Dados Ausentes em Problemas de Previsão: Uma Breve Revisão e Estudos de Casos

Giovanni Amormino da Silva Júnior, Alisson Marques Silva, Paulo Eduardo Maciel Almeida
Programa de Pós-Graduação em Modelagem Matemática e Computacional
Centro Federal de Educação Tecnológica de Minas Gerais - CEFET-MG
Belo Horizonte, MG, Brasil
Email: gigio_jr@hotmail.com, alisson@cefetmg.br, pema@cefetmg.br

***Resumo*—Uma grande dificuldade enfrentada no desenvolvimento de aplicações reais que utilizam fluxos de dados para resolver problemas de previsão são os dados ausentes. Apesar de existirem técnicas para reduzir os impactos ocasionados por este problema, a maioria dos sistemas não são modelados de forma preventiva para possibilitar o tratamento adequado deste tipo de ocorrência. Basicamente existem duas formas de lidar com os dados ausentes: i) exclusão, onde toda ou parte da amostra é removida ou ignorada; e ii) imputação, onde o valor ausente é substituído por zero, pela média da variável até a amostra com o problema ou por um valor estimado, onde a variável problemática tem seu valor estimado por algum modelo que, em alguns casos, pode levar em consideração outras variáveis e/ou valores anteriores. Neste contexto este artigo apresenta uma revisão da literatura abordando as principais metodologias utilizadas para tratar o problema de dados ausentes. Além disso, estudos de casos são apresentados para comparação de resultados e definição de situações onde cada tipo de abordagem por ser melhor empregada.**

***Keywords*—Dados Ausentes, Ausência Completamente Aleatória, Ausência Aleatória.**

I. Introdução

Uma grande dificuldade enfrentada no desenvolvimento de aplicações reais que utilizam fluxo de dados para resolver problemas de regressão e/ou classificação são os dados ausentes. Este problema consiste na existência de valores ausentes em uma base de dados e sua ocorrência pode se dar por vários motivos como sensores não confiáveis, observações incompletas, oclusão parcial do sinal desejado, interferência na comunicação, restrição da banda de transmissão de dados, dentre outras. Independente de sua causa, sua ocorrência tem sérias implicações para a extração de conhecimento para aplicações em previsão, identificação de sistemas e classificação [1], [2], [3], [4].

Na literatura é possível encontrar definições e interpretações ligeiramente diferentes sobre os tipos de dados ausentes. Contudo, no geral, os pesquisadores trabalham basicamente com três definições, classificadas de acordo com a forma do desaparecimento do dado, são elas: Ausência Completamente Aleatória (*Missing Completely at Random* – MCAR), Ausência Aleatória (*Missing at Random* – MAR) e Ausência Não Aleatória (*Not Missing at Random* – NMAR)[1] [3], [5], [6], [7], [8], [9], [10], [11].

[1]Também conhecida como *Missing Not Random* - MNAR.

No mecanismo MCAR a probabilidade de uma variável estar ausente não possui qualquer relação com as outras variáveis do conjunto de dados. Em outras palavras, todas as variáveis são independentes entre si e a probabilidade de um valor estar ausente é completamente aleatória. Dessa forma, a probabilidade da ausência de um valor é a mesma que a probabilidade de ausência dos demais valores. Este tipo de ausência é comum quando ocorrem falhas na coleta, um equipamento não funciona corretamente ou fica temporariamente indisponível, por exemplo [3], [5], [6], [7], [8], [11]. Estar em um cenário com mecanismo MCAR, apesar dos problemas ocasionados pela ausência, possui uma vantagem estatística que faz com que a análise permaneça imparcial, uma vez que não há relação entre os valores ausentes e outras variáveis ou amostras na base de dados. Segundo Osman *et. al* [11], este mecanismo pode ser representado estatisticamente por:

$$f(M|Y,\phi) = f(M|\phi)\forall Y,\phi, \tag{1}$$

onde Y e M denotam um vetor de valores de dados observados e um vetor de indicadores de ausências, respectivamente, ϕ é um parâmetro desconhecido e a função f indica a distribuição de probabilidade condicional.

Para o mecanismo MAR a probabilidade de ausência de uma variável continua aleatória e independente de seus valores, contudo essa ausência possui relação com os valores de outras variáveis da amostra [3], [8], [9], [10], [11]. Para Little *et. al* [7], os valores dos dados perdidos podem ser considerados como um efeito aleatório que pode ser previsto por outras variáveis no conjunto de dados. Para Kang [6], a probabilidade de que uma variável esteja ausente depende do conjunto de variáveis observadas, mas não está relacionada aos valores específicos das variáveis ausentes. Uma das consequências disso é fazer com que, em uma amostra, a probabilidade de uma variável estar ausente seja maior que a probabilidade das demais variáveis estarem ausentes [5]. É importante ressaltar que a ausência é condicional a outras variáveis, mas a ocorrência deste tipo de ausência é comum quando existem variáveis que dependem de processamentos ou seleções anteriores para definirem seus valores [9]. Segundo Osman *et. al* [11], este mecanismo pode ser representado estatisticamente por:

$$f(M|Y,\phi) = f(M|Y_{obs},\phi)\forall Y_{miss},\phi, \tag{2}$$

onde Y_{obs} e Y_{miss} são os componentes observados e ausentes da variável de destino Y. O parâmetro desconhecido ϕ pode ser estimado relacionando Y_{obs} com outras variáveis explicativas.

