Número 16 | 2025
https://doi.org/10.48552/ed16-104-124

Análise da Qualidade de Dados Sintéticos para Interfaces Neurais

João Paulo de Oliveira Macedo
joaomac@unicamp.br
Universidade Estadual de Campinas
ORCID 0009-0001-7361-7330

Gilmar Barreto
gilmar@unicamp.br
Universidade Estadual de Campinas
ORCID 0000-0003-2226-6558

Paulo Victor de Oliveira Miguel
pvictor@unicamp.br
Universidade Estadual de Campinas
ORCID 0000-0002-5332-7811

Resumo:
A geração de dados sintéticos é fundamental para o desenvolvimento de Interfaces Cérebro-Computador (ICC), especialmente em interfaces neurais não invasivas. Este artigo utiliza métodos estatísticos para avaliar a qualidade dos dados sintéticos gerados para ICCs, empregando métricas como matriz de correlação, distribuição univariada e bivariada. A matriz de correlação visualiza as relações entre variáveis, enquanto as distribuições analisam a capacidade do modelo em reproduzir as características estatísticas dos dados originais. A plataforma Mostly AI foi utilizada para gerar os dados, e diferentes modelos computacionais foram comparados para identificar o melhor desempenho. Os resultados mostraram que o Modelo Computacional 3 obteve os melhores resultados na geração de dados sintéticos. Além disso, a análise da Distribuição Cumulativa até os Registros mais Próximos (DCR) foi aplicada para avaliar a proximidade entre os dados sintéticos e reais, garantindo a confidencialidade dos dados originais.

Palavras-chave: Análise estatística, dados sintéticos, interfaces cérebro-computador, interfaces neurais.

Abstract:
The generation of synthetic data is fundamental for the development of Brain-Computer Interfaces (BCIs), especially in non-invasive neural interfaces. This article employs statistical methods to evaluate the quality of synthetic data generated for BCIs, utilizing metrics such as correlation matrix, univariate distribution, and bivariate distribution. The correlation matrix visualizes the relationships between variables, while the distributions assess the model's capability to reproduce the statistical characteristics of the original data. The Mostly

AI platform was used to generate the data, and different computational models were compared to identify the best performance. The results showed that Computational Model 3 achieved the best outcomes in generating synthetic data. Furthermore, the analysis of the Cumulative Distribution up to the Nearest Records (DCR) was applied to evaluate the proximity between synthetic and real data, ensuring the confidentiality of the original data.

Keywords: Brain-computer interfaces, neural interfaces, statistical analysis, synthetic data.

1. Geração de Dados Sintéticos

O desenvolvimento de sistemas baseados em Interfaces Cérebro-Computador (ICC) tem conquistado um espaço crescente nas últimas décadas, apresentando uma ampla gama de aplicações. Essas aplicações variam desde sistemas destinados à reabilitação de indivíduos com necessidades físicas até a criação de novos modelos e paradigmas de comunicação entre seres humanos e máquinas.

Entretanto, o avanço prático desses sistemas enfrenta desafios significativos, especialmente em relação à crescente demanda por dados de alta qualidade. A coleta de dados adequados é fundamental para o treinamento e a eficácia dos modelos de ICC, e a escassez de dados pode limitar seu desempenho.

Para enfrentar esses desafios, são empregados modelos computacionais que possibilitam a geração de dados sintéticos. Essas abordagens não apenas ajudam a mitigar a falta de dados, mas também podem melhorar a robustez e a generalização dos sistemas desenvolvidos. Assim, a utilização de técnicas avançadas de processamento de dados e aprendizado de máquina torna-se essencial para a evolução das interfaces cérebro-computador, promovendo um avanço significativo nas interações entre humanos e máquinas. Esses esforços são cruciais para a consolidação das ICC como ferramentas viáveis e eficazes em diversas áreas, ampliando as possibilidades de reabilitação, comunicação e interação em contextos variados (Habashi, Azab, Eldawlatly, & Aly, 2023).

2. Interfaces Neurais

As interfaces neurais têm conquistado destaque significativo nos últimos anos. Pesquisas indicam uma tendência crescente no desenvolvimento de sistemas que facilitam a interação entre o cérebro humano e dispositivos computacionais externos. Essas interfaces demonstram um potencial promissor, especialmente nas áreas terapêuticas, como a reabilitação médica, mas

também se expandem para setores de entretenimento e aplicações militares. Os diversos tipos de sistemas de Interface Cérebro-Computador (ICC) são projetados para registrar as ondas cerebrais de um indivíduo e transmiti-las a um sistema computacional, permitindo que ações sejam executadas com base nessas informações (Cengiz et al., 2024).

Atualmente, as principais interfaces neurais não invasivas utilizam a captação de sinais neurais por meio de sistemas de Eletroencefalografia (EEG), que registram as ondas cerebrais e as enviam para processamento computacional. Um sistema universal de captação de sinais de EEG foi desenvolvido para padronizar a forma como os estudos são conduzidos, demonstrando eficácia em diversas aplicações, conforme observado na Figura 1. Assim, os estudos são estruturados com base nas técnicas e procedimentos estabelecidos pela eletroencefalografia clínica, utilizando o Sistema Internacional 10/20. Essa padronização não apenas facilita a comparação entre pesquisas, mas também melhora a confiabilidade dos resultados obtidos.

Os sinais neurais dos seres humanos são traduzidos em sinais elétricos que variam em diferentes faixas de onda, refletindo a atividade cerebral em diversas condições, a Figura 2 ilustra essa ideia. Essas ondas são categorizadas em frequências distintas, cada uma associada a estados específicos de consciência e atividades cognitivas. As principais faixas de frequência incluem:

Delta (0,5 a 4 Hz): associadas ao sono profundo e à regeneração do corpo.

Theta (4 a 8 Hz): relacionadas a estados de relaxamento, meditação e criatividade.

Alpha (8 a 12 Hz): frequentemente observadas em estados de relaxamento e atenção calma.

Beta (12 a 30 Hz): ligadas a estados de alerta, concentração e atividade mental intensa.

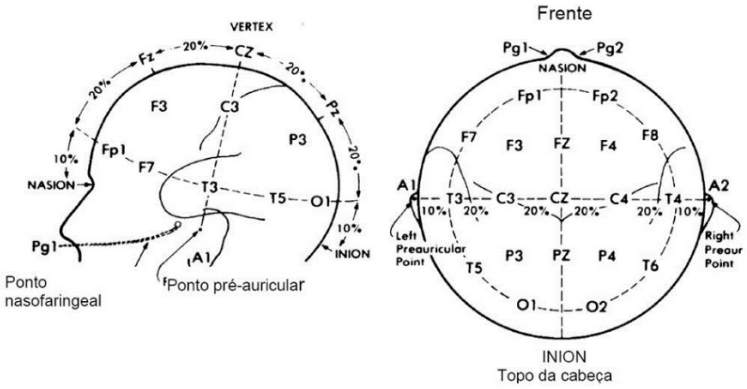
Gamma (acima de 30 Hz): associadas a processos cognitivos complexos, como a percepção e a consciência.

A amplitude do sinal também é crucial na interpretação dos dados de EEG, pois reflete a força da atividade elétrica registrada. Sinais de maior amplitude podem indicar uma maior sincronização de neurônios em uma determinada área do cérebro, enquanto sinais de menor amplitude podem sugerir uma atividade mais dispersa.

