CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ESTUDO DA CARDINALIDADE E DA RELAÇÃO ENTRE ATRIBUTOS NA IMPUTAÇÃO DE DADOS CATEGÓRICOS

Felipe de Ávila Tavares

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

Orientadores:
Jorge de Abreu Soares

Rio de Janeiro,

1 de outubro de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

T231 Tavares, Felipe de Ávila
Estudo da cardinalidade e da relação entre atributos na imputação de dados categóricos / Felipe de Ávila Tavares. — 2026.
88f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 86-88
Orientador: Jorge de Abreu Soares

1. Aprendizado de máquina. 2. Mineração de dados. 3. Algoritmos computacionais . 4. Redes neurais (Computação). I. Soares, Jorge de Abreu. (Orient.). II. Título.

CDD 006.3

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

ESTUDO DA CARDINALIDADE E DA RELAÇÃO ENTRE ATRIBUTOS NA IMPUTAÇÃO DE DADOS CATEGÓRICOS

Defesa de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Felipe de Ávila Tavares

Aprovada por:

Jorge de Abreu Soares, D.Sc. (orientador)

Diego Nunes Brandão, D.Sc.

Ronaldo Ribeiro Goldschmidt, D.Sc.

Rio de Janeiro,

1 de outubro de 2026

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me conceder força, sabedoria e perseverança ao longo desta caminhada, especialmente nos momentos de maior dificuldade.

À minha família, pelo apoio, incentivo e compreensão durante toda esta trajetória. Agradeço por estarem ao meu lado, acreditando em mim e me dando o suporte necessário para que eu pudesse chegar até aqui.

Ao meu orientador, Jorge de Abreu Soares, pela orientação, pelos ensinamentos e pelas contribuições ao longo do desenvolvimento deste trabalho. Agradeço também pela confiança e pela oportunidade de realizar esta pesquisa sob sua orientação.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro concedido durante parte do período de desenvolvimento desta pesquisa, contribuindo para a sua realização.

RESUMO

ESTUDO DA CARDINALIDADE E DA RELAÇÃO ENTRE ATRIBUTOS NA IMPUTAÇÃO DE DADOS CATEGÓRICOS

Felipe de Ávila Tavares

Orientadores:
Jorge de Abreu Soares

Resumo da defesa submetida 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 grau de mestre.

Esta dissertação investiga o problema da imputação de dados categóricos, com foco na análise comparativa de diferentes métodos de imputação e na influência das características estruturais das bases de dados sobre o desempenho dessas abordagens. O estudo considera três cenários distintos: bases simples, compostas predominantemente por atributos numéricos; bases mistas, contendo atributos numéricos e categóricos; e bases totalmente categóricas. A metodologia proposta envolve a geração artificial de valores ausentes segundo o mecanismo *Missing Completely at Random* (MCAR), em taxas de 10%, 20% e 30%, utilizando múltiplas sementes aleatórias para avaliar a estabilidade dos métodos analisados. Foram investigadas abordagens baseadas em KNN, Multi-Layer Perceptron (MLP) e Random Forest, considerando tanto a acurácia de imputação quanto o desempenho de classificadores treinados sobre os dados imputados. Além disso, o coeficiente V de Cramer foi utilizado para analisar a influência da correlação entre atributos categóricos no processo de imputação. Os resultados obtidos indicam que métodos baseados em modelos globais, especialmente Random Forest, apresentam maior robustez e estabilidade frente ao aumento da taxa de ausência e à complexidade estrutural das bases, enquanto o KNN demonstra maior sensibilidade à heterogeneidade e à redução da informação disponível. Os achados reforçam a importância de considerar características como cardinalidade e correlação entre atributos na escolha do método de imputação mais adequado.

Palavras-chave:
Imputação Categórica, Data-Centric Artificial Intelligence (DCAI), Pré-processamento de dados, Aprendizado de Máquina, Ciência de Dados.

Rio de Janeiro,

1 de outubro de 2026

ABSTRACT

ESTUDO DA CARDINALIDADE E DA RELAÇÃO ENTRE ATRIBUTOS NA IMPUTAÇÃO DE DADOS CATEGÓRICOS

Felipe de Ávila Tavares

Advisors:
Jorge de Abreu Soares

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

This dissertation investigates the problem of categorical data imputation, focusing on the comparative analysis of different imputation methods and on how dataset structural characteristics influence their performance. The study considers three distinct scenarios: simple datasets composed predominantly of numerical attributes, mixed datasets containing both numerical and categorical variables, and fully categorical datasets. The proposed methodology artificially generates missing values under the *Missing Completely at Random* (MCAR) mechanism at rates of 10%, 20%, and 30%, using multiple random seeds to evaluate the stability of the analyzed methods. Approaches based on KNN, Multi-Layer Perceptron (MLP), and Random Forest were investigated, considering both imputation accuracy and the performance of classifiers trained on the imputed data. In addition, Cramer's V coefficient was used to analyze the influence of correlation between categorical attributes in the imputation process. The obtained results indicate that global model-based methods, especially Random Forest, achieve greater robustness and stability against increasing missingness rates and dataset structural complexity, while KNN shows greater sensitivity to data heterogeneity and information loss. The findings reinforce the importance of considering characteristics such as cardinality and attribute correlation when selecting the most appropriate imputation method.

Key-words:
Categorical Imputation, Data-Centric Artificial Intelligence (DCAI), Data Preprocessing, Machine Learning, Data Science.

Rio de Janeiro,

1 de outubro de 2026

Sumário

I Introdução **2**

II Fundamentação Teórica **7**

II.1 Fundamentos de Imputação de Dados 7

II.2 Imputação de Dados Categóricos 8

II.3 Dados Faltantes: Morfologia da ausência 8

II.4 Causas de Dados Ausentes 10

II.5 Medidas Estatísticas e Métricas de Similaridade 10

II.5.1 Coeficiente Eta^2 10

II.5.2 V de Cramer 11

II.5.3 Distância de Gower 12

II.6 Métodos Utilizados nesta Pesquisa 13

II.6.1 K-Nearest Neighbors (KNN) 13

II.6.2 Random Forest (RF) 14

II.6.3 Multilayer Perceptron (MLP) 15

III Trabalhos relacionados **19**

III.1 Seleção dos Trabalhos 19

III.1.1 Critérios de Inclusão 20

III.1.2 Critérios de Exclusão 20

III.1.3 Processo de Seleção 20

III.2 Comparação dos Trabalhos Relacionados 21

III.2.1 Classificação dos Métodos de Imputação 28

III.2.2 Impactos nos Contextos de Aplicação 28

III.2.3 Tabela de Comparação dos Métodos de Imputação 29

IV Metodologia **33**

IV.1 Delineamento Experimental 33

IV.2 Notação Utilizada 35

IV.3 Conjuntos de Dados 35

IV.4 Preparação dos Dados 36
IV.5 Particionamento dos Dados 37
IV.6 Codificação e Normalização dos Atributos 37
IV.7 Geração Artificial de Valores Ausentes 38
IV.8 Métodos de Imputação 38
IV.8.1 Imputação com KNN 39
IV.8.2 Imputação com Random Forest 41
IV.8.3 Imputação com MLP 42
IV.9 Avaliação dos Métodos de Imputação 43
IV.9.1 Taxa de Acerto da Imputação 43
IV.9.2 Acurácia do Classificador 44
IV.9.3 Limitações da Métrica de Acurácia 44
IV.10 Repetições Experimentais e Agregação dos Resultados 45
IV.11 Síntese do Protocolo Experimental 45

V Resultados 47
V.1 Bases de Dados Utilizadas 47
V.2 Imputação via KNN 50
V.2.1 Análise dos Resultados do KNN 51
V.3 Imputação via MLP 61
V.4 Imputação via Random Forest 70
V.5 Discussão Comparativa dos Métodos de Imputação 79

VI Considerações Finais 84

Referências 86

Lista de Figuras

IV.1 Fluxo geral do procedimento experimental. 34

V.1 Desempenho das variações do KNN nas bases do cenário Simples. 52
V.2 Desempenho de classificadores treinados sobre dados imputados pelas variações do KNN no cenário Simples. 54
V.3 Desempenho das variações do KNN nas bases do cenário Misto. 56
V.4 Desempenho de classificadores treinados sobre dados imputados pelas variações do KNN no cenário Misto. 57
V.5 Desempenho das variações do KNN nas bases do cenário Categórico. 58
V.6 Desempenho de classificadores treinados sobre dados imputados pelas variações do KNN no cenário Categórico. 60
V.7 Desempenho do MLP nas bases do cenário Simples 63
V.8 Desempenho do MLP nas bases do cenário Misto 64
V.9 Desempenho do MLP nas bases do cenário Categórico 65
V.10 Desempenho dos classificadores MLP nas bases do cenário Simples 67
V.11 Desempenho dos classificadores MLP nas bases do cenário Simples 68
V.12 Desempenho dos classificadores MLP nas bases do cenário Simples 69
V.13 Desempenho do Random Forest nas bases do cenário Simples 72
V.14 Desempenho do Random Forest nas bases do cenário Misto 73
V.15 Desempenho do Random Forest nas bases do cenário Categórico 74
V.16 Desempenho dos classificadores RF nas bases do cenário Simples 76
V.17 Desempenho dos classificadores RF nas bases do cenário Misto 77
V.18 Desempenho dos classificadores RF nas bases do cenário Categórico 78
V.19 Comparação global da taxa de acerto da imputação entre Moda, KNN, MLP e Random Forest. 80
V.20 Comparação global do desempenho dos classificadores treinados sobre dados imputados pelos métodos avaliados. 81

Capítulo I Introdução

A presença de valores ausentes em bases de dados constitui um problema recorrente na análise estatística, na ciência de dados e em aplicações de aprendizado de máquina. A ausência de informações pode reduzir a quantidade de dados efetivamente disponível para análise, introduzir vieses e comprometer a validade de inferências e de modelos construídos a partir dos dados incompletos [Schafer, 1997; Little and Rubin, 2019]. Esse problema assume particular relevância em domínios como ciências sociais, saúde, finanças e inteligência artificial, nos quais a qualidade e a completude dos dados podem afetar diretamente a confiabilidade das análises e das decisões baseadas em dados.

Uma das estratégias utilizadas para lidar com esse problema é a *imputação de dados*, que consiste em estimar e atribuir valores aos elementos ausentes de uma base de dados, utilizando as informações disponíveis nas demais observações e variáveis. O objetivo não é simplesmente preencher lacunas, mas obter valores plausíveis que permitam preservar, tanto quanto possível, características relevantes dos dados e possibilitem a realização de análises subsequentes [Little and Rubin, 2019]. A escolha do método de imputação, entretanto, depende das características dos dados, do mecanismo responsável pela ausência e da natureza das variáveis envolvidas.

Desde os trabalhos pioneiros de Rubin [1976b], que estabeleceram as bases teóricas para a classificação dos mecanismos de ausência, *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Missing Not at Random* (MNAR), a imputação de dados consolidou-se como uma área relevante da estatística e da ciência de dados. Essa distinção é importante porque o mecanismo pelo qual os valores se tornam ausentes influencia as estratégias de tratamento possíveis e as condições sob as quais as análises realizadas sobre os dados incompletos podem ser consideradas válidas [Rubin, 1976b; Little and Rubin, 2019]. Nesta dissertação, o foco está no cenário MCAR, no qual a ocorrência da ausência é considerada completamente aleatória, permitindo concentrar a investigação nas características estruturais das bases de dados.

O presente trabalho concentra-se especificamente na imputação de variáveis categóricas. Variáveis categóricas representam atributos cujos valores pertencem a um conjunto discreto de categorias, como sexo, estado civil, profissão ou nível de escolaridade. Diferentemente das variáveis numéricas, essas categorias não necessariamente possuem uma relação de ordem ou uma distância quantitativa natural entre si. Consequentemente, métodos baseados diretamente em operações matemáticas ou métricas de distância desenvolvidas para dados contínuos não podem ser aplicados de maneira

direta a todos os problemas de imputação categórica. Mesmo quando as variáveis possuem uma ordenação natural, a distância entre categorias pode não ser representada adequadamente por uma escala numérica simples.

A natureza das variáveis categóricas introduz desafios adicionais ao processo de imputação. Uma imputação inadequada pode alterar a distribuição das categorias, modificar associações entre atributos e afetar análises multivariadas realizadas posteriormente [Schafer, 1997; Little and Rubin, 2019]. Assim, o problema não consiste apenas em recuperar corretamente cada valor ausente, mas também em produzir uma base imputada que permaneça coerente com as características relevantes da base original.

Nesse contexto, trabalhos relacionados têm explorado diferentes estratégias para a imputação de dados categóricos e mistos, utilizando abordagens estatísticas, métodos baseados em distância, técnicas de aprendizado de máquina e métodos de imputação múltipla. Estudos sobre dados categóricos e mistos mostram que a escolha do método pode depender da estrutura dos dados, da quantidade de variáveis envolvidas, do padrão de ausência e das relações existentes entre os atributos [Zimmermann et al., 2014; Memon et al., 2023; Schmitt and Rothlauf, 2024]. Também existem abordagens específicas para bases com grande quantidade de variáveis categóricas e diferentes padrões de ausência [Wilson and Lueck, 2014]. Esses trabalhos evidenciam que não existe, de maneira geral, um método de imputação que seja igualmente adequado para todas as configurações de dados.

Embora esses estudos tenham contribuído para o desenvolvimento e a avaliação de métodos de imputação, permanece uma questão relevante relacionada ao efeito das características da própria base de dados sobre o desempenho desses métodos. Em particular, ainda é necessário compreender de forma mais sistemática como características como o número de atributos, a cardinalidade das variáveis e as relações entre os atributos podem estar associadas à dificuldade de reconstrução dos valores ausentes. Essa perspectiva desloca parcialmente o foco da simples comparação entre algoritmos para a investigação das condições nas quais esses algoritmos apresentam melhor ou pior desempenho.

Neste trabalho, essas propriedades são denominadas *características estruturais dos dados*. Esse conceito compreende, principalmente, o número de atributos considerados, a cardinalidade das variáveis categóricas e as relações de associação existentes entre os atributos. A cardinalidade corresponde ao número de categorias distintas que uma variável pode assumir. Assim, uma variável com três categorias apresenta cardinalidade igual a três, enquanto uma variável com quinze categorias apresenta cardinalidade igual a quinze. Já as relações entre atributos representam o grau em que os valores de uma variável estão associados aos valores de outra. Neste estudo, a associações existentes entre os atributos é caracterizado por medidas adequadas ao tipo das variáveis, como o η^2 e o V de Cramer, conforme a natureza dos pares de atributos analisados.

A cardinalidade, entretanto, não deve ser interpretada isoladamente como sinônimo de dificuldade de imputação. Uma variável com maior número de categorias possui um espaço de possibilidades maior para a reconstrução de um valor ausente, mas a dificuldade efetiva também pode depender da distribuição dessas categorias e da quantidade de informação disponível nas demais variáveis. Uma categoria rara, por exemplo, pode apresentar comportamento diferente de uma categoria frequente, mesmo quando ambas pertencem a variáveis com a mesma cardinalidade. Da mesma forma, relações fortes entre atributos podem fornecer informações úteis para a reconstrução de valores ausentes. Portanto, investigar a cardinalidade isoladamente pode ser insuficiente para explicar o comportamento dos métodos de imputação, sendo necessário considerar conjuntamente diferentes características estruturais das bases.

Essa perspectiva também está relacionada à preservação das relações existentes nos dados. A imputação de valores ausentes pode ser considerada adequada não apenas quando recupera corretamente os valores removidos, mas também quando evita alterações indevidas na distribuição e nas associações entre os atributos. A preservação dessas relações é particularmente importante em análises multivariadas, nas quais o significado de uma variável pode depender de sua relação com outras variáveis da base [Little and Rubin, 2019]. Dessa forma, avaliar exclusivamente a proporção de valores corretamente imputados pode não ser suficiente para caracterizar a qualidade de um método.

Essa preocupação ganha relevância também no contexto da *Inteligência Artificial Centrada em Dados* (*Data-Centric Artificial Intelligence* – DCAI). Essa abordagem enfatiza a importância da qualidade, organização e adequação dos dados para o desenvolvimento de sistemas de inteligência artificial, deslocando parte da atenção tradicionalmente concentrada nos modelos para a melhoria sistemática dos próprios dados [Jakubik et al., 2024; Malerba and Pasquadibisceglie, 2024]. Sob essa perspectiva, métodos de imputação podem ser entendidos como parte das estratégias de preparação e melhoria dos dados, tornando relevante investigar não apenas seu desempenho na recuperação dos valores ausentes, mas também os efeitos produzidos sobre a estrutura e a utilidade da base resultante.

Apesar dos avanços existentes, permanecem limitações na literatura quanto à compreensão sistemática da influência das características estruturais sobre a imputação de dados categóricos. Muitos trabalhos concentram-se na proposição ou comparação de métodos, enquanto a influência conjunta do número de atributos, da cardinalidade e das relações entre variáveis recebe menor atenção [Zimmermann et al., 2014; Memon et al., 2023; Schmitt and Rothlauf, 2024]. Além disso, embora métricas de acurácia de imputação sejam amplamente utilizadas, a preservação das relações entre variáveis e os efeitos da imputação sobre tarefas subsequentes de análise ou modelagem também merecem consideração [Memon et al., 2023; Yang et al., 2025]. Nesse sentido, a lacuna investigada nesta

dissertação não está apenas na comparação entre diferentes métodos de imputação, mas principalmente na compreensão de *quais características estruturais dos dados condicionam o desempenho desses métodos.*

Diante desse cenário, a presente pesquisa busca compreender e caracterizar o problema da imputação de dados categóricos em diferentes configurações estruturais, formulando como questão central:

> **"Como a estrutura dos dados, considerando o número de atributos, a cardinalidade e as relações entre variáveis, influencia a dificuldade de imputação de valores categóricos?"**

A partir dessa questão, estabelece-se como hipótese de investigação que a dificuldade de imputação de valores categóricos pode variar de acordo com a complexidade estrutural dos dados, especialmente em função da cardinalidade, do número de atributos e da quantidade de informação disponível nas relações entre os atributos. Espera-se, adicionalmente, que esses efeitos não sejam uniformes entre os diferentes métodos de imputação, uma vez que cada abordagem explora de maneira distinta as informações presentes nas variáveis observadas.

Essa hipótese não implica assumir que a cardinalidade, isoladamente, determine o desempenho dos métodos. Ao contrário, parte-se da expectativa de que seus efeitos sejam condicionados pela estrutura conjunta da base, incluindo o número de atributos, a distribuição das categorias e as associações existentes entre os atributos. Dessa forma, a investigação busca verificar empiricamente em que medida essas características estão relacionadas à dificuldade de imputação e se determinados métodos apresentam maior robustez diante de diferentes configurações estruturais.

O objetivo central desta dissertação é analisar, de forma sistemática, os fatores que condicionam a dificuldade de imputação de dados categóricos, com ênfase no número de atributos, na cardinalidade e nas relações entre variáveis, estabelecendo fundamentos teóricos e evidências empíricas que contribuam para a compreensão do comportamento de diferentes métodos de imputação.

Para alcançar esse objetivo, a pesquisa busca: (i) caracterizar diferentes bases de dados segundo suas propriedades estruturais relevantes para o problema de imputação; (ii) avaliar o comportamento de diferentes métodos de imputação sob diferentes níveis de ausência de dados; (iii) investigar a influência do número de atributos e da cardinalidade sobre o desempenho dos métodos; (iv) analisar a relação entre a estrutura de associação dos atributos e a qualidade da imputação; e (v) verificar os impactos da imputação sobre características relevantes dos dados e sobre tarefas subsequentes de análise ou modelagem.

As principais contribuições desta dissertação estão, portanto, relacionadas à compreensão sistemática dos fatores estruturais que influenciam a imputação de dados categóricos. Em particular,

o trabalho contribui ao: (i) investigar empiricamente a influência da cardinalidade e do número de atributos sobre a dificuldade de imputação; (ii) analisar o papel das relações entre variáveis na capacidade de reconstrução dos valores ausentes; (iii) comparar o comportamento de diferentes métodos de imputação em bases com características estruturais distintas; e (iv) discutir o desempenho da imputação não apenas pela recuperação dos valores ausentes, mas também pelos efeitos produzidos sobre a estrutura dos dados e sobre análises subsequentes. Essas contribuições permitem deslocar a análise de uma perspectiva centrada exclusivamente na comparação de algoritmos para uma perspectiva que considera também as características da base de dados como possíveis determinantes do desempenho da imputação.

A relevância desta pesquisa se dá tanto no âmbito metodológico quanto em suas possíveis aplicações práticas. Em cenários reais, como pesquisas de opinião, registros clínicos e bancos de dados financeiros, a presença de valores ausentes pode comprometer análises e modelos construídos a partir de dados incompletos [Schafer, 1997; Little and Rubin, 2019]. Além disso, bases aplicadas frequentemente apresentam variáveis categóricas com diferentes níveis de cardinalidade e diferentes graus de associação entre seus atributos. Compreender como essas características se relacionam com a dificuldade de imputação pode fornecer subsídios para a avaliação e a seleção de métodos em diferentes configurações de dados.

É importante destacar que as conclusões desta pesquisa devem ser interpretadas à luz do delineamento experimental adotado. O estudo concentra-se no mecanismo MCAR e nas bases de dados e métodos definidos na metodologia, de modo que os resultados não devem ser automaticamente generalizados para mecanismos de ausência MAR ou MNAR, nem para todos os tipos de bases categóricas existentes. Essa delimitação permite controlar o mecanismo de ausência e concentrar a investigação na influência das características estruturais dos dados sobre o desempenho da imputação.

A estrutura desta dissertação está organizada da seguinte forma: o Capítulo II apresenta a fundamentação teórica, abordando os conceitos essenciais sobre valores ausentes, mecanismos de ausência e técnicas de imputação, com ênfase nas particularidades das variáveis categóricas. O Capítulo III discute os principais trabalhos relacionados, destacando lacunas, limitações e oportunidades de pesquisa identificadas na literatura. O Capítulo **??** descreve a metodologia adotada, detalhando o delineamento experimental, os critérios de seleção e caracterização das bases de dados, os métodos de imputação considerados e os indicadores utilizados para avaliação de desempenho. Por fim, o Capítulo VI apresenta as conclusões da dissertação, sintetizando os principais achados obtidos ao longo da pesquisa, discutindo suas contribuições teóricas e experimentais para o problema de imputação de dados categóricos, bem como suas limitações e perspectivas para trabalhos futuros.

Capítulo II Fundamentação Teórica

Neste capítulo, serão apresentados os principais conceitos e teorias relacionados à imputação de dados, com foco especial na imputação de dados categóricos. Discutiremos as técnicas amplamente utilizadas e como elas se posicionam no processo de descoberta de conhecimento em dados. Além disso, serão introduzidos conceitos gerais sobre a presença de dados faltantes e como diferentes métodos de imputação podem ser aplicados para lidar com essas ausências, com um foco detalhado em métodos específicos para imputação de variáveis categóricas.

II.1 Fundamentos de Imputação de Dados

A imputação de dados consiste no processo de substituição de valores ausentes por estimativas plausíveis, preservando a estrutura e a variabilidade dos dados originais [Rubin, 1987; van Buuren, 2018]. A escolha da técnica deve considerar o tipo de variável, a proporção de ausência e as características estruturais da base de dados.

A literatura apresenta diferentes famílias de métodos de imputação:

- Métodos simples, como imputação pela moda (para dados categóricos) ou média (para dados numéricos). Embora computacionalmente eficientes, podem distorcer distribuições e reduzir a variabilidade.
- Métodos estatísticos avançados, como *Multiple Imputation by Chained Equations* (MICE) e Expectation-Maximization (EM), que preservam incertezas associadas às estimativas [Little and Rubin, 2019].
- Métodos baseados em aprendizado de máquina, que utilizam algoritmos preditivos para estimar os valores ausentes considerando múltiplas variáveis simultaneamente. Nesse grupo, destacam-se KNN, Random Forest e Multilayer Perceptron (MLP), que combinam flexibilidade com alta capacidade de modelagem [Tang et al., 2021].

Quando aplicados a variáveis categóricas, esses métodos precisam lidar com desafios adicionais, como a inexistência de ordenação natural (em variáveis nominais) e a preservação da coerência semântica das categorias [Zimmermann et al., 2014].

II.2 Imputação de Dados Categóricos

Dados categóricos representam atributos qualitativos e podem ser nominais (sem ordem natural) ou ordinais (com hierarquia implícita). A imputação nesse contexto exige técnicas que respeitem a natureza discreta das variáveis, evitando imputações que introduzam categorias inexistentes ou distorçam relações estatísticas [Wu et al., 2012].

A preservação de métricas como a distribuição de frequência das categorias e medidas de associação (por exemplo, V de Cramer) é essencial para evitar viés em análises posteriores [Memon et al., 2023]. Além disso, medidas de similaridade adaptadas para dados categóricos, como distância de Gower, são amplamente utilizadas em métodos baseados em vizinhança [Yang et al., 2025].

II.3 Dados Faltantes: Morfologia da ausência

A ocorrência de dados faltantes é uma questão comum e complexa em diversas áreas de análise de dados. Dados faltantes, ou incompletos, referem-se a valores ausentes em variáveis de um conjunto de dados, podendo comprometer a integridade dos resultados e dificultar a interpretação das análises. Conforme destacado por [Rubin, 1976a], a correta compreensão da natureza desses dados é essencial para selecionar métodos de imputação que minimizem o impacto das ausências.

A presença de dados faltantes pode ocorrer por diversos motivos: desde falhas operacionais, como erros no processo de coleta de dados, até características próprias dos participantes, como respostas omitidas intencionalmente em questionários. As implicações de dados faltantes vão além da simples ausência, podendo alterar significativamente a representatividade da amostra e a robustez dos modelos analíticos. Isso torna fundamental não apenas o entendimento da morfologia dos dados faltantes, mas também a escolha de métodos que preservem as distribuições originais das variáveis e as interdependências entre elas [Graham, 2012].

Para tratar os dados faltantes de maneira adequada, [Rubin, 1976a] propôs três categorias que descrevem os mecanismos de ausência dos dados: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Missing Not at Random* (MNAR).

Missing Completely at Random (MCAR): Neste caso, a ausência de dados ocorre de forma totalmente aleatória e é independente de qualquer variável observada ou não observada. Isso significa que a probabilidade de um valor estar ausente é a mesma para todos os casos. Quando os dados são MCAR, a ausência não está relacionada às informações presentes ou ausentes na base, permitindo, sob essa hipótese, que métodos que desconsiderem as observações incompletas produzam estimativas não viesadas, embora com perda de eficiência devido à redução do tamanho amostral [Rubin, 1976a].

Na prática, a hipótese de MCAR pode ser investigada por meio de testes estatísticos, como o teste de Little, que verifica se os padrões de ausência observados são compatíveis com a hipótese de

ausência completamente aleatória [Little, 1988]. Além do teste estatístico, a análise deve considerar os padrões de ausência e o conhecimento sobre o processo de coleta dos dados. Portanto, um teste estatístico não deve ser interpretado isoladamente como uma comprovação definitiva do mecanismo de ausência.

No contexto desta pesquisa, entretanto, o mecanismo MCAR é controlado experimentalmente. As bases utilizadas são originalmente completas e os valores ausentes são introduzidos artificialmente na variável-alvo de maneira aleatória. Dessa forma, o mecanismo de ausência é conhecido a priori e não precisa ser inferido a partir dos dados. Essa estratégia permite avaliar os métodos de imputação sob diferentes proporções de ausência mantendo controlado o mecanismo responsável pela geração dos dados faltantes.

Missing at Random (MAR): Aqui, os dados faltantes dependem de outras variáveis observadas no conjunto de dados, mas não dos valores dos próprios dados faltantes, após consideradas as informações observadas. Por exemplo, em um estudo de saúde, é possível que as informações sobre altura estejam ausentes com maior frequência para pacientes de determinada faixa etária. Nesse caso, a ocorrência da ausência está associada à idade, que é uma variável observada.

Diferentemente do MCAR, a hipótese de MAR não pode, em geral, ser comprovada exclusivamente por meio de um teste estatístico aplicado aos dados observados, pois depende de informações sobre o processo que produziu a ausência [Rubin, 1976a]. Sua plausibilidade deve ser avaliada por meio da análise dos padrões de ausência, da relação entre a ocorrência da ausência e as variáveis observadas e, principalmente, do conhecimento sobre o processo de coleta dos dados.

Missing Not at Random (MNAR): A ausência dos dados está relacionada ao próprio valor que deveria estar presente, mesmo após considerar as informações observadas, o que torna a situação mais complexa. Por exemplo, um paciente pode optar por não informar o peso em um questionário devido a preocupações pessoais, e essa ausência pode estar relacionada ao próprio valor da variável.