Por fim, para o mecanismo NMAR, onde as características da ausência não se enquadram nos anteriores. No NMAR a probabilidade de uma variável estar ausente depende dos potenciais valores desta variável, ou seja, a ausência está diretamente relacionada ao próprio valor ausente ou a natureza da variável [5], [6], [8], [9], [10], [11]. Normalmente, a obtenção de estimativas precisas para este tipo de mecanismo tende a ser mais complexa, pois não se tem as informações necessárias para especificar corretamente modelos das ausências [9]. Segundo Osman *et. al* [11], este mecanismo pode ser representado estatisticamente por:

$$f(M, Y|\theta, \phi) = f(Y|\theta)f(M|Y, \phi), \tag{3}$$

onde θ é um parâmetro da distribuição de Y que é estimado a partir dos dados observados e ϕ é um parâmetro que caracteriza a distribuição do padrão de ausência.

Diante do exposto, o presente artigo tem por objetivo trazer uma revisão de literatura relacionada a dados ausentes e técnicas de manipulação de dados ausentes. Experimentos computacionais são realizados em bases MCAR e MAR para comparar três técnicas de manipulação de dados ausentes aplicadas a problemas de previsão. São avaliadas a técnica de substituição por zero, média e último valor.

Após esta breve introdução sobre dados ausentes a Seção II aborda algumas das principais técnicas para tratamento dos dados ausentes. A Seção III descreve a metodologia utilizada nos experimentos para avaliar as três técnicas de tratamento de dados em problemas de previsão. Nessa seção são detalhadas as bases de dados utilizadas e os algoritmos utilizados para gerar os dados ausentes, os algoritmos de previsão utilizados nos experimentos e a medida de erro adotada para avaliar o desempenho das técnicas e dos algoritmos. Os resultados computacionais são apresentados e discutidos na Seção IV. Por fim, a Seção V discorre com as considerações finais e as propostas de continuidade.

II. Técnicas de Manipulação dos Dados Ausentes

Há várias maneiras de tratar dados ausentes e, de uma forma intuitiva, é possível citar as seguintes maneiras: exclusão do registro, onde toda a amostra ou parte dela é descartada ou omitida, ou a inserção do valor. A exclusão é recomendada somente em casos onde a quantidade de amostras com dados ausentes é pequena, não haja uma relação entre as variáveis e a deleção não cause um grande impacto no sistema. É importante ressaltar que a remoção de amostras com dados ausentes pode ocasionar perda substancial de informações. Além disso, excluir variáveis da análise devido a ausência de alguns dados significaria utilizar as informações de forma ineficiente, ressaltando a importância de se utilizar métodos confiáveis para se estimar os valores que estão faltando [12]. Assim aplicações que dependem da base completa, como aplicações que trabalham com fluxo de dados, dados *on-line* ou que realizam processamento em tempo real para resolver problemas de regressão ou classificação, por exemplo, dependem destes dados para realizar seus procedimentos [2]. Contudo, nestes casos, a remoção das amostras não pode ser considerada. Por outro lado, a inserção pode ser feita de formas simples, substituindo o valor ausente por zero ou por uma média, por exemplo, ou estimando os valores ausentes utilizando modelos que possam identificar relações entre as variáveis para estimar novos valores, sendo este um exemplo de uma das formas mais complexas [2], [3].

Para Kang [6] o melhor método possível é ter um bom planejamento do estudo e coletar cuidadosamente os dados para evitar o problema. Contudo, o autor também considera não ser incomum existir uma quantidade considerável de dados ausentes em um estudo. É exatamente a partir dessa existência que se torna necessário o estudo dos métodos e técnicas descritos nesta seção. Na Figura 1 é apresentado um diagrama para auxiliar na seleção das técnicas de acordo com o contexto. Basicamente, o diagrama criado por Osman *et. al* [11] divide as técnicas em dois grupos: tradicionais e modernas. Utilizando os mesmos mecanismos para dados ausentes abordados neste trabalho, o autor indica que, para cenários MCAR ou MAR, com uma taxa de até 5% de ausência, seria mais indicado a utilização de técnicas tradicionais, como deleção ou inserção simples, enquanto que, para estes mesmos cenários, com uma taxa superior a 5% de ausência, seria mais indicado técnicas modernas, como imputação múltipla, modelos baseados em processo ou métodos de aprendizado de máquina. Muitos destes métodos serão abordados na sequência.

A. Técnicas de Deleção

Uma das técnicas mais comuns para este problema é a deleção ou omissão de amostras com valores ausentes, fazendo com que se trabalhe apenas com os dados disponíveis, sendo este um método padrão em muitos pacotes de regressão [6], [13], [14]. Muitos pacotes de *softwares* estatísticos a utilizam como solução padrão, apesar de alguns pesquisadores dizerem que esta técnica pode produzir resultados tendenciosos [6], [11], [13]. Por isso, esta técnica é aconselhável somente quando a deleção das amostras ausentes não representar um problema considerável pela quantidade de registros e o mecanismo a ser tratado for comprovadamente MCAR [6].

As duas principais técnicas de deleção são a exclusão da amostra e a eliminação da variável com valor ausente. A deleção da amostra simplesmente deleta ou ignora todas as variáveis da amostra que contém algum valor ausente. Esta técnica pode ocasionar um sério viés, especialmente quando há um grande número de valores ausentes e se o conjunto de dados original é muito pequeno. Por outro lado, a deleção da variável com valor ausente, apesar de também envolver deleção ou omissão de valores, se trata de uma técnica mais seletiva e menos radical quando comparada a técnica anterior. Neste caso, a amostra não é totalmente deletada e somente a variável com valor ausente é descartada do processamento, fazendo com que todas as demais possam ser devidamente

