levas, ou quando um entrevistado deixa de responder uma seção para fazê-lo posteriormente, e não retorna.

2.2.2 – Tipos de Ausência

Segundo Graham (2012), existem dois tipos de ausências de dados descritos na literatura: itens de dados ausentes ("*item nonresponse*") e ondas de dados ausentes ("*wave nonresponse*").

Itens de dados ausentes decorrem quando, em uma pesquisa de opinião, uma

parte das questões foi respondida; no entanto, ocorreu a presença de questões em branco. Diversos motivos podem fazer que o entrevistado não responda uma questão, tais como: (i) a não percepção de alguma questão em uma pesquisa por parte do entrevistado; (ii) o fato de o entrevistado não saber a resposta de alguma questão; ou (iii) o entrevistado decidir não responder a alguma questão por um motivo pessoal. Além disso, existem casos de perda de respostas ligadas às questões técnicas, tais como a falha na coleta ou no processamento dos dados, os quais também se enquadram no conceito de itens de dados ausentes (GRAHAM; CUMSILLE; SHEVOCK, 2012)

Já no caso de ocorrência de ondas de dados ausentes, a pesquisa de opinião é realizada através de ondas de perguntas, e ausência de respostas ocorre quando uma ou mais ondas de perguntas não são respondidas pelo entrevistado. Em alguns casos, pode ocorrer de o entrevistado retornar posteriormente para responder as questões ainda em aberto. Porém, em outros casos, o mesmo nunca retornará, o que pode ser denominado como atrito (GRAHAM, 2012).

Além do tipo, as ausências possuem outras características que serão descritas nas duas próximas seções: os padrões de ausência e os mecanismos de ausência.

2.2.3– Padrões de Ausência

Quando é realizada a análise de um conjunto de dados em busca de algum conhecimento não facilmente perceptível, esses dados são observados através de uma tabela. Neste sentido, alguns padrões de ausência de dados podem ser observados. Schafer e Graham (2002) definem que o reconhecimento desses padrões de ausência são um passo importante para que se possa escolher a melhor técnica de preenchimento desses valores e definem as seguintes categorias para estes padrões: (i) geral ou aleatório; e (ii) específico.

A ausência de dados de padrão geral ou aleatória, como a própria classificação sugere, está dispersa em qualquer registro do conjunto de dados. Com relação aos padrões de ausência específicos, estes são divididos em dois subtipos: (i) univariado; e (ii) monotônicos (SCHAFER; GRAHAM, 2002).

Nos padrões específicos classificados como univariados, observa-se uma distribuição de ausência dirigida a uma única variável, quando ocorre em um item Y mas o conjunto dos p outros itens $X_1 ,...,X_p$ continuam completos. Este padrão também deve

incluir situações onde *Y* representa um grupo de itens completamente observados ou completamente ausentes e em sua essência estão restritos a uma única variável do conjunto de dados. Já nos padrões monotônicos, a ausência é observada em mais de uma variável de tal forma que um grupo de itens $Y_1,...,Y_p$ quando ordenados, se Y_j contém valores ausentes então $Y_j+1,...,_p$ também conterão (SCHAFER; GRAHAM, 2002). Estes conceitos são apresentados na Figura 1.

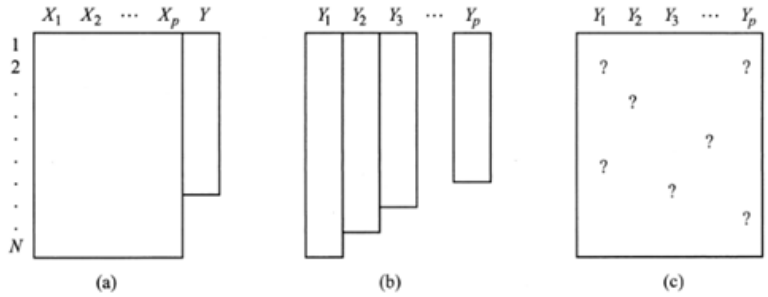

Figura 1 - Padrões de ausência de respostas em um conjunto de dados retangulares: (a) padrões univariados, (b) padrões monotônicos; (c) padrões gerais ou aleatórios. Fonte: (SCHAFER; GRAHAM, 2002).

2.2.4– Mecanismos de Ausência

Além de considerar a observação dos padrões de ausência existente no conjunto de dados, outra característica importante é o mecanismo de ausência dos dados, ou seja, o que causou a inexistência.

Schafer e Graham (2002) consideram o mecanismo de ausência de dados existente no conjunto de dados utilizado na avaliação experimental, argumentando que, conforme a literatura estatística, é sugerido que um mecanismo de ausência pode ser um processo pelo qual alguns dados são gerados e outros não. Segundo Little e Rubin (2002), o mecanismo de ausência possui relação com processo que ocasiona a não geração das ocorrências, e destacam a importância de conhecer a relação entre os dados ausentes e os valores subjacentes das variáveis do conjunto de dados .

O trabalho de Little e Rubin (2002) define os mecanismos de ausência da seguinte forma: sendo *Y* um conjunto de dados completo e *M* o conjunto de ausências, o mecanismo de ausência é caracterizado pela distribuição condicional de *M* dado *Y*, $f(M|Y,\phi)$ onde, ϕ significa parâmetros desconhecidos.

Rubin (1976) define três tipos de classificações para mecanismos de ausência:

completamente aleatório (*MCAR – Missing Completely At Random*), aleatório (*MAR – Missing At Random*) e não aleatório (*NMAR – Not Missing At Random*, ou *IM – Ignorable Missing*).

O mecanismo de ausência de dados é classificado como completamente aleatório quando o seu real motivo é desconhecido, não havendo qualquer relação com outro(s) atributo(s) do conjunto de dados, ou seja, quando *M* não depende dos valores de *Y*. Sendo assim, se $f(M|Y, \phi) = f(M| \phi)$ *para todo* Y, ϕ, o mecanismo de ausência de dados é classificado como completamente aleatório.

Se os valores ausentes dependem de algum(s) atributo(s) do conjunto de dados, este é classificado como aleatório. Seja Y_{obs} um subconjunto observado completo de *Y*, e Y_{mis} um subconjunto observado de dados ausentes, para que o mecanismo de ausência seja aleatório, Y_{mis} deve ser dependente de Y_{obs}, ou seja, $f(M|Y, \phi) = f(M|Y_{obs}\}, \phi)$ *para todo* Y_{mis}, ϕ.

O terceiro mecanismo de ausência definido por Rubin (1976) é chamado de não aleatório. Neste mecanismo, a distribuição de ausência *M* depende dos valores ausentes representados por Y_{mis}. Este mecanismo de ausência pode ser definido como inacessível, uma vez que o mesmo não pode ser mensurado e, portanto, não está disponível para avaliação (GRAHAM; DONALDSON, 1993).

Outros autores acrescentam que o mecanismo de ausência MNAR pode ter duas origens, as quais são classficadas da seguintes forma: (i) ausência dependendo das variáveis não observadas (MuOV), e: (ii) ausência dependendo do seu próprio valor (MIV), que pode acontecer quando uma variável possui um valor fora de sua faixa de representação (GARCIARENA; SANTANA, 2017).

2.2.5 – Soluções para o tratamento de dados ausentes

Vê-se nas referências literárias uma ampla gama de propostas de soluções para o tratamento de dados ausentes. No entanto, não existe consonância entre os autores, especialmente sobre a classificação das inúmeras técnicas de dados ausentes existentes. Os trabalhos sobre essa temática sugerem, de modo geral, uma classificação singular, com um nível de generalizado maior ou menor.

Uma classificação para os métodos de tratamento de dados ausentes é sugerida por alguns pesquisadores, sendo estes divididos em quatro grupos (BATISTA;

MONARD, 2003). Para o primeiro grupo, verifica-se a questão de ignorar e descartar dados (*Complete Case Analysis*) (BETHLEHEM, 2009). Para o segundo grupo, tem-se de descartar tabelas com propriedades de um grau elevado de dados ausentes. No terceiro grupo, leva-se em consideração fazer a estimativa de parâmetros. No quarto grupo faz-se a imputação dos dados ausentes.

Para Twala; Cartwright e Shepperd (2005), há três abordagens com relação aos dados ausentes, sendo estes: (i) Análise de dados completos; (ii) Imputação; (iii) Procedimentos fundamentados em modelos.

Outros autores argumentam que os modos de tratamento de dados ausentes podem ser baseados em cinco abordagens: (i) Ignorar as tabelas com valores ausentes; (ii) Preenchê-los manualmente; (iii) Fazer a substituição do valor ausente por uma constante ; (iv) Usar média ou moda; (v) Atribuir o valor mais provável (HRUSCHKA; HRUSCHKA; EBECKEN, 2003; HRUSCHKA; JR; EBECKEN, 2003).

Magnani e Montesi (2004) propõem uma taxonomia mais refinada, em quatro partes, ao considerar em sua primeira etapa os métodos convencionais, compreende-se que neste processo se trabalha com a remoção completa de casos e a remoção em pares. Na segunda parte tem-se a imputação. Nesta etapa faz-se a imputação global baseada no atributo ausente, não ausentes e imputação local. Na terceira etapa faz-se a estimativa de parâmetros e, na quarta parte, trabalha-se com o gerenciamento direto dos dados ausentes.

Verifica-se que de todas as propostas, Magnani e Montesi (2004) é quem faz o desdobramento dos atuais métodos de imputação, bem como estrutura as diferenças entre os métodos de complementação de dados ausentes. Todavia, é válido ressaltar que sua organização limita os métodos estatísticos, sintetizando-os somente ao de estimativas de parâmetros.

Alguns autores argumentam que os métodos de imputação de dados ausentes podem ser divididos em duas categorias principais: os métodos de imputação de dados ausentes globais e os métodos de imputação de dados ausentes locais (CHENG; LAW; SIU, 2012; FENG et al., 2015). A imputação global de dados ausentes inclui as estratégias que utilizam toda a estrutura de correlação global do conjunto de dados para imputar valores ausentes. Vários métodos de imputação atuais, como imputação iterativa (LITTLE; RUBIN, 2002; PEDREGOSA; VAROQUAUX; GRAMFORT, 2011; VAN BUUREN; GROOTHUIS-OUDSHOORN, [s.d.]) e a imputação utilizando o algoritmo Expectation Maximization (EM) (JUNNINEN et al., 2004; SCHNEIDER, 2001) estão incluídos nesta categoria (RAHMAN; ISLAM, 2013). Já a imputação local de dados

ausentes inclui as estratégias que usam apenas os registros semelhantes ao registro ausente para imputar valores ausentes. Os métodos de imputação, como imputação *hot-deck*, imputação com base em k-vizinhos mais próximo (KNNI) (BATISTA; MONARD, 2003), imputação com base em mínimos quadrados locais (LLSI) (KIM; GOLUB; PARK, 2005), LLSI iterativo (ILLSI) (CAI; HEYDARI; LIN, 2006) e o LLSI baseado em bicluster iterativo (IBLLS) (CHENG; LAW; SIU, 2012) são considerados nesta categoria.

A partir da taxonomia de Magnani e Montesi (2004), o autor Soares (2007) criou o termo “imputação híbrida”, apresentando uma taxonomia abrangente, sendo esta dividida em cinco partes:

(i) Métodos Convencionais: o problema é solucionado pela exclusão das observações/atributos com valores ausentes. Esta remoção pode ser realizada de três formas distintas, a saber: remoção completa de casos; remoção em pares (*Pairwise Deletion*) e ; remoção de colunas com valores ausentes;

(ii) Imputação: os métodos incluídos nesta categoria solucionam o problema estimando os valores ausentes a partir dos valores atuais existentes na base. A estimação dos valores pode ser realizada de acordo com um dos três paradigmas a seguir: imputação global baseada no atributo ausente; imputação global baseada nos atributos não ausentes e; imputação local;

(iii) Modelagem de dados: inclui todos os métodos que buscam criar um modelo para expressar de forma genérica as características dos dados e, a partir destes modelos, é realizada a estimação do valor ausente. A modelagem de dados é subdividida em: métodos de verossimilhança e modelos bayesianos;

(iv) Gerenciamento direto dos dados ausentes: nesta categoria estão todos os métodos que se apresentam robustos em relação aos valores ausentes, dispensando a imputação; e

(v) Métodos Híbridos: reúne os métodos que realizam a combinação de soluções e/ou o sequenciamento de tarefas com o objetivo de estimar o valor ausentes. Os métodos híbridos são agrupados em duas categorias: imputação múltipla e; imputação composta.

2.2.5.1 – Remoção de Casos

Uma vez investigados os mecanismos de ausência, algumas ações podem ser feitas para tratar a ausência dos dados, tal como a sua substituição por valores reais (RUBIN, 1988). No entanto, a maioria dos softwares estatísticos possui limitados recursos para o tratamento de dados ausentes, que consistem em: (i) remoção completa de casos (*Listwise Deletion*) ou (*Complete-Case Deletion*); (ii) remoção de pares *(Pairwise Deletion)*; e (iii) remoção de colunas com valores ausentes. A Figura 2 apresenta exemplos de remoção completa de casos, remoção de pares e remoção de colunas.

Base de Dados

ID	Genêro	Mão-de-obra	Vendas
1	M	23	343
2	F		280
3	M	35	332
4	F		272
5	F	20	300
6	M	26	326
7	M	30	259
8	M	33	297
N=8			

Exclusão completa de casos

	Genêro	Mão-de-obra	Vendas
1	M	23	343
3	M	35	332
5	F	20	300
6	M	26	326
7	M	30	259
8	M	33	297
N=6			

Base de Dados

ID	Genêro	Mão-de-obra	Vendas
1	M	23	343
2	F		280
3	M	35	332
4	F		272
5	F	20	300
6	M	26	326
7	M	30	259
8	M	33	297
N=8			

Remoção de pares

ID	Genêro	Mão-de-obra	Vendas
1	M	23	343
2	F	ignorado	280
3	M	35	332
4	F	ignorado	272
5	F	20	300
6	M	26	326
7	M	30	259
8	M	33	297
N=8	N=8	N=6	N=8

Base de Dados

ID	Genêro	Mão-de-obra	Vendas
1	M	23	343
2	F		280
3	M	35	332
4	F		272
5	F	20	300
6	M	26	326
7	M	30	259
8	M	33	297
N=8			

Remoção de colunas

ID	Genêro	Vendas
1	M	343
2	F	280
3	M	332
4	F	272
5	F	300
6	M	326
7	M	259
8	M	297
N=8		

Figura 2 - Exemplos de remoção completa de casos, remoção de pares e

remoção de colunas. Fonte: Elaborado pelo autor.

Nota: No método de exclusão completa de casos, as linhas 2 e 4 foram excluídas. No método de remoção de pares, observa-se que N (total de observações) varia entre as variáveis devido ao valor ausente ser ignorado. No método de remoção de colunas, a variável “mão-obra” foi excluída.

A remoção completa de casos consiste em uma simples op. Todo registro de um grupo de valores que possuir algum de seus aspectos é removido da amostra. Contudo, mesmo que se entenda que tal técnica possui certa facilidade, esta apresenta algumas barreiras, entre as quais: (i) A remoção pode descartar grande parte dos dados, tornando-os tendenciosos; (ii) A remoção também causa a poda de algumas regras; e (iii) O mecanismo de ausência dos dados deve ser completamente aleatório (MCAR).

A remoção em pares é uma variação da remoção completa de casos, pois considera na análise da variável todos os valores da coluna ignorando os dados ausentes. O benefício desta técnica é o de usar todos os dados disponibilizados na base de dados. As barreiras, de modo geral, estão em sua implementação, que é mais complexa do que a remoção completa.

Ferlin (2008) destaca que a realização da remoção completa de casos pode ocasionar algumas dificuldades na análise dos dados, tais como: (i) tornar a amostra muito pequena, perdendo a precisão; (ii) descartar grande parte dos dados, tornando-os tendenciosos, e para tarefas de classificação ou sumarização, também pode causar a poda de algumas regras. Semelhantemente, a remoção de pares também apresenta alguns riscos, como por exemplo: (i) a análise comparativa dentro de um estudo pode ser problemática, pois diferentes subconjuntos de casos são usados; e (ii) nem sempre a matriz de covariância pode ser definida, isto é, certos elementos podem assumir valores impossíveis dados outros elementos.

Já a remoção de colunas com valores ausentes, na situação onde ocorre valor ausente para o atributo em algum caso da base de dados, este atributo é removido para todos os casos. Este processo, em geral, resulta em uma significativa perda de informação, inviabilizando a sua aplicação. Além disso, é possível constatar que este método pode modificar a relação existente entre as colunas na base original e criar viés (FERLIN, 2008).

2.2.5.2 – Gerenciamento direto de casos

Existem métodos de classificação que são robustos em relação aos valores ausentes e dispensam o processo de imputação. O *Top-Down Induction of Decision Tree* (TDIDT) é um algoritmo robusto bem conhecido e utilizado como base para muitos algoritmos de indução de árvores de decisão, tais como: ID3 (QUINLAN, 1986), C4.5 (SALZBERG, 1994) e CART (BREIMAN et al., 1984).

O TDIDT produz regras de decisão de forma implícita numa árvore de decisão, a qual é construída por sucessivas divisões dos exemplos de acordo com os valores de seus atributos preditivos, um processo conhecido como particionamento recursivo (BRAMER, 2007).

A estrutura do algoritmo TDIDT tem como base três possibilidades sobre um conjunto de treinamento *T* contendo classes C_1, C_2, ..., C_k. Na primeira possibilidade *T* contém um ou mais objetos, sendo todos pertencentes à classe C_j. Assim, a árvore de decisão para *T* formada por um nó folha que identifica a classe C_j.

Já na segunda possibilidade o conjunto de treinamento *T* não possui objetos. A árvore de decisão também é constituída de um nó folha, mas a classe associada deve ser determinada através informação externa, tal como um conhecimento do domínio do problema.

Na terceira possibilidade, o conjunto de treinamento T contém exemplos pertencentes a mais de uma classe. Neste caso, *T* é dividido em *n* em subconjuntos que buscam ser coleções de exemplos com classes únicas. Para que ocorra esta divisão, é selecionado um atributo preditivo *A*, o qual possui um ou mais possíveis resultados O_1, O_2, ..., O_n. Em seguida, o conjunto de treinamento *T* é particionado em subconjuntos T_1, T_2, ..., T_n, onde T_i contém todos os exemplos de *T* que possuem resultado O_i para o atributo *A*. A árvore de decisão para *T* consiste de um nó de decisão identificando o teste sobre o atributo *A*, e uma aresta para cada possível resultado, ou seja, *n* arestas.

Desta forma, o mesmo algoritmo de indução de árvores de decisão considerando as três possibilidades é aplicado recursivamente para cada subconjunto de exemplos T*i*, com *i* variando de 1 até *n*. Assim, o algoritmo TDIDT realiza uma busca gulosa sobre um conjunto de atributos, tentando identificar os atributos que dividem o conjunto de exemplos em subconjuntos com classes únicas.

O algoritmo C4.5 pode ser aplicado em atributos categóricos (ordinais ou não-ordinais) como com atributos contínuos. Para lidar com atributos contínuos, o algoritmo

C4.5 define um limiar e então divide os exemplos de forma binária: aqueles cujo valor do atributo é maior que o limiar e aqueles cujo valor do atributo é menor ou igual ao limiar.

Dentre os algoritmos ID3, C4.5. e CART, a literatura evidencia que o algoritmo C4.5 possui um recurso que permite os valores para um determinado atributo sejam representados como '?'. Este recurso faz este algoritmo ser robusto para valores ausentes, uma vez que valores com a identificação '?' são tratados de forma especial e não são utilizados nos cálculos de ganho e entropia.

O algoritmo C4.5 foi aperfeiçoado para a versão chamada C.5.0 (para os sistemas operacionais Unix ou Linux) e uma versão comercial com o nome See5 para o sistema operacional Windows. O novo algoritmo C 5.0 apresentou as seguintes melhorias quando comparado com o seu antecessor: (i) C 5.0 é significativamente mais rápido; (ii) C 5.0 é mais eficiente em uso de memória; (iii) C 5.0 obtém resultados semelhantes ao C4.5 com árvores de decisão consideravelmente menores; (iv) o algoritmo C 5.0 passou a ter suporte para *boosting*, melhorando as árvores de decisão e aumentando a precisão; (v) C 5.0 permite ponderar diferentes casos e tipos de classificação incorreta; e (vi) o algoritmo C 5.0 possui opção que utiliza um algoritmo chamado *winnow* (uma técnica de aprendizado de máquina) que permite a limpeza automática de atributos e remove aqueles que podem não ser úteis (BUJLOW; RIAZ; PEDERSEN, 2015)

2.2.5.3 – Imputação de dados

2.2.5.3.1 – Introdução

A imputação de dados é formalmente definida como *"um campo de estudo que busca complementar dados ausentes com base em métodos estatísticos e de inteligência computacional" (GELMAN; HILL, 2006)*. Realizar uma imputação de dados consiste em substituir um valor ausente ou rejeitado por um valor "estimado" que seja viável, porém é um valor artificial.

Segundo Buuren (2012), a imputação pode ser influenciada por valores de outra tupla, para o mesmo atributo e para atributos diferentes, conforme exemplificado na

Figura 3.

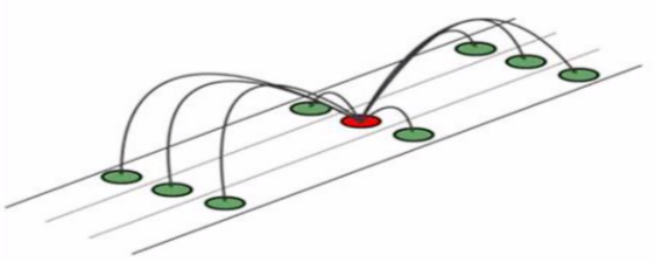

Figura 3 - Exemplo de imputação. Fonte:(BUUREN, 2012).

Já segundo os autores Little e Rubin (2002), imputações de dados podem ser definidas como médias ou extrações de uma distribuição preditiva dos valores omissos, as quais requerem uma metodologia de criação de uma distribuição preditiva para a imputação com base nos dados observados. Estes autores também definem duas abordagens genéricas para geração dessa distribuição: a modelagem explícita e a modelagem implícita.

Na modelagem explícita a distribuição preditiva possui como base um modelo estatístico formal (por exemplo normal multivariada), e, portanto, as suposições são explícitas. São exemplos de modelagem explícita: imputação por média, imputação por regressão linear, imputação por regressão estocástica (LITTLE; RUBIN, 2002).

Ao passo que o algoritmo é o foco na modelagem implícita, o que implica em um modelo subjacente. Ou seja, as suposições estão implícitas, mas ainda precisam ser cuidadosamente avaliadas para garantir que sejam razoáveis. São exemplos de modelagem implícita: imputação *hot-deck*, imputação por substituição e imputação *cold-deck* (LITTLE; RUBIN, 2002).

Quando lidamos com a imputação dos dados, Soares (2007) em sua taxonomia, classifica o processo de imputação em dois métodos: método simples e método híbrido. No método simples, apenas um valor proposto é considerado para cada valor ausente. Já no método híbrido, dois ou mais métodos de imputação simples são considerados no processo de imputação como um todo visando a melhoria do valor estimado (SOARES, 2007a).

Um exemplo do método de imputação híbrida e presente na ideia central deste trabalho de pesquisa é a imputação múltipla (RAGHUNATHAN, 2004). Neste método o conjunto de dados ausentes é substituído por mais de um conjunto de dados plausível e cada um destes, junto com o restante dos dados formam um conjunto de dados completo, que por sua vez, é analisado em separado por meio de estimativas

estatísticas (RAGHUNATHAN, 2004).

Soares (2007) propõe uma solução de imputação composta que permite avaliar o impacto de tarefas de seleção de atributos e agrupamento de dados precedendo a imputação de dados. Ferlin (2008) expande a solução proposta por Soares (2007) adotando a imputação em cascata.

2.2.5.3.2 – Imputação global

No processo de imputação global, também chamada de imputação simples ou única, apenas um único valor possível é gerado para cada campo ausente na base de dados. As técnicas de imputação global têm sido uma das ferramentas mais conhecidas e aceitas para tratar a não resposta (LITTLE; RUBIN, 2002). Técnicas de imputação global apresentam algumas vantagens sobre as técnicas de imputação múltipla. Por exemplo, técnicas de imputação global têm uma implementação mais fácil sem, ao contrário, sofrer uma perda significativa de eficiência em comparação com técnicas de imputação múltipla. Desta forma, destaca-se que as técnicas de imputação global podem ser divididas em duas categorias: aleatório e determinístico.

O uso de imputação pode causar sérios problemas de subestimação da verdadeira variação quando a proporção de dados perdidos é apreciável. Em geral, um método de imputação aleatória tem a vantagem de adicionar maior variabilidade através das imputações de que um método determinístico de imputação; isto é, as técnicas de imputação determinística global geralmente subestimam mais as variâncias do que técnicas de imputação aleatória global. No entanto, as técnicas determinísticas fornecem, em geral, estatísticas mais precisas do que aleatória.

Existem várias formas de realizar uma imputação global, onde com frequência se utilizam métodos estatísticos, tais como: média, moda, regressão simples, regressão logística, entre outros. No entanto, alguns autores ressaltam que o processo de imputação global oculta a incerteza do dado, fazendo com que os intervalos de confiança seja inválidos, uma vez que os valores estimados possuem como origem os dados existentes (CARTWRIGHT; SHEPPERD; SONG, 2004). Além disso, a imputação única estima valores com uma variância pequena, o que pode criar resultados com viés (CARTWRIGHT; SHEPPERD; SONG, 2004).

2.2.5.3.3 – Imputação global baseada no atributo ausente

A imputação global baseada no atributo com valores ausentes utiliza todos os valores existentes nas demais tuplas para preencher os valores ausentes e podem ser classificados de duas formas: (i) determinístico: a substituição dos valores ausentes é realizada a partir de valores centrais da distribuição estatística do atributo; e (ii) estocástico: realiza a introdução de uma perturbação na média com o objetivo de reduzir os efeitos de viés da média (MAGNANI; MONTESI, 2004).

A imputação baseada no atributo com valores ausentes do tipo determinístico é uma das técnicas mais utilizadas devido à sua simplicidade. No caso de atributos ausentes contínuos, estes são substituídos pela média dos valos do atributo; já para os atributos categóricos é utilizada a moda, que identifica o valor do atributo que possui maior frequência de ocorrência (BUSSAB; MORETTIN, 2010).

No entanto, apesar de sua simplicidade, esta estratégia apresenta duas desvantagens: (i) caso a base não esteja balanceada ou apresente uma grande quantidade de atributos com a ocorrência de valores extremos, a média sofre influência destes valores e sofrer viés, podendo não representar a realidade da base de dados; (ii) esta estratégia reduz a diversidade da amostra e, consequentemente, o seu desvio-padrão (SOARES, 2007a).

Já na imputação de baseada no atributo ausente estocástica, uma amostra é selecionada, na qual é introduzida uma perturbação. Esta perturbação pode adicionar ou remover um valor Δm da média, tentando reduzir sua distorção e melhorar o resultado do valor imputado (MAGNANI, 2004).

2.2.5.3.4 – Imputação global baseada nos demais atributos

A imputação global baseada nos demais atributos realiza a estimação dos valores ausentes a partir da possível relação existente entre os atributos da amostra. Um exemplo para esta categoria são as técnicas de regressão.

As regressões lineares simples e múltipla são as mais utilizadas para encontrarmos um modelo, geralmente uma função matemática (BUSSAB; MORETTIN, 2010). O modelo teórico ou populacional da regressão linear simples é representado

pela equação (1), a seguir:

$$Y = \alpha + \beta x + \mathcal{E}_i \quad (1)$$

Nesta equação, Y representa a variável dependente, ou seja, uma grandeza cujo valor depende de como a variável independente é manipulada, enquanto X representa a variável independente, que corresponde a uma grandeza que está sendo manipulada em um experimento. Já o β representa os parâmetros desconhecidos, assim como $\mathcal{E}_i$ representa o erro aleatório – definido como a diferença entre y e ŷ. Do fato de trabalharmos com dados amostrais, – representados pelos pares ordenados (x, y), os parâmetros α e β da reta teórica precisam ser estimados com base nesses pontos fornecidos pela amostra. Dessa forma, obtemos uma reta estimada da forma $\hat{y}_i = a + bx_i$, onde a letra *a* é estimativa do parâmetro α, e a letra *b* é a estimativa do parâmetro β. Já $\hat{y}_i$ representa o valor estimado da variável dependente e x_i o valor estimado da variável independente.

Embora seja possível traçar a reta no diagrama de dispersão, com a ajuda de uma régua tentando passar por entre os pontos plotados no gráfico, esta é uma forma subjetiva e sem suporte científico. No entanto, o método dos mínimos quadrados explica que a reta de regressão não é a que passa pelo maior número de pontos amostrais, mas a que melhor se ajusta aos dados (BUSSAB; MORETTIN, 2010).

O método dos mínimos quadrados se justifica pela minimização dos quadrados dos erros. A incerteza dos resultados fica evidente quando utilizamos os erros na equação, diferenciando y de ŷ. Os valores do coeficiente angular *b* e pelo intercepto vertical *a* que determinam a reta de regressão podem ser determinados pela equação 2 (BUSSAB; MORETTIN, 2010).

$$b = \frac{n \cdot \sum x_i \cdot y_i - (\sum x_i) \cdot (\sum y_i)}{n \cdot (\sum x_i^2) - (\sum x_i)^2} \qquad a = \bar{y} - b \cdot \bar{x} \quad (2)$$

sendo que:

$\bar{x}$ é a média dos valores de x_i, ou seja $x_i = \frac{\sum x_i}{n}$, e;

$\bar{y}$ é a média dos valores de y_i, ou seja $y_i = \frac{\sum y_i}{n}$.

A regressão múltipla possui princípios análogos à da regressão simples. No entanto, ao invés de somente uma variável independente, teremos várias independentes tentando explicar uma variável dependente. Desta forma, o modelo teórico ou populacional da regressão linear múltipla é representado pela equação (3) (BUSSAB; MORETTIN, 2010) a seguir:

$$Y = \alpha X + \beta_{x1} + \beta_{x2} + \beta_{x3} \ldots + \mathcal{E}_i \quad (3)$$

A preocupação geral do analista ao interpretar uma regressão linear múltipla é o resultado R^2 que indica a variabilidade total do modelo de regressão, e o p-valor (também chamado de significância) das variáveis independentes, que indicam se elas são significativas para explicar a variável dependente dentro do modelo estabelecido.

Tanto a aplicação do modelo de regressão linear simples como o modelos de regressão linear múltipla, pressupõem a verificação de alguns pressupostos descritos a seguir (WOOLDRIDGE, 2016): (i) os erros $\mathcal{E}_i$ são variáveis aleatórias de média zero e seguem uma distribuição normal; (ii) os erros $\mathcal{E}_i$ são variáveis aleatórias de variância constante (σ^2) – hipótese de homocedasticidade; (iii) as variáveis aleatórias E_1, E_2, ..., são independentes; (iv) As variáveis independentes x_1, x_2, ..., não são fortemente correlacionadas – hipótese de ausência de multicolinearidade entre as variáveis explicativas.

Soares (2007) expõe que a escolha da técnica de regressão baseada nos outros atributos apresenta alguns problemas, tais como: (i) em bases que apresentam muitos atributos ausentes, é possível existir valores ausentes nos atributos de entrada do algoritmo de regressão; e (ii) a regressão parte do princípio que o modelo escolhido é o melhor para todos os dados, mas nem sempre existe este modelo ideal ou é o mais adequando para representar o comportamento dos dados.

2.2.5.3.5 – Imputação local (*hot-deck).*

As técnicas com base em *hot-deck* são consideradas como um tipo de imputação local e preenchem um valor ausente de um atributo de um caso a partir dos valores observados para este mesmo atributo em outros casos do mesmo conjunto de dados (SCHAFER; GRAHAM, 2002).

A imputação de *hot-deck* envolve a substituição de valores ausentes de uma ou mais variáveis para um não respondente (chamado de destinatário) por valores observados de um respondente (o doador) que é semelhante ao não respondente com relação às características observadas por ambos os casos. Em alguns versões, o doador é selecionado aleatoriamente a partir de um conjunto de doadores potenciais, que chamamos de *pool* de doadores; chamamos esses métodos de métodos *hot-deck*

aleatórios. Em outras versões, um único doador é identificados e os valores são imputados a partir desse caso, geralmente o "vizinho mais próximo" com base em alguma métrica (por exemplo, um cálculo de distância); chamamos esses métodos de métodos *hot-deck* determinísticos, uma vez que não há aleatoriedade envolvido na seleção do doador. A Figura 4, a seguir, apresenta o um diagrama de fluxo da imputação *hot-deck* aleatória.