O mecanismo MNAR é particularmente difícil de ser identificado a partir dos dados observados, uma vez que a informação necessária para avaliar a dependência está justamente ausente. Dessa forma, sua análise normalmente exige hipóteses adicionais sobre o processo de ausência, modelos específicos ou análises de sensibilidade [Rubin, 1976a].

Assim, a distinção entre MCAR, MAR e MNAR não deve ser baseada exclusivamente em um teste estatístico. A identificação do mecanismo deve combinar evidências estatísticas, análise dos padrões de ausência e conhecimento sobre o processo de geração e coleta dos dados. No presente trabalho, a utilização de bases originalmente completas e a introdução artificial dos valores ausentes permitem controlar explicitamente o mecanismo MCAR empregado nos experimentos.

Além da questão de como os dados faltantes surgem, é importante considerar a frequência e a estrutura das ausências. Ausências que ocorrem de forma esparsa, em poucos casos, podem ser

tratadas de forma diferente de dados faltantes em grande volume ou em estruturas sistemáticas, como blocos inteiros de variáveis ausentes. Em situações onde há uma grande quantidade de dados faltantes, a imputação manual pode ser inviável, e técnicas automatizadas tornam-se essenciais [Lee et al., 2023].

II.4 Causas de Dados Ausentes

As causas que originam valores ausentes podem ser agrupadas em três grandes dimensões: operacionais, comportamentais e tecnológicas.

Causas operacionais incluem erros durante a coleta, armazenamento ou transmissão dos dados, como falhas de preenchimento, digitação ou sincronização entre sistemas. Mudanças de formato ou de protocolo de coleta ao longo do tempo também geram inconsistências e lacunas [Graham, 2012].

Causas comportamentais estão relacionadas às escolhas ou limitações dos respondentes. É comum, por exemplo, a omissão deliberada de respostas a perguntas sensíveis, levando a padrões sistemáticos de ausência frequentemente associados ao mecanismo MNAR [Rubin, 1976b].

Causas tecnológicas referem-se a problemas decorrentes de limitações de hardware ou software. Migrações de sistemas, erros de conversão de formatos e falhas de sensores podem gerar ausências pontuais ou contínuas.

Em muitos casos, a ausência é resultado da interação de múltiplos fatores, gerando mecanismos mistos que dificultam a modelagem. A identificação dessas causas, quando possível, contribui para a escolha de métodos de imputação mais adequados e para a interpretação correta dos resultados [Lee et al., 2023].

II.5 Medidas Estatísticas e Métricas de Similaridade

Nesta pesquisa, diferentes medidas estatísticas e métricas de similaridade são utilizadas para analisar associações entre variáveis e apoiar os processos de imputação de dados categóricos. Entre elas, destacam-se o coeficiente Eta², o V de Cramer e a distância de Gower, amplamente empregados em cenários envolvendo dados mistos e variáveis categóricas [Gower, 1971; Akoglu, 2018].

II.5.1 Coeficiente Eta²

O coeficiente Eta² (η^2) é uma medida estatística utilizada para quantificar a associação entre uma variável categórica e uma variável numérica. Essa medida é frequentemente empregada em análises de variância (ANOVA), representando a proporção da variabilidade total da variável numérica explicada pela variável categórica [Haykin, 2009].

O coeficiente pode ser definido como:

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

em que $SS_{between}$ representa a soma dos quadrados entre grupos e SS_{total} representa a soma total dos quadrados.

O termo "soma dos quadrados" refere-se ao cálculo das diferenças entre determinados valores e uma média de referência, elevando essas diferenças ao quadrado e somando os resultados. Dessa forma, em $SS_{between}$, são consideradas as diferenças entre a média de cada grupo e a média geral, enquanto em SS_{total} são consideradas as diferenças entre cada observação individual e a média geral.

A soma dos quadrados entre grupos pode ser expressa por:

$$SS_{between} = \sum_{j=1}^{g} n_j(\bar{Y}_j - \bar{Y})^2$$

em que g representa o número de grupos, n_j representa o número de observações pertencentes ao grupo j, $\bar{Y}_j$ representa a média da variável numérica no grupo j e $\bar{Y}$ representa a média geral da variável numérica.

Já a soma total dos quadrados é definida por:

$$SS_{total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

em que n representa o número total de observações, Y_i representa o valor da observação i e $\bar{Y}$ representa a média geral da variável.

Assim, o $SS_{between}$ mede quanto as médias dos grupos se afastam da média geral, enquanto o SS_{total} representa a variabilidade total existente nos valores da variável numérica. A razão entre essas duas quantidades resulta no coeficiente Eta**²**.

Os valores de η^2 variam entre 0 e 1. Valores próximos de 0 indicam que uma pequena parcela da variabilidade da variável numérica está associada à divisão em grupos, enquanto valores próximos de 1 indicam que uma parcela maior da variabilidade está associada às diferenças entre os grupos.

Nesta pesquisa, o coeficiente Eta**²** é utilizado para ponderar atributos numéricos no processo de imputação com KNN, atribuindo maior relevância às variáveis que apresentam maior associação com a variável categórica alvo.

II.5.2 V de Cramer

O V de Cramer é uma medida de associação entre variáveis categóricas baseada na estatística qui-quadrado (χ^2). Essa métrica é amplamente utilizada para avaliar o grau de associação entre atributos nominais [Akoglu, 2018].

O cálculo do V de Cramer parte de uma tabela de contingência, na qual são contabilizadas as frequências de ocorrência das combinações entre as categorias das duas variáveis analisadas. A estatística qui-quadrado compara as frequências observadas com as frequências que seriam esperadas caso as variáveis fossem independentes.

A estatística qui-quadrado é calculada por:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

em que O_{ij} representa a frequência observada na combinação da categoria i de uma variável com a categoria j da outra variável, enquanto E_{ij} representa a frequência esperada para essa mesma combinação caso as variáveis fossem independentes.

A frequência esperada é obtida a partir dos totais das linhas e das colunas da tabela de contingência, conforme:

$$E_{ij} = \frac{(\text{total da linha } i)(\text{total da coluna } j)}{n}$$

em que n corresponde ao número total de observações analisadas.

A estatística χ^2 quantifica, portanto, o quanto as frequências observadas se afastam das frequências esperadas sob a hipótese de independência. Quanto maior essa diferença, maior tende a ser o valor da estatística qui-quadrado.

A partir da estatística qui-quadrado, o V de Cramer é definido como:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

em que χ^2 representa a estatística qui-quadrado obtida na tabela de contingência, n corresponde ao número total de observações e k representa o menor número de categorias entre as duas variáveis analisadas.

Os valores do V de Cramer variam entre 0 e 1. Valores próximos de 0 indicam baixa associação entre as variáveis, enquanto valores maiores indicam maior grau de associação entre as variáveis categóricas.

Neste trabalho, o V de Cramer é utilizado como medida auxiliar para analisar a preservação das associações entre variáveis categóricas após o processo de imputação.

II.5.3 Distância de Gower

A distância de Gower foi proposta por [Gower, 1971] como uma métrica de similaridade adequada para bases contendo atributos mistos, incluindo variáveis numéricas, categóricas e ordinais.

Para duas observações i e j, a similaridade de Gower é calculada como:

$$S_{ij} = \frac{1}{p} \sum_{k=1}^{p} s_{ijk}$$

em que p representa o número total de atributos e s_{ijk} representa a similaridade parcial entre as observações i e j para o atributo k.

No caso de variáveis categóricas, a similaridade parcial assume valor 1 quando as categorias coincidem e 0 caso contrário. Para variáveis numéricas, a similaridade é calculada considerando a diferença normalizada entre os valores observados.

A distância de Gower é então obtida por:

$$d_{ij} = 1 - S_{ij}$$

Essa métrica é particularmente relevante em problemas de imputação categórica envolvendo bases heterogêneas, pois permite combinar atributos numéricos e categóricos em uma única medida de proximidade. Nesta pesquisa, a distância de Gower é utilizada na variante do KNN aplicada a bases mistas e categóricas.

II.6 Métodos Utilizados nesta Pesquisa

Nesta seção são apresentados os três algoritmos de imputação adotados neste estudo: K-Nearest Neighbors (KNN), Random Forest (RF) e Multilayer Perceptron (MLP), com ênfase nas adaptações necessárias para o tratamento de dados categóricos. Cada método é discutido quanto ao seu princípio de funcionamento, características relevantes e forma de aplicação no contexto de imputação sob mecanismo MCAR.

II.6.1 K-Nearest Neighbors (KNN)

O algoritmo *K-Nearest Neighbors* (KNN) é um método não paramétrico que baseia suas inferências no princípio da proximidade: instâncias semelhantes tendem a compartilhar valores semelhantes de atributos [Wu et al., 2012]. Para imputação de valores faltantes, o KNN identifica, para cada instância incompleta, os K registros mais próximos no espaço das variáveis observadas e estima o valor ausente a partir dos vizinhos mais similares.

Em dados categóricos, a definição da medida de similaridade é essencial. A distância de Gower [Gower, 1971], amplamente utilizada em dados mistos, é definida como:

$$d_{Gower}(i, j) = 1 - \frac{1}{p} \sum_{k=1}^{p} s_{ijk},$$

em que s_{ijk} representa o grau de similaridade entre as observações i e j para o atributo k, assumindo valor 1 se as categorias coincidirem e 0 caso contrário. Essa métrica permite combinar variáveis nominais e ordinais de forma consistente, atribuindo pesos iguais a todas as variáveis.

O processo de imputação ocorre substituindo o valor ausente pela categoria mais frequente (*moda*) entre os K vizinhos mais próximos. Alternativamente, é possível ponderar as categorias pela distância, atribuindo maior peso aos vizinhos mais similares.

O Algoritmo 1 apresenta o processo geral de imputação utilizando KNN. Inicialmente, para cada observação que apresenta um valor ausente, são identificadas as observações que possuem valores disponíveis para os atributos utilizados no cálculo da distância. Em seguida, a distância entre a observação incompleta e as demais observações é calculada. A partir dessas distâncias, são selecionados os K vizinhos mais próximos. Finalmente, as categorias observadas nesses vizinhos são utilizadas para determinar o valor a ser imputado, sendo escolhida a categoria mais frequente. Assim, o algoritmo utiliza a proximidade entre observações como principal fonte de informação para reconstruir o valor ausente.

Algoritmo 1: Imputação KNN para dados categóricos

Require: Conjunto de dados D, número de vizinhos K

1: **for** cada instância x_i com valor ausente em atributo A **do**
2: Calcular a distância $d(x_i, x_j)$ para todas as instâncias completas, usando a distância de Gower
3: Selecionar os K vizinhos mais próximos $\{x_{(1)}, \ldots, x_{(K)}\}$
4: Determinar a categoria c^* mais frequente em A entre os vizinhos
5: Atribuir $A_i = c^*$
6: **end for**
7: **return** Conjunto D com valores imputados

O pseudocódigo apresentado no Algoritmo 1 representa, portanto, uma sequência de quatro etapas principais: identificação das observações com valores ausentes, cálculo das distâncias, seleção dos vizinhos e determinação da categoria a ser imputada. A repetição do procedimento para cada observação incompleta resulta em uma base na qual os valores ausentes foram substituídos pelas categorias estimadas a partir das observações mais semelhantes.

As principais vantagens do KNN são sua simplicidade e aplicabilidade a diferentes tipos de dados. Contudo, o método é sensível à escolha de K, ao escalonamento das variáveis e pode ter desempenho reduzido em bases de alta dimensionalidade.

II.6.2 Random Forest (RF)

A *Random Forest* (RF) [Breiman, 2001] é um método baseado em um conjunto de árvores de decisão construídas de forma independente a partir de subconjuntos amostrados de dados e atributos. No contexto de imputação, a técnica *missForest* [Stekhoven and Bühlmann, 2012] adapta

o algoritmo para lidar iterativamente com dados incompletos.

O procedimento consiste em inicializar os valores ausentes (por exemplo, pela moda), treinar uma floresta para cada atributo com valores faltantes utilizando os demais como preditores e, em seguida, atualizar os valores imputados. Esse processo é repetido até a convergência dos valores ou até que a variação média entre iterações se torne mínima.

O Algoritmo 2 apresenta esse procedimento de forma simplificada. Primeiramente, os valores ausentes são inicializados, permitindo que o processo de treinamento seja iniciado. Para cada atributo que apresenta valores faltantes, esse atributo é tratado como variável resposta e os demais atributos são utilizados como variáveis preditoras. A Random Forest é então treinada utilizando as observações nas quais o atributo de interesse está disponível. Após o treinamento, o modelo estima os valores ausentes desse atributo. O procedimento é repetido para os demais atributos e, posteriormente, as novas imputações são utilizadas nas iterações seguintes. O processo é encerrado quando os valores imputados deixam de apresentar alterações relevantes entre as iterações.

Algoritmo 2: Imputação *missForest* para dados categóricos

Require: Conjunto de dados D com valores ausentes
1: Inicializar valores ausentes pela moda de cada atributo
2: **repeat**
3: **for** cada atributo A_j com valores ausentes **do**
4: Definir A_j como variável resposta e demais atributos como preditores
5: Treinar uma Random Forest classificatória usando instâncias completas
6: Prever valores faltantes em A_j e substituir as imputações anteriores
7: **end for**
8: **until** convergência dos valores imputados
9: **return** Conjunto D imputado

Dessa forma, o pseudocódigo pode ser interpretado como um processo iterativo de refinamento. A moda é utilizada apenas para fornecer uma primeira estimativa dos valores ausentes. A partir dessa inicialização, os modelos de Random Forest passam a utilizar as demais variáveis para obter estimativas mais informativas. A cada nova iteração, os valores imputados anteriormente podem ser substituídos pelas novas previsões, permitindo que o modelo refine progressivamente as estimativas.

A RF é particularmente vantajosa em cenários com variáveis categóricas e interações complexas, pois é capaz de capturar relações não lineares e lidar com dados ruidosos. Além disso, o uso de *bootstrap* e amostragem aleatória de atributos reduz o risco de *overfitting*. Sua limitação principal reside no custo computacional, que cresce com o número de árvores e a dimensionalidade dos dados.

II.6.3 Multilayer Perceptron (MLP)

O *Multilayer Perceptron* (MLP) é uma arquitetura de rede neural artificial composta por múltiplas camadas de neurônios interconectados, capaz de modelar relações complexas e não lineares

entre variáveis [Goodfellow et al., 2016]. O modelo é treinado por meio do algoritmo de retropropagação do erro (*backpropagation*), ajustando iterativamente os pesos das conexões internas para minimizar uma função de perda.

No contexto de imputação de dados categóricos, o MLP pode ser utilizado como um modelo supervisionado de classificação, no qual os valores ausentes são estimados a partir das demais variáveis observadas da base de dados. Para isso, os atributos categóricos precisam ser previamente codificados em representação numérica, geralmente utilizando técnicas como *one-hot encoding*.

A codificação *one-hot* representa cada categoria de uma variável por uma variável binária distinta. Por exemplo, considerando uma variável categórica *Cor*, composta pelas categorias "Azul", "Verde" e "Vermelho", cada observação pode ser representada por um vetor de três posições:

$$\text{Azul} = (1, 0, 0)$$

$$\text{Verde} = (0, 1, 0)$$

$$\text{Vermelho} = (0, 0, 1)$$

Assim, para cada observação, apenas a posição correspondente à sua categoria recebe o valor 1, enquanto as demais recebem o valor 0. Essa representação é adequada para variáveis categóricas nominais porque não estabelece uma ordem ou uma distância artificial entre as categorias. Por exemplo, a utilização direta dos valores 1, 2 e 3 para representar "Azul", "Verde" e "Vermelho" poderia fazer com que um modelo interpretasse que existe uma relação de ordem ou distância entre essas categorias, embora tal relação não exista.

No MLP, a codificação *one-hot* permite que as categorias sejam utilizadas como entradas numéricas da rede sem introduzir esse tipo de relação artificial. Os valores binários resultantes são então utilizados pelas camadas da rede para aprender padrões associados às demais variáveis.

Nesta pesquisa, o MLP é empregado para prever categorias ausentes utilizando as demais variáveis como atributos preditores. A arquitetura utilizada possui múltiplas camadas ocultas e função de ativação ReLU, permitindo capturar padrões não lineares e interações complexas entre variáveis.

A função de ativação ReLU (*Rectified Linear Unit*) é definida por:

$$\text{ReLU}(x) = \max(0, x).$$

Isso significa que valores negativos recebidos pelo neurônio são transformados em zero, enquanto valores positivos são mantidos. A função introduz uma não linearidade na rede, permitindo que o

MLP aprenda padrões e associações que não poderiam ser representados adequadamente por uma combinação exclusivamente linear das variáveis de entrada [Goodfellow et al., 2016].

O processo de imputação baseado no MLP é apresentado no Algoritmo 3. Inicialmente, as variáveis categóricas são transformadas para uma representação numérica adequada ao modelo. Em seguida, os valores ausentes são inicializados para permitir o treinamento. Para cada atributo que possui valores ausentes, o atributo é considerado como variável resposta e os demais atributos são utilizados como variáveis preditoras. O MLP é treinado utilizando as observações nas quais o valor da variável resposta está disponível. Após o treinamento, o modelo realiza a previsão das categorias correspondentes às observações com valores ausentes.

Algoritmo 3: Imputação baseada em MLP

Require: Conjunto de dados D com valores ausentes
1: Aplicar codificação às variáveis categóricas
2: Inicializar valores ausentes
3: **for** cada atributo com valores ausentes **do**
4: Definir o atributo como variável resposta
5: Treinar o modelo MLP utilizando instâncias completas
6: Prever os valores ausentes
7: Substituir os valores faltantes pelas previsões obtidas
8: **end for**
9: **return** Conjunto de dados imputado

O pseudocódigo do Algoritmo 3 pode ser dividido em quatro etapas principais. A primeira consiste na transformação dos dados categóricos para uma representação que possa ser processada pela rede neural. A segunda corresponde à definição dos valores iniciais dos dados faltantes. A terceira etapa consiste no treinamento do MLP utilizando as observações completas e na realização das previsões para as observações incompletas. Por fim, os valores ausentes são substituídos pelas categorias previstas pelo modelo.

O MLP apresenta elevada capacidade de modelagem e adaptação a relações complexas entre atributos, sendo especialmente útil em bases com alta dimensionalidade. Entretanto, seu desempenho depende da escolha adequada de hiperparâmetros, da quantidade de dados disponíveis e do pré-processamento aplicado às variáveis [Goodfellow et al., 2016].

Os três métodos descritos KNN, Random Forest e MLP, representam abordagens complementares para imputação de dados categóricos. O KNN enfatiza a similaridade entre instâncias por meio de métricas de distância, a Random Forest explora o uso de comitês de árvores de decisão capazes de capturar relações não lineares entre atributos, enquanto a MLP utiliza redes neurais artificiais para modelar padrões complexos presentes nos dados. A escolha dessas técnicas permite analisar diferentes estratégias de imputação sob o mecanismo MCAR, avaliando o equilíbrio entre simplicidade, robustez, capacidade de generalização e poder de modelagem em bases com características

distintas e diferentes proporções de valores ausentes.

Os conceitos apresentados neste capítulo fundamentam a abordagem experimental adotada nesta dissertação. A discussão sobre os mecanismos de ausência permite delimitar as condições sob as quais os experimentos são realizados, enquanto as medidas estatísticas e as métricas de similaridade fornecem recursos para representar as características dos atributos e suas associações. Os métodos KNN, Random Forest e MLP, por sua vez, exploram essas informações de maneiras distintas para estimar os valores ausentes.

A partir desses fundamentos, o próximo capítulo apresenta a revisão bibliográfica dos principais estudos relacionados à imputação de dados categóricos, com ênfase nos métodos e abordagens utilizados na literatura, bem como nas limitações identificadas que motivam o desenvolvimento desta pesquisa.

Capítulo III Trabalhos relacionados

Este capítulo apresenta uma revisão dos principais trabalhos relacionados ao tema de imputação de dados categóricos, com o objetivo de embasar as discussões e propostas desta dissertação. A análise dos estudos existentes, obtidos a partir de uma busca detalhada na base Scopus, visa fornecer uma visão abrangente do estado da arte na área e evidencia as lacunas que motivaram esta pesquisa.

III.1 Seleção dos Trabalhos

Para garantir que a revisão bibliográfica contemplasse as contribuições mais relevantes no campo da imputação de dados categóricos, foi utilizada a base de dados Scopus, reconhecida por sua ampla cobertura multidisciplinar e qualidade editorial, especialmente na área de Computação. A consulta foi construída a partir de expressões consagradas na literatura científica, buscando capturar estudos que discutissem especificamente métodos e técnicas voltados à imputação de variáveis categóricas. A string de busca utilizada foi a seguinte:

> *("categorical data imputation"OR "imputation methods for categorical data"OR "missing categorical data"OR "categorical variable imputation"OR "imputation techniques for categorical data"OR "categorical data missing values"OR "missing categorical data") AND NOT ("clinical data"OR "medical research")*

Essa formulação resultou em 45 artigos encontrados na Scopus em setembro de 2026. Para garantir a relevância dos estudos selecionados, foi realizada uma triagem inicial baseada na leitura dos *abstracts*, aplicando um critério de correlação com a temática da dissertação, classificado em três níveis:

Alta correlação: trabalhos que tratam diretamente de métodos ou análises de imputação de dados categóricos, com foco experimental ou teórico relevante para esta pesquisa;

Média correlação: estudos que abordam imputação de dados em geral, incluindo variáveis categóricas, porém com menor ênfase específica no tema central;

Baixa correlação: artigos cujo foco principal se afasta do tema de imputação categórica, abordando outras áreas ou tipos de dados, sendo descartados para análise detalhada.

Os artigos classificados com baixa correlação foram identificados por meio da leitura dos títulos e resumos, verificando a presença de termos e discussões diretamente relacionados à imputação de da-

dos categóricos. Artigos cujo foco principal se afastava do tema central, por exemplo, estudos sobre imputação exclusivamente numérica, métodos estatísticos gerais sem aplicação a dados categóricos, ou contextos clínicos e biomédicos, foram categorizados como de baixa correlação e, consequentemente, descartados da leitura completa. Esse processo permitiu concentrar a análise nos trabalhos mais pertinentes à pesquisa. A análise aprofundada concentrou-se nos trabalhos de alta correlação, garantindo que a revisão bibliográfica refletisse os avanços e desafios mais pertinentes à área de imputação de dados categóricos.

A formulação da consulta visou garantir que os resultados recuperados estivessem alinhados com o escopo da pesquisa, excluindo explicitamente estudos voltados ao domínio médico ou clínico, que geralmente apresentam características metodológicas e objetivos distintos dos propostos neste trabalho.

III.1.1 Critérios de Inclusão

Para selecionar os estudos mais pertinentes à análise, foram adotados os seguintes critérios de inclusão:

- Trabalhos que abordem direta e explicitamente técnicas de imputação de variáveis categóricas;
- Estudos que apresentem avaliação experimental ou análise empírica dos métodos propostos ou utilizados;
- Artigos veiculados em periódicos ou conferências de reconhecida relevância científica nas áreas de ciência de dados, estatística ou inteligência artificial.

III.1.2 Critérios de Exclusão

Com o objetivo de alinhar a seleção aos objetivos específicos da dissertação, foram definidos os seguintes critérios de exclusão:

- Trabalhos cujo foco principal seja a imputação de dados em contextos clínicos, biomédicos ou de saúde, identificados pelos termos *"clinical data"* e *"medical research"* presentes na string de busca;
- Estudos voltados exclusivamente à imputação de variáveis numéricas ou contínuas, sem qualquer abordagem voltada a dados categóricos;

III.1.3 Processo de Seleção

Após a execução da busca, os títulos e resumos dos 42 artigos retornados foram inicialmente inspecionados, com o objetivo de realizar uma triagem preliminar. Os estudos que se mostraram

aderentes aos critérios estabelecidos foram então analisados com maior profundidade por meio da leitura integral, permitindo a identificação de suas principais contribuições, limitações e metodologias empregadas. Os artigos selecionados serviram de base para a construção do capítulo de revisão de literatura, contextualizando a presente pesquisa dentro do estado da arte em imputação de dados categóricos.

III.2 Comparação dos Trabalhos Relacionados

A seguir apresentarei uma visão geral de diferentes trabalhos relacionados à imputação de dados categóricos, seguidos de uma comparação detalhada. Cada estudo é resumido de forma individual, destacando seus métodos, objetivos e contribuições. Em seguida, comparo esses trabalhos em termos de contexto de aplicação, eficiência e limitações.

1. **A comparison of imputation methods for categorical data (Memon et al., 2023)**:

   [Memon et al., 2023] realizaram uma análise abrangente de cinco métodos de imputação para lidar com valores ausentes em variáveis categóricas, frequentemente encontradas em bases de dados clínicos. Os métodos avaliados foram: *Mode*, *K-Nearest Neighbors* (KNN), *Random Forest* (RF), *Sequential Hot-Deck* (SHD) e *Multiple Imputation by Chained Equations* (MICE). O estudo utilizou dados de saúde materna extraídos de registros médicos em papel do Kawempe National Referral Hospital, em Uganda, com variáveis categóricas de três tipos: binárias, ordinais e multinomiais.

   O objetivo principal foi determinar qual método de imputação apresenta melhor desempenho em cenários de dados faltantes, considerando diferentes proporções de ausência (5% a 50%), todas sob o mecanismo MCAR (*Missing Completely at Random*). Para isso, os métodos foram avaliados em duas dimensões: (1) precisão na imputação, medida como a proporção de valores imputados que correspondem aos valores reais; e (2) acurácia preditiva em quatro classificadores (regressão logística, *Naïve Bayes*, máquinas de suporte vetorial e *Random Forest*) aplicados aos dados após imputação.

   Os resultados indicaram que o método KNN foi consistentemente o mais preciso na imputação de valores ausentes em todos os níveis de dados faltantes avaliados. O RF apresentou a segunda maior precisão em proporções menores de ausência (5% a 20%). Por outro lado, o SHD demonstrou o pior desempenho em todos os cenários testados. No entanto, ao considerar a acurácia preditiva, nenhum método se destacou como universalmente superior. Os resultados variaram amplamente entre os classificadores, mostrando que uma maior precisão na imputação nem sempre se traduz em uma melhoria na capacidade preditiva do modelo.

Além disso, o estudo utilizou o coeficiente de concordância de Kendall ($W = 0.853$, $p < 0.001$) para validar a consistência no desempenho dos métodos em diferentes proporções de dados faltantes. Esses achados reforçam a complexidade do problema de imputação e indicam que a escolha do método ideal depende do contexto de aplicação, como o tipo de análise subsequente e a proporção de valores ausentes.

Os autores destacaram a relevância do KNN para imputação em variáveis categóricas, mas enfatizaram a necessidade de futuras pesquisas para explorar o desempenho dos métodos em mecanismos MAR (*Missing At Random*) e MNAR (*Missing Not At Random*). O trabalho fornece insights importantes para pesquisadores que lidam com bancos de dados clínicos categóricos, especialmente em contextos de países em desenvolvimento onde registros em papel ainda são amplamente utilizados.

2. **A comparison of imputation methods using machine learning models (Suh & Song, 2023):**

   [Suh and Song, 2023] apresentam uma comparação detalhada entre métodos convencionais e modernos de imputação em bases de dados mistos (variáveis categóricas e numéricas) em diferentes cenários. Os métodos avaliados incluem imputação por média/moda, *K-Nearest Neighbors* (KNN), *Multiple Imputation by Chained Equations* (MICE), *missForest* e *mixgb*. Este último é baseado no *XGBoost*, destacando-se por sua capacidade de capturar relações complexas entre variável s e alta eficiência computacional.

   O estudo propõe uma nova métrica de avaliação, a *Imputation Performance Measure* (IPM), baseada na distância de Gower, que permite comparar métodos de imputação em dados mistos. Essa métrica combina erros de classificação para variáveis categóricas e erros quadráticos médios (RMSE) para variáveis numéricas. Os experimentos foram realizados em seis conjuntos de dados de tamanhos variados (1.059 a 56.420 observações), sob diferentes proporções de valores ausentes (10%, 20%, e 30%) e mecanismos MCAR (*Missing Completely at Random*).

   Os resultados demonstraram que o *mixgb* teve o melhor desempenho geral em cenários complexos com altas proporções de valores ausentes. Em conjuntos menores ou com baixos níveis de dados faltantes, o KNN foi competitivo em termos de precisão e tempo de execução. O *missForest* mostrou-se robusto, mas foi superado pelo *missRanger* em cenários de grandes volumes de dados, devido à maior eficiência computacional deste último. Os autores destacaram que o método *missRanger* é mais adequado quando os valores ausentes ocorrem de forma aleatória em apenas algumas colunas do conjunto de dados.

   Além disso, o trabalho ressalta que a escolha do método de imputação ideal depende das características específicas dos dados, como tamanho, proporção de valores ausentes e tipo de

variável. O estudo conclui que métodos baseados em aprendizado de máquina, como o *mixgb*, representam um avanço significativo em relação aos métodos tradicionais, mas requerem maior custo computacional.