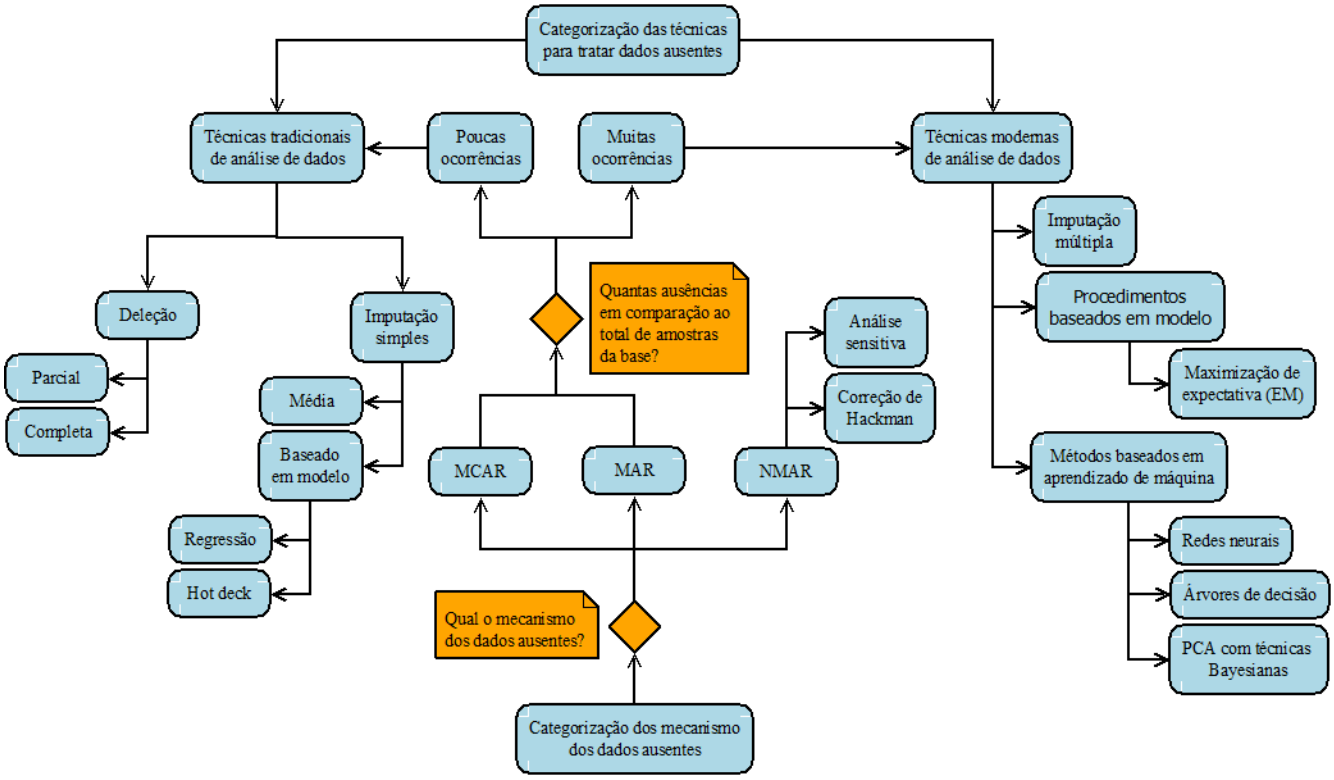

Figura 1. Diagrama de seleção para auxiliar na seleção de técnicas. Adaptado de [11].

computadas. Assim, esta técnica tenta minimizar a perda de dados e é eficaz quando o tamanho total da amostra é pequeno ou quando o número de amostras com dados ausentes é grande [6], [11], [13].

B. Técnicas de Inserção

Na literatura relacionada a dados ausentes é comum a utilização do termo imputação relacionado somente a métodos de estimativa e previsão de variáveis com valores ausentes. Contudo, para o presente trabalho, este termo também é considerado válido para os métodos mais simples, como substituição por zero ou pela média, por exemplo, uma vez que estes métodos também recorrem a inserção de algum valor em variáveis com valores ausentes. Amiri e Jensen [3] ressaltam que os métodos de inserção dos valores ausentes podem ser categorizados com base na técnica usada para aproximar os valores ausentes.

A substituição pela média, uma das formas simples de inserção, faz com que seja utilizado o valor médio de uma variável no lugar dos dados ausentes desta variável [6], [13]. Para Kang [6], embora teoricamente a média seja uma estimativa razoável para preencher o valor de uma variável com valor ausente por motivos aleatórios, na presença de uma grande quantidade de dados ausentes ou na falta de aleatoriedade destas ausências, a utilização da média pode acarretar resultados inconsistentes. Para Gelman e Hill [13], por outro lado, este método pode distorcer severamente a distribuição dos dados para esta variável, levando a complicações com medidas incluindo, notadamente, subestimações do desvio padrão e distorcendo relações entre as variáveis, tendendo para zero as estimativas da correlação. Osman *et. al* [11] ressalta que, embora a substituição pela média seja uma técnica relativamente fácil de ser utilizada, sendo esta sua vantagem, podem ocorrer resultados distorcidos devido a irregularidades na distribuição da amostra.

Outra técnica de inserção é a substituição do valor de uma variável ausente pelo último valor existente para a mesma variável [6], [13]. Apesar de se tratar de uma técnica simples, como esta técnica assume que os valores para uma variável não se alteraram para então poder substituir pelo último valor, a técnica pode produzir uma estimativa tendenciosa, fazendo com que seja aconselhável não a utilizar como abordagem primária [6].

A inserção *Hot-Deck* (HD) é uma técnica estatística popular e uma das técnicas mais amplamente utilizadas em avaliações de pesquisa. Esta técnica envolve encontrar um conjunto de dados semelhante ou muito próximo. A média deste conjunto é então calculada e são determinados os limites superior e inferior. A vantagem dessa técnica é sua simplicidade, contudo ela pode ser ineficiente em termos de computacionais. Outro problema é que a estimativa dos dados ausentes é baseada em um único conjunto de dados, resultando em subestimações do erro padrão e da variabilidade [11].

