UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Eduarda Costa Coppo

Geração de dados sintéticos para anonimização de dados de saúde por meio de redes adversárias generativas e uma função de perda customizada

Juiz de Fora
2024

Eduarda Costa Coppo

Geração de dados sintéticos para anonimização de dados de saúde por meio de redes adversárias generativas e uma função de perda customizada

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Saulo Moraes Villela

Coorientador: Prof. Dr. Marcelo Bernardes Vieira

Juiz de Fora

2024

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Coppo, Eduarda Costa.
Geração de dados sintéticos para anonimização de dados de saúde por meio de redes adversárias generativas e uma função de perda customizada / Eduarda Costa Coppo. – 2024.
48 f. : il.

Orientador: Saulo Moraes Villela
Coorientador: Marcelo Bernardes Vieira
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2024.

1. Geração de dados sintéticos. 2. Anonimização de dados de saúde. 3. Redes adversárias generativas. 4. Função de perda customizada. I. Villela, Saulo Moraes, orient. II. Vieira, Marcelo Bernardes, coorient. III. Título.

Eduarda Costa Coppo

Geração de dados sintéticos para anonimização de dados de saúde por meio de redes adversárias generativas e uma função de perda customizada

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação. Área de concentração: Ciência da Computação.

Aprovada em 11 de abril de 2024.

BANCA EXAMINADORA

Prof. Dr. Saulo Moraes Villela - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Marcelo Bernardes Vieira - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Heder Soares Bernardino
Universidade Federal de Juiz de Fora

Prof. Dr. Vinicius Layter Xavier

Juiz de Fora, 15/01/2024.

Documento assinado eletronicamente por **Saulo Moraes Villela**, **Professor(a)**, em 27/02/2025, às 14:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Vinicius Layter Xavier**, **Usuário Externo**, em 20/03/2025, às 18:16, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Heder Soares Bernardino**, **Professor(a)**, em 24/03/2025, às 10:53, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcelo Bernardes Vieira**, **Professor(a)**, em 24/03/2025, às 11:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1664662** e o código CRC **AA388B36**.

AGRADECIMENTOS

Primeiramente gostaria de agradecer aos meus pais, que sempre me incentivaram a seguir meus estudos e ofereceram todo o apoio possível para que eu os concluísse.

Aos meus orientadores, Saulo e Marcelo, pelo apoio, paciência e orientação durante o desenvolvimento deste trabalho.

Aos meus colegas da UFES e da UFJF, que estiveram comigo tanto em momentos de descontração e amizade quanto em momentos acadêmicos. O apoio e presença das pessoas com quem convivi durante esse período foi fundamental para meu crescimento pessoal e acadêmico.

"True education is a kind of never ending story — a matter of continual beginnings, of habitual fresh starts, of persistent newness." (J. R. R. Tolkien).

RESUMO

Dados de saúde podem apresentar vulnerabilidades por conterem informações privadas e sensíveis, as quais devem ser consideradas em contextos que exigem a manipulação desses dados. Uma das soluções para o problema de exposição de informações sensíveis é a geração de amostras sintéticas que representem adequadamente o conjunto de dados a ser estudado. Isso permitiria uma substituição da base de dados reais, isto é, a base de dados original, pelo novo conjunto de amostras sintéticas em estudos que propõe resolver alguma tarefa envolvendo essa base de dados. Entre os vários métodos de geração de dados sintéticos, a utilização de redes adversárias generativas (GANs) destaca-se no campo da geração de imagens. Para dados tabulares, embora os estudos ainda sejam limitados, as possibilidades são amplas e demonstram a flexibilidade dessas redes na geração de amostras de menor dimensionalidade. O método proposto baseia-se em uma arquitetura de GAN, complementada por um método de treinamento que emprega uma função de perda customizada e diferentes abordagens para sua aplicação, a fim de obter uma distribuição das amostras sintéticas o mais próxima possível à real, ou seja, preservando as características estatísticas dos dados reais, bem como correlações entre seus atributos. A principal hipótese é que a GAN, aliada ao método de treinamento proposto, é capaz de gerar dados cuja distribuição se aproxima da distribuição dos dados reais. Os resultados indicam que a utilização de uma função de perda baseada na aproximação de suas matrizes de covariância favorece a geração de dados sintéticos cujos atributos têm distribuição mais próxima aos atributos dos dados reais, fazendo com que esse conjunto de dados sintéticos possa ser utilizado nas aplicações requeridas por diversas tarefas de aprendizado de máquina.

Palavras-chave: Redes adversárias generativas. Aumento de dados. Dados tabulares.

ABSTRACT

Health data may present vulnerabilities by containing private and sensitive information, which must be considered in contexts that require the manipulation of such data. One solution to the problem of exposing sensitive information is the generation of synthetic samples that accurately represent the dataset to be considered, allowing it to be replaced in the works proposed for a specific task. Among the various methods for generating synthetic data, the use of generative adversarial networks (GANs) stands out in the field of image generation. For tabular data, although studies are still limited, the possibilities are vast and demonstrate the flexibility of these networks in generating samples of lower dimensionality. The proposed method is based on a GAN architecture, supplemented by a training method that employs a custom loss function and different approaches for its application. The goal is to obtain a distribution of the synthetic samples as faithful as possible to the real ones. The main hypothesis is that GAN, combined with the proposed training method, would be capable of generating data whose distribution closely approximates that of the real data. The results indicate that the use of a loss function, based on the approximation of two distributions, promotes the generation of more realistic data, which can be used in the applications required by various machine learning tasks.

Keywords: Generative adversarial networks. Data augmentation. Tabular data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BCE	*Binary Cross-Entropy*
GAN	*Generative Adversarial Network*
MSE	*Mean Squared Error*
ReLU	*Rectified Linear Unit*
VAE	*Variational Autoencoders*

SUMÁRIO

1 INTRODUÇÃO

A anonimização pode ser definida como o processo de transformar dados pessoais, anteriormente identificáveis, em irreconhecíveis ou irrecuperáveis, visando proteger a privacidade das pessoas a quem esses dados se referem. O principal objetivo desse processo é assegurar que as informações contidas em uma base de dados específica não possam ser vinculadas a indivíduos identificáveis, tornando-as seguras para serem compartilhadas e analisadas sem comprometer a privacidade dos envolvidos (EMAM; ARBUCKLE, 2013). A questão central da anonimização, no contexto deste trabalho, é: como proteger dados que contêm informações privadas enquanto se mantém sua qualidade para uso em diversas tarefas de aprendizado de máquina? A importância desta questão deriva do uso crescente de tecnologias de aprendizado de máquina e inteligência artificial que empregam dados de áreas sensíveis, como a saúde, especialmente para tarefas de diagnóstico (IBRAHIM; ABDULAZEEZ, 2021). Surge, então, a necessidade de compartilhar essas informações de maneira segura, criando um ambiente em que as tarefas de aprendizado de máquina propostas possam ser realizadas após o tratamento e pré-processamento dos dados a serem utilizados.