Além disso, a localização dos sensores no couro cabeludo, conforme definido pelo Sistema Internacional 10/20, permite a aquisição de dados de diferentes áreas do cérebro. Essa disposição é fundamental para a análise da atividade cerebral em regiões específicas, como o lobo frontal, que está associado ao controle executivo, e o lobo occipital, que é responsável pela visão (Miguel, 2010).

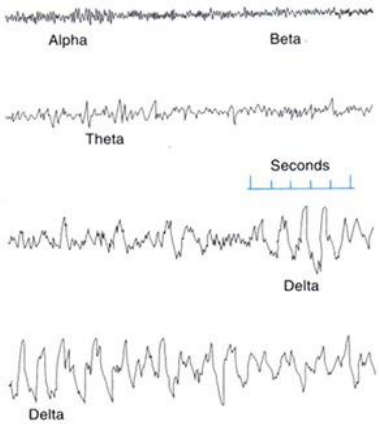
As condições do indivíduo durante a captação dos sinais também desempenham um papel importante. Fatores como estado emocional, nível de estresse, medicamentos, e até mesmo a qualidade do sono podem influenciar a atividade elétrica cerebral. Portanto, a interpretação dos dados de EEG deve considerar não apenas as características dos sinais, mas também o contexto em que foram coletados. Essa abordagem abrangente é essencial para um entendimento mais preciso da atividade neural e para o desenvolvimento de aplicações eficazes no campo das interfaces neurais. (Macedo, Barreto & Miguel, 2024).

Figura 1 – Sistema Internacional 10/20.



Fonte: Miguel (2010).

Figura 2 – Sinais cerebrais.



RÍTMOS DAS ONDAS CEREBRAIS

RITMO	DELTA	THETA	ALPHA	BETA
Componente em freqüência	< 4 Hz	4 a 7 Hz	8 a 13 Hz	> 13 Hz
Amplitude	100μV	**Criança:20μV** Adulto:10μV	Bebê: 20μV **Criança:75μV** Adulto:50μV	10 a 20 μV
Principal área do escalpo	Frontal	Temporal	Occipital Parietal	Frontal
Condição do Indivíduo	Sono profundo (adulto)	Sonolência	Repouso Olhos fechados	Repouso Olhos abertos

Fonte: Miguel (2010).

3. Métodos Estatísticos para Avaliação da Qualidade dos Dados

A utilização de dados sintéticos tem se tornado uma prática fundamental no desenvolvimento de interfaces neurais, permitindo a simulação de cenários complexos sem a necessidade de dados reais que, muitas vezes, são escassos ou difíceis de obter. No entanto, a eficácia dessa abordagem depende fortemente da qualidade e da precisão dos dados gerados. Para garantir que os dados sintéticos sejam representativos e úteis, é imprescindível a adoção de uma metodologia sistemática que integre uma base de dados sólida com métodos de avaliação rigorosos. Nesse contexto, a aplicação de técnicas estatísticas se revela essencial, pois não apenas valida os dados gerados, mas também possibilita ajustes e refinamentos que aumentam sua relevância e aplicabilidade em pesquisas e aplicações práticas (Macedo et al., 2024).

3.1. Matriz de Correlação

A utilização de uma matriz de correlação é uma abordagem eficaz para avaliar dados sintéticos gerados para aplicações em interfaces neurais. Este instrumento estatístico permite associar diversas variáveis dentro de um conjunto de análise, proporcionando uma visualização clara das relações entre elas.

Cada célula da matriz representa a correlação entre duas variáveis, permitindo identificar padrões e dependências que possam existir nos dados. A interpretação dessas correlações é fundamental para entender a estrutura dos dados sintéticos e sua validade em simulações e aplicações reais.

Além disso, as matrizes de correlação podem auxiliar no ajuste e na otimização dos modelos computacionais utilizados em ICC. Através da análise dessas correlações, pesquisadores e desenvolvedores podem identificar variáveis que exercem forte influência entre si, facilitando a seleção de características relevantes e a melhoria da eficácia dos sistemas desenvolvidos.

A correlação é uma medida numérica que varia em um intervalo de -1 a +1, amplamente utilizada em estatística para quantificar a relação entre duas variáveis. Comumente representada pela letra r, essa medida é conhecida como coeficiente de correlação linear, um conceito desenvolvido por Karl Pearson. Uma correlação positiva, onde $r > 0$, indica que ambas as variáveis tendem a aumentar ou diminuir juntas, ou seja, movem-se na mesma direção. Em contraste, uma correlação negativa, representada por $r < 0$, sugere que as variáveis se comportam de maneira oposta: quando uma variável aumenta, a outra tende a diminuir. Quando

$r = 0$, isso indica que não há uma relação linear discernível entre as variáveis analisadas. Os limites do intervalo de correlação são críticos, pois representam a intensidade da relação entre as variáveis. Um valor de r próximo de +1 ou -1 sugere uma forte correlação, enquanto valores próximos a 0 indicam uma relação fraca ou inexistente (Instituto Nacional de Ensino, 2023).

A correlação entre variáveis é frequentemente expressa pela coeficiente de correlação de Pearson, que é uma medida da relação linear entre duas variáveis. A fórmula é dada por:

$$r = \frac{\mathrm{n}(\Sigma \mathrm{xy}) - (\Sigma \mathrm{x})(\Sigma \mathrm{y})}{\{[\mathrm{n}\Sigma \mathrm{x}^2 - (\Sigma \mathrm{x})^2][\mathrm{n}\Sigma \mathrm{y}^2 - (\Sigma \mathrm{y})^2]\}^{\frac{1}{2}}}$$

Onde:

- r é o coeficiente de correlação de Pearson.
- n é o número de pares de dados.
- x e y são as variáveis cujos dados estão sendo analisados.
- ∑xy é a soma dos produtos das variáveis x e y.
- ∑x e ∑y são as somas das variáveis x e y, respectivamente.
- $\sum x^2$ e $\sum y^2$ são as somas dos quadrados das variáveis x e y, respectivamente.

3.2. Distribuição Univariada

A distribuição univariada refere-se à distribuição de uma única variável, sendo esse conceito de grande relevância para a estatística e a análise de dados. Essa análise permite compreender como um determinado valor de uma variável se distribui em um conjunto de dados. Os principais aspectos desse tipo de análise estão fundamentados na modelagem estatística, especialmente quando trabalhamos com dados sintéticos. Neste contexto, é crucial observar aspectos como a tendência central, a variabilidade e a forma da distribuição, pois esses elementos podem fornecer insights valiosos.

As distribuições univariadas podem ser categorizadas em distribuições discretas ou contínuas, dependendo da natureza dos dados analisados. Essa categorização é fundamental para a escolha dos métodos estatísticos apropriados e para a interpretação correta dos resultados (Thompson, 1999).

As principais características da distribuição univariada incluem a média, a mediana e a moda, que são indicadores fundamentais de tendência central. A média é definida como o valor médio dos dados, enquanto a mediana representa o ponto central da distribuição quando os dados estão

ordenados. A moda, por sua vez, é o valor que ocorre com maior frequência no conjunto de dados analisado.

Além das medidas de tendência central mencionadas, a variância e o desvio padrão são utilizados para descrever a dispersão dos dados em relação à média. A variância quantifica a média dos quadrados das diferenças entre cada valor e a média da distribuição, fornecendo uma noção da extensão da variabilidade. O desvio padrão, que é a raiz quadrada da variância, oferece uma medida mais intuitiva da dispersão, expressando a variação em unidades equivalentes aos dados originais.

Essas medidas estatísticas são cruciais para a compreensão da distribuição univariada, pois permitem analisar não apenas a centralidade dos dados, mas também a sua dispersão, contribuindo para uma interpretação mais robusta e informada dos resultados (Bastos & Duquia, 2007).