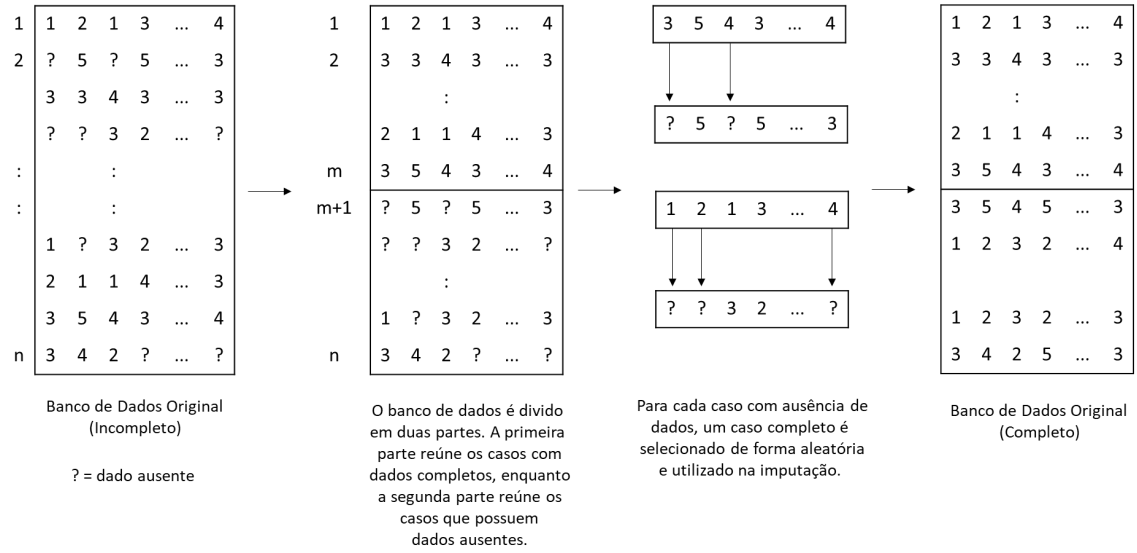

Figura 4. Diagrama de fluxo da imputação *hot-deck* aleatória. Adaptado de (SHEN; LAI, 2001).

O procedimento *hot-deck* é um método de ajuste de conjuntos de dados quando há dados ausentes. Atualmente, várias pesquisas que são realizadas usam muitas variações do procedimento de *hot-deck*. Algumas dessas variações não correspondem exatamente à definição do procedimento de *hot-deck*, mas combina em uma classe mais geral de procedimentos, dos quais o procedimento supracitado é apenas um subconjunto.

Apesar de a técnica não ser fundamentada em uma base teórica clara, o método busca atenuar o desvio, classificando a amostra, e sabe-se que este objetivo é substancialmente complexo para seu alcance.

Dentro do grupo, os objetos obedecem a um critério de similaridade, ou seja, uma característica que os torna mais parecidos. Desta forma, os componentes inclinam-se para a homogeneidade buscando agrupar os elementos similares. No entanto, é relevante que se tenha correlação entre os atributos classificadores – que foram a base para a geração dos grupos – e os atributos ausentes, por efeito de pena de aquisição

de resultados incorretos.

A classificação pode não se fundamentar tão somente em resultados da amostra. Para quem coleta os dados, observa-se os resultados frente aos entrevistados, por exemplo: sexo, raça ou faixa etária, dentre outros. Neste sentido, verifica-se algumas vantagens que conduzem o uso da técnica *hot-deck*, conforme Magnani (2004) expõe: (i) consegue-se uma redução de desvio sem a imposição de um modelo rígido; (ii) produção de um conjunto de dados limpo, sem valores ausentes; (iii) preservação da distribuição da população representada pela amostra.; (iv) Para alguns valores ausentes, nenhuma informação sobre imputação pode ser encontrada. Isto permite que outras técnicas de imputação possam ser usadas em conjunto; (v) pode-se usar uma técnica diferente para cada grupo gerado; (vi) Não precisa de um modelo robusto para prever valores ausentes; e (vii) não assume nenhuma distribuição em particular.

Um grupo de dados imputados não pode ser considerado como originalmente preenchido, uma vez que a imputação oculta a incerteza pertinente ao processo. Neste sentido, Ford (1983) relata que os dados imputados precisam estar, de modo geral, marcados, já que isto possibilita que os dados sejam novamente gerados segundo a escolha do analista de dados.

As vantagens dos métodos *hot-deck* são: (i) simplicidade conceitual; (ii) bom nível de medição de variáveis; e (iii) a geração de uma base de dados completa, a qual poderá ser analisada através de procedimentos analíticos convencionais (BROWN; KROS, 2003a). Já uma de suas desvantagens é em definir o que é similar, o que descarta a importância de realizar a comparação entre diferentes técnicas de agrupamento.

Já os métodos *cold-deck* são muito similares aos métodos *hot-deck*. Basicamente, o que os difere é que no *hot-deck* os dados utilizados para a substituição dos dados faltantes estão no próprio conjunto de dados. Enquanto isso, no *cold-deck*, estão em outro conjunto de dados (BROWN; KROS, 2003a).

2.2.5.3.6 – Imputação *Múltipla*

A imputação múltipla foi proposta como uma alternativa às técnicas de imputação

global (RUBIN, 1978). A imputação múltipla requer a construção de M (≥ 2) conjuntos de dados completos, que são obtidos substituindo cada dado faltante por M valores imputados, obtidos através do mesmo procedimento de imputação. Segundo Rubin (1988), a imputação múltipla pode ser definida como:

> "Um processo onde diversos bancos de dados completos são criados pela imputação de valores diferentes para refletir incerteza sobre o valor correto a imputar. No próximo passo, as bases são tratadas pelos procedimentos padrões de análise de bases completas. Por fim, as análises de cada base são combinadas produzindo o resultado final"

A definição acima pode ser detalhada da seguinte forma: um método específico estima *n* valores sugeridos para cada atributo de uma tupla que possua valores ausentes. Cada uma destas *n* sugestões preenche o atributo ausente com seu respectivo valor, simulando *n* conjuntos de dados, como se tivesse ocorrido *n* imputações globais. Estes *n* conjuntos de dados são avaliados levando-se em consideração a variância, a fim de gerar uma imputação consolidada dos conjuntos de dados (CARTWRIGHT; SHEPPERD; SONG, 2004).

Desta forma, a imputação múltipla considera a incerteza associada aos valores estimados, produzindo estimativas de atributos não enviesados (WAYMAN, 2003). Enfim, a imputação múltipla pode ser considerada como uma extensão da imputação global, onde cada atributo ausente é substituído por um conjunto de valores estimados, reduzindo assim a incerteza inerente ao processo de imputação (RUBIN, 1988)

Soares (2007) optou por incluir a imputação múltipla na categoria de métodos híbridos, ao invés de considerá-la como um tipo de imputação, uma vez que as múltiplas sugestões podem criadas a partir de qualquer método de imputação: convencional, simples, ou até mesmo por outro método híbrido.

2.2.5.3.7 – Imputação Composta

A imputação composta, proposta por Soares (2007), é uma técnica que se baseia

no conceito de estratégias e planos de imputação, e com isso, permite que a complementação de dados ausentes seja realizada a partir de um planejamento (*workflow*). Soares (2007) disserta acerca da possibilidade de melhorar a qualidade do dado imputado por meio da utilização de tarefas de aprenzidado de máquina, tais como seleção, agrupamento ou classificação, precedendo a imputação, com o objetivo de melhorar a qualidade do dado gerado.

A sequência de tarefas que precedem a imputação é denominada de uma estratégia, e um plano de imputação é a associação de um algoritmo a cada uma destas tarefas. Por exemplo, uma estratégia poderia ser: (i) selecionar atributos principais; (ii) agrupar casos; e (iii) imputar valores. Um possível plano para esta estratégia poderia ser: (i) selecionar atributos principais com o algoritmo *PCA*; (ii) agrupar casos com o algoritmo *K-Means*; e (iii) imputar valores pela média (SOARES, 2007a).

O método de imputação no qual uma tarefa de agrupamento é realizada antes da imputação apresentou ganhos quando comparado a técnicas que simplesmente imputavam dados. Neste sentido, Soares (2007) percebeu que este método deveria ser objeto de novas investigações, com a utilização de algoritmos ainda não testados.

Em seus experimentos, Soares (2007) utilizou dados de três diferentes bases do *UCI Machine Learning Repository* (LICHMAN, 2018) em bases numéricas. Para cada base de dados, foram geradas ausências que variavam de 10% a 50%, com saltos percentuais de 10%, e o processo de imputação se deu combinando as tarefas de seleção (utilizando o algoritmo *PCA*) e agrupamento (utilizando o algoritmo *K-Means*) precedendo a imputação de dados utilizando a média (e suas variantes), o algoritmo dos *k* vizinhos mais próximos (k-NN), e redes neuronais *back propagation*. Soares (2007) apresentou uma nova metodologia, que recebeu o nome de imputação composta, onde é realizada uma tarefa específica precedendo o processo de imputação.

2.2.5.3.8 – Imputação em cascata

A partir dos bons resultados obtidos por Soares (2007), através da realização de tarefas de agrupamento antes da imputação, Ferlin (2008) propôs um método de imputação em cascata. Nele o conjunto de dados é agrupado de acordo com sua

morfologia de ausência[2] e, posteriormente, é realizado o processo de imputação em cascata, onde cada grupo imputado é incorporado ao conjunto de valores presentes. Ou seja, o conjunto de dados completo C1 imputado pelo passo corrente é levado em consideração na imputação do conjunto de dados incompleto seguinte, gerando o conjunto de dados completo C2. Este processo é realizado até que todo o conjunto de dados seja imputado.

A partir da taxonomia de Soares (2007), Ferlin (2008) classificou esta metodologia como híbrida, nomeando-a como imputação em cascata. Em seus experimentos, este trabalho utilizou dados de cinco diferentes bases dados numéricas diferentes que pertencem à UCI Machine Learning Repository (LICHMAN, 2018). Os percentuais de ausência variaram de 10% até 30%. O algoritmo utilização para a realização do agrupamento foi o *Self-Organizing Map* (SOM) e o algoritmo *K-Nearest Neighbors* (*KNN*) foi utilizado para realizar a imputação dos dados.

Neste método, uma tarefa de agrupamento precede o processo de imputação. Os casos incompletos são distribuídos em grupos considerando como critério de alocação o conceito de morfologia da ausência neles existentes. A morfologia de ausência considera a distribuição espacial dos dados ausentes, ou seja, a forma como os valores presentes e ausentes nos atributos estão distribuídos nas tuplas (FERLIN, 2008)

Desta forma, uma análise da relação posicional da ausência de dados é utilizada para a realização do agrupamento de casos. A cada etapa de imputação, os grupos previamente complementados são reutilizados para a imputação dos grupos posteriores, criando um efeito cascata (FERLIN, 2008). A Figura 5 apresenta o método de imputação em cascata.

[2] Conceito proposto por Ferlin (2008) para descrever a distribuição de valores presentes e ausentes nos atributos de um conjunto de dados

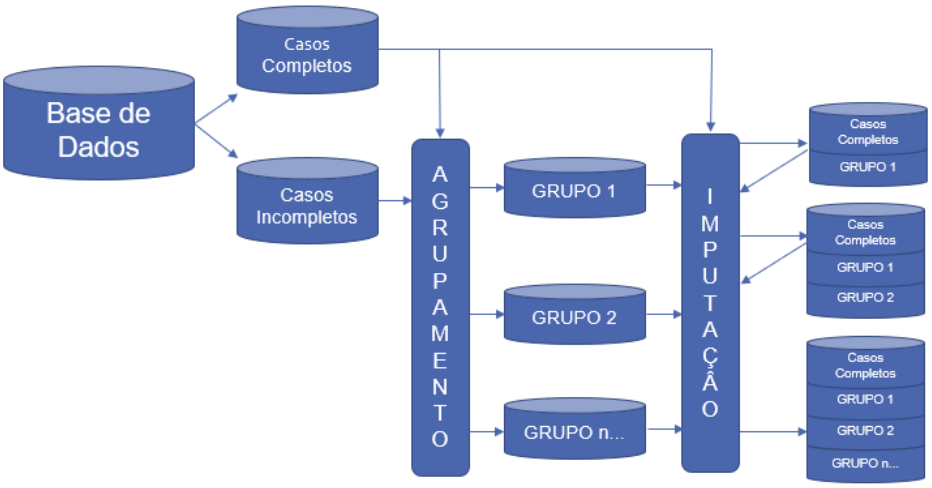

Figura 5 - Método de Imputação em Cascata. Fonte: Ferlin, 2008.

A morfologia de ausência permite que os métodos de agrupamento ou imputação não sejam pré-determinados. Cada grupo pode ter um método distinto de imputação, ou seja, grupos com padrões menos complexos podem utilizar métodos de imputação mais simples, enquanto grupos com padrões mais complexos podem ser imputados por métodos que exigem maior capacidade computacional (CASTANEDA et al., 2008).

2.2.5.3.9 – Imputação utilizando Ensemble

Tomadas de decisão em conjunto estão presentes no cotidiano da humanidade, tais como, em sistemas de eleição de representantes políticos, sindicais ou decisões de conselhos especializados. Ao buscar diferentes opiniões médicas, ler diferentes recomendações de um produto específico ou sobre uma oportunidade de trabalho, o ser humano busca construir uma decisão a partir do conjunto de diferentes visões.

A tomada de decisão em conjunto também pode ser realizada no aprendizado de máquina. Neste contexto, os algoritmos que possuem o objetivo de realizar decisões em conjunto são categorizados como do tipo *Ensemble* (ZHANG; MA, 2012).

Algoritmos do tipo *Ensemble* buscam criar conjuntos com diferentes componentes, os quais podem ser classificadores ou regressores. Logo, duas características importantes destes componentes precisam ser observadas: a precisão e a diversidade.

Um componente preciso é aquele que possui uma taxa de erro melhor que uma

adivinhação aleatória para os novos valores, como lançar uma moeda ao ar e adivinhar se o resultado será cara ou coroa (DIETTERICH, 2000). Se dois componentes apresentam diferentes taxas de erros para uma determinada predição, eles podem ser considerados distintos. Os conjuntos destes distintos componentes constituem a diversidade, uma importante propriedade para o sucesso da predição (DIETTERICH, 2000).

Zhang e Ma (2012) descrevem três etapas importantes para o sucesso de um sistema de tomada de decisão em conjunto: a seleção dos dados, também chamada de amostragem; os componentes de dados de treinamentos e combinação dos componentes.

A seleção dos dados precisa apoiar a diversidade para que a tomada de decisão em conjunto seja melhor. Uma estratégia para isto é a escolha de diferentes conjuntos de dados utilizando uma técnica chamada *Bootstrap*. Nela, os dados são sorteados de forma aleatória com repetição a fim de criar *N* distintos conjuntos de dados.

Um conjunto é bom se seus componentes individuais forem precisos e diversos. *Bagging* e *Boosting* são duas abordagens muito utilizadas para construir conjuntos precisos (ZHANG; MA, 2012).Tanto o Bagging quanto o Boosting usam técnicas de “reamostragem” para manipular os dados de treinamento.

Bagging manipula o conjunto de dados de treinamento original de N instâncias desenhando aleatoriamente com instâncias de substituição. Portanto, no conjunto de dados de treinamento resultante, algumas das instâncias originais podem aparecer várias vezes, enquanto outras podem desaparecer. O *Bagging* é frequente eficaz em algoritmos de aprendizagem "instáveis", como redes neurais e árvores de decisão, onde pequenas mudanças no conjunto de dados de treinamento levam a grandes mudanças nas previsões (LUO et al., 2016).

Boosting possui uma abordagem iterativa para geração do chamado componente forte, capaz de atingir baixas taxas de erro de treinamento a partir de componentes fracos, cada um sendo levemente mais preciso que uma escolha aleatória, ou seja, apresentando uma probabilidade de acerto pouco maior que 50% (ZHANG; MA, 2012).

Diferentemente da técnica *Bagging*, a seleção de dados de treinamento em *Boosting* não utiliza *Bootstrap*. Nesta técnica, modelos são treinados sequencialmente, ou seja, quando um treinamento é realizado, este gera valores precisos e valores com erro. O treinamento seguinte é direcionado aos valores gerados erroneamente pelo treinamento anterior, fazendo com que o aprendizado seja incremental (ZHANG; MA,

2012).

Uma nova abordagem sobre *Boosting* foi proposta por pesquisadores, denominada de *Adaptative Boosting* (*AdaBoost*). A ideia deste algoritmo e utilizar diferentes versões do mesmo conjunto de dados de treinamento com pesos associados a cada registro, a fim de induzir os modelos que são obtidos de forma sequencial. A cada iteração, os pesos são recalculados de acordo com as predições mais próximas e mais distantes dos valores reais gerados na execução anterior (FREUND; SCHAPIRE; ABE, 1999).

Semelhante ao *AdaBoost*, o algoritmo *Gradient Boosting* utiliza o mesmo conceito combinando componentes fracos de forma iterativa para a geração de um componente forte, mas, adicionalmente, utiliza um gradiente descendente para minimizar a função de perda. Este algoritmo busca construir um componente forte que maximize a correlação com o gradiente negativo da função ao de perda do comitê como um todo (NATEKIN; KNOLL, 2013).

Os resultados empíricos mostram que com pouco ou nenhum ruído de classificação, o algoritmo *Boosting* na maioria dos casos apresenta melhores resultados do que um único componente, e, às vezes, é mais preciso do que o algoritmo de *Bagging* (LUO et al., 2016). No entanto, em situações com ruído de classificação substancial, o algoritmo *Boosting* é frequentemente menos preciso do que um único componente porque ele muitas vezes apresenta *overfiting* em conjuntos de dados ruidosos (LUO et al., 2016).

2.2.5.3.10 – Imputação utilizando métodos estatísticos

Esta categoria de soluções para tratamento de valores ausentes engloba técnicas estatísticas e probabilísticas de obtenção de um modelo que consiga representar de forma genérica as características dos dados. As técnicas mais utilizadas desta categoria são a média, a moda e os modelos de regressão, além dos algoritmos de verossimilhança e os métodos bayesianos.

Os métodos estatísticos mais simples e mais utilizados na imputação de dados são o cálculo da média, no caso de atributos contínuos, e o cálculo da moda, quando se trata de atributos categóricos. Estes métodos apresentam como vantagem o eficiente desempenho em bases de dados com uma grande quantidade de dados. No entanto,

estes métodos podem gerar viés nos dados, além de reduzir o desvio-padrão da amostra.

Já os métodos de verossimilhança buscam estimar os parâmetros de uma função de distribuição estatística, a fim de encontrar um modelo que represente o conjunto de dados, criando condições de regredir qualquer valor ausente existente. Os algoritmos EM – Expectation-Maximization (DEMPSTER; LAIRD; RUBIN, 1977) e método de verossimilhança com informações completas (MYRTVEIT; STENSRUD; OLSSON, 2001) são os principais representantes desta categoria de método estatístico.

O algoritmo *EM* (*Expectation-Maximization*) é um método estatístico iterativo de verossimilhança, cujo objetivo é estimar os parâmetros de uma função de densidade (probabilística) de uma amostra. O algoritmo EM possui uma simples implementação e consegue encontrar com confiabilidade o máximo global através de passos estáveis e crescentes (DEMPSTER; LAIRD; RUBIN, 1977).

Já a técnica *full information maximum likelihood* (FIML) utiliza todas as observações para estimar os parâmetros do modelo e os seus erros padrão, sendo que alguns casos contribuem com mais informação do que outros (ENDERS, 2010). Neste sentido, se os dados estão completos, a função de verossimilhança a maximizar é para o elemento *i, conforme* representado pela equação (4), a seguir:

$$logL_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log\left|a_{i,j}\right| - \frac{1}{2}\,(Y_i - \mu)^T a_{1,j}^{-1}\,(Y_i - \mu)^T, \qquad (4)$$

onde, k é o número de variáveis, Y_i é o vetor das observações para o elemento i, μ é o vetor das médias populacionais e $a_{i,j}$ é a matriz das variâncias-covariâncias

No caso de existirem dados ausentes, a função de verossimilhança para o elemento *i* passa a ser dada pela seguinte expressão (5):

$$logL_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\left|b_{i,j}\right| - \frac{1}{2}\,(Y_i - \mu_i)^T b_{1,j}^{-1}\,(Y_i - \mu_i)^T, \qquad (5)$$

onde, k_i representa, para o elemento i, o número de variáveis com valores completos, e μi e $b_{i,j}$ são respectivamente o vetor das médias populacionais e a matriz das variâncias-covariâncias, calculados apenas com os dados disponíveis. Vale salientar que o cálculo da função $logL_i$ para a observação i depende apenas das variáveis e dos parâmetros para os quais esse elemento tem dados completos (ENDERS, 2010).

A função de verossimilhança final corresponde à soma de *N* funções de verossimilhança, para os *N* elementos, sendo dada pela equação (6).

$$logL\left(\mu, a_{i,j}\right) = \sum_{i=1}^{N} logL_i \quad (6)$$

Quando o mecanismo de ausência dos dados *é* MCAR ou MAR e os dados apresentam uma distribuição normal, o método FIML produz estimativas dos parâmetros, erros padrão e testes estatísticos que são consistentes e eficientes (PETERS; ENDERS, 2002).

Por fim, métodos bayesianos também podem ser utilizados na tarefa de imputação de dados ausentes. Eles possuem como base uma consistente teoria estatística, os seus algoritmos possuem uma discreta, porém importante presença no processo de imputação de valores ausentes.

A representação das redes bayesianas é feita por meio de um grafo acíclico direcionado, no qual os nós representam variáveis de um domínio, e os arcos representam a dependência condicional entre as variáveis. Para representar a força da dependência, são utilizadas probabilidades, associadas a cada grupo de nós pais-filhos na rede (PEARL, 2014).

Um exemplo de rede bayesiana é o problema de metástase de câncer (HRUSCHKA JR.; EBECKEN, 2002), conforme apresentado na Figura 6. Nela, os nós representam as variáveis (metástase, tumor cerebral, coma, aumento do nível de cálcio no plasma, fortes dores de cabeça), enquanto as conexões entre os nós representam a influência causal entre estas variáveis.

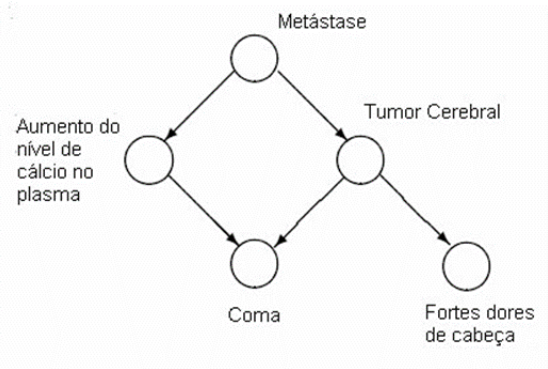

Figura 6 - Aplicação da rede bayesiana em um problema médico.
Fonte: Adaptado de (HRUSCHKA JR.; EBECKEN, 2002).

Quando os nós não estão conectados, significa que não existe influência causal entre as variáveis. Para as variáveis que estão conectadas, a intensidade da influência de cada conexão é definida pela probabilidade condicional $P(X_i|\Pi X_i)$, onde X_i é a *i*-ésima variável, e ΠX_i é o conjunto de nós-pai de X_i . Este exemplo demonstra como uma rede bayesiana pode ser utilizada como um instrumento de representação do conhecimento, permitindo que sejam realizadas inferências a partir de probabilidades.

2.2.5.3.11 – Imputação utilizando aprendizado de máquina.

A área de inteligência artificial ao longo dos anos desenvolveu diversas técnicas de aprendizado de máquina para múltiplas aplicações, incluindo a imputação de dados ausentes (READS, 2017). Atualmente, com o grande volume de documentos produzidos na web, vê-se a necessidade de ferramentas tecnológicas que ajudem a processar, extrair e resumir informações úteis (D ANDRÉA, 2006). Conforme Aranha e Passos (2006), um dos métodos, por exemplo, que surgiram para tentar resolver esse problema é a mineração de texto, que, entre outras técnicas, utiliza o aprendizado de máquina para minerar grandes coleções documentais em formato digital, a fim de extrair informações relevantes para os destinatários das informações

De acordo com Monard e Baranauskas (2003), o aprendizado de máquina é uma área que estuda como construir programas de computador que melhoram seu desempenho em alguma tarefa graças à experiência. Desta forma, baseia-se em ideias de várias disciplinas, como inteligência artificial, estatística e probabilidade, teoria da informação, psicologia e neurobiologia, teoria do controle e complexidade computacional.

Conforme Prati (2006), para usar a abordagem de aprendizagem, uma série de decisões deve ser considerada, incluindo a seleção do tipo de treinamento, a função objetivo a ser aprendida, sua representação e o algoritmo para aprender essa função a partir de exemplos de treinamento.

Para Araújo ; de Souza e Matheus (2020), os algoritmos de aprendizagem têm se mostrado úteis em diversos domínios de aplicação, como na mineração de dados em grandes bancos de dados que contêm regularidades implícitas, que podem ser descobertas de forma automatizada, em domínios pouco compreendidos e onde os humanos não possuem o conhecimento necessário para desenvolver algoritmos

eficazes, e em domínios onde os programas devem se adaptar dinamicamente para responder às mudanças nas condições do ambiente.

As técnicas de aprendizagem são classificadas como supervisionadas, não supervisionadas e por reforço. Na técnica supervisionada, o objetivo é aprender o mapeamento das respostas corretas aos dados de entrada fornecidos ao usuário. Desta forma, o sistema aprende o mapeamento de saída correto para cada padrão de entrada apresentado a ele (OLIVEIRA, 2020).

Na técnica não supervisionada, o aprendizado busca padrões previamente não detectados em um conjunto de dados com o mínimo de intervenção humana. Portanto, no aprendizado não supervisionado, a intervenção humana não é necessária para produzir um conjunto de dados previamente categorizado para ser apresentado ao algoritmo de aprendizado (OLIVEIRA, 2020).

Já a técnica de aprendizado por reforço consiste no treinamento de modelos de aprendizado de máquina para tomar uma sequência de decisões. Nesta técnica, a máquina aprende a atingir uma meta em um ambiente incerto e potencialmente complexo.

No aprendizado por reforço, a máquina utiliza tentativa e erro para encontrar uma solução para o problema. Para que a máquina realize o que o programador deseja, ela recebe recompensas pelos acertos ou penalidades pelos erros que executa. Desta forma, o objetivo da máquina é maximizar a recompensa total (LACERDA, 2020).

Monard e Baranauskas (2003) e Clementino (2020) mencionam os principais paradigmas do aprendizado de máquina, que são o modelo probabilístico, aprendizado simbólico e indução de regras, redes neurais, algoritmos baseados em evolução, aprendizado analítico e métodos híbridos.

O modelo probabilístico é um dos métodos de aprendizagem mais antigos, frequentemente usado para classificar diferentes objetos em classes previamente definidas com base em um conjunto de características. Exemplos disso são o modelo Bayesiano e o modelo Bayes ingênuo.

A aprendizagem simbólica e a indução de regras podem ser classificadas de acordo com a estratégia de aprendizagem subjacente na aprendizagem de rotina (memorização), por instruções, por analogia, por exemplos e por descoberta. Um exemplo de técnica desse tipo é o algoritmo de árvore de decisão e suas variações. Eles apresentam o resultado da classificação na forma de árvores de decisão ou um conjunto de regras de produção (MACENA; PIRES; PESSOA, 2020).

Redes neurais artificiais imitam neurônios humanos. Desta forma, o

conhecimento é representado por descrições simbólicas; o conhecimento é aprendido e lembrado por redes de neurônios artificiais interconectados por sinapses com pesos e unidades lógicas de limiar (ARAÚJO; DE SOUZA; MATEUS, 2020).

Algoritmos evolutivos imitam o processo de evolução na natureza. Identifica três categorias: genética, estratégias evolutivas e programação evolutiva. Os algoritmos genéticos imitam os princípios dos genes e usam operadores de mutação e crossover na população para selecionar os indivíduos mais adaptados e repetir essa operação em várias gerações até que o melhor indivíduo seja obtido. Em estratégias evolutivas, uma população de números reais é desenvolvida que codifica as soluções possíveis de um problema numérico e os tamanhos dos saltos. Desta forma, a seleção está implícita nas estratégias evolutivas (MEDINA, 2012).

A aprendizagem analítica representa o conhecimento como regras lógicas e faz um raciocínio sobre eles para a busca de testes, que são compilados em regras mais complexas para resolver problemas com um pequeno número de pesquisas (ARAÚJO; DE SOUZA; MATEUS, 2020).

Ressalta-se que, na prática, esses paradigmas de aprendizado de máquina costumam ser utilizados combinando diversas técnicas para aproveitar melhor as vantagens que cada uma apresenta e corrigir as fragilidades que teriam se fossem utilizadas individualmente.

2.3 – Agrupamento de dados

O agrupamento é uma das ações mais primitivas do homem. A procura por correspondência e distinções, de modo geral, aparece nas dinâmicas mais variadas, tais como a divisão de uma classe escolar para a realização de inúmeras atividades com os discentes por gêneros, a categorização de esportistas, por exemplo, infantil, júnior e sênior, dentre outros.

Logo, quando se refere às áreas de conhecimento, vê-se, também, de forma notória, o agrupamento e, neste sentido, a psiquiatria, biologia, psicologia, geologia, beneficiam-se desta técnica (JAIN; DUBES, 1988). Observa-se que na biologia há a classificação para animais e plantas, na medicina se classificam as doenças, já no marketing busca-se identificar pessoas com hábitos de compras semelhantes, e não

poderia ser diferente na Computação. A ação de agrupamento se depara com aplicações diretas, tais como o reconhecimento de padrões e análise de imagens, ou indiretas. No entanto, o objetivo fundamental de investigação desse processo permeia a aplicação do agrupamento no procedimento de descoberta de conhecimento em bases de dados, especialmente na fase de pré-processamento de dados.

Em seu conceito, a palavra “agrupar” possui o significado de reunir objetos que tenham características semelhantes. Levando-se em consideração que, geralmente, não conhecemos quais características fazem parte para tal delimitação, o agrupamento necessita ter por base algum parâmetro de similaridade. Bezerra (2006 p.44) determina o conceito de modelo de agrupamento como “um conjunto de grupos gerados a partir dos objetos de uma coleção”.

Desta forma, o intuito de se fazer o agrupamento é elevar a similaridade dos objetos de um determinado conjunto, e minimizá-la para objetos de grupos dissemelhantes. Logo, compreende-se que o êxito de um algoritmo de agrupamento encontra-se na escolha de uma boa medida de similaridade.

Ao se investigar os estudos sobre métodos de imputação de dados, é possível observar que a maioria das pesquisas estudam algoritmos para a realização da imputação global. Soares (2007) já observava este fato, realizando um estudo propondo uma abordagem chamada de imputação composta, abordada na subseção 2.3.3.7. Como exemplo dessas tarefas existem os algoritmos de aprendizado de máquina que realizam agrupamentos nos dados.

Os algoritmos de agrupamento surgiram como uma poderosa ferramenta para analisar com precisão o enorme volume de dados gerados por aplicativos modernos. Em particular, seu principal objetivo é categorizar os dados em grupos (*clusters*), de modo que os objetos sejam agrupados no mesmo *cluster* quando forem semelhantes, de acordo com métricas específicas (BANO; KHAN, 2018).

Fahad et. al. (2014) compararam, tanto do ponto de vista teórico quanto do empírico, diversos algoritmos de agrupamento e desenvolveram uma estrutura de categorização com base em suas principais propriedades com o objetivo de propor uma taxonomia para os algoritmos de agrupamento. Segundo estes autores, os algoritmos de agrupamento podem ser classificados em cinco categorias: *Partitioning Based Clustering, Hierarchical Based Clustering, Density Based Clustering, Grid Based Clustering* e *Model Based Clustering.*

Bano e Khan (2018), ao investigar os diferentes algoritmos de agrupamento, realizou uma categorização dos algoritmos de agrupamento muito similar à de Fahad

et. al (2014), acrescentando a categoria *Fuzzy Based Clustering*, incluindo o algoritmo Fuzzy C-Means (FCM) nesta categoria, diferente de Bano (2018), que havia classificado este algoritmo como *Partitioning Based Clustering.*

Já os autores Pandey e Shukla (2019), além de considerar estas seis categorias anteriormente definidas, acrescentaram uma nova categoria chamada *Graph Based Clustering*, incluindo nela algoritmos que realizam o agrupamento utilizando a teoria de grafos. A Tabela 1 apresenta as setes categorias consideradas nesta dissertação e os algoritmos incluídos em cada categoria (PANDEY; SHUKLA, 2019).

Tabela 1 - Categorias de agrupamento e algoritmos.