Uma das abordagens analisadas para a execução da anonimização desses dados, a ser discutida neste trabalho, é o uso de redes adversárias generativas (*generative adversarial networks* – GANs) visando ao aumento de dados (*data augmentation*), um método amplamente empregado na área de aprendizado de máquina que consiste em expandir a quantidade de dados inicialmente disponíveis (SHORTEN; KHOSHGOFTAAR, 2019). O aumento de dados adiciona novas amostras à base original, o que configura uma solução para o problema de desbalanceamento de classes, comum em tarefas de aprendizado de máquina. Além disso, pode-se optar por utilizar apenas os dados gerados pela GAN, uma solução para o problema de anonimização, em que apenas os dados que não se referem a pessoas reais comporiam a base de dados final. A arquitetura de uma GAN consiste em redes neurais formadas por dois modelos: um gerador e um discriminador, que competem entre si por meio de uma estratégia de aprendizado competitivo. O gerador é treinado para aprender a criar dados sintéticos cada vez mais realistas, gerando um novo conjunto de dados, enquanto o discriminador é aprimorado para distinguir entre dados sintéticos e reais.

O aprendizado adversário, ou competitivo, entre o gerador e o discriminador resulta em um aprimoramento contínuo da capacidade do gerador de criar dados sintéticos de alta qualidade visual (TRIASTCYN; FALTINGS, 2018). Assim, a geração de dados por meio de GANs apresenta uma solução para realizar tarefas de aprendizado de máquina com uma base de dados totalmente sintética, gerada a partir de um processo que visa extrair as características essenciais da base original para o treinamento de modelos destinados a resolver essas tarefas. Isso possibilita que os novos conjuntos gerados possam substituir os

originais sem comprometer o desempenho dos modelos.

O presente trabalho visa introduzir melhorias, principalmente nas etapas de treinamento e validação, bem como propor uma nova abordagem ao problema investigado por Piacentino e Angulo (2020), o de anonimização de dados através da geração de amostras sintéticas, geradas por uma rede especializada. Para tanto, desenvolveu-se um método de treinamento com o objetivo de ampliar os dados por meio da aplicação de GANs em conjunto com a inferência estatística, a fim de produzir novas amostras que se assemelham à distribuição da base de dados reais, obtendo resultados mais confiáveis. Este mesmo objetivo é perseguido na etapa de validação, onde são empregados métodos estatísticos, como a representação e comparação dos atributos através de *boxplots* para avaliar os resultados e tabelas que possam resumir os dados através de métricas clássicas como a média, a mediana e o desvio padrão. Além da validação estatística, a avaliação visual, comumente empregada em tarefas de geração de dados, também foi utilizada em conjunto, a fim de se ter parâmetros diversos para avaliar os resultados obtidos.

A utilização de GANs com o propósito de promover o aumento de dados e utilizá-lo para solucionar o problema de anonimização demonstra um significativo potencial em garantir a privacidade dos indivíduos representados nesses conjuntos, enquanto possibilita o uso seguro e eficaz das informações em tarefas de análise e modelagem. Os princípios de funcionamento das GANs, as etapas do processo de geração de dados sintéticos, as métricas para avaliar a qualidade desses dados, além de estudos de caso e resultados experimentais que evidenciam a eficácia e os desafios associados a essas estratégias, serão explorados em detalhes nos capítulos seguintes. Essa discussão contribuirá para uma melhor compreensão desse conjunto de métodos e suas implicações em cenários reais de aplicação, onde o aumento de dados pode ser benéfico.

1.1 DEFINIÇÃO DO PROBLEMA

O problema abordado neste trabalho divide-se em duas etapas principais: a geração de dados sintéticos por meio de GANs para o aumento de dados e a validação dos resultados utilizando métodos estatísticos e visuais. É essencial que os dados sintéticos gerados preservem ao máximo as características estatísticas, distribuições e correlações dos dados reais, garantindo assim que os modelos treinados com esses dados sintéticos mantenham sua eficácia na execução da tarefa. Na fase de validação, essa preservação de características e a capacidade de execução de tarefas são confirmadas por meio dos resultados e das análises estatísticas dos dados.

Para a primeira parte, enfrenta-se um desafio de aprendizado de máquina: dada uma base de dados, o objetivo é gerar novos dados a partir dela. Nesta tarefa de geração de dados sintéticos, o conjunto de dados em si serve como base de treinamento. No caso de dados tabulares, eles precisam ser normalizados e transformados para um formato de

representação visual após o pré-processamento e a eliminação de *outliers*, visto que um dos critérios de avaliação utilizados aqui é o critério visual. As amostras são padronizadas de modo que cada atributo esteja dentro de um intervalo fixo. Após a normalização e a transformação da base para o formato desejado, os dados são utilizados para treinar a GAN, cujo objetivo final é gerar uma nova base.

Para a segunda parte do problema, a validação é realizada por meio de análises estatísticas, inferência e métodos visuais. Este último visa acompanhar o treinamento da rede e oferecer um mecanismo de avaliação qualitativa, utilizado em conjunto com as análises estatísticas para comprovar a qualidade dos dados gerados, assegurando que estes reproduzam o mais fielmente possível a distribuição e as características das amostras originais da base.

1.2 OBJETIVOS

Os principais objetivos deste trabalho podem ser categorizados em objetivos relacionados ao uso de GANs para o aumento de dados e objetivos associados à metodologia de treinamento das GANs para otimizar tanto a geração de um novo conjunto de dados quanto a validação dos resultados obtidos.

No primeiro grupo, o objetivo principal é empregar o método de aumento de dados para gerar novas amostras, utilizando o estudo realizado por Piacentino e Angulo (2020) como referência.

Quanto à metodologia de treinamento, o objetivo principal é desenvolver um fluxo de operações na arquitetura da GAN e na formulação da função de perda que considere a distribuição e as características estatísticas dos dados de entrada, buscando assim melhorar a qualidade dos dados gerados pela rede e, consequentemente, de seus atributos.

1.3 CONTRIBUIÇÕES

As principais contribuições deste trabalho incluem:

- A geração de dados sintéticos que se assemelham aos dados reais, mas não contêm informações identificáveis;
- A introdução de uma função de perda customizada que emprega princípios de inferência estatística para melhorar a qualidade dos dados sintéticos;
- A aplicação de métodos estatísticos na validação dos dados gerados, proporcionando um recurso adicional para avaliar o desempenho do treinamento;
- A adoção de um mecanismo de treinamento que incorpora conceitos de janela (*window*) e passo (*stride*), semelhantes aos utilizados em convoluções, o que aprimora

a capacidade das GANs de reproduzir estruturas e relações nos dados, tornando o treinamento mais eficaz.

1.4 ORGANIZAÇÃO

Este trabalho está organizado em seis capítulos. O Capítulo 1 introduz os problemas abordados, suas definições, objetivos e as contribuições alcançadas. O Capítulo 2 expõe a fundamentação teórica, estabelecendo a base para os conceitos, ideias e métodos aplicados. O Capítulo 3 discorre sobre os trabalhos relacionados consultados, que contribuem para o desenvolvimento das soluções propostas e que se relacionam com este estudo. No Capítulo 4 são descritos os métodos propostos para a realização das investigações e experimentos. O Capítulo 5 relata os resultados alcançados nos experimentos conduzidos, e o Capítulo 6 apresenta as conclusões e considerações finais do estudo, além de sugerir possíveis trabalhos futuros e melhorias.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda a fundamentação teórica estudada para a realização deste trabalho, incluindo conceitos e explicações cruciais para o entendimento do problema proposto e sua solução. A Seção 2.1 discute a teoria das GANs, enquanto a Seção 2.2 explora as funções de perda e o desenvolvimento de funções customizadas, um aspecto fundamental para compreender a metodologia utilizada no treinamento do modelo.

