UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

DiT-MAE: Modelo *Transformer Autoencoder* Mascarado para Imputação de Dados Discretos para Questionários Psicométricos Educacionais

Guilherme Mendonça Freire

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Guilherme Mendonça Freire

DiT-MAE: Modelo *Transformer Autoencoder* Mascarado para Imputação de Dados Discretos para Questionários Psicométricos Educacionais

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Mariana Curi

USP – São Carlos
Fevereiro de 2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

M539d

Mendonça Freire, Guilherme
DiT-MAE: Modelo Transformer Autoencoder Mascarado para Imputação de Dados Discretos para Questionários Psicométricos Educacionais / Guilherme Mendonça Freire; orientadora Mariana Curi. -- São Carlos, 2025.
87 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Teoria de Resposta ao Item. 2. Redes Neurais. 3. Psicometria. I. Curi, Mariana, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Guilherme Mendonça Freire

DiT-MAE: Masked Autoencoder Transformer Model for Discrete Data Imputation for Educational Psychometric Questionnaires

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Mariana Curi

USP – São Carlos
February 2025

Este trabalho é dedicado ao direito à educação a todos.
Isto se resolvendo, o futuro é próspero e bem aventurado a todos.
Em especial, a todos que me acompanharam nessa jornada.

AGRADECIMENTOS

Agradeço, principalmente, a Daniele Magnavita de Alencar, minha esposa, Davi Magnavita, meu filho, Ana Maria Ferreira Mendonça Freire, minha mãe, Evandro Sena Freire (*In memoriam*), meu pai, Viviane Mendonça Freire, minha irmã, e Gustavo Mendonça Freire, meu irmão, pelo apoio incondicional desde o começo.

Agradeço, especialmente, a minha orientadora, Professora Dra. Mariana Curi, pelo apoio e paciência, e à Professora Sandra Fabbri, por ter-me ajudado na construção da minha carreira de pesquisador, juntamente com meus amigos de laboratório.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“As invenções são, sobretudo,
o resultado de um trabalho de teimoso.”
(Santos Dumont)

RESUMO

FREIRE, G. M. **DiT-MAE: Modelo *Transformer Autoencoder* Mascarado para Imputação de Dados Discretos para Questionários Psicométricos Educacionais**. 2025. 87 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Modelos de Teoria de Resposta ao Item (TRI) são bastante difundidos e aplicados em avaliações psicométricas para estimar comportamentos, habilidades, proficiências, aprendizado e distúrbios. Essa metodologia permite estimar traços latentes de indivíduos a partir das respostas a itens de um questionário ou prova. Recentemente, modelos da TRI foram incorporados em redes neurais profundas não supervisionadas, da família dos autoencoders, buscando maior flexibilidade e escalabilidade para análise de dados educacionais ou psicológicos de maior dimensão e complexidade. Os modelos do tipo autoencoders são caracterizados por duas redes neurais: uma de codificação que comprime os dados (encoder) e uma de decodificação que tenta reconstruir os dados originais (decoder). As atuais propostas incorporam um modelo de TRI no decoder, tornando essa parte da rede interpretável e capaz de estimar uma maior dimensão de habilidades e complexidade dos dados. No entanto, a grande limitação dessa metodologia é a incapacidade de lidar com a ausência de respostas para um ou mais itens do questionário ou prova. Dados faltantes são bastante comuns nas avaliações educacionais, seja por limitação de tempo de prova, ignorância do conteúdo avaliado no item ou, até mesmo, por planejamento de aplicação da prova de avaliação educacional. Os métodos de estimação usuais na TRI são capazes de lidar com dados faltantes e já foram muito investigados na literatura. Porém, no contexto das novas implementações com autoencoders, estes estudos são escassos, não tão eficazes e podem aumentar o nível de complexidade computacional demasiadamente. Esta pesquisa de trabalho implementou uma arquitetura de rede neural, utilizando modelos *Transformers* com mecanismos de *self-attention*, para imputar dados faltantes de avaliações educacionais e melhorar a qualidade das estimativas dos parâmetros do modelo de TRI e interpretabilidade dos resultados. Foram realizadas simulações para comparar com métodos tradicionalmente empregados na literatura. O novo método demonstrou desempenho melhor na redução da complexidade, tempo de execução e na estimativa de parâmetros da TRI para grandes conjuntos de dados e altas dimensões. Uma aplicação real também foi realizada com os dados do PISA.

Palavras-chave: Dados Faltantes, Teoria de Resposta ao Item, Modelo *Transformer*, Redes Neurais, *Variational Autoencoder*.

ABSTRACT

FREIRE, G. M. **DiT-MAE: Masked Autoencoder Transformer Model for Discrete Data Imputation for Educational Psychometric Questionnaires**. 2025. 87 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Item Response Theory (IRT) models are widely used and applied in psychometric assessments to estimate behaviors, skills, proficiencies, learning, and disorders. This methodology allows for the estimation of latent traits of individuals based on responses to items in a questionnaire or test. Recently, IRT models have been incorporated into unsupervised deep neural networks, from the autoencoder family, seeking greater flexibility and scalability to analyze larger and more complex educational or psychological data. Autoencoder models are characterized by two neural networks: an encoding network that compresses the data (encoder) and a decoding network that attempts to reconstruct the original data (decoder). Current proposals incorporate an IRT model into the decoder, making this part of the network interpretable and capable of estimating a larger dimension of skills and data complexity. However, the major limitation of this methodology is the inability to deal with the absence of responses to one or more items in the questionnaire or test. Missing data are quite common in educational assessments, whether due to time constraints, ignorance of the content assessed in the item, or even due to the planning of the educational assessment test application. The usual study methods in IRT are capable of dealing with missing data and have already been extensively investigated in the literature. However, in the context of new implementations with autoencoders, these studies are scarce, not so practical, and can significantly increase the computational complexity. This research work implemented a neural network architecture, using transformer models with self-attention mechanisms, to impute missing data from educational assessments and improve the quality of the estimates of the parameters of the IRT model and the interpretability of the results. Simulations were performed to compare with traditional methods used in the literature. The new method demonstrated better performance in reducing the complexity, execution time, and parameter estimation of TRIM for large data sets and high dimensions. A real application was also performed with PISA data.

Keywords: Item Response Theory, Missing Data, Neural Networks, Transformer Model, Variational Autoencoder.

LISTA DE ILUSTRAÇÕES

Figura 1 – Curva Característica do Item (CCI) (Fonte: Elaborado pelo autor) 32
Figura 2 – Modelo Logístico de 1 Parâmetro (Fonte: Elaborado pelo autor). 33
Figura 3 – Modelo Logístico de 2 Parâmetros (Fonte: Elaborado pelo autor). 34
Figura 4 – Modelo Logístico de 3 Parâmetros (Fonte: Elaborado pelo autor). 35
Figura 5 – (a) Representação gráfica do modelo compensatório e (b) representação gráfica do modelo parcialmente compensatório – superfície de resposta do item (RECKASE, 2006) 36
Figura 6 – Modelo não linear de um neurônio com w_{k0} para representar o viés b_k (HAYKIN, 1994) 43
Figura 7 – Função *Threshold* (HAYKIN, 1994) 44
Figura 8 – Função Sigmoide (HAYKIN, 1994) 44
Figura 9 – Aplicação da técnica gradiente descendente (Fonte: Elaborado pelo autor). . 46
Figura 10 – Rede neural no sentido direto do processamento (LECUN; BENGIO; HINTON, 2015). 46
Figura 11 – Rede neural no sentido reverso do processamento (LECUN; BENGIO; HINTON, 2015). 47
Figura 12 – Representação gráfica da rede neural *Variational Autoencoder*. Os vetores α e β representam os parâmetros das redes codificadora e decodificadora, respectivamente (Adaptado de Curi *et al.* (2019)) 48
Figura 13 – Mecanismos de atenção: (a) Atenção por Produto Escalar e (b) Atenção de Múltiplas Cabeças (VASWANI *et al.*, 2017). 52
Figura 14 – *Discrete Transformer Masked Autoencoder*. Adaptado de He *et al.* (2022). . 59
Figura 15 – Modelo *Transformer*. Adaptado de Vaswani *et al.* (2017). 60
Figura 16 – Arquitetura VAEQ proposta por Curi *et al.* (2019). 62
Figura 17 – Matriz de Confusão para dados faltantes com taxas de 10%, 25% e 50% porcentos. 65
Figura 18 – Sensibilidade, especificidade e acurácia para todos os itens com taxas de 10%, 25% e 50% de dados faltantes para 100 réplicas. 66
Figura 19 – Gráfico de dispersão das discriminações para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM. 68
Figura 20 – Gráfico de dispersão das dificuldades para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM. 69

Figura 21 – Gráfico de dispersão das habilidades para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM. 71
Figura 22 – Gráfico *boxplot* das habilidades para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM. 72
Figura 23 – Porcentagem de acertos da imputação do DiT-MAE para taxas de 10%, 25% e 50% de respostas ausentes. 73
Figura 24 – Gráfico de dispersão entre as estimativas das habilidades para para as três taxas de respostas ausentes do conjunto de dados de teste. 74
Figura 25 – Gráfico de dispersão entre as estimativas das habilidades para para as três taxas de respostas ausentes do conjunto de dados com respostas ausentes reais. 75
Figura 26 – Gráfico de dispersão para as estimativas das habilidades entre o VAEQ e o JML, variando a porcentagem de respostas ausentes para cada dimensão. . . 76
Figura 27 – Histograma e distribuição normal por habilidade das estimativas do VAEQ e taxas de dados faltantes para cada país, diferenciando países economicamente desenvolvidos. 77

LISTA DE ALGORITMOS

Algoritmo 1 – Algoritmo Robbins-Monro Metropolis-Hastings. Adaptado de Cai 2010 41

LISTA DE TABELAS

Tabela 1 – Resultado da acurácia binária para imputar dados faltantes (DF) para 3 e 21 dimensões no treino. 64
Tabela 2 – Correlação média entre valores reais e as estimativas do VAEQ e do JML para 100 réplicas. 67

LISTA DE ABREVIATURAS E SIGLAS

CBA	*Computer-Based Assesment*
CCI	Curva Característica do Item
CNN	*Convolutional Neural Network*
DAE	*Denoising Autoencoder*
DBN	*Deep Belief Network*
DiT-MAE	*Discrete Transformer Masked Autoencoder*
DNN	*Deep Neural Network*
EM	*Expectation-Maximization*
IM	Imputação Múltipla
JML	Méxima Verossimilhança Conjunta
MAR	*Missing At Random*
MAR	*Missing at Random*
MC	Monte Carlo
MCAR	*Missing Complete At Random*
MCAR	*Missing Completely at Random*
MCEM	*Monte Carlo Expectation-Maximization*
MCMC	*Markov Chain Monte Carlo*
MHRM	Metropolis-Hastings Robbins-Monro
MNAR	*Missing Not At Random*
MV	Máxima Verossimilhança
NLP	*Natural Language Processing*
PISA[1]	*Programme for International Student Assessment*
ReLU	*Rectified Linear Unit*
RMSE	Raiz do Erro Quadrático Médio
SVM	Support Vector Machine
TRI	Teoria de Resposta ao Item
TRIM	Teoria de Resposta ao Item Multidimensional
VAE	*Variational Autoencoder*
VI	*Variational Inference*
ViT-MAE	*Visual Transformer Masked Autoencoder*

LISTA DE SÍMBOLOS

J — Função de Custo

SUMÁRIO

1 INTRODUÇÃO 27

2 REFERENCIAL TEÓRICO 31
2.1 Teoria de Resposta ao Item (TRI) 31
2.1.1 Teoria de Resposta Ao Item Multidimensionais *35*
2.1.2 Dados Faltantes *36*
2.1.3 Métodos de Estimação *39*
2.2 Redes Neurais 42
2.2.1 Princípios da Rede Neural *42*
2.2.2 Variational Autoencoders *47*
2.2.3 Modelos Transformers *50*
2.2.4 Modelos de Deep Learning para Imputação de Dados Faltantes . . *51*

3 DESENVOLVIMENTO 55
3.1 Criação dos Dados Artificiais 55
3.2 Extração dos Dados Reais: PISA 2015 57
3.3 Arquitetura DiT-MAE 58

4 RESULTADOS 63
4.1 Resultados dos Dados Artificiais 63
4.2 Resultados dos Dados Reais (PISA 2015) 72

5 CONCLUSÃO 79
5.1 A Arquitetura DiT-MAE 79
5.2 Limitações do DiT-MAE 80
5.3 Trabalhos Futuros para O DiT-MAE 81

REFERÊNCIAS 83

CAPÍTULO

1

INTRODUÇÃO

As avaliações educacionais são ótimos recursos para obtenção de dados psicométricos dos indivíduos. Compostas por itens, normalmente de múltipla escolha, as avaliações geram dados que permitem analisar habilidades e proficiências dos indivíduos em determinado conteúdo. Por serem características não diretamente observáveis, essas habilidades são genericamente denominadas traços latentes. Para estimar traços latentes, a Teoria de Resposta ao Item (TRI) se destaca como a metodologia moderna para análises estatísticas de dados provenientes de avaliações educacionais (PASQUALI, 2011). Na educação, as políticas aplicadas têm buscado aprimorar as formas de avaliação que permitem identificar as habilidades ausentes nos alunos, na tentativa de suprir de forma mais eficaz as deficiências (ANDRADE; TAVARES; VALLE, 2000).

A TRI é um conjunto de modelos matemáticos que relacionam a probabilidade de resposta aos itens a habilidades dos indivíduos. O modelo mais comum na prática considera respostas dicotômicas e unidimensionalidade, ou seja, uma única habilidade influenciando as respostas aos itens. Apesar de muito usado, a adoção de um modelo unidimensional limita a precisão das estimativas e flexibilidade da avaliação (LINDEN; HAMBLETON, 2013; RECKASE, 2006). É intuitiva a percepção de que, na realidade, a resposta correta de um item de uma avaliação educacional envolve múltiplas habilidades. Mesmo que a prova ou questionário envolva um mesmo tema, digamos Matemática, há competências distintas que interferem no resultado, como cálculo algébrico, raciocínio geométrico e interpretação de dados, do enunciado. Considerar modelos unidimensionais simplifica demasiadamente essa complexidade, podendo comprometer a qualidade e eficácia dos resultados em alguns casos. Assim, modelos multidimensionais da TRI tornam-se mais adequados. A Teoria de Resposta ao Item Multidimensional (TRIM) estabelece uma relação probabilística entre a resposta do indivíduo aos itens e sua localização no espectro de traços latentes específicos, identificando quais habilidades da pessoa avaliada estão presentes e suas respectivas quantificações (RECKASE, 1997).

Tradicionalmente, as estimativas dos parâmetros de modelos TRIM são baseadas em

algoritmos de máxima verossimilhança marginal e *Expectation-Maximization* (EM), assumindo os traços latentes como efeitos aleatórios no modelo. Essa abordagem exige o cálculo de uma integral multidimensional, o que se torna um desafio computacional quando o número de traços latentes não é muito pequeno. Métodos de Monte Carlo (MC) dentro do algoritmo Metropolis-Hastings Robbins-Monro (MHRM) têm sido aplicados para obter aproximações numéricas dessa integral para modelos TRIM. Uma outra abordagem é assumir que os traços latentes são parâmetros fixos no modelo e realizar a estimação por Méxima Verossimilhança Conjunta (JML). No entanto, a utilidade desses métodos é confinada a contextos específicos e limitados. Isso surge da complexidade computacional de avaliar integrais de alta dimensão em funções de verossimilhança, particularmente ao lidar com a dimensionalidade crescente de características latentes (RECKASE, 2006; CHEN; LI; ZHANG, 2019; CAI, 2010).

Recentemente na literatura, os modelos TRIM estão sendo implementados utilizando redes neurais para contornar o problema de dimensionalidade alta (CURI *et al.*, 2019; WU *et al.*, 2020; CONVERSE; CURI; OLIVEIRA, 2019). A implementação de redes neurais e suas aplicações em redes neurais profundas tem desempenhado diversas funções de forma eficaz, que variam entre reconhecimento e classificação de imagens, áudio e texto. Entre esses modelos de rede neural profunda, os modelos probabilísticos com natureza bayesiana permitem maior flexibilidade para lidar com incertezas que ocorrem nos treinamentos dos vetores de peso em uma regressão linear, resultando em maior nível de inferência e a incorporação de um modelo estatístico (como o TRIM) numa rede neural profunda enriquece os resultados obtidos (HAYKIN, 1994; LECUN; BENGIO; HINTON, 2015; GOODFELLOW; BENGIO; COURVILLE, 2016; WANG; YEUNG, 2016). Os métodos para treinamento de algoritmos de *deep learning* permitem a estimação dos parâmetros do modelo TRIM aproveitando-se das vantagens das duas metodologias: uma estimação robusta num contexto de alta dimensão, de forma eficiente, relacionando probabilidade de resposta aos itens do questionário e variáveis latentes (MEADE; LAUTENSCHLAGER, 2004).

A aplicabilidade da metodologia proposta deve ser investigada em diferentes âmbitos das avaliações educacionais. No entanto, sua fundamentação teórica e sua utilidade em situações práticas ainda carecem de estudos aprofundados.

Simulações em diferentes cenários relacionando número de itens no questionário, dimensionalidade do espaço dos traços latentes e aplicações a dados reais são encontrados na literatura, com algumas propostas para o treinamento da rede. Faz-se necessário o uso de grandes volumes de dados e técnicas de inferência são adotadas para viabilizar a escalabilidade das redes e quantificação de incertezas. Em particular, os trabalhos apresentados por Curi *et al.* (2019) e Converse, Curi e Oliveira (2019), estendem a *Variational Autoencoder* (VAE) (KINGMA; WELLING, 2019) para incorporar na inferência uma matriz de habilidades necessárias para responder corretamente aos itens. A primeira parte da arquitetura (*encoder*) inferem-se as variáveis latentes que em seguida são entradas para a segunda parte da arquitetura (*decoder*) e estimação

das probabilidades de acerto dos itens. A VAE é uma arquitetura de rede neural da família dos modelos generativos (HAYKIN, 1994; GOODFELLOW; BENGIO; COURVILLE, 2016). Esta arquitetura utiliza a técnica de inferência variacional, *Variational Inference* (VI), para aproximar integrais intratáveis que surgem na inferência bayesiana (GRAVES, 2011; HOFFMAN *et al.*, 2013; BLEI; KUCUKELBIR; MCAULIFFE, 2017). VI adota a divergência de Kullback-Leibler (GOODFELLOW; BENGIO; COURVILLE, 2016; HAYKIN, 1994) para reduzir a distância entre as distribuições aproximada e real (ZHANG *et al.*, 2018; BÖHM; LANUSSE; SELJAK, 2019), neste estudo, associadas aos traços latentes condicionadas aos dados observados. Em conjunto com a TRIM, a arquitetura VAE tem sido utilizada para estimar as habilidades na saída do *encoder* e o *decoder* utilizado para estimar os parâmetros dos itens do modelo TRIM através dos pesos e vieses da rede neural (CURI *et al.*, 2019; CONVERSE; CURI; OLIVEIRA, 2019; CONVERSE *et al.*, 2021; URBAN; BAUER, 2021; WU *et al.*, 2020). Embora essa metodologia melhore drasticamente a eficiência da estimação para alta dimensão em comparação com os métodos tradicionais baseados em MVM e EM, a abordagem para lidar com valores ausentes é pouco explorada na literatura, com propostas ainda ineficazes.

Propostas simplórias envolvem a edição do conjunto de dados, eliminando linhas inteiras (*listwise deletion*) ou exclusão por pares (*pairwise deletion*) em que as variáveis de interesse não estão presentes. Outro método muito comum é realizar a média do atributo em específico (em um conjunto de dados, as linhas representam o número de registros e as colunas as variáveis características ou atributos) para imputar a variável de interesse faltante. Todavia, dados ausentes ocultam informações valiosas que podem ser distorcidas pela aplicação dos métodos simplistas acima descritos, comprometendo a fidedignidade dos resultados das análises estatísticas (LITTLE; RUBIN, 2019; GRAHAM, 2009; LITTLE *et al.*, 2014).

Redes neurais também estão sendo investigadas para abordar o problema de dados faltantes. Uma proposta é aplicar modelos de Redes Neurais Convolucionais (CNNs) para extrair recursos vizinhos para melhores resultados de imputação (GAD *et al.*, 2021). Porém, a abordagem frequentemente empregada é tratar dados ausentes como ruídos e utilizar técnicas de redução de ruído dentro das arquiteturas do Autoencoder para realizar inferências para os dados faltantes (CHEN *et al.*, 2015; LIN *et al.*, 2020; LIN; TSAI; ZHONG, 2022; PEREIRA *et al.*, 2020).

No âmbito do VAE com o *decoder* definido como um modelo TRIM, as propostas existentes na literatura para tratar dados faltantes adaptam a função de perda num primeiro momento de forma a ignorar os dados faltantes. Em seguida, utilizam a saída do VAE como imputação para os dados faltantes (LIU; WANG; XU, 2022; MONTECINO, 2023; VELDKAMP; GRASMAN; MOLENAAR, 2025).

Para o presente trabalho, também é proposta a imputação dos dados faltantes, mas fazendo uso do modelo estado da arte, *Transformer*. Um modelo reduzido de um *Transformer Autoencoder* mascarado é implementado para aprender o contexto das informações, eliminando

os ruídos. O modelo implementado trata os dados faltantes como ruídos e os mascara para que não participem do aprendizado. Finalizado o treinamento do modelo, os dados mascarados são imputados a partir dos dados observáveis. Após a etapa de imputação de dados, aplicam-se os resultados na arquitetura proposta por Curi *et al.* 2019 para estimar os parâmetros de item e pessoas da TRIM para avaliar a qualidade da imputação.

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta conceitos básicos de Teoria de Resposta ao Item, Dados faltantes e Redes Neurais. No Capítulo 3 são descritos o modelo para imputação de dados faltantes, a condução dos experimentos, a preparação dos dados artificiais e a preparação dos dados reais oriundos das respostas do PISA. Por fim, no Capítulo 4 são apresentados os resultados da imputação dos dados e as estimativas a partir do modelo VAEQ proposto por Curi *et al.* 2019 para validação dos dados imputados.

CAPÍTULO

2

REFERENCIAL TEÓRICO

O presente capítulo descreve os principais conceitos para conduzir esta pesquisa. A organização deste está dividida da seguinte forma: A Seção 2.1 apresenta os conceitos relacionados à TRI, TRI multidimensional e Dados Faltantes. Na Seção 2.2 são descritos os conceitos de Redes Neurais e as arquiteturas do *Variational Autoencoder* e *Transformers*. Por fim, nesta mesma seção, são apresentados trabalhos relacionados com a aplicação de redes neurais para imputação de dados faltantes.

2.1 Teoria de Resposta ao Item (TRI)

A Teoria de Resposta ao Item (TRI) é um paradigma para avaliação psicométrica, composto por modelos matemáticos que tentam explicar a relação entre traços latentes e sua manifestação em respostas dadas a itens em uma avaliação. Por traços latentes, entendem-se características ou atributos não observáveis que influenciam as respostas, desempenho e resultados de testes de escores, questionários e instrumentos avaliativos similares (ANDRADE; TAVARES; VALLE, 2000; ARAUJO; ANDRADE; BORTOLOTTI, 2009; EMBRETSON; REISE, 2013; RECKASE, 2006). Os primeiros modelos de TRI focam em itens de testes de natureza dicotômica. A resposta de um indivíduo j a um item i é denotada por uma variável binária U_{ij}, tal que: $U_{ij} = 0$ (se incorreta) e $U_{ij} = 1$ (se correta). A distribuição da variável U_{ij} é claramente Bernoulli e a probabilidade de resposta correta é função da habilidade do indivíduo, $\theta_j \in (-\infty, +\infty)$ e de parâmetros associados a item. Os mais comuns são os modelos logísticos de 1 (modelo de Rasch), 2 (modelo de Birnbaum) ou 3 parâmetros para cada item (ANDRADE; TAVARES; VALLE, 2000; LINDEN; HAMBLETON, 2013). No modelo de Birbaum, os parâmetros de cada item i, dificuldade (b) e discriminação (a), afetam a probabilidade de sucesso e são denotados, respectivamente, por $b_i \in (-\infty, +\infty)$ e $a_i \in (-\infty, +\infty)$, onde a's positivos geram condições com mais interpretabilidade na área educacional. A probabilidade de sucesso de um item i é representada por $P_i(\theta)$, ou seja, uma função de θ específica ao item

i, onde θ é o parâmetro de interesse que caracteriza o traço latente dos indivíduos e a_i e b_i são os parâmetros de propriedades do item. $P_i(\theta)$ não devem ser definidas por funções lineares, uma vez que representam probabilidades e devem ser limitadas no intervalo (0,1). Sendo assim, normalizações sigmoides (ou ogivais) são aplicadas a $P_i(\theta)$ (LINDEN; HAMBLETON, 2013).

Os modelos de TRI são ilustrados pela Curva Característica do Item (CCI) e é usual a suposição de independência local ou condicional. A CCI representa a relação entre alterações nos traços latentes e a probabilidade de uma resposta correta de cada item. A independência local está relacionada à suposição de que, para um dado nível de habilidade do indivíduo, as respostas aos itens são independentes entre si (ANDRADE; TAVARES; VALLE, 2000; ARAUJO; ANDRADE; BORTOLOTTI, 2009; LINDEN; HAMBLETON, 2013). Há diversos modelos na literatura que variam de acordo com o tipo de item e o tipo de processo de resposta, podendo ser acumulativo ou não acumulativo (LINDEN; HAMBLETON, 2013). A Figura 1 ilustra o gráfico da CCI de um modelo da TRI para itens dicotômicos e apenas um traço latente.