Na inserção por regressão as variáveis existentes são utilizadas para que se possa realizar uma previsão que será utilizada

para substituir o valor ausente, como se o valor previsto fosse o valor real da variável [6]. De uma forma geral, trata-se de uma ferramenta estatística que estima o relacionamento entre a entrada e saída ou entre um ponto dos dados e sua variável associada [11]. Primeiramente, é necessário ajustar um modelo de regressão definindo a variável de interesse como variável de resposta e outra variável relevante como covariável. Os coeficientes são estimados e, em seguida, os valores ausentes podem ser previstos pelo modelo ajustado [14]. Kang [6] ressalta que, embora esta técnica possua algumas vantagens sobre as técnicas de deleção descritas anteriormente por evitar alterar a distribuição dos dados e o desvio padrão, assim como a técnica de substituição pela média, nenhuma informação nova é adicionada. Por outro lado, Zhang [14] ressalta que a inserção por regressão em uma ou mais variáveis pode produzir valores mais inteligentes.

Métodos de inserção por meio da utilização de vizinhos próximos predizem valores para as variáveis com valores ausentes com base em instâncias completas próximas à variável problemática. Embora possuam boa precisão, sejam intuitivos e relativamente simples de serem utilizados, eles necessitam que o usuário insira a quantidade de vizinhos e comparam todas as instâncias para encontrar o número de vizinhos. Isso resulta em alta complexidade de tempo e problema de otimização local [3].

Há também técnicas que utilizam a máxima verossimilhança para tratar os dados ausentes. A estimativa de máxima verossimilhança identifica os valores dos parâmetros da população com a maior probabilidade de produzir os dados da amostra [9]. Após a estimativa dos parâmetros utilizando os dados disponíveis, os dados ausentes são estimados com base nos parâmetros que acabaram de ser estimados. De uma forma geral, esta técnica estima os valores das variáveis ausentes usando a distribuição condicional das outras variáveis [6]. O algoritmo EM (*Expectation Maximization*), por exemplo, pode ser utilizado para criar um novo conjunto de dados, no qual todos os valores omissos são inseridos com valores estimados pelos métodos de máxima verossimilhança [15].

Regularized Expectation Maximization (RegEM)[2], proposto por Schneider [12], teve como ponto de partida o algoritmo EM (*Expectation Maximization*) [15]. A partir dos dados incompletos, o algoritmo RegEM calcula as estimativas de máxima verossimilhança dos parâmetros de qualquer distribuição probabilística. Para dados Gaussianos, cuja distribuição probabilística pode ser parametrizada pela média e pela matriz de covariância, o EM inicia com estas duas e depois percorre as etapas alternadas de atribuição de valores ausentes e reestimativa da média e da matriz de covariância a partir conjunto de dados completo e de uma estimativa da matriz de covariância do erro de inserção do dado ausente. Na etapa de inserção, os valores ausentes são inseridos por meio da expectativa condicional obtida através dos valores disponíveis e a matriz de covariância do erro dos valores inseridos é estimada. Na etapa de estimação, a média e a matriz de covariância são reestimadas, considerando o erro de inserção condicional para a matriz de covariância. As etapas de inserção e estimativa são repetidas até que os valores inseridos, a média estimada e a matriz de covariância parem de se alterar [12].

Outra técnica interessante para tratar dados ausentes é a imputação múltipla, considerada uma alternativa à estimativa de máxima verossimilhança e que tem sido amplamente utilizada [6], [9]. Nesta técnica, em vez de substituir um único valor por outro, os valores ausentes são substituídos por um conjunto de valores plausíveis que contêm a variabilidade natural e a incerteza dos valores corretos [6]. Esta técnica começa com uma previsão dos dados ausentes utilizando os dados existentes das demais variáveis. Os valores ausentes são, então, substituídos pelos valores previstos e um conjunto de dados completo é criado e denominado conjunto de dados imputado. Como este processo é repetido dentro das iterações, são criados vários conjuntos de dados imputados, justificando assim o nome da técnica. Cada conjunto de dados produzido é então analisado utilizando os procedimentos de análise estatística padrão para dados completos, fornecendo vários resultados analíticos. Posteriormente, combinando estes resultados analíticos, é produzido um único resultado geral analítico. Além de restaurar a variabilidade natural dos valores ausentes, esta técnica incorpora a incerteza devido aos dados ausentes, o que resulta em uma inferência estatística válida [6]. Diferentemente da técnica de máxima verossimilhança, a imputação múltipla separa o tratamento de dados ausentes da análise estatística e, como a imputação normalmente emprega um modelo muito geral, um único conjunto de imputações geralmente pode suportar uma variedade de análises estatísticas [9]. A maior desvantagem desta técnica é sua complexidade que não envolve apenas a execução das análises, mas também a combinação dos resultados e o uso correto dos dados. Por outro lado, a imputação múltipla introduz a variabilidade para encontrar uma gama de respostas possíveis [11].

Árvore de decisão é um modelo preditivo para mapear dados e observações que são representados nos ramos da árvore. O objetivo é chegar a conclusões sobre o valor alvo, representado nas folhas das árvores. Esta técnica é um dos métodos de aprendizado supervisionado mais amplamente utilizados. A principal vantagem desta técnica é a interpretabilidade que possibilita que os dados sejam visualizados e que sua estrutura seja facilmente compreendida. Normalmente, o objetivo é encontrar a árvore de decisão ideal, minimizando o erro padrão [11].