2.1 REDES ADVERSÁRIAS GENERATIVAS

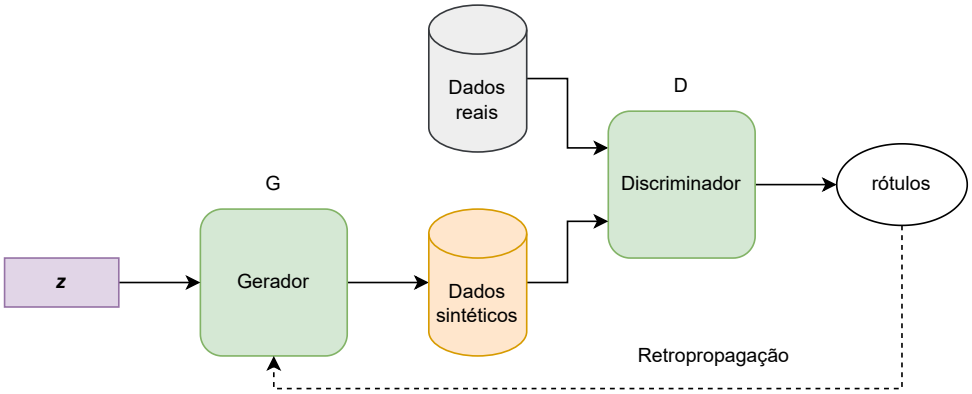
Goodfellow et al. (2014) introduziram uma arquitetura composta por duas redes neurais profundas em competição: o gerador e o discriminador. As GANs são modelos de aprendizado de máquina que adotam uma abordagem de aprendizado não supervisionado e têm sido aplicadas em uma vasta gama de domínios, possibilitando a criação de dados sintéticos utilizáveis para treinar modelos em situações onde os dados reais são escassos, de difícil acesso, ou, como no contexto deste trabalho, quando contêm informações sensíveis. As GANs são amplamente usadas para gerar dados sintéticos realistas em áreas como visão computacional (LEDIG et al., 2017), processamento de linguagem natural (YU et al., 2017) e ciência de dados, exemplificadamente na questão do aumento de dados (FRID-ADAR et al., 2018).

As GANs consistem em duas redes neurais: o gerador, ou rede generativa, responsável por criar amostras sintéticas a partir de ruído aleatório, e o discriminador, ou rede discriminativa, treinado para diferenciar entre as amostras reais, provenientes de um conjunto de dados, e as sintéticas, produzidas pelo gerador. O objetivo dessas redes é aprimorar continuamente seu desempenho por meio de uma competição entre elas, até que se atinja um equilíbrio. Neste ponto, o gerador consegue produzir dados sintéticos altamente realistas, e o discriminador, identificar com alta precisão a autenticidade dos dados. Assim, o gerador alcança o propósito de gerar dados de alta qualidade que se almeja ao treinar uma GAN.

A Figura 1 ilustra o processo de treinamento de uma GAN. Inicialmente, o gerador G, a partir de um vetor latente $\boldsymbol{z}$, também conhecido como vetor de ruído (*noise vector*), gera amostras sintéticas que são combinadas com dados reais para formar um conjunto de treinamento. O discriminador D recebe esse conjunto e classifica cada amostra, indicando se ela provém da base de dados reais ou se foi produzida pelo gerador. Por meio do algoritmo de retropropagação, G aprimora sua habilidade de gerar amostras cada vez mais realistas, enquanto D melhora sua capacidade de distinguir corretamente entre as amostras reais e as sintéticas. Este processo iterativo de competição promove melhorias contínuas em ambas as redes.

A formulação matemática da função de perda de uma GAN pode ser descrita como

Figura 1 – Ilustração simplificada do treinamento de uma GAN.



Fonte: Elaborada pela autora (2024).

um problema de otimização que envolve minimização e maximização, expressa por:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{dados}(x)}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \qquad (2.1)$$

onde $p_{\text{dados}(x)}$ representa a distribuição dos dados reais, z é o vetor latente, $G(z)$ denota a saída do gerador (os dados sintéticos gerados) e $D(x)$ é a saída do discriminador (os rótulos para cada amostra analisada).

Portanto, o objetivo de G é minimizar o termo $\mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$, isto é, reduzir a probabilidade de D classificar corretamente os dados sintéticos. Em outras palavras, G visa gerar amostras que induzam D a classificá-las como se originassem da base de dados reais. Já o objetivo de D é maximizar o termo $\mathbb{E}_{x\sim p_{\text{dados}(x)}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$, o que implica que D procura aumentar sua capacidade de distinguir entre amostras reais e sintéticas corretamente.

Um desafio comum no treinamento de GANs é o problema do colapso de modo (*mode collapse*), no qual o gerador começa a produzir um conjunto limitado de variações de amostras sem diversidade suficiente. Além disso, as GANs podem ser sensíveis às condições iniciais e aos hiperparâmetros, necessitando de um ajuste cuidadoso.

2.2 FUNÇÕES DE PERDA

As funções de perda têm um papel crucial no treinamento de modelos de aprendizado de máquina, servindo para quantificar a diferença entre as previsões do modelo e os valores reais dos dados. A modelagem adequada da função de perda é determinante para o sucesso do modelo, influenciando diretamente sua capacidade de aprender e se ajustar aos dados de forma eficiente (WANG et al., 2020). O objetivo principal de uma função de perda é fornecer um indicador quantitativo do desempenho do modelo durante o treinamento, visando à classificação correta, à regressão precisa de valores contínuos ou ao cumprimento de qualquer outra tarefa específica.

Entre as várias funções de perda adotadas em diferentes contextos, a entropia cruzada binária (*binary cross-entropy* – BCE) é amplamente empregada, especialmente em problemas de classificação binária, nos quais a saída deve ser uma probabilidade entre 0 e 1. A função é expressa por:

$$\text{BCE}(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right], \tag{2.2}$$

onde y_i representa o rótulo real da amostra i, $\hat{y}_i$ é o rótulo previsto pelo modelo para a mesma amostra i e n denota o número total de amostras no conjunto de treinamento.

Outra função de perda que será empregada nos experimentos é o erro quadrático médio (*mean squared error* – MSE), amplamente usado para medir o desempenho de modelos de regressão. O MSE é determinado pela média dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais, sendo expresso por:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2.3}$$

onde n representa o número de observações no conjunto de dados, y_i os valores reais para cada amostra i e $\hat{y}_i$ os valores previstos pelo modelo para cada amostra i.

A seleção da função de perda é determinada pelo tipo de tarefa e pelas especificidades do problema em questão. É crucial levar em conta a natureza dos dados, a distribuição dos rótulos, o equilíbrio entre as classes, entre outros fatores relevantes para a aplicação em estudo. Em certos casos, pode ser necessário customizar uma função de perda para atender às necessidades particulares do problema. Tal personalização pode envolver a inclusão de termos extras na função já existente ou a criação de uma função completamente nova, necessária, por exemplo, para enfrentar desequilíbrios de classes, penalizar determinados erros de previsão mais severamente ou adicionar restrições ao modelo.