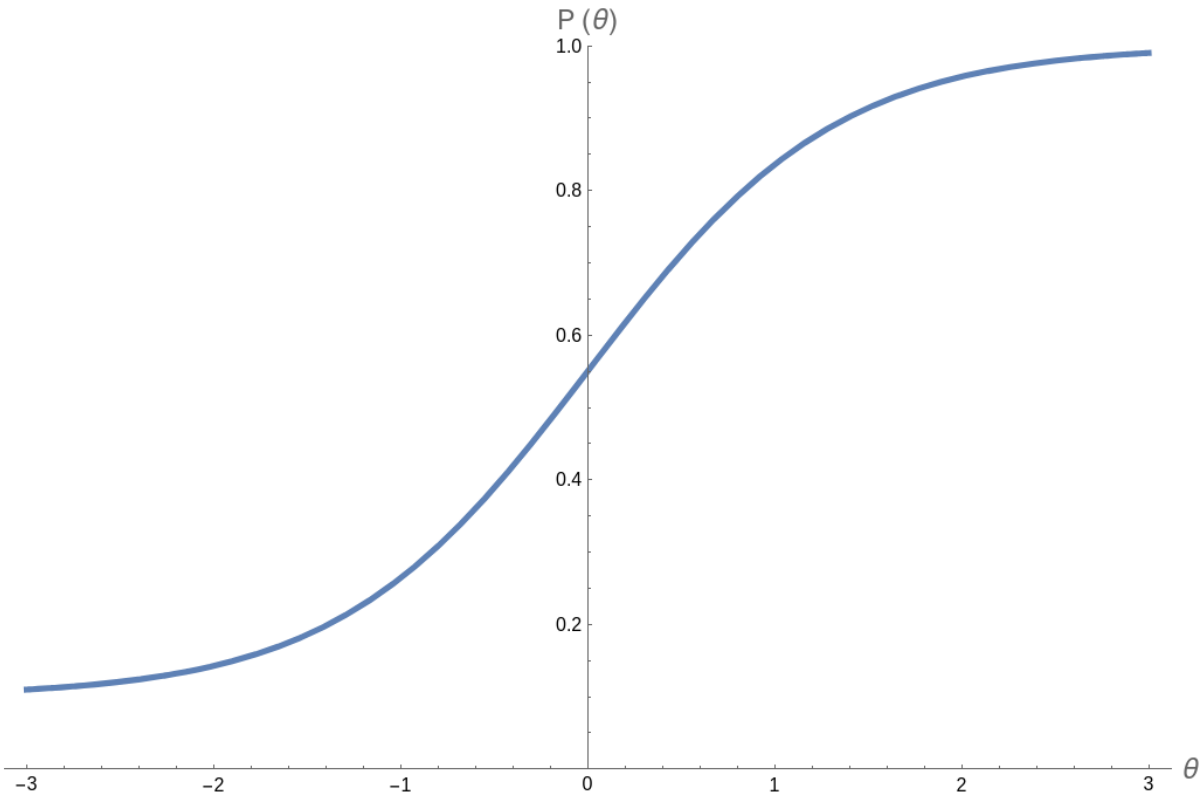

Figura 1 – Curva Característica do Item (CCI) (Fonte: Elaborado pelo autor)

.

De acordo com Andrade *et al.* (2000), os modelos de TRI variam basicamente com três fatores:

1. A natureza do item, dicotômicos ou não-dicotômicos;
2. A quantidade de traços latentes que está sendo medida, modelos unidimensionais ou multidimensionais;

3. O número de populações envolvidas.

O modelo de 1PL, ou modelo de Rasch, distingue as CCIs entre os itens apenas por 1 parâmetro: dificuldade (b) do item. Sua equação é dada por:

$$P_{ij} = P(U_{ij} = 1|\theta_j) = \frac{1}{1+e^{-(\theta_j - b_i)}}, \tag{2.1}$$

para $i = \{1,2,3,...,m\}$, tal que $i \leq m$ e $j = \{1,2,3,...,n\}$, tal que $j \leq n$, onde U_{ij} corresponde à resposta para o item i do indivíduo j, θ_j é a habilidade do indivíduo j e b_i é o parâmetro de dificuldade do item i. A Figura 2 mostra o modelo 1PL em que a dificuldade determina o nível da habilidade que corresponde a 50% da probabilidade de acerto (LINDEN; HAMBLETON, 2013). A dificuldade baixa determina maior probabilidade de acerto (linha azul) e a dificuldade alta determina menor probabilidade de acerto do item (linha verde) ao longo do traço latente, θ.

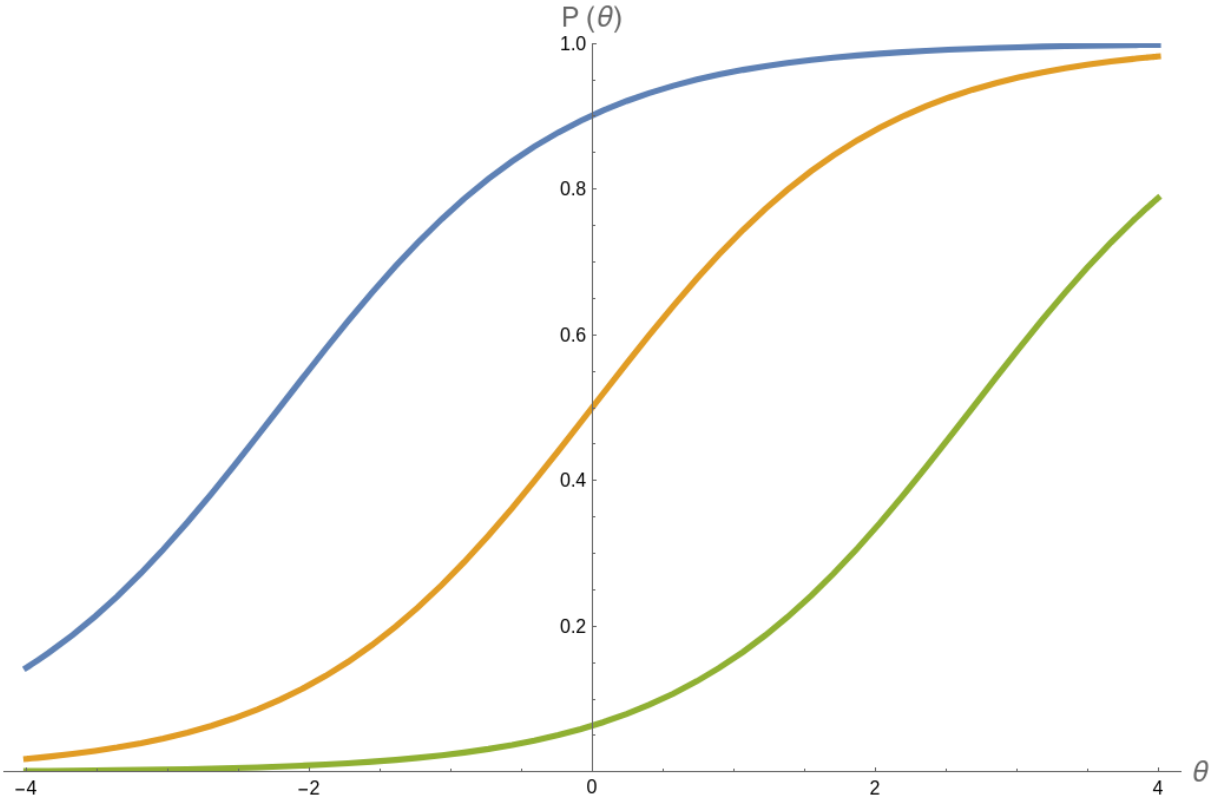

Figura 2 – Modelo Logístico de 1 Parâmetro (Fonte: Elaborado pelo autor).

A Equação 2.2 descreve a função do modelo logístico de 2 parâmetros, onde θ representa a habilidade, b_i é o parâmetro de dificuldade e a_i é o parâmetro de discriminação (LINDEN; HAMBLETON, 2013).

$$P_{ij} = P(U_{ij} = 1|\theta_j) = \frac{1}{1+e^{-a_i(\theta_j - b_i)}} \tag{2.2}$$

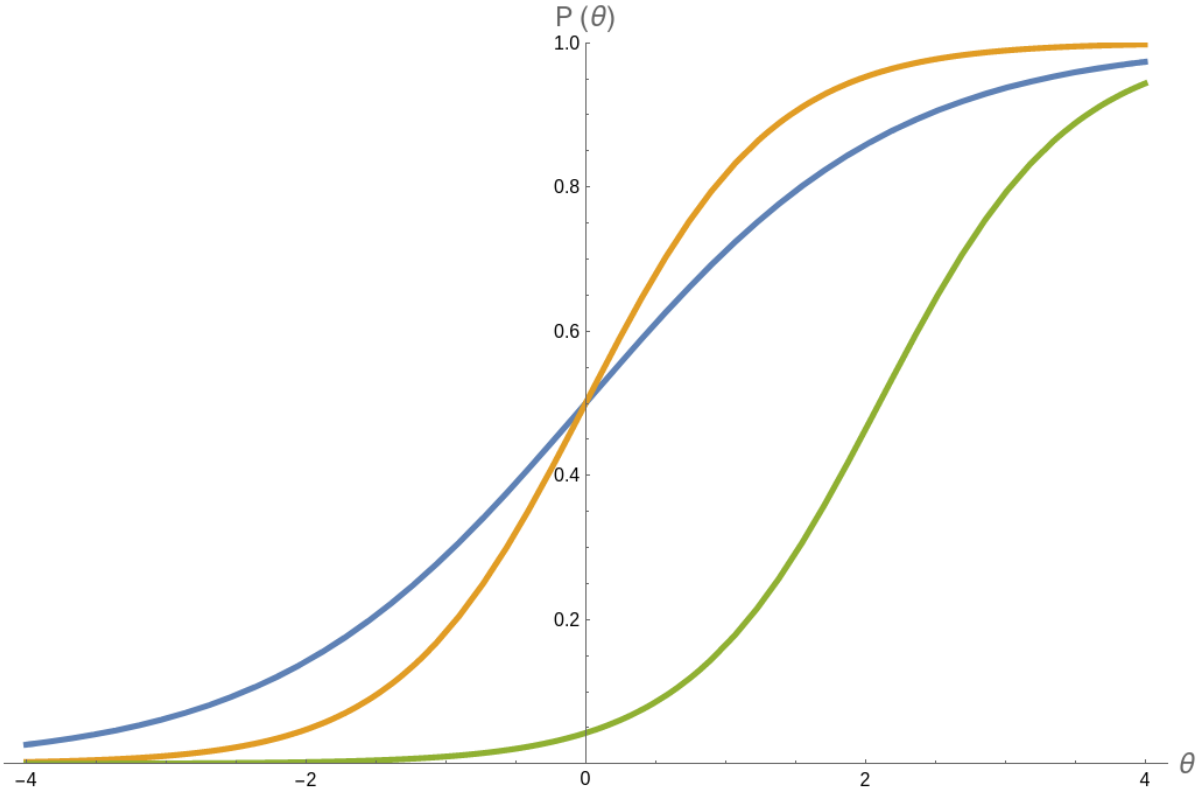

Figura 3 – Modelo Logístico de 2 Parâmetros (Fonte: Elaborado pelo autor).

A Figura 3 ilustra um exemplo de gráfico referente à Equação 2.2. As diferentes linhas no gráfico indicam diferentes itens. Para responder aos itens correspondentes às linhas azul e laranja, com probabilidade de 50% de acerto, o nível "0"(zero) de habilidade é necessário. Contudo, para se ter a mesma probabilidade de acerto para o item representado pela cor verde, é necessária uma habilidade maior. Ou seja, os parâmetros b_i e a_i influenciaram na posição e inclinação da curva, permitindo que alguns itens necessitem de um nível maior de habilidade para obter uma razoável probabilidade de acerto do item.

O modelo logístico de 3 parâmetros acrescenta uma terceira característica ao item, o acerto casual, que considera a probabilidade de resposta correta do indivíduo sem a habilidade necessária para responder ao item. A Equação do modelo 3PL é denotada por (LINDEN; HAMBLETON, 2013):

$$P_{ij} = P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \tag{2.3}$$

A Equação 2.3 modela a probabilidade do indivíduo conhecer a resposta correta ou a probabilidade de acerto casual igual ao valor de c_i. A Figura 2.3 ilustra 3 itens com parâmetros a e b iguais, porém o parâmetro c de acerto casual varia. Em caso do parâmetro c ser maior, aumenta a probabilidade de acerto para indivíduos com baixa habilidade, indicando que o item tem alternativas que permitem acertos por adivinhação. Itens com parâmetro c menor requerem um nível maior de habilidade para que o acerto ocorra. Se $c = 0$, o item se comporta como um

modelo 2PL (sem acerto ao acaso).

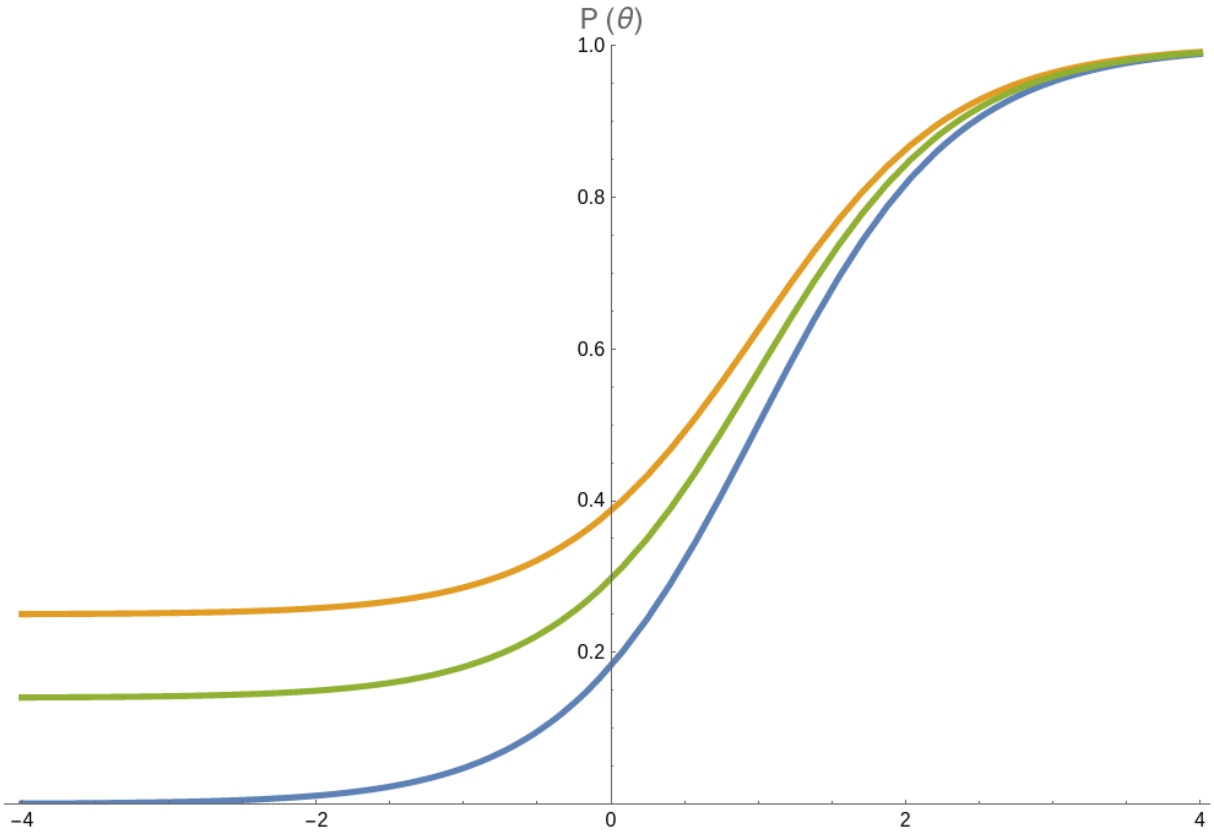

Figura 4 – Modelo Logístico de 3 Parâmetros (Fonte: Elaborado pelo autor).

Este trabalho de pesquisa se concentra no modelo dicotômico multidimensional (Subseção 2.1.1), uma generalização do modelo logístico de 2 parâmetros acima apresentado.

2.1.1 Teoria de Resposta Ao Item Multidimensionais

Os modelos de TRI multidimensionais estimam múltiplos traços latentes. Cada indivíduo j tem seu nível do traço latente l denotado por θ_{jl}. Os vários traços latentes do indivíduo são representados por um vetor $\theta_j = [\theta_{j1}, \theta_{j2}, \theta_{j3}, ..., \theta_{jM}]$ (RECKASE, 2009). Segundo Reckase (2006), diversos modelos de TRI multidimensional foram propostos, porém apenas duas formas básicas são apresentadas frequentemente na literatura e, particularmente, nas aplicações práticas: os modelos compensatórios e parcialmente compensatórios. O modelo compensatório possui a característica de que valores altos de um subconjunto de traços latentes, podem provocar valores menores para os demais. Adicionalmente, os M traços latentes podem se compensar de modo que dois indivíduos com vetores distintos de habilidades, θ_j e θ_{j*}, apresentem a mesma probabilidade de resposta correta. A Equação 2.4 é a forma logística do modelo compensatório, onde a_i é um vetor de parâmetros de discriminação do item i, de tamanho M. No geral, à medida que a localização na coordenada do vetor de traço latente altera, a taxa de probabilidade de resposta correta também é alterada. Note que o parâmetro d_i é escalar e está relacionado à dificuldade

do item. A Figura 5(a) é o gráfico relacionado a Equação 2.4, supondo dimensão dois para o vetor de traços latentes. Uma característica observável na superfície da Figura 5 é a dimensão θ_1 possuir uma inclinação mais rápida na superfície quando comparada com a dimensão θ_2. Isso ocorre devido aos diferentes elementos no vetor a: quanto maior seu valor, maior a inclinação.

$$P(u_i = 1|\theta_j) = \frac{e^{a_i'\theta_j + d_i}}{1 + e^{a_i'\theta_j + d_i}} \tag{2.4}$$

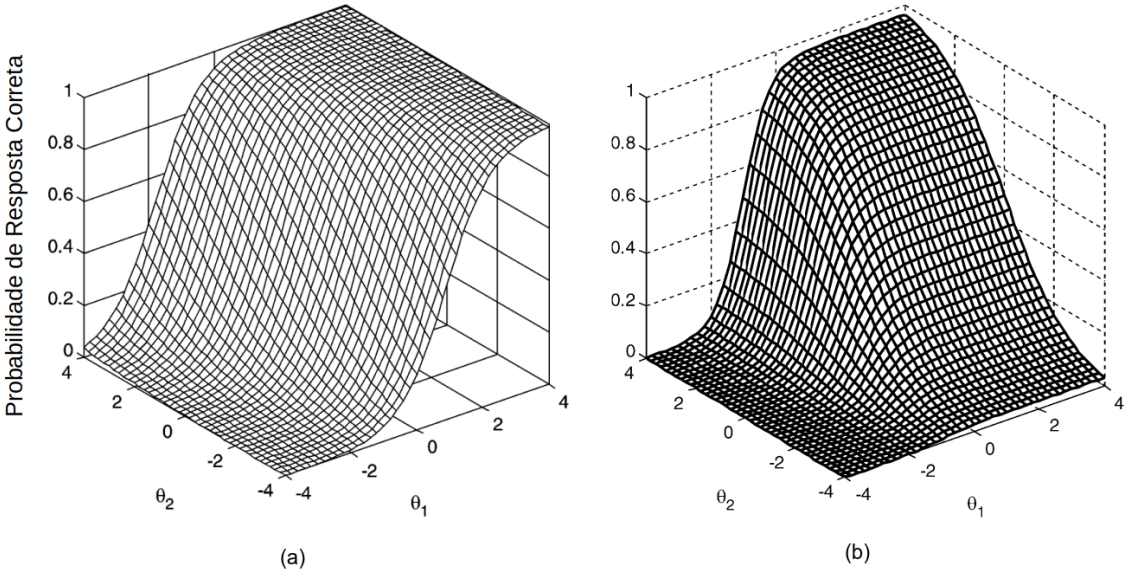

Figura 5 – (a) Representação gráfica do modelo compensatório e (b) representação gráfica do modelo parcialmente compensatório – superfície de resposta do item (RECKASE, 2006)

.

O modelo parcialmente compensatório possui a característica de um alto valor em uma habilidade não compensa um baixo valor em outra, como ocorre na Equação 2.4. O seu modelo é expresso na Equação 2.5, onde l representa o índice da coordenada das dimensões, b_{il} é o parâmetro que indica a dificuldade para desempenhar a tarefa relacionada à dimensão l, e a constante 1,7 está presente para que o comportamento do modelo se aproxime da ogiva normal. O modelo é ilustrado pela curvatura da superfície da Figura 5(b). Quando o parâmetro c é zero, a probabilidade de resposta correta para um item nunca poderá ser maior que o menor valor do produto. (RECKASE, 2006).

$$P(u_i = 1|\theta_j) = c_i + (1 - c_i)\prod_{l=1}^{m} \frac{e^{1,7a_{il}(\theta_{jl} - b_{il})}}{1 + e^{1,7a_{il}(\theta_{jl} - b_{il})}} \tag{2.5}$$

2.1.2 Dados Faltantes

Dados faltantes representam ums situação comum na análise estatística em diversos modelos de pesquisa e têm sido amplamente estudados e documentados na literatura. A investigação dos dados ausentes deve ser realizada a partir do mecanismo que causa a ausência desses dados.

Os diversos fatores que ocasionaram a perda de algumas observações podem estar associados às respostas observadas (LITTLE; RUBIN, 2019; ENDERS, 2022). Dessa forma, dados faltantes são considerados fenômenos probabilísticos e se dividem em três mecanismos: Dados faltantes completamente aleatórios (*Missing Complete At Random* (MCAR)), Dados faltantes aleatórios (*Missing At Random* (MAR)) e Dados faltantes não aleatórios (*Missing Not At Random* (MNAR)) (GRAHAM, 2009).

Com a finalidade de exemplificar melhor os três mecanismos, suponha R uma variável indicadora que define quais valores estão ausentes (R=0) e quais foram observados (R=1). O mecanismo de dados faltantes pode ser modelado por sua distribuição condicional aos dados completos, $P(R \mid Y_{com})$. O vetor de dados completos pode ser particionado em $Y_{com} = (Y_{obs}, Y_{mis})$, onde Y_{obs} são os dados observados e Y_{mis} são os dados faltantes. O mecanismo MAR (Dados Faltantes Aleatórios) é caracterizado pela distribuição de R independer de Y_{mis},

$$P(R|Y_{com}) = P(R|Y_{obs}). \tag{2.6}$$

Sendo assim, a ocorrência dos dados faltantes depende apenas dos dados observados e não dos dados faltantes. Para o caso especial MCAR (Dados Faltantes Completamente Aleatórios), a distribuição de dados faltantes é independente dos dados observados Y_{obs}, logo,

$$P(R|Y_{com}) = P(R). \tag{2.7}$$

Em outras palavras, o mecanismo MCAR pode ser definido como a distribuição de R não depende dos dados observados e faltantes, Y_{obs} e Y_{mis}, respectivamente. O último mecanismo é o MNAR (Dados Faltantes Não Aleatórios) em que R não depende dos dados observados, ou seja,

$$P(R|Y_{com}) = P(R|Y_{mis}). \tag{2.8}$$

Dessa forma, o mecanismo MNAR depende dos dados faltantes e é independente dos dados observados (RUBIN, 1976 apud SCHAFER; GRAHAM, 2002). Isso significa que o padrão de não resposta está diretamente relacionado aos valores que estão faltando, tornando a ausência de dados não aleatória.

No contexto da TRI em avaliação educacional, os dados faltantes ocorrem tipicamente porque os respondentes muitas vezes nem tentam responder a certos itens (como por exemplo, provas adaptativas ou desistências em testes longos). Se a probabilidade de falta de resposta depender apenas das variáveis observadas (como desempenho em itens anteriores), então o mecanismo pode ser tratado como MAR e os métodos tradicionais são viáveis para estimação.

Um exemplo de mecanismo não aleatório na área de avaliação educacional é o *Programme for International Student Assessment* (PISA[1]). De acordo com Kohler *et al.* 2015,

[1] https://www.oecd.org/en/about/programmes/pisa.html

Pohl *et al.* 2014, Pohl e Becker 2020, Robitzsch 2021 e Rutkowski 2011, as características que influenciam a probabilidade de dados faltantes em respostas de questionários do PISA são itens não respondidos por omissão (caso em que o aluno não tem conhecimento da resposta e não responde intencionalmente). Com isso, o mecanismo de dados faltantes que melhor representa as respostas do PISA é o MNAR (Dados Faltantes Não Aleatórios). Entendemos que o modelo apresentado neste trabalho de pesquisa, assim como os métodos tradicionais da TRI, não atende especificamente ao mecanismo MNAR. Essa abordagem ficará para ser explorada em trabalhos futuros.

Em Ciência de Dados, de forma geral, há soluções de fácil implementação para lidar com dados ausentes, frequentemente empregadas mas sem suporte empírico rigoroso. Exemplos desses métodos convencionais incluem imputação média e exclusão de unidade (LITTLE; SCHENKER, 1995; BARALDI; ENDERS, 2010; LITTLE *et al.*, 2014; SCHAFER; GRAHAM, 2002). Embora essas técnicas possam ser adequadas em situações específicas, elas podem comprometer a eficácia da análise estatística na estimativa de parâmetros (LITTLE; RUBIN, 2019).