A teoria dos conjuntos *Fuzzy-Rough* fornece uma excelente estrutura para lidar com a incerteza, possuindo algumas características desejáveis que a tornam uma boa escolha a ser utilizada como método de imputação. Os métodos *Fuzzy-Rough* não são essencialmente problemas de otimização, sendo assim, eles não iteram através de etapas de algoritmo. Isto é importante porque eles não precisam de um critério de parada. Eles também não dependem de parâmetros especificados pelo usuário. Outro motivo para utilizar técnicas *Fuzzy-Rough* é sua simplicidade e compreensibilidade. Eles simplesmente

[2]https://github.com/tapios/RegEM

calculam as semelhanças *fuzzy* das instâncias e tomam as decisões com base nelas. Eles podem trabalhar com facilidade e eficácia na presença de ruído, podendo também lidar com valores ausentes. Eles não precisam de estimativas iniciais para valores ausentes. Além disso, eles podem lidar facilmente com dados imprecisos [3].

Outra opção para imputação de valores ausentes são métodos baseados em Redes Neurais. Estes métodos geralmente definem um erro e tentam minimizá-lo iterativamente. Uma grande desvantagem destes métodos é sua complexidade de tempo. Como são métodos iterativos, eles também precisam de um limite predefinido para parar. Os métodos baseados em agrupamento são alternativas. Alguns desses métodos precisam de um número de grupos especificado pelo usuário para ser iniciado. Além disso, podem convergir para um mínimo local e para convergir para um mínimo global, várias repetições devem ser realizadas, o que é muito computacionalmente caro [3].

K-Nearest Neighbors Imput (KNNI) introduzido inicialmente por Troyanskaya *et. al* [16] em seu trabalho envolvendo dados ausentes em *microarrays* de DNA. O método baseado no KNN seleciona amostras com perfis de expressão semelhantes ao da amostra analisada para inserir os valores ausentes. Considerando, por exemplo, que a amostra A tem um valor ausente no experimento 1, este método encontrará K outras amostras, com o valor disponível no experimento 1 e com expressão mais próxima à de A nos $2 - N$ experimentos, onde N é o número total de experimentos. Então, uma média ponderada de valores no experimento 1 dos K mais próximos é utilizada como estimativa para o valor ausente na amostra. Com a utilização da média ponderada, a contribuição de cada amostra é ponderada pela similaridade de sua expressão com a do gene A [16]. Assim como o KNNI, o *Singular Value Decomposition Imput* (SVDI) foi introduzido por Troyanskaya *et. al* [16]. Este método utiliza a decomposição de valor singular para obter um conjunto de padrões de expressão mutuamente ortogonais que podem ser linearmente combinados para aproximar a expressão de todas as amostras no conjunto de dados.

Além dos citados, diversos outros métodos foram propostos para lidar com dados ausentes, como por exemplo: *Bayesian Principal Component Analysis* [17]; *Local Least Squares Imputation* [18]; *K-Means Imputation* [19]; *Fuzzy K-Means Imputation*[19]; *Global Most Comum Attribute* [20]; *Global Most Comum Average* [20]; *SWFKM* [21].

III. Materiais e Métodos

Esta seção detalha os conjuntos de dados e os modelos utilizados nos experimentos. Os métodos de avaliação dos resultados também são descritos nessa seção.

A. Bases de Dados Originais

Inicialmente, foram selecionadas duas bases de dados completas associadas a problemas de regressão (previsão). As bases de dados foram obtidas no Repositório da UCI [22] e são descritas a seguir:

- **DataSet 1**[3]: base de dados relacionada a testes aerodinâmicos e acústicos realizados pela NASA (*National Aeronautics and Space Administration*) compreendendo aerofólios de tamanhos diferentes NACA 0012 em várias velocidades de túnel de vento e ângulos de ataque. A base possui 1503 registros com 5 variáveis de entrada, sendo elas: frequência, ângulo, comprimento da corda, velocidade do fluxo livre, espessura do deslocamento lateral da sucção. O objetivo é prever o nível de pressão sonora em escala. Este conjunto de dados foi utilizado em [23], [24], [25].
- **DataSet 2**[4]: base de dados contendo um conjunto de dados de resistência à compressão do concreto. A base possui 1030 registros com 8 variáveis de entrada, sendo elas: a quantidade de cimento, a quantidade de escória de alto forno, a quantidade de cinzas volantes, a quantidade de água, a quantidade de superplastificante, a quantidade de agregado grosso, a quantidade de agregado fino e o número de dias. O objetivo é realizar a previsão da resistência à compressão do concreto. Este conjunto de dados foi utilizado em [26].

Destaca-se que para os experimentos os dados foram normalizados com valores entre 0 e 1.

B. Bases de Dados com Dados Ausentes

A partir das bases de dados completas descritas na Seção III-A foram criadas bases com dados ausentes para os cenários de MCAR e MAR.

1) Bases de Dados MCAR: Para criação da base MCAR de dados ausentes foi utilizado o Algoritmo 1. Este algoritmo recebe como entrada a base de dados completa e a taxa de ausência que a nova base deve ter. Com isso ele calcula o número de amostras que deve possuir dados ausentes a partir do número de amostras da base original e da taxa de ausência. Então, o algoritmo sorteia as amostras que possuirão dados ausentes. Por fim, ele faz uma iteração em cada amostra sorteada para conter um dado ausente e sorteia a variável que ficará com o valor ausente.

Para cada base de dados descrita na Seção III-A foram criadas 6 novas bases de dados com dados ausentes com taxas de ausência de 1%, 5%, 10%, 15%, 20% e 30%. A partir de cada base de dados com dados ausentes foram criadas três novas bases, uma substituindo os valores ausentes por zero, outra pela média dos últimos 5 valores e a outra pelo último valor disponível para imputar os dados ausentes. Portanto, foram criadas 18 bases de dados para os experimentos com MCAR.