Neste trabalho, além da utilização da BCE como função de perda principal, a medida conhecida como divergência de Kullback-Leibler (*Kullback-Leibler divergence* – KL *divergence*) como função de perda secundária teve um papel crucial no treinamento da GAN. Enquanto a BCE é aplicada de forma isolada em algumas fases do treinamento do discriminador para aprimorar a diferenciação entre dados reais e sintéticos, a introdução da divergência KL como uma função de perda adicional para o gerador estimula a geração de dados sintéticos que se assemelham mais à distribuição dos dados reais, aspecto explorado mais detalhadamente na Subseção **2.2.1** e na Seção 4.4. Além disso, o MSE também foi utilizado em alguns dos experimentos discutidos ao longo do trabalho.

2.2.1 *Kullback-Leibler Divergence*

Proposta por Kullback e Leibler (1951) como um meio de quantificar a diferença entre duas distribuições de probabilidade, a divergência KL, também conhecida como

entropia cruzada relativa, é uma medida que quantifica a diferença entre duas distribuições de probabilidade. Ela indica a quantidade média de informação perdida ao usar uma distribuição para aproximar a outra.

A formulação matemática da divergência KL envolve duas distribuições de probabilidade, P e Q. A equação calcula a média ponderada das diferenças entre as probabilidades atribuídas às duas distribuições para eventos correspondentes, na forma:

$$\mathrm{KL}(P \parallel Q) = \sum_i P(i) \ln \left(\frac{P(i)}{Q(i)} \right), \tag{2.4}$$

onde $P(i)$ é a probabilidade associada ao evento i na distribuição P e $Q(i)$ é a probabilidade associada ao mesmo evento i na distribuição Q. O somatório $\sum_i$ abrange todos os eventos possíveis. Quando a divergência é zero, isso indica que as duas distribuições são idênticas. Portanto, quanto maior a divergência, maior é a diferença entre as duas distribuições.

Essa medida possui, além da propriedade de não-negatividade, que assegura que a divergência sempre será maior ou igual a zero, a propriedade de não-simetria, significando que:

$$\mathrm{KL}(P \parallel Q) \neq \mathrm{KL}(Q \parallel P), \tag{2.5}$$

indicando que a divergência KL é sensível à ordem na qual as distribuições são comparadas.

A divergência KL também pode ser aplicada às matrizes de covariância de duas distribuições em análise. Para duas distribuições multivariadas, ambas com dimensão k, com matrizes de covariância Σ_P e Σ_Q, ela é expressa por:

$$\mathrm{KL}(P \parallel Q) = \frac{1}{2} \left(\mathrm{tr}(\Sigma_Q^{-1}\Sigma_P) + (\mu_Q - \mu_P)^T \Sigma_Q^{-1}(\mu_Q - \mu_P) - k + \ln \frac{\det(\Sigma_Q)}{\det(\Sigma_P)} \right), \tag{2.6}$$

onde tr indica o traço de uma matriz (a soma dos elementos da diagonal principal), μ_P e μ_Q são os vetores de média das distribuições P e Q, respectivamente, Σ_P e Σ_Q são as matrizes de covariância das distribuições P e Q, respectivamente, k é a dimensão dos vetores de média, e *det* refere-se ao determinante de uma matriz.

O formato da equação da divergência KL, considerando as covariâncias das distribuições, é crucial para este trabalho, pois foi adotado na função de perda da GAN. Esse aspecto será abordado com mais detalhes quando os procedimentos de treinamento e os resultados forem discutidos.

3 TRABALHOS RELACIONADOS

Este capítulo abrange toda a literatura consultada para a realização deste trabalho, assim como estudos correlatos, suas aplicações e limitações no contexto do problema em questão. A Seção 3.1 discute o emprego das GANs para a tarefa de anonimização de dados, incluindo os trabalhos referenciados sobre o tema e a arquitetura da rede neural adotada neste trabalho. A Seção 3.2 explora os principais métodos para avaliar os resultados gerados por GANs. Por fim, a Seção 3.3 discute a utilização de redes neurais em tarefas de inferência.

3.1 GANS PARA ANONIMIZAÇÃO DE DADOS

Piacentino e Angulo (2020) propuseram um método de anonimização utilizando GANs, visando gerar dados sintéticos que possam substituir os dados reais. Esse método busca manter a integridade estatística e preservar a privacidade, permitindo simultaneamente o uso dos dados para pesquisa e análise.

No âmbito da anonimização de dados de saúde, conforme desenvolvido por Piacentino e Angulo (2020), o gerador é treinado com um conjunto de dados reais, onde cada amostra representa os dados de um paciente. O objetivo é aprender as características e padrões dos dados reais para gerar novas amostras semelhantes. Entretanto, essas amostras sintéticas não estão ligadas aos pacientes reais, assegurando, assim, o anonimato.

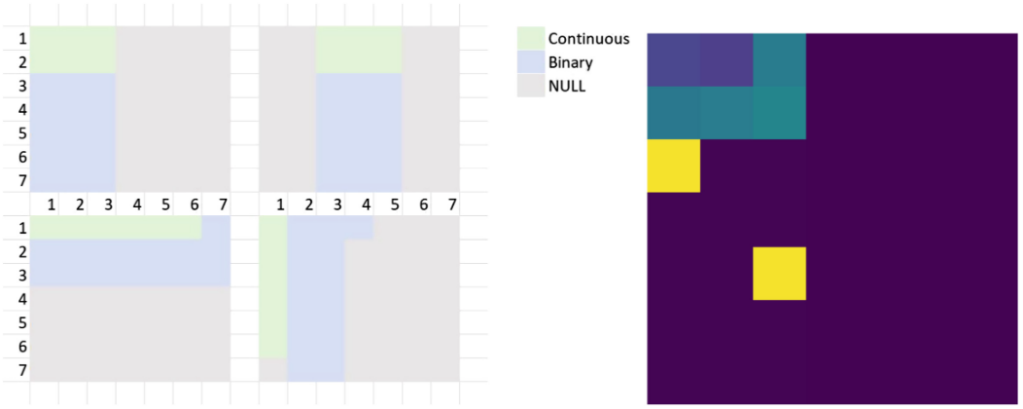
Os autores conduziram experimentos para avaliar a eficácia de sua abordagem proposta. Um conjunto de dados reais, consistindo em pacientes diagnosticados ou não com doenças da tireoide, foi utilizado para treinar uma GAN e gerar dados sintéticos que não contêm informações identificáveis dos pacientes. Sendo dados tabulares, foram convertidos em pequenas matrizes coloridas de tamanho 7×7, onde cada célula representa um atributo e cada matriz corresponde a uma amostra ou paciente. A Figura 2 ilustra essa representação dos dados em forma de matriz, onde cada célula representa um atributo e há uma escala de cores que evidencia a magnitude dos valores desses atributos, onde roxo representa valores mínimos, amarelo valores máximos e os tons de ver e azul valores intermediários. A figura demonstra que os atributos foram organizados de acordo com o tipo de dado, aspecto que será detalhado na Seção 4.2.