Avanços recentes no gerenciamento de dados ausentes abrangem técnicas como estimativas probabilísticas e estimativas bayesianas, conforme observado por Little e Rubin (2019) e Enders (2022). Entre esses métodos, dois surgiram como amplamente adotados, eficazes e com suporte de software prontamente disponível, conforme destacado por Allison (2009). Esses dois métodos são Imputação Múltipla (IM) e Máxima Verossimilhança (MV) (ALLISON, 2009; BARALDI; ENDERS, 2010; LITTLE *et al.*, 2014; SCHAFER; GRAHAM, 2002). De acordo com Baraldi e Enders(2010), IM e MV são os métodos de última geração para lidar com dados ausentes, considerando a relação entre a probabilidade de dados ausentes e o mecanismo de dados medidos proposto por Little e Rubin (2019). Essas abordagens exibem imparcialidade quando os dados ausentes seguem os padrões *Missing Completely at Random* (MCAR) ou *Missing at Random* (MAR). Além disso, eles superam as técnicas tradicionais mais simplórias, pois eliminam a necessidade de descartar quaisquer dados. No contexto da TRIM, a escolha entre IM e MV depende de vários fatores, incluindo o número de itens, características latentes, contagem de imputação e os métodos específicos selecionados para imputação e estimativa. No entanto, à medida que esses fatores aumentam, a complexidade computacional também aumenta.

De acordo com Cai (2010), a complexidade computacional associada aos métodos de IM e MV continua sendo um desafio, particularmente ao lidar com integrais de alta dimensão. Estratégias para mitigar essa complexidade incluem regras adaptativas para quadraturas gaussianas, aproximação de Laplace, algoritmos de *Monte Carlo Expectation-Maximization* (MCEM) e métodos de *Markov Chain Monte Carlo* (MCMC). No entanto, a viabilidade computacional se torna progressivamente elusiva à medida que a dimensionalidade do problema aumenta. Na verdade, a maioria das integrais de MV para variáveis latentes de alta dimensão é desencorajada, e até mesmo integrais de Monte Carlo baseadas em aproximações bayesianas podem consumir

muito tempo (CHEN; LI; ZHANG, 2019; HIPPEL; BARTLETT, 2021).

2.1.3 Métodos de Estimação

Nesta subseção são apresentadas as duas ferramentas, implementadas na linguagem R, para estimação dos parâmetros de modelos da TRIM com ou sem dados faltantes. São elas a méxima verossimilhança conjunta (JML) e o algoritmo Metropolis-Hastings Robbins-Monro (MHRM).

Joint Maximum Likelihood - JML

O *Joint Maximum Likelihood* (JML) é um método empregado para estimar parâmetros de distribuição de probabilidade, maximizando a função de verossimilhança de dados observados sob o modelo assumido. JML estima simultaneamente os parâmetros dos itens e das habilidades dos indivíduos em modelos da TRI ao maximizar a verossimilhança conjunta das respostas observadas (BOCK; GIBBONS, 2021; CHEN; LI; ZHANG, 2019). Nota-se que os parâmetros dos traços latentes dos indivíduos são considerados fixos nesta metodologia. As principais etapas do método JML são descritas a seguir:

(i) **Formulação da Função de Verossimilhança:** Cada unidade amostral supostamente segue uma determinada distribuição de probabilidades. Denotando a função de verossimilhança para a unidade amostral i como $L_i(\theta_i|\mathbf{X}_i)$, onde θ_i representa os parâmetros da distribuição de probabilidade suposta para i, e $\mathbf{X}_i$ representa os dados observados na unidade i;

(ii) **Combinando Probabilidades:** As funções de probabilidade individuais podem ser combinadas em uma função de probabilidade conjunta, gerando a verossimilhança conjunta $L_{\text{joint}}(\theta|\mathbf{X})$. $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ representa o vetor de parâmetros combinados e $\mathbf{X}$ expressa os dados observados na amostra total. A função de verossimilhança conjunta é definida como $L_{\text{joint}}(\theta|\mathbf{X}) = \prod_{i=1}^{n} L_i(\theta_i|\mathbf{X}_i)$, sendo n o número total de unidades na amostra;

(iii) **Maximizando a verossimilhança conjunta:** O objetivo principal é determinar os valores dos parâmetros θ^* que maximizam a função de verossimilhança conjunta $L_{\text{joint}}(\theta|\mathbf{X})$. Isso envolve resolver o problema de otimização $\theta^* = \arg\max_\theta L_{\text{joint}}(\theta|\mathbf{X})$.

(iv) **Técnicas de Otimização:** Para resolver o problema de otimização, várias técnicas de otimização numérica podem ser empregadas. Métodos comuns incluem abordagens baseadas em gradiente, como o algoritmo Newton-Raphson, ou algoritmos de otimização mais gerais, como o algoritmo Broyden–Fletcher–Goldfarb–Shanno (BFGS). O processo de otimização busca encontrar valores de parâmetros que satisfaçam a condição:

$$\frac{\partial L_{\text{joint}}(\theta|\mathbf{X})}{\partial \theta} = \mathbf{0}. \tag{2.9}$$

O método JML tem vantagens relacionadas à eficiência devido aos parâmetros compartilhadas entre as unidades amostrais e ao manuseio de dados de alta dimensão com muitos itens e traços latentes (CHEN; LI; ZHANG, 2019; LIU, 2020; MAYDEU-OLIVARES; CAI; HERNÁNDEZ, 2011). Como compartilha parâmetros comuns, fornece estimativas consistentes de parâmetros de itens e pessoas; permite análise comparativa direta de parâmetros de modelo, auxiliando na seleção de modelos e testes de hipóteses; e pode ser implementado usando algoritmos de otimização Riemanniana que são rápidos e estáveis (LIU, 2020; ROBITZSCH, 2021). Por outro lado, as desvantagens são que o método JML requer grandes tamanhos de amostra para garantir boas propriedades das estimativas e, à medida que os tamanhos de amostra aumentam, a complexidade computacional também aumenta. Ele também pode ser sensível a outliers ou especificações incorretas dos modelos e não leva em conta a incerteza dos parâmetros da pessoa, o que pode afetar os erros padrão e os índices de ajuste dos parâmetros do item (CHEN; LI; ZHANG, 2019; LIU, 2020; MAYDEU-OLIVARES; CAI; HERNÁNDEZ, 2011).

Metropolis-Hasting Robbins-Monro - MHRM

O método Metropolis-Hastings Robbins-Monro (MHRM) pertence à categoria de técnicas de Monte Carlo via Cadeia de Markov (MCMC) e é uma ferramenta valiosa para estimar parâmetros dentro de modelos complexos, particularmente aqueles que envolvem dados com alta dimensão. Ele encontra suas aplicações em Análise Fatorial Exploratória e em modelos de Teoria de Resposta ao Item (TRI). O algoritmo MHRM é adequado para tarefas computacionais de larga escala e oferece várias vantagens: **(a) Simulação de Monte Carlo Eficiente:** O método MHRM emprega eficientemente a simulação de Monte Carlo com um tamanho de iteração fixo e pequeno, tornando-o particularmente adepto ao manuseio de tarefas computacionalmente intensivas; (b) **Estimativa de Matriz de Informações de Parâmetros:** Além da estimativa de parâmetros, o método MHRM fornece uma estimativa da matriz de informações de parâmetros como um subproduto. Essas informações são importantes para estimativas de erro padrão dos estimadores e teste de qualidade de ajuste, aumentando a utilidade geral do método (CAI, 2010; MAYDEU-OLIVARES; CAI; HERNÁNDEZ, 2011).

O algoritmo MHRM opera por meio da geração de sequências de amostras de uma distribuição alvo. Esse processo envolve amostragem repetida com base em uma distribuição de proposta, e a aceitação ou rejeição é determinada pela razão Metropolis-Hasting. Posteriormente, o algoritmo Robbins-Monroe entra em ação como a segunda etapa do processo. Ele otimiza e atualiza iterativamente as estimativas de parâmetros, introduzindo ajustes de tamanho de passo na direção negativa do gradiente da função objetivo.

O Algoritmo 1 apresenta o processo de estimativas de parâmetros e amostras e é descrito nas etapas abaixo:

- **Inicialização:** Inicializar o valor dos parâmetros de θ_0 (linha 1 no Algoritmo 1);

Algoritmo 1 – Algoritmo Robbins-Monro Metropolis-Hastings. Adaptado de Cai 2010

1: Inicializar θ_0
2: $\theta_t = \mathbf{X}$
3: **para** $t = 1$ para T **faça** ▷ Iterações
4: Selecionar amostra θ' da distribuição da proposta $q(\theta'|\theta_t)$
5: Calcular probabilidade de aceitação $\alpha(\theta_t, \theta')$:
6: $\alpha(\theta_t, \theta') = \min\left(1, \frac{p(\theta')}{p(\theta_t)} \cdot \frac{q(\theta_t|\theta')}{q(\theta'|\theta_t)}\right)$
7: Gerar um número aleatório uniforme $u \sim U(0,1)$
8: **se** $u < \alpha(\theta_t, \theta')$ **então**
9: Set $\theta_{t+1} = \theta'$
10: Atualizar estimativa de Robbins-Monro:
11: $\theta_{t+1} = \theta_t + \alpha_t \cdot (f(\theta') - b_t)$
12: **senão**
13: Set $\theta_{t+1} = \theta_t$
14: **fim se**
15: **fim para**

- **Definir distribuição da proposta:** Escolha uma distribuição da proposta $q(\theta'|\theta_t)$ que sugira um novo valor de parâmetro θ' dado o valor atual $\theta_t = \mathbf{X}$, onde $\mathbf{x} \in \mathscr{E}$ (linha 2 no Algoritmo 1);

- **Iterar valor com base no gradiente dos dados observados:**

  - **Gerar candidato:** Gerar um valor de parâmetro candidato θ' a partir da distribuição da proposta (linha 4 no Algoritmo 1);

  - **Probabilidade de aceitação:** Calcule a probabilidade de aceitação $\alpha(\theta_t, \theta')$ usando a razão das densidades não normalizadas da distribuição alvo $p(\theta')$ e da distribuição proposta $q(\theta'|\theta_t)$ (linhas 5-7 no Algoritmo 1):

$$\alpha(\theta_t, \theta') = \min\left(1, \frac{p(\theta')}{p(\theta_t)} \cdot \frac{q(\theta_t|\theta')}{q(\theta'|\theta_t)}\right)$$

  - **Aceitar ou rejeitar:** Aceite o candidato θ' com probabilidade α, caso contrário, mantenha o valor atual θ_t (linhas 8-14 no Algoritmo 1);

  - **Atualização Robbins-Monro:** Se o candidato for aceito, atualize a estimativa do parâmetro usando a regra de atualização Robbins-Monro (linha 9-11 no Algoritmo 1):

$$\theta_{t+1} = \theta_t + a_t \cdot (f(\theta') - b_t)$$

    onde $f(\theta')$ é uma função relacionada à distribuição alvo, a_t é um tamanho de passo positivo e b_t é a estimativa atual de $f(\theta_t)$.

Em resumo, o método MHRM é uma abordagem poderosa para estimação de parâmetros em modelos complexos com dados de alta dimensão. Sua combinação de simulação de Monte

Carlo e otimização por meio do algoritmo Robbins-Monroe o torna um recurso valioso para pesquisadores e analistas envolvidos em tarefas que computam base de dados de grandes dimensões (CAI, 2010; MAYDEU-OLIVARES; CAI; HERNÁNDEZ, 2011).

2.2 Redes Neurais

As Redes Neurais são uma subárea da inteligência artificial e caracterizam-se por serem sistemas com uma topologia de rede composta por várias camadas de nós. Cada nó realiza um processamento simples que produz ativações de valores reais, as quais são repassadas para os nós subsequentes da próxima camada (SCHMIDHUBER, 2015). A ação de processamento individual que produz informações relativas e independentes em cada nó é o que o assemelha aos neurônios do cérebro humano (HAYKIN, 1994; GOODFELLOW; BENGIO; COURVILLE, 2016). Os nós, ou neurônios, são ativados a partir dos dados de entrada, que dependem da origem ou natureza do dado (textos, imagens ou áudios). Os nós seguintes são ativados a partir de conexões dos pesos dos nós anteriores. O aprendizado, inerente à rede, consiste em encontrar os pesos para executar a ação desejada e, a depender da complexidade do problema, a rede é composta por várias etapas computacionais para transformar diversos conjuntos de ativação (HAYKIN, 1994; SCHMIDHUBER, 2015).

2.2.1 Princípios da Rede Neural

Um nó é uma unidade de processamento de informação fundamental para compor uma rede neural. Cada nó é composto por conexões, um somatório das conexões e uma função de ativação. Uma conexão é iniciada a partir de um dado de entrada x e é multiplicada pelo peso w. O somatório das conexões é a etapa em que o produto do peso e da entrada, $w \times x$, é somado e ajustado pelo viés b. A função de ativação é aplicada para limitar a amplitude da saída dos dados (HAYKIN, 1994; GOODFELLOW; BENGIO; COURVILLE, 2016).

A Figura 6 ilustra o modelo de processamento de um neurônio. A entrada de dados pode ser composta por vários nós, um para cada sinal de entrada. Assim, x_j é descrito como um vetor de entrada, onde $x_j = \{x_1, x_2, ..., x_m\}$. Existe um peso $W_{kj} = \{w_{k1}, w_{k2}, ..., w_{km}\}$ para cada valor de $x_j \in \{1, 2, ..., m\}$, em que o primeiro subscrito da notação refere-se ao peso do neurônio correspondente e o segundo subscrito ao x_j de entrada. A Equação 2.10 formaliza o somatório dos produtos entre w_{kj} e x_j, produzindo o vetor u_k (HAYKIN, 1994):

$$u_k = \sum_{j=1}^{m} w_{kj} x_j. \tag{2.10}$$

O viés é um parâmetro externo ao nó. O número de vieses corresponde à mesma quantidade k de neurônios, formando um vetor $b_k = b_1, b_2, ..., b_k$. Após o somatório, o vetor u_k é ajustado por b_k, que pode incrementar ou decrementar um valor com o objetivo de ajustar a saída

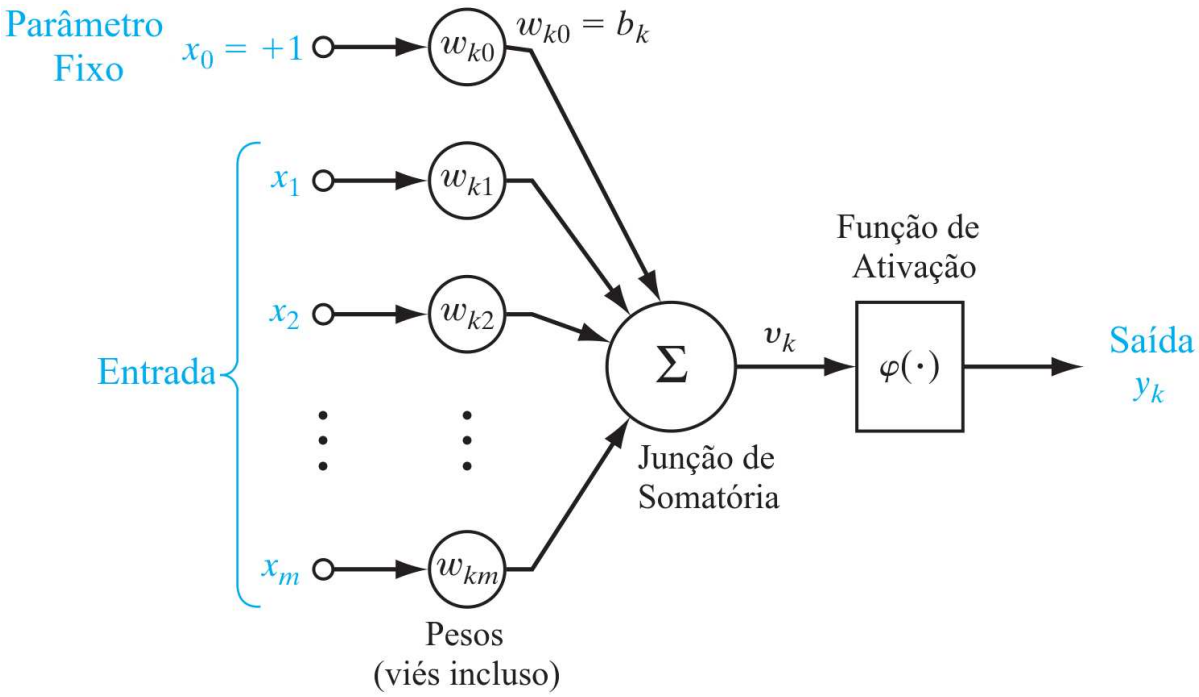

Figura 6 – Modelo não linear de um neurônio com w_{k0} para representar o viés b_k (HAYKIN, 1994)

do neurônio. O resultado desse ajuste é representado por v_k na Figura 6 e é transformado por uma função de ativação que produzirá uma saída que representa a ativação do neurônio em resposta ao dado de entrada. Essa transformação pode resultar em uma escala de [0, 1] , em alguns casos, [-1, 1], entre outras. As Equações 2.11 e 2.12 expressam o ajuste de u_k por b_k e a ativação de v_k, respectivamente (HAYKIN, 1994):

$$v_k = u_k + b_k, \tag{2.11}$$

$$y_k = \varphi(v_k). \tag{2.12}$$

Em termos gerais, a equação completa pode ser expressa por:

$$y_k = \varphi(\sum_{j=1}^{m} w_{kj} x_j + b_k). \tag{2.13}$$

A ativação de uma rede neural, expressa por φ na Equação 2.12, transforma a saída dos dados na representação desejada. A transformação pode resultar em valores contínuos ou discretos, dependendo da função selecionada. Uma das funções de ativação mais comuns é a função *threshold* (limiar), na qual o valor de saída é igual a 1 se a entrada for maior ou igual a 0, e 0 se for menor que 0. A Equação 2.14 apresenta a formalização matemática da função *threshold*, e a Figura 7 ilustra seu comportamento no gráfico (HAYKIN, 1994).

$$\varphi(v) = \begin{cases} 1 & se\ v \geq 0 \\ 0 & se\ v < 0 \end{cases} \tag{2.14}$$

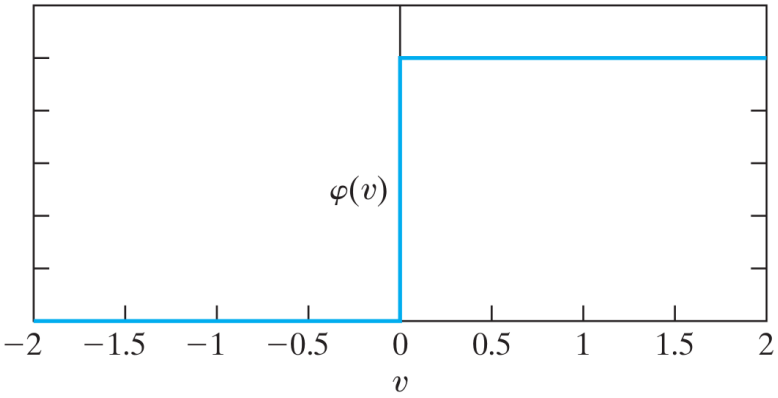

Figura 7 – Função *Threshold* (HAYKIN, 1994)

A função de ativação sigmoide possui um formato em S e é uma das mais comumente utilizadas em aplicações de redes neurais. A função é definida de tal forma que, para todo $x \in \mathbb{R}$, $f(x)$ é limitada entre 0 e 1. A Equação 2.15 descreve a função sigmoide:

$$\varphi(v) = \frac{1}{1+e^{-av}}, \tag{2.15}$$

onde a é o parâmetro de inclinação da função. À medida que v se aproxima de $+/-\infty$, o valor de $\varphi(v)$ tende a se tornar uma constante. As funções sigmoides têm a característica de serem deriváveis, enquanto as funções de *threshold* não são deriváveis. A Figura 8 ilustra o comportamento da função sigmoide no gráfico (HAYKIN, 1994).

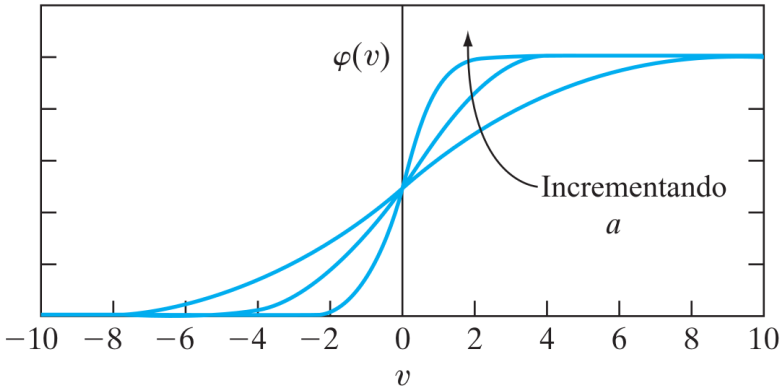

Figura 8 – Função Sigmoide (HAYKIN, 1994)

Entre outras funções de ativação, existem a função *Rectified Linear Unit* (ReLU), com limite inferiror no 0, e a função tangente hiperbólica, $tanh(v)$, com valores entre -1 e 1, denotadas pelas Equações 2.16 e 2.17, respectivamente (HAYKIN, 1994; GOODFELLOW; BENGIO; COURVILLE, 2016):

$$\varphi(v) = v^{+} = max(0, v), \tag{2.16}$$

$$\varphi(v) = tanh(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}}. \tag{2.17}$$

A próxima etapa do aprendizado do neurônio consiste em receber os dados de saída e avaliar se estão compatíveis com os dados de entrada. Caso estejam muito discrepantes, o erro é retroalimentado à junção de somatório, e os pesos e vieses são ajustados. Essa ação é conhecida como *feedback* e é o princípio fundamental para o processo de retropropagação, ou *backpropagation*, de uma rede neural.

Realizar *feedback* não é uma prática complexa para uma rede com apenas uma camada de saída, conhecida como Perceptron (HAYKIN, 1994). Entretanto, à medida que a rede cresce em profundidade e complexidade, ou seja, uma rede composta por múltiplas camadas de perceptrons, a retropropagação se torna mais complexa (HAYKIN, 1994; GOODFELLOW; BENGIO; COURVILLE, 2016). Os algoritmos de retropropagação têm o objetivo de ajustar os parâmetros internos da rede, que são utilizados para computar a informação em cada camada, com base nos dados provenientes da camada anterior (LECUN; BENGIO; HINTON, 2015). Em outras palavras, a retropropagação é a aplicação da técnica de gradiente descendente aos pesos da rede, computando as derivadas parciais de uma função de aproximação $F(w,x)$ em relação a todos os elementos do vetor de pesos ajustáveis w, dado um vetor de entrada x. Essas derivadas parciais são calculadas até que se atinja o custo mínimo dos parâmetros (HAYKIN, 1994). A Figura 9 ilustra o dado representado pelo ponto vermelho e a seta preta representando as derivadas parciais até atingir o custo mínimo, na parte inferior do gráfico.

Para computar o processamento direto de uma rede neural, cada camada constitui um módulo pelo qual os gradientes são retropropagados. Primeiro, é calculada a somatória do produto dos pesos w e das entradas x de cada nó da camada, $z = \sum wx$, e em seguida, é aplicada uma função não linear em z para produzir a saída do nó. A Figura 10 apresenta o processamento direto dos dados, com duas camadas ocultas e uma camada de saída. Cada camada é identificada com os subscritos i, j, k e l. Os vieses foram omitidos para fins de simplicidade (LECUN; BENGIO; HINTON, 2015).

O processamento reverso, a retropropagação, é a computação dos erros derivativos de cada camada. Com base na ilustração da Figura 11, na camada de saída, o erro derivativo em relação à saída do nó é a derivada da função de custo (J) :

$$\frac{\partial E}{\partial y_l} = y_l - t_l, \tag{2.18}$$

se $J = 0,5(y_l - t_l)^2$ e t_l é o valor real. Em seguida, é computado a erro derivativo em relação a z_l,

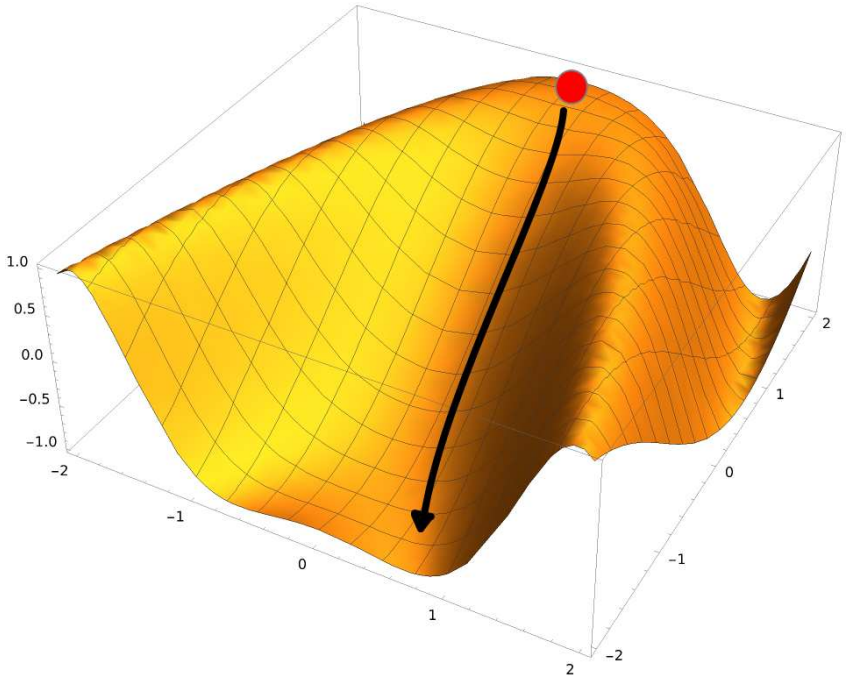

Figura 9 – Aplicação da técnica gradiente descendente (Fonte: Elaborado pelo autor).