Algoritmo de Agrupamento	Volume		Variedade		Velocidade
	Dado com Alta dimensão	Tratamento de ruído	Tipo de Dataset	Cluster shape	Complexidade
Agrupamento do tipo *Partition based*					
K-Means	Não	Alto	Numérico	Convexo	0 (knt)
K-Medoies	Não	Baixo	Categórico	Convexo	$0(k(n-k)^2)$
PAM	Não	Baixo	Numérico	Convexo	$0\ (k^3 * n^2)$
CLARA	Não	Baixo	Numérico	Convexo	$0(ks^2+k(n-k))$
CLARANS	Não	Baixo	Numérico	Convexo	$0(n^2)$
Agrupamento do tipo *Hierarchical based*					
BIRCH	Não	Baixo	Numérico	Convexo	0(n)
CURE	Sim	Alto	Numérico	Arbitrário	$0(n^2 \log n)$
ROKE	Sim	Baixo	Numérico/Categórico	Arbitrário	$0(n^2 \log n)$
Chameleon	Não	Baixo	Todos os tipos	Arbitrário	$0(n^2)$
ECHIDNA	Não	Baixo	Multivariado	Convexo	$0(nb(1+\log_b m)$
WARDS	Não	Baixo	Numérico	Arbitrário	----------
SNN	Não	Baixo	Categórico	Arbitrário	$0(n^2)$
Agrupamento do tipo *Density based*					
DBSCAN	Não	Baixo	Numérico	Arbitrário	0(nlogn)
OPTICS	Não	Baixo	Numérico	Arbitrário	0(nlogn)
Mean-shift	Não	Baixo	Numérico	Arbitrário	0 (kernel)
DENCLUE	Sim	Alto	Numérico	Arbitrário	0(log |d|)
GDBSCAN	Não	Baixo	Numérico	Arbitrário	----------
Agrupamento do tipo *Grid based*					
STING	Sim	Baixo	Espacial	Arbitrário	0(n)
CLIQUE	Sim	Médio	Numérico	Convexo	$0(n+k^2)$
Wave Cluster	Não	Alto	Espacial	Arbitrário	0(n)
OptiGrid	Sim	Alto	Espacial	Arbitrário	0(nd) até 0(nd-log n)
MAFIA	Não	Alto	Numérico	Arbitrário	$0(c^p + p^n)$
ENCLUS	Não	Alto	Numérico	Arbitrário	$0(nd+m^d)$
PROCLUS	Sim	Alto	Espacial	Arbitrário	0(n)
ORCLUS	Sim	Alto	Espacial	Arbitrário	$0(d^3)$
STIRR	Não	Baixo	Categórico	Arbitrário	0(n)
Agrupamento do tipo *Model based*					
COBWEB	Não	Médio	Numérico	Arbitrário	$0(n^2)$
SLINK	Não	Médio	Numérico	Arbitrário	$0(n^2)$
SOM	Sim	Baixo	Multivariado	Arbitrário	$0(n^2m)$
ART	Não	Alto	Multivariado	Arbitrário	(type+layer)
EM	Sim	Baixo	Espacial	Convexo	0(knp)

Agrupamento do tipo Fuzzy based					
FCM	Não	Alto	Numérico	Convexo	0(n)
FCS	Não	Alto	Numérico	Arbitrário	0 (kernel)
MM	Não	Baixo	Numérico	Arbitrário	$0(v^2n)$
Agrupamento do tipo Graph based					
CLICK	Não	Alto	Numérico/Categórico	Arbitrário	0(kf(v,e))
MST	Não	Alto	Numérico/Categórico	Arbitrário	0(e log v)

Fonte: Adaptado de Pandey e Shukla (2019).

Os algoritmos do tipo *Partitioning-based* determinam todos os clusters de forma imediata, dividindo os objetos de dados em várias partições, onde cada partição representa um cluster *(FAHAD et al., 2014)*. A ideia básica desse tipo de algoritmo de agrupamento é considerar o centro dos pontos de dados como o centro do cluster correspondente. Os algoritmos *K-means* e *K-medoids* são os mais famosos desse tipo para a realização de agrupamento (XU; TIAN, 2015).

O algoritmo *K-means* é o mais conhecido algoritmo da família *flat clustering* e considera que objetos a serem agrupados estão em uma representação vetorial (MANNING; RAGHAVAN; SCHUTZE, 2008). Através do cálculo dos centroides (centros de gravidade ou médias) de pontos de um conjunto de dados, o algoritmo realiza o agrupamento dos dados com base no cálculo iterativo das distâncias dos objetos aos centroides atuais (JAIN; DUBES, 1988), conforme apresentado na Figura 7.

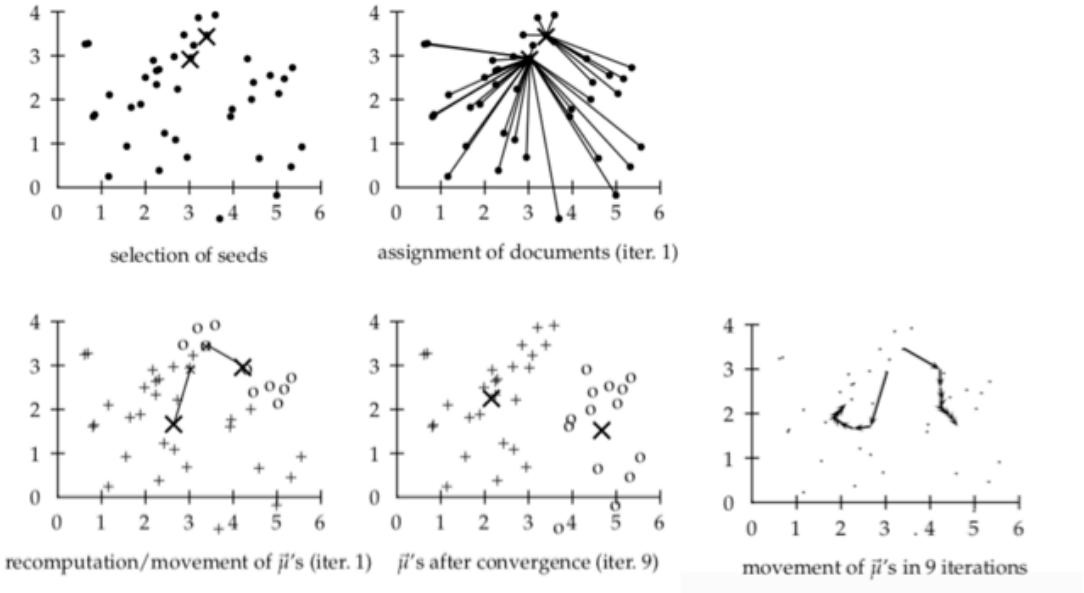

Figura 7 - Um exemplo de K-Means para K = 2. A posição dos dois centroides converge após nove iterações. Fonte: (MANNING; RAGHAVAN; SCHUTZE, 2008).

A ideia central do *K-means* é atualizar o centro do cluster que é representado pelo centro dos pontos de dados, pela computação analítica e pelo processo analítico que continuará até que alguns critérios de convergência sejam atendidos (FAHAD et. al. 2014).

Os métodos de agrupamento do tipo Fuzzy-Based dividem o conjunto de dados com base no valor discreto [0,1]. Em geral, esses algoritmos são adequados para situações de alta precisão com alta probabilidade, No entanto, apresenta baixa escalabilidade, baixo ótimo local, e o agrupamento é sensível ao parâmetro inicial definido e ao número de clusters necessários para a predefinição.

O algoritmo Fuzzy C-Means (FCM) é um algoritmo representativo do agrupamento do tipo Fuzzy-Based, que tem como base os conceitos de k-means para particionar o conjunto de dados em clusters. O algoritmo FCM é um método em que cada objeto é destinado a um cluster de acordo com um grau de confiança. Portanto, um objeto pode pertencer a mais de um cluster com diferentes graus de confiança (HUANG, 1997).

O algoritmo FCM busca encontrar o ponto mais característico de cada cluster, nomeado como o centro do cluster. Em seguida, computa o grau de associação para cada objeto nos clusters. O algoritmo FCM também minimiza a variação intracluster. No entanto, ele herda os problemas do K-means, pois o mínimo é apenas local e os agrupamentos finais dependem da escolha inicial dos pesos *(FAHAD et al., 2014)*.

O algoritmo FCM segue o mesmo princípio do algoritmo K-means, ou seja, pesquisa iterativa dos centros de cluster e a atualização das associações de objetos. A principal diferença é que, em vez de tomar uma decisão difícil sobre qual cluster o objeto deverá pertencer, ele atribui a um objeto um valor que varia de 0 a 1 para medir a probabilidade com que o objeto pertencerá a cada cluster. Uma regra *fuzzy* afirma que a soma do valor da associação de um objeto para todos os clusters deve ser 1. Quanto maior o valor da associação, maior a probabilidade de um objeto pertencer a esse cluster *(FAHAD et al., 2014)*.

Nos algoritmos tipo *Hierarchical-based*, a similaridade entre os objetos é organizada em uma estrutura hierárquica, ou seja, o relacionamento dos objetos (ou grupos) são representados por uma árvore, usualmente conhecida como dendrograma.

Objetos similares ficam em ramos próximos da árvore. O comprimento dos ramos reflete o grau de similaridade e registra a sequência de uniões e divisões do processo de agrupamento. A árvore retornada por estes métodos pode ser facilmente transformada em partições, bastando “cortar” o dendrograma em um certo nível (FAHAD

et. al. 2014).

Considerando o processo de construção da árvore, os algoritmos hierárquicos podem ser subdivididos em dois tipos: (i) divisivos (*top-down*), os quais partem de um grupo único conjunto inicial de objetos e, a cada passo, dividem os grupos gerados na iteração anterior em grupos menores, até que *n* grupos sejam formados,; e (ii) aglomerativos (*bottom-up*), os quais partem de *n* grupos (por exemplo, cada objeto da coleção é considerado um grupo) e, iterativamente, unem os grupos menores em grupos cada vez maiores, até que um único grupo que contém todos os objetos seja formado (BANO; KHAN, 2018).

Os algoritmos Density-based realizam cálculos que separam os objetos com base em suas regiões de densidade, conectividade e limite e estão intimamente relacionados aos vizinhos mais próximos do ponto (BANO; KHAN, 2018). Um cluster, definido como um componente denso conectado, cresce em qualquer direção que a densidade o direcione. Portanto, algoritmos baseados em densidade são capazes de descobrir grupos de contornos arbitrários (XU; TIAN, 2015). Além disso, estes algoritmos fornecem uma proteção natural contra ruídos (outliers). Os algoritmos DBSCAN, OPTICS, DBCLASD e DENCLUE usam esse método para filtrar ruídos (outliers) e descobrir grupos de formato arbitrário (FAHAD et al., 2014).

No grupo de algoritmos do tipo grid-based, espaço dos objetos de dados é dividido em grades. A principal vantagem dessa abordagem é o rápido tempo de processamento, uma vez que o algoritmo percorre o conjunto de dados uma vez para calcular os valores estatísticos das grades (XU; TIAN, 2015).

O desempenho deste método depende do tamanho da grade, que geralmente é muito menor que o tamanho da base de dados. No entanto, para distribuições altamente irregulares, o uso de uma grade uniforme pode não ser suficiente para obter a qualidade de agrupamento necessária ou o preenchimento necessário a um requisito. Os algoritmos STING, Wave-Cluster, CLIQUE e OptiGrid são exemplos típicos dessa categoria (FAHAD et al., 2014).

Os algoritmos do tipo *Model-based* otimizam o ajuste entre os dados fornecidos e alguns modelos matemáticos pré-definidos. Estes algoritmos tem como base o pressuposto de que os dados são gerados por uma mistura de distribuições de probabilidade subjacentes. Além disso, estes algoritmos permitem a determinação automática de clusters com base em estatísticas padrão, considerando a presença do ruído (outliers) e, assim, produzindo um método robusto de agrupamento (BANO; KHAN, 2018).

Existem duas principais abordagens no universo dos algoritmos do tipo *Model-based*: Abordagens de estatísticas e redes neurais. O *Expectation Maximization* (EM) é um algoritmo Model-based estatístico bastante conhecido que usa um modelo de densidade de mistura. Além dele, também temos os algoritmos COWEB e CLASSIT. Como exemplo de abordagem de redes neurais, temos o Self-Organizing Map (SOM) (BANO; KHAN, 2018).

Os métodos de agrupamento do tipo *Graph-Based* dividem o conjunto de dados com base no vértice relacionado. Os nós são definidos como os pontos de dados, e as bordas o relacionamento entre esses pontos. Em geral, este algoritmo é adequado para alta eficiência com alta precisão, mas a complexidade do tempo depende da complexidade do gráfico. Alguns algoritmos típicos desse tipo de algoritmo de agrupamento são CLICK e MST (PANDEY; SHUKLA, 2019).

.

3- Metodologia

Neste capítulo são abordados os assuntos relacionados à metodologia utilizada para a realização desta dissertação. Inicialmente é apresentado o processo de construção do descritor para a realização da busca no repositório de pesquisa científica, em seguida, os procedimentos adotados para seleção e avaliação dos artigos investigados.

3.1 – Busca na base de periódicos *Scopus*

A Para a realização de uma busca no repositório de pesquisa científica *Scopus* - serviço de indexação de citações científicas com base em assinaturas on-line - com o objetivo de encontrar trabalhos científicos que investigaram a técnica de imputação de dados *hot-deck*, objeto de estudo desta dissertação, foi construído um descritor (também chamado *string* de busca). Ele possui sete componentes, cada qual com um propósito específico. O primeiro, também principal, apresenta a seguinte descrição: TITLE-ABS-KEY ("*imputation*" OR "*missing data*") AND TITLE-ABS-KEY ("*hot-deck*" OR "*clustering algorithm*") OR TITLE-ABS-KEY (*"imputation"* AND *"clustering"*). Sua finalidade é buscar todos os artigos que abordem o tema imputação ou ausência de dados, realizando uma interseção com os temas *hot-deck* ou algoritmos de agrupamento.

No entanto, a avaliação das propostas dos periódicos retornados revelou que a combinação dos termos "*missing data"* e "*clustering"* seleciona artigos não incluídos na delimitação do objeto desta dissertação, uma vez que estes trabalhos não realizavam o processo de imputação de dados ausentes.

Desta forma, como solução para esta dificuldade, foi utilizada a combinação dos termos *"imputation"* e *"clustering"*. Entretanto, esta alteração retornou artigos que realizavam a imputação com o objetivo de melhorar o agrupamento, um processo diferente de uma imputação tipo *hot-deck*, e consequentemente, demandando classificação manual. No entanto, nem todos os autores que realizam o agrupamento antes de imputação classificam ou relacionam este procedimento com o termo "*hot-deck*", e estes artigos não poderiam ser descartados. Apesar de esta abordagem ser

mais trabalhosa, ela se revelou mais conservadora, no sentido de preservar a qualidade do estudo, alcançando os artigos que abordam o agrupamento dentro do contexto de imputação de dados ausentes, mesmo que o autor não utilize o termo *“hot-deck”*.

Após a formação do descritor principal, foram definidos cinco descritores almejando delimitar os resultados da busca dentro objeto de estudo, com o objetivo de excluir artigos não foram contemplados no objeto de pesquisa.

O primeiro descritor de exclusão definido foi: AND NOT TITLE-ABS-KEY (*"microarray"* OR *"RNA"* OR *"gene"* OR *"gene expression profiling"* OR *"genotype"* OR *"genomics"* OR *"genetics algorithms"* OR *"biology"* OR *"bioinformatics"* OR *"proteomics"* OR *"genetics"* OR *"human"* OR *"humans"* OR *"animals"*) . O objetivo deste trecho de descritor foi realizar a exclusão de periódicos da área de Genética e Biologia, pois são áreas que não estão incluídas no objeto de estudo desta dissertação.

Já o segundo descritor de exclusão definido foi: AND NOT TITLE-ABS-KEY (*"image"*), uma vez que não é objetivo desta dissertação investigar a imputação de dados dentro do contexto de processamento de imagens.

O terceiro descritor de exclusão definido foi: AND NOT TITLE-ABS-KEY (*"signal*" OR *"signal processing"* OR *"audio"* OR *"sound"*), com a finalidade de excluir os periódicos que estão presentes na área de multimídia, considerando o processamento de sinais e som.

Sobre o quarto descritor de exclusão: AND NOT TITLE-ABS-KEY (*"web"* OR *"text"*), desconsidera periódicos que abordem temas relacionados análise de textos na internet.

Como último descritor de exclusão, a expressão: AND NOT TITLE-ABS-KEY (*"game"* OR "games"), evita periódicos relacionados ao desenvolvimento e pesquisa de jogos.

Concluídos os descritores de exclusão, considerou-se os filtros relacionados aos tipos de publicação: AND (LIMIT-TO (SRCTYPE , *"j"*) OR LIMIT-TO (SRCTYPE , *"p"*)) AND (LIMIT-TO (PUBSTAGE , *"final"*)) AND (LIMIT-TO (DOCTYPE , *"ar"*) OR LIMIT-TO (DOCTYPE , *"cp"*)) AND (LIMIT-TO (LANGUAGE , *"English"*)) . Este descritor inclui as publicações em periódicos e artigos de conferências em estágio final de publicação, publicados no idioma inglês.

Desta forma, o descritor completo para a realização da busca na base *Scopus* ficou definido como: TITLE-ABS-KEY (*"imputation"* OR *"missing data"*) AND TITLE-ABS-KEY (*"hot-deck"* OR *"clustering algorithm"*) OR TITLE-ABS-KEY (*"imputation"* AND *"clustering"*) AND NOT TITLE-ABS-KEY (*"microarray"* OR *"RNA"* OR *"gene"*

OR *"gene expression profiling"* OR *"genotype"* OR *"genomics"* OR *"genetics algorithms"* OR *"genetic algorithm"* OR *"biology"* OR *"bioinformatics"* OR *"proteomics"* OR *"genetics"* OR *"human"* OR *"humans"* OR *"animals"*) AND NOT TITLE-ABS-KEY (*"image"*) AND NOT TITLE-ABS-KEY (*"signal"* OR *"signal processing"* OR *"wireless"* OR *"audio"* OR *"sound"*) AND NOT TITLE-ABS-KEY (*"web"* OR *"text"*) AND NOT TITLE-ABS-KEY (*"game"* OR *"games"*) AND (LIMIT-TO (SRCTYPE , *"j"*) OR LIMIT-TO (SRCTYPE , *"p"*)) AND (LIMIT-TO (PUBSTAGE , *"final"*)) AND (LIMIT-TO (DOCTYPE , *"ar"*) OR LIMIT-TO (DOCTYPE , *"cp"*)) AND (LIMIT-TO (LANGUAGE , *"English"*)) .

Segundo Wazlawick (2009), o primeiro passo de uma revisão sistemática é listar os títulos de periódicos e eventos relevantes para o tema de pesquisa e os títulos de periódicos gerais em computação que eventualmente possam ter algum artigo na área do tema de pesquisa. Em seguida, deve-se obter a lista de todos os artigos publicados nos últimos cinco ou mais anos nesses veículos (WAZLAWICK, 2009). Como o processo de seleção e leitura dos artigos exige uma significativa quantidade de tempo, foram realizadas buscas em três datas distintas, a fim de manter a lista dos artigos o mais atualizada possível.

A primeira busca foi realizada em 27/04/2020 e retornou 403 artigos. Já a segunda busca aconteceu em 10/07/2020 devolvendo 418 artigos. Por fim, em 25/10/2020 a terceira e última busca foi concluída, entregando 432 artigos.

Após análise inicial dos resumos de todas as publicações selecionadas conforme anteriormente descrito, 314 artigos foram escolhidos para acesso completo. Contudo, problemas de ordem técnica não permitiram o download de alguns artigos, tais como: *links* com erro, *links* que direcionavam para documentos distintos do descrito no resultado, diferentes *links* que retornavam para o mesmo artigo, além de alguns periódicos que somente permitiam acesso ao texto integral do artigo mediante o pagamento de uma assinatura.

3.2 – Classificação dos artigos

Organizado o conjunto de artigos alvo de investigação, entram em cena o terceiro e quarto passos da revisão sistemática, conforme sugerido por Wazlawick (2009), quais sejam: a seleção dessa lista aqueles títulos que tenham relação com o tema de pesquisa e a leitura dos resumos desses artigos e, em função da leitura, classificá-los como relevância “alta”, “média” ou “baixa”.

O critério para definição de prioridade dos artigos foi o seguinte: os de prioridade alta foram aqueles que, de uma forma clara, tratavam de imputação *hot-deck*, realizando estudos comparativos que apresentavam uma considerável fundamentação teórica sobre esta classe de imputação. Já os de prioridade média realizam o processo de agrupamento antes da imputação, mas não necessariamente mencionavam o termo “*hot-deck*” ou apresentavam fundamentação teórica sobre a imputação *hot-deck*. Por fim, os artigos de baixa prioridade tratavam de imputação e agrupamento, sem abordar o processo de agrupamento antes da imputação. A partir destes critérios, dos 314 artigos baixados, foram classificados 72 artigos com prioridade alta, 78 artigos com prioridade média e 171 artigos com prioridade baixa. Os artigos de prioridade alta (Apêndice A) e média (Apêndice B) foram analisados sob a orientação de doze de perguntas de pesquisa formuladas com o propósito de alcançar o objetivo desta dissertação, com resultados quantificados em uma planilha eletrônica.

3.3 – Análise dos artigos

Foram estabelecidas cinco perspectivas de avaliação, com a finalidade de avaliar artigos de acordo com o objetivo de pesquisa desta dissertação: (i) ausência de dados, (ii) imputação de dados, (iii) agrupamento de dados; (iv) tipo de estudo e; (v) reprodutibilidade. Cada uma destas perspectivas é descrita nas seções a seguir.

3.3.1 – Ausência de Dados

A primeira perspectiva de avaliação está relacionada à consideração da ausência dos dados nos artigos, na qual estão incluídos os padrões de ausência e os mecanismos de ausência.

Ao realizar um estudo sobre imputação de dados, é importante que o pesquisador compreenda o padrão de ausência de seus dados, que podem ser classificados em duas categorias: (i) geral ou aleatório ou; (ii) específico. Com relação aos padrões de ausência específicos, estes podem ser divididos em duas subcategorias: (i) univariados ou (ii) monotônicos (SCHAFER; GRAHAM, 2002).

Uma correta escolha de técnicas de imputação, assim como a avaliação da eficácia de algum método, dependem da identificação dos mecanismos de ausência nas bases de dados (SCHAFER; GRAHAM, 2002). Segundo Little e Rubin (2002), a identificação do mecanismo de ausência é importante para compreender o processo que ocasiona a ausência de dados e conhecer a relação entre os dados ausentes e os valores subjacentes das variáveis do conjunto de dados.

Desta forma, os artigos foram avaliados no sentido de considerarem em seus estudos a identificação dos três tipos de classificações para os mecanismos de ausência, conforme estabelecido por Rubin (1976): completamente aleatório (*MCAR – Missing Completely At Random*), aleatório (*MAR – Missing At Random*) e não aleatório (*NMAR – Not Missing At Random*, ou *IM – Ignorable Missing*).

3.3.2 – Taxonomia de imputação de dados

Como segunda perspectiva de avaliação, esta dissertação propõe uma taxonomia de imputação que possui como base a proposta feita por Soares (2007). Ela parte do princípio que, quando um pesquisador possui uma base de dados com ausência, ele precisa escolher entre três abordagens iniciais: (i) remover os casos que apresentam ausência de dados; (ii) gerenciar diretamente a ausência de dados ou; (iii) realizar um processo de imputação (SOARES, 2007b).

Dentro da perspectiva de imputação, a realização da revisão sistemática foi realizada inicialmente a partir da taxonomia proposta por Soares (2007). Conforme os

artigos foram analisados, identificou-se a necessidade de evoluir a taxonomia proposta com o objetivo de contemplar as novas abordagens tecnológicas que surgiram ao longo dos anos. Desta forma, pode-se afirmar que a taxonomia proposta neste estudo é um produto da revisão sistemática realizada.

Com relação a remoção dos casos ausentes, foram consideradas três abordagens distintas: (i) remoção completa de casos (*Listwise Deletion* ou *Complete-Case Deletion*); (ii) remoção de pares (*Pairwise Deletion*), e; (iii) remoção de colunas com valores ausentes (MAGNANI; MONTESI, 2004).

Já o gerenciamento direto de casos pode ser feito através de um algoritmo robusto capaz de gerenciar os dados ausentes nas bases de dados a partir de sua identificação, tal como os algoritmos C4.5 (SALZBERG, 1994) e C5.0. (PANG; GONG, 2009). Neste caso, o método ou algoritmo consegue resultados satisfatórios sem a efetiva complementação dos valores ausentes tratados.

A terceira abordagem é onde ocorre efetivamente o processo de imputação, ou seja, a substituição de um valor ausente ou rejeitado por um valor “estimado” que seja viável, porém artificial (GELMAN; HILL, 2006). Logo, o processo de imputação pode ser inicialmente categorizado sob duas óticas: (i) tipos de imputação; e (ii) métodos de imputação.

Sob a ótica de tipos de imputação, os processos podem ser: (i) global, também chamados de imputação simples ou única, na qual apenas um único valor possível é gerado para cada campo ausente na base de dados (LITTLE; RUBIN, 2002); (ii) híbridos, nos quais o processo de imputação global é precedido da aplicação de um ou mais algoritmos ou técnicas, visando a melhoria do valor estimado (SOARES, 2007b); e (iii) *hot-deck*, na qual os valores ausentes de uma ou mais variáveis para um não respondente (chamado de destinatário) são substituídos por valores observados de um respondente (o doador), o qual é semelhante ao não respondente com relação às características observadas por ambos os casos (ANDRIDGE; LITTLE, 2010).

A imputação global pode ser realizada de duas formas: (i) baseada no atributo com valores ausentes e; (ii) baseada nos demais atributos. A imputação baseada no atributo com valores ausentes utiliza todos os elementos existentes nas demais tuplas para preencher os que estão ausentes, e podem utilizar métodos determinísticos (por exemplo, o cálculo da média) ou estocásticos, nos quais é introduzida uma perturbação no calculo da média (MAGNANI; MONTESI, 2004). Já a imputação baseada nos demais atributos realiza a estimação dos valores ausentes a partir da possível relação existente entre os atributos da amostra utilizando, por exemplo, uma regressão linear (BUSSAB;

MORETTIN, 2010) ou outros tipos de regressão (SOARES, 2007b).

A imputação híbrida contempla todos os métodos que combinam dois ou mais técnicas de imputação global, e nesta categoria temos como principais exemplos: (i) imputação *hot-deck* e *cold-deck* (BROWN; KROS, 2003b); (ii) imputação múltipla (RUBIN, 1978); imputação composta (SOARES, 2007b); e imputação em cascata (FERLIN, 2008).

Já sob a ótica do métodos de imputação, os processos de imputação podem ser: (i) estatísticos para a estimação dos valores ausentes, e (ii) de aprendizado de máquina, nos quais são utilizados algoritmos que fornecem aos computadores a habilidade de aprender e identificar padrões sem serem explicitamente programados (SAMUEL, 1959).

Os métodos estatísticos pode ser classificados em três categorias: (i) simples, a qual inclui o cálculos estatísticos de tendência central, tais como a média, moda entre outros e os variados modelos de regressão (BUSSAB; MORETTIN, 2010); (ii) verossimilhança, que buscam estimar os parâmetros de uma função de distribuição estatística com o objetivo de encontrar um modelo que represente o conjunto de dados (por exemplo, os algoritmos EM – *Expectation-Maximization* (EM) (DEMPSTER; LAIRD; RUBIN, 1977) e *Full Information Maximum Likelihood* (FIML) (ARBUCKLE, 2012; PETERS; ENDERS, 2002); e (iii) bayesiano, com redes bayesianas representadas por meio de grafos acíclicos direcionados (PEARL, 2014).

Os métodos de aprendizado podem ser classificados em três categorias: (i) supervisionado, pela identificação de padrões em dados que possuem rótulos e estabelecer previsões que ajudam a otimizar a abordagem; (ii) não supervisionado, no qual o algoritmo fica encarregado de identificar padrões com o objetivo de estabelecer rótulos para os dados; e (iii) de reforço, onde o computador é estimulado a aprender com base em tentativas e erros, otimizando o processo na prática direta.

Uma vez que o tema desta dissertação foi delimitado para a abordagem de imputação, as abordagens remoção de casos e gerenciamento de casos não foram considerados na avaliação dos artigos. Por fim, é apresentado na Figura 8, seguir a taxonomia de imputação proposta nesta dissertação.

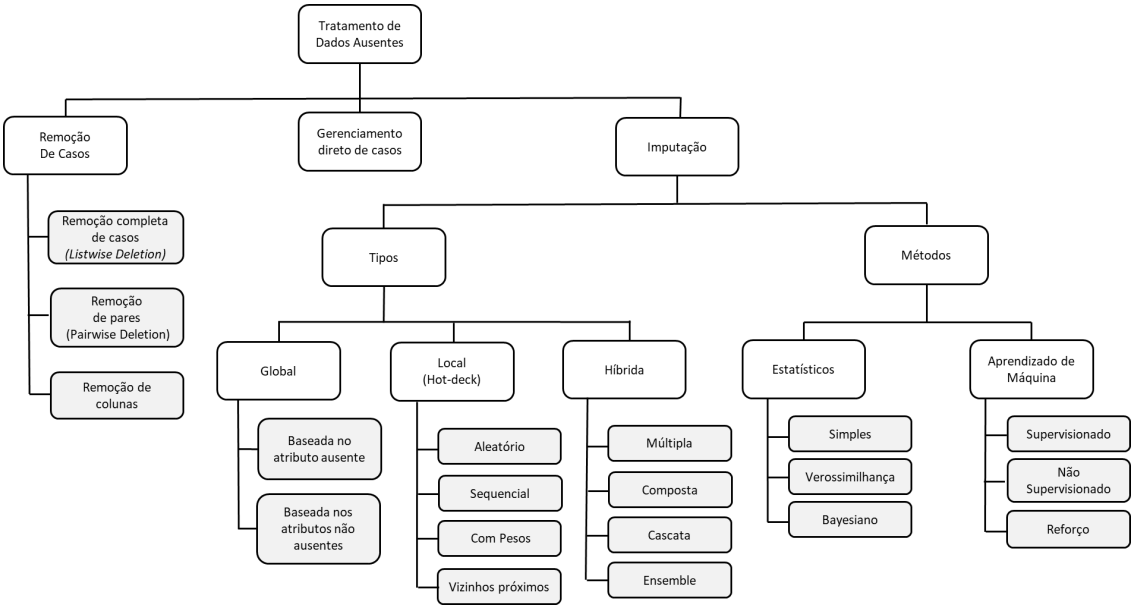

Figura 8 - Taxonomia de imputação proposta. Fonte: Elaborado pelo autor.

3.3.3 – Agrupamento de Dados

A terceira perspectiva de avaliação possui relação com a classificação dos algoritmos de agrupamento e foi realizada de acordo com a taxonomia definida por Pandey e Shukla (2019): (i) Agrupamento do tipo *Partitioning Based*; (ii) Algortimo do tipo *Hierarchical Based; (iii) Algoritmo do tipo Density Based;* (iv) Algortimo do tipo *Grid Based*; (v) Algortimo do tipo *Model Based;* (vi) Algoritmo do tipo *Fuzzy Based e;* (vii) Algoritmo do tipo *Graph Based.* A Figura 9 apresenta o esquema da taxonomia dos algoritmos de agrupamento destes autores.

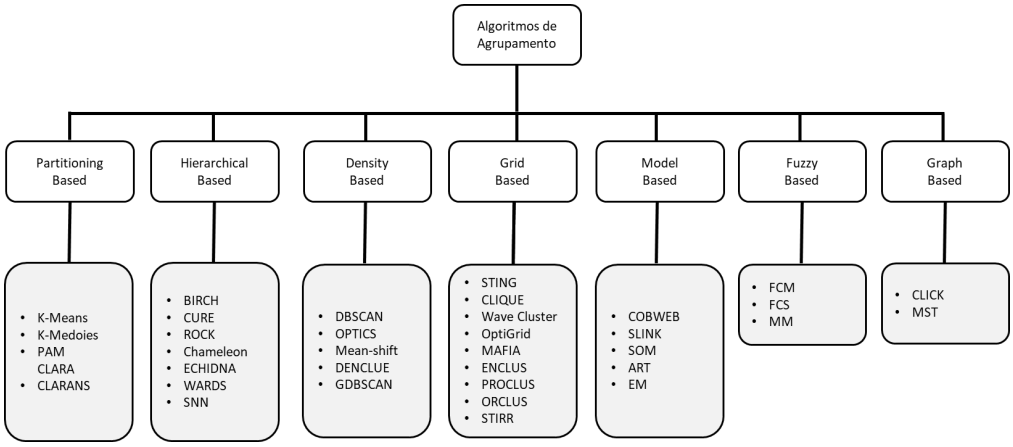

Figura 9 - Taxonomia dos algoritmos de agrupamento. Fonte: Adaptado de Pandey e Shukla (2019).

3.3.4 – Tipo de Estudo

Com o objetivo de avaliar o tipo de estudo dos artigos investigados, esta dissertação utiliza a taxonomia proposta por Carvalho et al. (2020) de acordo com o propósito de cada estudo. Ela é composta por três categorias de estudos: (i) descritivo, no qual o objetivo do estudo é analisar dados do passado e verificar as suas causas; (ii) preditivo, no qual o objetivo é tentar prever o que irá acontecer e; (iii) prescritivo, no qual o objetivo o objetivo é fornecer dicas (*insights*) para decisões sobre os eventos que estão acontecendo (CARVALHO et al., 2020).