As amostras foram normalizadas antes de sua transformação em representação visual, garantindo a padronização de cada célula. A normalização aplicada é descrita por:

$$\frac{x - \min(x)}{\max(x) - \min(x)} \cdot 254 + 1, \tag{3.1}$$

onde x é um atributo de uma amostra $\boldsymbol{x}$. Essa equação normaliza cada uma das amostras para um valor no intervalo entre 1 e 255. Após o treinamento e a geração de dados pelo modelo, a transformação inversa é aplicada para a escala original dos atributos.

Figura 2 – Representação visual dos dados tabulares. À esquerda estão possíveis configurações, onde os atributos contínuos e binários de cada amostra são organizados e os espaços marcados como NULL representam células vazias, em que não há nenhum atributo. À direita, o exemplo de uma amostra.



Fonte: Piacentino, Guarner e Angulo (2021).

A rede neural empregada neste trabalho foi obtida em um artigo de Lakhey (2019), publicado na plataforma *Medium*. Trata-se de uma rede sem camadas convolucionais, possuindo complexidade adequada para treinar matrizes de pequeno porte, tal como a representação visual dos dados tabulares. Esta mesma rede foi reutilizada no método proposto, e as modificações implementadas serão discutidas e ilustradas na Seção 4.1.

Os resultados do estudo indicaram que os dados sintéticos gerados preservaram as características estatísticas relevantes dos dados reais, ao mesmo tempo em que asseguraram o anonimato. Os autores concluem que essa abordagem representa uma promessa significativa para a anonimização de dados de saúde, ressaltando também a importância de mais pesquisas para aprimorar os métodos de geração de dados sintéticos e desenvolver maneiras eficazes de avaliar a utilidade e privacidade dos dados produzidos.

Além da arquitetura da GAN ter sido reutilizada neste trabalho com algumas modificações, a estratégia de padronização dos dados e a representação das amostras em sua representação visual também foram adotadas, mas utilizando matrizes de dimensões reduzidas, de tamanho 5×5, aspecto que será detalhado no Capítulo 4.

Park et al. (2018) discutem a importância da privacidade em uma sociedade em que dados pessoais são constantemente compartilhados e apontam que diversas soluções para o problema de anonimização apresentam duas limitações principais: informações privadas ainda podem ser vazadas se os atacantes possuírem algum conhecimento prévio ou outras fontes de informação e pode haver um impacto adverso na utilidade dos dados privatizados. Levando isso em consideração, os autores propõe um método chamado *table-*GAN, descrito como uma GAN estendida, com uma terceira rede neural além do gerador e do discriminador: a rede classificadora, ou classificador. O classificador é treinado com

os rótulos dos dados da base real e contribui para a integridade semântica das amostras sintéticas.

Outros estudos similares (YOON; DRUMRIGHT; SCHAAR, 2020) (HASHEMI et al., 2023) (RAJABI; GARIBAY, 2022) também abordam questões relacionadas à geração de conjuntos de dados sintéticos semelhantes aos dados reais. Cada artigo propõe uma abordagem específica para alcançar esse objetivo, podendo introduzir um modelo de GAN, funções de perda, um protocolo de treinamento, entre outros fatores que buscam alcançar um resultado satisfatório que atenda aos requisitos levantados pela tarefa. Dentre os modelos de GAN usados, está o *Wasserstein* GAN, ou WGAN, uma escolha motivada por ser mais estável durante o treinamento em comparação ao modelo mais simples de GAN, fazendo-o ser menos suscetível a problemas como colapso de modo, mencionado no Capítulo 2. Além do desenvolvimento de um protocolo de treinamento, os artigos também discutem a importância da validação dos dados sintéticos para garantir que esses mantenham as propriedades estatísticas e a privacidade dos dados reais.

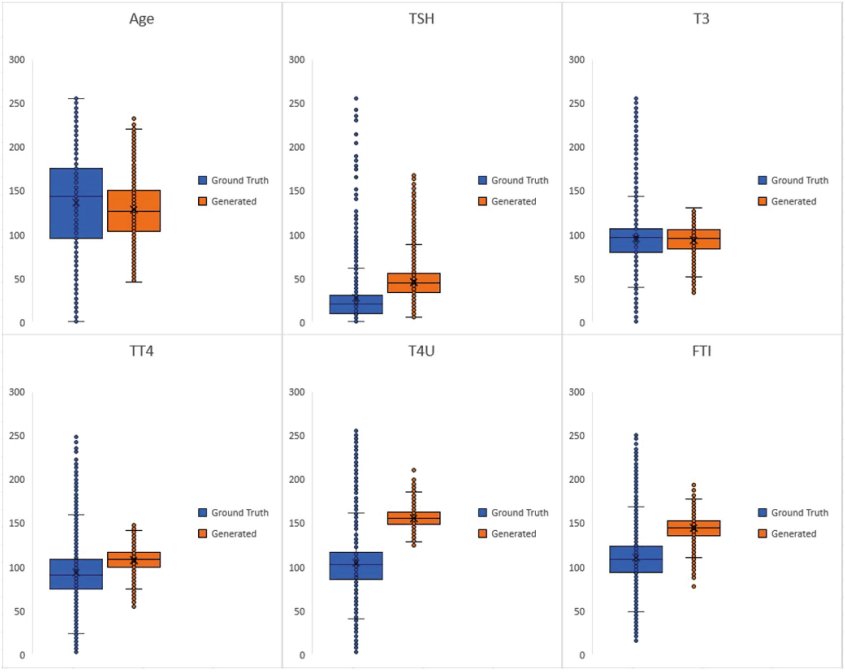
3.2 VALIDAÇÃO DOS RESULTADOS

Piacentino e Angulo (2020) empregaram ferramentas estatísticas e visuais para realizar uma avaliação qualitativa dos resultados. A utilização de *boxplots* para comparar as distribuições dos dados reais com os dados sintéticos, ilustrada na Figura 3, facilita a verificação da eficácia da GAN na geração de dados. É possível observar que os dados sintéticos apresentam uma distribuição próxima à dos dados da base original, porém com menor variância.

Juntamente com outros tipos de visualização, como gráficos que mostram os valores de perda por época e as amostras geradas durante o treinamento, é possível estimar a evolução dos resultados. A Figura 4, também retirada do artigo de Piacentino e Angulo (2020), demonstra a progressão das amostras produzidas pela rede generativa ao longo das épocas de treinamento. Nela, observa-se a primeira amostra gerada, um ruído inicial produzido pelo gerador. Com o avançar das épocas, o gerador aprimora sua capacidade de criar amostras cada vez mais similares àquelas da base de dados reais, evidenciado pela organização progressiva dos atributos nas representações visuais.