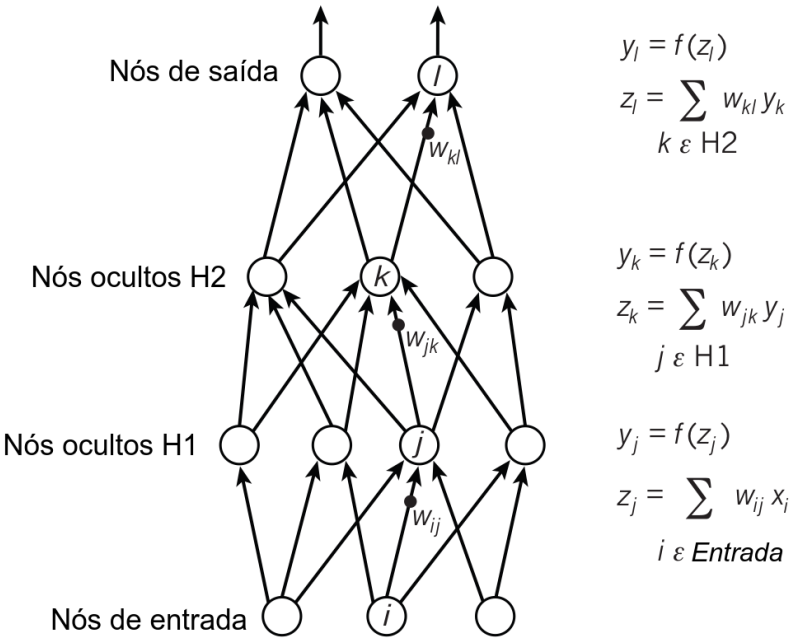

Figura 10 – Rede neural no sentido direto do processamento (LECUN; BENGIO; HINTON, 2015).

expressa por:

$$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l}\frac{\partial y_l}{\partial z_l}. \tag{2.19}$$

Na camada k, calcula-se a derivada do erro em relação à saída de cada nó $\partial E/\partial y_k$, que é o erro derivado da somatória do produto dos pesos e a derivada do erro de z_l (LECUN; BENGIO; HINTON, 2015):

$$\frac{\partial E}{\partial y_k} = \sum_{l \,\varepsilon\, saída} w_{kl} \frac{\partial E}{\partial z_l}. \tag{2.20}$$

A derivada é o erro retropropagado do nó na camada l para a camada k, como pode ser observado na Figura 11. Em seguida, é calculado o erro derivativo em relação a z_k (LECUN; BENGIO; HINTON, 2015):

$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}. \tag{2.21}$$

O mesmo processo é executado na camada H_1, no sentido de k para j, onde, conhecendo $\partial E/\partial z_k$, a derivada do erro do peso w_{jk} do nó j para a camada de entrada, é $y_j \partial E/\partial z_k$ (LECUN; BENGIO; HINTON, 2015).

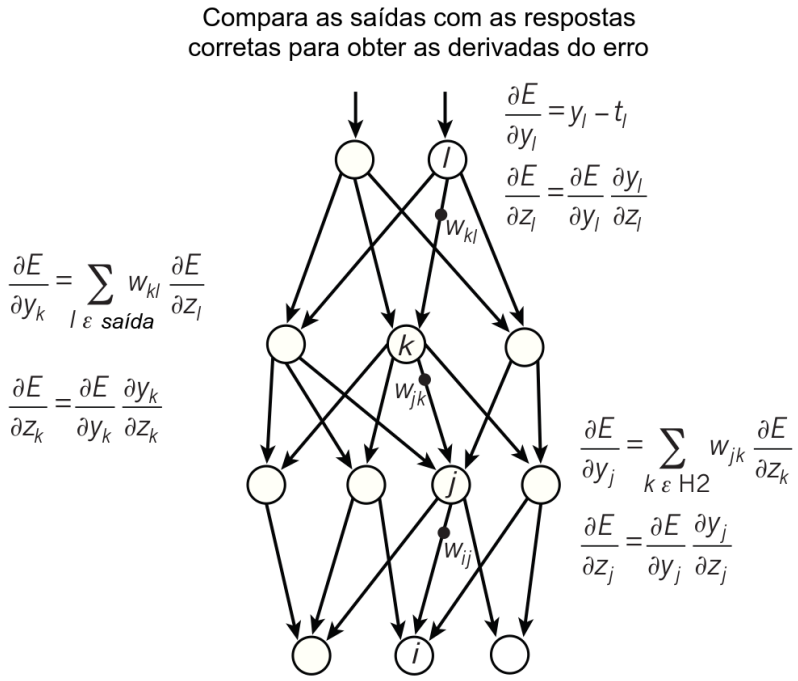

Figura 11 – Rede neural no sentido reverso do processamento (LECUN; BENGIO; HINTON, 2015).

2.2.2 Variational Autoencoders

Em redes neurais, podem-se encontrar diversas topologias de redes, com propósitos variados. As diferentes formas de conexão dos nós é um fator importante para o funcionamento da rede e sua relação com o algoritmo de aprendizado (HAYKIN, 1994; MIIKKULAINEN, 2010). A *Variational Autoencoder* é uma arquitetura de rede neural pertencente à classe das Redes Generativas (HAYKIN, 1994; GOODFELLOW; BENGIO; COURVILLE, 2016). A VAE, além de possuir uma topologia característica, implementa modelos probabilísticos e inferências bayesianas para treinar a rede, ou seja, realiza inferências e aprendizado em modelos diretos

de probabilidade, que visam tratar variáveis latentes contínuas em distribuições posteriores de grandes conjuntos de dados (KINGMA; WELLING, 2019). Uma peculiaridade de destaque da VAE é a de ser um método não supervisionado.

A arquitetura da VAE é composta por dois modelos, o modelo recognitivo (codificador ou encoder) e o modelo generativo (decodificador ou decoder). O modelo recognitivo transforma os dados de entrada em uma aproximação da distribuição a posteriori de variáveis aleatórias latentes, a partir da atualização dos parâmetros para maximizar as expectativas do aprendizado. O modelo generativo realiza o processo inverso, pois, com base na Equação de Bayes 1763, a decodificação é uma aproximação das variáveis latentes para aprender uma representação significativa dos dados de entrada(KINGMA; WELLING, 2019). A Figura 12 ilustra a estrutura de uma arquitetura VAE.

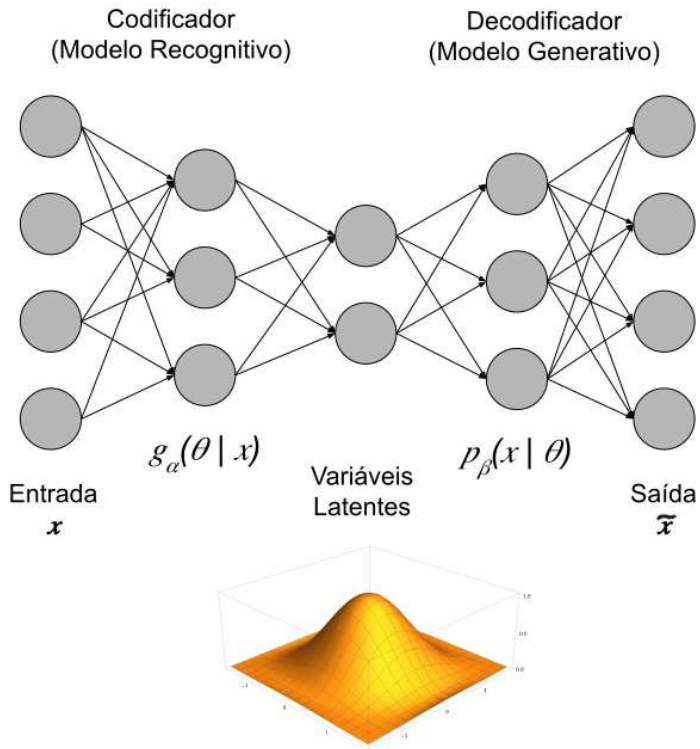

Figura 12 – Representação gráfica da rede neural *Variational Autoencoder*. Os vetores α e β representam os parâmetros das redes codificadora e decodificadora, respectivamente (Adaptado de Curi *et al.* (2019))

Com base na Figura 12, no modelo recognitivo, a distribuição a posteriori das variáveis latentes θ, dado as variáveis de entrada x, é inferida a partir da equação:

$$f(\theta|x) = \frac{p(x|\theta)f(\theta)}{p(x)}. \tag{2.22}$$

Devido ao fato de que $p(x) = \int p(x|\theta)f(\theta)d\theta$ se torna uma distribuição intratável a medida que o número de dimensões aumentam, faz-se necessário aproximar $f(\theta|x)$ por uma

função estocástica das variáveis de entrada e é denotada por $g(\theta|x)$, que depende dos parâmetros α do encoder para sua obtenção (KINGMA; WELLING, 2019):

$$g(\theta|x) \approx f(\theta|x). \tag{2.23}$$

A principal etapa proposta por Kingma e Welling (2019) para solucionar a aproximação da distribuição da Equação 2.23, é a parametrização de $g_\alpha(\theta|x)$ por meio de uma rede neural profunda em que α incluem os pesos e os vieses da rede. As Equações 2.24, 2.25 e 2.26 formalizam matematicamente o processo, onde a saída do encoder representa parâmetros de uma distribuição de probabilidade e viabilizam o modelo generativo decodificador (Equação 2.25). Para facilitar o processo de otimização no treinamento da rede, há necessidade de uma reparametrização, especificada em 2.26.

$$(\mu, log\ (\sigma)) = Encoder_\alpha(x), \tag{2.24}$$

$$g_\alpha(\theta|x) : \text{gera } \theta|x \sim \mathcal{N}(\mu, diag(\sigma)), \tag{2.25}$$

$$\theta = \mu + \sigma \times \varepsilon, \text{ em que } \varepsilon \sim \mathcal{N}(0,1). \tag{2.26}$$

A distribuição conjunta $f(x,\theta)$, usando a fórmula de Bayes, pode ser expressa por $f(x,\theta) = f(\theta)p(x|\theta)$, ou seja, o produto de uma distribuição a priori assumida para o traço latente e um decoder estocástico. O encoder aproxima a verdadeira distribuição a posteriori, conforme Equação 2.23, minimizando a Divergência de Kullback-Leibler (KL). Para qualquer estrutura definida no encoder, tem-se:

$$\begin{aligned} log\ p(x) &= \mathbb{E}_{g_\alpha(\theta|x)}\left[log\ p(x)\right] \\ &= \mathbb{E}_{g_\alpha(\theta|x)}\left[log\ [\frac{p(x,\theta)}{f(\theta|x)}\right]\right] \\ &= \mathbb{E}_{g_\alpha(\theta|x)}\left[log\left[\frac{p(x,\theta)}{g_\alpha(\theta|x)}\frac{g_\alpha(\theta|x)}{f(\theta|x)}\right]\right] \\ &= \mathbb{E}_{g_\alpha(\theta|x)}\left[log\left[\frac{p(x,\theta)}{g_\alpha(\theta|x)}\right]\right] + \mathbb{E}_{g_\alpha(\theta|x)}\left[log\left[\frac{g_\alpha(\theta|x)}{f(\theta|x)}\right]\right], \end{aligned} \tag{2.27}$$

onde:

$$\text{ELBO} = \mathbb{E}_{g_\alpha(\theta|x)}\left[log\left[\frac{p(x,\theta)}{g_\alpha(\theta|x)}\right]\right], \tag{2.28}$$

$$D_{KL}(g_\alpha(\theta|x)||f(\theta|x)) = \mathbb{E}_{g_\alpha(\theta|x)}\left[log\left[\frac{g_\alpha(\theta|x)}{f(\theta|x)}\right]\right]. \quad (2.29)$$

Portanto, a divergência KL tem dois objetivos, (a) determinar a distância (divergência) entre as posterior aproximada e a posterior real; e, (b) a lacuna entre ELBO e a verossimilhança marginal $log\ p(x)$. Quanto mais próxima $g_\alpha(\theta|x)$ está da posterior $f(\theta|x)$, com base na divergência KL, menor a lacuna (KINGMA; WELLING, 2019).

2.2.3 Modelos Transformers

Os modelos *Transformers* representam o estado da arte entre os modelos destinados a tarefas de Processamento de Linguagem Natural (*Natural Language Processing* (NLP)) e, mais recentemente, têm sido amplamente utilizados para classificação e reconhecimento de imagens. A arquitetura do *Transformer* oferece várias vantagens sobre as redes neurais recorrentes ou convolucionais tradicionais. Primeiramente, ela depende exclusivamente de mecanismos de atenção, o que possibilita um treinamento mais rápido e paralelizável. Em segundo lugar, ela atinge resultados superiores em tarefas de tradução automática, superando os modelos tradicionalmente utilizados. Por fim, lida de forma eficiente com grandes volumes de dados, como imagens, áudio e vídeo, por meio de mecanismos de atenção localizados e restritos (VASWANI *et al.*, 2017).

O modelo *Transformer* é composto por um codificador e um decodificador. O codificador consiste em 6 camadas idênticas, cada uma contendo um mecanismo de *multi-head self-attention* e uma subcamada de rede *feed-forward* totalmente conectada, que leva em consideração a posição de cada dado de entrada. Em cada subcamada, uma conexão residual é empregada, seguida pela normalização da camada. A normalização da camada segue o formato $LayerNorm(x + Sublayer(x))$, onde $Sublayer(x)$ é a função implementada pela subcamada correspondente. O codificador recebe uma sequência de representações de símbolos $x_i = x_1, x_2, ..., x_i$ e a converte em uma sequência de representações contínuas $z_i = z_1, z_2, ..., z_i$.

O decodificador, por sua vez, recebe z como entrada e gera uma sequência de símbolos $y_i = y_1, y_2, ..., y_i$. Assim como o codificador, o decodificador também é composto por 6 camadas idênticas, cada uma contendo 3 subcamadas. A terceira subcamada executa multi-head attention sobre a saída do codificador. Conexões residuais e normalização de camada também são aplicadas no decodificador, porém, a subcamada de *self-attention* é modificada para evitar o cálculo das posições subsequentes, garantindo que as previsões para a posição i dependam apenas de saídas conhecidas, anteriores a i. (VASWANI *et al.*, 2017).

De acordo com Bahdanau (2014), o mecanismo de atenção no *Transformer* estabelece dependências globais entre a entrada e a saída sem depender de recorrência. Especificamente, ele calcula uma soma ponderada dos valores com base em uma consulta e em um conjunto de chaves, em que os pesos são determinados por uma função Softmax aplicada ao produto escalar

entre a consulta e as chaves. Isso permite que o modelo se concentre nas partes mais relevantes da sequência de entrada ao gerar cada parte da sequência de saída. Esse mecanismo de atenção aprimora a qualidade da tradução ao permitir que o modelo lide melhor com dependências de longo alcance e reordenações de palavras, que são desafios comuns na tradução automática . (VASWANI *et al.*, 2017).

Os mecanismos de atenção podem ser classificados em dois tipos: atenção escalada por multiplicação escalar e *multi-head attention*. Um mecanismo de atenção escalada por multiplicação escalar utiliza consultas de entrada e chaves com dimensão d_k, além de valores com dimensão d_v. A multiplicação escalar entre a consulta e cada chave é computada, dividida por $\sqrt{d_k}$, e uma função Softmax é aplicada para obter os pesos associados aos valores. A função de atenção é calculada simultaneamente para um conjunto de consultas, que são compactadas em uma matriz Q. Da mesma forma, as chaves e os valores são compactados em matrizes K e V, respectivamente. A matriz de saídas é calculada como descrita na equação abaixo (VASWANI *et al.*, 2017):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.30}$$

Em vez de usar um único conjunto de chaves, valores e consultas de dimensão d_{model} para o mecanismo de atenção, são utilizadas projeções lineares aprendidas que os projetam h vezes nas dimensões d_k e d_v, respectivamente. A função de atenção é então executada em paralelo em cada um desses conjuntos projetados, produzindo valores de saída de dimensão d_v. Esses valores são concatenados e, em seguida, projetados novamente para gerar os valores finais. Com uma única cabeça de atenção, a capacidade do modelo de capturar informações provenientes de diferentes subespaços de representação em posições distintas seria limitada. O uso de múltiplas cabeças permite ao modelo atender a diferentes partes da entrada simultaneamente.

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1,...,head_h)W^O, \\ \text{onde } head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \tag{2.31}$$

onde as matrizes de parâmetros da projeção são $W_i^Q \in \mathbb{R}^{d_{model}*d_q}$, $W_i^K \in \mathbb{R}^{d_{model}*d_k}$, $W_i^V \in \mathbb{R}^{d_{model}*d_v}$ e $W^O \in \mathbb{R}^{hd_v*d_{model}}$ (VASWANI *et al.*, 2017). A figura 13 ilustra os mecanismos de atenção por multiplicação escalar e *multi-head*.

2.2.4 Modelos de Deep Learning para Imputação de Dados Faltantes

Redes neurais e técnicas de *Deep Learning* têm sido amplamente empregadas na imputação de dados faltantes. Essas redes oferecem a capacidade de adaptar tanto métodos tradicionais quanto modernos para estimar grandes volumes de dados, mantendo a eficiência computacional.

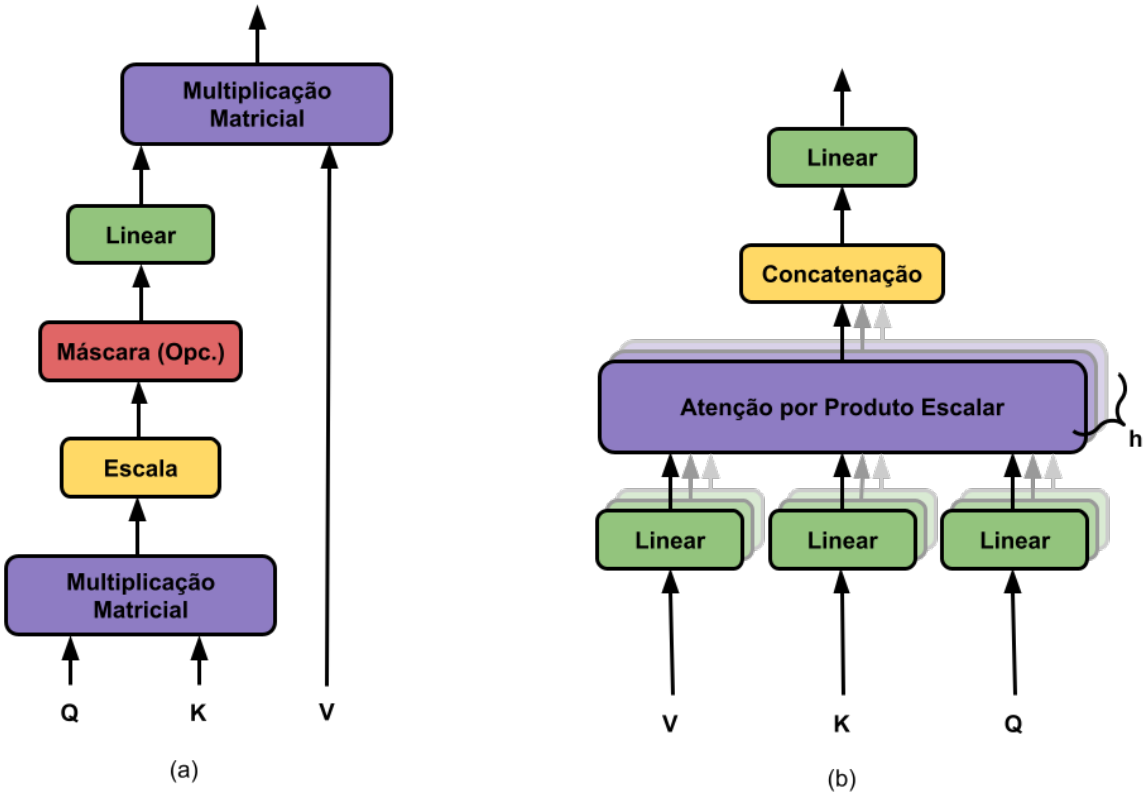

Figura 13 – Mecanismos de atenção: (a) Atenção por Produto Escalar e (b) Atenção de Múltiplas Cabeças (VASWANI *et al.*, 2017).

No estudo conduzido por Cheng et al. (2019), uma Rede Neural Profunda, (*Deep Neural Network* (DNN), foi empregada para processar diversos tipos de dados de entrada, como respostas de diferentes questionários e escalas utilizadas no diagnóstico de indivíduos com TDAH (Transtorno de Déficit de Atenção e Hiperatividade). Após o treinamento do modelo, os atributos com maior acurácia preditiva (itens dos questionários) foram selecionados para realizar a imputação das respostas ausentes. Em seguida, os dados imputados foram avaliados por meio de uma Support Vector Machine (SVM) para classificar se o indivíduo apresentava diagnóstico de TDAH ou ainda se estava em processo de desenvolvimento do transtorno.

A arquitetura utilizada por Chen *et al.* (2015); Lin, Tsai e Zhong (2022); e Lin *et al.* (2020), baseada e uma *Deep Belief Network* (DBN), funciona de forma semelhante a um *Autoencoder*, reduzindo o número de nós na camada de saída em comparação à camada de entrada, com o objetivo de categorizar de maneira mais eficaz a imputação dos dados faltantes. As DBNs permitem extrair informações relevantes sobre o contexto dos dados ausentes, facilitando, assim, uma imputação mais precisa e contextualizada.

No estudo proposto por Pereira *et al.* (2020), é apresentada uma arquitetura *Autoencoder* para a imputação de dados faltantes. Essa arquitetura é composta por duas partes principais: a

codificadora, responsável por aprender as informações dos dados de entrada e reduzi-las a um espaço latente, e a decodificadora, que reconstrói os dados a partir das informações aprendidas. Dessa forma, em casos de dados ausentes, o *Autoencoder* pode ser treinado utilizando apenas os dados observados. Para cada entrada incompleta, o modelo é capaz de inferir os valores ausentes com base nos padrões aprendidos dos dados observados durante o treinamento. Uma vez treinado, o *Autoencoder* pode ser utilizado para imputar valores ausentes. Para uma entrada com valores ausentes, o codificador processa os dados disponíveis para criar uma representação latente. Então, o decodificador utiliza as variáveis latentes para gerar uma saída completa, imputando, assim, os valores ausentes. Essa aplicação para o *Autoencoder* é também conhecida como *Denoising Autoencoder* (DAE) e tem sido aplicado em tarefas que exige imputação dados faltantes, ou seja, ruídos.

No estudo de Smieja et al. (2018), é proposta uma arquitetura onde os dados ausentes são representados por densidades de probabilidade condicional. Especificamente, cada ponto de dados incompleto é associado a um subespaço definido pelos atributos observados, enquanto os atributos não observados são modelados usando uma distribuição de probabilidade, como Modelos Gaussianos Mistos. As funções de ativação na primeira camada oculta da rede neural são modificadas para calcular o valor esperado da resposta do neurônio, com base na função de densidade de probabilidade que representa os dados ausentes. Dessa forma, a rede neural calcula a ativação média considerando os valores possíveis dos atributos ausentes. Essa abordagem permite que a rede seja treinada em conjuntos de dados que contenham amostras incompletas, uma vez que os parâmetros da função de densidade de probabilidade são aprendidos simultaneamente com os parâmetros da rede durante o processo de treinamento.

No estudo descrito por Gad *et al.* (2021), o modelo proposto utiliza uma arquitetura de Rede Neural Convolucional *Convolutional Neural Network* (CNN) 1D para imputar valores ausentes em dados meteorológicos. As CNNs realizam convoluções, que consistem em uma soma ponderada dos elementos vizinhos de uma matriz, com os pesos definidos por um filtro convolucional. Nesse contexto, os dados meteorológicos de entrada são representados como uma matriz 1D, enquanto um filtro 1D é aplicado para realizar as convoluções. O modelo aproveita a capacidade das CNNs de capturar padrões espaciais e temporais presentes nos dados meteorológicos, o que permite uma imputação mais eficaz dos valores ausentes.

Nesta seção foram apresentadas diversas aplicações de redes neurais para imputação de dados ausentes. A flexibilidade da rede neural mostra que pode ser aplicada em tarefas com dados de diferentes naturezas. Além disso, as redes oferecem diversas arquiteturas que podem ser adaptadas e exploradas para imputação de dados.

CAPÍTULO

3

DESENVOLVIMENTO

Neste capítulo de Desenvolvimento são descritos a criação dos dados artificiais para simular os experimentos do modelo, a extração dos dados reais a partir das respostas dos questionários do PISA 205, o modelo desenvolvido utilizando *Transformer* para imputação de dados e o processo de estimação de parâmetros de itens e pessoas utilizando a arquitetura de variational autoencoder com incorporação do modelo de TRI no decodoficador (VAEQ).

3.1 Criação dos Dados Artificiais

Esta seção descreve os métodos e processos para gerar dados artificiais e realizar o treinamento do modelo *Discrete Transformer Masked Autoencoder* (DiT-MAE) para imputar dados ausentes. Após a imputação, os parâmetros de itens e pessoas são estimados a partir da arquitetura VAEQ, descrito mais adiante, para fins de comparação e estudo da eficiência da proposta. Existem dois grupos de conjuntos de dados, cada grupo contendo um número diferente de itens e todos eles contendo 10.000 indivíduos. Os conjuntos de dados são divididos em: um grupo com 28 e 56 itens e variáveis latentes com 3 dimensões, e o segundo grupo com quatro conjuntos de dados de 90, 180, 270 e 360 itens e variáveis latentes com 21 dimensões. Nas subseções a seguir são descritas a criação de dados artificiais para os grupos de 3 e 21 dimensões junto com o seu algoritmo.