A análise descritiva é mais adequada quando o objetivo principal é saber o que aconteceu no passado. Este tipo de análise se inicia com a coleta de dados que podem ter origem em muitas fontes de dados, incluindo pesquisas, entre outras, e produz conclusões sobre esses dados. É uma fotografia do que ocorreu, a qual pode apontar para eventos que não estavam claros anteriormente (DAVENPORT; HARRIS, 2017).

Já o estudo preditivo possui o objetivo de, a partir dos dados existentes, utilizar a ciência da computação para antecipar possibilidades de acontecimentos no futuro. A previsão ganhou força nos últimos anos com o surgimento de novos métodos de aprendizagem de máquina que proporcionaram maior precisão nos resultados. O objetivo do estudo preditivo é construir modelos a partir de grandes volumes de dados (SINGH; REDDY, 2015).

A análise prescritiva é a aplicação de métodos de computação para otimizar as decisões em cada situação. É considerado o nível mais desafiador da área de análise de dados e, consequentemente, o estudo que proporciona os benefícios mais significativos. Ferramentas de análise prescritiva possibilitam a redução de riscos na tomada de decisões, proporcionam maior agilidade neste processo, além de diminuir o tempo gasto com tais diagnósticos (TSAI et al., 2015).

3.3.5 – Reprodutibilidade

A possibilidade de o leitor conseguir reproduzir todos os resultados apresentados nos artigos é algo importante para o método científico (MUNAFÒ et al., 2017). Por isso, a reprodutibilidade é uma questão fundamental para a evolução de área de pesquisa e

compõe uma perspectiva de avaliação desta dissertação.

O sucesso da reprodutibilidade ocorre quando os autores fornecem os conjuntos de dados, métodos, algoritmos ou código-fonte que permitam aos leitores reproduzir a avaliação experimental de artigos publicados. Além disso, uma abordagem comparativa que combina dados públicos, repositórios abertos de código fonte podem aumentar a reprodutibilidade dos artigos publicados.

Desta forma, esta perspectiva realiza a avaliação dos artigos considerando as seguintes questões: (i) O artigo utiliza um conjunto de dados público? (ii) O artigo apresenta o pseudocódigo de seus algoritmos? (iii) O artigo apresenta um repositório aberto de código-fonte? (iv) O artigo compara a acurácia dos valores estimados para os dados ausentes com os dados originais?

4- Resultados

Esta seção apresenta os resultados obtidos a partir da revisão sistemática proposta no Capítulo 3. Assim, na primeira parte deste capítulo, é apresentado o perfil das publicações, contendo sua evolução quantitativa ao longo dos anos, seus tipos, os periódicos que mais trataram a imputação *hot-deck*, bem como os autores mais citados. A segunda parte descreve a análise dos artigos com base na perspectiva de ausência de dados considerando os padrões e mecanismos de ausência. Já a terceira parte apresenta os resultados da análise dos artigos a partir da taxonomia de imputação proposta neste estudo. Com relação à taxonomia de agrupamento adotada nesta dissertação, os resultados obtidos a partir de sua utilização estão descritos na quarta parte deste capítulo. A quinta seção apresenta os resultados obtidos de acordo com a perspectiva de tipo de estudo. A avaliação com relação à reprodutibilidade dos experimentos está exposta na sexta parte deste capítulo. No fim, a sétima seção resume o conhecimento encontrado através da revisão sistemática.

4.1 – Perfil das Publicações

A revisão sistemática identificou 150 publicações entre os anos de 1986 e 2020. A Figura 10 apresenta a quantidade de publicações por ano, na qual é possível perceber um crescimento quantitativo ao longo nos anos.

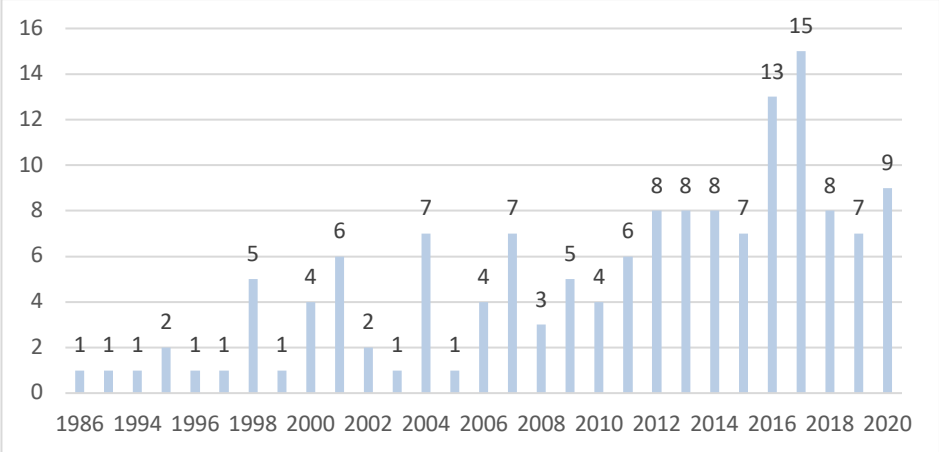

Figura 10 - Publicações por ano. Fonte: Scopus, 2020.

Distinguindo o tipo de publicação, ou seja, se são artigos de periódicos ou de conferência, a Figura 11 apresenta a quantidade de publicações por ano, a qual indica um crescimento quantitativo nos trabalhos publicados em conferências a partir do ano de 2006, isto é, se as apresentações de pesquisas nas conferências estão contribuindo cientificamente para a comunidade que se dedica a estudar o assunto.

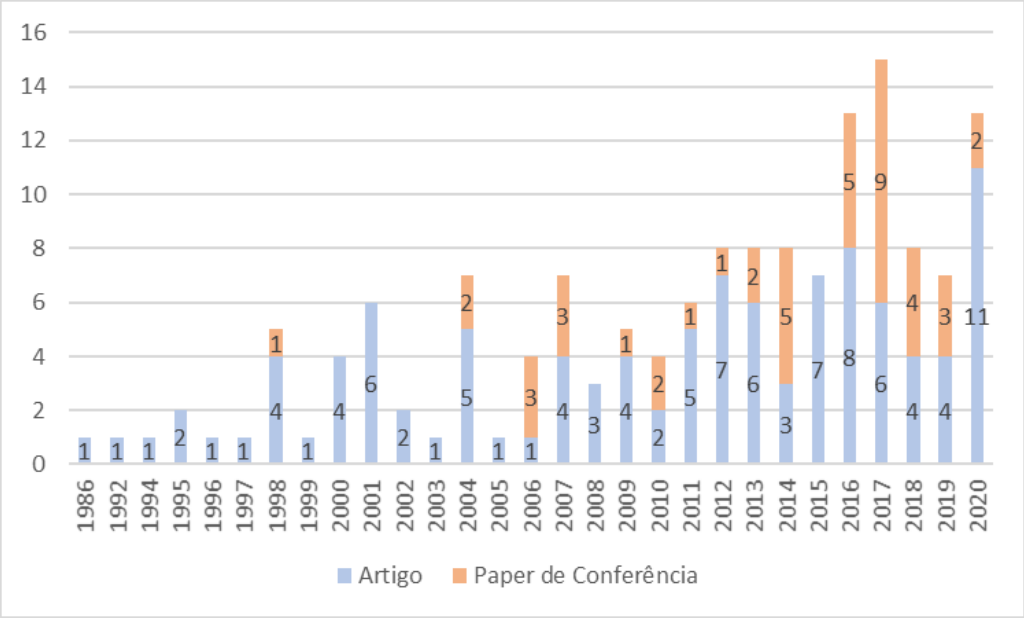

Figura 11 - Quantidade de publicações de artigos ou *papers* de conferência por ano.
Fonte: Scopus, 2020.

Sobre os periódicos que apresentaram o maior quantitativo de publicações relacionados a este tema de pesquisa, alvo da primeira pergunta de pesquisa: "**Quais são os principais periódicos que publicam artigos sobre imputação *hot-deck*?**", o periódico *"Journal of the American Statistical Association"* liderou o ranking com nove publicações, seguido do *"Computational Statistics and Data Analysis"* com seis publicações. Já os periódicos *"Biometrika"*, *"Canadian Journal of Statistics"*, *"Procedings of The International Society for Optical Engineering"* e *"Journal of Statistical Planning and Inference"* ficaram em terceiro lugar com três publicações cada. A Figura 12 apresenta o ranking dos dezoito primeiros periódicos, o que representa os periódicos que possuem mais de duas publicações. Logo, existe uma relevante contribuição de resultados de pesquisa na área de imputação *hot-deck* por meio dos periódicos de estatística. Entretanto, este tema é muito pulverizado entre diversos periódicos, pois vários periódicos apresentaram apenas uma publicação sobre este tema.

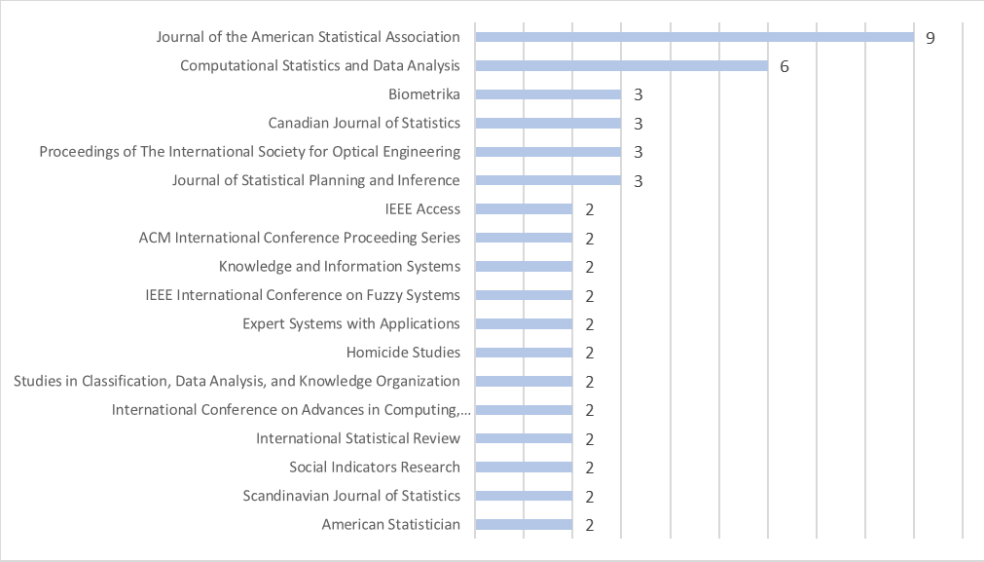

Figura 12 - Ranking mais frequentes de periódicos por quantidade de publicações.
Fonte: Scopus, 2020.

Com relação à quantidade de citações, a Tabela 2 apresenta o ranking dos 10 mais citados, o qual representa os autores que tiveram mais de 100 citações até a presente data. Nos parágrafos a seguir, são descritas as contribuições de alguns autores dentro da ótica da técnica de imputação *hot-deck*.

Tabela 2 - Ranking (10 primeiros) das publicações por quantidade de citações.

Ranking	Autores	Quant. de Citações
1	(MYERS, 2011)	245
2	(ROTH; SWITZER; SWITZER, 1999)	208
3	(REILLY; PEPE, 1995)	164
4	(STRIKE; EL EMAM; MADHAVJI, 2001)	158
5	(GOLD; BENTLER, 2000)	152
6	(SCHÄFER et al., 2012)	146
7	(SCHENKER; TAYLOR, 1996)	143
8	(RAO; SHAO, 1992)	142
9	(HUISMAN, 2000)	135
10	(LI et al., 2004)	107

Fonte: Scopus, 2020.

Segundo Myers (2011), o qual recebeu 245 citações, dados ausentes são um problema recorrente em pesquisas quantitativas na área de comunicação; no entanto,

as práticas de manipulação de dados ausentes encontrados na maioria dos trabalhos publicados nesta área deixam muito espaço para melhorias.

Este autor sugere a imputação *hot-deck* como uma solução prática para muitos problemas de dados ausentes, pois, embora apresente limitações, a técnica possui vantagens sobre os métodos de exclusão completa de casos (*Listwise Deletion*). Primeiramente, os procedimentos de *hot-deck* permitem a retenção da amostra completa de indivíduos, evitando a perda de casos incompletos e os subsequentes declínios no poder estatístico que são incorridos no resultado. Além disso, outras vantagens apontadas pelo autor são: (i) as imputações tendem a ser realistas, uma vez que têm como base valores observados em outro lugar; (i) as imputações não estarão fora da faixa de valores possíveis (uma possibilidade em imputações múltiplas); e (iii) a técnica *hot-deck* não exige a definição de um modelo explícito para a distribuição dos valores ausentes (por exemplo: distribuição normal, distribuição logística).

Myers (2011) apresenta uma relevante discussão sobre os padrões e percentuais de ausência, descrevendo as situações onde a aplicação da abordagem *hot-deck* é recomendada, conforme apresentado no Quadro 1, a seguir.

Quadro 1 - Alcance da aplicabilidade da abordagem hot-deck.

	Padrões de ausência		
Porcentagem de ausência de dados	**MCAR**	**MAR**	**MNAR**
1 até 5%	*Hot-deck recomendado*		
6 até10%			
11 até15%			
16 até 20%			Utilize *Maximum Likehood, Expectation Maximization* ou Imputação Múltipla

Fonte: Adaptado de Myers (2011).

Roth, Switzer e Switzer (1999), a partir dos resultados de seus experimentos, recomendam que os pesquisadores evitem utilizar a exclusão completa de casos, uma vez que este método pode eliminar uma grande quantidade de casos reais, criando um viés para a amostra. Os autores também recomendam que não seja utilizada a

substituição dos valores ausentes a partir do cálculo da média, uma vez que esta abordagem ignora a diferenças individuais e reduz a variabilidade da amostra.

De acordo com os resultados dos experimentos, observou-se que as técnicas de regressão e *hot-deck* aplicadas aos casos ausentes apresentaram os melhores resultados, uma vez que reconhecem as diferenças individuais entre os sujeitos no processo de imputação (ROTH; SWITZER; SWITZER, 1999). Uma vantagem da imputação *hot-deck* é ela reproduzir com maior precisão a variabilidade (por exemplo, desvio padrão) dentro das variáveis. Esta técnica reproduz melhor a variância porque as pontuações que ela toma emprestado das respostas reais incluem um componente de erro dessas pontuações. Já as abordagens de regressão com nenhum erro inserido podem não incluir essa proporção de variação em sua estimativa de dados ausentes, mas realizam um bom trabalho ajudando a reproduzir muitas estimativas de covariância.

Strike, El Emam e Madhavi (2001) apresentaram uma extensa simulação onde foram avaliadas diferentes técnicas para o tratamento de dados ausentes no contexto de modelagem de custo de software. As técnicas avaliadas foram a exclusão completa de casos, imputação através do cálculo de média e oito variações de imputação utilizando *hot-deck*. De acordo com os resultados dos pesquisadores, o melhor desempenho (mínimo viés e maior precisão) pode ser obtido através da imputação *hot-deck* com distância euclidiana e uma padronização Z-score.

Gold e Bentler (2000) realizaram estudos experimentais comparando o desempenho entre as técnicas de imputação *hot-deck*, regressão estocástica e Expectation Maximization. Sobre um mesmo conjunto de dados, foram simuladas ausências aleatórias com variações no tamanho da amostra, características de distribuição e proporção de dados excluídos. No que diz respeito à técnica *hot-deck*, os autores recomendaram a sua utilização em casos onde o tamanho da amostra for pequeno e a proporção de dados ausentes estiver entre 8% e 16%.

Uma relevante característica dos estudos experimentais sobre algoritmos de imputação é a realização de estudos comparativos, ou seja, estudos que confrontam o desempenho entre diferentes algoritmos sobre distintos conjuntos de dados. Estudo comparativos apresentam consideráveis contribuições acadêmicas na área de imputação. Logo, as Figuras 4.5 e 4.6 apresentam os resultados relacionados à segunda pergunta de pesquisa: *"***Quantos estudos comparativos com a utilização de *hot-deck* foram realizados?***"* De acordo com a Figura 13, 87 artigos avaliados, o que corresponde a 58%, realizaram estudos comparativos.

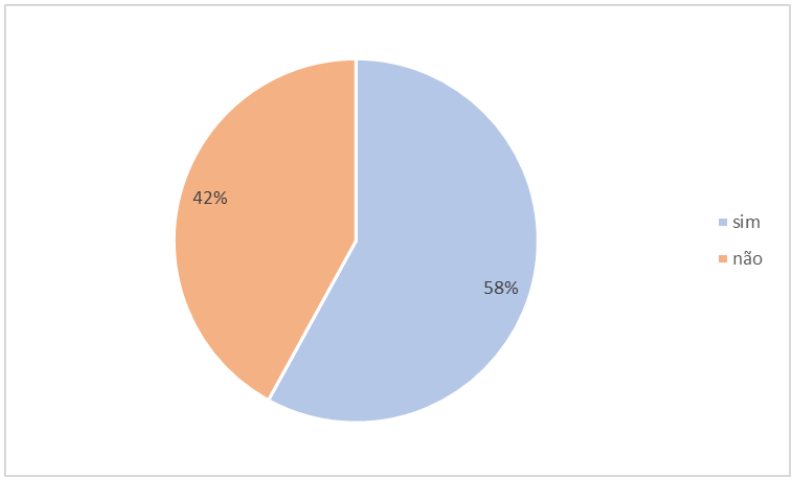

Figura 13 - Proporção de estudos comparativos. Nota: n =150. Fonte: Scopus, 2020.

A partir da Figura 14, a seguir, podemos observar ao longo dos anos um crescimento na quantidade de estudos comparativos nesta área de pesquisa.

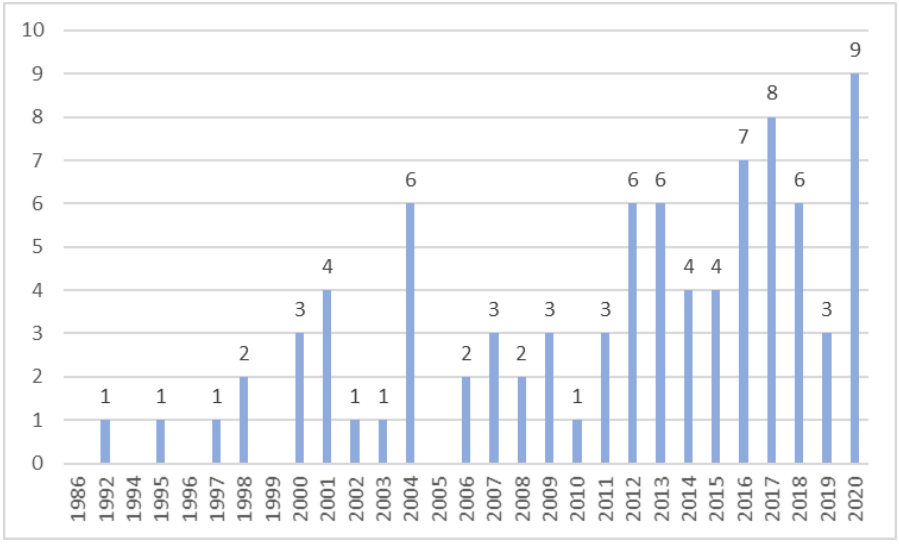

Figura 14 - Quantidade de estudos comparativos por ano. Fonte: Scopus, 2020.

4.2 – Perspectiva de ausência de dados.

Tratando-se da perspectiva de ausência de dados, objeto da terceira pergunta de pesquisa: *"***Os padrões de ausências são adequadamente identificados nos estudos experimentais?***"*, no conjunto de artigos avaliados, somente o artigo "*Software Cost Estimation with Incomplete Data*" dos autores Strike, El Emam e Madhavi (2001)

realizou os experimentos simulando os padrões de ausência univariado e monotônico.

A análise a seguir focaliza os mecanismos de ausência de dados dos conjuntos utilizados nos experimentos, questões abordadas na quarta pergunta de pesquisa: *"***Os mecanismos de ausências são adequadamente identificados nos estudos experimentais?***"*. De acordo com a Figura 15, somente 37% dos estudos realizaram esta avaliação. Logo, a importante recomendação de Little e Rubin (2002) sobre a identificação dos mecanismos de ausência não foi considerado por 63% dos artigos investigados.

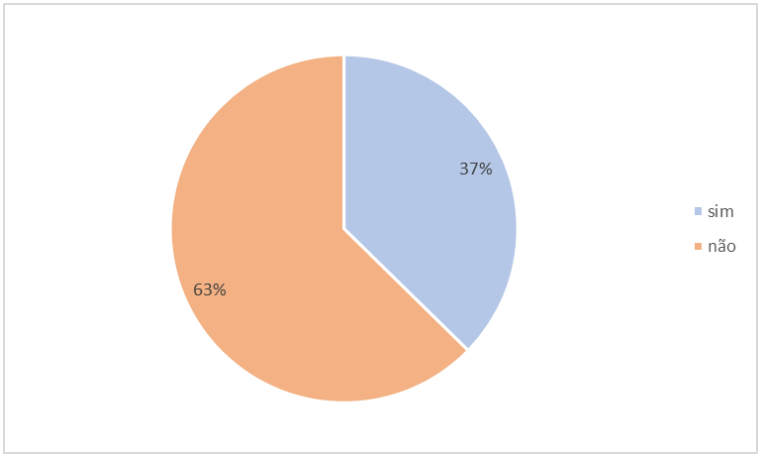

Figura 15 - Proporção dos artigos que avaliaram os mecanismos de ausência nos datasets utilizados nos experimentos. Nota: n =150. Fonte: Scopus, 2020.

A Figura 16 apresenta os artigos que avaliaram e os que não analisaram os mecanismos de ausência ao longo dos anos. A partir desta figura podemos observar que, apesar de estar ocorrendo um crescimento quantitativo de estudos sobre imputação *hot-deck*, uma considerável parte dos artigos que compõem este crescimento não realizou a identificação dos mecanismos de ausência, indicando uma necessidade de melhoria qualitativa sobre estes estudos.

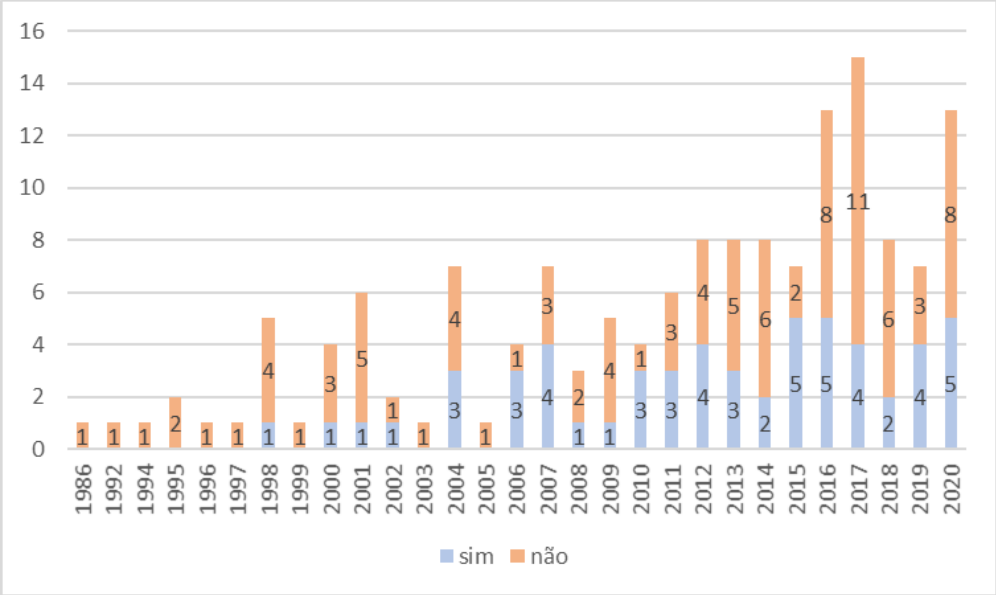

Figura 16 - Quantidade de artigos que identificaram os mecanismos de ausência nos datasets utilizados nos experimentos por ano. Fonte: Scopus, 2020.

Ao realizar uma observação mais detalhada, considerando apenas os artigos que identificaram mecanismos de ausência, a Figura 17 indica que 64% dos artigos realizam experimentos apenas considerando um único mecanismo (MAR ou MCAR ou NMAR), 15% consideram dois mecanismos de ausência (MCAR e MAR ou MCAR e NMAR), e 21% consideraram os três mecanismos de ausência (MAR, MCAR e NMAR). Este fato expõe uma necessidade de novos estudos que considerem ao menos mais de um mecanismo de ausência em seus experimentos.

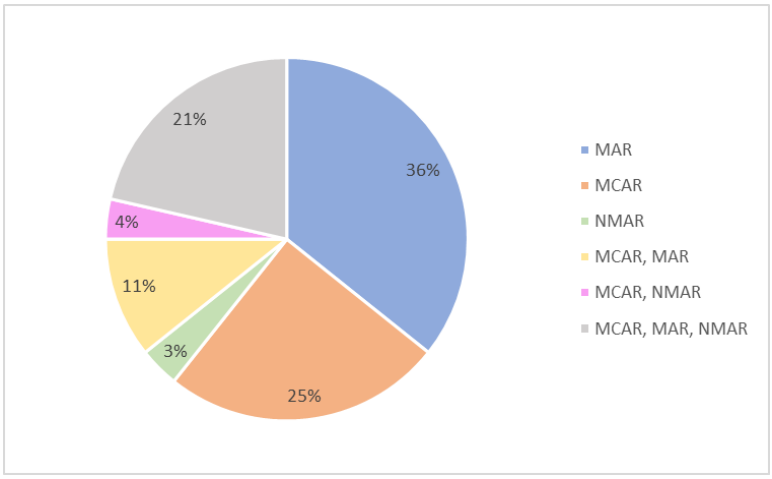

Figura 17 - Proporção dos mecanismos de ausência avaliados nos conjuntos de dados. Nota: n =150. Fonte: Scopus, 2020.

Ao observar as avaliações de mecanismos de ausência ao longo dos anos (Figura 18), é possível perceber um pequeno aumento dos estudos que identificam os três mecanismos de ausência a partir do ano de 2014.

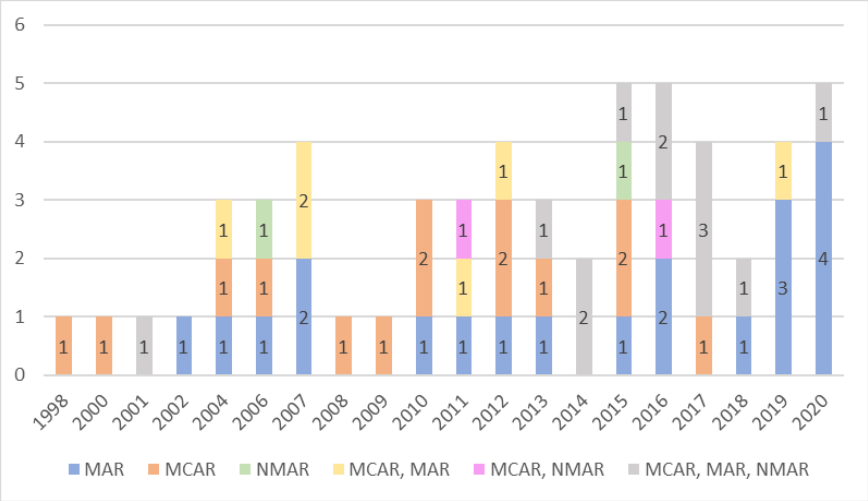

Figura 18 - Quantidade de mecanismos de ausência avaliados por ano. Fonte: Scopus, 2020.

4.3 – Perspectiva de imputação

As Figuras a seguir descrevem os resultados da avaliação dos artigos de acordo com a taxonomia de imputação proposta. A Figura 19 apresenta o quantitativo de artigos para tipos de imputação global e híbrido. Com relação à quinta pergunta de pesquisa: *"***Ao longo dos anos está ocorrendo uma maior utilização de *hot-deck* em estudos de imputação híbrida?***"*, é possível observar um aumento nos estudos híbridos que utilizam imputação *hot-deck* a partir do ano 2001.

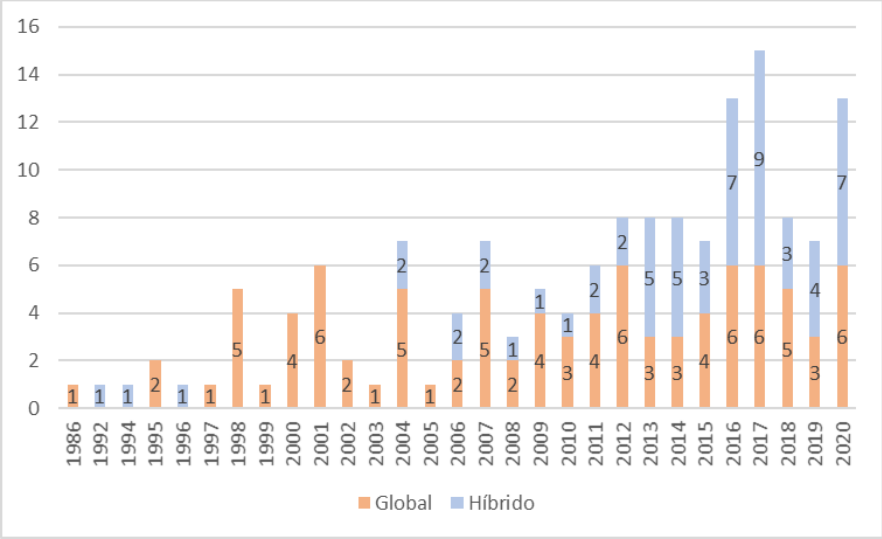

Figura 19 - Quantidade de artigos por tipo de imputação por ano. Fonte: Scopus, 2020.

Dentro da categoria imputação global, os artigos que realizaram a imputação baseada nos valores do atributo que apresenta ausência utilizaram, em maior proporção, o cálculo da média para atributos contínuos e a moda para atributos categóricos. Somente um artigo calculou a mediana do atributo categórico. Acerca dos artigos que realizaram a imputação baseada nos demais atributos, destacam-se as regressões linear e logística. A Figura 20 apresenta o resultado para estas categorias, na qual se observa um equilíbrio entre a utilização das duas categorias de imputação global.

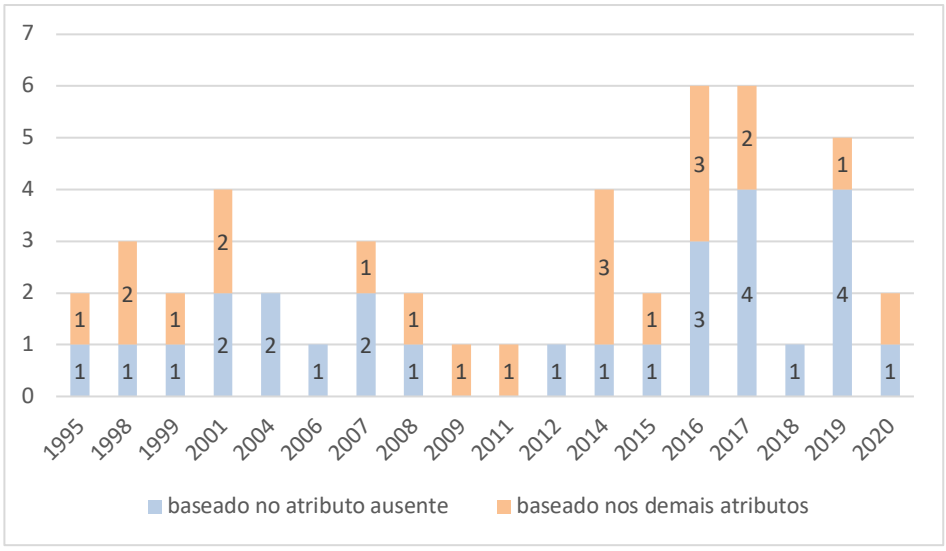

Figura 20 - Quantidade de artigos que utilizaram a imputação global baseada no atributo ausente ou baseada nos demais atributos por ano. Fonte: Scopus, 2020.

Estabelecer um critério para a classificação dos métodos incluídos na categoria imputação híbrida consiste em um desafio devido à grande diversidade de combinações de diferentes técnicas. Neste contexto, a imputação múltipla realizada através de algoritmos de agrupamento é o objeto da sexta pergunta de pesquisa: *"***Quantos estudos de imputação híbrida utilizaram a técnica *hot-deck*? E destes estudos, quantos realizam a Imputação Múltipla?***"*. De acordo com a Figura 21, foram identificados 59 artigos que utilizaram a técnica *hot-deck*, o que corresponde a 39% dos artigos investigados.