3.3. Distribuição Bivariada

As análises bivariadas são essenciais para compreender as relações entre pares de variáveis, frequentemente representadas como x e y. Essas análises permitem a construção, por exemplo, de elipses de confiança, que são regiões estatísticas que facilitam a predição de novas observações em uma população, com um nível de probabilidade específico. A representação gráfica dessas elipses, comumente utilizada em gráficos de dispersão, oferece uma visualização intuitiva das interações entre as variáveis em estudo.

A correlação entre as variáveis x e y é interpretada com base na forma e na inclinação da elipse de confiança. Quando a elipse apresenta um formato circular, isso indica a ausência de correlação entre as variáveis. Por outro lado, à medida que a elipse se alonga, a correlação entre x e y tende a aumentar. O grau de alongamento da elipse pode classificar a correlação como moderada ou alta; uma elipse que se afunila em torno de um ângulo de 45º entre os eixos x e y sugere uma forte correlação positiva.

Além das elipses de confiança, diversas abordagens podem ser empregadas na análise bivariada, que envolve a comparação de duas variáveis. Essas técnicas permitem estabelecer relações entre elas, determinando a intensidade e a direção dessa relação. Além disso, é possível identificar tendências e padrões nos dados, testar hipóteses de causalidade e associação, e realizar previsões sobre os valores das variáveis (Malhado, Ramos, Carneiro, Azevedo, & Martins Filho, 2008).

Essas análises são especialmente relevantes no contexto das interfaces neurais, onde a compreensão das relações entre diferentes parâmetros, como a atividade elétrica cerebral e a resposta motora, é fundamental. A análise bivariada também pode elucidar as interações entre sinais elétricos provocados por um mesmo estímulo. Tal compreensão oferece insights valiosos para o desenvolvimento de tecnologias que aprimoram a interação entre humanos e máquinas, além de proporcionar uma visão mais profunda sobre o funcionamento do cérebro. Essa abordagem pode expandir nosso conhecimento sobre as diferentes manifestações neurais em resposta a estímulos semelhantes, contribuindo para a inovação na área de interfaces neurais e na geração de dados sintéticos.

4. Dados Neurais e a Geração de Dados Sintéticos

Para o modelamento da geração de dados sintéticos destinados às Interfaces Neurais, utilizou-se a plataforma computacional Mostly AI. Esta plataforma é capaz de gerar dados que preservam a similaridade em relação aos dados originais, assegurando a manutenção de aspectos estatísticos fundamentais, como a matriz de correlação, a distribuição univariada e a distribuição bivariada. Essas características justificam a aderência dos dados sintéticos à estrutura dos dados originais.

A plataforma Mostly AI gera dados sintéticos de alta fidelidade, preservando a privacidade dos dados originais. Utilizando técnicas avançadas de aprendizado de máquina, a plataforma cria conjuntos de dados que mantêm características estatísticas e de distribuição muito próximas aos dados reais. Essa capacidade possibilita um treinamento e validação confiáveis de modelos de inteligência artificial, sendo especialmente relevante para o desenvolvimento de tecnologias como Interfaces Cérebro-Computador, onde a disponibilidade de dados robustos e representativos é essencial.

Neste trabalho, optou-se por gerar dados sintéticos na mesma quantidade que os dados originais. Como base de dados real, utilizamos as informações fornecidas pela Universidade de Tübingen, Instituto de Psicologia Médica e Neurobiologia Comportamental, referentes ao desafio da II Competição de Interfaces Cérebro-Computador, realizada em 2002 (Blankertz, 2002).

Os dados utilizados neste estudo foram coletados de um sujeito saudável, que foi instruído a mover um cursor para cima e para baixo na tela do computador, enquanto seus potenciais corticais eram monitorados. Durante a gravação, o participante recebeu feedback visual sobre seus potenciais corticais lentos (Cz-Mastoides), onde a positividade cortical resultava em um

movimento descendente do cursor, enquanto a negatividade cortical causava um movimento ascendente. Cada tentativa tinha a duração de 6 segundos, com a tarefa visualmente apresentada por meio de um alvo destacado na parte superior ou inferior da tela, indicando a negatividade ou positividade do sinal, desde 0,5 segundos até o final da tentativa. O feedback visual foi disponibilizado entre 2 e 5,5 segundos, com um intervalo total de 3,5 segundos utilizado para o treinamento e teste. Com uma taxa de amostragem de 256 Hz e uma duração de gravação de 3,5 segundos, cada tentativa resultou em 896 amostras por canal. Para a competição de Interfaces Cérebro-Computador, os dados de eletroencefalografia foram coletados utilizando o amplificador PsyLab EEG8, que possui uma faixa de medição de ±1000 microvolts (μV). O sinal foi digitalizado por placas de computador PCIM-DAS1602, com resolução de 16 bits, e amostrado a 256 amostras por segundo. Os dados foram registrados em seis posições diferentes no escalpo do participante, seguindo o sistema internacional 10/20 de posicionamento de eletrodos. O conjunto de dados coletado, que serviu de base para a competição, permitiu a avaliação e desenvolvimento de algoritmos de processamento de sinais cerebrais. Especificamente, foram registrados 268 ensaios em dois dias diferentes, sendo 168 no primeiro dia e 100 no segundo. As matrizes "Traindata_0.txt" e "Traindata_1.txt" contêm dados de 135 ensaios da classe 0 e 133 ensaios da classe 1, com dimensões de 135x5377 e 133x5377, respectivamente. A classe representa o sinal cortical positivo (0) e negativo (1). Para este estudo, foi selecionado um subconjunto da matriz "Traindata_0.txt", com dimensões de 134x98, visando validar a eficiência do modelo de geração de dados sintéticos desenvolvido na plataforma Mostly AI, sem comprometer o desempenho computacional.

4.1. Base de Dados

A utilização da plataforma Mostly AI demonstra um significativo potencial para a ampliação de bases de dados, oferecendo um ambiente computacional acessível tanto para pesquisadores experientes quanto para aqueles com menos formação técnica em modelagem e avaliação de dados. Para a geração de dados sintéticos, foi inserido um banco de dados real extraído do arquivo "Traindata_0.txt", denominado "Base de Dados". O objetivo dessa abordagem é desenvolver um modelo computacional otimizado (rede neural) para a geração eficaz de dados sintéticos.

A base de dados consiste nos sinais neurais extraídos de um indivíduo submetido a um ensaio laboratorial. Cada valor apresentado na Figura 3 representa um estímulo neural, que não está relacionado a uma tarefa específica, mas sim a uma parte do conjunto de excitações cerebrais

que resultam no movimento do cursor do mouse na tela. É importante destacar que os intervalos de amostragem são determinados pelo tipo de equipamento utilizado; quanto maior a resolução do dispositivo, maior será a quantidade de dados amostrados, resultando em uma base de dados mais rica e detalhada.

Nesse contexto, a ampliação de dados sintéticos não é um processo trivial; requer um olhar crítico sobre o processo e, principalmente, sobre os dados resultantes por parte do pesquisador. A plataforma Mostly AI utiliza uma combinação robusta de algoritmos para garantir a geração e o controle dos dados sintéticos, o que pode facilitar o gerenciamento das informações e, ao mesmo tempo, gerar insights valiosos para a verificação da qualidade dos dados gerados artificialmente.

Figura 3 – Parte da Base de Dados dos sinais neurais.