Para gerar os dados artificiais, é necessário definir o valor das variáveis correspondentes ao número de indivíduos (i), quantidade de itens (j), número de dimensões (d) e a matriz Q de habilidades, cujos elementos identificam qual (ou quais) variáveis latentes são exigidas para responder adequadamente cada item. Lembrando que pelo menos uma das habilidades precisa ter o valor 1 para cada item. Dessa forma, de acordo com o número de dimensões, a matriz de habilidades é definida como:

$$Q = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1j} \\ q_{21} & q_{22} & \cdots & q_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ q_{i1} & q_{i2} & \cdots & q_{ij} \end{bmatrix}, \text{ onde cada } q_{ij} \in \{0,1\} \text{ e } \sum_{n=1}^{j} q_{ij} \geq 1 \text{ para todo } i \in \{1,2,...,J\}.$$

O próximo passo é gerar valores das discriminações reais (a_{real}) a partir da distribuição logNormal com média $\mu = 0$, desvio padrão $\sigma = 0,4$ e valores mínimo de 0,6 e máximo de 2,3 desejados. Sendo assim, a distribuição dos parâmetros de discriminação reais adotada é:

$$A \sim LogN(\mu, \sigma^2).$$

Ressalta-se que a quantidade de valores reais produzidos equivale ao total de itens j multiplicado pelo número de dimensões d e subtraindo a quantidade de elementos iguais a zero na matriz Q.

As dificuldades reais (b_{real}) são geradas a partir de uma distribuição Normal com $\mu = 0$ e $\sigma = 1$, conforme a definição abaixo:

$$B \sim N(\mu, \sigma).$$

O total de valores reais gerados é igual à quantidade de itens j.

Para os valores dos traços latentes (θ), primeiro cria-se uma matriz identidade I_d, em que d representa as dimensões do espaço das habilidades. Em seguida, é realizada a geração dos valores a partir de uma Normal Multivariada Aleatória com dimensões i indivíduos e d habilidades. Abaixo, a descrição formal para a simulação dos $\theta's$ e o ajuste dos valores com a média:

$$\theta = z, \text{ onde } z \sim N(0, I_d)$$

A partir do a_{real} é gerada a matriz de a inicializada com zeros (0), com as mesmas dimensões de Q em que, para cada posição de $q_{ij} = 1 \Rightarrow a_{ij} = 1$, mantendo os zeros nas suas devidas posições. O mesmo é realizado para os b_{real}, fazendo uma cópia com b, e com θ_{real}, criando uma cópia de θ. Os valores reais de a_{real}, b_{real} e θ_{real} são preservados e utilizados para fins de comparação com as estimativas da arquitetura VAEQ, enquanto a, b e θ serão utilizados para gerar as respostas aos itens simuladas com base na equação 2.4.

Com os dados simulados prontos, é preciso forçar situações de dados ausentes para todos os conjuntos de dados. Isso envolve simular dados faltantes, mascarando alguns valores das respostas dentro das porcentagens de 10%, 25% e 50%. Os valores ausentes podem assumir o valor de -1 em um conjunto de dados discretos.

3.2 Extração dos Dados Reais: PISA 2015

Esta seção explica como são organizados e selecionados os dados de respostas do PISA 2015. O PISA é um questionário que possui dados faltantes por delineamento da própria aplicação da avaliação (*missing by design*) (OECD, 2016). O termo de *missing by design* refere-se a dados ausentes de forma proposital e planejada, que não são coletados para todas as variáveis ou para todos os indivíduos. No caso do PISA, o questionário é projetado com *missing por design* devido às seguintes características:

1. **Estrutura de amostragem rotativa (questionários por blocos):** O PISA utiliza uma metodologia chamada de amostragem em matrizes, onde diferentes blocos de perguntas são distribuídos entre os estudantes. Isso significa que cada estudante responde apenas a uma parte do questionário, e não a todas as perguntas. Dessa forma, evita-se sobrecarregar os estudantes com um questionário longo e cansativo, mas ainda se obtém uma amostra representativa de respostas para todas as perguntas ao nível da população.

2. **Separação por temas:** Os questionários do PISA cobrem diversos tópicos contextuais, como fatores socioeconômicos, práticas de ensino, ambiente escolar, ambiente residencial, etc. Esses temas podem variar entre países diferentes ou dentro de um mesmo país. Para não sobrecarregar os estudantes com temas menos relevantes para alguns países ou contextos específicos, certas perguntas são aplicadas apenas em determinados subgrupos.

3. **Modularidade internacional:** Alguns módulos do questionário são opcionais e podem ou não ser aplicados dependendo do país participante. Cada país pode optar por coletar informações adicionais em módulos específicos, levando a um conjunto de dados com lacunas deliberadas entre países.

Com base nessa informação, para realizar os experimentos do presente trabalho a partir das respostas de provas cognitivas do PISA 2015, os dados foram extraídos das respostas dos questionários cognitivos de ciência [1]. Fez-se necessário identificar as questões de provas respondidas por um mesmo grupo de alunos. Para essa escolha, foram considerados alunos que responderam os questionários por meio do computador (*Computer-Based Assesment* (CBA)). Entre as provas respondidas, sendo 12 disponibilizadas, chamadas de *survey*, foi selecionada a *survey* de código S01 por apresentar a maior quantidade de itens para alunos de diversos países. O processo é realizado a partir do Anexo A do PISA 2015 *Technical Report* 2016:

1. Extração dos códigos de itens respondidos pelo computador dentro do *survey* S01;

2. Campos dos países e códigos dos estudantes são adicionados juntamente com os códigos de itens para identificar os alunos que responderam a esse conjunto de itens;

[1] https://webfs.oecd.org/pisa/PUF_SPSS_COMBINED_CMB_STU_COG.zip

3. A partir dos códigos de itens é extraída a matriz de referência, contendo as habilidades necessárias para responder as questões. As habilidades encontram-se no relatório técnico e são divididas em 3 habilidades principais: Competência, Conhecimento e Sistema. Cada uma dessas são divididas em 3 subcategorias. Para no nosso estudo, utilizamos a habilidade de conhecimento, dividida em Procedural, Conteúdo e Epistêmico.

Após agrupadas as respostas dos alunos para o mesmo *survey*, selecionam-se os países desejados para imputação. Para treinar e testar a imputação do nosso modelo, foram selecionados quatro países, sendo eles Brasil, Canadá, Colômbia e Estados Unidos. Os dados do PISA, quando selecionado um único país, oferecem poucos indivíduos que responderam à prova toda, sem resposta faltante, para treinamento do modelo. Por esse motivo, foram selecionados mais de um país para agregar maior quantidade de informação para aprendizagem do modelo. Dessa forma, é possível analisar o desempenho do modelo no que diz respeito ao aprendizado do contexto de cada indivíduo, durante o treinamento. Do total de respostas, selecionam-se apenas os alunos que não omitiram nenhuma resposta e são separadas 80% para treinamento, 10% para teste e 10% para validação.

A etapa final consiste em avaliar o modelo na imputação de respostas faltantes reais. Isso quer dizer, selecionar um grupo de alunos que não responderam parte das questões. Foram selecionados os indivíduos que omitiram entre 1 a 6 respostas, ou seja, não responderam itens aleatórios dentro da mesma prova.

3.3 Arquitetura DiT-MAE

Esta seção descreve o modelo adaptado a partir do trabalho proposto por He *et al.* (2022) e implementado por Gosthipaty e Paul (2020) que apresenta um modelo *Trasnformer* mascarado para classificação e reconhecimento de imagens. O estudo apresenta uma abordagem simples de aprendizado auto supervisionado para visão computacional usando *autoencoders* mascarados (MAE). A abordagem consiste em mascarar aleatoriamente partes da imagem de entrada e reconstruir os *pixels* ausentes usando uma arquitetura codificador-decodificador assimétrica. O codificador opera apenas no subconjunto visível da imagem (sem máscara), enquanto o decodificador reconstrói a imagem original a partir da representação latente e partes mascaradas. Os autores descobriram que mascarar 75% da imagem de entrada é crucial para criar uma tarefa auto supervisionada significativa. A abordagem permite o treinamento de grandes modelos, melhorando a precisão e o desempenho.

No contexto dos questionários de avaliação educacional, é comum observar uma quantidade significativa de respostas ausentes. Em alguns casos, esse processo de ausência é intencionalmente imposto pelo desenho da aplicação do questionário, uma vez que nem todos os itens são apresentados a todos os participantes. Para lidar com essa situação, a arquitetura proposta por He *et al.* (2022) foi adaptada para processar dados binários no codificador. O processo de imputação

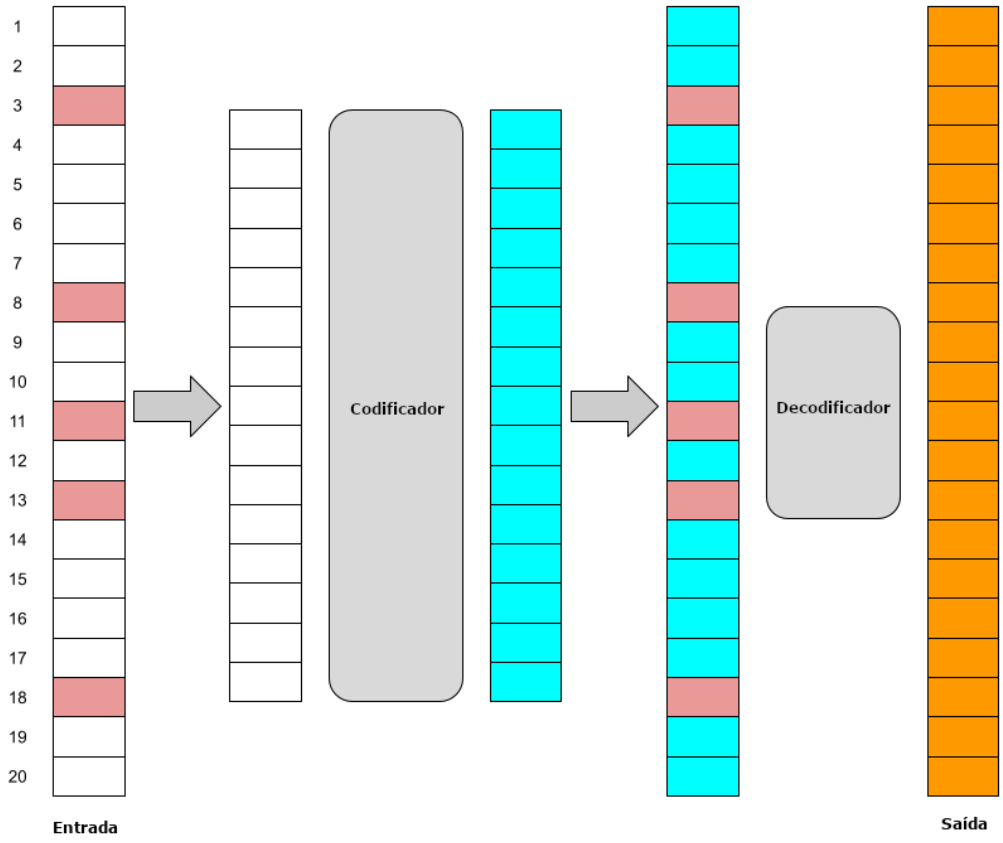

Figura 14 – *Discrete Transformer Masked Autoencoder*. Adaptado de He *et al.* (2022).

se baseia nas respostas existentes dos alunos para treinamento do modelo. Diferentemente dos *patches* de imagem, que contêm uma grande quantidade de informações, o mascaramento de dados discretos foi testado em proporções menores em relação a imagens, de: 10%, 25% e 50%.

A Figura 14 exibe o modelo capaz de processar dados discretos como entrada, aqui proposto. O modelo consiste em 18 itens de dados de entrada, representados por retângulos. Os retângulos vermelhos indicam respostas ausentes. O modelo DiT-MAE treina o codificador usando apenas os dados observados, ou seja, excluindo os valores ausentes. Os valores aprendidos e ausentes do codificador são combinados e enviados ao decodificador para reconstruir a entrada original.

Um recurso essencial no modelo de He *et al.* (2022), é a atualização dinâmica dos *embeddings* posicionais durante cada ciclo de treinamento. O *embedding* posicional desempenha um papel crítico nas NLP's (VASWANI *et al.*, 2017), auxiliando no aprendizado da semântica de sequências textuais. Na implementação descrita por Gosthipaty e Paul (2020), os *embeddings* posicionais são particularmente importantes para a compreensão do contexto dos *patches* em dados de imagem. O *patch* de entrada é transformado em uma matriz e alimentado em uma camada densa, sendo então propagado para uma camada de *embedding* posicional. A cada etapa de treinamento, o *embedding* posicional desmascarado é atualizado, permitindo ao modelo

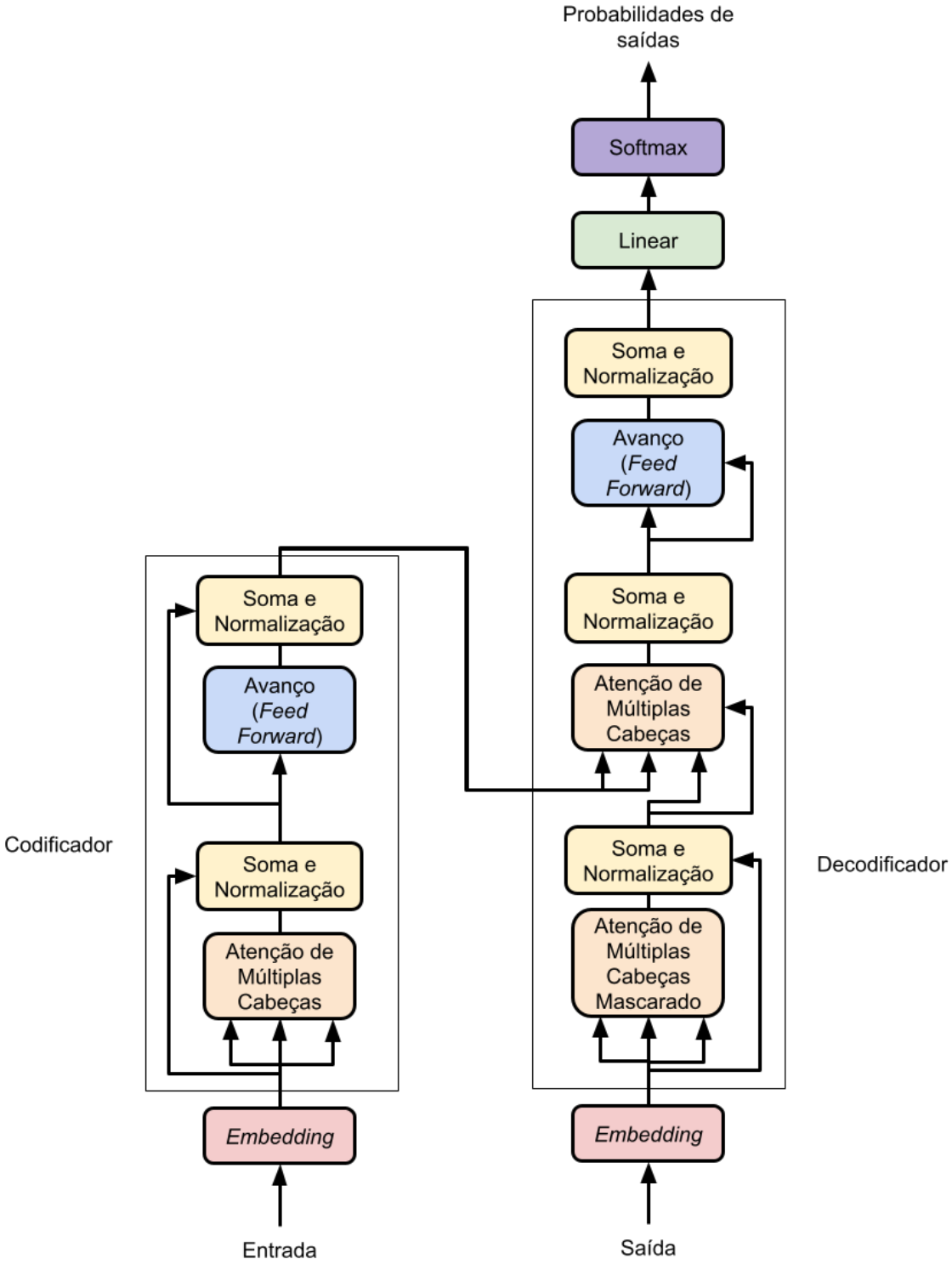

Figura 15 – Modelo *Transformer*. Adaptado de Vaswani *et al.* (2017).

capturar as relações espaciais entre diferentes elementos nos dados. Precisa-se salientar que as informações incorporadas em cada *patch* são cruciais para o sucesso do aprendizado. Mesmo com até 95% de mascaramento, o modelo demonstra uma aprendizagem eficaz, aproveitando os *embeddings* posicionais para a aprendizagem contextual (HE *et al.*, 2022).

A Figura 15 ilustra as etapas de processamento do codificador e decodificador. No codificador, os dados de entrada são incorporados em uma camada de *embedding* para depois serem computados em um mecanismo de Atenção de Múltiplas Cabeças (*Multi-Head Attention*). Logo em seguida, uma camada de soma e normalização dos dados é realizada e processada por uma camada de *Feed Forward* e, novamente, mais uma camada de soma e normalização. No decodificador, diferenciando-se do codificador, existe uma etapa anterior em que os dados mascarados são processados e seus pesos e viés enviados como entrada para o mecanismo de Atenção de Múltiplas Cabeças.

No nosso modelo, os valores aprendidos passam por atualizações principalmente na primeira camada, especificamente na camada densa de projeção. Essa escolha de design é influenciada pela natureza dos valores binários, que carregam informações limitadas em comparação com as informações espaciais presentes nos *patches* de imagem ou a semântica das sequências textuais, para o caso de tarefas de NLP's. Em uma sequência de entrada binária, onde as respostas ausentes podem variar de posição, o modelo adapta seu aprendizado com base no contexto fornecido pelas respostas não mascaradas de cada aluno. Nosso código está disponível no repositório Github[2].

A característica distintiva do nosso modelo reside em sua capacidade de aprender com as diferentes posições e configurações de respostas ausentes dentro de uma sequência binária. Como o contexto de cada aluno é considerado a partir das respostas não mascaradas, o processo de aprendizado do modelo é influenciado de maneira única para cada indivíduo. No entanto, nosso modelo pode enfrentar desafios na estimativa precisa de parâmetros ao lidar com uma alta proporção de dados mascarados, divergindo da abordagem apresentada por He *et al.* (2022). Isso ocorre porque os dados binários carregam menos informações em comparação com dados de sequência de texto ou de imagem, conforme discutido anteriormente.

O próximo passo consiste em substituir os valores ausentes do conjunto de dados original pelos dados imputados gerados pelo modelo DiT-MAE. Isso garantirá a obtenção de um conjunto de dados completo e confiável, apto para ser utilizado em etapas subsequentes. O conjunto de dados imputado poderá, então, ser utilizado para aplicação do modelo VAEQ, um *Variational Autoencoder* adaptado para estimar parâmetros da TRIM, proposto por Curi *et al.* (2019). O codificador da arquitetura VAEQ possui uma dimensão de entrada equivalente ao número de itens a serem estimados, uma camada oculta e uma camada latente cujo tamanho corresponde ao número de habilidades a serem inferidas.

O decodificador tem um papel fundamental na estimativa dos parâmetros dos itens, especialmente no que diz respeito à discriminação (a) e à dificuldade (b). Nessa arquitetura, o decodificador deriva esses parâmetros a partir dos pesos e vieses associados às conexões entre a camada latente e a camada de saída. A função de ativação utilizada no decodificador é a função sigmoide, conforme definida pela Equação 2.2. Para a estimativa dos parâmetros dos itens, o

[2] https://github.com/guilhermemfreire/DiT-MAE/tree/main

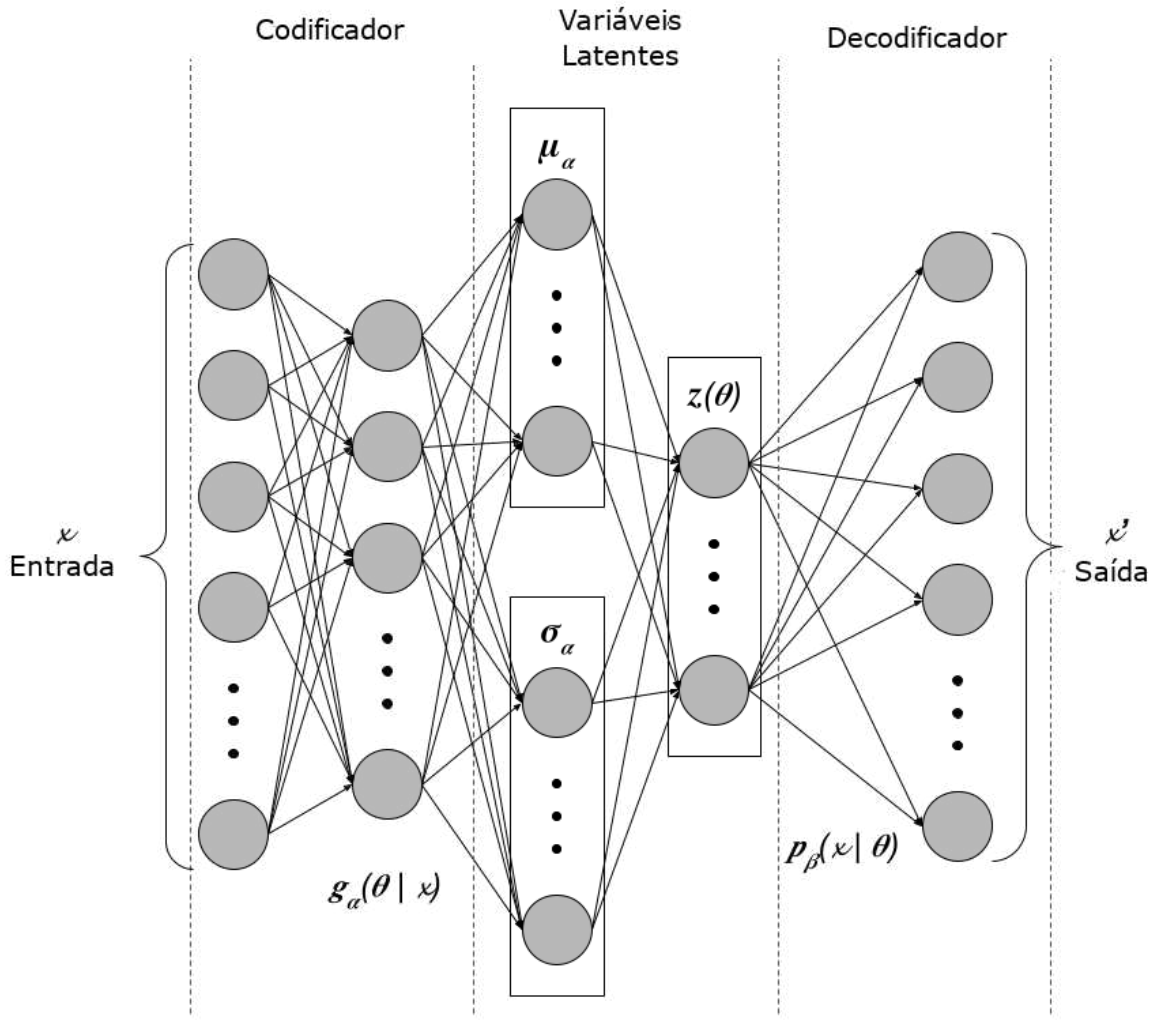

Figura 16 – Arquitetura VAEQ proposta por Curi *et al.* (2019).

decodificador não possui uma camada oculta e é estruturado com uma camada de saída que incorpora uma função de restrição. Essa função de restrição ativa uma matriz Q (Q-Matrix), cujos elementos são iguais a 0 ou 1, dependendo da necessidade de certas habilidades para responder corretamente ao item. Um valor de 1 indica que a habilidade (um nó na camada latente) é necessária para a resposta correta ao item (um nó na camada de saída), enquanto 0 indica que a habilidade não é necessária para responder corretamente ao item (CURI *et al.*, 2019). A arquitetura do VAEQ é ilustrada na Figura 16, que fornece uma representação visual da estrutura e do fluxo do modelo.

Vale destacar que o decodificador definido dessa forma equivale ao MIRT, exatamente como descrito na Equação 2.4: os pesos do decoder equivalem aos parâmetros de discriminação, os viéses equivalem ao parâmetro de dificuldade e a saída dos nós da camada latente aos traços latentes do indivíduo.

CAPÍTULO

4

RESULTADOS

O presente capítulo descreve os resultados e análises obtidas dos experimentos realizados. A Seção 4.1 apresenta os resultados oriundos dos dados artificiais. A Seção 4.2 descreve os resultados de imputação a partir de um conjunto de dados reais, as respostas de questionários cognitivos do PISA 2015.