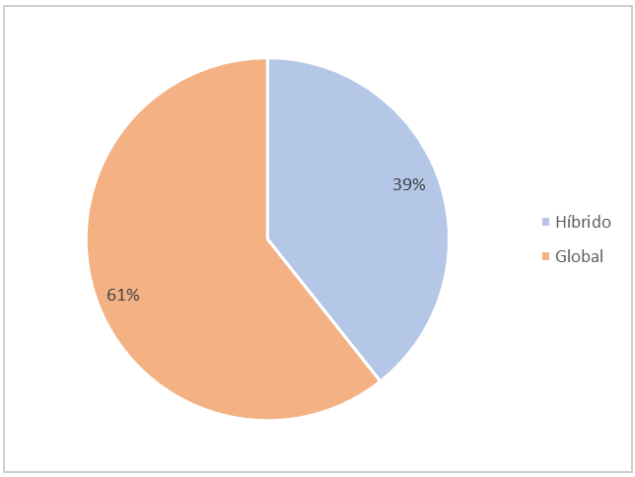

Figura 21 - Proporção dos artigos por tipo de imputação. Nota: n = 150. Fonte: Scopus, 2020.

Dentre estes 59 artigos, 39 (66%) artigos utilizaram métodos híbridos diversos e 20 artigos realizaram a imputação múltipla utilizando algoritmos de agrupamento, conforme apresentado na Figura 22.

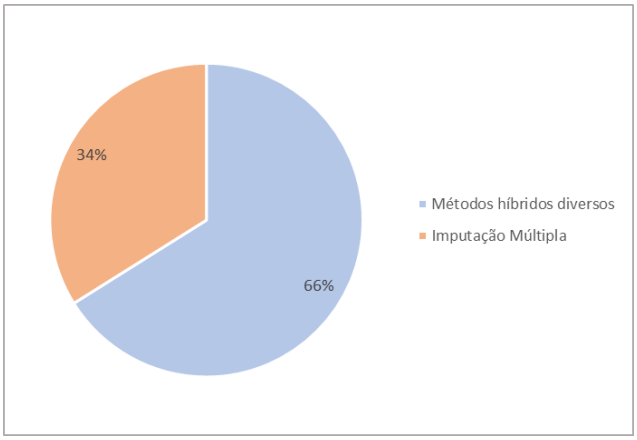

Figura 22 - Proporção dos artigos que realizam a imputação múltipla. Nota: n = 59 Fonte: Scopus, 2020.

Os algoritmos utilizados na imputação múltipla são apresentados na tabela 3, a qual apresenta uma diversidade de técnicas utilizadas.

Tabela 3 - Lista de autores que realizam experimentos com a Imputação Múltipla e respectivos algoritmos de agrupamento utilizados.

Autores	**Algoritmos utilizados na Imputação Múltipla**
(SILVA-RAMÍREZ; PINO-MEJÍAS; LÓPEZ-COELLO, 2015)	*Artificial neural networks*
(DING; ROSS, 2010)	*Expectation–maximization*
(ZHANG; FANG, 2016) (NIKFALAZAR et al., 2017)	*Fuzzy C-Means*
(HEITJAN; LANDIS, 1994) (DURRANT, 2009) (GOMES et al., 2013) (SULLIVAN; ANDRIDGE, 2015) (SUN et al., 2020)	*Hot-Deck Bayesian Bootstrap*
(PENNY; ASHRAF; DUFFY, 2007) *(ZHANG et al., 2017a)*	*Hot-Deck Multiple Imputation*
(PAIK; LARSEN, 2013) (STEINER et al., 2016)	*Markov Chain Monte Carlo*
(YANG et al., 2019)	*Multiple Imputation by Chained Equations*
(ALBAYRAK; TURHAN; KURT, 2017)	*Maximum Likelihood Estimation*
(KANG; KOEHLER; LARSEN, 2012)	*Multinominal Model*
(SCHENKER; TAYLOR, 1996) (ZANDBERG; HUISMAN, 2019)	*Predictive Mean Matching*
(LORENZO-SEVA; VAN GINKEL, 2016)	Predictive Mean Matching, *Hot-Deck Multiple Imputation*
(KHUSNULKHOTIMAH; SUPRAJITNO, 2018)	*Self-organized map*

Fonte: Scopus, 2020.

Ao avaliar os artigos através das categorias de métodos de imputação, dentro da categoria métodos estatísticos (Figura 23), os simples (os quais incluem média, moda, mediana e regressões) correspondem a 71% de artigos avaliados; já os bayesianos correspondem a 17% dos artigos, enquanto os de verossimilhança correspondem a apenas 12% dos artigos avaliados. Aparentemente, os métodos estatísticos simples ainda são muito utilizados devido a sua simplicidade de cálculo. No entanto, de acordo com Soares (2007), estes métodos podem gerar viés ou reduzir a diversidade da amostra.

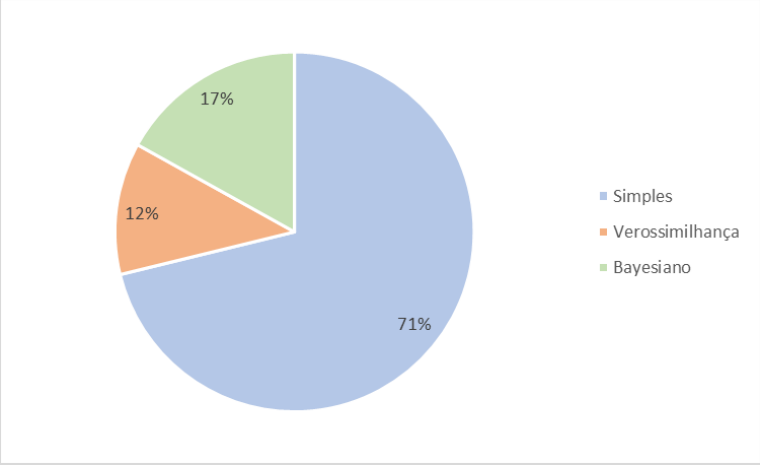

Figura 23 - Proporção de artigos que utilizam métodos estatísticos. Nota: n=59. Fonte: Scopus, 2020.

A Figura 24 apresenta os métodos estatísticos utilizados ao longo dos anos. De acordo com esta figura, observa-se a presença dos métodos bayesianos a partir do ano 2004 e dos de verossimilhança a partir do ano de 2010.

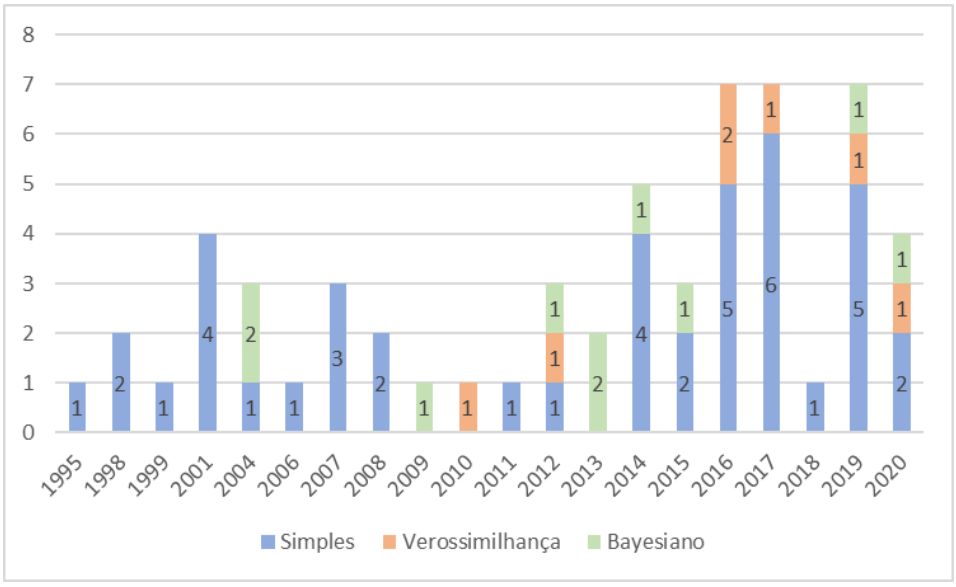

Figura 24 - Quantidade de artigos que utilizam métodos estatísticos por ano. Fonte: Scopus, 2020.

Com o objetivo de responder à sétima pergunta de pesquisa: *"***Quantos estudos de imputação híbrida utilizaram métodos estatísticos?***"*, as perspectivas tipos de imputação e métodos de imputação foram combinadas. Os métodos estatísticos simples apresentaram o maior percentual de utilização com 76%, conforme Figura 25. Ao observar o gráfico ao longo dos anos (Figura 26), percebe-se que os métodos estatísticos simples prevalecem no universo da imputação global.

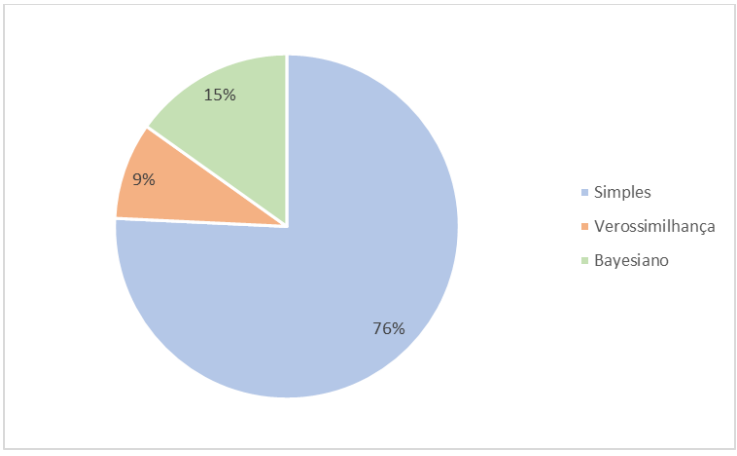

Figura 25 - Métodos Estatísticos utilizados nos estudos que realizam imputação global. Nota: n = 33. Fonte: Scopus

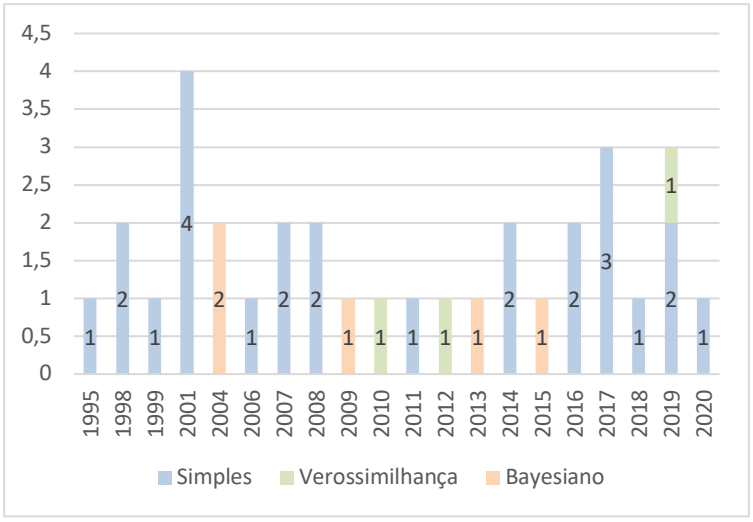

Figura 26 - Métodos Estatísticos utilizados nos estudos que realizam imputação global ao longo dos anos. Fonte: Scopus

Considerando apenas os artigos que realizaram imputação do tipo híbrida, a resposta para a sétima pergunta de pesquisa é o total de 26 artigos que utilizaram métodos estatísticos. De acordo com a Figura 27, 17 artigos (66%) com a abordagem de imputação híbrida utilizaram métodos estatísticos simples.

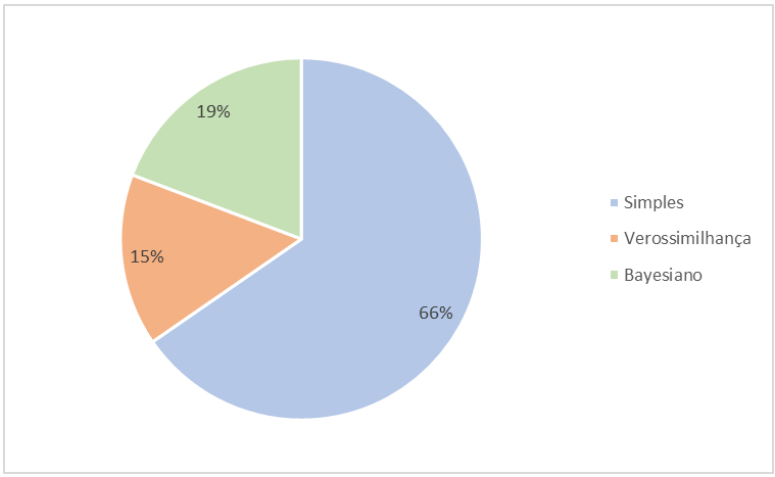

Figura 27 - Métodos Estatísticos utilizados nos estudos que realizam imputação híbrida. Nota: n = 26. Fonte: Scopus

Ao observar o gráfico ao longo dos anos (Figura 28), percebe-se que os métodos estatísticos simples também prevalecem no universo da imputação híbrida.

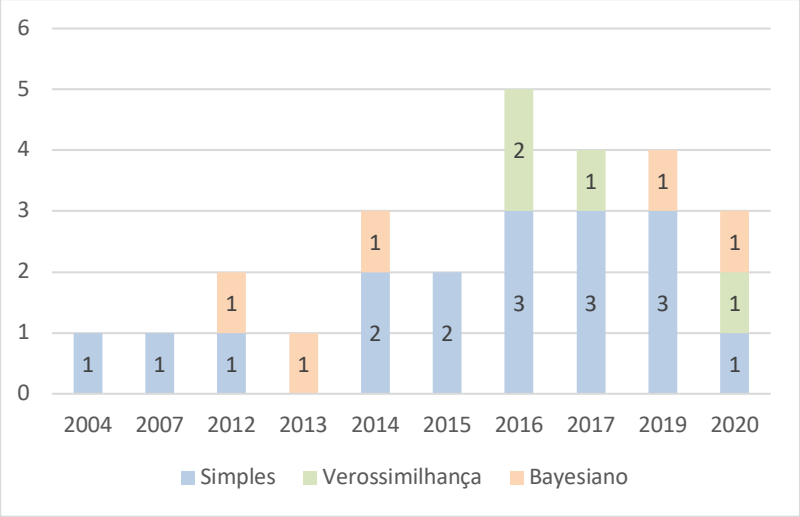

Figura 28 - Métodos Estatísticos utilizados nos estudos que realizam imputação híbrida ao longo dos anos. Fonte: Scopus

No que concerne ao aprendizado de máquina, a Figura 29 apresenta a sua utilização na imputação *hot-deck* a partir do ano 2000. No contexto deste tipo de imputação, os algoritmos de aprendizado supervisionado foram os preferidos, sendo o *K-Nearest Neighbor* utilizado na maioria dos casos. Referente aos algoritmos de aprendizado não supervisionado, a maioria dos casos utilizou redes neurais artificiais. Não foi possível identificar trabalhos que optaram por algoritmos de aprendizado por reforço no contexto da imputação *hot-deck*.

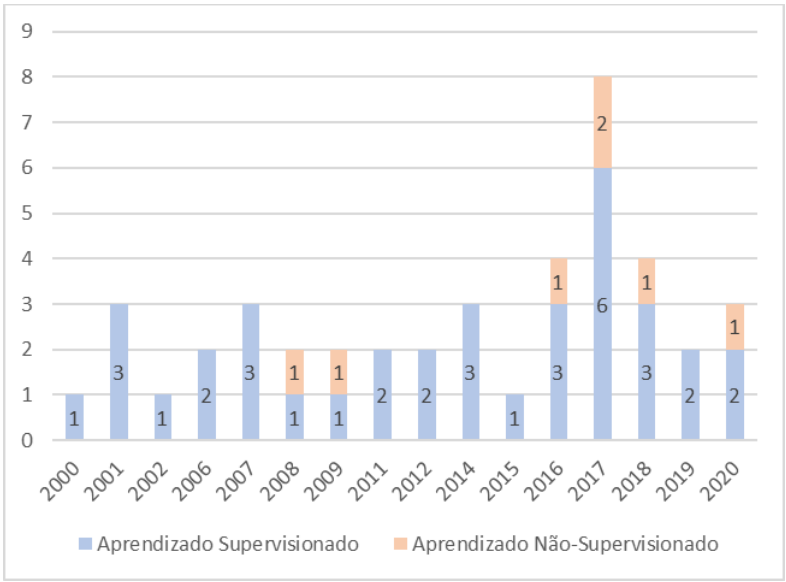

Figura 29 - Quantidade de artigos que utilizam algoritmos de aprendizado de máquina por ano. Fonte: Scopus, 2020.

4.4 – Perspectiva de agrupamento.

A perspectiva de agrupamento avalia os algoritmos utilizados para realizar o agrupamento que precede a imputação *hot-deck*. A avaliação dos artigos utiliza a taxonomia de Kamlesh e Diwakar (2019) para classificar os algoritmos de agrupamento.

A Figura 30 apresenta a proporção dos artigos de acordo com cada categoria com o propósito de responder a oitava pergunta de pesquisa*: "***Qual categoria de algoritmos de agrupamento é a mais utilizada na imputação *hot-deck*?**". Dos 72% dos Algoritmos utlizados, uma considerável proporção é do tipo *Partitioning Based*. Outras categorias possuem pequena participação nos estudos experimentais, tais como *Fuzzy Based* com 12%, *Model Based* com 10%, *Hierarquical Based* e *Density Based* com 3% cada uma. A revisão sistemática realizada não encontrou artigos que utilizassem algoritmos do tipo *Grid Based* ou *Graph Based*.

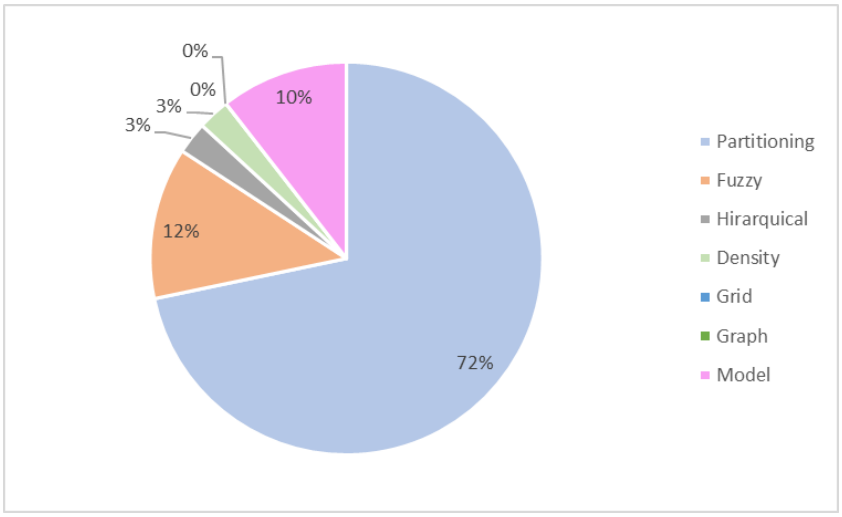

Figura 30 - Proporção de artigos por tipo de algoritmo de agrupamento. Nota: n = 150.
Fonte: Scopus, 2020.

A Figura 31 apresenta o quantitativo de artigos ao longo dos anos. Ela apresenta a predominância dos algoritmos do tipo *Partitioning Based*. Apesar dos algoritmos do tipo *Partitioning Based* serem os mais representativos, é possível observar um discreto aumento na diversidade de tipos a partir do ano de 2001.

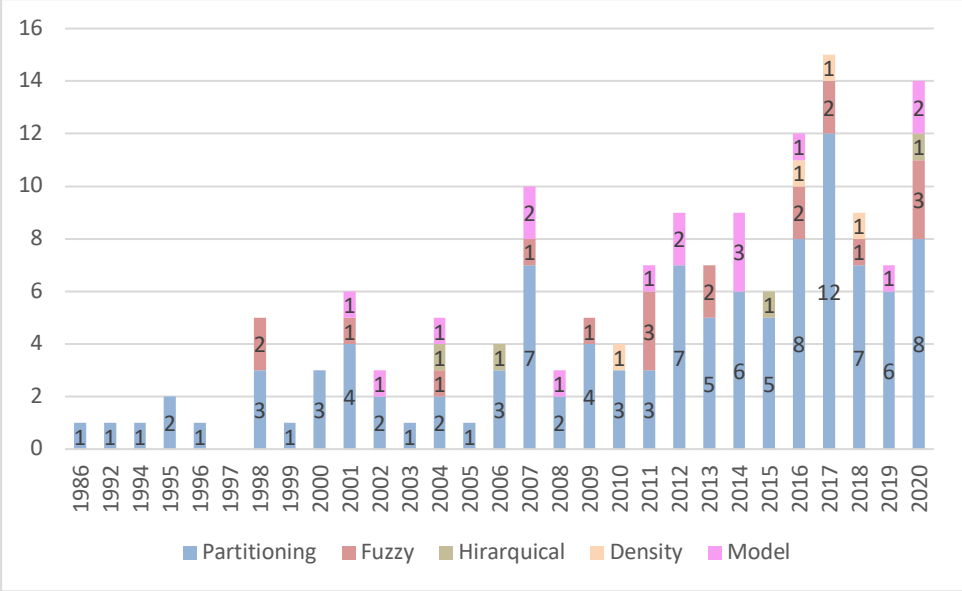

Figura 31 - Quantidade de artigos por tipo de agrupamento por ano. Fonte: Scopus, 2020.

Com relação à pergunta de pesquisa: "**Quais são os principais algoritmos utilizados na etapa de agrupamento que precede a imputação?***",* ao observar o universo dos algoritmos do tipo *Partitioning Based*, os do tipo *Random hot-deck* (RHD), *K-Nearest-Neighbor* (KNN) e *K-means* correspondem a 75% dos algoritmos utilizados nos estudos experimentais. A Figura 32 apresenta a proporção dos algoritmos de agrupamento por categoria.

Sobre a categoria do tipo *Fuzzy Based*, esta foi representada pelos algoritmos *Fuzzy C-Means* (FCM) com 89% e *Fuzzy K-Means* (FKM) com 11%. Já a categoria *Hierarquical Based* foi composta por três algoritmos de agrupamento hierárquicos: (i) *Hierarchical Agglomerative Clustering* (HAC); (ii) *Bayesian Hierarquical Model* (BHM) e *Bayesian Network* (BN).

De acordo com a Figura 32, os algoritmos *Self-Organized Map* (SOM), *Density-Based Spatial Clustering and Application with Noise* (DBSCAN) e *Gaussian Mixture Model* (GMM) compõem a categoria de agrupamento *Density Based.*

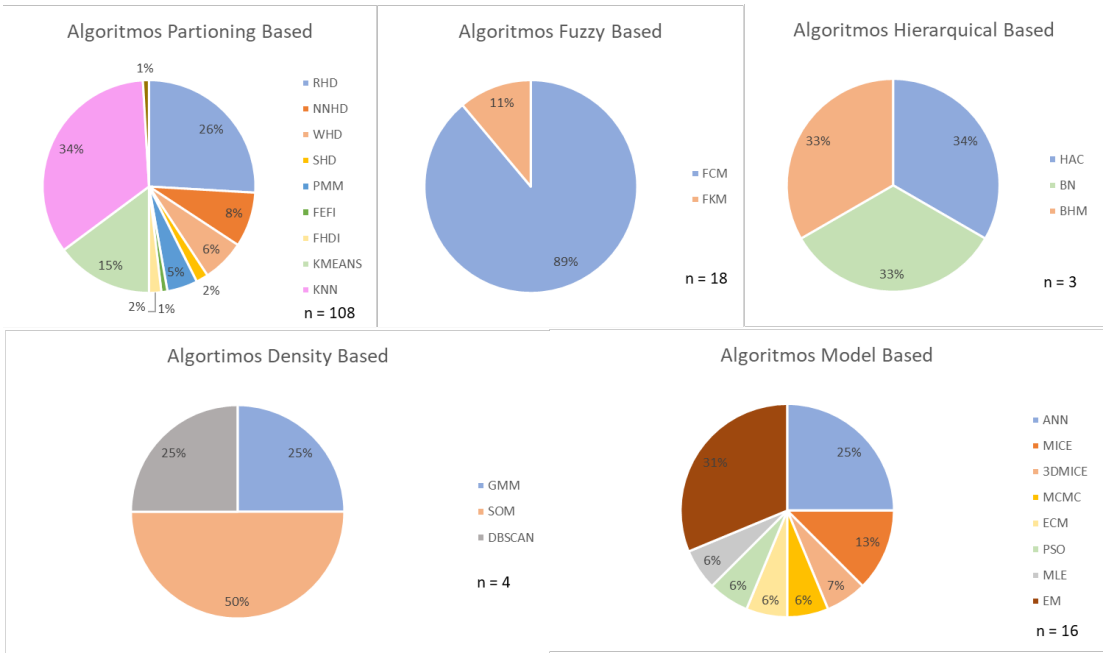

Figura 32 - Proporção dos algoritmos de agrupamentos por categoria. Fonte: Scopus, 2020.

Uma pequena quantidade de algoritmos de agrupamento do tipo *Model Based* foi encontrada, mas com uma considerável variedade. Os algoritmos que apresentaram maior utilização foram as redes neurais artificiais (*Artificial Neural Networks* – ANN) com 25% e o algoritmo *Multiple Imputation by Chained Equations* (MICE) com 31%.

4.5 – Perspectiva de tipo de estudo.

Tratando-se da perspectiva de tipo de estudo, os artigos investigados foram classificados em descritivo, preditivo ou prescritivo. É possível observar na Figura 33 uma predominância de estudos descritivos, poucos estudos preditivos com um discreto crescimento a partir do ano de 2014 e nenhum artigo prescritivo.

De uma forma geral, processos de imputação de dados inserem-se na etapa de pré-processamento de dados. Busca-se construir um conjunto de dados sem ausência com o objetivo de melhorar o resultado de uma análise descritiva ou dos modelos computacionais preditivos. Já estudos prescritivos ocorrem quando já existe uma maturidade no processo de predição e o analista ou pesquisador está em busca de dicas (*insights*).

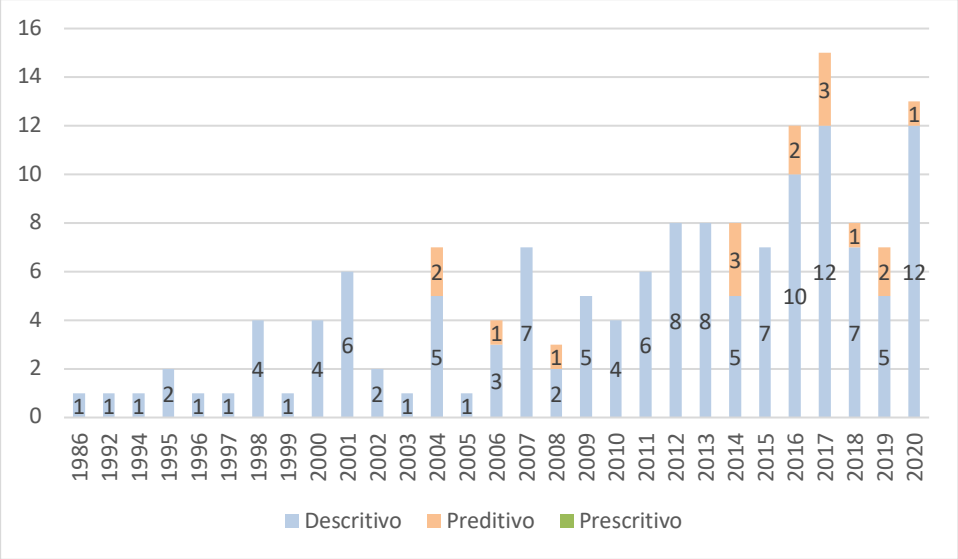

Figura 33 – Quantidade de publicações por tipo de estudo por ano. Fonte: Scopus, 2020.

Neste contexto, observam-se estudos valendo-se da remoção completa de casos (*Listwise Deletion)*, que inspiram a décima pergunta de pesquisa: *"***Quantos artigos realizaram a remoção completa de casos? O estudo ser descritivo, preditivo ou prescritivo impacta na escolha pela remoção completa de casos?***"*.

Dentre os artigos investigados, foram encontrados somente seis artigos interessados em saber se técnicas de imputação apresentam melhores resultados que a realização da remoção completa de casos, todos eles do tipo descritivo. De acordo com a literatura, existem riscos relacionados à remoção completa de casos, tais como a perda de precisão, redução da amostra e criação de vieses (FERLIN, 2008), os quais podem inviabilizar a sua utilização em estudo preditivos ou prescritivos.

4.6 – Perspectiva de reprodutibilidade

Entende-se que a possibilidade de reproduzir os estudos experimentais é uma questão de grande importância para a comunidade científica, pois permite que a pesquisa acadêmica se desenvolva através da contribuição e o diálogo entre diferentes pesquisadores (MUNAFÒ et al., 2017). Desta forma, a possibilidade de se reproduzir os estudos experimentais foi objeto da décima primeira pergunta de pesquisa: *"***Os estudos realizados podem ser reproduzidos? Estes estudos apresentaram pseudocódigo,**

código em repositório aberto ou conjunto de dados público? Os experimentos realizaram a comparação dos resultados com o conjunto de dados original?*"*.

A Figura 34 aponta que somente 30% dos artigos apresentaram pseudocódigos dos algoritmos utilizados nos experimentos, 42% utilizaram conjuntos de dados públicos, apenas 1% dos artigos apresentou código fonte em repositório aberto e 45% dos artigos compararam os resultados de seus experimentos com o conjunto de dados original. Estes resultados demonstram uma necessidade de melhoria dos estudos no que tange à possibilidade de reprodutibilidade de seus experimentos.

Apresentar o código fonte em um repositório aberto, como por exemplo, no GitHub[3] ou em notebooks Jupyter[4], consiste na forma eficiente de apresentar a reprodutibilidade de seus experimentos para a comunidade científica. No entanto, este recurso foi o menos utilizado nos artigos (somente 1%). Isto pode significar que a pesquisa científica ocorre frequentemente de forma isolada, fazendo com que a sinergia de pesquisa ocorra através de textos contidos nos artigos, sendo necessário que os pesquisadores busquem a realização de parcerias para a realização de pesquisas conjuntas. Não que isto esteja errado. Todavia, será que o desenvolvimento tecnológico poderia ocorrer de forma mais acelerada caso os métodos experimentais fossem mais explícitos? Esta pergunta traz o contexto para a importância de reprodutibilidade.

[3] https://github.com/
[4] https://jupyter.org/

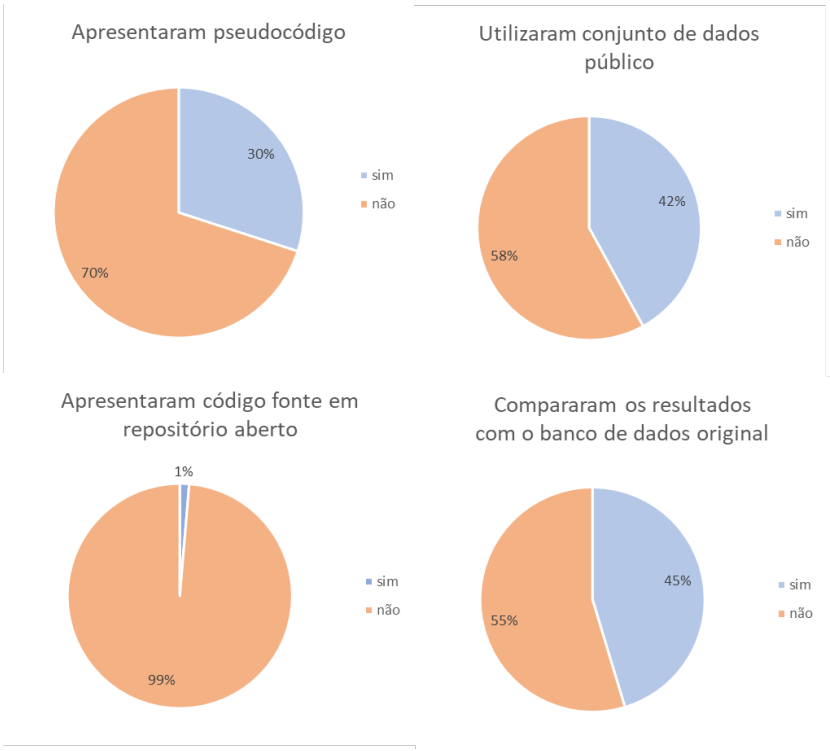

Figura 34 - Proporção dos artigos que apresentaram pseudocódigos, conjunto de dados público e código fonte em repositório aberto. Nota: n =150. Fonte: Scopus, 2020.

Esta dissertação estruturou a reprodutibilidade nos experimentos de imputação a partir de quatro componentes: (i) pseudocódigo; (ii) conjunto de dados público; (iii) código fonte em repositório aberto; e (iv) comparação com o banco de dados original. Avaliou-se sua utilização ao longo dos anos com o objetivo de identificar se ocorre alguma evolução no desenvolvimento dos artigos. A observação da Figura 35 permite observar um aumento da reprodutibilidade dos artigos a partir do ano 2012, com um aumento da publicação dos pseudocódigos e da utilização de conjunto de dados públicos.

Utilizar um conjunto de dados público em seus estudos que esteja disponível para a comunidade científica é muito importante para a reprodutibilidade dos experimentos, uma vez que cada conjunto de dados possui características específicas. No contexto de imputação de dados, a avaliação das características das bases de dados é fundamental para decidir quais técnicas ou algoritmos utilizar nos experimentos. Se um pesquisador não utiliza um conjunto de dados público e o não disponibiliza para a comunidade cientifica, ele dificulta a reprodutibilidade de seus experimentos por outros pesquisadores.