3. **Two-stage imputation method to handle missing data for categorical response variable (Kim et al., 2023):**

   [Kim et al., 2023] propõem um método de imputação em duas etapas para lidar com dados faltantes em variáveis categóricas de resposta, utilizando uma combinação de seleção de variáveis e imputação multivariada. No primeiro estágio, o algoritmo de seleção de variáveis *Boruta* é aplicado para identificar os preditores mais relevantes. No segundo estágio, os dados faltantes são imputados com base no método MICE (*Multiple Imputation by Chained Equations*), utilizando as variáveis selecionadas.

   Os autores avaliaram o método em dois cenários: simulações com dados assimétricos gerados via cópula de Clayton e análises com dados reais do *National Health and Nutrition Examination Survey* (NHANES III). As métricas de precisão mostraram que o método proposto supera as abordagens convencionais, especialmente na presença de taxas moderadas a altas de ausência de dados (10% a 40%) sob o mecanismo MAR (*Missing At Random*).

   O primeiro estágio, baseado no *Boruta*, seleciona variáveis levando em consideração relações multivariadas, o que reduz o conjunto de variáveis necessárias para a imputação. O segundo estágio aplica técnicas específicas para diferentes tipos de variáveis: regressão logística para dados binários, regressão polinomial para variáveis categóricas com múltiplos níveis e *Predictive Mean Matching* (PMM) para variáveis numéricas.

   Nos experimentos de simulação, o método apresentou maior precisão em comparação com o MICE padrão, mostrando-se eficiente em contextos de dados correlacionados e assimétricos. Na aplicação com dados reais, o método também foi mais preciso do que abordagens sem seleção de variáveis. Isso ilustra como a seleção prévia melhora a qualidade da imputação ao reduzir o viés e a variância associados aos estimadores.

   O trabalho conclui que o método de imputação em duas etapas oferece ganhos significativos em precisão e eficiência computacional, sendo especialmente útil para cenários onde variáveis categóricas são predominantes. Futuras extensões incluem a aplicação em dados longitudinais e contextos econômicos.

4. **Missing Categorical Data Imputation and Individual Observation Level Imputation (Zimmermann et al., 2014):**

   [Zimmermann et al., 2014] propõem uma nova metodologia para imputação de dados categó-

ricos ausentes, utilizando regressão logística multinomial combinada com a busca pelo modo da distribuição multinomial. O objetivo principal é superar as limitações das técnicas tradicionais, que tendem a super-representar categorias dominantes ao atribuir valores "mais prováveis". O método busca preservar a estrutura esperada dos dados ausentes condicionada aos valores observados, respeitando as proporções esperadas de cada categoria.

A metodologia é composta por quatro etapas principais: (1) identificação de determinantes significativos para a variável com valores ausentes; (2) criação de perfis baseados nas combinações únicas dos determinantes; (3) estimativa das probabilidades condicionais para cada categoria utilizando regressão logística multinomial; e (4) distribuição das contagens imputadas entre os indivíduos em cada perfil, maximizando a probabilidade conjunta na distribuição multinomial.

O artigo apresenta uma aplicação prática da metodologia utilizando dados demográficos da República Tcheca, relacionados à distribuição educacional da população entre 1995 e 2011. A variável "educação"foi selecionada como variável de interesse com valores ausentes significativos nos anos de 2010 e 2011. A imputação foi realizada com base em determinantes como gênero, estado civil, idade e causa da morte. Os resultados mostraram que o método proposto foi capaz de preservar tendências temporais nos dados imputados, mesmo em cenários com proporções significativas de dados ausentes.

A abordagem demonstrou vantagens sobre técnicas convencionais, como imputação por moda e máxima verossimilhança, ao reduzir vieses associados à super-representação de categorias dominantes. Os autores destacam que o método é aplicável a diversos contextos, como dados de status social, e que futuros estudos devem explorar sua generalização para diferentes tipos de mecanismos de dados ausentes (MAR e MNAR).

5. **Missing Categorical Data Imputation Approach Based on Similarity (Wu et al., 2012):**

   [Wu et al., 2012] propõem um algoritmo inovador denominado MIBOS (*Missing Categorical Data Imputation Based on Similarity*) para imputação de dados categóricos ausentes. O método combina a teoria de conjuntos aproximados (*rough sets*) com a abordagem de vizinhos mais próximos, adaptando-a para lidar com variáveis categóricas. Diferentemente de métodos paramétricos, como o KNN, o MIBOS é um algoritmo não paramétrico que elimina a necessidade de escolher o número de vizinhos (K), o que o torna mais flexível para diferentes configurações de dados.

   O algoritmo define um modelo de similaridade entre objetos com dados incompletos e constrói uma matriz de similaridade para identificar os conjuntos de objetos mais semelhantes

(*nearest undifferentiated sets*). Durante o processo de imputação, os valores imputados são imediatamente utilizados nas iterações subsequentes, permitindo um aumento progressivo na completude dos dados. O MIBOS também possui uma variante chamada MIBOS-DA (*Descending Ascending*), que utiliza estratégias de ordenação descendente e ascendente para melhorar a precisão e a taxa de completude dos dados imputados.

Experimentos foram conduzidos em três conjuntos de dados do repositório UCI: Zoo, SPECT Heart e Voting. Os dados foram corrompidos artificialmente com taxas de ausência variando de 5% a 70%, sob o mecanismo MCAR (*Missing Completely at Random*). Comparações entre o MIBOS, MIBOS-DA, KNN e imputação por moda revelaram que o MIBOS-DA superou os outros métodos em termos de taxa de completude, acurácia absoluta e relativa, especialmente em cenários com altas taxas de dados ausentes. O MIBOS-DA também apresentou melhor eficiência computacional em relação ao KNN, embora com maior custo de tempo do que o MIBOS padrão.

Os autores concluem que o MIBOS-DA, combinado com a imputação por moda, é uma abordagem robusta e eficiente para lidar com dados categóricos ausentes, sendo especialmente útil em contextos de mineração de dados onde a taxa de valores ausentes é alta.

6. **Software Cost Estimation: A State-of-the-Art Statistical and Visualization Approach for Missing Data (Chatzipetrou, 2019):**

   [Chatzipetrou, 2019] investiga o impacto de diferentes técnicas de imputação de dados faltantes em modelos de estimativa de custos de software, um problema comum em projetos de desenvolvimento de software onde conjuntos de dados categóricos frequentemente apresentam valores ausentes. O estudo compara cinco métodos: regressão logística multinomial (MLR), exclusão por lista (*Listwise Deletion*, LD), imputação por média (*Mean Imputation*, MI), maximização da expectativa (EM) e imputação por regressão (RI).

   Os dados analisados foram extraídos do repositório ISBSG, contendo 152 projetos categorizados por variáveis como tipo de plataforma, linguagem de programação e banco de dados utilizado. As métricas de avaliação incluem precisão, viés e dispersão de erro. Além disso, a metodologia incorpora técnicas de visualização, como curvas REC (*Regression Error Characteristic*) e RROC (*Regression Receiver Operating Characteristic*), que permitem analisar graficamente o desempenho preditivo e os padrões de erro de cada técnica.

   Os resultados indicaram que o método MLR apresentou o melhor desempenho geral em termos de precisão, menor viés e menor dispersão dos erros. Ele se destacou particularmente na capacidade de lidar com valores ausentes em variáveis categóricas e melhorar a acurácia preditiva dos modelos de regressão linear utilizados para estimar o esforço de trabalho. Em

contrapartida, a imputação por média (MI) foi a técnica menos eficaz, demonstrando maior viés e menor precisão em todos os cenários avaliados.

A análise gráfica revelou que as curvas REC do MLR dominavam as de outros métodos, indicando consistentemente maior precisão para diferentes níveis de tolerância de erro. Além disso, a análise RROC destacou que os modelos baseados em MLR apresentaram menor propensão a superestimações e subestimações em relação aos outros métodos testados.

O artigo conclui que a combinação de métodos estatísticos e visuais oferece uma abordagem robusta para comparar técnicas de imputação, com destaque para a superioridade do MLR em cenários de dados faltantes. Futuras pesquisas são recomendadas para avaliar a eficácia dessas técnicas em outros contextos, como modelos de estimativa por analogia e dados longitudinais.

7. **Comparison of Imputation Methods for Categorical Real-World Prostate Cancer Data with Natural Order (Schmitt & Rothlauf, 2024):**

   [Schmitt and Rothlauf, 2024] investigam métodos de imputação para dados categóricos ordinais com valores ausentes em um conjunto de dados clínicos reais de pacientes com câncer de próstata. O estudo compara técnicas de imputação, incluindo imputação pela moda, *k-nearest neighbors* (KNN) e métodos baseados em MICE (*Multiple Imputation by Chained Equations*) usando modelos de regressão logística proporcional (*mice_polr*) e florestas aleatórias (*mice_rf*).

   O conjunto de dados utilizado foi coletado pelo Cancer Registry of Rhineland-Palatinate, na Alemanha, contendo 16.709 observações e seis variáveis categóricas ordinais (idade, tempo de sobrevivência, escore de Gleason, classificação TNM, níveis de PSA e graduação tumoral). Para fins experimentais, valores ausentes foram introduzidos artificialmente com proporções de 10%, 20%, 30% e 50% seguindo o mecanismo MCAR (*Missing Completely At Random*).

   Os resultados indicam que o método *mice_rf* teve o melhor desempenho geral, alcançando maior precisão em todas as proporções de valores ausentes. O *KNN* foi o segundo melhor método, enquanto o *mice_polr* apresentou o pior desempenho. A análise revelou que a precisão dos métodos de imputação aumenta quando o número de categorias é baixo, as proporções das categorias são relativamente uniformes, ou a maioria das observações se concentra em uma única categoria dominante. Em contrapartida, a precisão diminui significativamente quando há muitas categorias com distribuições desiguais, como observado nas variáveis "idade"e "escore de Gleason".

   Os autores também destacaram que o método MCAR, usado para gerar os dados ausentes, pode ser inadequado para dados clínicos com dependências estruturais entre variáveis. Nesse contexto, os mecanismos MAR (*Missing At Random*) ou MNAR (*Missing Not At Random*)

seriam mais representativos. O trabalho conclui que o *mice_rf* é uma abordagem robusta para dados categóricos ordinais, mas reconhece a necessidade de estudos futuros que explorem a influência de mecanismos MAR e MNAR.

8. **EKNN: Ensemble Classifier Incorporating Connectivity and Density into kNN with Application to Cancer Diagnosis (Mahfouz et al., 2020)**

   [Mahfouz et al., 2020] propõem um novo classificador de ensemble denominado EKNN, que combina o tradicional *k-Nearest Neighbors* (kNN) com quatro modelos de decisão baseados em conectividade e densidade. O objetivo do estudo é melhorar o desempenho do kNN em conjuntos de dados pequenos e desbalanceados, comuns em diagnósticos médicos baseados em microarrays de expressão gênica.

   O método EKNN utiliza uma *KNN-table* construída na fase de treinamento para armazenar os vizinhos mais próximos de cada amostra, reduzindo o tempo de classificação e permitindo ajustes dinâmicos. O ensemble é composto pelo kNN tradicional e quatro novos modelos: (1) KDN (*Density-based kNN*), (2) KLN-Avg (*Average Linkage*), (3) KLN-Min (*Single Linkage*) e (4) KLN-Max (*Complete Linkage*). Esses modelos exploram estatísticas da *KNN-table* para aumentar a robustez do classificador frente a ruídos e desbalanceamento de classes.

   Os experimentos foram conduzidos em cinco conjuntos de dados de câncer e sete bases de dados padrão, comparando o EKNN com métodos clássicos de classificação, como redes neurais, árvores de decisão, *Random Forest*, *Bagging* e *AdaBoost*. Os resultados mostraram que o EKNN superou o kNN tradicional e obteve melhor desempenho do que vários métodos de ensemble, especialmente em conjuntos de dados pequenos e desbalanceados.

   A principal contribuição do artigo é a introdução de um classificador de ensemble que melhora a precisão preditiva sem aumentar significativamente o tempo de classificação. Além disso, o estudo sugere que o EKNN pode ser adaptado para trabalhar com grandes volumes de dados ou fluxos contínuos. No entanto, o método tem limitações, como maior complexidade na fase de treinamento e a necessidade de ajustes para bases de dados muito grandes.

   O trabalho é relevante para pesquisadores que lidam com classificação em conjuntos de dados pequenos e desbalanceados, especialmente em aplicações médicas. O uso da *KNN-table* permite reduzir a influência de amostras ruidosas e melhorar a imputação de dados categóricos em cenários de aprendizado supervisionado.

Os trabalhos analisados oferecem uma visão abrangente das técnicas de imputação de dados categóricos, evidenciando avanços e limitações específicas em diferentes contextos de aplicação. A seguir, apresentamos uma comparação dos métodos utilizados.

III.2.1 Classificação dos Métodos de Imputação

Os métodos empregados nos estudos podem ser agrupados em três categorias principais, com características específicas:

Métodos Simples e Estatísticos: A imputação por moda, utilizada em [Memon et al., 2023], [Wu et al., 2012] e [Schmitt and Rothlauf, 2024], é computacionalmente eficiente, mas apresenta limitações significativas, como a super-representação de categorias dominantes. O método *Listwise Deletion* (LD), avaliado em [Chatzipetrou, 2019], é eficaz para preservar distribuições estatísticas, porém resulta em perda de dados, restringindo sua aplicabilidade em contextos com altas proporções de valores ausentes.

Métodos Paramétricos e Não Paramétricos: Técnicas como *K-Nearest Neighbors* (KNN), amplamente utilizadas ([Memon et al., 2023], [Suh and Song, 2023], [Wu et al., 2012], [Schmitt and Rothlauf, 2024]), mostraram-se eficazes em dados de menor dimensão ou com taxas moderadas de ausência. Por outro lado, a regressão logística multinomial ([Zimmermann et al., 2014], [Chatzipetrou, 2019]) destaca-se em contextos onde as categorias têm naturalidade ordinal, proporcionando maior precisão, mas com custo computacional elevado.

Métodos Avançados e Baseados em Machine Learning: Abordagens como *mixgb* (Suh and Song [2023]) e *mice_rf* (Schmitt and Rothlauf [2024]) demonstraram excelente desempenho em cenários com alta complexidade e maior proporção de valores ausentes. Esses métodos superaram as técnicas tradicionais em precisão, especialmente quando o conjunto de dados apresenta relações complexas entre variáveis. Contudo, os custos computacionais são consideravelmente maiores, limitando sua aplicação em bases de dados muito grandes ou com restrições de recursos.

III.2.2 Impactos nos Contextos de Aplicação

Dados Categóricos Ordinais: Trabalhos como [Memon et al., 2023] e [Schmitt and Rothlauf, 2024] destacaram a relevância de métodos robustos para bases de dados categóricos ordinais. Nesses cenários, *mice_rf* e regressão logística multinomial mostraram superioridade em manter a consistência dos dados e as distribuições originais. Além disso, o *EKNN*, com sua capacidade de lidar com dados pequenos e desbalanceados, oferece uma alternativa útil para situações em que os dados clínicos são limitados ou desbalanceados, sem comprometer o tempo de classificação.

Bases de Dados Mistos: Estudos como [Suh and Song, 2023] e [Kim et al., 2023] ampliaram o escopo para dados categóricos e numéricos, combinando técnicas multivariadas e aprendizado de máquina para lidar com complexidade e assimetrias nos dados. Essas abordagens fornecem uma base sólida para aplicações mais generalistas. O *mixgb* e o *mice_rf* foram destacados nesses contextos por sua flexibilidade e precisão, enquanto o *EKNN* pode ser uma opção interessante em cenários onde a distribuição dos dados apresenta alta variação e desbalanceamento.

Mineração de Dados e Análise Preditiva: Métodos como *MIBOS-DA*, proposto por [Wu et al., 2012], e *mixgb* ([Suh and Song, 2023]) destacam-se pela precisão em cenários de alta dimensionalidade e taxas elevadas de ausência de dados. Essas técnicas são especialmente úteis para análise preditiva, mas requerem alto custo computacional. O *EKNN*, por outro lado, oferece uma solução eficiente em termos de tempo de execução, especialmente em conjuntos de dados desbalanceados, equilibrando precisão e desempenho computacional para tarefas de mineração de dados.

III.2.3 Tabela de Comparação dos Métodos de Imputação

A Tabela III.1 apresenta um resumo comparativo dos métodos de imputação discutidos nos trabalhos relacionados, mapeando os métodos utilizados, seus contextos de aplicação, principais vantagens e limitações.

Tabela III.1: Caracterização dos trabalhos relacionados quanto às bases de dados, mecanismos de ausência, métodos de imputação e características estruturais.

Trabalho	**Tipo de Base**	**Mecanismo de Ausência**	**Métodos de Imputação**	**Características Estruturais**	**Nº de Atributos**	**Cardinalidade**
[Memon et al., 2023]	Clínica, categórica	MCAR	Mode, KNN, RF, SHD, MICE	5.036 registros maternos; variáveis binárias, ordinais e multinomiais; dados do Hospital Nacional de Referência de Kawempe, Uganda.	30	2–5 categorias
[Suh and Song, 2023]	Mista	MCAR	Mode, KNN, MICE, missForest, missRanger, mixgb	Seis bases: Milk, Nwtsco, Bike, Adult, Bank e WeatherAUS, com diferentes tamanhos e proporções de atributos categóricos e numéricos. 1.059 a 56.420 observações.	7–21	3–10 atributos categóricos
[Kim et al., 2023]	Categórica, simulada e real	MAR	Boruta + MICE; regressão logística, regressão politômica e predictive mean matching	Dados simulados por cópula de Clayton, com variáveis correlacionadas e distribuições não normais. 50 variáveis: 20 binárias, 15 com três categorias e 15 com quatro. Aplicação real com dados do NHANES III.	50 na simulação; 15 na aplicação	2–4 categorias na simulação; resposta binária na aplicação
[Zimmermann et al., 2014]	Demográfica, categórica	Não especificado / observacional	Regressão logística multinomial e distribuição multinomial	Dados demográficos da República Tcheca, de 1995 a 2011. Educação como variável de interesse, utilizando gênero, estado civil, idade e causa da morte como determinantes.	Não especificado	Categorias educacionais
[Wu et al., 2012]	Categórica	MCAR	MIBOS, MIBOS-DA, KNN, Moda	Três bases UCI: Zoo, SPECT Heart e Voting. Ausência artificial de 5% a 70%. O MIBOS utiliza similaridade entre objetos e não requer definição de K.	Variável entre as bases	Variável entre as bases

Continua na próxima página

Trabalho	**Tipo de Base**	**Mecanismo de Ausência**	**Métodos de Imputação**	**Características Estruturais**	**Nº de Atributos**	**Cardinalidade**
[Chatzipetrou, 2019]	Mista, software / estimativa de custos	Não especificado	MLR, LD, MI, EM, RI	152 projetos do repositório ISBSG, com variáveis como plataforma, linguagem de programação e banco de dados. Avaliação por erro, viés, dispersão e curvas REC e RROC.	Não especificado	Variável conforme os atributos
[Schmitt and Rothlauf, 2024]	Categórica ordinal, clínica	MCAR	Moda, KNN, mice_polr, mice_rf	Base real de câncer de próstata do Cancer Registry of Rhineland-Palatinate, Alemanha. Dados predominantemente categóricos ordinais, relacionados à classificação de risco e ao intervalo entre diagnóstico e óbito.	Não especificado	Variável conforme os atributos ordinais
[Mahfouz et al., 2020]	Microarray, alta dimensionalidade	Não aplicável	EKNN: KNN, KDN, KLN-Avg, KLN-Min, KLN-Max	Classificação de câncer em dados de expressão gênica. Alta dimensionalidade, muitos genes irrelevantes, poucas amostras e desbalanceamento entre classes. Referência metodológica para KNN, não estudo específico de imputação.	Alta dimensionalidade	Variável conforme o problema

A comparação realizada evidencia que a escolha do método de imputação depende fortemente do contexto de aplicação, das características dos dados e dos objetivos analíticos. Enquanto métodos simples, como imputação por moda, ainda são amplamente utilizados pela sua simplicidade, técnicas avançadas baseadas em aprendizado de máquina, como *mixgb* e *mice_rf*, representam o estado da arte, proporcionando maior precisão em cenários complexos.

Entretanto, desafios permanecem, especialmente em cenários onde os mecanismos MAR e MNAR prevalecem. Além disso, a eficiência computacional de métodos avançados continua sendo uma barreira para sua adoção em larga escala. Essas lacunas reforçam a necessidade de desenvolver soluções que conciliem robustez e eficiência para lidar com dados categóricos ausentes.

A análise dos trabalhos relacionados evidencia avanços importantes na aplicação de métodos estatísticos e de aprendizado de máquina para imputação de dados faltantes. No entanto, observa-se que grande parte das pesquisas concentra-se em bases numéricas ou mistas, havendo relativamente menos estudos direcionados especificamente à imputação de variáveis categóricas. Além disso, poucos trabalhos investigam de forma sistemática como características estruturais dos dados, como cardinalidade das categorias, número de atributos e intensidade das relações entre variáveis, impactam a dificuldade de imputação e a preservação das propriedades estatísticas originais. Também são limitadas as análises comparativas envolvendo diferentes métodos em bases heterogêneas sob múltiplos níveis de ausência. Nesse contexto, esta dissertação busca contribuir para o estado da arte ao analisar experimentalmente diferentes abordagens de imputação categórica, considerando não apenas a acurácia da imputação, mas também os impactos estruturais e preditivos decorrentes do processo de preenchimento dos dados ausentes.

Capítulo IV Metodologia

Este capítulo apresenta o delineamento experimental utilizado para avaliar métodos de imputação de dados categóricos sob diferentes condições de estrutura dos dados. O estudo foi desenvolvido com o objetivo de analisar como características dos conjuntos de dados, especialmente o tipo dos atributos preditores, a quantidade de categorias da variável-alvo e a proporção de valores ausentes, podem estar associadas ao desempenho dos métodos de imputação.

A avaliação foi conduzida por meio da introdução controlada de valores ausentes segundo o mecanismo MCAR (*Missing Completely At Random*) exclusivamente na variável categórica utilizada como alvo da imputação. Foram consideradas três proporções de ausência e diferentes *seeds*, permitindo repetir o experimento sob diferentes partições dos dados e diferentes posições dos valores removidos.

Foram avaliados três métodos de imputação baseados em paradigmas distintos: K-Nearest Neighbors (KNN), Random Forest (RF) e Multilayer Perceptron (MLP). A escolha desses métodos não tem como finalidade estabelecer uma hierarquia entre eles, mas contemplar diferentes formas de aprendizagem. O KNN utiliza a proximidade entre instâncias, o Random Forest emprega um conjunto de árvores de decisão e o MLP utiliza uma rede neural artificial. Essa diversidade permite analisar o comportamento da imputação diante de diferentes estruturas de aprendizagem.

A avaliação considerou duas perspectivas complementares. A primeira corresponde à taxa de acerto da imputação, obtida pela comparação entre os valores originalmente observados e os valores recuperados pelo método. A segunda corresponde à acurácia do classificador, utilizada para verificar o impacto das imputações sobre uma tarefa preditiva posterior.

IV.1 Delineamento Experimental

O experimento foi estruturado em etapas sequenciais. Inicialmente, os conjuntos de dados foram submetidos ao pré-processamento e particionamento em conjuntos de treinamento e teste. Em seguida, somente os valores da variável-alvo pertencentes ao conjunto de treinamento foram artificialmente removidos. Os métodos de imputação foram então aplicados ao conjunto de treinamento, utilizando os valores ainda conhecidos como referência para aprendizagem.

Após a imputação, o conjunto de treinamento completo, contendo os valores originalmente

observados e os valores imputados, foi utilizado em uma tarefa de classificação. O conjunto de teste permaneceu inalterado durante todo o processo e foi utilizado para avaliar o desempenho preditivo.

A Figura IV.1 apresenta, de forma geral, o fluxo experimental adotado.

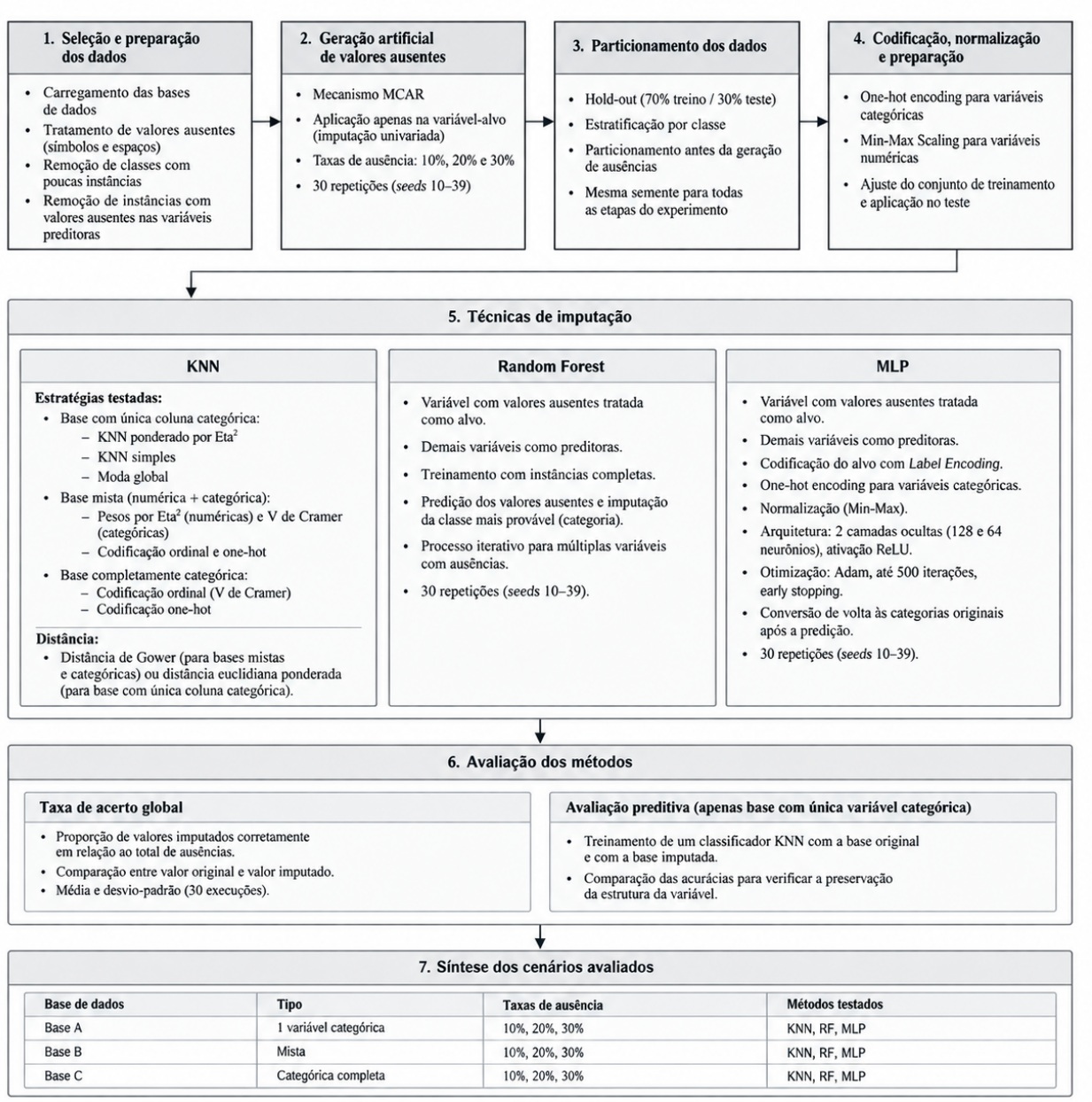


Figura IV.1: Fluxo geral do procedimento experimental.

O protocolo experimental pode ser resumido nas seguintes etapas:

1. leitura e preparação dos conjuntos de dados;

2. identificação da variável-alvo;

3. remoção de instâncias inadequadas ao experimento;

4. particionamento estratificado em treinamento e teste;

5. codificação e normalização dos atributos preditores;
6. introdução artificial de valores ausentes na variável-alvo do conjunto de treinamento;
7. aplicação dos métodos de imputação;
8. cálculo da taxa de acerto da imputação;
9. treinamento do classificador com os valores imputados;
10. avaliação do classificador no conjunto de teste;
11. repetição do procedimento para diferentes proporções de ausência e *seeds*.

IV.2 Notação Utilizada

Para facilitar a descrição dos algoritmos e dos pseudocódigos, considera-se X a matriz de atributos preditores e y a variável-alvo categórica que contém os valores a serem imputados.

Cada instância é representada por x_j, enquanto A_i representa o i-ésimo atributo. O valor do atributo A_i na instância x_j é representado por x_{ji}. O conjunto de índices das instâncias cujo valor da variável-alvo permanece conhecido é representado por I_{obs}, enquanto I_{miss} representa o conjunto de índices nos quais o valor da variável-alvo foi artificialmente removido.