4.1 Resultados dos Dados Artificiais

Nesta seção apresentamos os resultados do algoritmo DiT-MAE para respostas ausentes em simulações com dados artificiais. Avaliamos o algoritmo imputando respostas ausentes para 10%, 25% e 50% de diferentes conjuntos de números de itens para 10.000 indivíduos. A organização do conjunto de dados é dividida em dois grupos com base em suas dimensões - um grupo tem 3 dimensões, enquanto o outro tem 21 dimensões. O grupo tridimensional contém conjuntos de dados com 28 e 56 itens, enquanto o grupo 21-dimensional contém conjuntos de dados com 90, 180, 270 e 360 itens. Primeiro, apresentamos os resultados para imputação de dados usando o DiT-MAE. Em seguida, estimamos os parâmetros de itens e pessoas utilizando o VAEQ proposto por Curi *et al.* (2019) para compararmos com os resultados gerados com os métodos JML e MHRM, métodos tradicionalmente aplicados para estimar parâmetros da TRIM.

O primeiro processo foi imputar dados ausentes a partir dos dados artificiais. A avaliação de desempenho baseou-se em 100 réplicas dos dados com 10.000 indivíduos, cada uma. Durante os experimentos, notamos que, à medida que as dimensões e o número de itens aumentavam, o tempo de imputação aumentava. Assim, a primeira réplica foi utilizada para treinamento e as demais restantes foram utilizadas para a previsão do modelo. Após passar pela imputação pelo algoritmo DiT-MAE, os conjuntos de dados com 10%, 25% e 50% de valores ausentes apresentaram excelente precisão binária e perda mínima. A Tabela 1 exibe a acurácia binária do DiT-MAE para dois conjuntos diferentes de dimensões e apresenta valores que, quando o conjunto de dados possui baixa dimensão, menor quantidade de itens e uma pequena porcentagem

de valores ausentes, o modelo pode prever com uma acurácia binária de até 98%. No entanto, para conjuntos de dados com um número maior de itens e dimensões e maior taxa de dados faltantes, a precisão de imputação diminui para aproximadamente 74%, oferecendo, assim, resultados satisfatórios.

Tabela 1 – Resultado da acurácia binária para imputar dados faltantes (DF) para 3 e 21 dimensões no treino.

DF	**3 Dimensões**		**21 Dimensões**			
	28 Itens	**56 Itens**	**90 Itens**	**180 Itens**	**270 Itens**	**360 Itens**
10%	0.98	0.96	0.86	0.88	0.78	0.77
25%	0.97	0.94	0.84	0.79	0.76	0.76
50%	0.88	0.91	0.79	0.78	0.77	0.74

Para a segunda etapa, é necessário manipular os dados imputados para que possam ser avaliados utilizando a arquitetura VAEQ. Os resultados das imputações estão representados valores contínuas e, portanto, é necessária a conversão para valores discretos binários. Em seguida, os dados ausentes imputados são preservados, enquanto os dados não mascarados, que foram, também, reconstruídos, são substituídos por seus valores originais correspondentes. Essa abordagem é fundamental, pois a arquitetura DiT-MAE foi projetada para aprendizado auto-supervisionado em dados mascarados. Dessa forma, a reconstrução de dados observados pode resultar em imprecisões, já que o modelo não foi treinado para esse contexto específico.

Outra razão que pode implicar na imprecisão do modelo ao reconstruir os dados observados é que, sendo o DiT-MAE uma forma de *autoencoder* de eliminação de ruído (DAE), foi projetado principalmente para dados contínuos e para um maior número de dados espaciais. Os dados discretos, por outro lado, podem sofrer uma perda maior de informação na reconstrução. Este fenômeno foi observado em nossa pesquisa, onde a acurácia na reconstrução de dados faltantes mostrou-se superior quando comparada à reconstrução de dados não mascarados.

A Figura 17 mostra a matriz de confusão para o percentual de imputação correta de dados ausentes. A imputação para conjunto de dados com 3 dimensões possui a precisão em 90% de acerto, inclusive aumentando as taxas de dados faltantes. Para conjuntos de dados com maior dimensão, a precisão de acerto na imputação reduz em torno de 20% quando aumenta a taxa de dados faltantes para 50%. Porém, ainda é uma precisão satisfatória, atingindo um mínimo de 74% de acerto em conjuntos de dados de 21 dimensões para conjuntos de itens com tamanho de 180, 270 e 360 itens, conforme ilustrado na Figura 17.

O gráfico de barras da Figura 18 mostra a relação entre sensibilidade, especificidade e precisão para diferentes conjuntos de dados imputados com porcentagens variáveis de dados ausentes (10%, 25% e 50%). Como ilustrado na imagem, à medida que a porcentagem de dados ausentes aumenta, a sensibilidade, a especificidade e a acurácia diminuem para todos os conjuntos de dados imputados. Como previsto, o conjunto de 28 itens apresenta a maior sensibilidade.

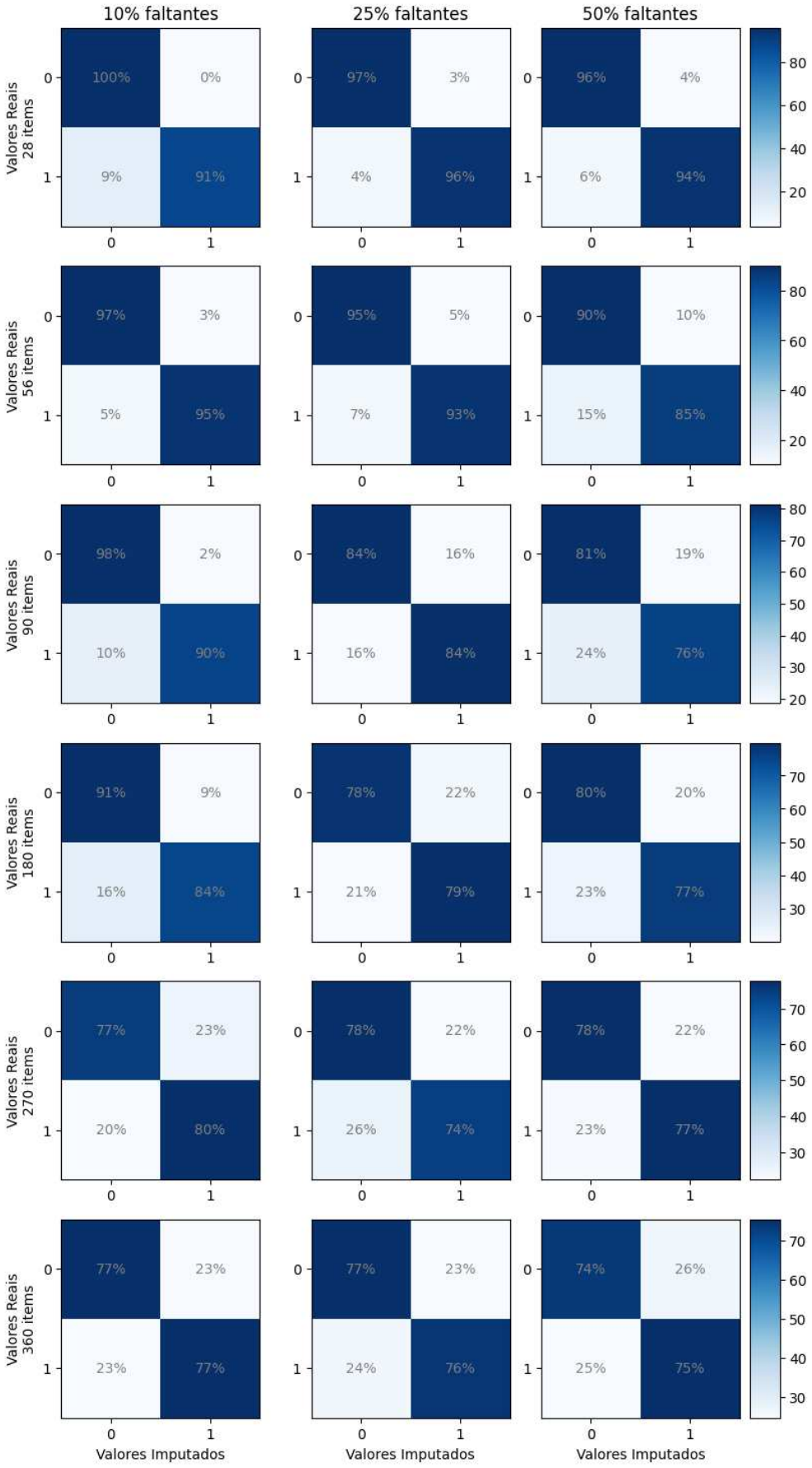

Figura 17 – Matriz de Confusão para dados faltantes com taxas de 10%, 25% e 50% porcentos.

A imputação de dados com o menor número de itens geralmente tem o melhor desempenho em termos de sensibilidade, especificidade e precisão, independentemente da porcentagem de dados ausentes, como apresentado na imagem. À medida que o número de itens aumenta, o desempenho dos métodos de imputação de dados diminui, com o método de 360 itens tendo a menor sensibilidade, especificidade e precisão em todas as porcentagens de dados ausentes. O número maior de dimensões também prejudica as taxas de eficácia da imputação. Entretanto, ainda é um valor satisfatório para grandes quantidades de itens e maior dimensionalidade. Isso vai permitir futuras investigações em que possam ser explorados hiperparâmetros para melhorar a imputação.

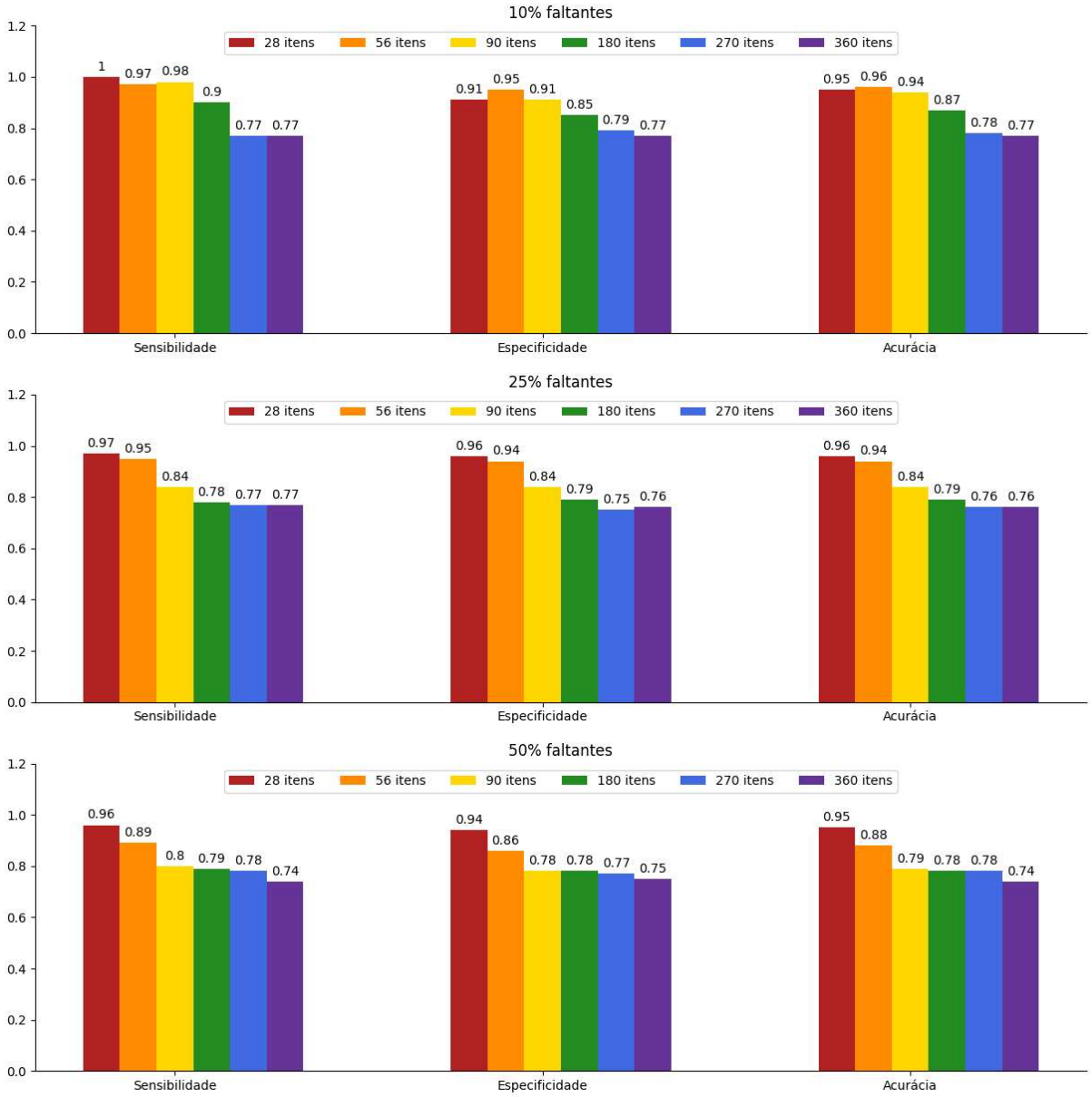

Figura 18 – Sensibilidade, especificidade e acurácia para todos os itens com taxas de 10%, 25% e 50% de dados faltantes para 100 réplicas.

A arquitetura VAEQ é aplicada para estimar parâmetros de itens e indivíduos na segunda etapa do processo. A tabela 2 mostra a correlação entre a média de 100 réplicas dos valores

Tabela 2 – Correlação média entre valores reais e as estimativas do VAEQ e do JML para 100 réplicas.

MD	3 Dimensões				21 Dimensões							
	28 Itens		56 Itens		90 Itens		180 Itens		270 Itens		360 Itens	
	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML
10%	0.9527	0.6885	0.9622	0.6660	0.6917	0.4302	0.8774	0.5923	0.9045	0.5103	0.9339	0.4203
25%	0.9514	0.7521	0.9589	0.6112	0.6481	0.4482	0.8198	0.5836	0.8773	0.5310	0.9218	0.4242
50%	0.8982	0.7416	0.8818	0.5933	0.4775	0.4174	0.7087	0.5016	0.7286	0.5964	0.7660	0.5591

(a) Discriminação

MD	3 Dimensões				21 Dimensões							
	28 Itens		56 Itens		90 Itens		180 Itens		270 Itens		360 Itens	
	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML
10%	0.9981	0.9653	0.9969	0.9660	0.9732	0.7849	0.9722	0.7701	0.9750	0.8765	0.9696	0.9433
25%	0.9985	0.9534	0.9952	0.9519	0.9725	0.7927	0.9695	0.8122	0.9682	0.8360	0.9652	0.8756
50%	0.9912	0.9636	0.9897	0.9507	0.9609	0.8476	0.9562	0.8645	0.9614	0.8505	0.9461	0.8370

(b) Dificuldades

MD	3 Dimensões				21 Dimensões							
	28 Itens		56 Itens		90 Itens		180 Itens		270 Itens		360 Itens	
	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML	VAEQ	JML
10%	0.8265	0.7555	0.8821	0.8513	0.7078	0.6050	0.7942	0.7124	0.8190	0.7844	0.8396	0.8379
25%	0.8236	0.7189	0.8785	0.8093	0.6701	0.5747	0.7589	0.6750	0.7890	0.7525	0.8184	0.7434
50%	0.8114	0.6349	0.8626	0.7390	0.5926	0.5032	0.7116	0.5918	0.7448	0.6631	0.7497	0.7318

(c) Habilidades

reais dos parâmetros e as respectivas estimativas para conjuntos de dados de diferentes números de itens e dimensões com valores ausentes de 10%, 25% e 50%, usando os métodos VAEQ e JML. Com uma observação, o método JML permite realizar as estimativas mesmo com dados faltantes. Para o VAEQ o pré-processamento de imputação por meio da arquitetura DiT-MAE, se faz necessária. A alta correlação entre os valores verdadeiros e estimados é um indicativo de que o algoritmo VAEQ se adequa à análise IRT.

O JML foi o método que apresentou o menor tempo de execução para realizar os experimentos com 100 réplicas de dados de entrada e produzir as estimativas dos parâmetros de itens e pessoas. Entretanto, também realizaram-se experimentos com o método MHRM com apenas um conjunto de dados de cada grupo de dimensão, devido à alta demanda computacional. Os conjuntos de dados escolhidos foram os com menor número de itens de cada grupo, pois o tempo de execução do MHRM é consideravelmente alto à medida que o número de itens e dimensões aumenta.

O gráfico de dispersão da Figura 19 descreve a correlação entre valores verdadeiros de discriminação e valores estimados (a). Conforme mostrado no gráfico, dos três métodos em avaliação, VAEQ, JML e MHRM, os métodos VAEQ e MHRM apresentaram os melhores resultados de ajuste. No entanto, devido à alta dimensionalidade do conjunto de dados, o tempo gasto para produzir as estimativas utilizando o método MHRM torna-se inviável. O método VAEQ demandou mais tempo para imputar os dados ausentes em um pré-processo utilizando o algoritmo DiT-MAE. Para este conjunto de dados, o tempo necessário varia entre 10 e 13 horas. Com os dados imputados, as estimativas dos parâmetros pelo VAEQ são realizadas em 1 hora

e 20 minutos, enquanto o método MHRM exigiu um tempo total de 9 dias para imputação e estimativa. As estimativas da discriminação produzidas pela arquitetura VAEQ apresentaram-se mais dispersas, quando comparadas ao método MHRM. Isso ocorre devido ao aumento de itens e dimensões que influenciaram na imputação dos dados faltantes, pois a porcentagem de acurácia diminui mais de 10% como apresentada na Tabela 1.

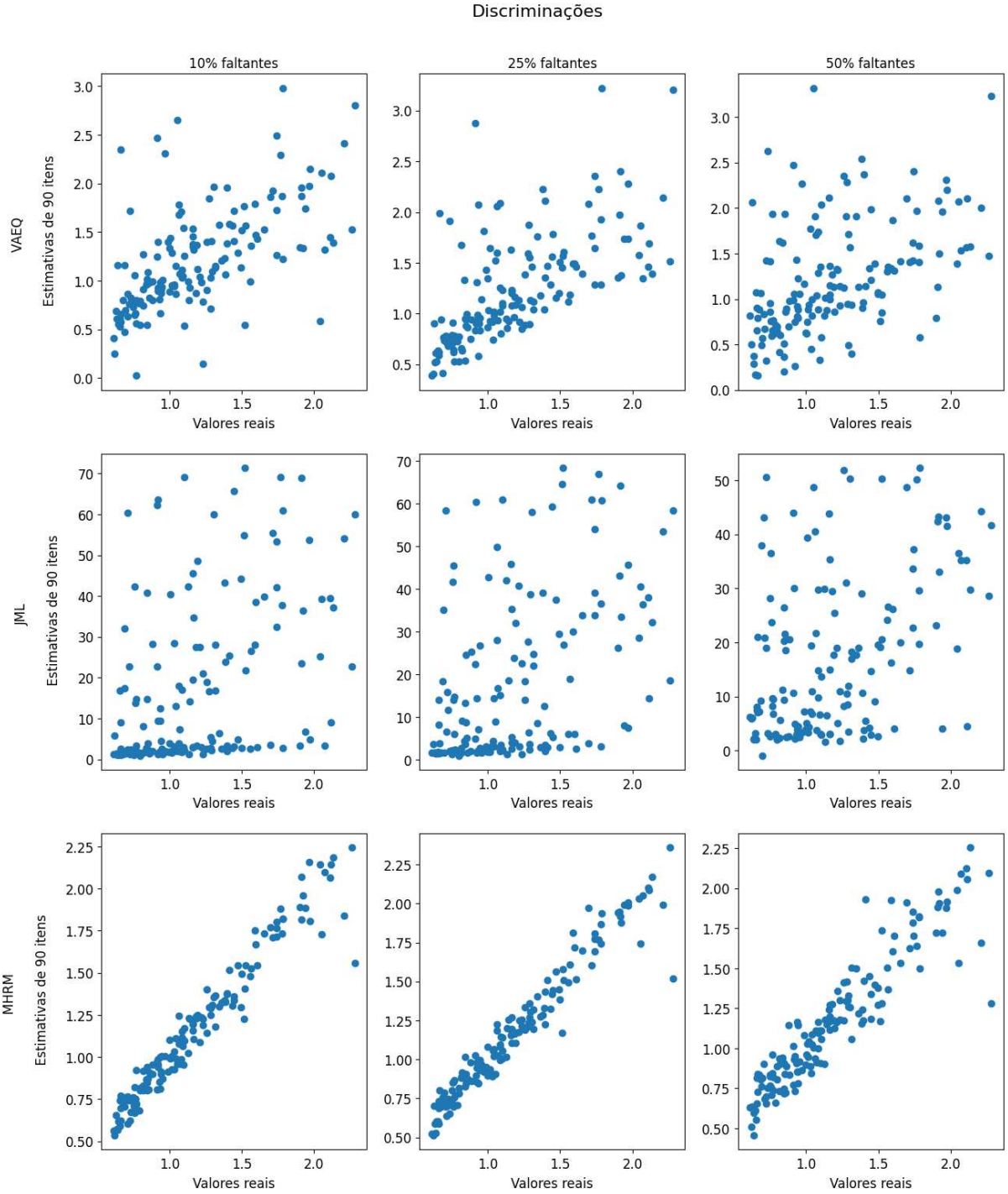

Figura 19 – Gráfico de dispersão das discriminações para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM.

A Figura 20 apresenta os resultados das dificuldades (b) para os três métodos. Geralmente,

as estimativas do parâmetro *b* pelo algoritmo VAEQ exibem um gráfico menos disperso em comparação com o parâmetro *a*, pois o parâmetro de discriminação é fortemente influenciado pela quantidade de habilidades presentes. A dificuldade de um item permanece relativamente constante em, inclusive, diferentes populações, desde que não seja significativamente alterada (RECKASE, 2006; LINDEN; HAMBLETON, 2013).

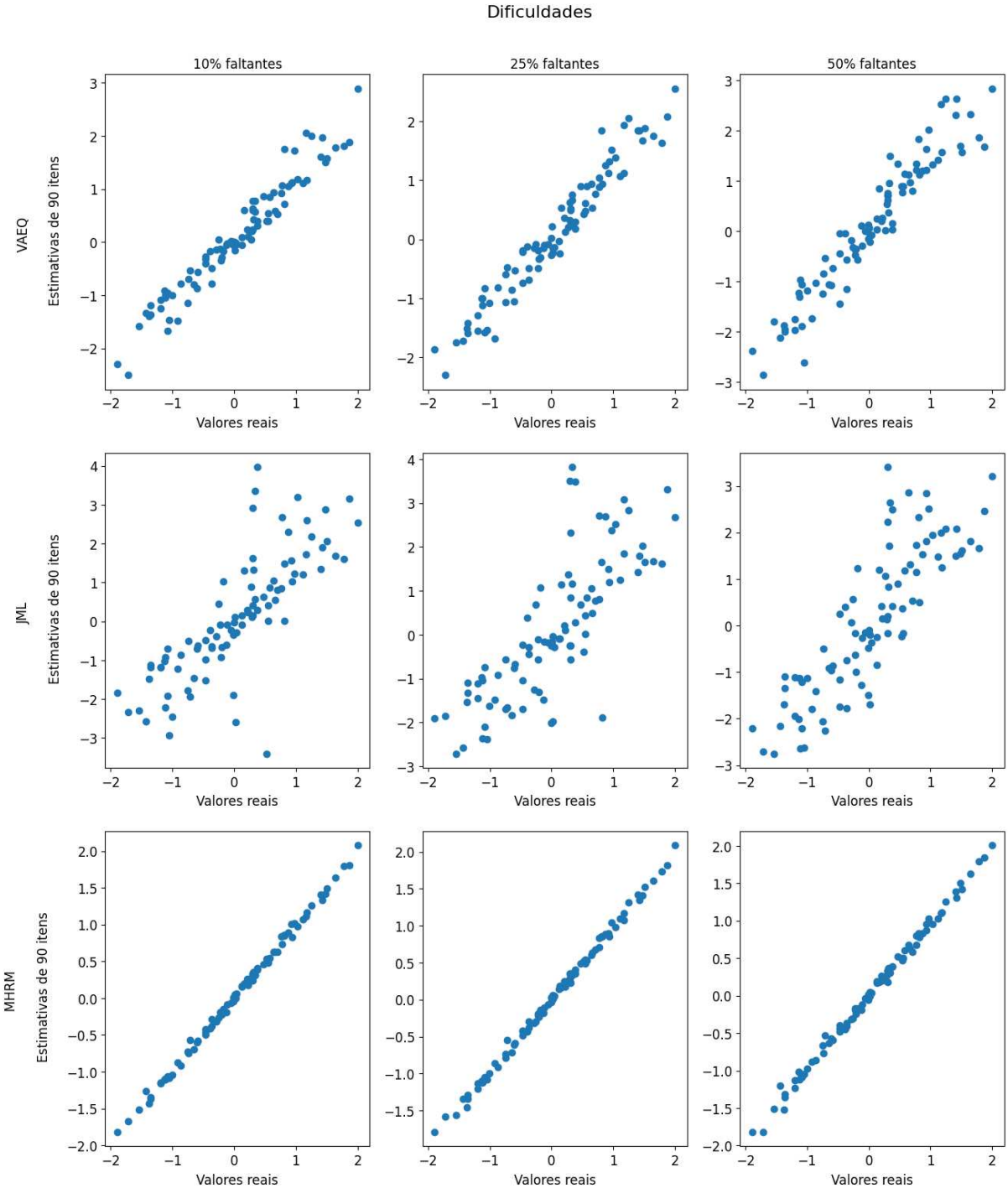

Figura 20 – Gráfico de dispersão das dificuldades para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM.

Isso contrasta com a discriminação, que pode variar de acordo com a diferença do

item entre vários níveis de habilidade. Dessa forma, pequenas mudanças na habilidade levam a grandes mudanças na probabilidade de uma resposta correta. Essa influência pode ser observada na Equação 2.2.