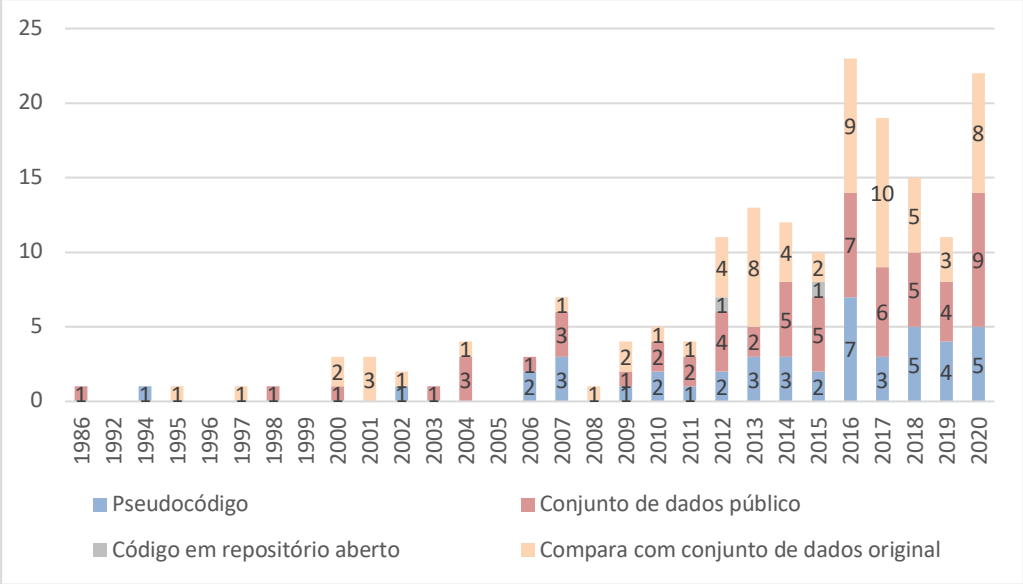

Figura 35 - Reprodutibilidade dos artigos por ano. Fonte: Scopus, 2020.

Desta forma, considerando os benefícios proporcionados à comunidade científica quando um pesquisador utiliza conjuntos de dados públicos, foi feita a décima segunda pergunta de pesquisa: "**Quais os principais conjuntos de dados utilizados nos estudos experimentais de imputação *hot-deck*?**".

De acordo com a Figura 36, que responde à esta pergunta de pesquisa, os conjuntos de dados públicos *Iris Plant*, *Wine*, Pima *Indians Diabetes* e *Breast Cancer Wisconsin* lideram o ranking de utilização em experimentos. A relação completa de todos os conjuntos de dados utilizados nos artigos investigados pode ser encontrada no Apêndice C desta dissertação.

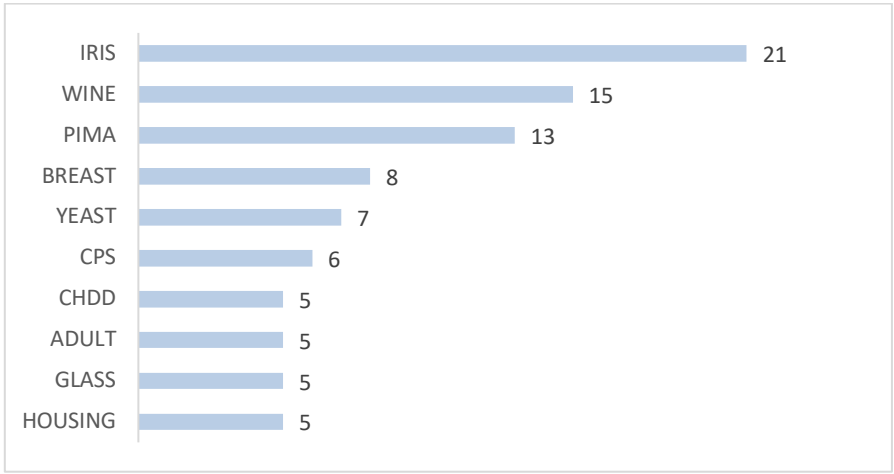

Figura 36 - Ranking dos 10 conjunto de dados públicos por quantidade de vezes que foram utilizados. Fonte: Scopus, 2020.

A Tabela 4 apresenta as principais características destes conjunto de dados mais utilizados: (i) o tema relativo ao conjunto de dados para a compreensão de seu contexto; (ii) a quantidade de tuplas, também chamada de casos; e (iii) a quantidade de atributos presentes em cada conjunto de dados, ou seja, as suas variáveis.

Tabela 4 - Principais características dos quatro conjuntos de dados mais utilizados.

Nome	**Tema**	**Quant. Tuplas**	**Quant. Atributos**
Iris Plant	Características de plantas	150	4
Wine	Análises químicas de vinhos da Itália	178	13
Pima Indians Diabetes	Mulheres com diabetes da tribo Pima do Arizona	768	8
Breast Cancer Wisconsin	Pacientes com câncer de mama	699	9

Fonte: Elaborado pelo autor.

O conjunto de dados *Iris Plants* registra as medidas de comprimento e largura de pétalas e caule de três tipos de plantas: *Virginica, Versicolor* e *Setosa*. Este conjunto possui 150 tuplas distribuída em três classes com 50 representantes cada. Nenhum registro apresenta valores ausentes. O atributo classe armazena o tipo da planta, em função das medidas de caule e pétalas. Já o atributo *sepal length* armazena o comprimento do caule enquanto o *sepal width* sua largura. Em relação as pétalas, *petal lenght* representa o comprimento e o *petal width* a largura das mesas.

Já o conjunto de dados *Wine* é composto de dados sobre as análises químicas de vinhos provenientes da mesma região da Itália, mas com origem em três cultivadores distintos. As análises determinam as quantidades de treze componentes encontrados em cada um dos três tipos de vinhos. Este conjunto possui 13 atributos contínuos e três classes que indicam o tipo do vinho. A primeira classe e formada por 59 casos, enquanto a segunda possui 71 casos. Por fim, a terceira classe possui 48 casos. Os atributos deste conjunto são: *Alcohol, Malic, acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines* e *Proline*.

O conjunto de dados *Pima Indians Diabetes* armazena dados de mulheres com diabetes e mais de 21 anos da tribo Pima do Arizona, EUA. Seus registros possuem um atributo de classificação que indica se a paciente possui ou não ao a doença diabetes.

Seus atributos são: *age* (idade), *blood pressure* (pressão sanguínea diastólica), *body mass* (índice de massa corporal[5]), *glucose concentration* (nível de glicose no sangue duas horas após a ingestão de glicose concentrada de um teste de tolerância glicose), *pedigree function* (função de características hereditárias da diabetes), *pregnancy times* (número de vezes que a paciente engravidou), *serum insulin* (nível de insulina no sangue), e *skin fold thickness* (espessura da pele do tríceps).

O conjunto *Wisconsin Breast Cancer* apresenta dados sobre o diagnóstico de câncer de mama relativos a pacientes que tiveram a doença nos meses de janeiro e outubro de 1989; fevereiro, abril e agosto de 1990; e janeiro, junho e novembro de 1991. Este conjunto possui nove atributos contínuos e duas classes que indicam uma neoplasia benigna ou maligna. Dos 699 casos existente, 682 não possuem ausência de dados, dos quais 239 indicam pacientes que apresentam neoplasia de mama maligna. Os demais atributos são: *Uniformity of Cell Size, Clump Thickness, Bland Chromatin, Uniformity of Cell Shape, Marginal Adhesion, Mitoses, Bare Nuclei, Normal Nucleoli, Single Epithelial Cell Size*.

As características destes quatro conjuntos de dados que lideram o ranking dos mais utilizados foram apresentadas com o objetivo de ilustrar a existência de uma variedade de conjuntos de dados com distintos contextos, variáveis e casos. É muito importante para o pesquisador conhecer os resultados dos algoritmos avaliados em diversos conjunto de dados. Assim, a Figura 37 apresenta a quantidade de conjuntos de dados utilizados nos experimentos, na qual se percebe um crescimento quantitativo a partir do ano de 2011.

[5] IMC (Índice de Massa Corporal) = peso/altura^2

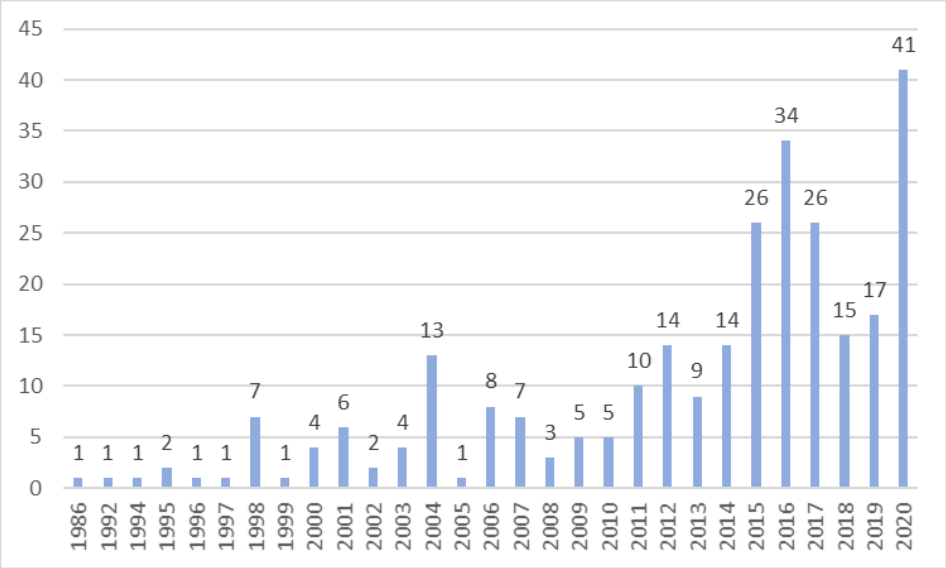

Figura 37 - Quantidade de conjuntos de dados utilizados nos experimentos por ano.
Fonte: Scopus, 2020.

Os estudos sobre imputação de dados que utilizam a técnica *hot-deck* avaliam o desempenho dos algoritmos em uma quantidade cada vez maior de conjunto de dados. Como exemplo observado nesta revisão sistemática, temos o estudo *"Single imputation with multiplayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns"*, no qual os autores desenvolveram modelos de imputação automatizada com base em redes neurais artificiais que foram testados em dezoito conjuntos de dados públicos (SILVA-RAMÍREZ; PINO-MEJÍAS; LÓPEZ-COELLO, 2015).

Além deste, no artigo *"Takagi-Sugeno Modeling of Incomplete Data for Missing Value Imputation with the Use of Alternate Learning"* , os autores desenvolveram um modelo que realiza o grupamento da categoria *Fuzzy* antes da imputação e avaliaram o seu desempenho em doze conjuntos de dados públicos (LAI; ZHANG; LIU, 2020). Segundo os autores, os resultados experimentais em vários conjuntos e diferentes simulações de mecanismos e porcentagem de ausência demonstraram a eficácia da estratégia e do método propostos.

4.7– Conhecimento encontrado

Esta seção apresenta um resumo do conhecimento encontrado através da revisão sistemática norteada pelas perguntas de pesquisas realizadas. O quadro 2 descreve as perguntas de pesquisa, as respostas e os comentários sobre o conhecimento encontrado.

Quadro 2 - Resumo do conhecimento encontrado.

Pergunta de Pesquisa	Resposta	Comentário
Quais são os principais periódicos que publicam artigos sobre imputação *hot-deck*?	*Journal of the American Statistical Association, Computational Statistics and Data Analysis, Biometrika, Canadian Journal of Statistics, Procedings of The International Society for Optical Engineering* e *Journal of Statistical Planning and Inference.*	Os periódicos relacionados à área de estatística apresentam grandes contribuições para a comunidade acadêmica. No entanto, o tema imputação *hot-deck* é pulverizado em diferentes áreas de pesquisa.
Quantos estudos comparativos com a utilização de *hot-deck* foram realizados?	Do total de 150 artigos investigados, 87 artigos (58% do total) realizaram estudos comparativos.	É possível observar um crescimento, ao longo dos anos, na quantidade de estudos comparativos sobre imputação *hot-deck*.
Os padrões de ausências são adequadamente identificados nos estudos experimentais?	Somente foi encontrado um artigo (STRIKE; EL EMAM; MADHAVJI, 2001) que simulou os padrões de ausência em seus experimentos.	É preciso comunicar à comunidade científica sobre a necessidade de considerar os padrões de ausência em seus experimentos.
Os mecanismos de ausências são adequadamente identificados nos estudos experimentais?	Somente 37% dos artigos investigados avaliaram os mecanismos de ausência.	O crescimento quantitativo de estudos sobre imputação *hot-deck* ao longo dos anos não é acompanhando de um crescimento qualitativo com relação à avaliação dos mecanismos de ausência.

Pergunta de Pesquisa	Resposta	Comentário
Ao longo dos anos está ocorrendo uma maior utilização de *hot-deck* em estudos de imputação híbrida?	Ocorre um aumento nos estudos de imputação híbrida que utilizam imputação *hot-deck* a partir do ano 2001.	É possível que o aumento dos estudos de imputação hídrica aconteça devido aos avanços tecnológicos na capacidade de processamento dos computadores.
Quantos estudos de imputação híbrida utilizaram a técnica *hot-deck*? Destes estudos, quantos realizaram a Imputação Múltipla?	Foram identificados 59 estudos de imputação híbrida que utilizaram a técnica *hot-deck*, o que corresponde a 39% dos artigos investigados. Dentre estes 59 artigos, 39 (66%) artigos realizaram métodos de imputação híbrida diversa e 20 artigos realizaram a imputação múltipla utilizando algoritmos de agrupamento	Classificar os estudos híbridos é um desafio devido à diversidade de combinações de técnicas e algoritmos.
Quantos estudos de imputação híbrida utilizaram métodos estatísticos?	Do total de 59 estudos de imputação híbrida 26 artigos (44%) recorrem a métodos estatísticos.	Os métodos estatísticos simples prevalecem tanto em estudos de imputação híbrida quanto em imputação global.
Qual categoria de algoritmos de agrupamento é a mais utilizada na imputação *hot-deck*?	A categoria de agrupamento do tipo *Partitioning Based* é a mais utilizada. Do total de algoritmos utilizados para realizar o agrupamento, 72% estão incluídos nesta categoria.	É necessário que os algoritmos de outras categorias de agrupamento sejam objeto de novas investigações.
Quais são os principais algoritmos utilizados na etapa de agrupamento que precede a imputação?	Os algoritmos *Random hot-deck*, *K-Nearest-Neighbor* e *K-means* são os principais algoritmos utilizados na etapa de agrupamento que precede a imputação.	Sobre os algoritmos de aprendizado de máquina, os algoritmos de aprendizado supervisionado ainda possuem uma utilização predominante.

Pergunta de Pesquisa	Resposta	Comentário
Quantos artigos realizaram a remoção completa de casos? O estudo ser descritivo, preditivo ou prescritivo impacta na escolha pela remoção completa de casos?	Dentre o total de artigos investigados, foram encontrados somente seis artigos que buscavam saber se técnicas de imputação apresentavam melhores resultados que a realização da remoção completa de casos, todos eles do tipo descritivo.	Os riscos relacionados à remoção completa de casos, tais como: a perda de precisão, redução da amostra e criação de viés, podem inviabilizar a sua utilização em estudo preditivos ou prescritivos.
Os estudos realizados podem ser reproduzidos? Estes estudos apresentaram pseudocódigo, código em repositório aberto ou conjunto de dados público? Os experimentos realizaram a comparação dos resultados com o conjunto de dados original?	Somente 30% dos artigos investigados apresentaram pseudocódigos dos algoritmos utilizados nos experimentos; 42% utilizaram conjunto de dados público; apenas 1% dos artigos apresentou código fonte em repositório aberto; e 45% dos artigos compararam os resultados de seus experimentos com o conjunto de dados original.	É preciso comunicar à comunidade acadêmica sobre a necessidade de melhoria dos estudos no que diz respeito à possibilidade de reprodutibilidade de seus experimentos. Entretanto, observou-se um crescimento da reprodutibilidade dos artigos a partir do ano de 2012.
Quais os principais conjuntos de dados utilizados nos estudos experimentais utilizadas na imputação *hot-deck*?	Os conjuntos de dados públicos *Iris Plant*, *Wine*, *Pima Indians Diabetes* e *Breast Cancer Wisconsin* lideram o ranking de utilização em experimentos de imputação *hot-deck*.	Notou-se um crescimento na quantidade de conjunto de dados utilizados a partir do ano de 2011.

Fonte: Elaborado pelo autor.

5- Considerações Finais

Nesta seção são apresentadas uma análise retrospectiva sobre a pesquisa realizada, as considerações sobre os resultados encontrados e os próximos passos para a realização de trabalhos futuros.

5.1 – Análise Retrospectiva

Neste trabalho foi apresentada uma revisão sistemática sobre técnicas de imputação *hot-deck* com o objetivo de avaliar como ocorre a evolução dos estudos sobre este tema longo dos anos. Buscou-se também verificar a qualidade e a reprodutibilidade dos mesmos e identificar novos caminhos de pesquisa para trabalhos futuros.

Foram realizadas buscas no repositório de pesquisa científica Scopus com o objetivo de encontrar trabalhos científicos que investigaram a técnica objeto de estudo desta dissertação, as quais retornaram 432 documentos. Deste total, foram baixados 321, os quais foram classificados 72 artigos com prioridade alta, 78 artigos com prioridade média e 171 artigos com prioridade baixa.

Os artigos de prioridade média e alta foram analisados sob a orientação de doze de perguntas de pesquisa formuladas com o propósito de alcançar o objetivo desta dissertação. Esta análise baseou-se em uma taxonomia proposta por este estudo, estendida de Soares (2007), que buscou classificar, ordenar e estabelecer hierarquias para as técnicas de imputação. Além desta, também foi utilizada uma taxonomia de agrupamento definida por Kamlesh e Diwakar (2019).

5.2 – Considerações finais sobre os resultados

A revisão sistemática identificou uma considerável contribuição acadêmica dos periódicos de estatística. No entanto, o tema imputação *hot-deck* se encontra pulverizado em diferentes áreas de pesquisa. Este cenário revela-se desafiador para

esta área de pesquisa, uma vez alguns estudos investigam e comparam o desempenho entre distintas abordagens de imputações e diferentes algoritmos, enquanto outros aplicam a técnica de *hot-deck* em seus problemas específicos. Classificar estes dois grupos de estudos não foi possível por meio de um descritor de busca, o que tornou necessária a leitura e compreensão dos objetivos de pesquisa de cada artigo, procedimento que permitiu identificar um crescimento, ao longo dos anos, na quantidade de estudos comparativos sobre imputação *hot-deck.*

Apesar de a identificação dos padrões e mecanismos de ausência ser fundamental para a elaboração de adequadas estratégias de imputação, uma pequena proporção dos artigos realizou estas identificações. Além disso, foi possível observar que ocorre um aumento de estudos sobre imputação *hot-deck* ao longo dos anos. Entretanto, não se percebe uma melhoria com relação à identificação dos padrões e mecanismos de ausência, fato que precisa ser comunicado à comunidade científica.

No que diz respeito à possibilidade de reprodutibilidade dos experimentos, observou-se o seu crescimento ao longo dos anos. Todavia, o cenário de reprodutibilidade precisa melhorar. Desta forma, recomenda-se que futuros artigos estejam comprometidos com a difusão de uma cultura de reprodutibilidade experimental que pode ser realizada pela utilização preferencial de bases públicas de dados, apresentação de pseudocódigos dos algoritmos e a publicação de códigos-fonte dos experimentos em repositórios abertos.

A literatura sobre imputação *hot-deck* demonstra a sua capacidade de gerar estimativas precisas e com a maior variabilidade, uma vez que esta abordagem reconhece as diferenças individuais entre os sujeitos no processo de imputação. Considerando esta vantagem, o presente estudo destaca que a abordagem *hot-deck* pode ser algo mais amplo que utilizar os seus algoritmos clássicos, tais como: *Random hot-deck* ou *Nearest Neighbor*, uma vez que existe um vasto conjunto de outros algoritmos de agrupamento que podem ser estudados e investigados, para os quais ainda existem poucos artigos publicados.

5.3 – Trabalhos futuros

O presente estudo identificou algumas oportunidades de trabalhos futuros. Observou-se um aumento do interesse na imputação híbrida com diferentes combinações de técnicas e algoritmos. Logo, realizar uma investigação mais detalhada sobre as técnicas de imputação híbrida e encontrar similaridades entre estas combinações parece ser uma oportunidade de trabalho futuro.

Compreender que os métodos estatísticos simples ainda prevalecem tanto em estudos de imputação híbrida quanto de imputação global, mesmo com a literatura acadêmica evidenciando as suas limitações, abre um campo de possibilidades de experimentos que utilizem métodos estatísticos bayesianos e de verossimilhança.

Situação similar também ocorre com os algoritmos de agrupamento, uma vez que os do tipo *Partitioning Based* são os mais utilizados. Existe um considerável universo de algoritmos das outras categorias de agrupamento que podem ser objeto de novos estudos experimentais.

Considerando o universo dos estudos comparativos, é preciso investigar quais são os principais métodos de medição da acurácia da imputação, e partir das características mais frequentes entre os conjuntos de dados, propor os métodos mais adequados.

APÊNDICE A – Lista de artigos com prioridade alta

Título	Referência
A comparison of four imputation procedures in a two-variable prediction system	(HEGAMIN-YOUNGER; FORSYTH, 1998)
A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck	(SULLIVAN; ANDRIDGE, 2015)
A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction	(VAZIFEHDAN; MOATTAR; JALALI, 2019)
A likelihood-based constrained algorithm for multivariate normal mixture models	(INGRASSIA, 2004)
A martingale representation for matching estimators	(ABADIE; IMBENS, 2009)
A Monte Carlo Analysis of Missing Data Techniques in a HRM Setting	(ROTH; SWITZER, 1995)
A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria	(AIEB et al., 2019)
A new imputation method for small software project data sets	(SONG; SHEPPERD, 2007)
A new multivariate imputation method based on Bayesian networks	(NILOOFAR; GANJALI, 2014)
A study of missing data imputation in predictive modeling of a wood-composite manufacturing process	(STEINER et al., 2016)
Additive integer-valued data envelopment analysis with missing data: A multi-criteria evaluation approach	(CHEN et al., 2020)
Alternative methods for CPS income imputation	(DAVID et al., 1986)
Analysis of a pilot study for amelioration of itching in liver disease: When is a failed trial not a failure?	(MCGEE; BERGASA, 2006)
Balanced k-nearest neighbour imputation	(HASLER; TILLÉ, 2016)
Balanced repeated replication for stratified multistage survey data under imputation	(SHAO; CHEN; CHEN, 1998)
Benchmarking k-nearest neighbour imputation with homogeneous Likert data	(JÖNSSON; WOHLIN, 2007)
Calibrated hot deck imputation for numerical data under edit restrictions	(DE WAAL; COUTINHO; SHLOMO, 2017)
Can earnings equations estimates improve CPS hot-deck imputations?	(BISHOP; FORMBY; THISTLE, 2003)
Confidence intervals for marginal parameters under imputation for item nonresponse	(QIN; RAO; REN, 2008)
Doubly Robust Inference for the Distribution Function in the Presence of Missing Survey Data	(BOISTARD; CHAUVET; HAZIZA, 2016)
Empirical likelihood-based hot deck imputation methods	(XUE; LAZAR, 2012)
Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators	(CONTI; MARELLA; SCANU, 2008)
Experimental analysis of methods for imputation of missing values in databases	(FARHANGFAR; KURGAN; PEDRYCZ, 2004)
Fractional regression hot deck imputation weight adjustment	(PAIK; LARSEN, 2013)
Fully efficient estimation of coefficients of correlation in the presence of imputed survey data	(CHAUVET; HAZIZA, 2012)
Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data	(MYERS, 2011)
GRAFT, a complete system for data fusion	(ALUJA-BANET; DAUNIS-I-ESTADELLA; PELLICER, 2007)
Handling incomplete quality-of-life data	(SHEN; LAI, 2001)
Handling item nonresponse in the U.S. component of the IEA reading literacy study	(WINGLEE et al., 2001)

Título	Referência
Improvement of prognostic models for ESRD mortality by the bootstrap method with random hot deck imputation	(LIN; YANG; CHIANG, 2014)
Imputation of missing item responses: Some simple techniques	(HUISMAN, 2000)
Imputation: Methods, simulation experiments and practical examples	(NORDHOLT, 1998)
Intelligent imputation technique for missing values	(ALJUAID; SASI, 2016b)
Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors	(SKINNER; RAO, 2002)
Jackknife variance estimation for nearest-neighbor imputation	(CHEN; SHAO, 2001)
Jackknife variance estimation under imputation for estimators using poststratification information	(YUNG; RAO, 2000)
Jackknife variance estimation with survey data under hot deck imputation	(RAO; SHAO, 1992)
Measuring disclosure risk for multimethod synthetic data generation	(LARSEN; HUCKETT, 2010)
Methods for Addressing Missing Data with Applications from ACS Exams	(BRANDRIET; HOLME, 2015)
Mining trauma injury data with imputed values	(PENNY; CHESNEY, 2009)
Missing behavior data in longitudinal network studies: the impact of treatment methods on estimated effect parameters in stochastic actor oriented models	(ZANDBERG; HUISMAN, 2019)
Missing data imputation and its effect on the accuracy of classification	(HUNT, 2017)
Missing Data Imputation for MIMIC-III using Matrix Decomposition	(YANG et al., 2019)
Missing data imputation, matching and other applications of random recursive partitioning	(IACUS; PORRO, 2007)
Missing Data in Homicide Research	(RIEDEL; REGOECZI, 2004)
Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques	(ROTH; SWITZER; SWITZER, 1999)
Missing Data Problems in the SHR: Imputing Offender and Relationship Characteristics	(FOX, 2004)
Missing Value Analysis of Numerical Data using Fractional Hot Deck Imputation	(CHRISTOPHER et al., 2019)
Missing-values adjustment for mixed-type data	(TARSITANO; FALCONE, 2011)
Multiple imputation of missing values in exploratory factor analysis of multidimensional scale	(LORENZO-SEVA; VAN GINKEL, 2016)
Neural network imputation: An experience with the National Resources Inventory Survey	(MAITI; MILLER; MUKHOPADHYAY, 2008)(BANKHOFER; JOENSSEN, 2014)
On limiting donor usage for imputation of missing data via hot deck methods	(BANKHOFER; JOENSSEN, 2014)
Performance evaluation of imputation based on Bayesian Networks	(NILOOFAR; GANJALI; FARID ROHANI, 2013)
Probabilistic neural network based categorical data imputation	(NISHANTH; RAVI, 2016)
Proper imputation techniques for missing values in data sets	(ALJUAID; SASI, 2016a)
Research and implementation of animations evaluation system	(HUANG; DU, 2017)
Robust and automatic data cleansing method for short-term load forecasting of distribution feeders	(HUYGHUES-BEAUFOND et al., 2020)
Semi-empirical likelihood confidence intervals for the differences of quantiles with missing data	(QIN; ZHANG, 2009)
Simultaneous Edit-Imputation for Continuous Microdata	(KIM et al., 2015)
Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns	(SILVA-RAMÍREZ; PINO-MEJÍAS; LÓPEZ-COELLO, 2015)
Smoothed jackknife empirical likelihood inference for ROC curves with missing data	(YANG; ZHAO, 2015)
Software cost estimation with incomplete data	(STRIKE; EL EMAM; MADHAVJI, 2001)

Título	Referência
Statistical data fusion for cross-tabulation	(KAMAKURA; WEDEL, 1997)
The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. first results from the multicare cohort study	(SCHÄFER et al., 2012)
The Pre- and Post-1997 well-being of Hong Kong residents	(SHEN; CHOY, 2005)
The use of hot deck imputation to compare performance of further education colleges	(PENNY; ASHRAF; DUFFY, 2007)
Treatments of missing data: A Monte carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization	(GOLD; BENTLER, 2000)
Using multiple imputation to address missing values of hierarchical data	(ZHANG et al., 2017a)
Variance estimation for nearest neighbor imputation for US Census long form data	(KIM; FULLER; BELL, 2011)
Variance estimation in two-stage cluster sampling under imputation for missing data	(HAZIZA; RAO, 2010)
Variance Estimation of Imputed Estimators of Change for Repeated Rotating Surveys	(BERGER; ESCOBAR, 2017)
When data goes missing: Methods for missing score imputation in biometric fusion	(DING; ROSS, 2010)

APÊNDICE B – Lista de artigos com prioridade média

Título	Referência
A classifier ensemble approach for the missing feature problem	(NANNI; LUMINI; BRAHNAM, 2012)
A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data	(LI; GU; ZHANG, 2010)
A mean score method for missing and auxiliary covariate data in regression models	(REILLY; PEPE, 1995)
A new iterative fuzzy clustering algorithm for multiple imputation of missing data	(NIKFALAZAR et al., 2017)
A proposal for a two-step sampling design to oversample units responding to prescribed characteristics	(ANDREIS; BONETTI, 2018)
A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets	(HUANG; ZHU, 2002)
A Rough Set Approach to Data Imputation and Its Application to a Dissolved Gas Analysis Dataset	(WATADA et al., 2016)
Adaptive logistic group Lasso method for predicting the no-reflow among the multiple types of high-dimensional variables with missing data	(YANG et al., 2016)
Administrative data informed donor imputation in the Australian Census of Population and Housing	(FARNELL; DARBY, 2020)
An agglomerative clustering methodology for data imputation	(YENDURI, 2006)
An efficient approach for imputation and classification of medical data values using class-based clustering of medical records	(YELIPE; PORIKA; GOLLA, 2018)
An enhanced fuzzy K-means clustering with application to missing data imputation	(MIGDADY; AL-TALIB, 2018)
Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level	(ZHANG et al., 2017b)
Assessing secular trends in blood pressure: A multiple-imputation approach	(HEITJAN; LANDIS, 1994)
Bayesian networks for imputation	(DI ZIO et al., 2004)
Bootstrap methods for imputed data from regression, ratio and hot-deck imputation	(MASHREGHI; LÉGER; HAZIZA, 2014)
Calibrated hot-deck donor imputation subject to edit restrictions	(COUTINHO; WAAL; SHLOMO, 2013)
Cluster-Based Best Match Scanning for Large-Scale Missing Data Imputation	(YU et al., 2017)
Clustering-based missing value imputation for data preprocessing	(ZHANG et al., 2006)
Clustering-based multiple imputation via gray relational analysis for missing data and its application to aerospace field	(TIAN et al., 2013)
Computational aspects of nonparametric bayesian analysis with applications to the modeling of multiple binary sequences	(QUINTANA; NEWTON, 2000)
Convergence of random k-nearest-neighbour imputation	(DAHL, 2007)
Evolving clustering based data imputation	(GAUTAM; RAVI, 2014)
Fractional hot deck imputation	(KIM; FULLER, 2004)
Fractional imputation for incomplete two-way contingency tables	(KANG; KOEHLER; LARSEN, 2012)
Fuzzy cluster analysis with missing values	(TIMM; KRUSE, 1998)
Fuzzy c-means classifier with deterministic initialization and missing value imputation	(ICHIHASHI et al., 2007)
Improving data quality through high precision gender categorization	(MULLER; TE; JAIN, 2017)
Imputation for multisource data with comparison and assessment techniques	(CASLETON; OSTHUS; VAN BUREN, 2018)
Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates	(DURRANT, 2009)
Imputation of incomplete data using adaptive ellipsoids with linear regression	(YAO; WENG, 2015)

Título	Referência
Instance driven clustering for the imputation of missing data in KDD	(ILANGO; VIJAYAKUMAR; BABU, 2014)
Integrating WLI fuzzy clustering with grey neural network for missing data imputation	(KUPPUSAMY; PARAMASIVAM, 2017)
Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario	(DI NUOVO, 2011)
Missing data imputation using decision trees and fuzzy clustering with iterative learning	(NIKFALAZAR et al., 2020)
Missing value imputation techniques depth survey and an imputation Algorithm to improve the efficiency of imputation	(THIRUKUMARAN; SUMATHI, 2012)
Multiple imputation methods for handling missing data in cost-effectiveness analyses that use data from hierarchical studies: An application to cluster randomized trials	(GOMES et al., 2013)
Multiple-vs Non-or Single-Imputation Based Fuzzy Clustering for Incomplete Longitudinal Behavioral Intervention Data	(ZHANG; FANG, 2016)
Nonparametric bayesian analysis for assessing homogeneity in k × l Contingency tables with fixed right margin totals	(QUINTANA, 1998)
Partially parametric techniques for multiple imputation	(SCHENKER; TAYLOR, 1996)
Performance evaluation of imputation methods for incomplete datasets	(YENDURI; IYENGAR, 2007)
Performance evaluation of predictive models for missing data imputation in weather data	(DORESWAMY; GAD; MANJUNATHA, 2017)
Privacy-preserved big data analysis based on asymmetric imputation kernels and multiside similarities	(CHEN et al., 2016)
Relational data clustering with incomplete data	(HATHAWAY et al., 2001)
Scaling out big data missing value imputations: Pythia vs. Godzilla	(ANAGNOSTOPOULOS; TRIANTAFILLOU, 2014)
Semi-empirical likelihood inference for the ROC curve with missing data	(YANG; ZHAO, 2015)
Sharing confidential data for algorithm development by multiple imputation	(VERWER; VAN DEN BRAAK; CHOENNI, 2013)
Survey Data Fusion	(ALUJA-BANET; THIÓ, 2001)
Takagi-Sugeno Modeling of Incomplete Data for Missing Value Imputation with the Use of Alternate Learning	(LAI; ZHANG; LIU, 2020)
The Role of CPS Nonresponse in the Measurement of Poverty	(HOKAYEM; BOLLINGER; ZILIAK, 2015)
Towards missing data imputation: A study of fuzzy K-means clustering method	(LI et al., 2004)
Variance estimation when donor imputation is used to fill in missing values	(BEAUMONT; BOCCI, 2009)
A Clustering-Based Approach for Data-Driven Imputation of Missing Traffic Data	(KU et al., 2016)
A Comparative Study on TIBA Imputation Methods in FCMdd-Based Linear Clustering with Relational Data	(YAMAMOTO et al., 2011a)
A Missing Data Imputation Approach Using Clustering and Maximum Likelihood Estimation	(ALBAYRAK; TURHAN; KURT, 2017)
A Novel Fuzzy Rough Clustering Parameter-based missing value imputation	(RAJA; SASIREKHA; THANGAVEL, 2020)
A study on a fuzzy clustering for mixed numerical and categorical incomplete data	(FURUKAWA; OHNISHI; YAMANOI, 2013)
A study on missing values imputation using K-Harmonic means algorithm: Mixed datasets	(ANWAR et al., 2019)
Adaptive SOMMI (Self Organizing Map Multiple Imputation) base on Variation Weight for Incomplete Data	(KHUSNULKHOTIMAH; SUPRAJITNO, 2018)
Clustering Data with the Presence of Missing Values by Ensemble Approach	(PATTANODOM; IAM-ON; BOONGOEN, 2016)
DBSCANI: Noise-Resistant Method for Missing Value Imputation	(PURWAR; SINGH, 2016)

Título	Referência
Design Space Exploration of The KNN Imputation on FPGA	(AL-ZOUBI; TATAS; KYRIACOU, 2018)
Estimation of Tree Lists from Airborne Laser Scanning Using Tree Model Clustering and k-MSN Imputation	(LINDBERG et al., 2013)
FCMdd-type Linear Fuzzy Clustering for Incomplete Non-Euclidean Relational Data	(YAMAMOTO et al., 2011b)
Improving Missing Values Imputation in Collaborative Filtering With User-Preference Genre and Singular Value Decomposition	(INSUWAN; SUKSAWATCHON; SUKSAWATCHON, 2014)
K-Means Clustering With Incomplete Data	(WANG et al., 2019)
K-Nearest Temperature Trends: A Method for Weather Temperature Data Imputation	(KIANI; SALEEM, 2017)
k-POD A Method for k-Means Clustering of Missing Data	(CHI; CHI; BARANIUK, 2016; RAJA; SASIREKHA; THANGAVEL, 2020)
Missing Data Imputation: A Fuzzy K-means Clustering Algorithm over Sliding Window	(LIAO et al., 2009)
Missing value imputation using a fuzzy clustering-based EM approach	(RAHMAN; ISLAM, 2016)
Missing Value Imputation using Hybrid K-Means and Association Rules	(CHHABRA; VASHISHT; RANJAN, 2018)
Missing value imputation using unsupervised machine learning techniques	(RAJA; THANGAVEL, 2020)
A hybrid approach for the stratified mark-specific proportional hazards model with missing covariates and missing marks, with application to vaccine efficacy trials	(SUN et al., 2020)
Asymptotic theory and inference of predictive mean matching imputation using a super population model framework	(YANG; KIM, 2020)
Best Fit Missing Value Imputation (BFMVI) Algorithm for Incomplete Data in the Internet of Things	(AGBO; QIN; HILL, 2020)
Completion of multiview missing data based on multi-manifold regularised non-negative matrix factorisation	(JING-TAO; QIU-YU, 2020)
Fuzzy case-based-reasoning-based imputation for incomplete data in software engineering repositories	(ABNANE; IDRI; ABRAN, 2020)
Optimize Neural Network Algorithm of Missing Value Imputation for Clustering Chocolate Product Type Following “STEAMS” Methodology	(CHEN; CHEN, 2020)

APÊNDICE C – Quadro com a relação de conjuntos de dados utilizados nos experimentos.