A variável k representa o número de vizinhos considerados pelo KNN. O símbolo $\hat{y}_j$ representa o valor estimado pelo método de imputação para a instância j.

Essa notação é utilizada nos procedimentos apresentados ao longo deste capítulo, permitindo distinguir os valores originalmente observados dos valores artificialmente ocultados e posteriormente estimados.

IV.3 Conjuntos de Dados

Foram utilizados conjuntos de dados com diferentes características estruturais, permitindo analisar o comportamento dos métodos em contextos distintos. A seleção buscou contemplar diferentes combinações entre atributos numéricos e categóricos e diferentes quantidades de categorias na variável-alvo.

Para fins experimentais, os conjuntos de dados foram organizados em cenários de acordo com a natureza dos atributos preditores e da variável-alvo. Nos conjuntos em que existem atributos categóricos, esses atributos são tratados como categóricos no processo de codificação, sem imposição de uma ordenação artificial entre suas categorias.

A variável-alvo é sempre uma variável categórica no procedimento de imputação. Mesmo nos casos em que a variável possui representação numérica no conjunto de dados original, ela é consi-

derada como uma variável categórica quando suas diferentes classes são utilizadas como categorias a serem recuperadas.

Os cenários avaliados são apresentados na Tabela IV.1.

Tabela IV.1: Síntese dos cenários avaliados.

Cenário	**Características dos dados**	**Objetivo da avaliação**
Uma variável categórica alvo	Variável-alvo categórica e atributos preditores numéricos	Avaliar a imputação quando a informação auxiliar é predominantemente numérica.
Dados mistos	Variável-alvo categórica, com atributos preditores numéricos e categóricos	Avaliar a imputação na presença simultânea de diferentes tipos de atributos.
Dados categóricos	Variável-alvo categórica e atributos preditores categóricos	Avaliar o comportamento dos métodos quando a informação auxiliar é predominantemente categórica.

Além do tipo dos atributos, a quantidade de categorias da variável-alvo constitui uma característica importante para a análise. O número de categorias influencia a dificuldade potencial da tarefa de imputação, uma vez que aumenta a quantidade de possibilidades que o método deve distinguir.

A tabela detalhada dos conjuntos de dados, contendo quantidade de instâncias, atributos e categorias da variável-alvo, é apresentada juntamente aos resultados experimentais.

IV.4 Preparação dos Dados

Os conjuntos de dados foram carregados a partir dos arquivos utilizados nos experimentos. Durante a leitura, os símbolos utilizados para representar valores ausentes foram convertidos para valores nulos. Para atributos textuais, foram removidos espaços excedentes no início e no final dos valores.

Quando necessário, foram removidas as instâncias pertencentes a classes com quantidade insuficiente de observações para a realização da divisão estratificada. Especificamente, foram desconsideradas classes que apresentavam apenas uma instância, uma vez que não seria possível garantir sua representação adequada nos conjuntos de treinamento e teste.

Também foram removidas instâncias que apresentavam valores ausentes nos atributos preditores. Essa decisão está relacionada à implementação utilizada nos experimentos, na qual os métodos recebem uma matriz de atributos preditores completa. A ausência artificial utilizada para avaliar a imputação é introduzida exclusivamente na variável-alvo.

Assim, o problema experimental analisado consiste na existência de valores ausentes na variável categórica alvo, enquanto os atributos utilizados como informação auxiliar permanecem observados.

IV.5 Particionamento dos Dados

Após o pré-processamento, os dados foram divididos em conjuntos de treinamento e teste utilizando a estratégia *hold-out*. Foram destinadas 70% das instâncias ao treinamento e 30% ao teste.

A divisão foi realizada de maneira estratificada em relação à variável-alvo, buscando manter, tanto quanto possível, a distribuição das classes nos dois subconjuntos. O particionamento foi realizado antes da introdução dos valores ausentes.

O conjunto de teste não sofre remoção artificial de valores. Dessa forma, seus valores verdadeiros permanecem conhecidos durante toda a execução do experimento e são utilizados somente na avaliação da tarefa preditiva posterior.

O particionamento utiliza a mesma *seed* empregada em cada repetição experimental. Portanto, a alteração da *seed* pode modificar tanto a composição dos conjuntos de treinamento e teste quanto as instâncias selecionadas posteriormente para receber valores ausentes. Para os métodos Random Forest e MLP, a *seed* também é utilizada para controlar a aleatoriedade interna dos respectivos modelos.

Não foi utilizada validação cruzada no particionamento externo. No caso específico do MLP, existe uma divisão interna para validação decorrente do mecanismo de *early stopping*, descrito posteriormente.

IV.6 Codificação e Normalização dos Atributos

Os atributos categóricos utilizados como preditores foram transformados por meio da técnica *one-hot encoding*. Nesse procedimento, cada categoria de um atributo categórico é representada por uma variável binária, permitindo utilizar essas informações nos métodos que operam sobre representações numéricas.

O codificador é ajustado utilizando somente os dados de treinamento. Posteriormente, o mesmo codificador é aplicado ao conjunto de teste. Essa estratégia evita que as informações presentes no conjunto de teste participem do processo de ajuste da representação dos dados.

Os atributos numéricos e os atributos resultantes da codificação foram posteriormente normalizados por meio do *Min-Max Scaling*. O ajuste do normalizador também é realizado exclusivamente sobre o conjunto de treinamento, sendo posteriormente aplicado ao conjunto de teste.

Não foi utilizada codificação ordinal para os atributos categóricos. Essa decisão é particularmente importante porque uma codificação ordinal atribuiria valores numéricos às categorias e poderia introduzir uma relação de ordem que não necessariamente existe entre elas. Por exemplo, representar três categorias nominais como 0, 1 e 2 poderia fazer com que uma distância numérica interpretasse a categoria 0 como mais próxima da categoria 1 do que da categoria 2, ainda que não

exista tal relação no domínio.

Assim, para as técnicas que utilizam a representação codificada dos dados, as variáveis categóricas são tratadas, para fins de representação, sem a imposição de uma ordenação artificial. Essa escolha também mantém uma representação consistente entre os métodos KNN, Random Forest e MLP.

No cálculo da distância de Gower, entretanto, os atributos categóricos são convertidos internamente em códigos numéricos apenas para possibilitar a comparação de igualdade entre categorias. Esses códigos não são utilizados como valores ordinais e, portanto, não introduzem uma distância baseada na magnitude do código.

IV.7 Geração Artificial de Valores Ausentes

A ausência de dados foi introduzida artificialmente segundo o mecanismo MCAR (*Missing Completely At Random*). Nesse mecanismo, a ocorrência da ausência não depende dos valores observados ou não observados da variável analisada [Little and Rubin, 1987].

A introdução artificial dos valores ausentes foi realizada exclusivamente na variável-alvo do conjunto de treinamento. Dessa forma, os atributos preditores permanecem disponíveis para que os métodos possam utilizá-los como informação auxiliar na imputação.

Foram avaliadas três proporções de ausência: 10%, 20% e 30% das instâncias do conjunto de treinamento. A quantidade efetiva de instâncias removidas é obtida pela conversão da proporção em um número inteiro de observações.

Para cada combinação entre conjunto de dados e proporção de ausência, foram realizadas 30 repetições experimentais, utilizando *seeds* de 10 a 39. Em cada repetição, uma nova divisão entre treinamento e teste e uma nova seleção das instâncias que recebem valores ausentes são determinadas pela *seed* correspondente.

Esse procedimento permite reduzir a dependência dos resultados em relação a uma única partição dos dados ou a uma única configuração de valores ausentes. Os resultados finais são apresentados pela média e pelo desvio-padrão obtidos nas 30 repetições.

IV.8 Métodos de Imputação

Foram avaliados três métodos de imputação: KNN, Random Forest e MLP. A escolha foi realizada buscando contemplar diferentes paradigmas de aprendizagem, e não estabelecer uma comparação baseada em uma única família de algoritmos.

O KNN é um método baseado em instâncias e utiliza a proximidade entre observações para estimar o valor ausente. O Random Forest utiliza um conjunto de árvores de decisão para realizar a classificação da variável-alvo. O MLP utiliza uma rede neural artificial capaz de aprender relações

não lineares entre os atributos preditores e a variável-alvo [Breiman, 2001] [Haykin, 2009] [Goodfellow et al., 2016].

Todos os métodos recebem a mesma divisão entre treinamento e teste e são avaliados sob as mesmas proporções de ausência e *seeds*. Dessa forma, as diferenças observadas podem ser analisadas considerando o mesmo protocolo experimental.

IV.8.1 Imputação com KNN

O KNN estima o valor ausente a partir das instâncias conhecidas mais próximas. No presente estudo, foi utilizado $k = 5$, de modo que cinco vizinhos são selecionados para cada instância com valor ausente.

Para cada instância pertencente a I_{miss}, são calculadas as distâncias em relação às instâncias pertencentes a I_{obs}. Após a identificação dos cinco vizinhos mais próximos, o valor imputado corresponde à moda dos valores da variável-alvo desses vizinhos.

Foram avaliadas três configurações do KNN: distância Euclidiana simples, distância ponderada por Eta² e distância de Gower.

IV.8.1.1 KNN com Distância Euclidiana

Na primeira configuração, os atributos preditores são representados após a aplicação do *one-hot encoding* e da normalização Min-Max. A distância entre duas instâncias é calculada pela distância Euclidiana:

$$d(x_j, x_l) = \sqrt{\sum_{i=1}^{p} (x_{ji} - x_{li})^2}, \tag{IV.1}$$

em que p representa o número de atributos após a codificação, e x_{ji} e x_{li} representam os valores do atributo i nas instâncias j e l, respectivamente.

Após o cálculo das distâncias, são selecionadas as cinco instâncias mais próximas. A categoria mais frequente entre esses vizinhos é utilizada como valor imputado.

IV.8.1.2 KNN Ponderado por Eta²

A segunda configuração utiliza o coeficiente Eta² para ponderar os atributos de acordo com sua associação estatística com a variável-alvo.

Para cada atributo A_i, o Eta² é calculado a partir da razão entre a variabilidade entre os grupos definidos pelas categorias da variável-alvo e a variabilidade total do atributo:

$$\eta_i^2 = \frac{SS_{between}}{SS_{total}}, \tag{IV.2}$$

em que $SS_{between}$ representa a variabilidade entre os grupos e SS_{total} representa a variabilidade total do atributo.

O termo $SS_{between}$ é calculado por:

$$SS_{between} = \sum_{c=1}^{C} n_c(\bar{A}_{ic} - \bar{A}_i)^2, \tag{IV.3}$$

em que C é o número de categorias da variável-alvo, n_c é a quantidade de observações pertencentes à categoria c, $\bar{A}_{ic}$ é a média do atributo A_i dentro da categoria c e $\bar{A}_i$ é a média geral do atributo.

Já a variabilidade total é dada por:

$$SS_{total} = \sum_{j=1}^{n}(A_{ij} - \bar{A}_i)^2. \tag{IV.4}$$

Na implementação, o Eta² é calculado para cada coluna da representação utilizada pelo KNN. Portanto, quando um atributo categórico é transformado em múltiplas colunas por *one-hot encoding*, cada coluna binária recebe seu próprio valor de Eta².

Após o cálculo dos pesos, cada atributo da matriz normalizada é multiplicado pelo respectivo valor de Eta²:

$$x' * ji = \eta_i^2 x * ji. \tag{IV.5}$$

A distância Euclidiana é então calculada sobre essa representação ponderada:

$$d_\eta(x_j, x_l) = \sqrt{\sum_{i=1}^{p}(x' * ji - x' * li)^2}. \tag{IV.6}$$

Dessa forma, os atributos com valores maiores de Eta² possuem maior influência sobre a representação utilizada no cálculo das distâncias, enquanto atributos com valores próximos de zero possuem menor influência.

Após a obtenção dos cinco vizinhos mais próximos, a categoria mais frequente entre eles é utilizada para realizar a imputação.

IV.8.1.3 KNN com Distância de Gower

A terceira configuração utiliza a distância de Gower, proposta por Gower [Gower, 1971], especialmente adequada para situações que envolvem simultaneamente atributos numéricos e categóricos.

Diferentemente das configurações anteriores, a distância de Gower é calculada sobre os atributos em sua representação original, sem aplicação de *one-hot encoding*. Para atributos numéricos, a

contribuição da distância é normalizada pelo intervalo observado do atributo:

$$d_{ijl}^{num} = \frac{|x_{ji} - x_{li}|}{\max(A_i) - \min(A_i)}. \tag{IV.7}$$

Para atributos categóricos, a contribuição é definida pela igualdade ou diferença entre as categorias:

$$d_{ijl}^{cat} = \begin{cases} 0, & \text{se } x_{ji} = x_{li}, \\ 1, & \text{caso contrário.} \end{cases} \tag{IV.8}$$

A distância final corresponde à média das contribuições dos atributos:

$$d_G(x_j, x_l) = \frac{1}{p} \sum_{i=1}^{p} d_{ijl}. \tag{IV.9}$$

Após o cálculo das distâncias, são selecionados os cinco vizinhos mais próximos e a moda de suas categorias é utilizada para preencher o valor ausente.

No tratamento dos atributos categóricos, os códigos numéricos utilizados na implementação têm apenas a função de permitir a comparação de igualdade entre categorias. Portanto, a distância não considera, por exemplo, que o código 2 esteja mais distante do código 1 do que do código 0. O que importa é somente se as categorias são iguais ou diferentes.

IV.8.2 Imputação com Random Forest

O Random Forest é um método de aprendizagem baseado na combinação de múltiplas árvores de decisão [Breiman, 2001]. No presente trabalho, o algoritmo é utilizado como um classificador supervisionado para estimar diretamente os valores ausentes da variável-alvo.

Para cada execução, as instâncias com valores conhecidos da variável-alvo são utilizadas para treinar o modelo. Em seguida, as instâncias cujo valor foi artificialmente removido são submetidas ao modelo treinado, que produz uma previsão para cada uma delas.

Formalmente, considerando I_{obs} como o conjunto de instâncias observadas e I_{miss} como o conjunto de instâncias com valores ausentes, o modelo é ajustado por:

$$RF(X_{I_{obs}}, y_{I_{obs}}), \tag{IV.10}$$

e posteriormente utilizado para estimar:

$$\hat{y} * I * miss = RF(X_{I_{miss}}). \tag{IV.11}$$

Foi utilizado um conjunto de 200 árvores de decisão, configurado por meio do parâmetro

`n_estimators=200`. A aleatoriedade do modelo é controlada pela *seed* correspondente à repetição experimental.

A imputação ocorre, portanto, em uma única etapa: o Random Forest é treinado uma vez com os valores conhecidos e utilizado uma vez para prever os valores ausentes.

É importante destacar que essa implementação não corresponde ao algoritmo missForest. No missForest, o procedimento de imputação é iterativo e envolve sucessivas atualizações dos valores imputados até que um critério de convergência seja atingido. No presente trabalho, não há iterações sucessivas, atualização de valores previamente imputados ou cálculo de convergência. Consequentemente, não é utilizado um critério de convergência no procedimento baseado em Random Forest.

IV.8.3 Imputação com MLP

O Multilayer Perceptron (MLP) é uma rede neural artificial composta por camadas de neurônios capazes de modelar relações não lineares entre as variáveis de entrada e a variável-alvo [Haykin, 2009] [Goodfellow et al., 2016].

Assim como no Random Forest, o MLP é utilizado como um classificador supervisionado para estimar os valores ausentes da variável-alvo.

Os atributos preditores são previamente transformados por *one-hot encoding* quando categóricos e posteriormente normalizados pelo método Min-Max. O modelo é então treinado exclusivamente com as instâncias cujo valor da variável-alvo permanece conhecido.

IV.8.3.1 Codificação da Variável-Alvo

Diferentemente dos atributos preditores, a variável-alvo não é transformada por *one-hot encoding*. Para o treinamento do MLP, as categorias da variável-alvo são convertidas em códigos inteiros por meio do *Label Encoding*.

Essa transformação não tem a finalidade de estabelecer uma ordem entre as categorias. Os códigos são utilizados apenas como identificadores das classes que serão previstas pelo classificador.

O codificador é ajustado exclusivamente sobre os valores conhecidos da variável-alvo. Após o treinamento, as previsões produzidas pelo MLP são convertidas novamente para as categorias originais por meio da operação inversa do *Label Encoding*. Dessa forma, os valores utilizados para preencher as posições ausentes mantêm a mesma representação categórica utilizada originalmente no conjunto de dados.

IV.8.3.2 Arquitetura e Treinamento

A arquitetura utilizada possui duas camadas ocultas, com 128 e 64 neurônios, respectivamente. A função de ativação utilizada nas camadas ocultas é a ReLU (*Rectified Linear Unit*), definida por:

$$ReLU(z) = \max(0, z). \tag{IV.12}$$

A utilização da ReLU permite introduzir não linearidade no modelo e manter valores negativos zerados na saída da função de ativação.

O treinamento utiliza o algoritmo Adam (*Adaptive Moment Estimation*) [Kingma and Ba, 2015]. O número máximo de iterações foi definido como 500.

Também foi utilizado o mecanismo de *early stopping*, com 'n_iter_no_change=20'. Dessa forma, o treinamento pode ser interrompido antes de atingir o limite de 500 iterações quando o desempenho no conjunto interno de validação deixa de apresentar melhora suficiente durante 20 iterações consecutivas.

Como o parâmetro *early stopping* foi ativado sem especificação explícita da fração de validação, utiliza-se a configuração padrão do algoritmo, correspondente a 10% das instâncias fornecidas ao treinamento do MLP. Essa parcela é utilizada internamente para monitorar a evolução do modelo e decidir quando interromper o treinamento. Ela não corresponde ao conjunto de teste utilizado na avaliação final.

A camada de saída é definida automaticamente de acordo com o número de classes da variável-alvo. Em problemas multiclasse, existe uma unidade de saída associada a cada classe. Para problemas binários, o `MLPClassifier` utiliza uma unidade de saída. Assim, a quantidade de unidades de saída está diretamente relacionada à quantidade de categorias que o modelo precisa distinguir.

Após o treinamento, o MLP realiza a previsão somente sobre as instâncias pertencentes a I_{miss}. As classes previstas são então convertidas dos códigos numéricos para suas categorias originais e utilizadas para preencher as posições ausentes.

IV.9 Avaliação dos Métodos de Imputação

A avaliação dos métodos foi realizada sob duas perspectivas: a taxa de acerto da imputação e a acurácia do classificador.

IV.9.1 Taxa de Acerto da Imputação

A taxa de acerto da imputação mede diretamente a capacidade do método de recuperar os valores que foram artificialmente removidos. Como o experimento utiliza valores originalmente conhecidos, é possível comparar cada valor imputado com seu respectivo valor verdadeiro.

Seja I_{miss} o conjunto de instâncias nas quais o valor da variável-alvo foi removido. A taxa de acerto da imputação é calculada por:

$$TAI = \frac{1}{|I_{miss}|} \sum_{j \in I_{miss}} \mathbb{I}(y_j = \hat{y}_j) \times 100, \tag{IV.13}$$

em que y_j representa o valor original da variável-alvo, $\hat{y}_j$ representa o valor imputado e $\mathbb{I}(\cdot)$ é uma função indicadora que assume valor 1 quando os valores são iguais e 0 caso contrário.

A utilização da expressão "taxa de acerto da imputação" permite diferenciar essa medida da acurácia utilizada para avaliar o classificador em uma tarefa preditiva posterior.

Para cada combinação de conjunto de dados e proporção de ausência, a taxa de acerto da imputação é calculada nas 30 repetições experimentais. Os resultados são posteriormente apresentados pela média e pelo desvio-padrão.

IV.9.2 Acurácia do Classificador

Além da comparação direta entre valores originais e imputados, foi analisado o efeito da imputação sobre uma tarefa de classificação posterior.

Para essa avaliação, foi utilizado um classificador KNN com $k = 5$. Inicialmente, o classificador é treinado utilizando o conjunto de treinamento completo, isto é, antes da introdução dos valores ausentes. Esse resultado constitui a referência correspondente aos dados sem perda da variável-alvo.

Em seguida, para cada método de imputação, o mesmo classificador KNN é treinado utilizando o conjunto de treinamento após a imputação. Em ambos os casos, a avaliação é realizada sobre o mesmo conjunto de teste, que permanece completo e inalterado.

A acurácia do classificador é calculada por:

$$ACC = \frac{1}{|I_{test}|} \sum_{j \in I_{test}} \mathbb{I}(y_j = \hat{y}_j) \times 100. \tag{IV.14}$$

Essa avaliação não mede diretamente a preservação de associações entre atributos. Seu objetivo é verificar o impacto das diferentes estratégias de imputação sobre o desempenho preditivo posterior. Assim, a comparação entre a acurácia obtida com os dados completos e aquela obtida após a imputação fornece uma medida complementar à taxa de acerto da imputação.

IV.9.3 Limitações da Métrica de Acurácia

Embora a taxa de acerto da imputação seja adequada para a comparação direta entre os valores verdadeiros e imputados em um experimento controlado, sua interpretação deve considerar a distribuição das classes.

Em conjuntos de dados desbalanceados, uma estratégia que favoreça as classes majoritárias pode apresentar uma taxa de acerto elevada mesmo apresentando desempenho limitado na recuperação de classes menos frequentes. O mesmo problema pode ocorrer na avaliação da acurácia do classificador.

Por esse motivo, os resultados não devem ser interpretados exclusivamente a partir de um único valor agregado. A análise considera conjuntamente as características dos conjuntos de dados, a quantidade de categorias da variável-alvo, as proporções de ausência avaliadas e a variabilidade observada entre as diferentes repetições.

A ausência de métricas específicas para avaliar o desempenho por classe, como medidas balanceadas ou métricas de desempenho individual por categoria, constitui uma limitação do protocolo experimental. Dessa forma, os resultados obtidos por meio da taxa de acerto e da acurácia devem ser interpretados dentro das características de distribuição de cada conjunto de dados.

IV.10 Repetições Experimentais e Agregação dos Resultados

O experimento foi repetido utilizando 30 *seeds*, numeradas de 10 a 39. Para cada *seed*, o conjunto de dados é particionado novamente, a seleção das instâncias que recebem valores ausentes é realizada novamente e, nos métodos que possuem componentes aleatórios, a mesma *seed* controla a aleatoriedade do modelo.

Para cada conjunto de dados foram avaliadas as proporções de ausência de 10

Ao final, são calculadas a média e o desvio-padrão das métricas obtidas nas repetições. A média representa o desempenho médio observado no conjunto de execuções, enquanto o desvio-padrão permite verificar a variabilidade dos resultados em função das diferentes partições e configurações de ausência.

IV.11 Síntese do Protocolo Experimental

A Tabela IV.2 apresenta uma síntese das principais configurações utilizadas no experimento.

Tabela IV.2: Síntese do protocolo experimental.

Elemento	**Configuração**
Variável submetida à ausência	Variável-alvo categórica
Mecanismo de ausência	MCAR (*Missing Completely At Random*)
Proporções de ausência	10%, 20% e 30% do conjunto de treinamento
Número de repetições	30 *seeds*, de 10 a 39
Particionamento	*hold-out*, 70% treinamento e 30% teste
Estratificação	Sim, em relação à variável-alvo
Codificação categórica	*one-hot encoding* para os atributos preditores
Normalização	*Min-Max Scaling*
KNN	$k = 5$, com distância Euclidiana, Eta² e Gower
Random Forest	200 árvores
MLP	Camadas ocultas com 128 e 64 neurônios, ReLU e Adam
MLP – treinamento	Máximo de 500 iterações e *early stopping*
MLP – tolerância de parada	20 iterações sem melhora suficiente
Avaliação da imputação	Taxa de acerto da imputação
Avaliação posterior	Acurácia de classificador KNN com $k = 5$
Agregação	Média e desvio-padrão das 30 repetições

O protocolo apresentado permite comparar os métodos sob condições experimentais controladas, mantendo constantes a forma de particionamento, as proporções de ausência, o conjunto de teste e o procedimento de avaliação. A variação das *seeds*, dos conjuntos de dados e das proporções de ausência permite observar o comportamento dos métodos em diferentes condições sem atribuir os resultados a uma única realização do experimento.

Por fim, ressalta-se que a avaliação por taxa de acerto da imputação e por acurácia do classificador representa duas perspectivas distintas. A primeira verifica diretamente se o valor ausente foi recuperado corretamente, enquanto a segunda avalia o efeito das imputações sobre uma tarefa preditiva posterior. A utilização conjunta dessas medidas permite uma análise mais abrangente do comportamento dos métodos avaliados.

Capítulo V Resultados

V.1 Bases de Dados Utilizadas

A condução dos experimentos desta dissertação foi estruturada de forma a contemplar diferentes níveis de complexidade no problema de imputação categórica. Para isso, foram selecionadas nove bases de dados amplamente utilizadas na literatura, majoritariamente provenientes do UCI Machine Learning Repository, permitindo a análise sistemática do comportamento dos métodos sob diferentes configurações estruturais.

A escolha das bases não foi realizada de forma arbitrária, mas orientada por três dimensões principais: (i) o número de variáveis categóricas presentes no conjunto de dados, (ii) a cardinalidade dessas variáveis e (iii) o grau de associação entre os atributos. Essas dimensões refletem diretamente a dificuldade do problema de imputação categórica, conforme discutido no Capítulo II.

De modo geral, as bases foram organizadas em três grandes grupos conceituais. O primeiro grupo é composto por bases com estrutura simplificada, nas quais há uma única variável categórica a ser imputada, sendo as demais variáveis predominantemente numéricas. Nesse cenário, a imputação depende essencialmente da capacidade dos métodos em explorar relações entre atributos contínuos e a variável categórica, caracterizando um problema de menor complexidade estrutural.

O segundo grupo contempla bases com múltiplas variáveis categóricas, nas quais passam a existir relações entre atributos qualitativos. Nesses casos, o problema de imputação torna-se mais complexo, pois os métodos precisam capturar dependências entre diferentes variáveis categóricas, além de lidar com possíveis interações com atributos numéricos.

Por fim, o terceiro grupo inclui bases compostas majoritariamente ou integralmente por variáveis categóricas. Nesse cenário, a imputação passa a depender quase exclusivamente das relações entre atributos qualitativos, tornando-se particularmente sensível à cardinalidade das variáveis e ao grau de associação entre elas. Esse tipo de estrutura representa o caso mais desafiador considerado neste trabalho.

Além da estrutura das variáveis, outro fator relevante considerado na seleção das bases foi a diversidade em termos de cardinalidade. Foram incluídas bases com variáveis categóricas de baixa, média e alta cardinalidade, permitindo analisar como o aumento do número de categorias impacta o desempenho dos métodos de imputação. Conforme observado nos resultados experimentais, a

cardinalidade mostrou-se um dos fatores mais determinantes na dificuldade do problema.

Outro aspecto importante refere-se ao tamanho das bases e à distribuição das classes. Conjuntos de dados menores e com classes desbalanceadas tendem a apresentar maior sensibilidade ao processo de geração artificial de valores ausentes, especialmente sob o mecanismo MCAR. Nesses casos, a remoção aleatória de instâncias pode comprometer significativamente a representatividade de determinadas categorias, influenciando diretamente o desempenho dos métodos de imputação.

A utilização de múltiplas bases com características distintas permitiu avaliar os métodos de forma abrangente, evitando conclusões restritas a cenários específicos. Dessa forma, os resultados apresentados nas seções subsequentes refletem não apenas o comportamento dos algoritmos em uma base isolada, mas sim padrões consistentes observados ao longo de diferentes contextos estruturais.

Essa abordagem possibilita uma análise mais robusta do problema, evidenciando que o desempenho da imputação categórica é fortemente dependente das propriedades dos dados, e não apenas do método empregado.

Tabela V.1: Caracterização das bases de dados utilizadas nos experimentos.