Base de Dados

Valor 1	Valor 2	Valor 3	Valor 4	Valor 5
0.000.000.500.000.000	0.000.000.300.000.000	_RARE_	0.000.000.500.000.000	0.000.000.400.000.000
0.000.000.200.000.000	0.000.000.300.000.000	0.000.000.200.000.000	0.000.000.400.000.000	0.000.000.500.000.000
RARE	0.000.000.400.000.000	0.000.000.300.000.000	0.000.000.700.000.000	0.000.000.600.000.000
0.000.000.700.000.000	0.000.000.200.000.000	0.000.000.400.000.000	0.000.000.600.000.000	0.000.000.200.000.000
0.000.000.200.000.000	0.000.000.300.000.000	0.000.000.100.000.000	0.000.000.600.000.000	0.000.000.200.000.000
0.000.000.600.000.000	0.000.000.400.000.000	0.000.000.400.000.000	0.000.000.100.000.000	0.000.000.200.000.000
0.000.000.500.000.000	0.000.000.500.000.000	0.000.000.200.000.000	0.000.000.700.000.000	0.000.000.700.000.000
0.000.000.700.000.000	0.000.000.400.000.000	0.000.000.600.000.000	0.000.000.500.000.000	0.000.000.500.000.000

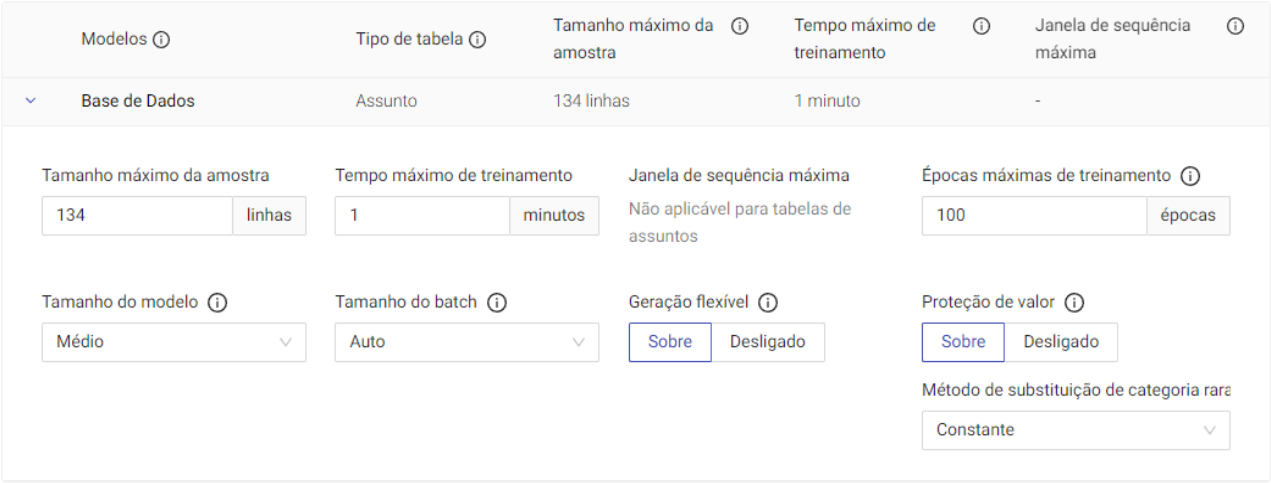
Fonte: Elaboração própria.

4.2. Modelo Computacional 1

O primeiro modelo computacional implementado (Figura 4) visou otimizar tanto o tempo de processamento quanto o esforço computacional necessário para a execução do modelo. Após o treinamento do Modelo Computacional 1, os resultados de desempenho obtidos foram os seguintes:

Distribuição Geral: 48,3%
Distribuição Univariada: 64,4%
Distribuição Bivariada: 32,2%

Figura 4 - Configuração do Modelo Computacional 1

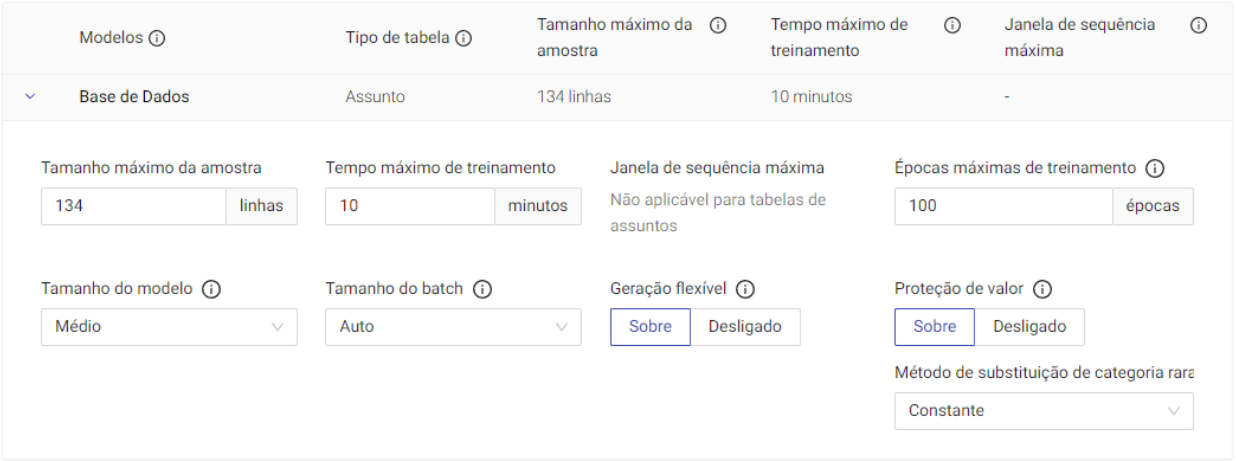


Fonte: Elaboração própria.

4.3. Modelo Computacional 2

O segundo modelo computacional implementado caracterizou-se pela ampliação do tempo de processamento de 1 para 10 minutos. Os demais parâmetros foram mantidos constantes, conforme ilustrado na Figura 5.

Figura 5 - Configuração do Modelo Computacional 2.



Fonte: Elaboração própria.

Como resultado do treinamento, o Modelo Computacional 2 apresentou os seguintes desempenhos:

Distribuição Geral: 49,1%

Distribuição Univariada: 65,4%

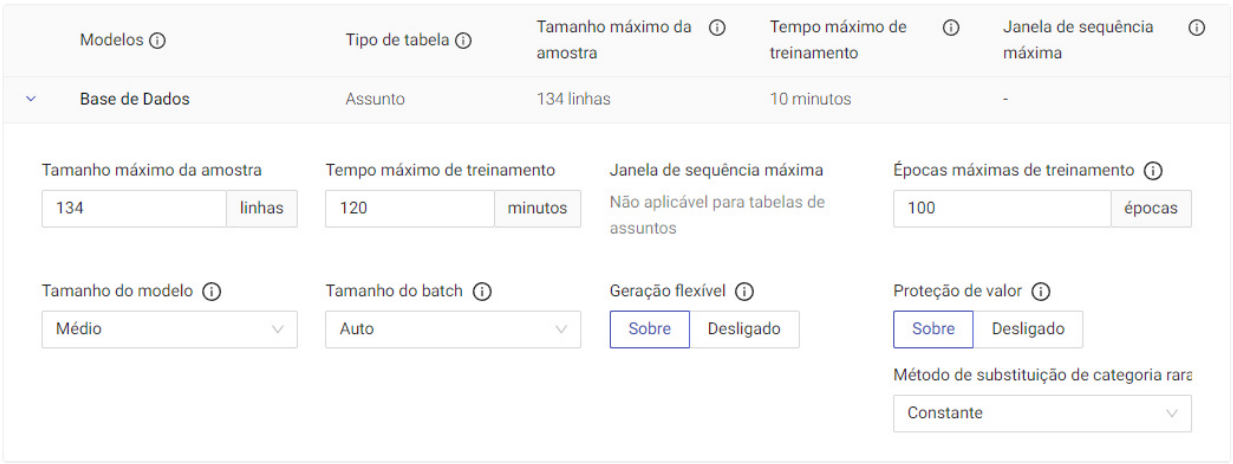
Distribuição Bivariada: 32,7%

Esses valores de distribuição indicam que o aumento do tempo de treinamento, de 1 para 10 minutos, contribuiu para um nível de acurácia superior, especialmente na métrica de Distribuição Univariada.