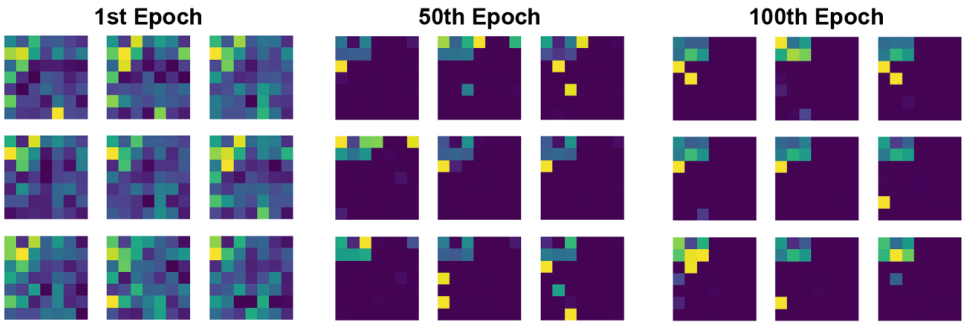
Este trabalho também incluirá a análise das amostras geradas pela rede ao longo das épocas e a distribuição dos atributos contínuos dos dados reais em comparação com os dados sintéticos produzidos pelo gerador. Adicionalmente, serão apresentados gráficos dos valores de perda por época, fornecendo uma visão clara de como as duas redes adversárias evoluem ao longo do treinamento.

Figura 3 – *Boxplots* com a distribuição dos dados reais e sintéticos.



Fonte: Piacentino, Guarner e Angulo (2021).

Figura 4 – Evolução das amostras geradas ao longo das épocas de treinamento.



Fonte: Piacentino, Guarner e Angulo (2021).

3.3 INFERÊNCIA COM REDES NEURAIS

A inferência em estatística, ciência de dados e aprendizado de máquina envolve a estimativa de propriedades não diretamente observáveis nos dados com base em informações acessíveis por meio deles. Em outras palavras, consiste em usar as informações disponíveis sobre os dados para inferir a distribuição que os gerou (WASSERMAN, 2004). A inferência estatística é aplicável em uma ampla gama de áreas, incluindo robótica (LANILLOS et al., 2021), processamento de linguagem natural (FEDER et al., 2022) e computação gráfica (SUN et al., 2019).

O trabalho de Liu et al. (2018) introduz uma abordagem que utiliza redes neurais profundas para estimar matrizes de covariância de medições de sensores em problemas de

estimação de estado. O método de inferência profunda para estimação de covariância (*deep inference for covariance estimation* – DICE), desenvolvido pelos autores, gera estimativas de trajetória mais acuradas a partir de leituras de sensores, resultando em uma aproximação mais precisa ao calcular o estado. Ao estimar as covariâncias diretamente a partir dos dados brutos dos sensores, o método propõe uma forma de melhorar a precisão da inferência de estado em uma variedade de aplicações, desde robótica até redes de sensores.

Considerando o objetivo principal do protocolo de treinamento desenvolvido neste trabalho, que procura gerar dados sintéticos semelhantes aos dados reais, a relevância do artigo citado está na necessidade de modelar e estimar a estrutura presente nos dados de referência, os dados reais, neste caso. A matriz de covariância é uma medida essencial para entender a variabilidade e as relações entre as diferentes variáveis de um conjunto de dados. Ao aprender modelos de ruído gaussiano através do método proposto no artigo, é possível capturar as características complexas do ruído nos dados reais, permitindo com que o objetivo principal deste trabalho seja alcançado.

4 ABORDAGEM PROPOSTA

Este capítulo apresenta os métodos propostos para o problema de anonimização de dados utilizando GANs. A Seção 4.1 descreve, tanto visual quanto textualmente, a estrutura das camadas da arquitetura da rede neural adotada neste trabalho, explicando seus componentes e a razão de sua utilização. A Seção 4.2 introduz as bases de dados utilizadas. A Seção 4.3 examina em detalhes o processo de pré-processamento dos dados reaisizado antes da etapa de treinamento. A Seção 4.4 aborda a execução do treinamento, focando nas funções de perda empregadas e em como foram aplicadas tanto à rede quanto aos dados. Por último, a Seção 4.5 discute os métodos de validação empregados para avaliar os resultados obtidos e identificar possíveis aprimoramentos na rede e no processo de treinamento.

4.1 ARQUITETURA DA REDE

A rede generativa adversária utilizada neste trabalho possui uma arquitetura simplificada, sem o uso de camadas convolucionais. Na realidade, redes convolucionais não podem ser aplicadas em dados tabelados porque estes não possuem vizinhança implícita, passíveis da aplicação do produto de convolução. Optou-se pelo uso de camadas totalmente conectadas nas quais a descoberta das inter-relações entre os diferentes campos é parte do aprendizado.

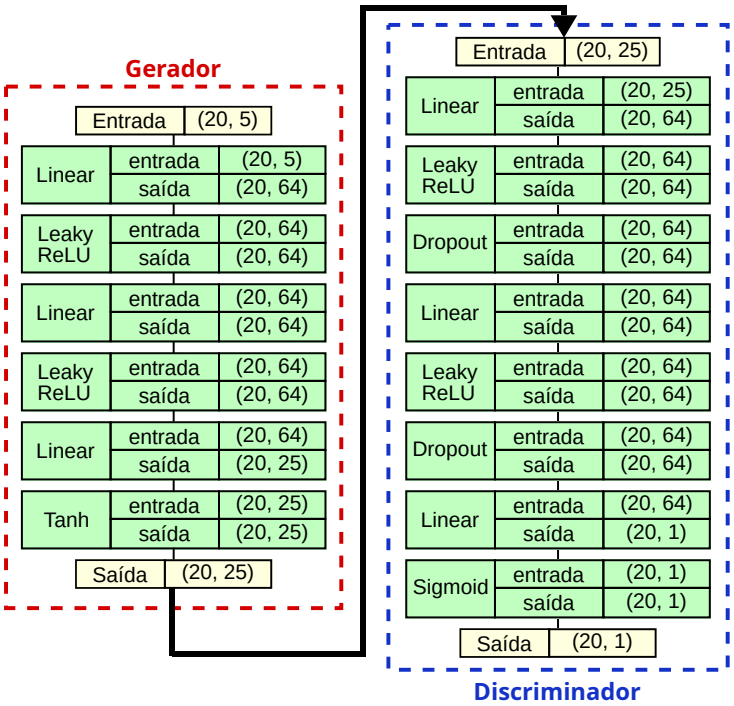
A arquitetura da GAN abrange duas redes neurais distintas: o gerador (*generator*) e o discriminador (*discriminator*). O gerador tem a função de criar amostras sintéticas a partir de um espaço latente ou ruído, enquanto o discriminador tem o objetivo de diferenciar as amostras reais, provenientes do conjunto de dados, das sintéticas, criadas pelo gerador. Ambas as redes incluem camadas densamente conectadas (*fully-connected layers*) e empregam funções de ativação Leaky ReLU (*Leaky Rectified Linear Unit*) para introduzir não-linearidade.

O gerador consiste de uma série de camadas densas, cada uma processando um vetor de entrada amostrado aleatoriamente do espaço latente. Estes vetores são submetidos a transformações lineares seguidas por funções de ativação Leaky ReLU, permitindo ao gerador aprender a converter o espaço latente em um espaço de características mais complexo. A última camada do gerador utiliza a função de ativação tangente hiperbólica (*tanh*) para produzir as amostras sintéticas dentro do intervalo de -1 a $+1$.