Cenário	**Base**	**Instâncias**	**Card.**	**Entropia**	**V de Cramér**	**Eta^2**
Simples	Wine	178	3	0.9886	–	0.4382
Simples	Image Segmentation	2310	7	1.0000	–	0.5362
Simples	Pen-Based	10992	10	0.9997	–	0.5058
Mista	Student	649	17	0.8540	0.1039	0.1799
Mista	Obesity	2111	7	0.9983	0.3159	0.1946
Mista	Bank	45211	12	0.8523	0.1710	0.0391
Categórica	Balance	625	3	0.8316	0.2852	–
Categórica	Nursery	12960	5	0.7393	0.2002	–
Categórica	Car	1728	4	0.6029	0.1961	–

A Tabela V.1 caracteriza as nove bases de dados utilizadas nos experimentos, organizadas em três cenários Simples (atributos exclusivamente numéricos), Mista (atributos numéricos e categóricos) e Categórica (atributos exclusivamente categóricos) e descritas por seu número de instâncias, cardinalidade do problema (número de classes, k) e três medidas complementares. A Entropia normalizada ($H_{norm} = -\sum_i p_i \log_2 p_i \,/\, \log_2 k$, variando de 0 a 1) quantifica o grau de balanceamento da distribuição das classes: valores próximos de 1 indicam classes com frequências semelhantes (como Image Segmentation, com $H_{norm} = 1{,}0000$, e Pen-Based e Obesity, ambas acima de 0,99), enquanto valores mais baixos denunciam desbalanceamento acentuado caso de Car ($H_{norm} = 0{,}6029$), cuja classe majoritária (*unacc*) domina mais de 70% das instâncias. As duas últimas colunas medem, de forma resumida, a força da relação entre os atributos preditores e a classe: o V de Cramér (média sobre todas as variáveis categóricas de cada base) captura essa associação quando os preditores são nominais, enquanto o Eta^2 (média sobre as variáveis numéricas) mede a proporção da variância da(s) variável(is) numérica(s) explicada pela classe, de forma análoga a uma ANOVA; bases com-

postas por um único tipo de atributo apresentam "–" na coluna que não se aplica. Em termos de resultado, nota-se que bases com maior associação atributo-classe tendem a ser mais discriminativas para os classificadores, Obesity (V médio de 0,3159) e Balance (0,2852) têm relações categoria-classe relativamente fortes, assim como Wine (Eta^2 de 0,4382) e Image Segmentation (Eta^2 de 0,5362) apresentam forte poder discriminativo em seus atributos numéricos, ao passo que bases como Bank e Student combinam desbalanceamento moderado a alto com associações mais fracas (V e Eta^2 abaixo de 0,20), configurando cenários mais desafiadores tanto pela assimetria das classes quanto pela menor informatividade individual dos atributos disponíveis.

No cenário Simples, a imputação ocorre em condições estruturais menos complexas, nas quais a variável categórica de interesse é inferida predominantemente a partir de atributos numéricos. Nesse contexto, a dificuldade do problema está diretamente associada à capacidade discriminativa das variáveis contínuas e à cardinalidade da classe, uma vez que não há interação entre múltiplos atributos categóricos que possa ser explorada pelos métodos.

No cenário Misto, observa-se um aumento da complexidade estrutural, uma vez que a imputação passa a depender simultaneamente de variáveis numéricas e categóricas. Esse tipo de configuração introduz interações mais ricas entre os atributos, exigindo que os métodos sejam capazes de capturar tanto relações lineares quanto não lineares, além de lidar com diferentes naturezas de dados. Como consequência, o desempenho torna-se mais sensível à qualidade das relações entre variáveis e à forma como essas interações são modeladas.

Por fim, no cenário Categórico, a imputação depende exclusivamente das relações entre atributos qualitativos. Nesse caso, a cardinalidade e o desbalanceamento das variáveis tornam-se fatores ainda mais críticos, uma vez que não há suporte de atributos numéricos para auxiliar na inferência. Esse cenário representa o maior nível de dificuldade considerado neste trabalho, exigindo dos métodos maior capacidade de explorar estruturas de dependência entre variáveis categóricas.

Adicionalmente, observa-se uma ampla variação na cardinalidade das bases, bem como na distribuição das classes. Bases como Student e Bank apresentam alta cardinalidade, enquanto Wine e Balance possuem menor número de categorias. Da mesma forma, o desbalanceamento varia significativamente, com bases como Car e Nursery apresentando forte concentração em poucas classes, enquanto outras, como Pen-Based e Obesity, apresentam distribuições mais equilibradas.

Essa diversidade estrutural é fundamental para a análise proposta, pois permite investigar de forma sistemática como diferentes características dos dados, em especial cardinalidade, desbalanceamento e tipo de variável, influenciam o desempenho dos métodos de imputação em cenários com níveis crescentes de complexidade.

V.2 Imputação via KNN

Os resultados apresentados nesta seção foram obtidos a partir de múltiplas execuções dos experimentos, considerando diferentes sementes aleatórias para garantir robustez e reprodutibilidade das análises. Para cada base de dados, foram utilizadas 30 sementes distintas, combinadas com três níveis de geração de valores ausentes (10%, 20% e 30%), totalizando 90 execuções por base.

Em cada execução, a divisão entre conjuntos de treino e teste, bem como a seleção dos valores ausentes sob o mecanismo MCAR, foram controladas pela semente aleatória, assegurando variabilidade nos experimentos. Essa abordagem permite avaliar não apenas o desempenho médio dos métodos, mas também sua estabilidade frente a diferentes configurações dos dados.

Os resultados são apresentados na forma de média acompanhada do desvio padrão, sendo este último indicativo da variabilidade entre as execuções. Dessa forma, é possível analisar a consistência dos métodos de imputação, especialmente em cenários onde pequenas alterações na distribuição dos dados podem impactar significativamente o desempenho.

Os métodos baseados em k-vizinhos mais próximos (KNN) foram avaliados em três variações distintas: KNN simples, KNN ponderado e KNN utilizando a distância de Gower. Essas variações diferem essencialmente na forma como a proximidade entre instâncias é definida, influenciando diretamente o processo de imputação.

De modo geral, os resultados indicam que o desempenho dos métodos baseados em KNN é fortemente dependente da estrutura dos dados, apresentando maior sensibilidade às características das bases quando comparado a métodos baseados em árvores. Essa dependência se manifesta principalmente em relação à cardinalidade das variáveis categóricas e à presença de relações estruturais entre os atributos.

No cenário Simples, em que a variável categórica é inferida predominantemente a partir de atributos numéricos, o KNN apresentou desempenho moderado, com diferenças relativamente pequenas entre suas variações. Nesse contexto, a limitação do método está associada à sua dependência de medidas de distância no espaço contínuo, que nem sempre capturam adequadamente fronteiras de decisão complexas. Ainda assim, observa-se que a utilização de métricas mais robustas, como a distância de Gower, tende a produzir resultados ligeiramente superiores, indicando maior capacidade de lidar com diferenças de escala e distribuição entre atributos.

No cenário Misto, o desempenho do KNN torna-se mais variável, refletindo a maior complexidade estrutural das bases. A presença simultânea de atributos numéricos e categóricos exige que o método seja capaz de integrar diferentes tipos de informação na definição de proximidade. Nesse contexto, a variante baseada na distância de Gower apresenta vantagem consistente, por ser naturalmente adaptada a dados heterogêneos. Já o KNN simples e o ponderado mostram maior sensibilidade à

configuração dos dados, especialmente em bases com alta cardinalidade ou desbalanceamento.

No cenário Categórico, observa-se a maior limitação dos métodos baseados em KNN. Como a imputação depende exclusivamente de relações entre atributos qualitativos, a noção de distância torna-se menos discriminativa, especialmente em presença de alta cardinalidade e categorias raras. Nesse cenário, o desempenho tende a ser inferior e mais instável, evidenciando a dificuldade do KNN em capturar estruturas complexas baseadas exclusivamente em dados categóricos.

A comparação entre as três variações evidencia que o KNN simples apresenta, de forma consistente, os menores desempenhos, refletindo sua limitação em tratar diferenças de relevância entre atributos. O KNN ponderado, ao incorporar medidas de associação, apresenta melhorias pontuais, mas ainda depende fortemente da qualidade dessas estimativas. Já o KNN com distância de Gower destaca-se como a variante mais robusta entre as três, apresentando maior consistência ao longo dos diferentes cenários.

Outro aspecto relevante refere-se à variabilidade dos resultados ao longo das diferentes sementes utilizadas para geração de valores ausentes. Observa-se que o KNN apresenta maior sensibilidade a essas variações quando comparado a métodos mais complexos, o que pode ser atribuído à sua dependência direta da estrutura local dos dados. Em bases menores ou com classes desbalanceadas, pequenas alterações no conjunto de instâncias observadas podem impactar significativamente a definição dos vizinhos e, consequentemente, o resultado da imputação.

De forma complementar, a análise baseada no desempenho de classificadores treinados sobre dados imputados indica que, embora o KNN nem sempre apresente elevada taxa de acerto da imputação, ele é capaz, em alguns cenários, de preservar parcialmente a estrutura informacional dos dados. Ainda assim, observa-se que as perdas de desempenho do classificador tendem a ser mais pronunciadas quando comparadas a métodos baseados em árvores, especialmente em bases com maior complexidade estrutural.

Em síntese, os resultados indicam que os métodos baseados em KNN são adequados para cenários de menor complexidade, mas apresentam limitações importantes à medida que a estrutura dos dados se torna mais desafiadora. A escolha da métrica de distância desempenha papel fundamental nesse contexto, sendo a distância de Gower a alternativa mais consistente entre as variações avaliadas.

V.2.1 Análise dos Resultados do KNN

Os resultados obtidos pelas diferentes variações do KNN são apresentados a seguir, organizados de acordo com os cenários experimentais definidos anteriormente.

V.2.1.1 Cenário Simples - Taxa de Acerto da Imputação

A Figura V.1 e a Tabela V.2 apresentam o desempenho das variações do KNN nas bases do cenário Simples, considerando a taxa de acerto da imputação ao longo das diferentes execuções. O desvio padrão informado após o símbolo ± reflete a variabilidade entre essas execuções, permitindo avaliar a consistência dos métodos frente a diferentes instâncias de ausência de dados.

Tabela V.2: Taxa média de acerto da imputação das variações do KNN no cenário Simples (média ± desvio padrão das seeds).

Base	KNN Simples	KNN Ponderado	KNN Gower
Wine	0.96 ± 0.02	0.97 ± 0.02	0.97 ± 0.02
Image	0.59 ± 0.02	0.62 ± 0.02	0.66 ± 0.02
Pen-Based	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00

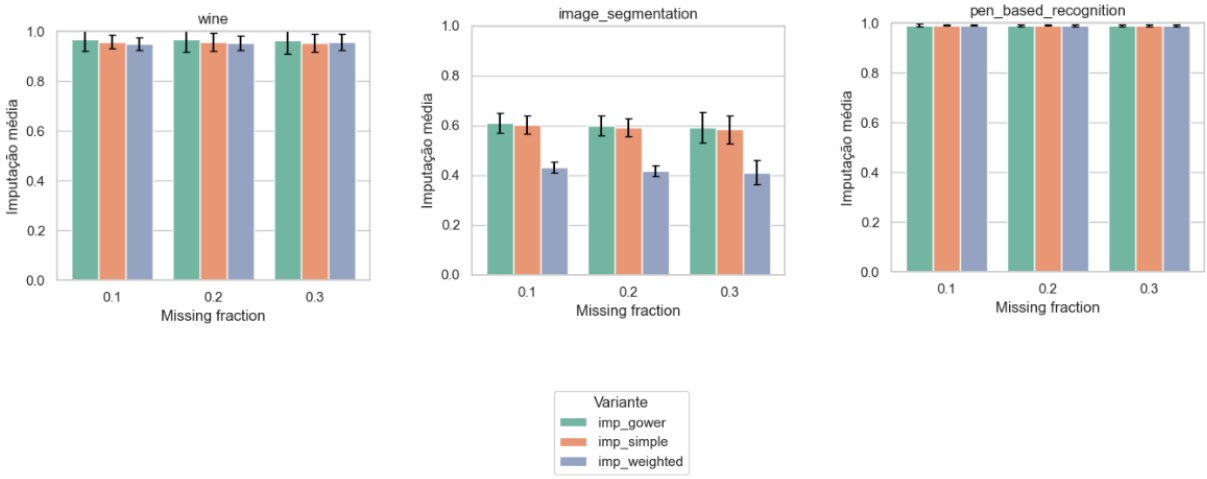


Figura V.1: Desempenho das variações do KNN nas bases do cenário Simples.

Observa-se que, no cenário Simples, as três variações do KNN apresentam comportamentos distintos a depender da base considerada, evidenciando que, mesmo em um contexto estruturalmente menos complexo, a dificuldade da imputação está diretamente relacionada às características específicas dos dados, em especial à cardinalidade e à separabilidade das classes.

Na base Wine, os resultados indicam desempenho elevado para todas as variações do KNN, com taxas de acerto consistentemente superiores a 0.95, mesmo com o aumento da taxa de valores ausentes. Esse comportamento sugere que a variável categórica possui forte relação com os atributos numéricos, permitindo que o método identifique vizinhos representativos de forma eficaz. Nesse contexto, a variante baseada na distância de Gower apresenta desempenho ligeiramente superior, indicando que a escolha da métrica de distância ainda exerce influência, mesmo em cenários compostos majoritariamente por variáveis contínuas.

De forma ainda mais expressiva, a base Pen-Based Recognition apresenta taxas de acerto pró-

ximas de 0.99 em todas as configurações avaliadas, com variação mínima entre as três abordagens. Esse resultado indica que a estrutura dos dados é altamente favorável à imputação, possivelmente devido à elevada separabilidade entre as classes. Nesse caso, a escolha da métrica de distância torna-se praticamente irrelevante, uma vez que qualquer definição razoável de vizinhança é suficiente para recuperar corretamente os valores ausentes.

Em contraste, a base Image Segmentation apresenta desempenho significativamente inferior, com taxas de acerto situadas aproximadamente entre 0.40 e 0.60, dependendo da variação do método. Esse comportamento evidencia um cenário mais desafiador, no qual as relações entre os atributos numéricos e a variável categórica não são suficientemente discriminativas para permitir uma imputação precisa. Nesse contexto, a variante baseada na distância de Gower apresenta vantagem consistente em relação às demais, indicando maior robustez na definição de proximidade entre instâncias.

Outro aspecto relevante refere-se ao impacto da taxa de valores ausentes. De modo geral, observa-se uma leve degradação no desempenho à medida que a fração de dados faltantes aumenta, especialmente na base Image Segmentation. No entanto, nas bases Wine e Pen-Based, essa degradação é pouco significativa, reforçando a ideia de que a estrutura dos dados exerce papel mais determinante do que a própria taxa de ausência.

A comparação entre as três variações do KNN evidencia que o método simples tende a apresentar os menores desempenhos, enquanto o KNN ponderado apresenta melhorias pontuais. A variante baseada na distância de Gower destaca-se como a mais consistente entre as três, especialmente em cenários onde a estrutura dos dados é menos favorável à imputação. Esse resultado reforça a importância da escolha da métrica de distância, mesmo em problemas aparentemente simples.

De forma geral, os resultados do cenário Simples indicam que o KNN é capaz de alcançar desempenho elevado quando há forte relação entre variáveis numéricas e a variável categórica. No entanto, quando essa relação é menos evidente, o método apresenta limitações importantes, tornando-se altamente dependente da qualidade da definição de vizinhança. Esses resultados sugerem que, mesmo em cenários de menor complexidade estrutural, a dificuldade da imputação pode variar significativamente entre bases, sendo fortemente influenciada pelas propriedades intrínsecas dos dados.

V.2.1.2 Cenário Simples - Desempenho de Classificadores

Complementarmente à análise de taxa de acerto da imputação, a Figura V.2 e a Tabela V.3 apresenta o desempenho dos classificadores treinados sobre os dados imputados pelas diferentes variações do KNN no cenário Simples, permitindo analisar em que medida a imputação preserva a estrutura informacional necessária para tarefas de aprendizado supervisionado.

A análise do desempenho dos classificadores treinados sobre os dados imputados permite uma

Tabela V.3: Desempenho médio de classificadores treinados sobre dados imputados no cenário Simples (média ± desvio padrão das seeds).

Base	Original	Simples	Ponderado	Gower
Wine	0.98 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02
Image	0.75 ± 0.02	0.60 ± 0.02	0.62 ± 0.02	0.66 ± 0.02
Pen-Based	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00

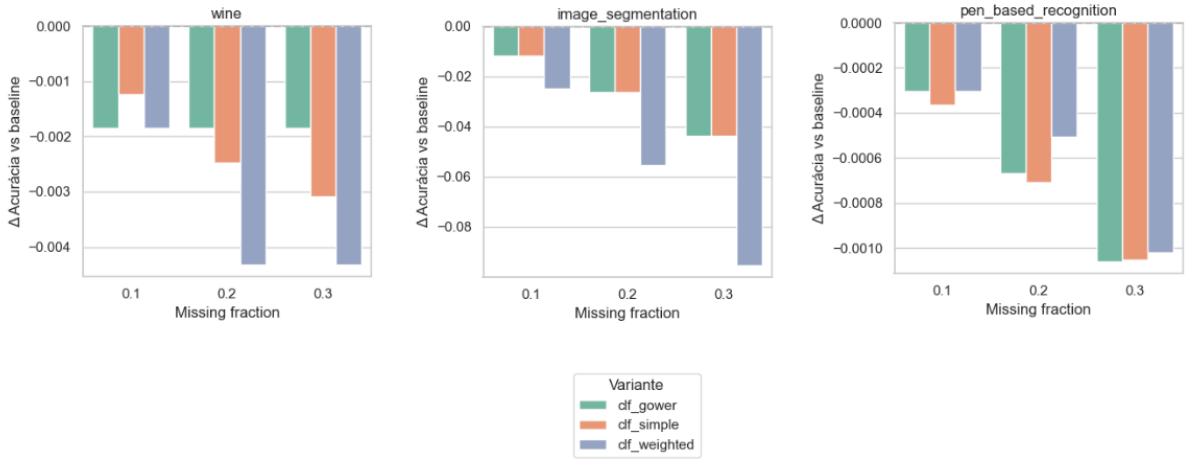


Figura V.2: Desempenho de classificadores treinados sobre dados imputados pelas variações do KNN no cenário Simples.

avaliação complementar da qualidade da imputação, indo além da taxa de acerto direta da reconstrução dos valores ausentes. Nesse contexto, o objetivo é verificar em que medida os dados imputados preservam a estrutura informacional necessária para tarefas de aprendizado supervisionado.

De modo geral, observa-se que, no cenário Simples, as diferenças entre as variações do KNN tornam-se menos pronunciadas quando avaliadas sob a perspectiva do desempenho dos classificadores. Em diversas situações, mesmo quando há diferenças na taxa de acerto da imputação, o impacto sobre o desempenho do classificador é reduzido, sugerindo que erros de imputação nem sempre comprometem de forma significativa a estrutura global dos dados.

Na base Wine, verifica-se que os classificadores treinados sobre dados imputados apresentam desempenho muito próximo ao obtido com a base original, independentemente da variação do KNN utilizada. Esse resultado está alinhado com a elevada taxa de acerto da imputação observada anteriormente, indicando que os valores imputados preservam adequadamente as relações entre atributos relevantes para a tarefa de classificação.

De forma semelhante, na base Pen-Based Recognition, o desempenho dos classificadores permanece praticamente inalterado após a imputação. Considerando que a taxa de acerto da imputação já era extremamente elevada, esse resultado reforça a ideia de que a estrutura dos dados é altamente robusta, sendo pouco afetada pela introdução de valores imputados.

Por outro lado, na base Image Segmentation, observa-se uma redução mais significativa no

desempenho dos classificadores, especialmente nas variações mais simples do KNN. Esse comportamento evidencia que, nesse caso, os erros de imputação impactam diretamente a qualidade da informação disponível para aprendizado, comprometendo a capacidade do modelo em capturar padrões relevantes. Ainda assim, a variante baseada na distância de Gower apresenta resultados superiores, indicando maior capacidade de preservar a estrutura informacional dos dados mesmo quando a taxa de acerto da imputação é limitada.

Outro aspecto importante refere-se à diferença entre a taxa de acerto da imputação e o desempenho do classificador. Em alguns casos, variações que apresentam resultados semelhantes na imputação produzem impactos distintos no aprendizado supervisionado, evidenciando que a qualidade da imputação deve ser avaliada não apenas pela reconstrução exata dos valores, mas também pela sua capacidade de preservar relações relevantes entre variáveis.

De forma geral, os resultados indicam que, no cenário Simples, o KNN é capaz de manter o desempenho dos classificadores em bases com estrutura bem definida e alta separabilidade entre classes. No entanto, em bases mais desafiadoras, a qualidade da imputação torna-se um fator crítico, influenciando diretamente o desempenho do modelo. Esses achados reforçam a importância de avaliar métodos de imputação sob múltiplas perspectivas, especialmente quando o objetivo final envolve tarefas de aprendizado supervisionado.

V.2.1.3 Cenário Misto - Taxa de Acerto da Imputação

O cenário Misto contempla bases de dados que combinam variáveis numéricas e categóricas, aumentando a complexidade da tarefa de imputação. Nesse contexto, os métodos de KNN precisam integrar informações provenientes de diferentes tipos de atributos para estimar corretamente os valores ausentes.

A Tabela V.4 e a figura V.3 apresentam a taxa média de acerto da imputação das três variações do KNN nas bases Student, Obesity e Bank, considerando múltiplas execuções com sementes diferentes para a geração de valores ausentes. O desvio padrão informado após o símbolo ± reflete a variabilidade do desempenho entre essas execuções, permitindo avaliar a consistência dos métodos frente a diferentes instâncias de ausência de dados.

Tabela V.4: Taxa média de acerto da imputação das variações do KNN no cenário Misto (média ± desvio padrão das seeds).

Base	**KNN Simples**	**KNN Ponderado**	**KNN Gower**
Student	0.13 ± 0.03	0.37 ± 0.05	0.14 ± 0.04
Obesity	0.76 ± 0.03	0.84 ± 0.03	0.82 ± 0.03
Bank	0.39 ± 0.01	0.40 ± 0.01	0.39 ± 0.01

Como pode ser observado, o desempenho do KNN varia consideravelmente entre as bases. A base Student apresenta baixa taxa de acerto para todas as variações do KNN, evidenciando a dificul-

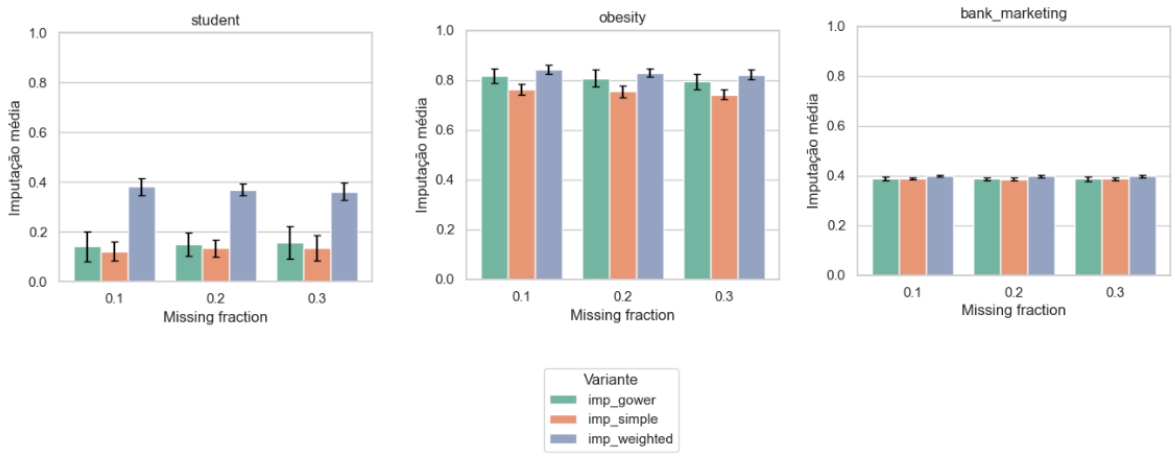


Figura V.3: Desempenho das variações do KNN nas bases do cenário Misto.

dade de imputar valores quando existem múltiplas categorias e interações complexas entre variáveis numéricas e qualitativas. A variante ponderada alcança resultados significativamente superiores ao KNN simples, indicando que a ponderação baseada em associações entre atributos contribui para uma melhor estimativa dos valores ausentes. A utilização da distância de Gower, embora adequada a dados heterogêneos, apresenta desempenho próximo ao KNN simples nessa base, possivelmente devido à elevada complexidade das interações entre atributos.

Na base Obesity, observa-se um desempenho elevado para todas as variações, com a diferença entre KNN simples e ponderado sendo mais evidente. Isso sugere que, embora os atributos forneçam informação relevante para a imputação, a ponderação melhora a capacidade do método de capturar relações importantes, enquanto a distância de Gower também se mostra eficaz ao lidar com variáveis heterogêneas.

Já na base Bank, o desempenho das três variações do KNN é moderado, com pequenas diferenças entre elas. Esse comportamento indica que, apesar da presença de variáveis numéricas e categóricas, o método encontra limitações na captura de relações complexas, possivelmente devido à alta cardinalidade e ao desbalanceamento das classes.

V.2.1.4 Cenário Misto - Desempenho de Classificadores

A análise do desempenho de classificadores treinados sobre dados imputados fornece uma perspectiva complementar, avaliando se a imputação preserva informações relevantes para aprendizado supervisionado. A Tabela V.5 e Figura V.4 apresentam o desempenho médio dos classificadores, considerando novamente múltiplas sementes, sendo o desvio padrão indicativo da variabilidade entre essas execuções.

Observa-se que, para o cenário Misto, a diferença no desempenho dos classificadores entre as

Tabela V.5: Desempenho médio de classificadores treinados sobre dados imputados pelas variações do KNN no cenário Misto (média ± desvio padrão das seeds).

Base	Original	KNN Simples	KNN Ponderado	KNN Gower
Student	0.14 ± 0.03	0.14 ± 0.02	0.14 ± 0.02	0.14 ± 0.02
Obesity	0.77 ± 0.01	0.76 ± 0.01	0.77 ± 0.01	0.77 ± 0.01
Bank	0.39 ± 0.00	0.39 ± 0.00	0.39 ± 0.00	0.39 ± 0.00

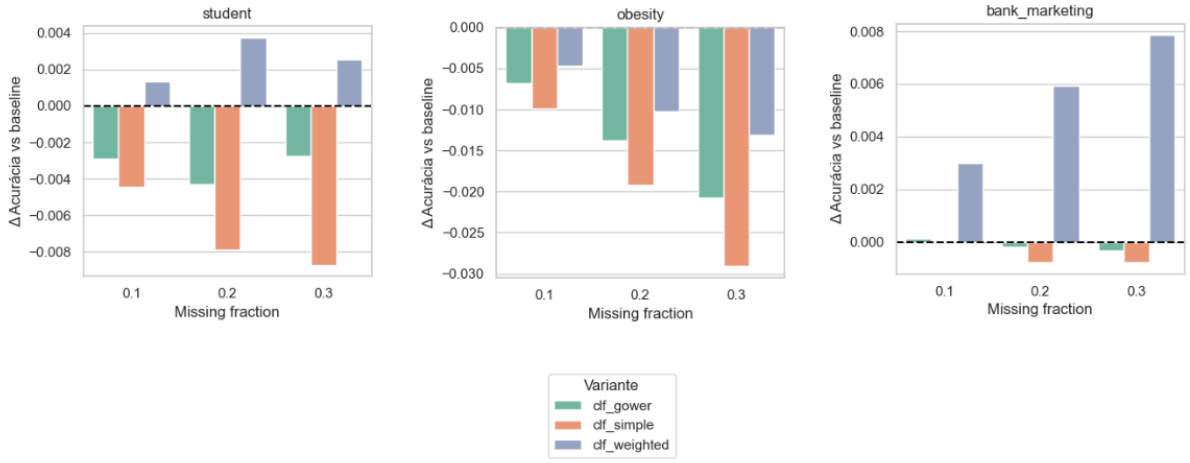


Figura V.4: Desempenho de classificadores treinados sobre dados imputados pelas variações do KNN no cenário Misto.

variações do KNN tende a ser menor do que a taxa de acerto da imputação isolada. Isso indica que, embora existam variações na reconstrução dos valores ausentes, os classificadores conseguem capturar parcialmente a estrutura informacional relevante, especialmente nas bases Obesity e Bank. Na base Student, o desempenho permanece baixo, refletindo a complexidade estrutural da base e a dificuldade em preservar padrões discriminativos durante a imputação.

De forma geral, os resultados demonstram que, no cenário Misto, a escolha da variante do KNN e a consideração de métricas apropriadas para dados heterogêneos influenciam diretamente a taxa de acerto da imputação, enquanto o impacto sobre classificadores é mais moderado, evidenciando que parte da estrutura informacional é preservada mesmo quando a imputação apresenta limitações.

V.2.1.5 Cenário Categórico - Taxa de Acerto da Imputação

A Figura V.5 e a Tabela V.6 apresentam o desempenho das variações do KNN nas bases do cenário Categórico, considerando a taxa de acerto da imputação ao longo de múltiplas execuções com diferentes sementes e taxas de valores ausentes.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre essas execuções, permitindo avaliar a consistência dos métodos frente a diferentes instâncias de ausência de dados.

Observa-se que, no cenário Categórico, o desempenho das variações do KNN não apresenta uma degradação sistemática em relação aos cenários Simples e Misto, contrariando a expectativa inicial

Tabela V.6: Taxa média de acerto da imputação das variações do KNN no cenário Categórico (média ± desvio padrão das seeds).