2) Bases de Dados MAR: O Algoritmo 2 foi utilizado para gerar as bases de dados ausentes MAR. Este algoritmo recebe a base de dados original, a taxa de ausência para a variável com maior propensão e a taxa de ausência para as demais variáveis. O algoritmo seleciona uma variável da base para ser a variável com maior propensão a possuir valor ausente e

[3]http://archive.ics.uci.edu/ml/datasets/airfoil+self-noise

[4]https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength

Data: baseOriginal, taxaAusencia
Result: Base de dados MCAR
Calcula número de amostras e variáveis de entrada da base original
Calcula a quantidade de amostras a receber dado ausente
Sorteia as amostras que receberão dado ausente
Para *cada amostra sorteada*
  Sorteia uma variável
  Variável sorteada recebe dado ausente
Fim

Algorithm 1: Gerador de bases MCAR

calcula a probabilidade das demais variáveis estarem ausentes com base no número de variáveis e na taxa mínima de ausência estabelecida. Depois, são iniciados dois vetores para controle dos sorteios, um zerado e um com o valor de cada índice. Com isso, o algoritmo faz uma iteração em cada variável para que cada uma ocupe uma quantidade de posições no vetor de sorteio relativa à sua porcentagem probabilística de ausência. Por fim, ao iterar em cada amostra da base de dados original, o algoritmo seleciona aleatoriamente um índice do vetor de sorteio e, caso seja sorteado o índice de uma variável, a variável fica com o valor ausente.

O Algoritmo 2 foi utilizado com taxas de ausência de 5x1, 10x1, 10x5, 20x5, 20x10, 30x5 e 30x10 para criação de 7 bases de dados com dados ausentes MAR, onde o primeiro dígito se refere a taxa de ausência máxima e o segundo se refere a taxa de ausência mínima. Após este procedimento, a partir de cada base de dados com dados ausente geradas foram geradas três novas bases, uma substituindo os valores ausentes por zero, outra pela média dos últimos 5 valores e a outra pelo último valor disponível para imputar os dados ausentes. Portanto, foram criadas 21 bases de dados para os experimentos com MAR.

C. Algoritmos de Validação e Medida de Erro

Os experimentos computacionais foram realizados com três sistemas *fuzzy* evolutivos: eTS [27], eMG [28] e ALMMo [29]. Os sistemas *fuzzy* evolutivos consistem em modelos inteligentes com aprendizado incremental e contínuo. Os modelos iniciam-se com uma amostra de dados de entrada, estimam a saída, ajustam seus parâmetros e evoluem sua estrutura a cada nova amostra apresentada [28]. Nos experimentos computacionais realizados foi simulado o processamento *online*, isto é, durante a estimação dos parâmetros e avaliação do desempenho os modelos evoluem sua estrutura para todas as amostras do conjunto de dados. Portanto, os resultados apresentados se referem ao desempenho dos modelos na previsão de todas as amostras do conjunto de dados.

Os parâmetros dos modelos evolutivos foram definidos como se segue: para o eTS foram utilizados os valores de 750 para o Omega, modelos não fixos e objetivo de otimização global; para o eMG foram utilizados os valores de $0,01$ para o α básico, $0,05$ para o λ, 40 para o tamanho da janela; para o ALMMo foram utilizados *OnlineLearning* para o *status* e *Regression* para o *goal*.

Data: baseOriginal, taxaAusenciaMax, taxaAusenciaMin
Result: Base de dados MCAR
Calcula número de amostras e variáveis de entrada da base original
Cria uma lista de sorteio de tamanho 100
Sorteia uma variável para ter taxa de ausência máxima
Calcula a taxa de ausência das demais variáveis
Para *cada variável da base original*
  if *a variável for a sorteada para taxa máxima* **then**
    Sorteia posições da lista de sorteio usando a taxaAusenciaMax para definir a quantidade
  else
    Sorteia posições da lista de sorteio usando a taxa de ausência calculada para definir a quantidade
  end
  Altera os valores das posições selecionadas da lista de sorteio com o valor da variável
Fim
Para *cada amostra da base original*
  Sorteia aleatoriamente uma posição da lista de sorteio
  if *a posição sorteada não for igual a zero* **then**
    Obtém a variável pelo valor sorteado
    Variável sorteada recebe dado ausente
  end
Fim

Algorithm 2: Gerador de bases MAR

Para avaliar o desempenho dos algoritmos e consequentemente das técnicas de imputação foi utilizada como medida de erro o RMSE (*Root Mean Square Error*), descrito na Eq. 4, calculado utilizando o quadrado da diferença entre a saída obtida e a real.

$$RMSE = \frac{1}{H}\sum_{h=1}^{H}\sqrt{(\hat{y}^{[h]} - y^{[h]})^2} \tag{4}$$

IV. Experimentos, Resultados e Discussões

Esta seção detalha os experimentos realizados e apresenta os resultados obtidos. Inicialmente são apresentados os experimentos realizados empregando as 18 bases de dados MCAR seguido dos experimentos com as 21 bases MAR. As primeiras bases a serem executadas pelos algoritmos definidos na Seção III foram as originais para se obter os valores que serão a base para toda a análise dos demais resultados, seguida das bases MCAR e, por fim, as bases MAR.

A. Experimentos MCAR

A Tabela I ilustra os resultados obtidos para as bases MCAR geradas a partir da base de dados 1 com taxas de ausência de 1%, 5%, 10%, 15%, 20% e 30% e para a base original. Conforme pode ser verificado, para os cenários com 1% e 5% de ausência, a substituição por zero obteve resultados melhores do que os da base original, contudo, o resultado foi piorando à medida que a taxa de ausência aumentava e, em comparação com a base original, sua proximidade máxima foi de 1 casa decimal. Para as bases que utilizaram a média como

solução foram obtidos resultados com proximidade de até três casas decimais no cenário com 1% de ausência e de duas casas decimais no cenário com 20%. Por fim, para as bases que utilizaram a substituição pelo último valor foi obtido o mesmo resultado da base original em dois dos três algoritmos utilizados nos testes, para o cenário com 1% de ausência, uma diferença de apenas 0,0001 no terceiro algoritmo, para os cenários com 1% e 20% de ausências, e um resultado melhor que o da base original com o primeiro algoritmo, para o cenário com 20% de ausência. Diferente da substituição por zero na qual os resultados pioraram gradativamente conforme o aumento na taxa de ausência, nos experimentos com bases obtidas utilizando a média e o último valor os resultados oscilaram e tiverem na maioria dos casos resultados similares.