Sigla	Descrição	Instâncias	Homepage	Público
102FLOWER	102 Category Flower Dataset		http://www.robots.ox.ac.uk/~vgg/data/flowers/102/	sim
17FLOWER	17 Category Flower Dataset		http://www.robots.ox.ac.uk/~vgg/data/flowers/17/	sim
ABALONE	Abalone Dataset	4177	https://archive.ics.uci.edu/ml/datasets/abalone	sim
ACPH	Australian Census of Population and Housing		https://www.abs.gov.au/websitedbs/censushome.nsf/4a256353001af3ed4b2562bb00121564/census	sim
ACS	American Community Survey		https://www.census.gov/programs-surveys/acs/data.html	sim
ADULT	Adulto Dataset	48842	http://archive.ics.uci.edu/ml/datasets/Adult	sim
API	Academic Performance Index of the California Department of Education		https://www.cde.ca.gov/re/pr/api.asp	sim
ARCENE	Arcene Data Set	900	https://archive.ics.uci.edu/ml/datasets/Arcene	sim
AUTOMPG	Auto MPG Data Set	398	https://archive.ics.uci.edu/ml/datasets/Auto+MPG	sim
AVILA	Avila Data Set	20867	https://archive.ics.uci.edu/ml/datasets/Avila	sim
BANKNOTE	banknote Authentication Data Set	1372	https://archive.ics.uci.edu/ml/datasets/banknote+authentication	sim
BODYFAT	Bodyfat: Body Measures	100	https://rdrr.io/cran/Lock5Data/man/BodyFat.html	sim
BREAST	Breast Cancer Wisconsin (Original) Data Set	669	https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29	sim
CADD	Coronary artery disease data from Danish heart clinic	238	https://rdrr.io/cran/gRbase/man/data-cad.html	sim
CARE	Car Evaluation Data Set	1728	https://archive.ics.uci.edu/ml/datasets/Car+Evaluation	sim
CHDD	Cleveland Heart Disease Data Set	303	https://archive.ics.uci.edu/ml/datasets/heart+disease	sim
CHESS	Chess (King-Rook vs. King-Pawn) Data Set	3196	https://archive.ics.uci.edu/ml/datasets/Chess+(King-Rook+vs.+King-Pawn)	sim
CONTRACEPTIVE	Contraceptive Method Choice Data Set	1473	https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice	sim

Fonte: Elaborado pelo autor.

Sigla	Descrição	Instâncias	Homepage	Público
CPS	Current Population Survey		https://www.census.gov/programs-surveys/cps/data.html	sim
CREDIT	Credit Approval Dataset	690	https://archive.ics.uci.edu/ml/datasets/Credit+Approval	sim
CSTDS	Concrete Slump Test Data Set	103	https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test	sim
DERMATOLOGY	Dermatology Data Set	366	https://archive.ics.uci.edu/ml/datasets/dermatology	sim
DHDS	Dutch Housing Demand Survey		https://www.cbs.nl/en-gb/our-services/methods/surveys/korte-onderzoeksbeschrijvingen/netherlands-housing-survey--woon--	sim
DIASTASIS	Digital Era Statistical Indicators			não
DOW	Dow Jones Index Data Set	750	https://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index	sim
ECG	ECG Heartbeat Categorization Dataset	109446	https://www.kaggle.com/shayanfazeli/heartbeat	sim
ECHP	The European Community Household Panel Survey		https://ec.europa.eu/eurostat/web/microdata/european-community-household-panel	sim
ECOLI	Ecoli Dataset	336	https://archive.ics.uci.edu/ml/datasets/Ecoli	sim
ESRD	CMS End-Stage Renal Disease		https://www.cms.gov/Medicare/End-Stage-Renal-Disease/ESRDGeneralInformation/Data	sim
FATALITIES	US traffic fatalities panel data for the “lower 48” US states (i.e., excluding Alaska and Hawaii), annually for 1982 through 1988	336	https://www.rdocumentation.org/packages/AER/versions/1.2-9/topics/Fatalities	sim
FFS	Financia Farm Survey of Statistics Canada		https://www150.statcan.gc.ca/n1/en/catalogue/21F0008X	sim
FOREST	Forest Fires Data Set	517	https://archive.ics.uci.edu/ml/datasets/Forest+Fires	sim
GENE	Gene Expression Data Set	72	https://www.kaggle.com/crawford/gene-expression	sim
GERMAN	Statlog (German Credit Data) Data Set	1000	https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)	sim
GLASS	Glass Identification Data Set	214	https://archive.ics.uci.edu/ml/datasets/glass+identification	sim
HAYES	Hayes-Roth Dataset	132	https://archive.ics.uci.edu/ml/datasets/Hayes-Roth	sim

Fonte: Elaborado pelo autor.

Sigla	Descrição	Instâncias	Homepage	Público
HDMA	The Boston HMDA Data Set	2381	https://rdrr.io/cran/Ecdat/man/Hmda.html	sim
HOUSING	The Boston Housing Dataset	506	https://www.kaggle.com/schirmerchad/bostonhoustingmlnd	sim
ILPD	ILPD (Indian Liver Patient Dataset) Data Set	583	https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)	sim
IONOSPHERE	Ionosphere Data Set	351	https://archive.ics.uci.edu/ml/datasets/Ionosphere	sim
IRIS	Iris Dataset	150	https://archive.ics.uci.edu/ml/datasets/iris	sim
ISBSG	International Software Benchmarking Standads Group	1238	https://www.isbsg.org/software-project-data/	sim
ISTANBUL	ISTANBUL STOCK EXCHANGE Data Set	536	https://archive.ics.uci.edu/ml/datasets/ISTANBUL+STOCK+EXCHANGE	sim
JAIN	Jain (2D) Data Set	372	http://cs.joensuu.fi/sipu/datasets/	sim
LALONDE	Lalonde propensity score matching	445	http://sekhon.berkeley.edu/matching/lalonde.html	sim
LED	LED Display Domain Data Set	NA	https://archive.ics.uci.edu/ml/datasets/LED+Display+Domain	sim
LENSES	Lenses Dataset	24	https://archive.ics.uci.edu/ml/datasets/Lenses	sim
LIVER	Liver Disorders Dataset	345	https://archive.ics.uci.edu/ml/datasets/liver+disorders	sim
LUNG	Lung Cancer Data Set	32	http://archive.ics.uci.edu/ml/datasets/Lung+Cancer	sim
MESOTHELIOMA	Mesothelioma Disease Data Set	324	https://archive.ics.uci.edu/ml/datasets/Mesotheliomaâ€™s+disease+data+set+	sim
MICE	Mice Protein Expression	1080	https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression	sim
MIMIC	MIMIC-III Critical Care Database		https://mimic.physionet.org/about/mimic	sim
MSU	Michigan State University			não
MU284	The MU284 population	284	https://rdrr.io/cran/sampling/man/MU284.html	sim
MULTIPLE	Multiple Features Dataset	2000	https://archive.ics.uci.edu/ml/datasets/Multiple+Features	sim
MUSHROOM	Mushroom Data Set	8124	https://archive.ics.uci.edu/ml/datasets/Mushroom	sim

Fonte: Elaborado pelo autor.

Sigla	Descrição	Instâncias	Homepage	Público
NCHS	National Center for Health Statistics		https://www.cdc.gov/nchs/index.htm	sim
NLS	National Longitudinal Survey		https://www.bls.gov/nls/	sim
NMES1988	NMES1988: Demand for Medical Care in NMES 1988	4406	https://rdrr.io/cran/AER/man/NMES1988.html	sim
NOAA	National Oceanic and Atmospheric Administration		https://www.ncdc.noaa.gov/cdo-web/	sim
NRI	National Resources Inventory		https://www.nrcs.usda.gov/wps/portal/nrcs/main/national/technical/nra/nri/	sim
NSH	Patients treated at the North Staffordshire Hospital			não
NURSELY	Nursery Data Set	12960	https://archive.ics.uci.edu/ml/datasets/nursery	sim
OMAS	Ohio Medicaid Survey		https://grc.osu.edu/OMAS	sim
OMID	Breast Cancer Dataset of OMid Hospital of Mashahad	217		não
ONSUK	Office for National Statistics UK		https://www.ons.gov.uk/census	sim
PENDIGITS	Pen-Based Recognition of Handwritten Digits Data Set	10992	https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits	sim
PIMA	The Pima Indians	768	http://networkrepository.com/pima-indians-diabetes.php	sim
PSID	Panel Study on Income Dynamics		https://psidonline.isr.umich.edu/	sim
RANDOM	Random generated			não
RCDI	Research and Development in Canadian Industry		https://www.statcan.gc.ca/eng/survey/business/4201	sim
SEEDS	Seeds Data Set	210	https://archive.ics.uci.edu/ml/datasets/seeds	sim
SEGMENT	Image Segmentation Data Set	2310	https://archive.ics.uci.edu/ml/datasets/Image+Segmentation	sim
SENSORLESS	Dataset for Sensorless Drive Diagnosis Data Set	58509	https://archive.ics.uci.edu/ml/datasets/dataset+for+sensorless+drive+diagnosis	sim

Fonte: Elaborado pelo autor.

Sigla	Descrição	Instâncias	Homepage	Público
SES	The new Dutch Structure of Earnings Survey			não
SIMULATED	Simulated Function Dataset			não
SINA	Sina Weibo Platform		https://weibo.com/overseas	sim
SIPP	Survey of Income and Program Participation		https://www.census.gov/programs-surveys/sipp.html	sim
SKIN	Skin Segmentation Data Set	245057	https://archive.ics.uci.edu/ml/datasets/skin+segmentation	sim
SOLAR	Solar Flare Data Set	1389	https://archive.ics.uci.edu/ml/datasets/Solar+Flare	sim
SONAR	Connectionist Bench (Sonar, Mines vs. Rocks) Data Set	208	http://networkrepository.com/sonar.php	sim
SOYBEAN	Soybean (Large) Data Set	307	https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)	sim
SPECT	SPECT Heart Data Set	267	https://archive.ics.uci.edu/ml/datasets/SPECT+Heart	sim
STATLOG	Statlog (Australian Credit Approval) Data Set	690	http://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval)	sim
THYROID	Thyroid Disease Data Set	7200	http://archive.ics.uci.edu/ml/datasets/thyroid+disease	sim
TICTAC	Tic-Tac-Toe Endgame Data Set	958	https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame	sim
UKLF	Uk Labour Force Survey		https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/labourforcesurvey	sim
UKPN	UK Power Networks		https://innovation.ukpowernetworks.co.uk/open-data/	sim
USCM	US Census of Manufactures		https://www.census.gov/econ/www/mancen.html	sim
WAVEFORM21	Waveform Database Generator (Version 1) Data Set	5000	https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+1%29	sim
WDBC	Breast Cancer Wisconsin (Diagnostic) Data Set	569	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)	sim
WEB	Data extrected from Web			não
WES	Workplace and Employee Survey		https://crdcn.org/datasets/wes-workplace-and-employee-survey	sim
WINE	Wine Chemical Analysis	178	http://archive.ics.uci.edu/ml/datasets/Wine	sim

Fonte: Elaborado pelo autor

Sigla	Descrição	Instâncias	Homepage	Público
WIRELESS	Wireless Indoor Localization Data Set	2000	https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization	sim
WPBC	The Wisconsin Prognostic Breast Cancer	198	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)	sim
YEAST	Yeast Data Set	1484	https://archive.ics.uci.edu/ml/datasets/Yeast	sim
ZOO	Zoo Data Set	101	https://archive.ics.uci.edu/ml/datasets/Zoo	sim

Fonte: Elaborado pelo autor.

Referências

ABADIE, A.; IMBENS, G. **A martingale representation for matching estimators**. Cambridge, MA: National Bureau of Economic Research, fev. 2009.

ABNANE, I.; IDRI, A.; ABRAN, A. Fuzzy case-based-reasoning-based imputation for incomplete data in software engineering repositories. **Journal of Software: Evolution and Process**, v. 32, n. 9, set. 2020.

AGBO, B.; QIN, Y.; HILL, R. **Best fit missing value imputation (BFMVI) algorithm for incomplete data in the internet of things**Proceedings of the 5th International Conference on Internet of Things, Big Data and Security. **Anais**... In: 5TH INTERNATIONAL CONFERENCE ON INTERNET OF THINGS, BIG DATA AND SECURITY. SCITEPRESS - Science and Technology Publications, 7 maio 2020

AIEB, A. et al. A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria. **Heliyon**, v. 5, n. 2, p. e01247, 21 fev. 2019.

AL-ZOUBI, A.; TATAS, K.; KYRIACOU, C. **Design space exploration of the KNN imputation on FPGA**2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST). **Anais**... In: 2018 7TH INTERNATIONAL CONFERENCE ON MODERN CIRCUITS AND SYSTEMS TECHNOLOGIES (MOCAST). IEEE, 7 maio 2018

ALBAYRAK, M.; TURHAN, K.; KURT, B. **A missing data imputation approach using clustering and maximum likelihood estimation**2017 Medical Technologies National Congress (TIPTEKNO). **Anais**... In: 2017 MEDICAL TECHNOLOGIES NATIONAL CONGRESS (TIPTEKNO). IEEE, 12 out. 2017

ALJUAID, T.; SASI, S. **Proper imputation techniques for missing values in data sets**2016 International Conference on Data Science and Engineering (ICDSE). **Anais**... In: 2016 INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ENGINEERING (ICDSE). IEEE, 23 ago. 2016a

ALJUAID, T.; SASI, S. **Intelligent imputation technique for missing values**2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). **Anais**... In: 2016 INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATIONS AND INFORMATICS (ICACCI). IEEE, 21 set. 2016b

ALLISON, P. D. Missing data. **Missing data**, 2001.

ALUJA-BANET, T.; DAUNIS-I-ESTADELLA, J.; PELLICER, D. GRAFT, a complete system for data fusion. **Computational statistics & data analysis**, v. 52, n. 2, p. 635-649, out. 2007.

ALUJA-BANET, T.; THIÔ, S. Survey Data Fusion. **Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique**, v. 72, n. 1, p. 20-36, out. 2001.

ANAGNOSTOPOULOS, C.; TRIANTAFILLOU, P. **Scaling out big data missing value**

imputations: Pythia vs. godzilla Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. **Anais**... In: KDD' ' '14: THE 20TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. New York, NY, USA: ACM, 24 ago. 2014

ANDREIS, F.; BONETTI, M. A proposal for a two-step sampling design to oversample units responding to prescribed characteristics. **Environmental and ecological statistics**, v. 25, n. 1, p. 139-154, mar. 2018.

ANDRIDGE, R. R.; LITTLE, R. J. A. A Review of Hot Deck Imputation for Survey Non-response. **International statistical review = Revue internationale de statistique**, v. 78, n. 1, p. 40-64, abr. 2010.

ANWAR, T. et al. **A study on missing values imputation using K-Harmonic means algorithm: Mixed datasets**INTERNATIONAL CONFERENCE ON SCIENCE AND APPLIED SCIENCE (ICSAS) 2019. **Anais**...: AIP Conference Proceedings. In: INTERNATIONAL CONFERENCE ON SCIENCE AND APPLIED SCIENCE (ICSAS) 2019. AIP Publishing, 2019

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v. 5, n. 2, 2006.

ARAÚJO, S. M. A.; DE SOUZA, F. S. H.; MATEUS, G. R. Alocação de Recursos para Redes Virtuais com Seleção de Método de Resolução via Aprendizado de Máquina. **Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**, p. 211, 2020.

ARBUCKLE, J. L. Ibm amos 21 user's guide. chicago: SPSS Inc. Retrieved fromftp. **public. dhe. ibm. com/software/analytics/spss/documentation/amos/21.0/en/Manuals/IBM_SPSS_Amos_Users_Guide. pdf**, 2012.

BANKHOFER, U.; JOENSSEN, D. W. On limiting donor usage for imputation of missing data via hot deck methods. In: SPILIOPOULOU, M.; SCHMIDT-THIEME, L.; JANNING, R. (Eds.). **Data analysis, machine learning and knowledge discovery**. Studies in classification, data analysis, and knowledge organization. Cham: Springer International Publishing, 2014. p. 3-11.

BANO, S.; KHAN, M. N. A. A survey of data clustering methods. **International Journal of Advanced Science and Technology**, v. 113, p. 133-142, 30 abr. 2018.

BATISTA, G. E. A. P. A.; MONARD, M. C. An analysis of four missing data treatment methods for supervised learning. **Applied Artificial Intelligence**, v. 17, n. 5-6, p. 519-533, maio 2003.

BEAUMONT, J.-F.; BOCCI, C. Variance estimation when donor imputation is used to fill in missing values. **Canadian Journal of Statistics**, v. 37, n. 3, p. 400-416, set. 2009.

BERGER, Y. G.; ESCOBAR, E. L. Variance estimation of imputed estimators of change for repeated rotating surveys. **International Statistical Review**, v. 85, n. 3, p. 421-438, dez. 2017.

BETHLEHEM, J. Applied survey methods: A statistical perspective. **Applied survey methods: A statistical perspective**, 2009.

BISHOP, J. A.; FORMBY, J. P.; THISTLE, P. D. Can earnings equations estimates improve CPS hot-deck imputations? **Journal of labor research**, v. 24, n. 1, p. 153-159, mar. 2003.

BOISTARD, H.; CHAUVET, G.; HAZIZA, D. Doubly robust inference for the distribution function in the presence of missing survey data. **Scandinavian Journal of Statistics**, v. 43, n. 3, p. 683-699, set. 2016.

BRAMER, M. Principles of data mining. **Principles of data mining**, 2007.

BRANDRIET, A.; HOLME, T. Methods for Addressing Missing Data with Applications from ACS Exams. **Journal of chemical education**, v. 92, n. 12, p. 2045-2053, 8 dez. 2015.

BREIMAN, L. et al. **Classification and regression trees**. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

BROWN, M. L.; KROS, J. F. Data mining and the impact of missing data. **Industrial Management & Data Systems**, v. 103, n. 8, p. 611-621, 2003a.

BROWN, M. L.; KROS, J. F. The impact of missing data on data mining. In: WANG, J. (Ed.). **Data mining: opportunities and challenges**. [s.l.] IGI Global, 2003b. p. 174-198.

BUJLOW, T.; RIAZ, T.; PEDERSEN, J. M. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. **International Journal of Computer Applications**, v. 117, n. 16, p. 18-21, 2015.

BUSSAB, W. O.; MORETTIN, P. A. Estatística básica. **Estatística básica**, 2010.

BUUREN, S. **Flexible imputation of missing data**. [s.l.] Chapman and Hall/CRC, 2012. v. 20125245

CAI, Z.; HEYDARI, M.; LIN, G. Iterated local least squares microarray missing value imputation. **Journal of Bioinformatics and Computational Biology**, v. 4, n. 5, p. 935-957, out. 2006.

CARDOSO, V. et al. Systematic review of mixed methods: method of research for the incorporation of evidence in nursing. **Texto & Contexto - Enfermagem**, v. 28, 2019.

CARTWRIGHT, M. H.; SHEPPERD, M. J.; SONG, Q. **Dealing with missing software project data**Proceedings. 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (IEEE Cat. No. 03EX717). **Anais**...2004

CARVALHO, L. et al. On the relevance of data science for flight delay research: a systematic review. **Transport Reviews**, p. 1-30, 24 dez. 2020.

CASLETON, E.; OSTHUS, D.; VAN BUREN, K. Imputation for multisource data with comparison and assessment techniques. **Applied Stochastic Models in Business and**

Industry, v. 34, n. 1, p. 44-60, jan. 2018.

CASTANEDA, R. et al. **Aprimorando processos de imputacão multivariada de dados com workflows**Proceedings of the 23rd Brazilian symposium on Databases. **Anais**...2008

CHAUVET, G.; HAZIZA, D. Fully efficient estimation of coefficients of correlation in the presence of imputed survey data. **Canadian Journal of Statistics**, v. 40, n. 1, p. 124-149, mar. 2012.

CHEN, B.-W. et al. Privacy-preserved big data analysis based on asymmetric imputation kernels and multiside similarities. **Future Generation Computer Systems**, v. 78, n. 78, p. 859-866, nov. 2016.

CHEN, C. et al. Additive integer-valued data envelopment analysis with missing data: A multi-criteria evaluation approach. **Plos One**, v. 15, n. 6, p. e0234247, 11 jun. 2020.

CHEN, J.; SHAO, J. Jackknife Variance Estimation for Nearest-Neighbor Imputation. **Journal of the American Statistical Association**, v. 96, n. 453, p. 260-269, mar. 2001.

CHEN, M.; CHEN, C. Optimize Neural Network Algorithm of Missing Value Imputation for Clustering Chocolate Product Types Following“ STEAMS” Methodology. **CATA**, p. 230, 2020.

CHENG, K. O.; LAW, N. F.; SIU, W. C. Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. **Pattern recognition**, v. 45, n. 4, p. 1281-1289, abr. 2012.

CHHABRA, G.; VASHISHT, V.; RANJAN, J. **Missing Value Imputation using Hybrid K-Means and Association Rules**2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). **Anais**... In: 2018 INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATION CONTROL AND NETWORKING (ICACCCN). IEEE, 12 out. 2018

CHI, J. T.; CHI, E. C.; BARANIUK, R. G. *k* -POD: A Method for*k* -Means Clustering of Missing Data. **The American Statistician**, v. 70, n. 1, p. 91-99, 2 jan. 2016.

CHIU, H. Y.; SEDRANSK, J. A bayesian procedure for imputing missing values in sample surveys. **Journal of the American Statistical Association**, v. 81, n. 395, p. 667-676, set. 1986.

CHRISTOPHER, S. Z. et al. **Missing Value Analysis of Numerical Data using Fractional Hot Deck Imputation**2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). **Anais**... In: 2019 3RD INTERNATIONAL CONFERENCE ON INFORMATICS AND COMPUTATIONAL SCIENCES (ICICOS). IEEE, 29 out. 2019

CLEMENTINO, T. Modelo para avaliação da Qualidade do Valor Ambiental Percebido: adoção de Aprendizagem de Máquina para auxílio à tomada de decisões estéticas em projetos de embalagens ecologicamente orientadas. 2020.

CONEGLIAN, C. S.; GONÇALEZ, P. R. V. A.; SANTARÉM SEGUNDO, J. E. O profissional da informação na era do big data. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 50, p. 128, 6 set. 2017.

CONTI, P. L.; MARELLA, D.; SCANU, M. Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. **Computational statistics & data analysis**, v. 53, n. 2, p. 354-365, dez. 2008.

COUTINHO, W.; WAAL, T. DE; SHLOMO, N. Calibrated Hot-Deck Donor Imputation Subject to Edit Restrictions. **Journal of official statistics**, v. 29, n. 2, p. 299-321, 1 set. 2013.

D ANDRÉA, C. Estratégias de produção e organização de informações na web: conceitos para a análise de documentos na internet. **Ciência da Informação**, v. 35, n. 3, p. 39-44, dez. 2006.

DA SILVA, E. B. Agrupamento Semi-Supervisionado de Documentos XML. **Agrupamento Semi-Supervisionado de Documentos XML**, 2006.

DAHL, F. A. Convergence of random -nearest-neighbour imputation. **Computational statistics & data analysis**, v. 51, n. 12, p. 5913-5917, ago. 2007.

DAVENPORT, T.; HARRIS, J. Competing on analytics: Updated, with a new introduction: The new science of winning. **Competing on analytics: Updated, with a new introduction: The new science of winning**, 2017.

DAVID, M. et al. Alternative methods for CPS income imputation. **Journal of the American Statistical Association**, v. 81, n. 393, p. 29-41, mar. 1986.

DE WAAL, T.; COUTINHO, W.; SHLOMO, N. Calibrated hot deck imputation for numerical data under edit restrictions. **Journal of Survey Statistics and Methodology**, v. 5, n. 3, p. 372-397, set. 2017.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data Via the*EM* Algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 39, n. 1, p. 1-22, set. 1977.

DI NUOVO, A. G. Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario. **Expert systems with applications**, v. 38, n. 6, p. 6793-6797, jun. 2011.

DI ZIO, M. et al. Bayesian networks for imputation. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, v. 167, n. 2, p. 309-322, maio 2004.

DIETTERICH, T. G. Ensemble Methods in Machine Learning. In: GOOS, G.; HARTMANIS, J.; VAN LEEUWEN, J. (Eds.). **Multiple Classifier Systems**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. v. 1857p. 1-15.

DING, Y.; ROSS, A. **When data goes missing: methods for missing score imputation in biometric fusion** (B. V. K. Vijaya Kumar, S. Prabhakar, A. A. Ross,

Eds.)Biometric Technology for Human Identification VII. **Anais**...: SPIE Proceedings. In: SPIE DEFENSE, SECURITY, AND SENSING. SPIE, 5 abr. 2010

DORESWAMY; GAD, I.; MANJUNATHA, B. R. **Performance evaluation of predictive models for missing data imputation in weather data**2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). **Anais**... In: 2017 INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATIONS AND INFORMATICS (ICACCI). IEEE, 13 set. 2017

DURRANT, G. B. Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. **International journal of social research methodology**, v. 12, n. 4, p. 293-304, out. 2009.

DYBA, T.; KITCHENHAM, B. A.; JORGENSEN, M. Evidence-based software engineering for practitioners. **IEEE Software**, v. 22, n. 1, p. 58-65, jan. 2005.

ENDERS, C. K. Applied missing data analysis. **Applied missing data analysis**, 2010.

FAHAD, A. et al. A survey of clustering algorithms for big data: taxonomy and empirical analysis. **IEEE transactions on emerging topics in computing**, v. 2, n. 3, p. 267-279, set. 2014.

FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. **Experimental analysis of methods for imputation of missing values in databases** (K. L. Priddy, Ed.)Intelligent Computing: Theory and Applications II. **Anais**...: SPIE Proceedings. In: DEFENSE AND SECURITY. SPIE, 12 abr. 2004

FARNELL, J.; DARBY, P. Administrative data informed donor imputation in the Australian Census of Population and Housing. **Statistical Journal of the IAOS**, p. 1-8, 1 jan. 2020.

FENG, X. et al. Automatic instance selection via locality constrained sparse representation for missing value estimation. **Knowledge-Based Systems**, v. 85, p. 210-223, set. 2015.

FERLIN, C. **Imputação Multivariada: Uma Abordagem em Cascata**. Doctoral dissertation—[s.l.] COPPE/UFRJ, 1 ago. 2008.

FONTES, N. R.; DA SILVA, G.; DE ALMEIDA, J. W. R. UTILIZACÃO DO BIG DATA PARA OBTER VANTAGENS COMPETITIVAS. **Revista Cientfica on-line-Tecnologia, Gestão e Humanismo**, v. 6, n. 1, 2016.

FORD, B. L. An overview of hot-deck procedures. **Incomplete data in sample surveys**, v. 2, p. 185-207, 1983.

FOX, J. A. Missing data problems in the SHR. **Homicide studies**, v. 8, n. 3, p. 214-254, ago. 2004.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. **Journal-Japanese Society For Artificial Intelligence**, v. 14, n. 771-780, p. 1612, 1999.

FURUKAWA, T.; OHNISHI, S.; YAMANOI, T. **A study on a fuzzy clustering for mixed numerical and categorical incomplete data**2013 International Conference on Fuzzy Theory and Its Applications (iFUZZY). **Anais**... In: 2013 INTERNATIONAL CONFERENCE ON FUZZY THEORY AND ITS APPLICATIONS (IFUZZY). IEEE, 6 dez. 2013

GARCIARENA, U.; SANTANA, R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. **Expert systems with applications**, v. 89, p. 52-65, dez. 2017.

GAUTAM, C.; RAVI, V. **Evolving clustering based data imputation**2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]. **Anais**... In: 2014 INTERNATIONAL CONFERENCE ON CIRCUIT, POWER AND COMPUTING TECHNOLOGIES (ICCPCT). IEEE, 20 mar. 2014

GELMAN, A.; HILL, J. **Data Analysis Using Regression and Multilevel/Hierarchical Models**. Cambridge: Cambridge University Press, 2006.

GOLD, M. S.; BENTLER, P. M. Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. **Structural Equation Modeling: A Multidisciplinary Journal**, v. 7, n. 3, p. 319-355, jul. 2000.

GOMES, M. et al. Multiple imputation methods for handling missing data in cost-effectiveness analyses that use data from hierarchical studies: an application to cluster randomized trials. **Medical Decision Making**, v. 33, n. 8, p. 1051-1063, 1 ago. 2013.