Em relação às estimativas das habilidades, o VAEQ apresentou uma linearidade mais consistente nos resultados para todas as taxas de dados faltantes testadas. Entretanto, pode-se perceber que, à medida que aumenta o número de dados faltantes, as três abordagens apresentam maior dispersão no centro do eixo x. Na Figura 21 ilustramos a comparação entre os valores reais das habilidades e as estimativas do VAEQ, JML e MHRM nos cenários de 10%, 25% e 50% de valores ausentes.

Como apresentado na Figura 21, a densidade dos pontos entre os valores reais e as estimativas do VAEQ está mais bem distribuída. No método JML, os pontos apresentam maior dispersão nas extremidades do gráfico. Para o método MHRM, a dispersão é mais visível no centro do gráfico, apresentando maior achatamento dos pontos, sendo observado pela redução da escala do eixo y, quando comparado com os demais métodos.

Os gráficos *boxplot* na Figura 22 ilustram melhor as diferenças entre os três métodos de avaliação (VAEQ, JML e MHRM) a partir da média das 100 réplicas de Raiz do Erro Quadrático Médio (RMSE) e Vieses para as diferentes porcentagens de respostas ausentes (10%, 25% e 50%).

Em relação ao RMSE, atentando-se para as diferentes escalas para cada método, o VAEQ apresenta uma mediana mais equilibrada e dentro de um intervalo menor de variação à medida que aumenta a quantidade de dados faltantes. Para os métodos JML e MHRM a variação da mediana é mais aparente. Essa alta variação aparente, ilustrada no gráfico de RMSE para o JML e o MHRM, implica que esses métodos se tornam menos precisos nas estimativas das habilidades, quando a taxa de dados ausentes aumenta.

Para todas as abordagens, a quantidade de valores atípicos (*outliers*) é visivelmente maior após o terceiro quartil. O maior número de valores atípicos além do quarto quartil é justificado pela presença maior do número de dados faltantes. Isso pode indicar que os métodos de avaliação são menos confiáveis ou consistentes quando uma porcentagem maior de respostas está faltando. Entretanto, o método MHRM mantém pontuações mais estáveis, com uma ligeira diminuição no nível de 50% de respostas ausentes, indicando sensibilidade moderada às diferentes taxas de dados ausentes.

Em relação ao gráfico de Viés, para o VAEQ, o viés aumenta conforme a porcentagem de dados ausentes aumenta, indo de quase zero em 10% para um viés visivelmente maior em 50%. Isso sugere que o VAEQ pode estar subajustado (*unferfitted*), não sendo capaz de reduzir o erro de treinamento e previsão.

Para os demais métodos, JML e MHRM, o viés é quase semelhante para todas as taxas de dados faltantes, o que significa que tende a prever levemente os valores verdadeiros quando

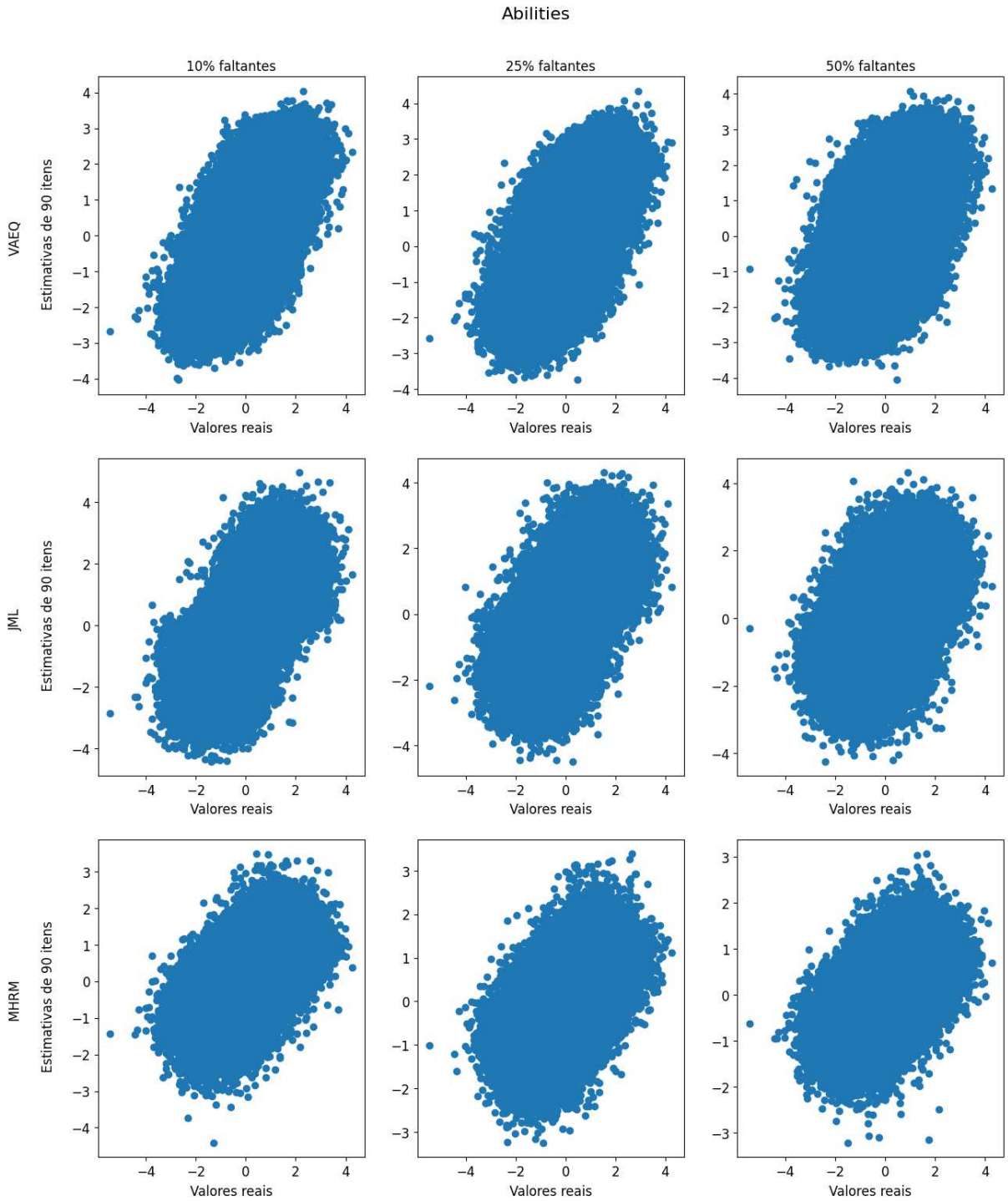

Figura 21 – Gráfico de dispersão das habilidades para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM.

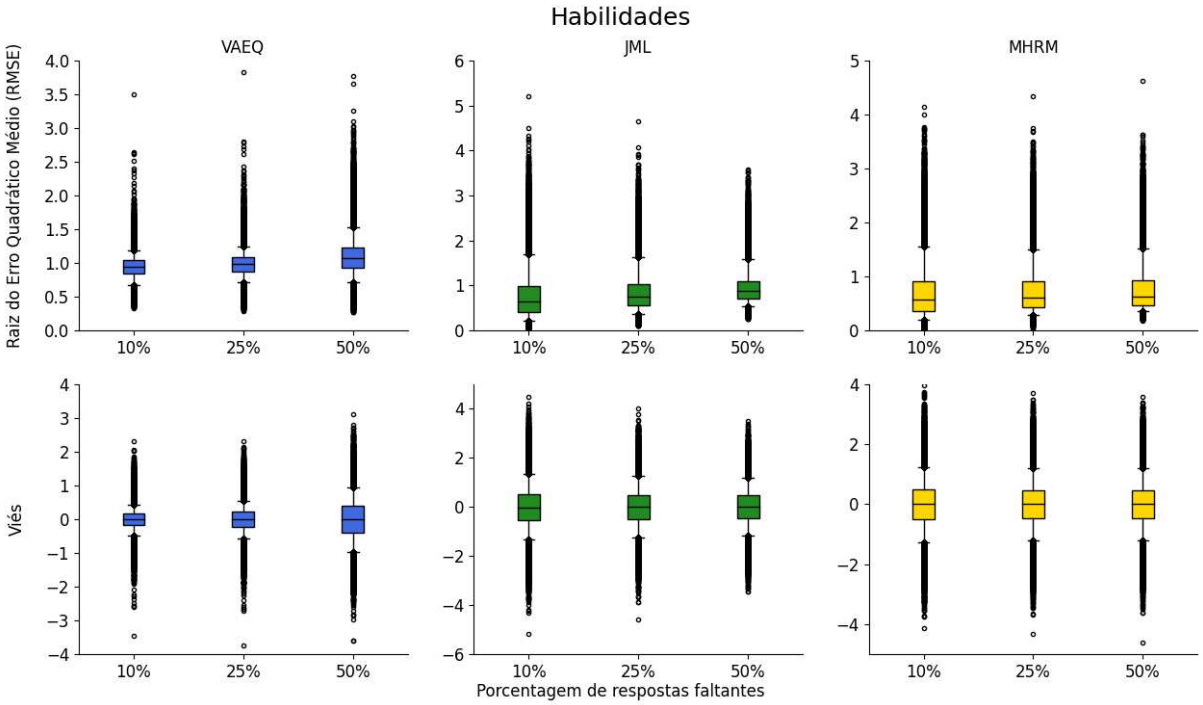

Figura 22 – Gráfico *boxplot* das habilidades para 90 itens e 21 dimensões entre os valores reais e as estimativas do VAEQ, JML e MHRM.

metade dos dados está ausente. Isso sugere que JML e MHRM são mais robustos a problemas de viés em comparação com VAEQ ao lidar com porcentagens maiores de dados ausentes.

4.2 Resultados dos Dados Reais (PISA 2015)

Nesta subseção, são apresentados os resultados obtidos do modelo DiT-MAE para imputação das respostas do PISA 2015. A etapa inicial envolveu a seleção de quatro países com níveis variados de desenvolvimento econômico. É amplamente reconhecido que países com maior desenvolvimento econômico tendem a alocar mais recursos financeiros para políticas educacionais em comparação com países menos desenvolvidos economicamente. Os países selecionados para análise incluem Brasil, Canadá, Colômbia e Estados Unidos. Esses países têm um número substancial de alunos participando do teste OECD-PISA. Esse recurso garante um conjunto de dados experimentais diversificado, permitindo comparações sobre o desempenho do modelo VAEQ relacionado a interpretações de parâmetros em diferentes contextos educacionais.

O conjunto de dados usado para treinar, validar e testar o modelo DiT-MAE compreende 2.154 alunos com quatro dimensões (habilidades). A primeira habilidade é necessária para responder a três itens; a segunda habilidade, para quatro itens; a terceira habilidade, para oito itens; e a quarta habilidade, para três itens. As habilidades são contexto local e relacionadas à segurança e aos perigos das condições ambientais de vida.

A partir deste conjunto de dados, as respostas ausentes são simuladas e também conhe-

cidas como dados mascarados em nosso algoritmo proposto. Isso é feito para o treinamento do modelo a partir dos dados do PISA 2015. O cenário para treinamento, teste e validação foi estabelecido para que cada pessoa, com 18 itens, tenha o mesmo número de respostas ausentes. O processo de imputação durante o treinamento e a validação produziu resultados satisfatórios. O desempenho do modelo foi avaliado em três cenários de respostas ausentes. Para um cenário com 10% de respostas ausentes (2 dados ausentes em 18 itens para cada indivíduo), o processo de validação retornou uma perda de 4% e uma precisão binária de 98%. No caso de 25% de respostas ausentes (4 dados ausentes em 18 itens), a validação produziu uma perda de 10% e uma precisão binária de 95%. Sob a condição de 50% de respostas ausentes (9 dados ausentes em 18 itens), a validação exibiu uma perda de 17% e uma precisão binária de 93%.

Essas métricas demonstram a capacidade do modelo de imputar valores ausentes sob diferentes graus de escassez de dados de forma eficaz. A perda computada e a precisão binária demonstram a robustez e a confiabilidade do modelo DiT-MAE no tratamento de vários níveis de dados ausentes.

Para análises mais claras dos resultados de imputação, uma matriz de confusão compara o número de sucessos do processo de imputação usando o modelo DiT-MAE. A tabela de tabulação cruzada mostra apenas os resultados de valores ausentes simulados em comparação com os valores originais no conjunto de dados de teste e pode ser visualizada na Figura 23.

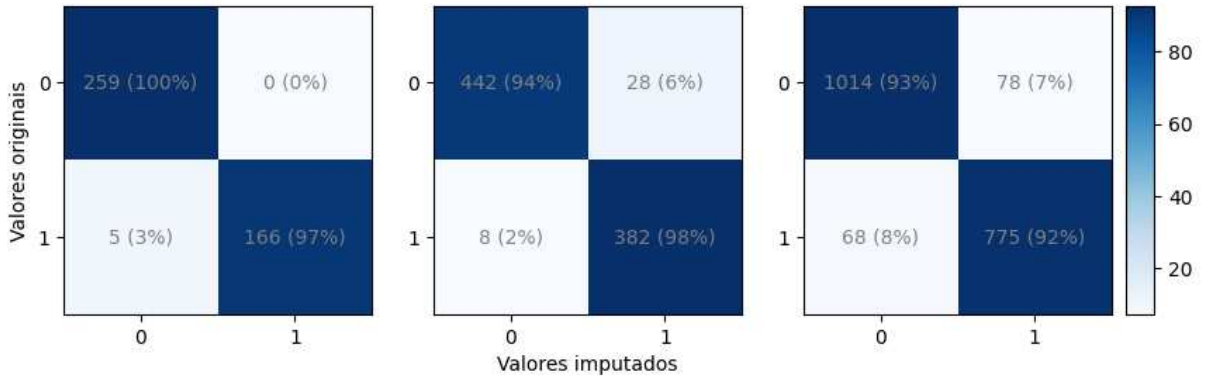

Figura 23 – Porcentagem de acertos da imputação do DiT-MAE para taxas de 10%, 25% e 50% de respostas ausentes.

O segundo passo na avaliação do processo de imputação envolve estimar os parâmetros de item e pessoa usando o modelo VAEQ e o método JML. Para os dados de entrada do VAEQ, é realizada a substituição apenas das respostas mascaradas pelas respostas imputadas pelo DiT-MAE. Para o método JML, utilizamos o conjunto de dados de testes mascarados para avaliar o seu desempenho ao estimar os parâmetros de itens e pessoas.

Conforme elucidado anteriormente, é essencial reconhecer que a saída do DiT-MAE pode não ser totalmente precisa para dados não mascarados. Uma consideração significativa é que o DiT-MAE opera como uma forma de DAE (*Denoising Autoencoder*), projetado principalmente para dados contínuos. No entanto, ao lidar com dados discretos, a aplicabilidade da mesma

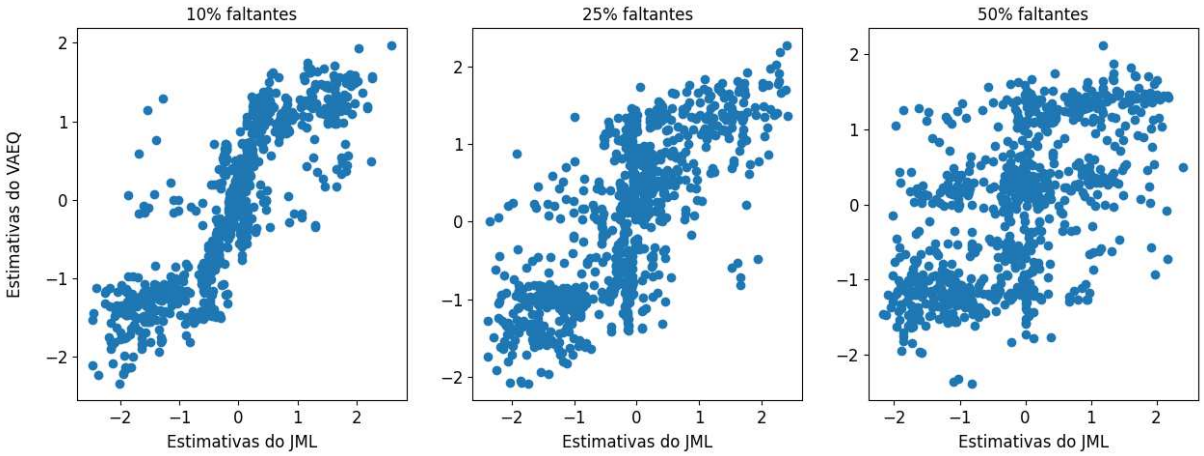

Figura 24 – Gráfico de dispersão entre as estimativas das habilidades para para as três taxas de respostas ausentes do conjunto de dados de teste.

abordagem de redução de ruído pode não produzir resultados ideais. Além disso, a arquitetura e a metodologia de treinamento do DiT-MAE podem não ser explicitamente otimizadas para a reconstrução precisa de dados discretos. As complexidades dos dados discretos podem representar desafios na captura e reconstrução precisas dos padrões subjacentes, levando a discrepâncias potenciais entre os dados não mascarados originais e reconstruídos.

A Figura 24 ilustra o gráfico de dispersão entre as estimativas das habilidades do modelo VAEQ e o método JML para o conjunto de dados de teste. Devido ao conjunto de dados de teste ser pequeno, contendo 860 indivíduos, foi decidido analisar apenas as estimativas das habilidades.

O gráfico de 10% de respostas ausentes tem a maior densidade de pontos em que são mais fortemente agrupados em torno de uma linha de tendência linear, pois há menos influência dos dados faltantes. À medida que a porcentagem de dados faltantes aumenta, para 25% e 50%, os pontos tornam-se mais dispersos.

Essas diferenças refletem o impacto do aumento de dados faltantes na comparação entre as estimativas JML e VAEQ. Com mais valores ausentes, os dados se tornam mais esparsos e espalhados, tornando mais desafiador discernir padrões ou tendências claras no relacionamento entre as duas variáveis. Entretanto, a partir dos gráficos de 25% e 50% de respostas ausentes, percebe-se que as estimativas do método JML começam a se concentrar verticalmente, bem próximo de zero, no eixo x. Enquanto as estimativas do VAEQ estão dispersas horizontalmente em diferentes intervalos. Pode-se concluir que, com o aumento de dados faltantes, o VAEQ ainda consegue estimar valores contínuos de habilidades para diferentes respondentes dos itens.

A avaliação também inclui a imputação de respostas faltantes reais do conjunto de dados do PISA, selecionando pessoas com um intervalo de 1 a 6 respostas faltantes. As respostas faltantes reais estão presentes no conjunto de respostas de estudantes que, propositalmente, não responderam aos itens por não ter o conhecimento necessário ou por tempo insuficiente. O

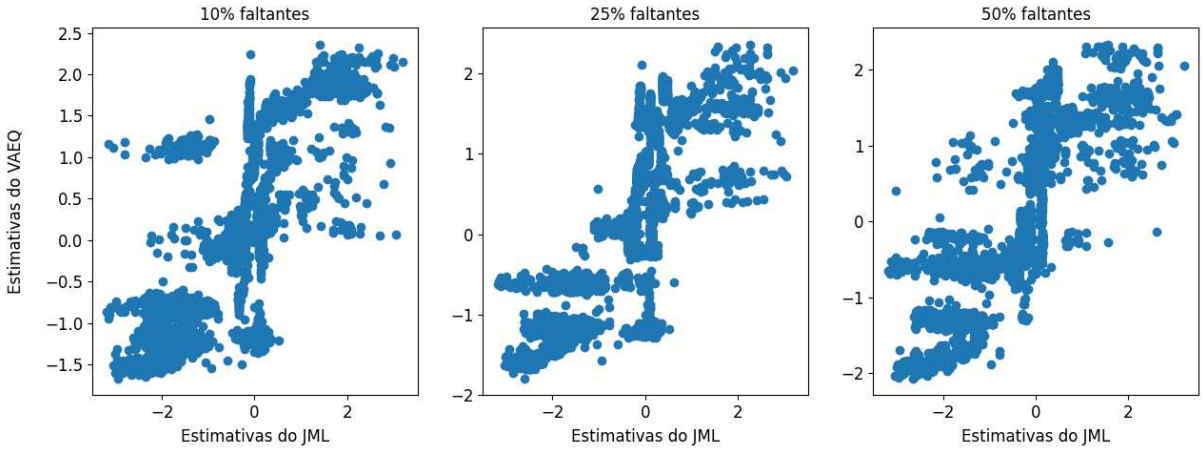

Figura 25 – Gráfico de dispersão entre as estimativas das habilidades para para as três taxas de respostas ausentes do conjunto de dados com respostas ausentes reais.

tamanho do conjunto de dados é de 1.427 pessoas e visa à comparação dos parâmetros estimados entre o modelo VAEQ e o método JML (Figura 25).

A Figura 25 deixa perceptível e reforça a nossa conclusão sobre o comportamento do método JML. Em todos os gráficos, as estimativas produzem uma distribuição pouco linear, exceto para o gráfico com menor taxa de dados faltantes. Nos gráficos de 10% e 25% ainda é visível uma tendência linear do VAEQ e, devido ao método JML ser mais sensível à presença de dados faltantes, produzindo densidades de pontos verticalmente em único e pequeno intervalo do eixo x, os gráficos apresentam aglomerações de pontos (*clusters*) não aparentando uma linearidade crescente. Em relação ao método JML, é importante observar que o pacote R *mirtjml*[1] não produz valores de imputação. Em vez disso, ele se concentra em fornecer parâmetros de item e pessoa.

Esta análise comparativa visa avaliar a consistência e a concordância entre os resultados de imputação gerados pelo modelo VAEQ e o método JML. Como o JML é um método bem conhecido e aplicado para estimativas de parâmetros de itens e pessoas, os gráficos de dispersão oferecem uma base de comparação entre o VAEQ na estimativa de habilidades latentes sob diferentes taxas de dados ausentes.

Uma outra conclusão observada pode ser vista na Figura 26, nos gráficos de dispersão relacionando estimativas de traços latentes VAEQ e JML, mas separados por dimensão. A figura mostra que quanto maior o número de itens relacionados a uma dimensão de habilidade, melhor a correlação entre os dois métodos. De todas as quatro habilidades, a terceira habilidade está associada a um número maior de itens, o que resulta em uma correlação maior entre as estimativas VAEQ e JML. Uma diminuição na correlação também é perceptível com o aumento na porcentagem de valores ausentes.

Uma característica notável observada no modelo VAEQ durante a estimativa de parâme-

[1] https://CRAN.R-project.org/package=mirtjml

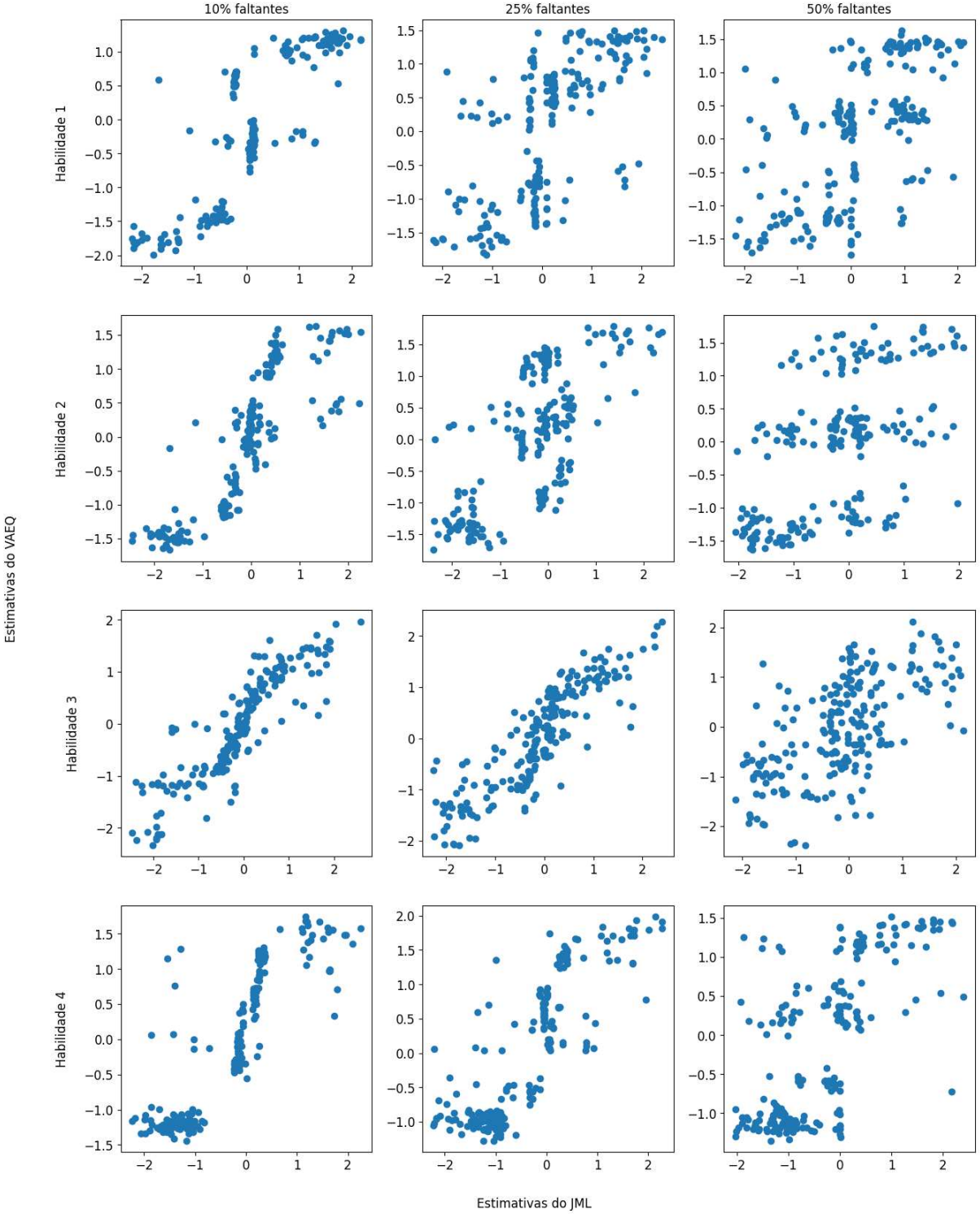

Figura 26 – Gráfico de dispersão para as estimativas das habilidades entre o VAEQ e o JML, variando a porcentagem de respostas ausentes para cada dimensão.

tros é sua capacidade de reter informações relacionadas a distinções de grupo. Como mencionado anteriormente, países com maior desenvolvimento econômico normalmente alocam mais financiamento para políticas educacionais. Ao estimar as habilidades de uma pessoa usando o método JML, esse tipo de informação é frequentemente perdido, e os parâmetros θ são produzidos em uma distribuição normal. No entanto, o modelo VAEQ demonstra a capacidade de manter tais informações durante a estimativa dos parâmetros θ.