Por sua vez, o discriminador analisa amostras tanto do conjunto real quanto do gerador. Sua estrutura também é formada por camadas densas que processam as entradas, aplicando transformações lineares seguidas por funções de ativação Leaky ReLU. A camada de saída do discriminador usa uma função de ativação sigmoide para fornecer a probabilidade estimada de cada amostra ser real ou sintética.

A Figura 5 apresenta a configuração das redes do gerador e do discriminador, mostrando detalhadamente cada camada, juntamente com o tamanho das entradas e saídas.

Figura 5 – Ilustração da arquitetura das redes do gerador e do discriminador.



Fonte: Elaborada pela autora (2024).

4.2 BASE DE DADOS

Dado o contexto do objetivo deste trabalho, que é a anonimização de dados sensíveis, a base de dados sobre doenças da tireoide (QUINLAN, 1987), também utilizada por Piacentino e Angulo (2020), foi escolhida para a realização dos experimentos. Trata-se de uma base que contém informações relacionadas à disfunção da tireoide. O problema consiste em determinar se um paciente tem uma tireoide funcionando normalmente, se tem hipertireoidismo ou se tem hipotireoidismo. A base de dados conta com 7200 casos, sendo 166 (2,3%) de hipertireoidismo e 368 (5,1%) de hipotireoidismo. Os outros 6666 casos (92,6%) pertencem ao grupo com níveis normais de tireoide. Além disso, a base é composta por vinte e um atributos, sendo 15 binários e 6 contínuos, utilizados para determinar em qual das três classes pertence o diagnóstico do paciente.

A Tabela 1 resume informações relevantes sobre esta base, listando todos os seus atributos e indicando se pertencem a domínios de valores binários ou contínuos.

Tabela 1 – Atributos da base de doenças da tireoide e seus domínios de valores estimados pela rede.

Atributo	Domínio	Atributo	Domínio
Age	[1, 97]	Query_hyperthyroid	{0, 1}
Sex	{0, 1}	Lithium	{0, 1}
On_thyroxine	{0, 1}	Goitre	{0, 1}
Query_on_thyroxine	{0, 1}	Tumor	{0, 1}
On_antithyroid_medication	{0, 1}	Hypopituitary	{0, 1}
Sick	{0, 1}	Psych	{0, 1}
Pregnant	{0, 1}	TSH	{0, 1}
Class	{1, 2, 3}	T3	[0,0005, 0,18]
Thyroid_surgery	{0, 1}	TT4	[0,0020, 0,6]
I131_treatment	{0, 1}	T4U	[0,017, 0,233]
Query_hypothyroid	{0, 1}	FTI	[0,0020, 0,642]

Fonte: Elaborada pela autora (2024).

4.3 PRÉ-PROCESSAMENTO

No que tange ao pré-processamento dos dados, inicialmente adotou-se a metodologia empregada por Piacentino e Angulo (2020), com modificações, acréscimos e ajustes finos conforme os resultados preliminares indicaram. Dado o caráter tabular dos dados, optou-se por representar esses dados em pequenas matrizes de tamanho 5×5 para facilitar sua visualização, de maneira similar ao procedimento descrito no artigo citado. Igualmente, a normalização especificada pela Equação (3.1) foi aplicada aos valores de cada atributo para que estes ficassem dentro de um intervalo de 1 a 255, conforme ilustrado na Figura 6, que exibe nove amostras da base de dados. As cores representam a magnitude do valor, variando do azul como valor mínimo, passando pelo verde como valor médio, até a cor amarela como valor máximo.

4.3.1 Remoção de *Outliers*

Para a remoção dos *outliers*, adotou-se o método interquartil. O intervalo interquartil (IQR) é uma medida estatística de dispersão, calculada pela diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), definida como:

$$\text{IQR} = \text{Q3} - \text{Q1}. \tag{4.1}$$

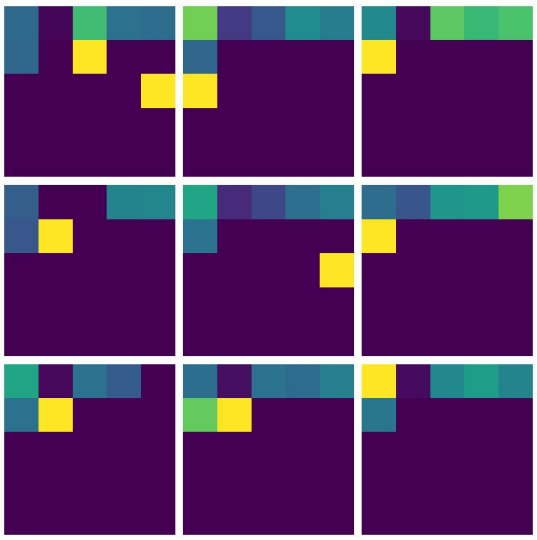
Amostras situadas fora do intervalo:

$$[\text{Q1} - 1.5 \times \text{IQR}, \text{Q3} + 1.5 \times \text{IQR}] \tag{4.2}$$

são consideradas *outliers* e, portanto, são excluídas.

A Figura 7 ilustra os *boxplots* dos atributos contínuos antes e após a remoção dos *outliers*, evidenciando as distribuições. Observa-se que alguns atributos, particularmente o

Figura 6 – Conjunto de 9 amostras aleatórias tiradas da base de dados. As cores representam a magnitude do valor, variando do azul como valor mínimo, passando pelo verde como valor médio, até a cor amarela como valor máximo.



Fonte: Elaborada pela autora (2024).

TSH, apresentam amostras que excedem significativamente o intervalo interquartil. Ao considerar essa característica dos dados, a faixa para exclusão dos *outliers* foi estabelecida de modo a preservar o máximo de amostras possível. Portanto, o intervalo selecionado foi o de $[30, 95]$.

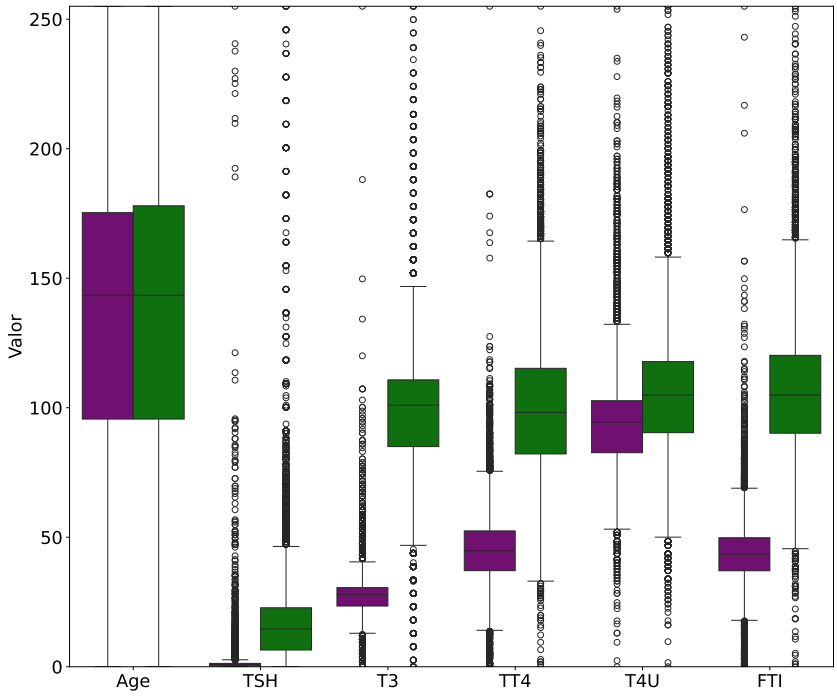
É possível observar através da Figura 7 que após a remoção dos *outliers* os atributos T3, TT4, T4U e FTI passaram a ter valores mais homogêneos entre si, eliminando valores esparsos observados principalmente nos atributos T3, TT4 e FTI. Além disso, o atributo TSH, que possuía uma distância grande entre os valores de algumas amostras, preservou a característica de sua distribuição, com amostras com valores distantes da mediana, mas sem um salto grande entre eles, como havia antes da remoção dos *outliers*.