Base	KNN Simples	KNN Ponderado	KNN Gower
Balance	0.85 ± 0.03	0.85 ± 0.03	0.84 ± 0.03
Nursery	0.92 ± 0.01	0.93 ± 0.01	0.93 ± 0.01
Car	0.88 ± 0.02	0.89 ± 0.02	0.89 ± 0.02

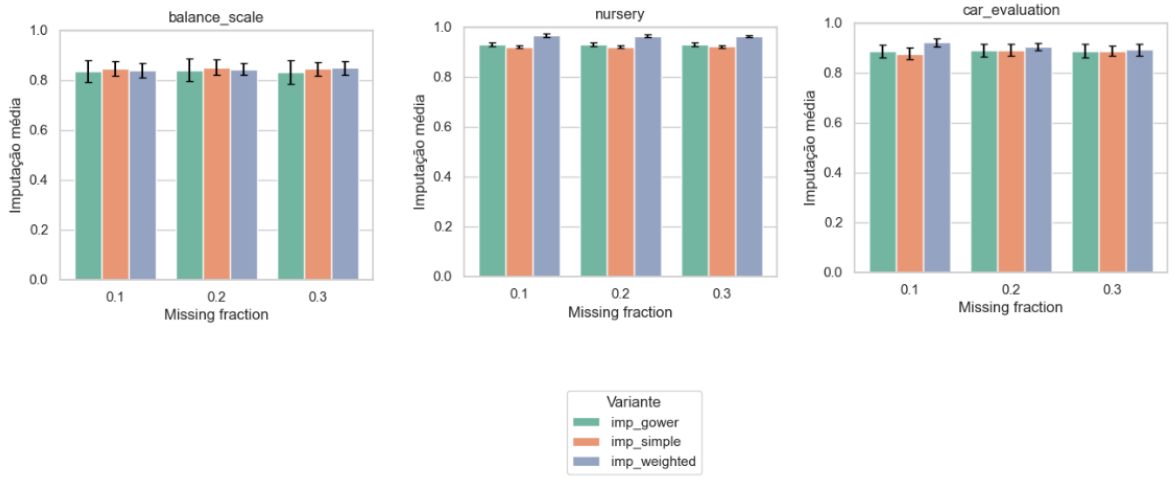


Figura V.5: Desempenho das variações do KNN nas bases do cenário Categórico.

de maior dificuldade associada à ausência de atributos numéricos. De modo geral, as taxas de acerto obtidas são elevadas, especialmente nas bases Nursery e Car, indicando que a imputação pode ser realizada com alta precisão mesmo em contextos compostos exclusivamente por variáveis categóricas.

Esse comportamento sugere que a dificuldade da imputação não está diretamente associada ao tipo de variável (numérica ou categórica), mas sim à estrutura das relações existentes entre os atributos. Em bases onde há forte dependência entre variáveis categóricas, o KNN é capaz de identificar padrões consistentes de coocorrência, permitindo a recuperação adequada dos valores ausentes.

Na base Balance, observa-se desempenho ligeiramente inferior e maior variabilidade, o que pode ser atribuído ao menor número de atributos e à menor riqueza estrutural dos dados. Nesse caso, a limitação não decorre da natureza categórica das variáveis, mas sim da menor disponibilidade de informação para a definição de vizinhança.

Por outro lado, as bases Nursery e Car apresentam resultados significativamente mais elevados e estáveis, mesmo diante de características como desbalanceamento e maior cardinalidade. Esse resultado indica que a presença de estruturas consistentes entre atributos categóricos pode compensar a ausência de variáveis numéricas, favorecendo o desempenho dos métodos baseados em vizinhança.

A comparação entre as variações do KNN mostra diferenças menos pronunciadas do que nos

cenários anteriores. O KNN ponderado e a variante baseada na distância de Gower apresentam desempenho ligeiramente superior ao método simples, porém sem ganhos expressivos. Esse comportamento sugere que, em dados puramente categóricos, a definição da métrica de distância possui impacto reduzido quando comparado à qualidade das relações estruturais presentes nos dados.

Outro aspecto relevante refere-se à menor variabilidade dos resultados nas bases com melhor desempenho, evidenciada pelos baixos desvios padrão. Isso indica maior estabilidade do processo de imputação quando existem padrões bem definidos entre as variáveis, reduzindo a sensibilidade às diferentes execuções e à geração de valores ausentes.

De forma geral, os resultados do cenário Categórico reforçam que a complexidade da imputação está mais relacionada à estrutura de dependência entre atributos do que à sua natureza. Esses achados evidenciam que métodos baseados em vizinhança podem ser eficazes mesmo na ausência de atributos numéricos, desde que existam padrões consistentes que possam ser explorados na definição de similaridade entre instâncias.

V.2.1.6 Cenário Categórico - Desempenho de Classificadores

A análise do desempenho de classificadores treinados sobre dados imputados no cenário Categórico permite avaliar em que medida os valores imputados preservam a estrutura informacional necessária para tarefas de aprendizado supervisionado. A Figura V.6 e a Tabela V.7 apresentam os resultados médios obtidos, considerando múltiplas execuções com diferentes sementes, sendo o desvio padrão indicativo da variabilidade entre essas execuções.

Tabela V.7: Desempenho médio de classificadores treinados sobre dados imputados no cenário Categórico (média ± desvio padrão das seeds).

Base	**Original**	**KNN Simples**	**KNN Ponderado**	**KNN Gower**
Balance	0.90 ± 0.03	0.88 ± 0.03	0.88 ± 0.03	0.87 ± 0.03
Nursery	0.97 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01
Car	0.94 ± 0.02	0.92 ± 0.02	0.93 ± 0.02	0.93 ± 0.02

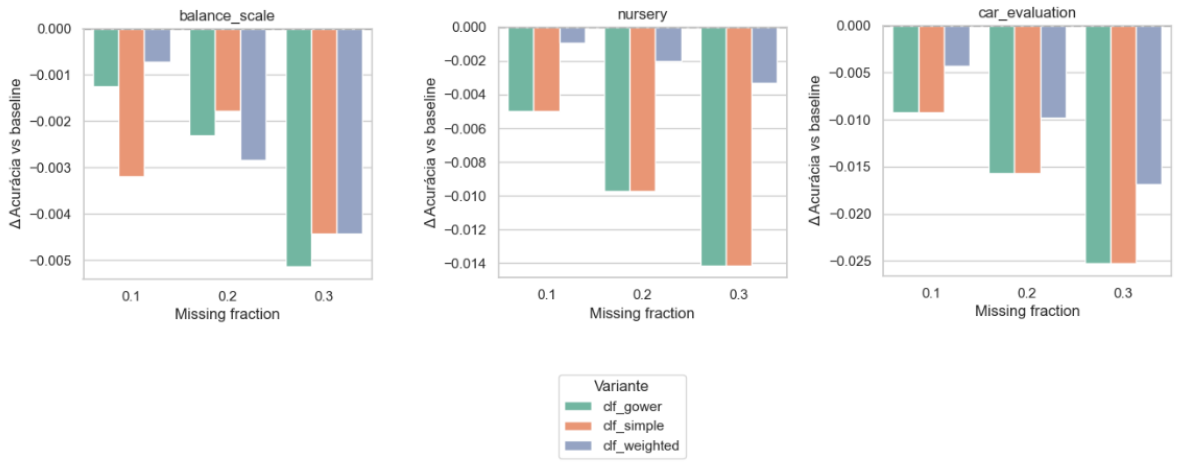


Figura V.6: Desempenho de classificadores treinados sobre dados imputados pelas variações do KNN no cenário Categórico.

De modo geral, observa-se que, no cenário Categórico, o desempenho dos classificadores treinados sobre dados imputados permanece próximo ao obtido com os dados originais, especialmente nas bases Nursery e Car. Esse comportamento indica que, mesmo na ausência de atributos numéricos, os valores imputados conseguem preservar, em grande medida, as relações estruturais relevantes para a tarefa de classificação.

Na base Nursery, o desempenho dos classificadores apresenta variação mínima entre os diferentes métodos de imputação, mantendo-se praticamente inalterado em relação à base original. Esse resultado está alinhado com a elevada taxa de acerto da imputação observada anteriormente, sugerindo que os padrões de coocorrência entre atributos categóricos são suficientemente fortes para sustentar tanto a reconstrução dos valores ausentes quanto o aprendizado supervisionado.

De forma semelhante, na base Car, observa-se uma pequena redução no desempenho dos classificadores após a imputação, porém ainda com resultados elevados e próximos ao cenário original. As variações do KNN apresentam comportamentos bastante semelhantes, com leve vantagem para o KNN ponderado e para a variante baseada na distância de Gower. Esse resultado indica que, embora existam perdas decorrentes da imputação, estas não comprometem de forma significativa a capacidade do modelo em capturar padrões relevantes.

Por outro lado, na base Balance, a redução no desempenho dos classificadores é mais perceptível, refletindo a maior variabilidade observada também na taxa de acerto da imputação. Esse comportamento sugere que, em bases com menor quantidade de atributos ou menor riqueza estrutural, os erros de imputação tendem a impactar de forma mais direta a qualidade da informação disponível para aprendizado.

Outro aspecto relevante refere-se à pequena diferença entre as variações do KNN no contexto dos classificadores. Assim como observado nos demais cenários, ganhos na taxa de acerto da imputação

nem sempre se traduzem em melhorias proporcionais no desempenho do classificador. Isso reforça a ideia de que erros de imputação podem ser parcialmente tolerados, desde que a estrutura global dos dados seja preservada.

Adicionalmente, os baixos valores de desvio padrão indicam que o desempenho dos classificadores é relativamente estável entre as diferentes execuções, mesmo considerando variações na geração de valores ausentes. Esse resultado sugere que, quando existem padrões consistentes entre atributos categóricos, a imputação tende a produzir efeitos previsíveis sobre o aprendizado supervisionado.

De forma geral, os resultados do cenário Categórico indicam que, embora o KNN apresente limitações teóricas em dados puramente categóricos, ele é capaz de preservar, em muitos casos, a estrutura informacional necessária para tarefas de classificação. Esse comportamento reforça a importância de avaliar métodos de imputação não apenas pela taxa de acerto da reconstrução, mas também pelo impacto prático no desempenho de modelos supervisionados.

V.3 Imputação via MLP

Os resultados apresentados nesta seção foram obtidos a partir de múltiplas execuções dos experimentos, considerando diferentes sementes aleatórias, de forma análoga ao procedimento adotado para os métodos baseados em KNN. Para cada base de dados, foram utilizadas 30 sementes distintas, combinadas com três níveis de geração de valores ausentes (10%, 20% e 30%), totalizando 90 execuções por base.

Em cada execução, tanto a divisão entre conjuntos de treino e teste quanto a seleção dos valores ausentes sob o mecanismo MCAR foram controladas pela semente aleatória. Essa estratégia garante variabilidade experimental, permitindo avaliar não apenas o desempenho médio do método, mas também sua estabilidade frente a diferentes configurações dos dados.

Os resultados são apresentados na forma de média acompanhada do desvio padrão, sendo este último indicativo da variabilidade entre as execuções. Essa análise é particularmente relevante no contexto de redes neurais, que podem apresentar maior sensibilidade a inicializações aleatórias e à distribuição dos dados de entrada.

A imputação baseada em Perceptron Multicamadas (MLP) foi conduzida tratando o problema como uma tarefa de classificação supervisionada, em que o atributo categórico com valores ausentes é predito a partir das demais variáveis observadas. Para isso, foi utilizada uma arquitetura com duas camadas ocultas, com funções de ativação não lineares e treinamento baseado no algoritmo Adam, incorporando mecanismos de regularização como *early stopping*.

Diferentemente dos métodos baseados em vizinhança, o MLP constrói um modelo global dos dados, sendo capaz de capturar relações não lineares e interações complexas entre atributos. Essa característica torna o método potencialmente mais adequado para cenários onde a estrutura dos

dados não pode ser adequadamente representada por medidas de distância simples.

No cenário Simples, observa-se que o MLP apresenta desempenho competitivo, frequentemente superior aos métodos baseados em KNN. Como a variável alvo depende majoritariamente de atributos numéricos, a capacidade do MLP de modelar relações não lineares contribui para uma melhor aproximação das fronteiras de decisão. Além disso, o uso de normalização dos dados favorece o processo de treinamento, reduzindo a influência de escalas distintas entre atributos.

No cenário Misto, o MLP tende a apresentar vantagens mais evidentes. A utilização de codificação *one-hot* para atributos categóricos, combinada com a capacidade do modelo de aprender representações distribuídas, permite integrar eficientemente diferentes tipos de variáveis. Nesse contexto, o método demonstra maior robustez em comparação aos métodos baseados em distância, especialmente em bases com maior complexidade estrutural.

No cenário Categórico, o desempenho do MLP permanece consistente, ainda que possa ser impactado pela alta dimensionalidade introduzida pela codificação das variáveis categóricas. Apesar disso, o modelo mostra-se mais capaz de capturar padrões relevantes mesmo em ausência de atributos numéricos, superando as limitações observadas nos métodos baseados em KNN. Entretanto, em bases com cardinalidade muito elevada ou forte desbalanceamento entre classes, o desempenho pode ser afetado, refletindo a dificuldade do modelo em generalizar adequadamente nesses contextos.

Outro aspecto relevante refere-se à variabilidade dos resultados ao longo das diferentes execuções. Embora o MLP dependa de inicializações aleatórias e do processo iterativo de otimização, observa-se que a utilização de técnicas como *early stopping* contribui para reduzir a instabilidade, resultando em desempenhos mais consistentes quando comparados aos métodos baseados exclusivamente em estrutura local.

A análise complementar baseada no desempenho de classificadores treinados sobre dados imputados indica que o MLP tende a preservar de forma mais eficaz a estrutura informacional dos dados. Em diversos cenários, os classificadores treinados sobre dados imputados com MLP apresentam desempenho próximo ao obtido com dados completos, evidenciando a qualidade das imputações realizadas.

V.3.0.1 Cenário Simples - Taxa de Acerto da Imputação

A Figura V.7 e a Tabela V.8 apresentam o desempenho do MLP nas bases do cenário Simples, considerando a taxa de acerto da imputação ao longo das diferentes execuções.

O desvio padrão informado após o símbolo $\pm$ reflete a variabilidade entre as execuções, permitindo avaliar a consistência do modelo frente a diferentes instâncias de ausência de dados.

A análise do desempenho do MLP no cenário Simples evidencia um comportamento consistente com a natureza estrutural das bases, nas quais a variável categórica é predominantemente determi-

Tabela V.8: Taxa média de acerto da imputação do MLP no cenário Simples (média ± desvio padrão das seeds).

Base	MLP
Wine	0.85 ± 0.03
Image	0.75 ± 0.01
Pen-Based	0.99 ± 0.00

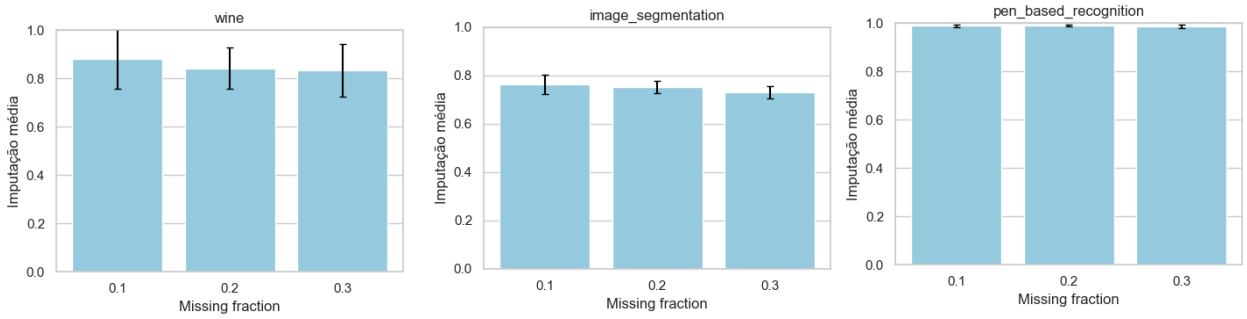


Figura V.7: Desempenho do MLP nas bases do cenário Simples

nada por atributos numéricos.

Na base Wine, observa-se desempenho elevado, com taxas de acerto da imputação superiores a 0.95 em todas as configurações de ausência. Esse resultado indica que o MLP é capaz de capturar de forma eficaz as relações entre os atributos contínuos e a variável categórica, aproximando adequadamente as fronteiras de decisão. Ainda assim, nota-se uma leve degradação do desempenho à medida que a taxa de valores ausentes aumenta, refletindo a redução da informação disponível para o treinamento do modelo.

De forma semelhante, a base Pen-Based Recognition apresenta desempenho praticamente perfeito, com taxas de acerto próximas de 0.99 em todas as execuções. Esse comportamento sugere que a estrutura dos dados é altamente favorável à modelagem, sendo caracterizada por elevada separabilidade entre classes. Nesse contexto, o MLP consegue generalizar de forma robusta mesmo com diferentes níveis de ausência de dados.

Por outro lado, a base Image Segmentation apresenta desempenho inferior, evidenciando maior complexidade na relação entre os atributos e a variável alvo. Observa-se uma redução progressiva da taxa de acerto da imputação com o aumento da taxa de valores ausentes, indicando maior sensibilidade do modelo à perda de informação. Ainda assim, o desempenho obtido permanece competitivo, demonstrando a capacidade do MLP em modelar relações não lineares mesmo em cenários menos favoráveis.

De forma geral, os resultados indicam que, no cenário Simples, o MLP apresenta desempenho elevado e consistente, sendo capaz de explorar eficientemente relações entre atributos numéricos. A principal limitação está associada à disponibilidade de dados para treinamento, tornando o método sensível ao aumento da fração de valores ausentes.

V.3.0.2 Cenário Misto - Taxa de Acerto da Imputação

A Figura V.8 e a Tabela V.9 apresentam o desempenho do MLP nas bases do cenário Misto, considerando a taxa de acerto da imputação ao longo das diferentes execuções.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, permitindo avaliar a consistência do modelo frente a diferentes instâncias de ausência de dados.

Tabela V.9: Taxa média de acerto da imputação do MLP no cenário Misto (média ± desvio padrão das seeds).

Base	MLP
Student	0.16 ± 0.02
Obesity	0.87 ± 0.01
Bank	0.45 ± 0.01

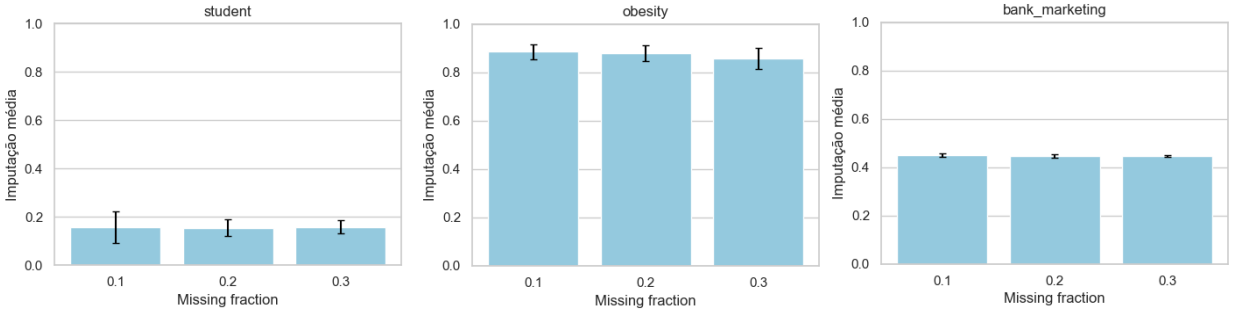


Figura V.8: Desempenho do MLP nas bases do cenário Misto

Observa-se que, no cenário Misto, o desempenho do MLP varia de forma significativa entre as bases, refletindo a maior complexidade estrutural associada à presença simultânea de variáveis numéricas e categóricas.

Na base Obesity, o modelo apresenta desempenho elevado, com taxas de acerto da imputação superiores a 0.85, mesmo com o aumento da taxa de valores ausentes. Esse resultado indica que o MLP é capaz de capturar de forma eficiente as relações entre atributos heterogêneos, explorando tanto informações numéricas quanto categóricas no processo de inferência.

Por outro lado, na base Bank, o desempenho é moderado, com taxas de acerto em torno de 0.45. Esse comportamento pode ser atribuído à alta cardinalidade e ao desbalanceamento das classes, que dificultam o processo de generalização do modelo. Ainda assim, observa-se estabilidade nos resultados, evidenciada pelos baixos valores de desvio padrão.

Na base Student, observa-se o cenário mais desafiador, com taxas de acerto da imputação significativamente inferiores. Esse resultado reflete a elevada complexidade estrutural da base, caracterizada por múltiplas variáveis categóricas com alta cardinalidade e relações não triviais entre atributos. Nesse contexto, mesmo a capacidade do MLP de modelar relações não lineares mostra-se limitada.

De modo geral, verifica-se uma leve degradação no desempenho à medida que a taxa de valores ausentes aumenta, indicando que a redução da informação disponível impacta diretamente o processo de aprendizado do modelo. Esse efeito é mais pronunciado nas bases mais complexas, como Student e Bank.

Em síntese, os resultados do cenário Misto indicam que o MLP apresenta boa capacidade de adaptação a dados heterogêneos, superando limitações observadas em métodos baseados em distância. No entanto, seu desempenho permanece dependente da qualidade das relações entre atributos e é sensível a fatores como cardinalidade e desbalanceamento das classes.

V.3.0.3 Cenário Categórico - Taxa de Acerto da Imputação

A Figura V.9 e a Tabela V.10 apresentam o desempenho do MLP nas bases do cenário Categórico, considerando a taxa de acerto da imputação ao longo das diferentes execuções.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, permitindo avaliar a consistência do modelo frente a diferentes instâncias de ausência de dados.

Tabela V.10: Taxa média de acerto da imputação do MLP no cenário Categórico (média ± desvio padrão das seeds).

Base	**MLP**
Balance	0.86 ± 0.05
Nursery	0.99 ± 0.00
Car	0.96 ± 0.02

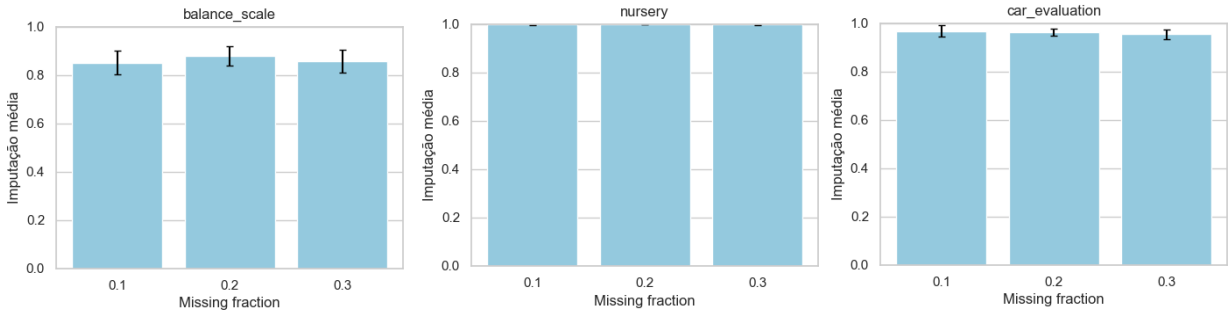


Figura V.9: Desempenho do MLP nas bases do cenário Categórico

Observa-se que, no cenário Categórico, o MLP apresenta desempenho elevado e, em geral, bastante consistente entre as diferentes bases, evidenciando sua capacidade de modelar relações entre atributos puramente qualitativos.

Na base Nursery, o modelo atinge desempenho praticamente perfeito, com taxas de acerto próximas de 1.00 e variabilidade nula entre as execuções. Esse resultado indica que existem padrões muito fortes de coocorrência entre as variáveis categóricas, os quais são facilmente capturados pelo MLP, mesmo com a presença de valores ausentes.

De forma semelhante, a base Car apresenta taxas de acerto elevadas, em torno de 0.95, com baixa variabilidade. Esse comportamento sugere que, embora haja alguma perda de desempenho com o aumento da taxa de valores ausentes, o modelo é capaz de preservar de forma robusta as relações estruturais entre os atributos.

Na base Balance, observa-se desempenho também elevado, porém com maior variabilidade quando comparado às demais bases do cenário. Esse resultado pode ser atribuído à menor quantidade de atributos disponíveis, o que reduz a quantidade de informação útil para a imputação e torna o modelo mais sensível às variações introduzidas pelas diferentes execuções.

De modo geral, verifica-se uma leve degradação no desempenho à medida que a taxa de valores ausentes aumenta, embora esse efeito seja menos pronunciado do que nos cenários anteriores. Esse comportamento indica que, quando existem relações consistentes entre variáveis categóricas, o MLP consegue manter um desempenho estável mesmo com redução da informação disponível.

Em síntese, os resultados do cenário Categórico indicam que o MLP é altamente eficaz na imputação de dados puramente categóricos, sendo capaz de capturar padrões complexos de dependência entre atributos. Ainda assim, fatores como a quantidade de atributos e a riqueza estrutural dos dados continuam desempenhando papel relevante no desempenho do modelo.

V.3.0.4 Cenário Simples - Desempenho de Classificadores

A Figura V.10 e a Tabela V.11 apresentam o desempenho dos classificadores treinados sobre os dados imputados pelo MLP no cenário Simples, permitindo avaliar em que medida a imputação preserva a estrutura informacional necessária para tarefas de aprendizado supervisionado.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, evidenciando a estabilidade dos resultados frente a diferentes instâncias de geração de valores ausentes.

Tabela V.11: Desempenho médio de classificadores treinados sobre dados imputados pelo MLP no cenário Simples (média ± desvio padrão das seeds).

Base	**Original**	**MLP**
Wine	0.96 ± 0.03	0.96 ± 0.03
Image	0.79 ± 0.01	0.78 ± 0.01
Pen-Based	0.99 ± 0.00	0.99 ± 0.00

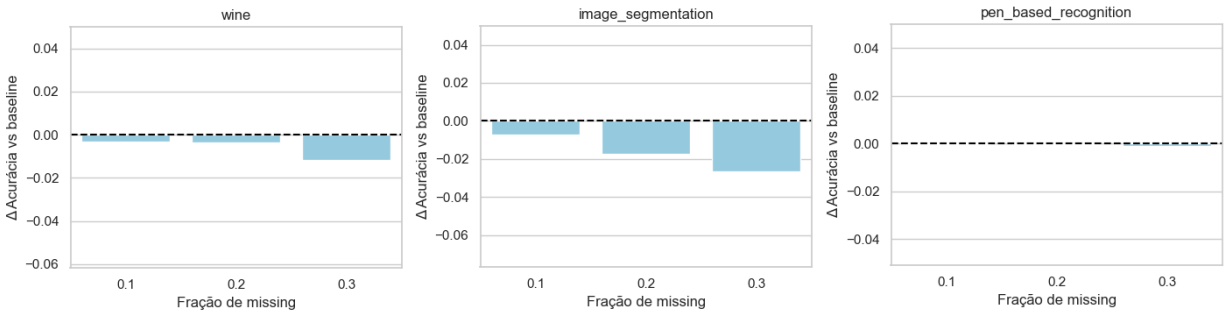


Figura V.10: Desempenho dos classificadores MLP nas bases do cenário Simples

De modo geral, observa-se que, no cenário Simples, o MLP é capaz de preservar adequadamente o desempenho dos classificadores, especialmente em bases com estrutura bem definida e elevada separabilidade entre classes.

Na base Wine, o desempenho dos classificadores treinados sobre dados imputados permanece praticamente idêntico ao obtido com os dados originais. Esse resultado indica que o MLP consegue manter as relações relevantes entre os atributos, mesmo após o processo de imputação, o que está alinhado com a boa taxa de acerto da imputação observada anteriormente.

De forma semelhante, na base Pen-Based Recognition, o desempenho dos classificadores permanece inalterado, com taxas de acerto próximas de 0.99 em todas as configurações. Esse comportamento reforça que a estrutura dos dados é altamente robusta e que o MLP é capaz de preservar integralmente a informação necessária para a tarefa de classificação.

Por outro lado, na base Image Segmentation, observa-se uma leve redução no desempenho dos classificadores após a imputação, acompanhando a menor taxa de acerto da imputação observada nesse cenário. Ainda assim, essa redução é relativamente pequena, indicando que, mesmo quando não reconstrói perfeitamente os valores ausentes, o MLP consegue preservar parcialmente a estrutura informacional dos dados.

Outro aspecto relevante refere-se à proximidade entre os resultados obtidos com dados originais e imputados. Em contraste com métodos baseados em distância, observa-se que o MLP tende a introduzir menor degradação no desempenho dos classificadores, evidenciando sua capacidade de aprender representações mais robustas dos dados.

De forma geral, os resultados indicam que, no cenário Simples, o MLP não apenas apresenta bom desempenho de imputação, mas também preserva de maneira eficaz a qualidade dos dados para tarefas de aprendizado supervisionado, sendo particularmente eficiente em bases com forte relação entre atributos numéricos e a variável categórica.