4.4. Modelo Computacional 3

Para o Modelo Computacional 3 a principal diferença observada foi a ampliação do tempo total necessário para a conclusão do treinamento, que chegou a um total de até 120 minutos. Os parâmetros específicos utilizados nessa configuração estão apresentados detalhadamente na Figura 6.

Figura 6 - Configuração do Modelo Computacional 3.



Fonte: Elaboração própria.

Após os ciclos de treinamento, o Modelo Computacional 3 apresentou os seguintes resultados de desempenho:

Distribuição Geral: 49,4%
Distribuição Univariada: 65,4%
Distribuição Bivariada: 33,3%

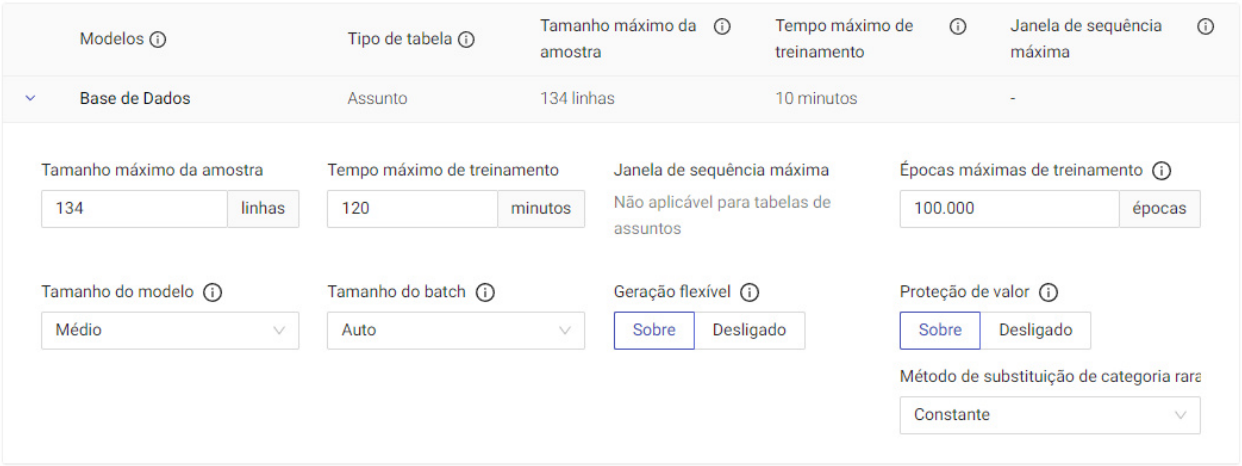
Esses resultados mostram uma melhoria significativa em comparação aos resultados iniciais do Modelo 1. Destaca-se especialmente o aumento da Distribuição Univariada, que alcançou

65,4%, indicando que o Modelo 3 se tornou mais eficaz em realizar previsões precisas ao considerar cada variável preditora individualmente.

4.5. Modelo Computacional 4

Para aprimorar ainda mais o desempenho dos modelos computacionais, no modelo Computacional 4, os autores optaram por aumentar consideravelmente o número de interações de treinamento, passando de 100 para 100.000, conforme ilustrado na Figura 7. Essa estratégia foi adotada com o objetivo de maximizar o potencial do modelo. Contudo, apesar desse incremento significativo nas interações, o resultado do Modelo 4 não mostrou melhorias relevantes e, na verdade, foi inferior ao desempenho do Modelo 3, que até então apresentava os melhores resultados.

Figura 7 – Configuração do Modelo Computacional 4.



Fonte: Elaboração própria.

Os principais desempenhos do Modelo 4, após a ampliação das interações, foram:

Distribuição Geral: 47,8%
Distribuição Univariada: 63,7%
Distribuição Bivariada: 31,8%

Mesmo com o tempo de treinamento mantido em 120 minutos, a tentativa de aumentar drasticamente o número de interações não resultou em avanços no desempenho do Modelo 4. Ao contrário, observou-se uma diminuição em relação aos resultados anteriores. Essa análise

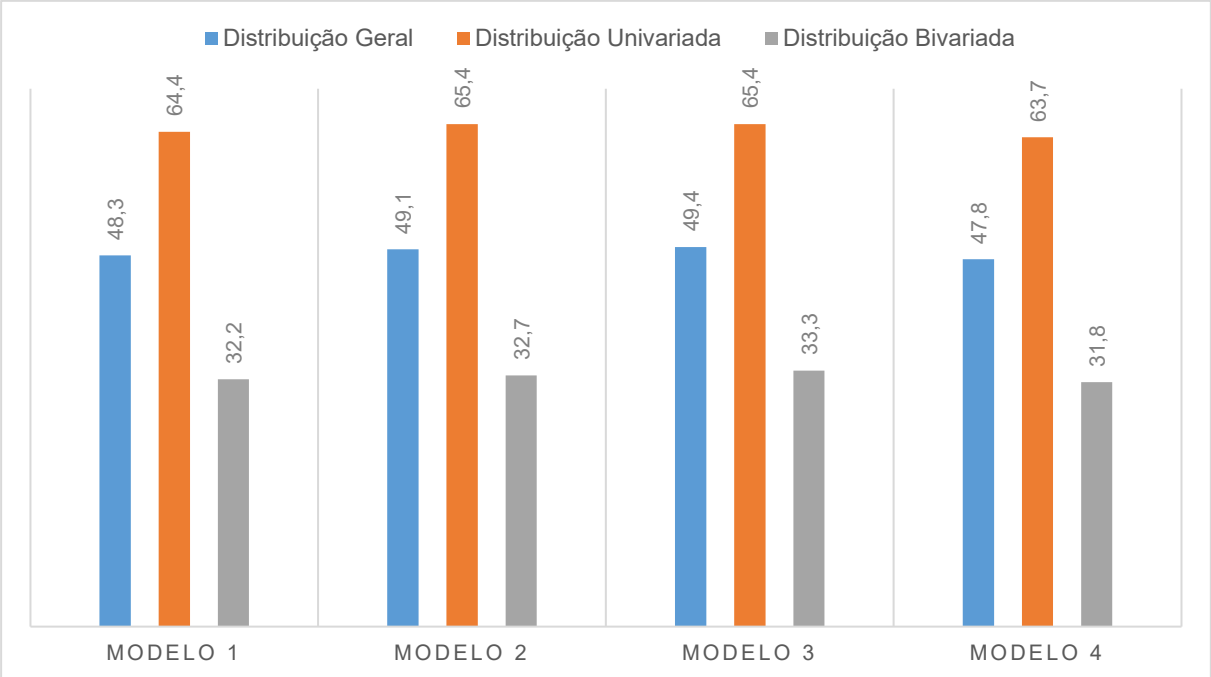
sugere que, para este conjunto de dados e configurações específicas, simplesmente aumentar o número de interações não é uma abordagem eficaz para melhorar a precisão do modelo. Fatores como a arquitetura do modelo, a seleção de hiperparâmetros e a qualidade dos dados de treinamento podem influenciar mais significativamente o desempenho.