Tabela I
RESULTADO PARA O CENÁRIO MCAR NAS BASES GERADAS DA BASE ORIGINAL 1

Bases	**eTS**			**eMG**			**ALMMo**		
Original	0,1272			0,1294			0,1264		
Ausência	Zero	Média	Último	Zero	Média	Último	Zero	Média	Último
1%	0,0561	0,1273	0,1272	0,0560	0,1293	0,1294	0,0644	0,1265	0,1265
5%	0,1068	0,1316	0,1317	0,1098	0,1342	0,1343	0,1029	0,1300	0,1308
10%	0,1508	0,1289	0,1309	0,1366	0,1307	0,1315	0,1377	0,1281	0,1287
15%	0,1477	0,1315	0,1313	0,1456	0,1330	0,1314	0,1449	0,1278	0,1270
20%	0,1916	0,1247	0,1223	0,1660	0,1434	0,1427	0,1613	0,1310	0,1265
30%	0,2221	0,1311	0,1300	0,2125	0,1573	0,1333	0,2008	0,1305	0,1281

A Tabela II mostra os resultados obtidos para as bases MCAR geradas a partir da base de dados 2. Neste cenário, a substituição por zero obteve apenas dois resultados melhores que o da base original no cenário com 1% de ausência nos três algoritmos testados e todos os demais com proximidade de apenas uma casa decimal. Para a substituição utilizando a média dos últimos cinco valores os resultados similares aos obtidos com a base original, mantendo uma proximidade de até três casas decimais. Para a substituição pelo último valor, foram obtidos resultados próximos aos obtidos pela substituição pela média.

Tabela II
RESULTADO PARA O CENÁRIO MCAR NAS BASES GERADAS DA BASE ORIGINAL 2

Bases	**eTS**			**eMG**			**ALMMo**		
Original	0,1333			0,1301			0,1255		
Ausência	Zero	Média	Último	Zero	Média	Último	Zero	Média	Último
1%	0,1288	0,1319	0,1311	0,1485	0,1300	0,1300	0,1250	0,1254	0,1253
5%	0,2448	0,1303	0,1350	0,1730	0,1316	0,1348	0,1323	0,1292	0,1301
10%	0,1810	0,1310	0,1319	0,1895	0,1332	0,1303	0,1417	0,1256	0,1285
15%	0,1564	0,1298	0,1304	0,1838	0,1510	0,1342	0,1468	0,1265	0,1309
20%	0,1590	0,1428	0,1327	0,2015	0,1357	0,1675	0,1484	0,1323	0,1272
30%	0,2498	0,1352	0,1522	0,2753	0,1330	0,1395	0,1693	0,1309	0,1351

B. Experimentos MAR

A Tabela III ilustra os resultados obtidos para as bases MAR geradas a partir da base de dados 1 com taxas de ausência de 5x1, 10x1, 10x5, 20x5, 20x10, 30x5 e 30x10. Neste cenário, a substituição por zero continuou obtendo resultados com proximidade de apenas uma casa decimal dos resultados da base original. Ressalta-se que seu pior caso no cenário de ausência 30x5 foi muito inferior a todos os outros obtidos até então. Para a utilização da média, foram obtidos resultados melhores e resultados com proximidade de três casas decimais em pelo menos dois dos algoritmos testados. Para a utilização do último valor, foram obtidos alguns resultados ligeiramente inferiores que os obtidos pela utilização da média, outros também com três casas decimais de proximidade dos resultados da base original e um resultado equivalente no cenário de 10x1 com o algoritmo ALMMo.

Tabela III
RESULTADO PARA O CENÁRIO MCAR NAS BASES GERADAS DA BASE ORIGINAL 1

Bases	**eTS**			**eMG**			**ALMMo**		
Original	0,1272			0,1294			0,1264		
Ausência	Zero	Média	Último	Zero	Média	Último	Zero	Média	Último
5x1	0,1323	0,1275	0,1273	0,1529	0,1292	0,1296	0,1284	0,1276	0,1265
10x1	0,1337	0,1275	0,1315	0,1358	0,1449	0,1450	0,1285	0,1274	0,1264
10x5	0,1164	0,1194	0,1280	0,0919	0,1258	0,1276	0,0944	0,1278	0,1275
20x5	0,1016	0,1281	0,1279	0,1096	0,1277	0,1301	0,0926	0,1290	0,1266
20x10	0,1534	0,1309	0,1362	0,1498	0,1292	0,1290	0,1465	0,1278	0,1277
30x5	0,4800	0,1417	0,1520	0,4467	0,1390	0,1546	0,4092	0,1356	0,1450
30x10	0,1483	0,1441	0,1375	0,1256	0,1419	0,1335	0,1271	0,1392	0,1329

A Tabela IV mostra os resultados obtidos para as bases MAR geradas a partir da base de dados 2. Para os resultados obtidos pela substituição por zero, todos os valores obtidos foram inferiores que os obtidos pela base original e a proximidade máxima foi de uma casa decimal. Para a utilização da média apesar de conseguir uma proximidade de duas a três casas, a maioria dos resultados foi inferior ao da base original. Para a utilização do último valor todos os resultados obtidos foram inferiores que os resultados da base original, tendo a maior proximidade com duas casas decimais.