V.3.0.5 Cenário Misto - Desempenho de Classificadores

A Figura V.11 e a Tabela V.12 apresentam o desempenho dos classificadores treinados sobre os dados imputados pelo MLP no cenário Misto, permitindo avaliar o impacto da imputação na preservação da estrutura informacional dos dados.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, evidenciando a estabilidade dos resultados frente a diferentes instâncias de geração de valores ausentes.

Tabela V.12: Desempenho médio de classificadores treinados sobre dados imputados pelo MLP no cenário Misto (média ± desvio padrão das seeds).

Base	Original	MLP
Student	0.14 ± 0.02	0.15 ± 0.02
Obesity	0.77 ± 0.01	0.77 ± 0.01
Bank	0.39 ± 0.00	0.41 ± 0.00

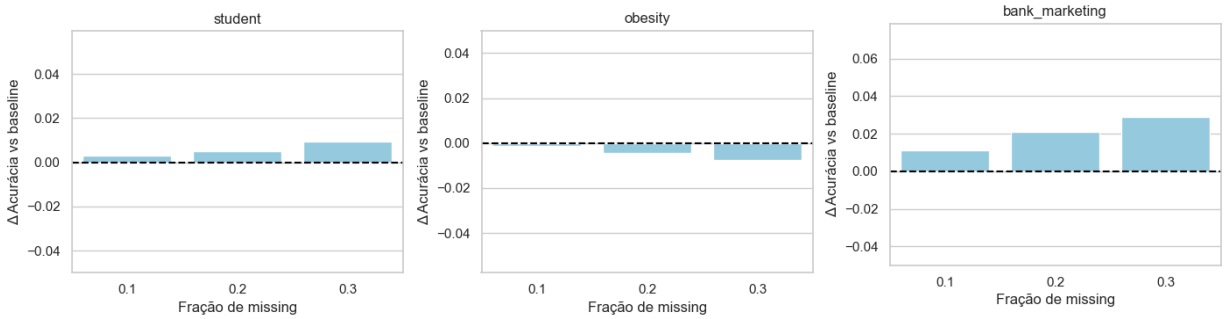


Figura V.11: Desempenho dos classificadores MLP nas bases do cenário Simples

De modo geral, observa-se que, no cenário Misto, o impacto da imputação sobre o desempenho dos classificadores é relativamente moderado, mesmo em bases onde a taxa de acerto da imputação não é elevada.

Na base Obesity, o desempenho dos classificadores treinados sobre dados imputados permanece praticamente idêntico ao obtido com os dados originais. Esse resultado indica que o MLP consegue preservar de forma eficaz a estrutura informacional dos dados, mesmo diante da presença de variáveis heterogêneas.

Na base Bank, observa-se um comportamento particularmente interessante, no qual o desempenho dos classificadores treinados sobre dados imputados supera ligeiramente o obtido com os dados originais. Esse resultado sugere que o MLP pode atuar como um mecanismo de regularização, suavizando ruídos presentes nos dados e favorecendo o aprendizado do modelo.

Por outro lado, na base Student, o desempenho permanece baixo tanto para os dados originais quanto para os imputados, refletindo a elevada complexidade estrutural da base. Ainda assim, observa-se uma leve melhoria após a imputação, indicando que o MLP é capaz de preservar parcialmente as relações relevantes entre os atributos, mesmo em cenários mais desafiadores.

Outro aspecto relevante refere-se à discrepância entre a taxa de acerto da imputação e o desempenho dos classificadores. Em particular, observa-se que, mesmo quando a imputação apresenta desempenho limitado, o impacto sobre o aprendizado supervisionado pode ser reduzido. Esse comportamento evidencia que a preservação da estrutura global dos dados é mais relevante do que a reconstrução exata dos valores ausentes.

De forma geral, os resultados indicam que, no cenário Misto, o MLP apresenta boa capacidade de preservar a qualidade dos dados para tarefas de classificação, mesmo em contextos de maior complexidade estrutural. Esse comportamento reforça a robustez do método em cenários heterogêneos e destaca sua vantagem em relação a abordagens baseadas exclusivamente em estrutura local.

V.3.0.6 Cenário Categórico - Desempenho de Classificadores

A Figura V.12 e a Tabela V.13 apresentam o desempenho dos classificadores treinados sobre os dados imputados pelo MLP no cenário Categórico, permitindo avaliar em que medida a imputação preserva a estrutura informacional necessária para tarefas de aprendizado supervisionado.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, evidenciando a estabilidade do método frente a diferentes configurações de ausência de dados.

Tabela V.13: Desempenho médio de classificadores treinados sobre dados imputados pelo MLP no cenário Categórico (média ± desvio padrão das seeds).

Base	**Original**	**MLP**
Balance	0.90 ± 0.03	0.89 ± 0.03
Nursery	0.97 ± 0.01	0.96 ± 0.01
Car	0.94 ± 0.02	0.93 ± 0.02

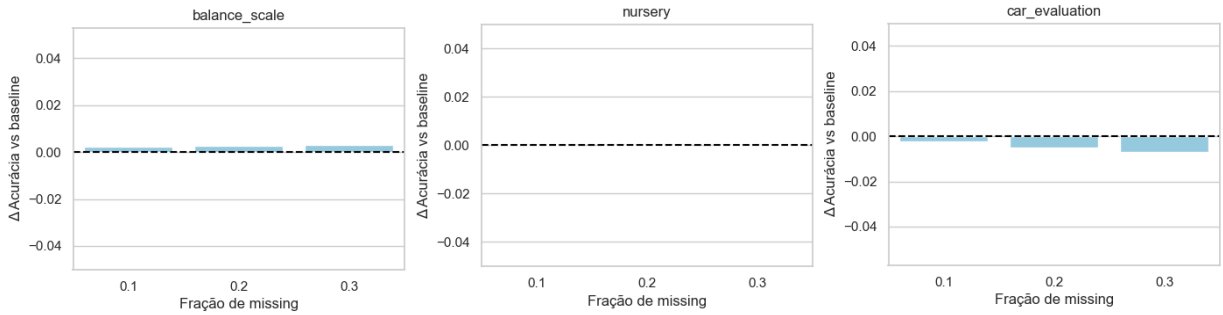


Figura V.12: Desempenho dos classificadores MLP nas bases do cenário Simples

A análise do desempenho dos classificadores no cenário Categórico indica que o MLP é capaz de preservar, de forma consistente, a estrutura informacional dos dados, mesmo em contextos compostos exclusivamente por variáveis qualitativas.

De modo geral, observa-se que o desempenho dos classificadores treinados sobre dados imputados permanece bastante próximo ao obtido com os dados originais, com reduções pouco expressivas nas três bases analisadas. Esse comportamento sugere que, embora a imputação não seja perfeita, os padrões de dependência entre atributos categóricos são adequadamente capturados pelo modelo.

Na base Nursery, os resultados evidenciam elevada estabilidade, com desempenho praticamente inalterado após a imputação. Esse resultado está alinhado com a alta taxa de acerto da imputação observada anteriormente, indicando que o MLP é capaz de aprender de forma eficaz as relações entre atributos categóricos, mesmo após a transformação via codificação *one-hot*.

De forma semelhante, na base Car, observa-se uma pequena redução no desempenho dos classificadores, porém mantendo valores elevados e próximos ao cenário original. Esse comportamento indica que o modelo consegue preservar as relações mais relevantes para a tarefa de classificação, mesmo diante da introdução de valores imputados.

Na base Balance, a diferença entre o desempenho original e o obtido após a imputação é ligeiramente mais perceptível, embora ainda moderada. Esse resultado pode ser associado à menor quantidade de atributos disponíveis, o que reduz a capacidade do modelo de capturar padrões mais robustos, tornando a imputação mais sensível a variações nos dados.

Outro aspecto relevante refere-se à baixa variabilidade dos resultados, evidenciada pelos desvios padrão reduzidos. Esse comportamento indica que o MLP apresenta boa estabilidade mesmo em cenários com diferentes instâncias de ausência de dados, beneficiando-se do uso de técnicas de regularização como o *early stopping*.

De forma geral, os resultados do cenário Categórico indicam que o MLP consegue manter o desempenho dos classificadores em níveis próximos aos dados completos, evidenciando sua capacidade de modelar relações complexas entre variáveis categóricas. Esses achados reforçam que, mesmo em ausência de atributos numéricos, métodos baseados em redes neurais podem produzir imputações de alta qualidade, preservando de maneira eficaz a estrutura necessária para tarefas de aprendizado supervisionado.

V.4 Imputação via Random Forest

Os resultados apresentados nesta seção foram obtidos a partir de múltiplas execuções dos experimentos, considerando diferentes sementes aleatórias, de forma análoga ao procedimento adotado para os métodos anteriores. Para cada base de dados, foram utilizadas 30 sementes distintas, combinadas com três níveis de geração de valores ausentes (10%, 20% e 30%), totalizando 90 execuções

por base.

Em cada execução, tanto a divisão entre conjuntos de treino e teste quanto a seleção dos valores ausentes sob o mecanismo MCAR foram controladas pela semente aleatória. Essa estratégia permite avaliar não apenas o desempenho médio do método, mas também sua estabilidade frente a diferentes configurações dos dados.

Os resultados são apresentados na forma de média acompanhada do desvio padrão, sendo este último indicativo da variabilidade entre as execuções. Essa análise é particularmente relevante no contexto de métodos baseados em árvores, que, embora mais estáveis que modelos como redes neurais, ainda incorporam aleatoriedade em seu processo de construção.

A imputação baseada em Random Forest (RF) foi conduzida tratando o problema como uma tarefa de classificação supervisionada, em que o atributo categórico com valores ausentes é predito a partir das demais variáveis observadas. O modelo é composto por um conjunto de árvores de decisão construídas a partir de subconjuntos aleatórios dos dados e dos atributos, sendo a predição final obtida por meio de votação majoritária.

Diferentemente dos métodos baseados em vizinhança, o Random Forest constrói um modelo global e não linear dos dados, sendo capaz de capturar interações complexas entre atributos sem a necessidade de pré-processamentos intensivos, como normalização. Além disso, sua estrutura baseada em múltiplas árvores contribui para reduzir o risco de sobreajuste e aumentar a robustez do modelo.

V.4.0.1 Cenário Simples - Taxa de Acerto da Imputação

A Figura V.13 e a Tabela V.14 apresentam o desempenho do Random Forest nas bases do cenário Simples, considerando a taxa de acerto da imputação ao longo das diferentes execuções.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, permitindo avaliar a consistência do modelo frente a diferentes instâncias de ausência de dados.

Tabela V.14: Taxa média de acerto da imputação do Random Forest no cenário Simples (média ± desvio padrão das seeds).

Base	**RF**
Wine	0.97 ± 0.03
Image	0.74 ± 0.03
Pen-Based	0.99 ± 0.00

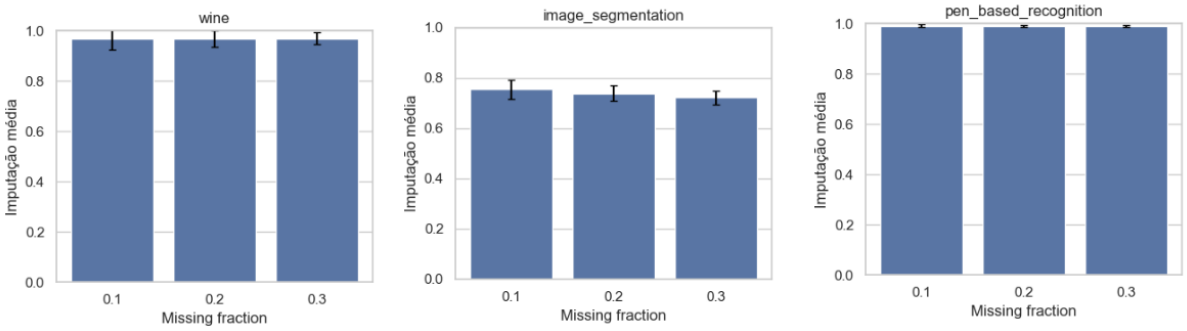


Figura V.13: Desempenho do Random Forest nas bases do cenário Simples

A análise do desempenho do Random Forest no cenário Simples evidencia um comportamento bastante consistente e competitivo, alinhado à capacidade do modelo de capturar relações não lineares entre atributos numéricos.

Na base Wine, observa-se desempenho elevado, com taxas de acerto médias próximas de 0.97. Esse resultado indica que o Random Forest é altamente eficaz em modelar as relações entre os atributos contínuos e a variável categórica, conseguindo reconstruir com precisão os valores ausentes. Ainda que haja uma leve variação entre as execuções, refletida no desvio padrão, o desempenho geral permanece robusto mesmo com o aumento da taxa de ausência.

De forma semelhante, a base Pen-Based Recognition apresenta desempenho praticamente perfeito, com taxas de acerto próximas de 0.99 e variabilidade mínima. Esse comportamento sugere que a estrutura dos dados é altamente favorável à modelagem por árvores de decisão, com padrões bem definidos e elevada separabilidade entre classes. Nesse contexto, o Random Forest consegue generalizar de forma extremamente eficaz.

Por outro lado, a base Image Segmentation apresenta desempenho inferior, com taxa de acerto média em torno de 0.74. Esse resultado evidencia maior complexidade na relação entre os atributos e a variável alvo, tornando o processo de imputação mais desafiador. Observa-se também uma maior sensibilidade à perda de informação, com degradação do desempenho à medida que a taxa de valores ausentes aumenta.

De modo geral, os resultados indicam que, no cenário Simples, o Random Forest apresenta desempenho elevado e consistente, sendo particularmente eficaz em bases com boa separabilidade entre classes. Sua principal vantagem está na capacidade de modelar relações não lineares de forma robusta, enquanto sua limitação se manifesta em cenários com maior complexidade estrutural ou menor quantidade de informação disponível para treinamento.

V.4.0.2 Cenário Misto - Taxa de Acerto da Imputação

A Figura V.14 e a Tabela V.15 apresentam o desempenho do Random Forest nas bases do cenário Misto, considerando a taxa de acerto da imputação ao longo das diferentes execuções.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, permitindo avaliar a consistência do modelo frente a diferentes instâncias de ausência de dados.

Tabela V.15: Taxa média de acerto da imputação do Random Forest no cenário Misto (média ± desvio padrão das seeds).

Base	RF
Student	0.37 ± 0.06
Obesity	0.93 ± 0.02
Bank	0.43 ± 0.01

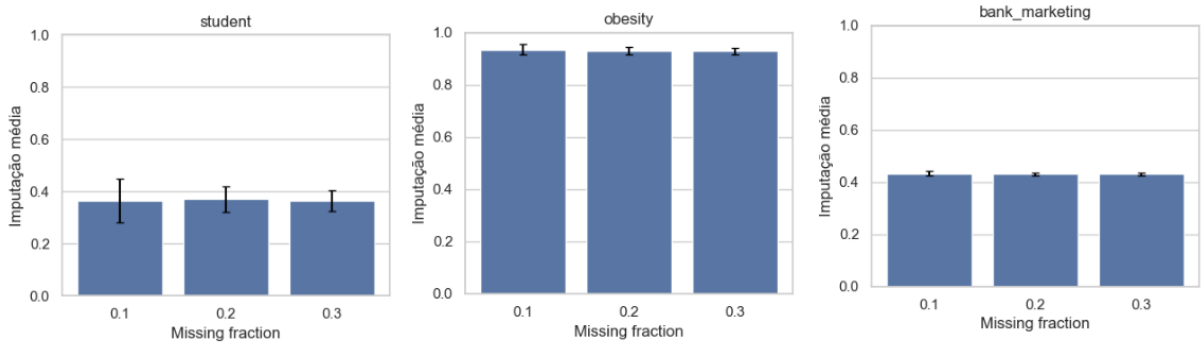


Figura V.14: Desempenho do Random Forest nas bases do cenário Misto

Observa-se que, no cenário Misto, o desempenho do Random Forest apresenta variação significativa entre as bases, refletindo a complexidade associada à presença simultânea de variáveis numéricas e categóricas.

Na base Obesity, o modelo apresenta desempenho elevado, com taxa de acerto média em torno de 0.93 e baixa variabilidade entre as execuções. Esse resultado indica que o Random Forest é altamente eficaz em capturar relações entre atributos heterogêneos, explorando de forma eficiente tanto variáveis numéricas quanto categóricas no processo de imputação. A estrutura baseada em árvores favorece a modelagem de interações complexas, contribuindo para o alto desempenho observado.

Por outro lado, na base Bank, o desempenho é moderado, com taxas de acerto próximas de 0.43. Esse comportamento pode ser atribuído à maior complexidade estrutural da base, incluindo fatores como alta cardinalidade de atributos categóricos e possível desbalanceamento entre classes. Ainda assim, observa-se baixa variabilidade nos resultados, indicando boa estabilidade do método.

Na base Student, observa-se o cenário mais desafiador, com taxa de acerto média de aproximadamente 0.37 e maior variabilidade entre as execuções. Esse resultado sugere que o Random Forest encontra dificuldades em capturar padrões consistentes nessa base, possivelmente devido à presença

de múltiplas variáveis categóricas com relações não lineares e menos evidentes. A maior dispersão dos resultados indica também maior sensibilidade à configuração dos dados e à geração dos valores ausentes.

De modo geral, verifica-se uma tendência de leve degradação no desempenho à medida que a taxa de valores ausentes aumenta, refletindo a redução da informação disponível para o treinamento das árvores. Esse efeito é mais pronunciado nas bases mais complexas, como Student e Bank.

Em síntese, os resultados do cenário Misto indicam que o Random Forest apresenta elevada capacidade de adaptação a dados heterogêneos, sendo particularmente eficaz em bases com relações bem definidas entre atributos. No entanto, seu desempenho pode ser impactado por fatores como alta cardinalidade, desbalanceamento e complexidade estrutural, especialmente em cenários onde os padrões são menos evidentes.

V.4.0.3 Cenário Categórico - Taxa de Acerto da Imputação

A Figura V.15 e a Tabela V.16 apresentam o desempenho do Random Forest nas bases do cenário Categórico, considerando a taxa de acerto da imputação ao longo das diferentes execuções.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, permitindo avaliar a consistência do modelo frente a diferentes instâncias de ausência de dados.

Tabela V.16: Taxa média de acerto da imputação do Random Forest no cenário Categórico (média ± desvio padrão das seeds).

Base	RF
Balance	0.88 ± 0.04
Nursery	1.00 ± 0.00
Car	0.97 ± 0.01

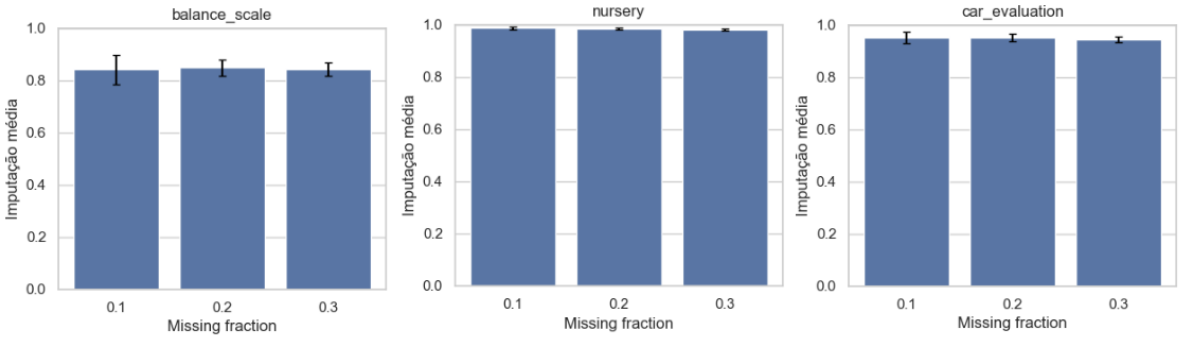


Figura V.15: Desempenho do Random Forest nas bases do cenário Categórico

Observa-se que, no cenário Categórico, o Random Forest apresenta desempenho elevado e consistente entre as diferentes bases, evidenciando sua capacidade de modelar relações entre atributos puramente qualitativos.

Na base Nursery, o modelo atinge desempenho praticamente perfeito, com taxa de acerto média igual a 1.00 e variabilidade nula entre as execuções. Esse resultado indica que existem padrões extremamente fortes de dependência entre os atributos categóricos, os quais são facilmente capturados pelo conjunto de árvores de decisão. Nesse contexto, o Random Forest consegue reconstruir os valores ausentes com alta precisão, independentemente da taxa de ausência considerada.

De forma semelhante, a base Car apresenta taxa de acerto elevada, em torno de 0.97, com baixa variabilidade. Esse comportamento sugere que, embora haja alguma perda de desempenho com o aumento da taxa de valores ausentes, o modelo é capaz de preservar de forma robusta as relações estruturais entre os atributos categóricos.

Na base Balance, observa-se desempenho também elevado, porém com maior variabilidade quando comparado às demais bases do cenário. Esse resultado pode ser atribuído à menor quantidade de atributos disponíveis, o que reduz a redundância informacional e torna o processo de imputação mais sensível às variações introduzidas pelas diferentes execuções.

De modo geral, verifica-se uma leve degradação no desempenho à medida que a taxa de valores ausentes aumenta, embora esse efeito seja menos pronunciado em comparação aos cenários Simples e Misto. Esse comportamento indica que, quando existem padrões consistentes entre variáveis categóricas, o Random Forest consegue manter desempenho elevado mesmo com redução da informação disponível.

Em síntese, os resultados do cenário Categórico indicam que o Random Forest é altamente eficaz na imputação de dados puramente categóricos, sendo capaz de capturar padrões complexos de dependência entre atributos. Sua estrutura baseada em múltiplas árvores favorece a robustez e a estabilidade dos resultados, tornando-o uma abordagem particularmente adequada para esse tipo de cenário.

V.4.0.4 Cenário Simples - Desempenho de Classificadores

A Figura V.16 e a Tabela V.17 apresentam o desempenho dos classificadores treinados sobre os dados imputados pelo Random Forest no cenário Simples, permitindo avaliar em que medida a imputação preserva a estrutura informacional necessária para tarefas de aprendizado supervisionado.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, evidenciando a estabilidade dos resultados frente a diferentes instâncias de geração de valores ausentes.

Tabela V.17: Desempenho médio de classificadores treinados sobre dados imputados pelo Random Forest no cenário Simples (média ± desvio padrão das seeds).

Base	**Original**	**RF**
Wine	0.96 ± 0.03	0.96 ± 0.03
Image	0.79 ± 0.01	0.78 ± 0.01
Pen-Based	0.99 ± 0.00	0.99 ± 0.00

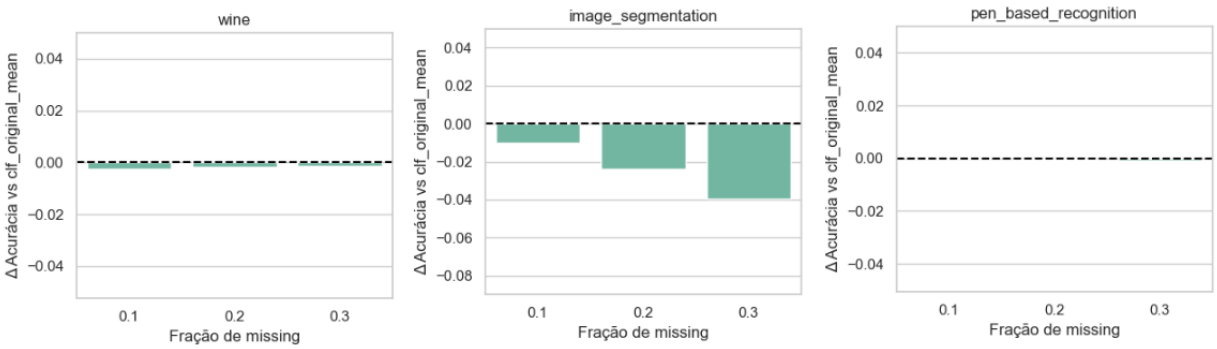


Figura V.16: Desempenho dos classificadores RF nas bases do cenário Simples

De modo geral, observa-se que, no cenário Simples, o Random Forest é capaz de preservar de forma eficaz o desempenho dos classificadores, mantendo resultados bastante próximos aos obtidos com os dados originais.

Na base Wine, o desempenho dos classificadores treinados sobre dados imputados permanece praticamente idêntico ao cenário original. Esse resultado está alinhado com a elevada taxa de acerto da imputação observada anteriormente, indicando que o Random Forest consegue reconstruir de forma consistente as relações entre os atributos e a variável alvo, sem comprometer a qualidade dos dados para tarefas de classificação.

De forma semelhante, na base Pen-Based Recognition, o desempenho permanece inalterado, com taxas de acerto próximas de 0.99 em todas as execuções. Esse comportamento reforça a ideia de que a estrutura dos dados é altamente robusta e que o Random Forest é capaz de preservar integralmente a informação relevante, mesmo após o processo de imputação.

Por outro lado, na base Image Segmentation, observa-se uma leve redução no desempenho dos classificadores após a imputação, acompanhando a menor taxa de acerto da imputação observada nesse cenário. Ainda assim, essa redução é pouco expressiva, indicando que, mesmo quando não reconstrói perfeitamente os valores ausentes, o Random Forest consegue preservar a maior parte da estrutura informacional dos dados.

Outro aspecto relevante refere-se à estabilidade dos resultados. Os baixos valores de desvio padrão indicam que o Random Forest apresenta comportamento consistente frente a diferentes configurações de ausência de dados, beneficiando-se de sua natureza baseada em múltiplas árvores e amostragem aleatória.

De forma geral, os resultados indicam que, no cenário Simples, o Random Forest não apenas apresenta elevada taxa de acerto da imputação, mas também preserva de maneira eficaz a qualidade dos dados para tarefas de aprendizado supervisionado, sendo particularmente eficiente em bases com forte relação entre atributos numéricos e a variável categórica.

V.4.0.5 Cenário Misto - Desempenho de Classificadores

A Figura V.17 e a Tabela V.18 apresentam o desempenho dos classificadores treinados sobre os dados imputados pelo Random Forest no cenário Misto, permitindo avaliar o impacto da imputação na preservação da estrutura informacional dos dados.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, evidenciando a estabilidade dos resultados frente a diferentes instâncias de geração de valores ausentes.

Tabela V.18: Desempenho médio de classificadores treinados sobre dados imputados pelo Random Forest no cenário Misto (média ± desvio padrão das seeds).

Base	Original	RF
Student	0.14 ± 0.02	0.18 ± 0.02
Obesity	0.77 ± 0.01	0.78 ± 0.01
Bank	0.39 ± 0.00	0.42 ± 0.00

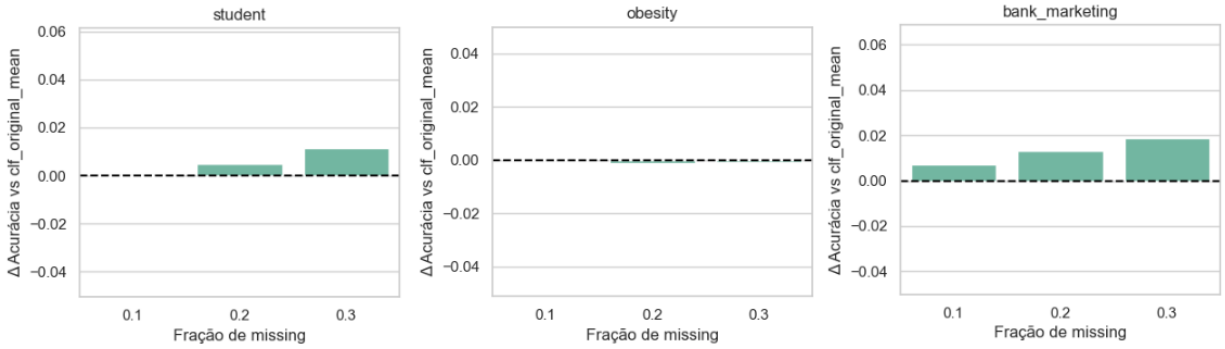


Figura V.17: Desempenho dos classificadores RF nas bases do cenário Misto

De modo geral, observa-se que, no cenário Misto, o Random Forest apresenta impacto positivo ou neutro no desempenho dos classificadores, mesmo em bases onde a taxa de acerto da imputação não é elevada.

Na base Obesity, o desempenho dos classificadores treinados sobre dados imputados é praticamente idêntico ao obtido com os dados originais, com uma leve melhoria. Esse resultado indica que o Random Forest consegue preservar de forma eficaz a estrutura informacional dos dados, mesmo diante da presença de variáveis heterogêneas.

Na base Bank, observa-se uma melhoria mais evidente no desempenho após a imputação, com acurácias superiores às obtidas com os dados originais. Esse comportamento sugere que o Random Forest pode atuar como um mecanismo de regularização, suavizando ruídos presentes nos dados e favorecendo o processo de aprendizado dos classificadores.