5. Análise de Resultados

Neste estudo, foram investigados quatro modelos distintos com o propósito de avaliar seu desempenho na geração de dados sintéticos que replicam as características dos dados coletados em um levantamento sobre a utilização de Interfaces Neurais não invasivas para o controle do cursor de um mouse em uma tela de computador. A análise busca compreender a eficácia de cada modelo na tarefa de simular dados que se assemelham aos monitorados durante a operação das interfaces.

Os resultados obtidos para cada um dos modelos são apresentados na Figura 8, proporcionando uma visão detalhada de suas respectivas performances e contribuindo para a discussão sobre a viabilidade de diferentes abordagens na geração de dados sintéticos.

Figura 8 – Resultados para cada modelo computacional.



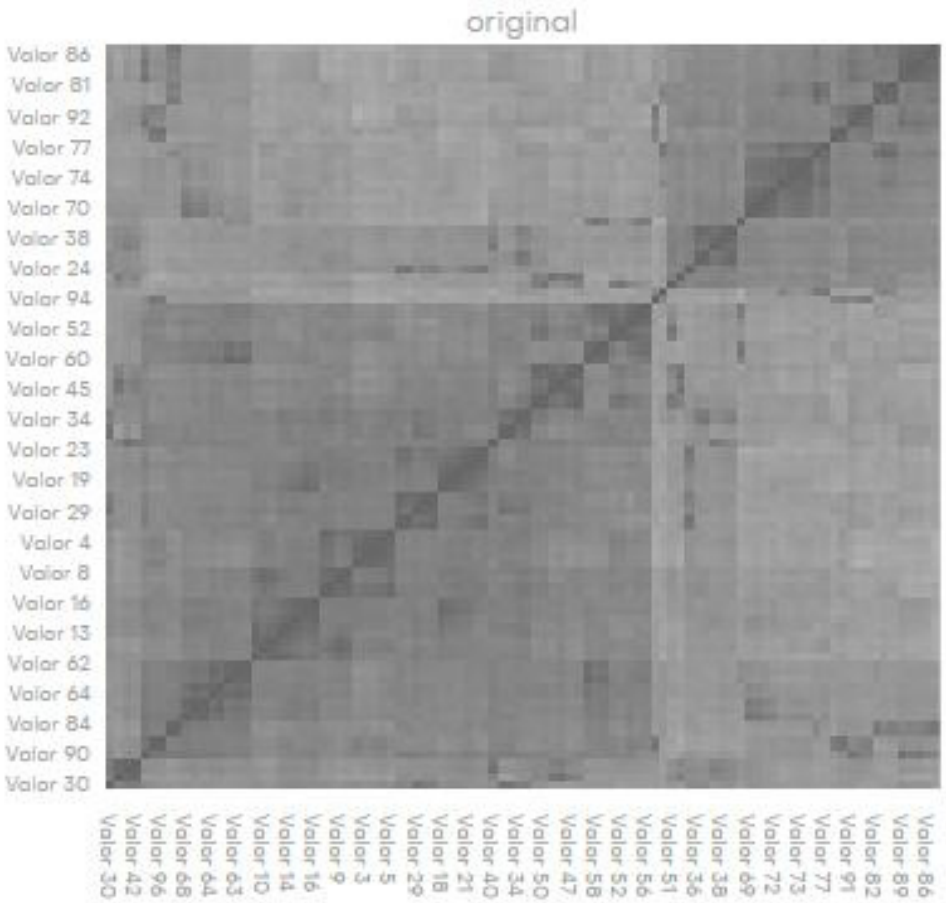
Fonte: Elaboração própria.

Este estudo avaliou três métricas principais para analisar o desempenho de modelos na geração de dados sintéticos: Distribuição Geral, Distribuição Univariada e Distribuição

Bivariada. A Distribuição Geral fornece uma visão abrangente da eficácia dos modelos em reproduzir características dos dados reais. A Distribuição Univariada analisa a capacidade dos modelos de prever variáveis individuais, permitindo uma avaliação detalhada da acurácia. Já a Distribuição Bivariada examina as interações entre pares de variáveis, essencial para entender influências conjuntas.

As variações de desempenho entre os modelos foram atribuídas às diferentes parametrizações utilizadas no treinamento das redes neurais, afetando a aprendizagem de padrões tanto univariados quanto bivariados. Essa análise comparativa oferece insights valiosos para a seleção e aprimoramento de abordagens em pesquisas futuras e aplicações práticas. Para a geração de dados sintéticos voltados a Interfaces Neurais, a adoção dessas métricas é crucial para garantir a produção de dados relevantes.

Figura 9 – Matriz de correlação dos dados originais.



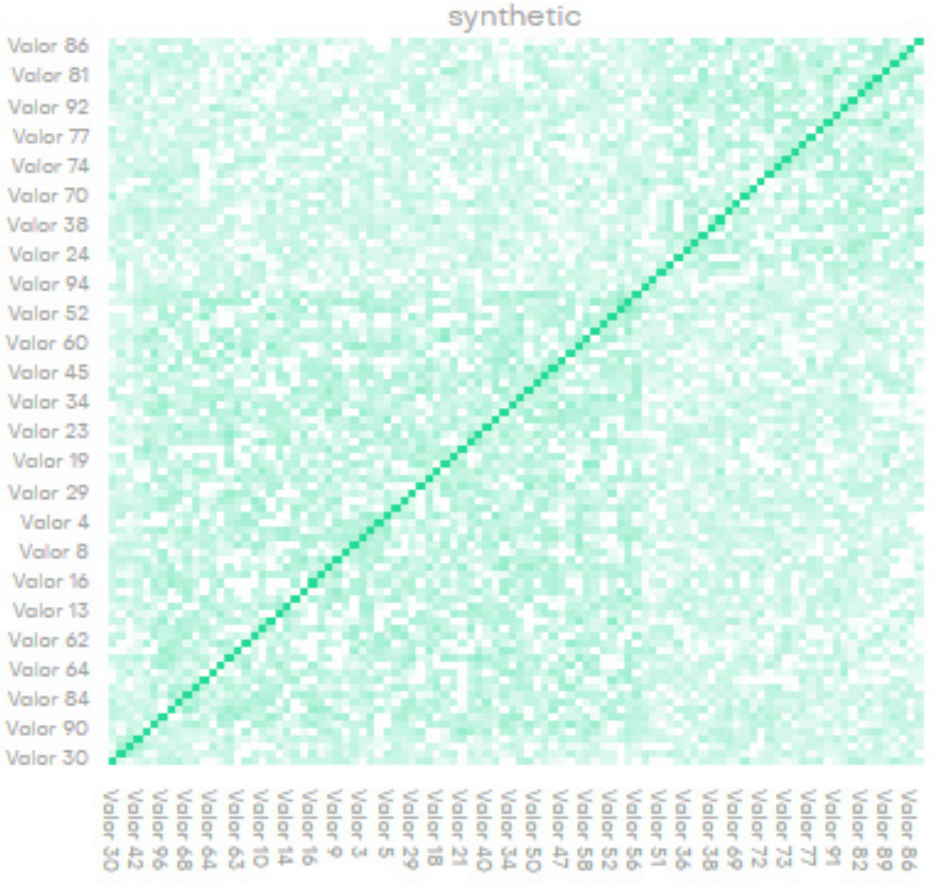
Fonte: Elaboração própria.