A figura 27 ilustra essa característica distintiva pela média representada pela linha pontilhada vertical, selecionando deliberadamente países com desenvolvimento econômico variável para investigar esse comportamento. Mais uma vez, a terceira dimensão, Habilidade 3, fornece uma ilustração mais precisa devido à sua associação com um número maior de itens.

Essa capacidade do modelo VAEQ de preservar e refletir distinções de grupo acrescenta uma camada valiosa de interpretabilidade aos parâmetros estimados, contribuindo para uma compreensão mais diferenciada das habilidades latentes em diferentes contextos econômicos.

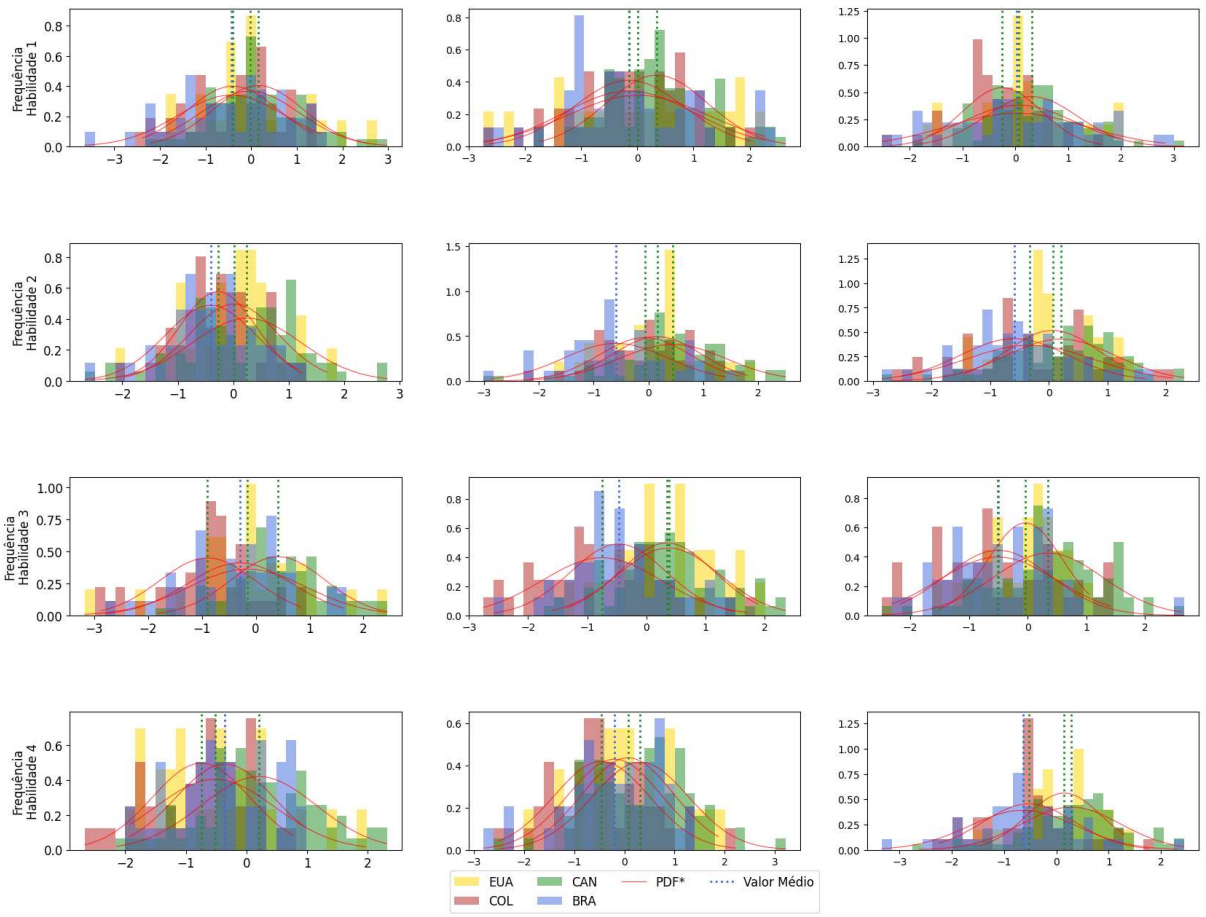

Figura 27 – Histograma e distribuição normal por habilidade das estimativas do VAEQ e taxas de dados faltantes para cada país, diferenciando países economicamente desenvolvidos.

CAPÍTULO

5

CONCLUSÃO

Este capítulo apresenta as conclusões sobre a arquitetura DiT-MAE, seu papel no processo de imputação de dados faltantes binários, sendo parte de um processo de estimação de parâmetros do modelo matemático da Teoria de Resposta ao Item. Aqui, são apresentadas as limitações do DiT-MAE e bem como os trabalhos futuros para melhorar a arquitetura e o processo de estimação de parâmetros da TRI utilizando a abordagem de redes neurais. O capítulo está organizado da seguinte forma: a Seção 5.1 descreve os alcances e a eficácia da arquitetura na imputação de dados ausentes binários. A Seção 5.2 destaca as limitações da arquitetura. A Seção 5.3 apresenta as ideias que podem ser acrescentadas para investigação em trabalhos futuros.

5.1 A Arquitetura DiT-MAE

A arquitetura DiT-MAE foi implementada com o intuito de resolver os problemas de respostas ausentes dicotômicas para itens de questionários educacionais. A implementação teve como referência o trabalho desenvolvido por He *et al.* (2022) em que foi utilizado um modelo de *Transformer* reduzido e mascarado (*Visual Transformer Masked Autoencoder* (ViT-MAE)) para classificação e reconhecimento de imagens. A proposta era que o modelo aprendesse com poucas partes visíveis da imagem, sendo capaz de uma reconstrução da imagem. Isso ocorre, pois a arquitetura consegue aprender com a parte da imagem visível, já que as imagens são ricas em informações espaciais.

Com base nessa abordagem, o ViT-MAE foi adaptado para receber e processar dados binários de respostas dicotômicas de questionários educacionais. Nesses questionários, é comum que o estudante omita a resposta de alguns itens pelo próprio delineamento da aplicação da avaliação, pela falta de conhecimento sobre o assunto (habilidade) ou por falta de tempo para responder corretamente. Assim, o conjunto de respostas de um estudante torna-se o contexto que a arquitetura irá aprender para imputar as respostas ausentes.

Como apresentado no Capítulo de 4, o DiT-MAE consegue imputar dados com altas taxas de acertos, para diferentes cenários de dados faltantes. Também foram testados diferentes tamanhos de conjuntos de dados e dimensões. O DiT-MAE apresentou bom desempenho e escalabilidade, chegando a apresentar 74% de acerto para grandes conjuntos de dados.

No que tange sobre parâmetros da TRIM, o processo de imputação demonstrou melhores resultados para as estimativas de habilidades. O método de imputação proposto com o modelo VAEQ gerou resultados superiores ao método JML e MHRM, que são os métodos tradicionais propostos pela literatura para conjuntos de dados com alta dimensão. A piora no desempenho dos métodos com o aumento das taxas de valores ausentes é esperada, pois há menos informações para estimar e obter resultados. Outro fator importante, identificado neste trabalho de pesquisa, é a redução do tempo para processamento das estimativas e a redução da complexidade computacional, que são pontos críticos nos métodos JML e MHRM quando lidam com grandes quantidades de dados e altas dimensões. Além do mais, as estimativas produzidas pelo JML e MHRM perdem informações estatísticas que poderiam ser utilizadas para a distinção de grupos populacionais, como apresentado na Seção 4.2 no Capítulo 4.

5.2 Limitações do DiT-MAE

Como ressaltado na seção anterior, o DiT-MAE foi implementado com base em um algoritmo para aprender informações de imagens. Devido à natureza dos dados utilizados neste trabalho de pesquisa, ocorre a perda de informação ou não existe informação necessária para melhorar a imputação dos dados, principalmente em conjuntos de dados pequenos, como pode ser observado nos experimentos reais das respostas de questionários do PISA na Seção 4.2, em que um cenário com maior taxa de respostas ausentes influencia no resultado final da imputação.

Foi observado durante os experimentos que a arquitetura DiT-MAE não consegue reconstruir os dados não mascarados, de forma que a maior parte está incorreta quando comparada ao valor original. Um dos motivos que podem levar a esse comportamento é o fato do DiT-MAE ser uma forma de *autoencoder* de eliminação de ruído (DAE), projetado principalmente para dados contínuos. Os dados discretos, por outro lado, podem não ser adequados para reconstrução utilizando a mesma abordagem. Nesse caso, é lógico afirmar que, com entradas discretas e binárias, existe uma redução de informação que poderia agregar no aprendizado da arquitetura.

Outro fator que implica nesse comportamento é o projeto e a finalidade do DiT-MAE, pois foi projetado especificamente para reconstruir as respostas ausentes em um conjunto de entrada que foi parcialmente mascarado. Quando aplicado a dados não mascarados, o modelo pode não ter sido treinado para lidar com tais entradas, levando a reconstruções imprecisas. O DiT-MAE destina-se principalmente ao aprendizado auto-supervisionado em dados mascarados, e usá-lo diretamente em dados não mascarados pode não produzir resultados precisos.

5.3 Trabalhos Futuros para O DiT-MAE

Para trabalhos futuros, podem ser explorados os mecanismos de autoatenção relativa. Uma das principais características do *Transformer* é o mecanismo de aprendizado posicional, desenvolvido principalmente para o processamento de linguagens naturais (NLP), pois a ordem das palavras pode influenciar no contexto de uma frase. Entretanto, para um conjunto de dados de entrada binário, contendo respostas ausentes, a posição destas não é crucial, já que as respostas ausentes podem variar para cada conjunto de respostas de alunos diferentes.

Em relação à arquitetura, avaliar hiperparâmetros que podem agregar no aprendizado do DiT-MAE levando em consideração a natureza dos dados de entrada. Identificar mecanismos apropriados que enriqueçam o aprendizado, tanto o número de camadas latentes necessárias quanto as funções de ativação e otimização. Incluir mais informações que podem enriquecer o contexto informacional para cada estudante, pois o PISA oferece informações sobre o aluno, sobre ambiente residencial, renda familiar, acesso à tecnologia, que podem agregar informação, em que a arquitetura foi projetada para lidar com espaços ricos de informação.

No que diz respeito à reconstrução de dados não mascarados, seria necessário investigar uma forma em que estes valores contribuíssem ou influenciassem diretamente na reconstrução das saídas da arquitetura. Dessa forma, amenizaria a baixa taxa de acerto, quando comparados com os valores originais.

Com a finalidade de melhorar o modelo, existem trabalhos que incorporam a um modelo *Transformer*, apesar da sua natureza *autoencoder*, os mecanismos de uma VAE (*Variational Autoencoder*). Esse tipo de arquitetura permite estimar valores probabilísticos a partir de uma distribuição normal em um espaço latente. Assim, durante o treinamento, o espaço latente produz amostras aleatórias até que a melhor estimativa seja produzida.

Para finalizar, é necessário comparar os parâmetros da TRIM, estimados pelo processo de duas etapas, DiT-MAE e VAEQ, com parâmetros reais de questionários educacionais para avaliar e melhorar as estimativas deste do nosso processo. Com os parâmetros reais, teremos uma base confiável para entendermos e direcionarmos melhor esse trabalho de pesquisa.

REFERÊNCIAS

ALLISON, P. D. Missing data. **The SAGE handbook of quantitative methods in psychology**, Thousand Oaks, CA, p. 72–89, 2009. Citado na página 38.

ANDRADE, D. F. d.; TAVARES, H. R.; VALLE, R. d. C. Teoria da resposta ao item: conceitos e aplicações. **ABE, São Paulo**, 2000. Citado nas páginas 27, 31 e 32.

ARAUJO, E. A. C. d.; ANDRADE, D. F. d.; BORTOLOTTI, S. L. V. Teoria da Resposta ao Item. **Revista da Escola de Enfermagem da USP**, scielo, v. 43, p. 1000–1008, 12 2009. ISSN 0080-6234. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342009000500003&nrm=iso>. Citado nas páginas 31 e 32.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014. Citado na página 50.

BARALDI, A. N.; ENDERS, C. K. An introduction to modern missing data analyses. **Journal of school psychology**, Elsevier, v. 48, n. 1, p. 5–37, 2010. Citado na página 38.

BAYES, T. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. **Philosophical transactions of the Royal Society of London**, The Royal Society London, n. 53, p. 370–418, 1763. Citado na página 48.

BLEI, D. M.; KUCUKELBIR, A.; MCAULIFFE, J. D. Variational inference: A review for statisticians. **Journal of the American statistical Association**, Taylor & Francis, v. 112, n. 518, p. 859–877, 2017. Citado na página 29.

BOCK, R. D.; GIBBONS, R. D. **Item response theory**. [S.l.]: John Wiley & Sons, 2021. Citado na página 39.

BÖHM, V.; LANUSSE, F.; SELJAK, U. Uncertainty quantification with generative models. **arXiv preprint arXiv:1910.10046**, 2019. Citado na página 29.

CAI, L. High-dimensional exploratory item factor analysis by a metropolis–hastings robbins–monro algorithm. **Psychometrika**, Springer, v. 75, p. 33–57, 2010. Citado nas páginas 17, 28, 38, 40, 41 e 42.

CHEN, Y.; LI, X.; ZHANG, S. Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. **Psychometrika**, Springer, v. 84, p. 124–146, 2019. Citado nas páginas 28, 39 e 40.

CHEN, Z.; LIU, S.; JIANG, K.; XU, H.; CHENG, X. A data imputation method based on deep belief network. In: IEEE. **2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing**. [S.l.], 2015. p. 1238–1243. Citado nas páginas 29 e 52.

CHENG, S.; LIU, Q.; CHEN, E.; HUANG, Z.; HUANG, Z.; CHEN, Y.; MA, H.; HU, G. Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In: **Proceedings of the 28th ACM International Conference on Information and Knowledge Management**. [S.l.: s.n.], 2019. p. 2397–2400. Citado na página 52.

CONVERSE, G.; CURI, M.; OLIVEIRA, S. Autoencoders for educational assessment. In: SPRINGER. **International Conference on Artificial Intelligence in Education**. [S.l.], 2019. p. 41–45. Citado nas páginas 28 e 29.

CONVERSE, G.; CURI, M.; OLIVEIRA, S.; TEMPLIN, J. Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. **Machine learning**, Springer, v. 110, n. 6, p. 1463–1480, 2021. Citado na página 29.

CURI, M.; CONVERSE, G. A.; HAJEWSKI, J.; OLIVEIRA, S. Interpretable variational autoencoders for cognitive models. In: IEEE. **2019 International Joint Conference on Neural Networks (IJCNN)**. [S.l.], 2019. p. 1–8. Citado nas páginas 15, 28, 29, 30, 48, 61, 62 e 63.

EMBRETSON, S. E.; REISE, S. P. **Item response theory**. [S.l.]: Psychology Press, 2013. ISBN 978-08-0585-303-2. Citado na página 31.

ENDERS, C. K. **Applied missing data analysis**. [S.l.]: Guilford Publications, 2022. Citado nas páginas 37 e 38.

GAD, I.; HOSAHALLI, D.; MANJUNATHA, B.; GHONEIM, O. A. A robust deep learning model for missing value imputation in big ncdc dataset. **Iran Journal of Computer Science**, v. 4, p. 67–84, 2021. Citado nas páginas 29 e 53.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016. Citado nas páginas 28, 29, 42, 44, 45 e 47.

GOSTHIPATY, A. R.; PAUL, S. **Masked image modeling with Autoencoders**. 2020. <https://keras.io/examples/vision/masked_image_modeling/>. Accessed: (17 oct 2024). Citado nas páginas 58 e 59.

GRAHAM, J. W. Missing data analysis: Making it work in the real world. **Annual review of psychology**, Annual Reviews, v. 60, n. 1, p. 549–576, 2009. Citado nas páginas 29 e 37.

GRAVES, A. Practical variational inference for neural networks. **Advances in neural information processing systems**, v. 24, 2011. Citado na página 29.

HAYKIN, S. S. **Neural networks: A comprehensive foundation**. [S.l.]: IEEE Press Book, 1994. Citado nas páginas 15, 28, 29, 42, 43, 44, 45 e 47.

HE, K.; CHEN, X.; XIE, S.; LI, Y.; DOLLÁR, P.; GIRSHICK, R. Masked autoencoders are scalable vision learners. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 16000–16009. Citado nas páginas 15, 58, 59, 60, 61 e 79.

HIPPEL, P. T. von; BARTLETT, J. W. Maximum likelihood multiple imputation: Faster imputations and consistent standard errors without posterior draws. **Statistical Science**, Institute of Mathematical Statistics, v. 36, n. 3, p. 400–420, 2021. Citado na página 39.

HOFFMAN, M. D.; BLEI, D. M.; WANG, C.; PAISLEY, J. Stochastic variational inference. **Journal of Machine Learning Research**, v. 14, n. 5, 2013. Citado na página 29.

KINGMA, D. P.; WELLING, M. An introduction to variational autoencoders. **arXiv preprint arXiv:1906.02691**, 2019. Citado nas páginas 28, 48, 49 e 50.

KÖHLER, C.; POHL, S.; CARSTENSEN, C. H. Investigating mechanisms for missing responses in competence tests. **Psychological Test and Assessment Modeling**, PABST Science Publishers, v. 57, n. 4, p. 499, 2015. Citado na página 37.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. Citado nas páginas 15, 28, 45, 46 e 47.

LIN, J.; LI, N.; ALAM, M. A.; MA, Y. Data-driven missing data imputation in cluster monitoring system based on deep neural network. **Applied Intelligence**, Springer, v. 50, n. 3, p. 860–877, 2020. Citado nas páginas 29 e 52.

LIN, W.-C.; TSAI, C.-F.; ZHONG, J. R. Deep learning for missing value imputation of continuous data and the effect of data discretization. **Knowledge-Based Systems**, Elsevier, v. 239, p. 108079, 2022. Citado nas páginas 29 e 52.

LINDEN, W. J. V. D.; HAMBLETON, R. K. **Handbook of modern item response theory**. [S.l.]: Springer Science & Business Media, 2013. ISBN 978-0-387-94661-0. Citado nas páginas 27, 31, 32, 33, 34 e 69.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. [S.l.]: John Wiley & Sons, 2019. v. 793. Citado nas páginas 29, 37 e 38.

LITTLE, R. J.; SCHENKER, N. Missing data. In: **Handbook of statistical modeling for the social and behavioral sciences**. [S.l.]: Springer, 1995. p. 39–75. Citado na página 38.

LITTLE, T. D.; JORGENSEN, T. D.; LANG, K. M.; MOORE, E. W. G. On the joys of missing data. **Journal of pediatric psychology**, Oxford University Press, v. 39, n. 2, p. 151–162, 2014. Citado nas páginas 29 e 38.

LIU, T.; WANG, C.; XU, G. Estimating three-and four-parameter mirt models with importance-weighted sampling enhanced variational auto-encoder. **Frontiers in Psychology**, Frontiers Media SA, v. 13, p. 935419, 2022. Citado na página 29.

LIU, Y. A riemannian optimization algorithm for joint maximum likelihood estimation of high-dimensional exploratory item factor analysis. **Psychometrika**, Springer, v. 85, n. 2, p. 439–468, 2020. Citado na página 40.

MAYDEU-OLIVARES, A.; CAI, L.; HERNÁNDEZ, A. Comparing the fit of item response theory and factor analysis models. **Structural Equation Modeling: A Multidisciplinary Journal**, Taylor & Francis, v. 18, n. 3, p. 333–356, 2011. Citado nas páginas 40 e 42.

MEADE, A. W.; LAUTENSCHLAGER, G. J. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. **Organizational Research Methods**, v. 7, n. 4, p. 361–388, 2004. Disponível em: <https://doi.org/10.1177/1094428104268027>. Citado na página 28.

MIIKKULAINEN, R. Topology of a neural network. In: ____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 988–989. ISBN 978-0-387-30164-8. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_837>. Citado na página 47.

MONTECINO, C. E. E. **Using VAE for Incomplete Educational Data**. Tese (PhD thesis) — University of São Paulo, 2023. Disponível em <https://www.teses.usp.br/teses/disponiveis/104/104131/tde-24082023-102049/en.php>. Citado na página 29.

OECD. **PISA 2015 Results (Volume I)**. [s.n.], 2016. 468 p. Disponível em: <https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/technical-report/PISA2015_TechRep_Final.pdf>. Citado na página 57.

PASQUALI, L. **Psicometria: Teoria dos Testes na Psicologia e na Educação**. 5. ed. [S.l.]: Vozes, 2011. ISBN 8532628893; 9788532628893. Citado na página 27.

PEREIRA, R. C.; SANTOS, M. S.; RODRIGUES, P. P.; ABREU, P. H. Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. **Journal of Artificial Intelligence Research**, v. 69, p. 1255–1285, 2020. Citado nas páginas 29 e 52.

POHL, S.; BECKER, B. Performance of missing data approaches under nonignorable missing data conditions. **Methodology**, v. 16, n. 2, p. 147–165, 2020. Citado na página 38.

POHL, S.; GRÄFE, L.; ROSE, N. Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. **Educational and Psychological Measurement**, Sage Publications Sage CA: Los Angeles, CA, v. 74, n. 3, p. 423–452, 2014. Citado na página 38.

RECKASE, M. D. The past and future of multidimensional item response theory. **Applied Psychological Measurement**, SAGE PUBLICATIONS, INC. 2455 Teller Road, Thousand Oaks, CA 91320, v. 21, n. 1, p. 25–36, 1997. Citado na página 27.

____. 18 multidimensional item response theory. **Handbook of Statistics**, Elsevier, v. 26, p. 607–642, 2006. Citado nas páginas 15, 27, 28, 31, 35, 36 e 69.

____. Multidimensional item response theory models. In: ____. **Multidimensional Item Response Theory**. New York, NY: Springer, 2009. p. 79–112. ISBN 978-0-387-89976-3. Disponível em: <https://doi.org/10.1007/978-0-387-89976-3_4>. Citado na página 35.

ROBITZSCH, A. On the treatment of missing item responses in educational large-scale assessment data: An illustrative simulation study and a case study using pisa 2018 mathematics data. **European Journal of Investigation in Health, Psychology and Education**, MDPI, v. 11, n. 4, p. 1653–1687, 2021. Citado nas páginas 38 e 40.

RUBIN, D. B. Inference and missing data. **Biometrika**, Oxford University Press, v. 63, n. 3, p. 581–592, 1976. Citado na página 37.

RUTKOWSKI, L. The impact of missing background data on subpopulation estimation. **Journal of Educational Measurement**, Wiley Online Library, v. 48, n. 3, p. 293–312, 2011. Citado na página 38.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. **Psychological methods**, American Psychological Association, v. 7, n. 2, p. 147, 2002. Citado nas páginas 37 e 38.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural networks**, Elsevier, v. 61, p. 85–117, 2015. Citado na página 42.

ŚMIEJA, M.; STRUSKI, Ł.; TABOR, J.; ZIELIŃSKI, B.; SPUREK, P. Processing of missing data by neural networks. **Advances in neural information processing systems**, v. 31, 2018. Citado na página 53.

URBAN, C. J.; BAUER, D. J. A deep learning algorithm for high-dimensional exploratory item factor analysis. **Psychometrika**, Springer, v. 86, n. 1, p. 1–29, 2021. Citado na página 29.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017. Citado nas páginas 15, 50, 51, 52, 59 e 60.

VELDKAMP, K.; GRASMAN, R.; MOLENAAR, D. Handling missing data in variational autoencoder based item response theory. **British Journal of Mathematical and Statistical Psychology**, Wiley Online Library, v. 78, n. 1, p. 378–397, 2025. Citado na página 29.

WANG, H.; YEUNG, D.-Y. Towards bayesian deep learning: A framework and some existing methods. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 28, n. 12, p. 3395–3408, 2016. Citado na página 28.

WU, M.; DAVIS, R. L.; DOMINGUE, B. W.; PIECH, C.; GOODMAN, N. Variational item response theory: Fast, accurate, and expressive. **arXiv preprint arXiv:2002.00276**, 2020. Citado nas páginas 28 e 29.

ZHANG, C.; BÜTEPAGE, J.; KJELLSTRÖM, H.; MANDT, S. Advances in variational inference. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 41, n. 8, p. 2008–2026, 2018. Citado na página 29.


ICMC
USP
SÃO CARLOS