Tabela IV
RESULTADO PARA O CENÁRIO MCAR NAS BASES GERADAS DA BASE ORIGINAL 2

Bases	**eTS**			**eMG**			**ALMMo**		
Original	0,1333			0,1301			0,1255		
Ausência	Zero	Média	Último	Zero	Média	Último	Zero	Média	Último
5x1	0,1456	0,1443	0,1424	0,1585	0,1323	0,1341	0,1295	0,1278	0,1297
10x1	0,1429	0,1327	0,1411	0,1376	0,1446	0,1584	0,1257	0,1270	0,1301
10x5	0,1803	0,1576	0,1394	0,2127	0,1346	0,1329	0,1507	0,1320	0,1293
20x5	0,1673	0,1399	0,1510	0,1714	0,1300	0,1326	0,1403	0,1265	0,1266
20x10	0,2150	0,2201	0,1593	0,1783	0,1293	0,1344	0,1493	0,1338	0,1390
30x5	0,2059	0,1281	0,1446	0,2227	0,1739	0,1405	0,1576	0,1334	0,1373
30x10	0,4461	0,1796	0,1787	0,2828	0,1639	0,1770	0,2406	0,1245	0,1355

V. CONCLUSÃO

Este artigo discorre com a revisão da literatura sobre dados ausentes e apresenta experimentos computacionais utilizando sistemas *fuzzy* evolutivo e três técnicas para tratar dados ausentes aplicadas em problemas de previsão. De acordo com a revisão realizada é possível verificar a complexidade do tema abordado. Cada mecanismo de dados ausente merece uma atenção e um tratamento especial para que seu impacto nos resultados sejam os menores possíveis. Além disso, é importante lembrar que não há uma solução única que resolva totalmente o problema ou que seja aconselhável em todos os casos, o que ressalta a importância da escolha das técnicas para

tratar cada mecanismos da melhor forma o problema, seja por meio de uma ou mais técnicas em conjunto.

Os experimentos computacionais foram realizados utilizando três algoritmos *fuzzy* evolutivo (eTS, eMG e ALMMo), três técnicas de imputação de dados ausentes (substituição por zero, substituição pela média e substituição pelo último valor) e duas bases de dados para problemas de previsão. Os resultados dos experimentos mostraram que as técnicas de substituição pela média e pelo último valor conseguem resultados similares aos obtidos com a base completa na maioria dos cenários analisados. Por outro lado, os resultados obtidos com a técnica de substituição por zero tentem a piorar com o aumento da taxa de ausência dos dados.

Como trabalhos futuros estão tanto o estudo de outras possíveis formas de geração de bases com dados ausentes, quanto a análise de outras técnicas e algoritmos que solucionem este problema, com o objetivo de se gerar um modelo próprio.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores também são gratos ao CEFET-MG pelo apoio.

REFERÊNCIAS

[1] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 863–866.

[2] A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, no. 5, pp. 692–709, 2007.

[3] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," *Neurocomputing*, vol. 205, pp. 152–164, 2016.

[4] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, vol. 7, pp. 11 651–11 667, 2019.

[5] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art." *Psychological methods*, vol. 7, no. 2, p. 147, 2002.

[6] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, p. 402, 2013.

[7] T. D. Little, K. M. Lang, W. Wu, and M. Rhemtulla, "Missing data," *Developmental psychopathology*, pp. 1–37, 2016.

[8] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen, "Missing data and multiple imputation in clinical epidemiological research," *Clinical epidemiology*, vol. 9, p. 157, 2017.

[9] C. K. Enders and A. N. Baraldi, "Missing data handling methods," *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, pp. 139–185, 2018.

[10] R. W. Krause, M. Huisman, C. Steglich, and T. A. Sniiders, "Missing network data a comparison of different imputation methods," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 159–163.

[11] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, "A survey on data imputation techniques: water distribution system as a use case," *IEEE Access*, vol. 6, pp. 63 279–63 291, 2018.

[12] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of climate*, vol. 14, no. 5, pp. 853–871, 2001.

[13] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[14] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[16] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[17] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[18] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for dna microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, 2004.

[19] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: a study of fuzzy k-means clustering method," in *International Conference on Rough Sets and Current Trends in Computing*. Springer, 2004, pp. 573–579.

[20] J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse, and X. Zheng, "Handling missing attribute values in preterm birth data sets," in *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*. Springer, 2005, pp. 342–351.

[21] Z. Liao, X. Lu, T. Yang, and H. Wang, "Missing data imputation: a fuzzy k-means clustering algorithm over sliding window," in *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 3. IEEE, 2009, pp. 133–137.

[22] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[23] T. F. Brooks, D. S. Pope, and M. A. Marcolini, "Airfoil self-noise and prediction," 1989.

[24] K. Lau, R. López, E. Oñate, E. Ortega, R. Flores, M. Mier-Torrecilla, S. Idelsohn, C. Sacco, and E. González, "A neural networks approach for aerofoil noise prediction," *Master thesis*, 2006.

[25] R. Lopez, E. Balsa-Canto, and E. Oñate, "Neural networks for variational problems in engineering," *International Journal for Numerical Methods in Engineering*, vol. 75, no. 11, pp. 1341–1360, 2008.

[26] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete research*, vol. 28, no. 12, pp. 1797–1808, 1998.

[27] P. P. Angelov and D. P. Filev, "An approach to online identification of takagi-sugeno fuzzy models," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 484–498, 2004.

[28] A. Lemos, W. Caminhas, and F. Gomide, "Multivariable gaussian evolving fuzzy modeling system," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 1, pp. 91–104, 2010.

[29] P. P. Angelov, X. Gu, and J. C. Príncipe, "Autonomous learning multimodel systems from data streams," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 4, pp. 2213–2224, 2017.