O Modelo Computacional 3 destacou-se entre os analisados, apresentando os melhores resultados gerais e, portanto, foi escolhido para a geração dos dados sintéticos. Com sua

implementação, foi criada uma matriz com dimensões de 134x98, refletindo a quantidade de registros dos dados originais. A escolha do Modelo 3 baseou-se em uma avaliação detalhada, considerando a Distribuição Univariada, Bivariada e a capacidade de replicar características estatísticas dos dados reais.

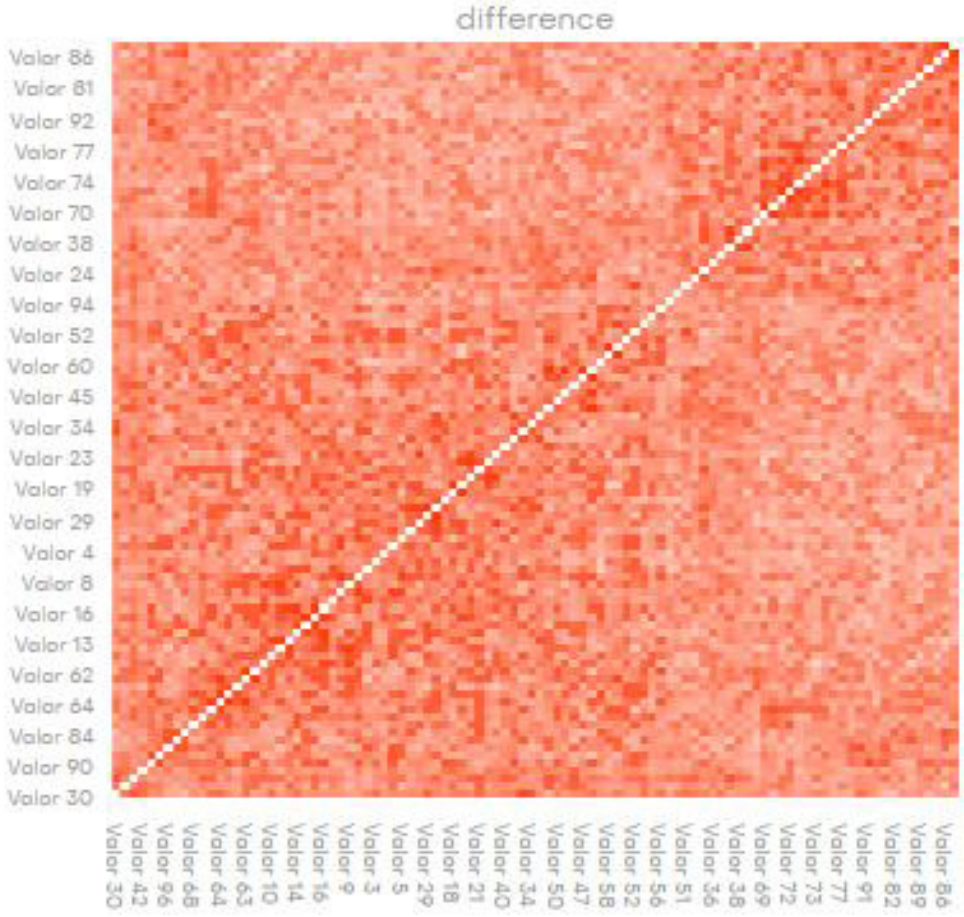
As matrizes de correlação, apresentadas nas Figuras 9, Figura 10 e Figura 11, ilustram as relações entre as variáveis. Cada célula da matriz representa o coeficiente de correlação entre duas variáveis, variando de -1 a 1, indicando a força e a direção da relação linear. Foram apresentadas três matrizes de correlação: dos dados originais, dos dados sintéticos e a diferenciação entre ambas. Essas matrizes são fundamentais para avaliar a qualidade dos dados sintéticos gerados, assegurando que as relações estatísticas essenciais sejam mantidas para utilização eficaz nas interações planejadas.

Figura 10 – Matriz de correlação dos dados sintéticos.



Fonte: Elaboração própria.

Figura 11 – Diferença entre a matriz de correlação dos dados originais e a matriz de correlação dos dados sintéticos.



Fonte: Elaboração própria.

No contexto da análise das distribuições univariada e bivariada neste estudo, foram examinados os dados ilustrados nas Figuras 12 e 13. O modelo sintético, representado pelas linhas verdes, demonstrou um desempenho significativo na distribuição univariada, enquanto os dados originais são apresentados em cinza. Os dados sintéticos acompanham de forma semelhante as variações dos dados originais, indicando uma boa aderência ao sistema real.

Em relação à distribuição bivariada, a análise revelou que os dados originais tendem a apresentar maior densidade entre variáveis próximas, enquanto os dados sintéticos mostram uma tendência a serem mais espaçados. Essa diferença pode estar relacionada à estrutura computacional utilizada, sugerindo que esse parâmetro poderia ser melhor explorado em pesquisas futuras.

Quando os dados são comparados em distâncias maiores, como entre os valores 1 e 93 ou 1 e 92, a divergência entre os dados originais e sintéticos se torna menos perceptível. Isso indica

que, na distribuição bivariada, os dados originais se agrupam mais estreitamente entre valores de variáveis próximas, com maior espaçamento entre variáveis distantes. Em contrapartida, os dados sintéticos tendem a ser espaçados tanto entre variáveis próximas, como 95 e 96 ou 31 e 32, por exemplo.

Figura 12 – Distribuição univariada dos dados originais e sintéticos.

Fonte: Elaboração própria.

Figura 13 – Distribuição bivariada dos dados originais e sintéticos.

Fonte: Elaboração própria.

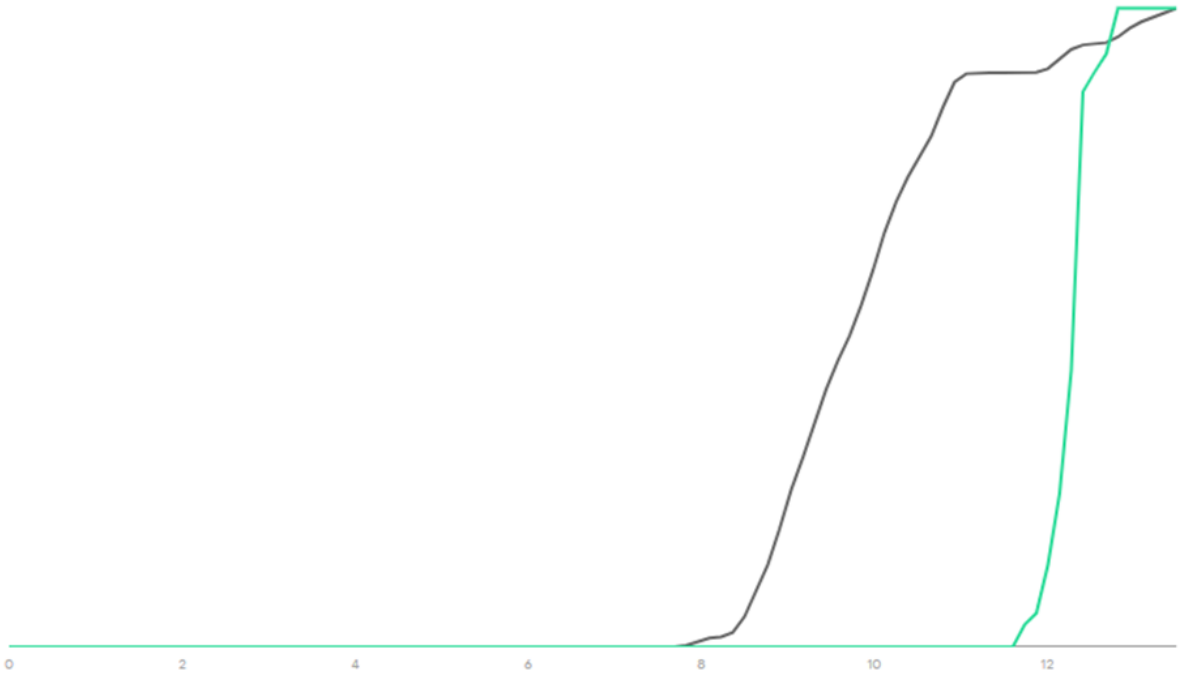
Os dados sintéticos devem estar próximos, mas não "excessivamente próximos" (iguais), dos dados originais, a fim de garantir a confidencialidade e a variabilidade das amostras. Essa

proximidade pode ser avaliada empiricamente por meio da medição das distâncias entre os registros sintéticos e seus correspondentes mais próximos nos dados originais.