GRAHAM, J. W. Missing data: Analysis and design. **Missing data: Analysis and design**, 2012.

GRAHAM, J. W.; CUMSILLE, P. E.; SHEVOCK, A. E. Methods for handling missing data. **Handbook of Psychology, Second Edition**, v. 2, 2012.

GRAHAM, J. W.; DONALDSON, S. I. Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. **Journal of Applied Psychology**, v. 78, n. 1, p. 119, 1993.

HASLER, C.; TILLÉ, Y. Balanced*k* -nearest neighbour imputation. **International Journal of Statistics and Applications**, v. 50, n. 6, p. 1310-1331, nov. 2016.

HATHAWAY, R. J. et al. **Relational data clustering with incomplete data** (K. L. Priddy, P. E. Keller, P. J. Angeline, Eds.)Applications and Science of Computational Intelligence IV. **Anais**...: SPIE Proceedings. In: AEROSPACE/DEFENSE SENSING, SIMULATION, AND CONTROLS. SPIE, 21 mar. 2001

HAZIZA, D.; RAO, J. N. K. Variance Estimation in Two-Stage Cluster Sampling under Imputation for Missing Data. **Journal of statistical theory and practice**, v. 4, n. 4, p. 827-844, dez. 2010.

HEGAMIN-YOUNGER, C.; FORSYTH, R. A Comparison of Four Imputation Procedures in a Two-Variable Prediction System. **Educational and psychological measurement**,

v. 58, n. 2, p. 197-210, abr. 1998.

HEITJAN, D. F.; LANDIS, J. R. Assessing Secular Trends in Blood Pressure: A Multiple-Imputation Approach. **Journal of the American Statistical Association**, v. 89, n. 427, p. 750-759, set. 1994.

HOKAYEM, C.; BOLLINGER, C.; ZILIAK, J. P. The role of CPS nonresponse in the measurement of poverty. **Journal of the American Statistical Association**, v. 110, n. 511, p. 935-945, 3 jul. 2015.

HOLLY, C.; SALMOND, S.; SAIMBERT, M. (EDS.). **Comprehensive systematic review for advanced practice nursing**. New York, NY: Springer Publishing Company, 2017.

HRUSCHKA, E. R.; HRUSCHKA, E. R.; EBECKEN, N. F. F. Evaluating a Nearest-Neighbor Method to Substitute Continuous Missing Values. In: GEDEON, T. (TOM) D.; FUNG, L. C. C. (Eds.). **AI 2003: Advances in Artificial Intelligence: 16th Australian Conference on AI, Perth, Australia, December 3-5, 2003. Proceedings**. Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. v. 2903p. 723-734.

HRUSCHKA, E. R.; JR, E. H.; EBECKEN, N. F. F. A Nearest-Neighbor Method as a Data Preparation Tool for a Clustering Genetic Algorithm. **SBBD**, p. 319, 2003.

HRUSCHKA JR., E. R.; EBECKEN, N. F. F. Missing values prediction with K2. **Intelligent Data Analysis**, v. 6, n. 6, p. 557-566, 27 dez. 2002.

HUANG, X.; DU, M. Research and implementation of animations evaluation system. **Cluster computing**, v. 20, n. 2, p. 1047-1062, jun. 2017.

HUANG, X.; ZHU, Q. A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets. **Pattern recognition letters**, v. 23, n. 13, p. 1613-1622, nov. 2002.

HUANG, Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. **DMKD**, v. 3, n. 8, p. 34-39, 1997.

HUISMAN, M. Imputation of missing item responses: Some simple techniques. **Quality and Quantity**, v. 34, n. 4, p. 331 - 351, 1 jan. 2000.

HUNT, L. A. Missing data imputation and its effect on the accuracy of classification. In: PALUMBO, F.; MONTANARI, A.; VICHI, M. (Eds.). **Data Science**. Studies in classification, data analysis, and knowledge organization. Cham: Springer International Publishing, 2017. p. 3-14.

HUYGHUES-BEAUFOND, N. et al. Robust and automatic data cleansing method for short-term load forecasting of distribution feeders. **Applied energy**, v. 261, p. 114405, mar. 2020.

IACUS, S. M.; PORRO, G. Missing data imputation, matching and other applications of random recursive partitioning. **Computational statistics & data analysis**, v. 52, n. 2, p.

773-789, out. 2007.

ICHIHASHI, H. et al. **Fuzzy c-Means Classifier with Deterministic Initialization and Missing Value Imputation**2007 IEEE Symposium on Foundations of Computational Intelligence. **Anais**... In: 2007 IEEE SYMPOSIUM ON FOUNDATIONS OF COMPUTATIONAL INTELLIGENCE. IEEE, 1 abr. 2007

ILANGO, P.; VIJAYAKUMAR, K.; BABU, M. R. Instance driven clustering for the imputation of missing data in KDD. **International Journal of Communication Networks and Distributed Systems**, v. 12, n. 1, p. 69, 2014.

INGRASSIA, S. A likelihood-based constrained algorithm for multivariate normal mixture models. **Statistical methods & applications**, v. 13, n. 2, set. 2004.

INSUWAN, W.; SUKSAWATCHON, U.; SUKSAWATCHON, J. **Improving missing values imputation in collaborative filtering with user-preference genre and singular value decomposition**2014 6th International Conference on Knowledge and Smart Technology (KST). **Anais**... In: 2014 6TH INTERNATIONAL CONFERENCE ON KNOWLEDGE AND SMART TECHNOLOGY (KST). IEEE, 30 jan. 2014

JAIN, A. K.; DUBES, R. C. Algorithms for clustering data. **Englewood Cliffs: Prentice Hall, 1988**, 1988.

JING-TAO, S.; QIU-YU, Z. Completion of multiview missing data based on multi-manifold regularised non-negative matrix factorisation. **Artificial Intelligence Review**, v. 53, n. 7, p. 5411-5428, out. 2020.

JÖNSSON, P.; WOHLIN, C. Benchmarking k-nearest neighbour imputation with homogeneous Likert data. **Empirical Software Engineering**, v. 11, n. 3, p. 463-489, 9 jul. 2007.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric environment**, v. 38, n. 18, p. 2895-2907, jun. 2004.

KAMAKURA, W. A.; WEDEL, M. Statistical Data Fusion for Cross-Tabulation. **Journal of Marketing Research**, v. 34, n. 4, p. 485-498, nov. 1997.

KANG, S.-S.; KOEHLER, K. J.; LARSEN, M. D. Fractional imputation for incomplete two-way contingency tables. **Metrika**, v. 75, n. 5, p. 581-599, jul. 2012.

KHUSNULKHOTIMAH, B.; SUPRAJITNO, H. Adaptive SOMMI (Self Organizing Map Multiple Imputation) base on Variation Weight for Incomplete Data. **2018 International Conference on Sustainable Information Engineering and Technology (SIET)**, p. 82, 2018.

KIANI, K.; SALEEM, K. **K-Nearest Temperature Trends: A Method for Weather Temperature Data Imputation**Proceedings of the 2017 International Conference on Information System and Data Mining' ' - ICISDM ’17. **Anais**... In: THE 2017 INTERNATIONAL CONFERENCE. New York, New York, USA: ACM Press, 1 abr. 2017

KIM, H.; GOLUB, G. H.; PARK, H. Missing value estimation for DNA microarray gene

expression data: local least squares imputation. **Bioinformatics**, v. 21, n. 2, p. 187-198, 15 jan. 2005.

KIM, H. J. et al. Simultaneous Edit-Imputation for Continuous Microdata. **Journal of the American Statistical Association**, v. 110, n. 511, p. 987-999, 3 jul. 2015.

KIM, J. K.; FULLER, W. Fractional hot deck imputation. **Biometrika**, v. 91, n. 3, p. 559-578, 1 set. 2004.

KIM, J. K.; FULLER, W. A.; BELL, W. R. Variance estimation for nearest neighbor imputation for US Census long form data. **Annals of Applied Statistics**, v. 5, n. 2A, p. 824-842, jun. 2011.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering – A systematic literature review. **Information and software technology**, v. 51, n. 1, p. 7-15, jan. 2009.

KITCHENHAM, B. A.; BUDGEN, D.; BRERETON, O. P. **The value of mapping studies – A participant-observer case study**: Electronic workshops in computing. In: 14TH INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING (EASE). BCS Learning & Development, 1 abr. 2010

KU, W. C. et al. A clustering-based approach for data-driven imputation of missing traffic data. In 2016 IEEE Forum on Integrated and Sustainable Transportation Systems (FISTS). p. 1-6, 2016.

KUPPUSAMY, V.; PARAMASIVAM, I. Integrating WLI fuzzy clustering with grey neural network for missing data imputation. **International Journal of Intelligent Enterprise**, v. 4, n. 1, p. 103-127, 2017.

LACERDA, M. C. C. DASH sobre OpenFlow: estimando métricas de QoS a partir da rede. 2020.

LAI, X.; ZHANG, L.; LIU, X. Takagi-Sugeno Modeling of Incomplete Data for Missing Value Imputation With the Use of Alternate Learning. **IEEE Access**, v. 8, p. 83633-83644, 2020.

LAROSE, D. T.; LAROSE, C. D. **Data mining and predictive analytics**. [s.l.] John Wiley & Sons, 2015.

LARSEN, M. D.; HUCKETT, J. C. **Measuring disclosure risk for multimethod synthetic data generation**2010 IEEE Second International Conference on Social Computing. **Anais**... In: 2010 IEEE SECOND INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING (SOCIALCOM). IEEE, 20 ago. 2010

LI, D. et al. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method. In: TSUMOTO, S. et al. (Eds.). **Rough sets and current trends in computing**. Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. v. 3066p. 573-579.

LI, D.; GU, H.; ZHANG, L. A fuzzy c-means clustering algorithm based on nearest-

neighbor intervals for incomplete data. **Expert systems with applications**, v. 37, n. 10, p. 6942-6947, out. 2010.

LIAO, Z. et al. **Missing Data Imputation: A Fuzzy K-means Clustering Algorithm over Sliding Window**2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. **Anais**... In: 2009 SIXTH INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY. IEEE, 14 ago. 2009

LICHMAN, M. UCI Machine Learning Repository. Irvine, University of California, Irvine, School of Information and Computer Sciences.(2013). 2018.

LIN, T.-R.; YANG, C.-J.; CHIANG, I.-J. **Improvement of prognostic models for ESRD mortality by the bootstrap method with random hot deck imputation**2014 IEEE International Conference on Granular Computing (GrC). **Anais**... In: 2014 IEEE INTERNATIONAL CONFERENCE ON GRANULAR COMPUTING (GRC). IEEE, 22 out. 2014

LINDBERG, E. et al. Estimation of Tree Lists from Airborne Laser Scanning Using Tree Model Clustering and k-MSN Imputation. **Remote sensing**, v. 5, n. 4, p. 1932-1955, 19 abr. 2013.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with Missing Data**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2002.

LORENZO-SEVA, U.; VAN GINKEL, J. R. Multiple Imputation of missing values in exploratory factor analysis of multidimensional scales: estimating latent trait scores. **Anales de Psicología**, v. 32, n. 2, p. 596, 3 abr. 2016.

LUO, W. et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. **Journal of Medical Internet Research**, v. 18, n. 12, p. e323, 16 dez. 2016.

MACENA, J.; PIRES, F.; PESSOA, M. Operaçao Lovelace: uma abordagem ludica para introduçao de aprendizagem em algoritmos. **XIX SBGames**, 2020.

MAGNANI, M.; MONTESI, D. **A new reparation method for incomplete data in the context of supervised learning**International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. **Anais**... In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY: CODING AND COMPUTING, 2004. PROCEEDINGS. ITCC 2004. IEEE, 5 abr. 2004

MAITI, T.; MILLER, C. P.; MUKHOPADHYAY, P. K. Neural network imputation: An experience with the national resources inventory survey. **Journal of Agricultural, Biological and Environmental Statistics**, v. 13, n. 3, p. 255-269, set. 2008.

MANCINI, M. Estudos de Revisão Sistemática: um guia para síntese criteriosa da evidência científica. 2007. 2017.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008.

MASHREGHI, Z.; LÉGER, C.; HAZIZA, D. Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. **Canadian Journal of Statistics**, v. 42, n. 1, p. 142-167, mar. 2014.

MCGEE, M.; BERGASA, N. V. Analysis of a pilot study for amelioration of itching in liver disease. **The American Statistician**, v. 60, n. 4, p. 303-308, nov. 2006.

MEDINA, J. V. DE. **Algoritmos Genéticos e Redes Kohonen na Complementação de Dados Ausentes**. Undergraduate thesis—[s.l.] Universidade do Estado do Rio de Janeiro - UERJ, 20 out. 2012.

MIGDADY, H.; AL-TALIB, M. M. An enhanced fuzzy K-means clustering with application to missing data imputation. **Electronic Journal of Applied Statistical Analysis**, v. 11, n. 2, p. 674-686, 2018.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. Sistemas Inteligentes, Rezende, SO. p. 89-114, 2003.

MORENO, J. O valor económico da informacão na sociedade em rede. **Observatorio (OBS*)**, v. 9, n. 2, p. 1-28, 2015.

MULLER, D.; TE, Y.-F.; JAIN, P. **Improving data quality through high precision gender categorization**2017 IEEE International Conference on Big Data (Big Data). **Anais**... In: 2017 IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA). IEEE, 11 dez. 2017

MUNAFÒ, M. R. et al. A manifesto for reproducible science. **Nature human behaviour**, v. 1, n. 1, p. 0021, 10 jan. 2017.

MUNN, Z. et al. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. **BMC Medical Research Methodology**, v. 18, n. 1, p. 143, 19 nov. 2018.

MYERS, T. A. Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data. **Communication methods and measures**, v. 5, n. 4, p. 297-310, out. 2011.

MYRTVEIT, I.; STENSRUD, E.; OLSSON, U. H. Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. **IEEE Transactions on Software Engineering**, v. 27, n. 11, p. 999-1013, 2001.

NANNI, L.; LUMINI, A.; BRAHNAM, S. A classifier ensemble approach for the missing feature problem. **Artificial Intelligence in Medicine**, v. 55, n. 1, p. 37-50, maio 2012.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in neurorobotics**, v. 7, p. 21, 4 dez. 2013.

NIGHTINGALE, A. A guide to systematic literature reviews. **Surgery (Oxford)**, v. 27, n. 9, p. 381-384, set. 2009.

NIKFALAZAR, S. et al. **A new iterative fuzzy clustering algorithm for multiple**

imputation of missing data2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). **Anais**... In: 2017 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE). IEEE, 9 jul. 2017

NIKFALAZAR, S. et al. Missing data imputation using decision trees and fuzzy clustering with iterative learning. **Knowledge and information systems**, v. 62, n. 6, p. 2419-2437, jun. 2020.

NILOOFAR, P.; GANJALI, M. A new multivariate imputation method based on Bayesian networks. **Journal of applied statistics**, v. 41, n. 3, p. 501-518, 4 mar. 2014.

NILOOFAR, P.; GANJALI, M.; FARID ROHANI, M. R. Performance evaluation of imputation based on Bayesian Networks. **Sankhya B**, v. 75, n. 1, p. 90-111, maio 2013.

NISHANTH, K. J.; RAVI, V. Probabilistic neural network based categorical data imputation. **Neurocomputing**, v. 218, p. 17-25, dez. 2016.

NORDHOLT, E. S. Imputation: methods, simulation experiments and practical examples. **International Statistical Review / Revue Internationale de Statistique**, v. 66, n. 2, p. 157, ago. 1998.

OLIVEIRA, E. N. F. Uma abordagem semissupervisionada para classificação de pastagens usando séries temporais de NDVI. 2020.

PAIK, M.; LARSEN, M. D. Fractional regression hot deck imputation weight adjustment. **Communications in Statistics - Simulation and Computation**, v. 42, n. 7, p. 1514-1532, ago. 2013.

PANDEY, K. K.; SHUKLA, D. A study of clustering taxonomy for big data mining with optimized clustering MapReduce model. **decision making**, v. 1, n. 5, p. 30, 2019.

PANG, S.; GONG, J. C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. **Systems Engineering - Theory & Practice**, v. 29, n. 12, p. 94-104, dez. 2009.

PATTANODOM, M.; IAM-ON, N.; BOONGOEN, T. **Clustering data with the presence of missing values by ensemble approach**2016 Second Asian Conference on Defence Technology (ACDT). **Anais**... In: 2016 SECOND ASIAN CONFERENCE ON DEFENCE TECHNOLOGY (ACDT). IEEE, 21 jan. 2016

PEARL, J. Probabilistic reasoning in intelligent systems: networks of plausible inference. **Probabilistic reasoning in intelligent systems: networks of plausible inference**, 2014.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A. Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, v. 12, p. 2825-2830, 2011.

PENNY, K.; CHESNEY, T. Mining trauma injury data with imputed values. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, v. 2, n. 4, p. 246-254, nov. 2009.

PENNY, K. I.; ASHRAF, M. Z.; DUFFY, J. C. **The use of hot deck imputation to compare performance of further education colleges**2007 29th International Conference on Information Technology Interfaces. **Anais**... In: 2007 29TH INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY INTERFACES. IEEE, 25 jun. 2007

PETERS, C. L. O.; ENDERS, C. A primer for the estimation of structural equation models in the presence of missing data: Maximum likelihood algorithms. **Journal of Targeting, Measurement and Analysis for Marketing**, v. 11, n. 1, p. 81-95, ago. 2002.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and software technology**, v. 64, p. 1-18, ago. 2015.

PRATI, R. C. Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos. **Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos**, 2006.

PURWAR, A.; SINGH, S. K. DBSCANI: Noise-Resistant Method for Missing Value Imputation. **Journal of Intelligent Systems**, v. 25, n. 3, p. 431-440, 1 jul. 2016.

QIN, Y.; RAO, J. N. K.; REN, Q. Confidence intervals for marginal parameters under imputation for item nonresponse. **Journal of statistical planning and inference**, v. 138, n. 8, p. 2283-2302, ago. 2008.

QIN, Y. S.; ZHANG, J. C. Semi-empirical likelihood confidence intervals for the differences of quantiles with missing data. **Acta mathematica Sinica, English series**, v. 25, n. 5, p. 845-854, maio 2009.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, v. 1, n. 1, p. 81-106, mar. 1986.

QUINTANA, F. A. Nonparametric Bayesian Analysis for Assessing Homogeneity in k × l Contingency Tables with Fixed Right Margin Totals. **Journal of the American Statistical Association**, v. 93, n. 443, p. 1140, set. 1998.

QUINTANA, F. A.; NEWTON, M. A. Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. **Journal of Computational and Graphical Statistics**, v. 9, n. 4, p. 711-737, 2000.

RAGHUNATHAN, T. E. What do we do with missing data? Some options for analysis of incomplete data. **Annual review of public health**, v. 25, p. 99-117, 2004.

RAHMAN, M. G.; ISLAM, M. Z. kDMI: A Novel Method for Missing Values Imputation Using Two Levels of Horizontal Partitioning in a Data set. In: MOTODA, H. et al. (Eds.). **Advanced data mining and applications**. Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. v. 8347p. 250-263.

RAHMAN, M. G.; ISLAM, M. Z. Missing value imputation using a fuzzy clustering-based EM approach. **Knowledge and information systems**, v. 46, n. 2, p. 389-422, fev. 2016.

RAJA, P. S.; SASIREKHA, K.; THANGAVEL, K. A Novel Fuzzy Rough Clustering Parameter-based missing value imputation. **Neural Computing and Applications**, v. 32, n. 14, p. 10033-10050, jul. 2020.

RAJA, P. S.; THANGAVEL, K. Missing value imputation using unsupervised machine learning techniques. **Soft computing**, v. 24, n. 6, p. 4361-4392, mar. 2020.

RAO, J. N. K.; SHAO, J. Jackknife variance estimation with survey data under hot deck imputation. **Biometrika**, v. 79, n. 4, p. 811-822, 1992.

READS, S. Inteligência Artificial: Compreender em Que Consiste a IA e o Que Implica a Aprendizagem das Máquinas. **Inteligência Artificial: Compreender em Que Consiste a IA e o Que Implica a Aprendizagem das Máquinas**, 2017.

REILLY, M.; PEPE, M. S. A mean score method for missing and auxiliary covariate data in regression models. **Biometrika**, v. 82, n. 2, p. 299-314, 1995.

RIEDEL, M.; REGOECZI, W. C. Missing data in homicide research. **Homicide studies**, v. 8, n. 3, p. 163-192, ago. 2004.

ROTH, P. L.; SWITZER, F. S. A monte carlo analysis of missing data techniques in a HRM setting. **Journal of management**, v. 21, n. 5, p. 1003-1023, out. 1995.

ROTH, P. L.; SWITZER, F. S.; SWITZER, D. M. Missing data in multiple item scales: A monte carlo analysis of missing data techniques. **Organizational Research Methods**, v. 2, n. 3, p. 211-232, jul. 1999.

RUBIN, D. B. Inference and missing data. **Biometrika**, v. 63, n. 3, p. 581-592, 1976.

RUBIN, D. B. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. **Proceedings of the survey research methods section of the American Statistical Association**, p. 20, 1978.

RUBIN, D. B. **An overview of multiple imputation**Proceedings of the survey research methods section of the American statistical association. **Anais**...1988

SALZBERG, S. L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. **Machine learning**, v. 16, n. 3, p. 235-240, set. 1994.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210-229, jul. 1959.

SCHÄFER, I. et al. The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. First results from the multicare cohort study. **BMC Health Services Research**, v. 12, p. 89, 3 abr. 2012.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. **Psychological methods**, v. 7, n. 2, p. 147-177, jun. 2002.

SCHENKER, N.; TAYLOR, J. M. G. Partially parametric techniques for multiple imputation. **Computational statistics & data analysis**, v. 22, n. 4, p. 425-446, ago.

1996.

SCHNEIDER, T. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. **Journal of climate**, v. 14, n. 5, p. 853-871, mar. 2001.

SHAO, J.; CHEN, Y.; CHEN, Y. Balanced Repeated Replication for Stratified Multistage Survey Data under Imputation. **Journal of the American Statistical Association**, v. 93, n. 442, p. 819-831, jun. 1998.

SHEN, S. M.; CHOY, S. T. B. The Pre- and Post-1997 Well-Being of Hong Kong Residents. In: SHEK, D. T. L.; CHAN, Y. K.; LEE, P. S. N. (Eds.). **Quality-of-Life Research in Chinese, Western and Global Contexts**. Social indicators research series. Dordrecht: Springer Netherlands, 2005. v. 25p. 231-258.

SHEN, S. M.; LAI, Y. L. Handling incomplete quality-of-life data. **Social Indicators Research**, v. 55, n. 2, p. 121-166, 2001.

SILVA-RAMÍREZ, E.-L.; PINO-MEJÍAS, R.; LÓPEZ-COELLO, M. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. **Applied soft computing**, v. 29, p. 65-74, abr. 2015.

SINGH, D.; REDDY, C. K. A survey on platforms for big data analytics. **Journal of big data**, v. 2, n. 1, p. 8, 2015.

SKINNER, C. J.; RAO, J. N. K. Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors. **Journal of statistical planning and inference**, v. 102, n. 1, p. 149-167, mar. 2002.

SOARES, J. Pré-Processamento em mineração de dados: Um Estudo Comparativo em Complementação. Tese de Doutorado. Engenharia de Sistemas e Computação. UFRJ 2007. 2007a.

SOARES, J. A. Pré-processamento em mineração de dados: Um estudo comparativo em complementação. **Rio de Janeiro, RJ**, 2007b.

SONG, Q.; SHEPPERD, M. A new imputation method for small software project data sets. **Journal of Systems and Software**, v. 80, n. 1, p. 51-62, jan. 2007.

SOUSA, L. M. M.; FIRMINO, C. F. Revisões da literatura científica: tipos, métodos e aplicações em enfermagem. **Revista Portuguesa de Enfermagem de Reabilitação**, v. 1, n. 1, p. 45-54, 2018.

STEINER, S. et al. A Study of Missing Data Imputation in Predictive Modeling of a Wood-Composite Manufacturing Process. **Journal of Quality Technology**, v. 48, n. 3, p. 284-296, jul. 2016.

STRIKE, K.; EL EMAM, K.; MADHAVJI, N. Software cost estimation with incomplete data. **IEEE Transactions on Software Engineering**, v. 27, n. 10, p. 890-908, 2001.

SULLIVAN, D.; ANDRIDGE, R. A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. **Computational statistics & data analysis**, v. 82, p. 173-185, fev. 2015.

SUN, Y. et al. A hybrid approach for the stratified mark-specific proportional hazards model with missing covariates and missing marks, with application to vaccine efficacy trials. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 69, n. 4, p. 791-814, ago. 2020.

TARSITANO, A.; FALCONE, M. Missing-Values Adjustment for Mixed-Type Data. **Journal of probability and statistics**, v. 2011, p. 1-20, 2011.

THIRUKUMARAN, S.; SUMATHI, A. **Missing value imputation techniques depth survey and an imputation Algorithm to improve the efficiency of imputation**2012 Fourth International Conference on Advanced Computing (ICoAC). **Anais**... In: 2012 FOURTH INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING (ICOAC). IEEE, 13 dez. 2012

TIAN, J. et al. Clustering-based multiple imputation via gray relational analysis for missing data and its application to aerospace field. **Thescientificworldjournal**, v. 2013, p. 720392, 2 maio 2013.

TIMM, H.; KRUSE, R. **Fuzzy cluster analysis with missing values**1998 Conference of the North American Fuzzy Information Processing Society - NAFIPS (Cat. No.98TH8353). **Anais**... In: 1998 CONFERENCE OF THE NORTH AMERICAN FUZZY INFORMATION PROCESSING SOCIETY - NAFIPS (CAT. NO.98TH8353). IEEE, 1998

TSAI, C.-W. et al. Big data analytics: a survey. **Journal of big data**, v. 2, n. 1, p. 21, dez. 2015.

TWALA, B.; CARTWRIGHT, M.; SHEPPERD, M. **Comparison of various methods for handling incomplete data in software engineering databases**2005 International Symposium on Empirical Software Engineering, 2005. **Anais**... In: 2005 INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING, 2005. IEEE, 2005

VAN BUUREN, S.; GROOTHUIS-OUDSHOORN, K. Multivariate imputation by chained equations in RJ Stat. [s.d.].

VAZIFEHDAN, M.; MOATTAR, M. H.; JALALI, M. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. **Journal of King Saud University - Computer and Information Sciences**, v. 31, n. 2, p. 175-184, abr. 2019.

VERWER, S.; VAN DEN BRAAK, S.; CHOENNI, S. **Sharing confidential data for algorithm development by multiple imputation** (A. Szalay et al., Eds.)Proceedings of the 25th International Conference on Scientific and Statistical Database Management - SSDBM. **Anais**... In: THE 25TH INTERNATIONAL CONFERENCE. New York, New York, USA: ACM Press, 29 jul. 2013

WANG, S. et al. K-Means Clustering With Incomplete Data. **IEEE access : practical**

innovations, open solutions, v. 7, p. 69162-69171, 2019.

WANG, Y. et al. **Energy efficient neural networks for big data analytics**Proceedings of the conference on Design, Automation & Test in Europe. **Anais**...2014

WATADA, J. et al. **A rough set approach to data imputation and its application to a dissolved gas analysis dataset**2016 Third International Conference on Computing Measurement Control and Sensor Network (CMCSN). **Anais**... In: 2016 THIRD INTERNATIONAL CONFERENCE ON COMPUTING MEASUREMENT CONTROL AND SENSOR NETWORK (CMCSN). IEEE, 20 maio 2016

WAYMAN, J. C. **Multiple imputation for missing data: What is it and how can I use it**Annual Meeting of the American Educational Research Association, Chicago, IL. **Anais**...2003

WAZLAWICK, R. S. **Metodologia de Pesquisa para Ciência da Computação**. [s.l.] Elsevier, 2009.

WINGLEE, M. et al. Handling item nonresponse in the U.S. component of the IEA reading literacy study. **Journal of Educational and Behavioral Statistics**, v. 26, n. 3, p. 343-359, set. 2001.

WOHLIN, C. et al. On the reliability of mapping studies in software engineering. **Journal of Systems and Software**, v. 86, n. 10, p. 2594-2610, out. 2013.

WOHLIN, C. **Guidelines for snowballing in systematic literature studies and a replication in software engineering**Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE' ' ’14. **Anais**... In: THE 18TH INTERNATIONAL CONFERENCE. New York, New York, USA: ACM Press, 13 maio 2014

WOOLDRIDGE, J. M. Introductory econometrics: A modern approach. **Introductory econometrics: A modern approach**, 2016.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. **Annals of Data Science**, v. 2, n. 2, p. 165-193, jun. 2015.

XUE, Y.; LAZAR, N. A. Empirical likelihood-based hot deck imputation methods. **Journal of nonparametric statistics**, v. 24, n. 3, p. 629-646, set. 2012.

YAMAMOTO, T. et al. A Comparative Study on TIBA Imputation Methods in FCMdd-Based Linear Clustering with Relational Data. **Advances in Fuzzy Systems**, v. 2011, p. 1-10, 2011a.

YAMAMOTO, T. et al. **FCMdd-type linear fuzzy clustering for incomplete non-Euclidean relational data**2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011). **Anais**... In: 2011 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE). IEEE, 27 jun. 2011b

YANG, H.; ZHAO, Y. Smoothed jackknife empirical likelihood inference for ROC curves with missing data. **Journal of multivariate analysis**, v. 140, p. 123-138, set. 2015.

YANG, S.; KIM, J. K. Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. **Scandinavian Journal of Statistics**, v. 47, n. 3, p. 839-861, set. 2020.

YANG, X. et al. **Adaptive logistic group Lasso method for predicting the no-reflow among the multiple types of high-dimensional variables with missing data**2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). **Anais**... In: 2016 7TH IEEE INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND SERVICE SCIENCE (ICSESS). IEEE, 26 ago. 2016

YANG, X. et al. **Missing Data Imputation for MIMIC-III using Matrix Decomposition**2019 IEEE International Conference on Healthcare Informatics (ICHI). **Anais**... In: 2019 IEEE INTERNATIONAL CONFERENCE ON HEALTHCARE INFORMATICS (ICHI). IEEE, 10 jun. 2019

YAO, L.; WENG, K.-S. Imputation of incomplete data using adaptive ellipsoids with linear regression. **Journal of Intelligent & Fuzzy Systems**, v. 29, n. 1, p. 253-265, 23 set. 2015.

YELIPE, U.; PORIKA, S.; GOLLA, M. An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. **Computers & Electrical Engineering**, v. 66, p. 487-504, fev. 2018.

YENDURI, S. **An agglomerative clustering methodology for data imputation**Third International Conference on Information Technology: New Generations (ITNG'06). **Anais**... In: THIRD INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY: NEW GENERATIONS (ITNG'06). IEEE, 10 abr. 2006

YENDURI, S.; IYENGAR, S. S. Performance evaluation of imputation methods for incomplete datasets. **International Journal of Software Engineering and Knowledge Engineering**, v. 17, n. 01, p. 127-152, fev. 2007.

YU, W. et al. **Cluster-Based Best Match Scanning for Large-Scale Missing Data Imputation**2017 3rd International Conference on Big Data Computing and Communications (BIGCOM). **Anais**... In: 2017 3RD INTERNATIONAL CONFERENCE ON BIG DATA COMPUTING AND COMMUNICATIONS (BIGCOM). IEEE, 10 ago. 2017

YUNG, W.; RAO, J. N. K. Jackknife Variance Estimation under Imputation for Estimators Using Poststratification Information. **Journal of the American Statistical Association**, v. 95, n. 451, p. 903-915, set. 2000.

ZANDBERG, T.; HUISMAN, M. Missing behavior data in longitudinal network studies: the impact of treatment methods on estimated effect parameters in stochastic actor oriented models. **Social network analysis and mining**, v. 9, n. 1, p. 8, dez. 2019.

ZHANG, C. et al. **Clustering-based Missing Value Imputation for Data Preprocessing**2006 IEEE International Conference on Industrial Informatics. **Anais**... In: 2006 IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL INFORMATICS. IEEE, 16 ago. 2006

ZHANG, C.; MA, Y. Ensemble machine learning: methods and applications. **Ensemble**

machine learning: methods and applications, 2012.

ZHANG, Y. et al. Using multiple imputation to address missing values of hierarchical data. **Journal of Modern Applied Statistical Methods**, v. 16, n. 1, p. 744-752, 1 maio 2017a.

ZHANG, Z. et al. Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level. **Journal of hydrology**, v. 553, p. 384-397, out. 2017b.

ZHANG, Z.; FANG, H. **Multiple- vs Non- or Single-Imputation based Fuzzy Clustering for Incomplete Longitudinal Behavioral Intervention Data.**2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). **Anais**... In: 2016 IEEE FIRST INTERNATIONAL CONFERENCE ON CONNECTED HEALTH: APPLICATIONS, SYSTEMS AND ENGINEERING TECHNOLOGIES (CHASE). IEEE, jun. 2016

ZOBEL, J. Writing for computer science. **Writing for computer science**, 2004.