Na base Student, embora o desempenho absoluto permaneça baixo, observa-se um ganho em relação ao cenário original. Esse resultado é particularmente relevante, pois indica que, mesmo em cenários mais desafiadores e com baixa taxa de acerto da imputação, o Random Forest é capaz

de preservar ou até melhorar parcialmente a estrutura informacional relevante para a tarefa de classificação.

Outro aspecto importante refere-se à discrepância entre a taxa de acerto da imputação e o desempenho dos classificadores. Assim como observado para o MLP, verifica-se que uma imputação com desempenho moderado não implica necessariamente em pior desempenho em tarefas supervisionadas. Isso ocorre porque a preservação da estrutura global dos dados pode ser mais relevante do que a reconstrução exata dos valores ausentes.

Além disso, a baixa variabilidade dos resultados, evidenciada pelos reduzidos desvios padrão, reforça a estabilidade do Random Forest frente a diferentes instâncias de geração de valores ausentes.

De forma geral, os resultados indicam que, no cenário Misto, o Random Forest apresenta elevada robustez, sendo capaz de preservar, e em alguns casos melhorar, o desempenho dos classificadores, mesmo em contextos de maior complexidade estrutural.

V.4.0.6 Cenário Categórico - Desempenho de Classificadores

A Figura V.18 e a Tabela V.19 apresentam o desempenho dos classificadores treinados sobre os dados imputados pelo Random Forest no cenário Categórico, permitindo avaliar em que medida a imputação preserva a estrutura informacional necessária para tarefas de aprendizado supervisionado.

O desvio padrão informado após o símbolo ± reflete a variabilidade entre as execuções, evidenciando a estabilidade do método frente a diferentes configurações de ausência de dados.

Tabela V.19: Desempenho médio de classificadores treinados sobre dados imputados pelo Random Forest no cenário Categórico (média ± desvio padrão das seeds).

Base	Original	RF
Balance	0.90 ± 0.03	0.90 ± 0.03
Nursery	0.97 ± 0.01	0.97 ± 0.01
Car	0.94 ± 0.02	0.94 ± 0.02

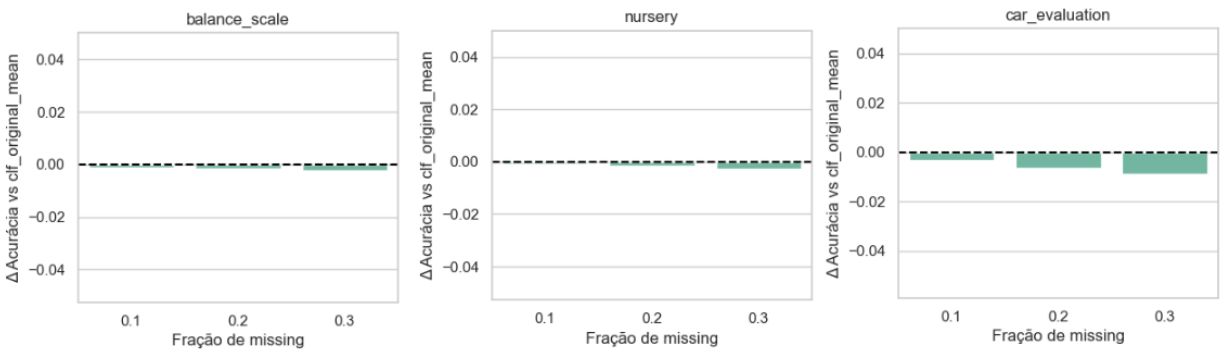


Figura V.18: Desempenho dos classificadores RF nas bases do cenário Categórico

A análise do desempenho dos classificadores no cenário Categórico indica que o Random Forest é capaz de preservar de forma consistente a estrutura informacional dos dados, mesmo em contextos compostos exclusivamente por variáveis qualitativas.

De modo geral, observa-se que o desempenho dos classificadores treinados sobre dados imputados permanece praticamente idêntico ao obtido com os dados originais nas três bases analisadas. Esse comportamento sugere que, embora a imputação não seja perfeita em todos os casos, os padrões de dependência entre atributos categóricos são adequadamente preservados pelo modelo.

Na base Nursery, os resultados evidenciam elevada estabilidade, com desempenho inalterado após a imputação. Esse resultado está diretamente relacionado à alta taxa de acerto da imputação observada anteriormente, indicando que o Random Forest consegue capturar com precisão as relações entre atributos categóricos e preservar integralmente a informação relevante para a tarefa de classificação.

De forma semelhante, na base Car, não se observa perda de desempenho após a imputação, mantendo-se acurácias elevadas e consistentes entre as execuções. Esse comportamento reforça a capacidade do Random Forest de modelar interações entre variáveis categóricas de maneira robusta.

Na base Balance, o desempenho também permanece estável, com valores praticamente idênticos ao cenário original. Ainda que essa base apresente maior variabilidade na taxa de acerto da imputação, esse efeito não se traduz em perda de desempenho dos classificadores, indicando que as relações mais relevantes para a tarefa supervisionada são preservadas.

Outro aspecto importante refere-se à baixa variabilidade dos resultados, evidenciada pelos reduzidos desvios padrão. Esse comportamento reforça a robustez do Random Forest, que, por sua natureza baseada em múltiplas árvores e amostragem aleatória, tende a produzir resultados estáveis mesmo em diferentes configurações de ausência de dados.

De forma geral, os resultados do cenário Categórico indicam que o Random Forest consegue manter o desempenho dos classificadores em níveis equivalentes aos dados completos, evidenciando sua eficácia na imputação de dados categóricos. Esses achados reforçam que o método é capaz de preservar a estrutura global dos dados, sendo uma abordagem confiável para cenários compostos exclusivamente por variáveis qualitativas.

V.5 Discussão Comparativa dos Métodos de Imputação

Com o objetivo de sintetizar os resultados obtidos nos diferentes cenários experimentais, os Grids V.19 e V.20 apresentam uma comparação consolidada entre os métodos avaliados. Além dos métodos KNN, MLP e Random Forest, foi incluída a imputação pela Moda como baseline, permitindo comparar o desempenho de abordagens mais sofisticadas com uma estratégia simples amplamente utilizada na literatura.

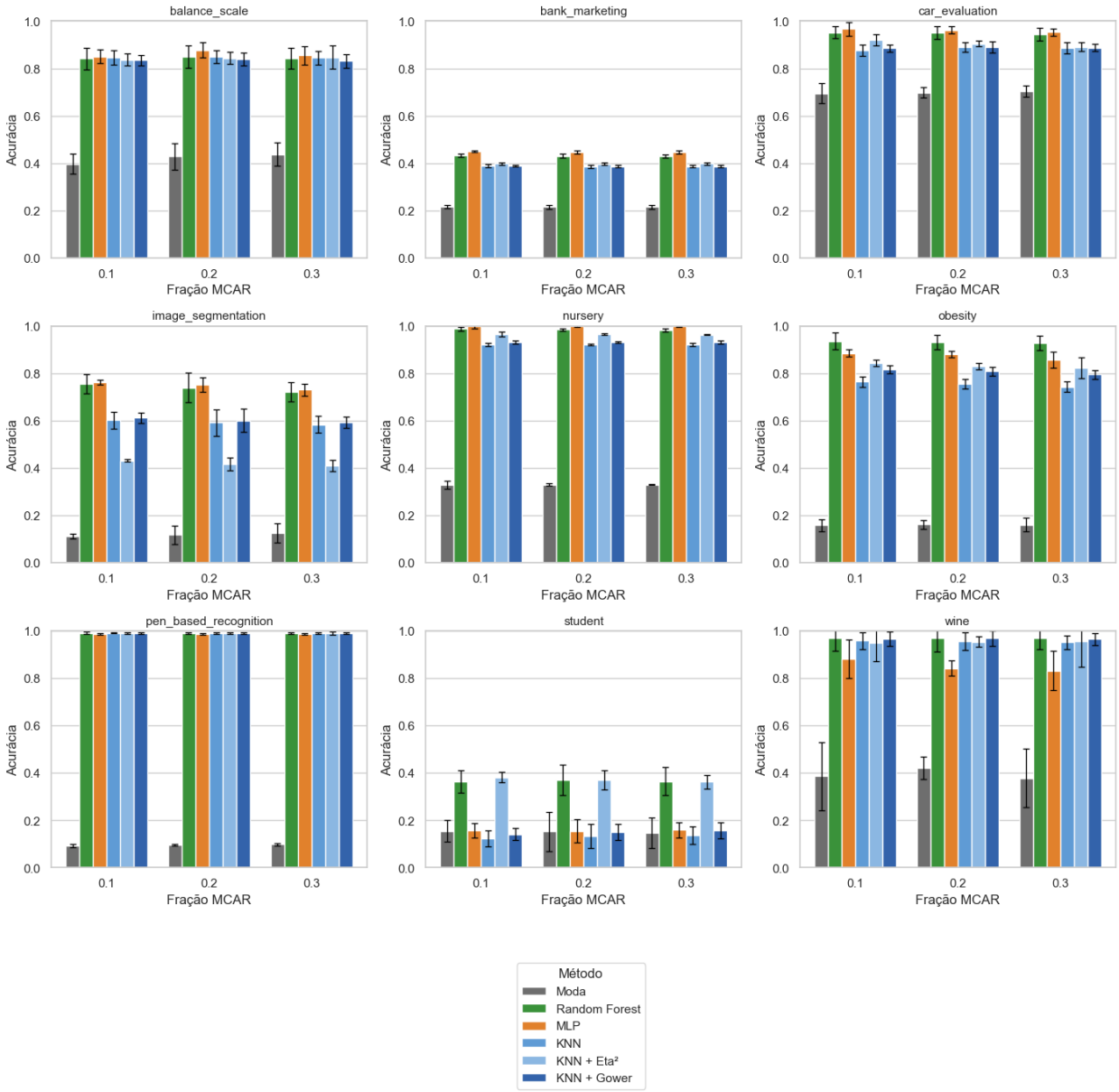


Figura V.19: Comparação global da taxa de acerto da imputação entre Moda, KNN, MLP e Random Forest.

Esta seção apresenta uma análise comparativa dos métodos de imputação avaliados Moda, KNN, MLP e Random Forest, considerando tanto a taxa de acerto da imputação quanto o desempenho dos classificadores treinados sobre os dados imputados. A utilização conjunta desses dois critérios permite analisar não apenas a capacidade de reconstrução dos valores ausentes, mas também o impacto prático da imputação em uma tarefa supervisionada.

De forma geral, observa-se que o aumento da taxa de valores ausentes está associado à redução do desempenho de todos os métodos avaliados, embora em intensidades distintas. Esse comportamento decorre da diminuição da informação disponível para inferência, afetando tanto a qualidade da imputação quanto o desempenho dos modelos de classificação.

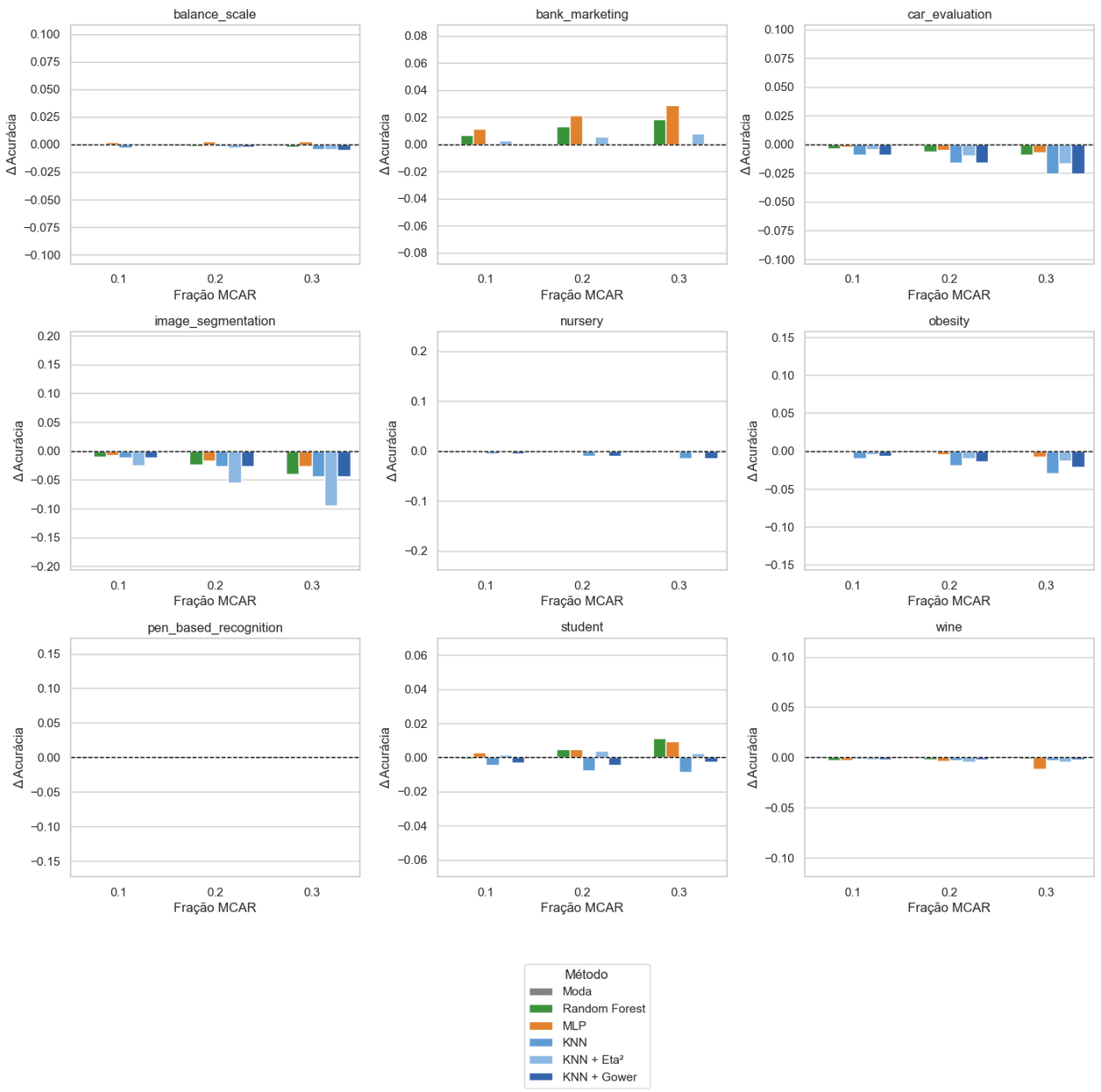


Figura V.20: Comparação global do desempenho dos classificadores treinados sobre dados imputados pelos métodos avaliados.

Como esperado, a imputação pela Moda apresentou, em geral, as menores taxas de acerto da reconstrução dos valores ausentes, por desconsiderar qualquer relação entre os atributos e substituir todos os valores faltantes pela categoria mais frequente. Ainda assim, os resultados mostram que, em diversas bases, essa estratégia produziu classificadores com desempenho relativamente próximo ao obtido com os dados completos. Esse comportamento evidencia que uma baixa taxa de acerto da imputação nem sempre implica uma degradação proporcional do desempenho preditivo, reforçando a importância de avaliar ambos os aspectos de forma complementar.

Entre os métodos avaliados, as abordagens baseadas em KNN apresentaram maior sensibilidade ao aumento da taxa de ausência. Como dependem diretamente da similaridade entre instâncias,

a remoção de valores reduz a confiabilidade da definição de vizinhança, tornando o processo de imputação progressivamente mais difícil à medida que aumenta a proporção de dados ausentes.

Em contraste, os métodos baseados em modelos, representados pelo Random Forest e pelo MLP, mostraram maior estabilidade ao longo dos diferentes níveis de ausência. Nos experimentos realizados, o Random Forest apresentou, de forma consistente, elevada taxa de acerto da imputação e pequena degradação no desempenho dos classificadores, enquanto o MLP apresentou desempenho competitivo, especialmente em bases estruturalmente mais complexas.

A comparação entre os dois grids evidencia um resultado importante: a taxa de acerto da imputação, isoladamente, não constitui um indicador suficiente da qualidade do método. Em diversos cenários, métodos com taxa de acerto moderada preservaram adequadamente a estrutura informacional necessária para manter o desempenho dos classificadores. Da mesma forma, diferenças relativamente grandes na taxa de acerto da imputação nem sempre se refletiram em diferenças equivalentes na etapa de classificação.

Os resultados também sugerem que características estruturais das bases influenciam significativamente o comportamento dos métodos. Neste trabalho, a associação entre atributos categóricos foi caracterizada por meio do coeficiente V de Cramer. Bases que apresentaram maiores níveis de associação entre atributos tenderam a produzir melhores resultados de imputação, indicando que uma maior redundância informacional favorece a inferência dos valores ausentes.

Por outro lado, bases com menor associação entre atributos apresentaram maior dificuldade para todos os métodos avaliados, especialmente aqueles baseados em similaridade, como o KNN. Nesses cenários, a ausência de relações consistentes entre as variáveis reduz a informação disponível para reconstrução dos valores faltantes.

A cardinalidade dos atributos também mostrou associação com o desempenho observado. Em geral, bases contendo atributos com maior número de categorias apresentaram maior dificuldade de imputação. Entretanto, esse comportamento não deve ser atribuído exclusivamente à cardinalidade, uma vez que as bases utilizadas diferem simultaneamente em outras características, como número de atributos, tamanho da amostra, distribuição das classes e domínio do problema. Dessa forma, os resultados sugerem que esses fatores atuam de maneira conjunta, não sendo possível isolar o efeito da cardinalidade a partir dos experimentos realizados.

Essa interação entre múltiplas características estruturais torna particularmente desafiadores os cenários que combinam alta cardinalidade, baixa associação entre atributos e maior complexidade dos dados. Por outro lado, bases com menor cardinalidade e maior associação entre variáveis tenderam a apresentar desempenho mais elevado para praticamente todos os métodos avaliados.

Considerando conjuntamente a qualidade da imputação e a preservação do desempenho dos classificadores, os resultados indicam que o Random Forest apresentou o comportamento mais con-

sistente entre os métodos analisados, enquanto o MLP também apresentou desempenho competitivo em diversos cenários. O KNN mostrou resultados satisfatórios em bases estruturalmente mais simples, porém revelou maior sensibilidade ao aumento da taxa de ausência.

Em síntese, os resultados obtidos sugerem que a escolha do método de imputação deve considerar as características estruturais da base de dados, especialmente o grau de associação entre atributos e a complexidade do problema. A inclusão da imputação pela Moda como baseline reforça essa análise ao fornecer uma referência simples de comparação, permitindo evidenciar os ganhos obtidos pelos métodos mais sofisticados. Embora os experimentos indiquem que fatores estruturais influenciam significativamente o desempenho da imputação, estudos adicionais controlando individualmente essas características são necessários para estabelecer relações causais entre elas e o desempenho observado.

Capítulo VI Considerações Finais

Esta dissertação teve como objetivo investigar como características estruturais de conjuntos de dados categóricos influenciam a dificuldade da imputação de valores ausentes. Em particular, buscou-se compreender o impacto de fatores como número de atributos, cardinalidade das variáveis e grau de associação entre os atributos sobre o desempenho de diferentes métodos de imputação, contribuindo para uma lacuna identificada na literatura, que tradicionalmente concentra esforços na comparação de algoritmos, mas dedica menor atenção à influência das propriedades intrínsecas dos dados sobre a qualidade da imputação.

Para atingir esse objetivo, foi conduzido um estudo experimental baseado em múltiplas bases de dados categóricas e mistas, nas quais valores ausentes foram inseridos artificialmente segundo o mecanismo MCAR em diferentes proporções. Os métodos K-Nearest Neighbors (KNN), Random Forest (RF) e Multilayer Perceptron (MLP) foram avaliados quanto à capacidade de recuperar os valores originais removidos, utilizando múltiplas execuções e diferentes métricas para caracterização das bases e avaliação dos resultados.

Os resultados obtidos demonstraram que o desempenho dos métodos de imputação não depende exclusivamente do algoritmo empregado, mas também das características estruturais dos dados. Observou-se que atributos com elevada cardinalidade tendem a apresentar maior dificuldade de imputação, uma vez que o aumento do número de categorias reduz a probabilidade de identificação correta do valor ausente. De forma semelhante, bases com menor associação entre atributos mostraram-se mais desafiadoras, indicando que a existência de relações informacionais entre variáveis constitui um fator importante para o sucesso dos métodos avaliados.

As análises também evidenciaram que diferentes algoritmos respondem de maneira distinta às características das bases. Enquanto métodos baseados em vizinhança dependem fortemente da similaridade entre instâncias, abordagens baseadas em modelos preditivos demonstram maior capacidade de explorar padrões complexos presentes nos dados. Entretanto, nenhum método apresentou superioridade absoluta em todos os cenários analisados, reforçando a hipótese de que a dificuldade da imputação está relacionada tanto ao algoritmo quanto à estrutura dos dados.

Como contribuição científica, este trabalho propõe uma mudança de perspectiva na análise da imputação categórica, deslocando o foco da simples comparação entre métodos para a investigação dos fatores que tornam determinados problemas de imputação mais ou menos difíceis. Nesse con-

texto, medidas como cardinalidade, número de atributos e associação entre variáveis mostraram-se relevantes para explicar parte da variabilidade observada nos resultados dos algoritmos. Além disso, o estudo fornece evidências empíricas que podem auxiliar pesquisadores e profissionais na escolha de estratégias de imputação mais adequadas às características de seus conjuntos de dados.

Entre as limitações deste trabalho, destaca-se a utilização exclusiva do mecanismo MCAR para geração dos valores ausentes. Embora essa escolha tenha permitido controlar experimentalmente o processo de remoção dos dados e isolar os efeitos das características estruturais das bases, cenários reais frequentemente apresentam mecanismos MAR e MNAR, que introduzem desafios adicionais. Outra limitação refere-se ao conjunto de algoritmos investigados, que, apesar de representativos e amplamente utilizados na literatura, não abrangem todas as abordagens contemporâneas de imputação.

Como trabalhos futuros, sugere-se ampliar a investigação para mecanismos de ausência MAR e MNAR, permitindo avaliar se os padrões observados permanecem válidos em cenários mais próximos de aplicações reais. Também é recomendada a inclusão de novos métodos de imputação, especialmente abordagens baseadas em aprendizado profundo e modelos generativos. Adicionalmente, estudos futuros podem explorar a construção de métricas capazes de estimar previamente a dificuldade de imputação de uma base de dados, auxiliando na seleção automática de métodos e estratégias de tratamento de valores ausentes.

Por fim, conclui-se que a imputação de dados categóricos deve ser compreendida não apenas como um problema algorítmico, mas também como um problema dependente das características estruturais dos dados. Os resultados obtidos reforçam a importância de considerar propriedades como cardinalidade, dimensionalidade e associação entre atributos durante o planejamento de estratégias de imputação, contribuindo para o avanço do entendimento teórico e prático dessa área de pesquisa.

Referências

Akoglu, L. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, 2018.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Chatzipetrou, P. Software cost estimation: A state-of-the-art statistical and visualization approach for missing data. *International Journal of Service Science, Management, Engineering, and Technology*, 10(3):14–29, 2019.

de Abreu Soares, J. *Pré-processamento em Mineração de Dados: Um Estudo Comparativo em Complementação*. Tese de doutorado, COPPE/UFRJ, 2007.

Ferlim, A. O., Santos, P. A., and Carvalho, A. C. P. L. F. Imputação de dados faltantes utilizando regressão em cascata. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 205–210, Fortaleza, CE, Brasil. SBC, 2019.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

Graham, J. *Missing Data: Analysis and Design*. Statistics for Social and Behavioral Sciences. Springer, New York, NY, 2012.

Haykin, S. *Neural Networks and Learning Machines*. Pearson, 3rd edition, 2009.

Jakubik, J., Vössing, M., Kühl, N., Walk, J., and Satzger, G. Data-centric artificial intelligence. *Business & Information Systems Engineering*, 66(4):507–515, 2024.

Kim, J.-M., Lee, K.-J., and Lee, S.-J. Two-stage imputation method to handle missing data for categorical response variable. *Communications for Statistical Applications and Methods*, 30(6):577–587, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Lee, K. J., Carlin, J. B., Simpson, J. A., and Moreno-Betancur, M. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the mcar/mar/mnar classification. *International Journal of Epidemiology*, 52(4):1268 – 1275, 2023.

Little, R. J. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019.

Little, R. J. A. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1st edition, 1987.

Mahfouz, M. A., Shoukry, A., and Ismail, M. A. Eknn: Ensemble classifier incorporating connectivity and density into knn with application to cancer diagnosis. *Artificial Intelligence in Medicine*, 104:101985, 2020.

Malerba, D. and Pasquadibisceglie, V. Data-centric ai. *Journal of Intelligent Information Systems*, 62:1493–1502, 2024.

Memon, S. M., Wamala, R., and Kabano, I. H. A comparison of imputation methods for categorical data. *Informatics in Medicine Unlocked*, 42:101382, 2023.

Rubin, D. B. Biometrika trust inference and missing data, 1976a.

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

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

Schafer, J. L. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, 1997.

Schmitt, S. and Rothlauf, F. Comparison of imputation methods for categorical real-world prostate cancer data with natural order. *Studies in Health Technology and Informatics*, 316:1800–1804, 2024.

Stekhoven, D. J. and Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Suh, H. and Song, J. A comparison of imputation methods using machine learning models. *Communications for Statistical Applications and Methods*, 30(3):331–341, 2023.

Tang, F., Ishwaran, H., and Ghosal, S. A survey on missing data in machine learning: Issues and techniques. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(2):e1443, 2021.

van Buuren, S. *Flexible Imputation of Missing Data*. CRC press, 2018.

Wilson, M. D. and Lueck, K. Working with missing data: Imputation of nonresponse items in categorical survey data with a non-monotone missing pattern. *Journal of Applied Mathematics*, 2014:1–9, 2014.

Wu, S., Feng, X., Han, Y., and Wang, Q. Missing categorical data imputation approach based on similarity. In *Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2827–2832. IEEE, 2012.

Yang, S., Zheng, Y., Lin, Y., and Zhang, X. Imputation of missing values in mixed-type data using ensemble learning. *Scientific Reports*, 15(1):11768, 2025.

Zimmermann, P., Mazouch, P., and Tesárková, K. H. Missing categorical data imputation and individual observation level imputation. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(6):1527–1534, 2014.

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO/DIVULGAÇÃO DE DOCUMENTO ELETRÔNICO NO CATÁLOGO DO SISTEMA DE BIBLIOTECAS DO CEFET/RJ E NO REPOSITÓRIO INSTITUCIONAL

Na qualidade de titular dos direitos de autor da publicação abaixo citada, de acordo com a Lei nº 9610/98, autorizo, a partir da presente data, o CEFET/RJ, a disponibilizar gratuitamente, sem ressarcimento de direitos autorais, conforme permissões assinadas abaixo, no Catálogo Online do Sistema de Bibliotecas e no Repositório Institucional do Cefet/RJ, ou qualquer outro formato de disseminação da informação que a instituição implemente, no formato PDF, para fins de leitura, impressão e/ou download pela internet, a título de divulgação da produção científica gerada por esta Instituição.
Declaro, ainda, que a versão definitiva do trabalho a ser disponibilizada é de inteira responsabilidade do(a) autor(a) e do(a) orientador(a), que atestam sua conformidade com a versão aprovada e autorizam sua divulgação nos meios institucionais do CEFET/RJ, eximindo a Instituição de qualquer responsabilidade quanto ao conteúdo, à integridade ou a eventuais alterações posteriores ao depósito.

1- Tipo de trabalho: [ ] Projeto Final (Graduação) [ ] Especialização
[X] Dissertação [ ] Tese

2- Identificação do trabalho/autor(es):
Curso: Mestrado em Ciência da Computação
Título: ESTUDO DA CARDINALIDADE E DA RELAÇÃO ENTRE ATRIBUTOS NA IMPUTAÇÃO DE DADOS CATEGÓRICOS
Autor(es): Felipe de Ávila Tavares
Orientador: Jorge de Abreu Soares
Co-orientador: N/A

3- Informações de acesso ao trabalho:

Este trabalho é confidencial? [ ] Sim [X] Não
Pode ser liberado para publicação? [X] Total [ ] Parcial [ ] Não
Justifique:_____
_____
Em caso de publicação parcial, assinale as permissões:
[ ] Sumário
[ ] Capítulos. Especifique: _____
[ ] Bibliografia
[ ] Outras partes do trabalho: _____

Rio de Janeiro, 28/09/2026

Documento assinado digitalmente
gov.br
FELIPE DE AVILA TAVARES
Data: 30/09/2026 23:15:59-0300
Verifique em https://validar.iti.gov.br

_____
Assinatura do(s) autor(es)

Documento assinado digitalmente
gov.br
JORGE DE ABREU SOARES
Data: 01/10/2026 09:09:31-0300
Verifique em https://validar.iti.gov.br

_____
Assinatura do orientador

_____
Assinatura do co-orientador