Para fundamentar essa avaliação, emprega-se o conceito de Distribuição Cumulativa até os Registros mais Próximos (DCR), que é utilizado em estatística para analisar a distribuição de valores e suas respectivas proximidades.

As distâncias calculadas podem ser comparadas com as distâncias observadas nos dados originais, que servem como referência para definir o que significa estar "excessivamente próximo". Na visualização apresentada na Figura 14, as distâncias para os dados sintéticos são indicadas em verde, enquanto as distâncias para os dados originais aparecem em cinza. Uma linha verde significativamente à esquerda da linha cinza nos gráficos de densidade cumulativa indicaria que os dados gerados estão muito próximos dos registros reais, sinalizando um potencial risco à confidencialidade dos dados.

Figura 14 – Distribuição Cumulativa até os Registros mais Próximos (DCR).



Fonte: Elaboração própria.

6. Conclusões

A pesquisa apresentada neste artigo evidencia o potencial significativo da geração de dados sintéticos no desenvolvimento de Interfaces Cérebro-Computador (ICCs), especialmente em interfaces neurais não invasivas. Utilizando a plataforma Mostly AI, foram gerados dados sintéticos que replicam as características estatísticas de dados reais de eletroencefalografia (EEG) obtidos durante uma tarefa de controle de cursor via ICC.

A qualidade dos dados sintéticos foi avaliada por meio de métricas estatísticas, incluindo a matriz de correlação e as distribuições univariada e bivariada. A matriz de correlação facilitou a visualização das inter-relações entre variáveis, enquanto as análises de distribuição confirmaram a capacidade do modelo em reproduzir as propriedades estatísticas dos dados originais. Diferentes modelos computacionais foram treinados e comparados, com o Modelo Computacional 3 apresentando o desempenho mais promissor. Esse resultado destaca a importância da otimização dos parâmetros de treinamento para obter dados sintéticos mais precisos.

Além disso, a análise da Distribuição Cumulativa até os Registros mais Próximos (DCR) foi utilizada para assegurar que os dados sintéticos não fossem idênticos aos dados reais, protegendo a confidencialidade das informações e garantindo a variabilidade do modelo.

Esta pesquisa contribui para o campo das ICC ao demonstrar a viabilidade da utilização de dados sintéticos para o desenvolvimento e aprimoramento de sistemas. A geração de dados sintéticos de alta qualidade pode mitigar os desafios associados à escassez de dados reais, acelerando o progresso na área. Os métodos e resultados apresentados indicam oportunidades para futuras investigações, incentivando a exploração de diferentes plataformas e modelos de geração de dados sintéticos, além do desenvolvimento de novas métricas para avaliação de qualidade.

A aplicação de dados sintéticos em ICCs tem o potencial de impulsionar a criação de sistemas mais robustos, eficientes e personalizados, ampliando as possibilidades em áreas como saúde, comunicação e entretenimento. Esses dados sintéticos podem ser utilizados no treinamento e validação de modelos de inteligência artificial para ICCs, onde a disponibilidade de dados robustos e representativos é essencial. Além disso, permitem o teste de diferentes algoritmos de processamento de sinais cerebrais e a ampliação de bases de dados, oferecendo um ambiente computacional acessível para pesquisa. A utilização de dados sintéticos também garante a privacidade dos dados originais durante a colaboração e o compartilhamento,

contribuindo para a criação de sistemas de ICC mais eficazes e personalizados em diversos setores.

Referências

Bastos, J. L. D., & Duquia, R. P. (2007). Medidas de dispersão: os valores estão próximos entre si ou variam muito. *Scientia Medica*, *17*(1), 40-44.

Cengiz, K., J, S., K, U. P., L, G. H., & M, D. K. S. (2024). Development of a calibration-free brain-computer interface utilizing common spatial patterns and artificial neural networks for EEG signal analysis. In *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-6). Istanbul, Turkiye. https://doi.org/10.1109/HORA61326.2024.10550748

Habashi, A. G., Azab, A. M., Eldawlatly, S., & Aly, G. M. (2023). Motor imagery classification enhancement using generative adversarial networks for EEG spectrum image generation. *In 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 354-359). L'Aquila, Italy. https://doi.org/10.1109/CBMS58004.2023.00243

Instituto Nacional de Ensino (Ed.). (2023). *Estatística aplicada*. Retrieved from https://www.institutoine.com.br/arquivos/estatistica_aplicada_6010077c83f2b.pdf.

Macedo, J. P. de O., Barreto, G., & Miguel, P. V. de O. (2024). Introdução ao estudo de geração de dados sintéticos para aplicação em interfaces neurais interativas. In *XIV Rede de Investigadores da Qualidade (Riqual)*.

Malhado, C. H. M., Ramos, A. de A., Carneiro, P. L. S., Azevedo, D. M. M. R., & Martins Filho, R. (2008). Distribuição univariada e bivariada das características de crescimento de bubalinos de corte no Brasil. *Revista Científica de Produção Animal*, 10(1), 69-80. http://www.alice.cnptia.embrapa.br/alice/handle/doc/578730

Miguel, P. V. O. (2010). *ECOLIG - O Protocolo Semiótico para Comunicação Homem-Máquina que Utiliza Interfaces do tipo Cérebro-Computador* (Tese de doutorado). Universidade Estadual de Campinas (UNICAMP). Recuperado de https://repositorio.unicamp.br/Busca/Download?codigoArquivo=467129

Thompson, K. M. (1999). Developing Univariate Distributions from Data for Risk Analysis. *Human and Ecological Risk Assessment: An International Journal*, *5*(4), 755–783. https://doi.org/10.1080/10807039.1999.9657739.

Authors Profiles

João Paulo de Oliveira Macedo is a postgraduate student at the Faculty of Electrical and Computer Engineering - State University of Campinas - Brazil. He is currently a professor at CEETEPS - Brazil and a Teaching Support Professional for Research and Extension at State University of Campinas. His research interests are in the areas of synthetic data, intelligent systems and artificial intelligence**.**

Gilmar Barreto has received a PhD. from State University of Campinas - Brazil. He is currently a professor at the State University of Campinas - Brazil. Her research interests are in the areas of fuzzy systems, multivariable systems, control, multiobjective optimization, electrochemistry and data modeling.

Paulo Victor de Oliveira Miguel has received a PhD. from State University of Campinas - Brazil. He is currently a professor at the State University of Campinas - Brazil. Her research interests are in the areas of synthetic data, artificial intelligence, computational semiotics, neuroscience and education.