4.4 TREINAMENTO

O treinamento da rede foi realizado utilizando a função de perda BCE em combinação com uma função de perda customizada baseada na divergência KL. Essa combinação teve como objetivo melhorar tanto o desempenho quanto a qualidade dos resultados produzidos pela rede.

A BCE é frequentemente empregada na rede discriminadora, sendo particularmente adequada para problemas de classificação binária, já que o discriminador tem como função diferenciar entre amostras reais e sintéticas. Durante o treinamento do gerador, a BCE é utilizada nas saídas do discriminador em relação aos rótulos reais, visando maximizar a probabilidade de que o discriminador classifique incorretamente as amostras sintéticas

Figura 7 – *Boxplot* de cada atributo contínuo com (roxo) e sem *outliers* (verde).



Fonte: Elaborada pela autora (2024).

como sendo reais. Esta estratégia incentiva o gerador a criar amostras cada vez mais convincentes. Adicionalmente, no treinamento do gerador, a BCE é empregada em conjunto com uma função de perda customizada que incorpora a divergência KL para medir a discrepância entre a distribuição das amostras reais fornecidas ao gerador e das por ele geradas. O objetivo é minimizar essa discrepância, permitindo que o gerador produza amostras cada vez mais próximas da distribuição estatística dos dados reais.

4.4.1 Inferência Estatística no Treinamento

Liu et al. (2018) discutem a estimativa de covariância em problemas de inferência de estado, com o objetivo de desenvolver um método baseado em aprendizado profundo para aprender modelos de ruído gaussiano que representem de maneira fidedigna a estrutura de covariância em um sistema.

A covariância é uma medida estatística que indica a relação entre variáveis distintas em um sistema. A estimativa precisa da covariância é crucial para diversas tarefas de inferência de estado, como rastreamento de objetos, filtro de Kalman e controle adaptativo. No entanto, a modelagem eficaz da covariância pode ser um desafio, dada a complexidade das interações entre variáveis e a presença de ruído.

O método introduzido, denominado DICE, emprega redes neurais profundas para aprender modelos de ruído gaussiano que captam as nuances da covariância em um

sistema. Essa abordagem envolve treinar uma rede neural para aprender a função de mapeamento entre as entradas do sistema e os parâmetros da distribuição gaussiana que melhor representa o ruído observado nos dados.

A rede neural é treinada utilizando um conjunto de dados composto por exemplos de entradas do sistema e suas covariâncias correspondentes. Durante o processo de treinamento, a rede é otimizada para minimizar a diferença entre as covariâncias estimadas pela rede e as covariâncias reais.

Os experimentos conduzidos pelos autores demonstraram que o método proposto consegue estimar covariâncias com precisão em diversos cenários, superando métodos convencionais em termos de acurácia e robustez, particularmente em situações envolvendo não-linearidades, ruídos e distorções.

No contexto deste trabalho, o artigo oferece uma perspectiva relevante sobre a implementação de uma rede que considera uma covariância fornecida *a priori*. Neste caso, a função de perda seria projetada para produzir uma matriz de covariância estimando a diferença entre os dados reais e os dados sintéticos, visando uma aproximação máxima dos últimos em relação aos primeiros. Inicialmente, este desafio é um problema de maximização, conforme expresso por:

$$\arg\max \sum_{i=1}^{n} p(e_i \mid R_i), \tag{4.3}$$

que pode ser reformulado como um problema de minimização da forma:

$$\arg\min \sum_{i=1}^{n} -\log(p(e_i \mid R_i)). \tag{4.4}$$

4.5 VALIDAÇÃO

Borji (2022) examina os prós e contras de várias métricas de avaliação e os resultados alcançados por GANs. Devido à complexidade da tarefa, muitas dessas métricas devem ser aplicadas conjuntamente para que os resultados possam ser adequadamente interpretados, permitindo conclusões baseadas em múltiplos fatores. Com isso em mente e levando em consideração algumas dessas métricas, a avaliação dos resultados deste estudo é realizada sob duas perspectivas comumente adotadas: uma qualitativa e outra quantitativa.

Sob a perspectiva qualitativa, busca-se avaliar os resultados com base nas amostras geradas e em gráficos que permitam uma comparação direta e visual entre os dados sintéticos e os reais. Serão apresentadas amostras produzidas ao longo das épocas, gráficos de perda por época para os modelos gerador e discriminador, visando avaliar a convergência do modelo, e *boxplots* dos valores dos atributos contínuos.

A comparação dos atributos por meio de histogramas facilita a visualização da distribuição e similaridade entre os dados sintéticos e os reais, além de permitir a identificação de diferenças nas formas das distribuições e verificar se o modelo conseguiu reproduzir

corretamente picos, assimetrias, tendências e outros padrões dos dados reais. Os dados sintéticos tendem a seguir a distribuição dos dados reais, mantendo seus valores dentro de uma faixa mais restrita.

Quanto à perspectiva quantitativa, optou-se pela utilização da distância de início de *Fréchet* (*Fréchet inception distance* – FID), uma métrica amplamente usada para estimar a qualidade dos resultados de GANs (HEUSEL et al., 2017). A FID é dada por:

$$\text{FID} = \sqrt{|\mu_P - \mu_Q|^2 + \mu_P^T \mu_P - 2(\mu_P^T \mu_Q) + \mu_Q^T \mu_Q + \sigma_P^2 + \sigma_Q^2 - 2\sqrt{\sigma_P^2 \sigma_Q^2}}, \qquad (4.5)$$

onde μ_P e μ_Q são as médias das distribuições dos dados reais e sintéticos, respectivamente, e σ_P e σ_Q são as variâncias das distribuições dos dados reais e sintéticos, respectivamente.

Além da FID, uma tabela é elaborada contendo informações de média, desvio padrão e mediana para comparar os dados reais e sintéticos gerados pelos modelos treinados com os diferentes parâmetros adotados neste trabalho.

5 RESULTADOS EXPERIMENTAIS

Este capítulo apresenta os resultados experimentais de maneira detalhada junto à validação e ao desenvolvimento de novos métodos com base nas conclusões obtidas a cada iteração experimental, isto é, a cada etapa de experimentos. A Seção 5.1 detalha os resultados alcançados após o treinamento de um modelo *baseline* e a Seção 5.2 explora um método de treinamento que utiliza o tamanho da janela e o passo para introduzir uma nova função de perda, além de abordar um treinamento que incorpora um parâmetro adicional e uma nova modificação na função de perda.

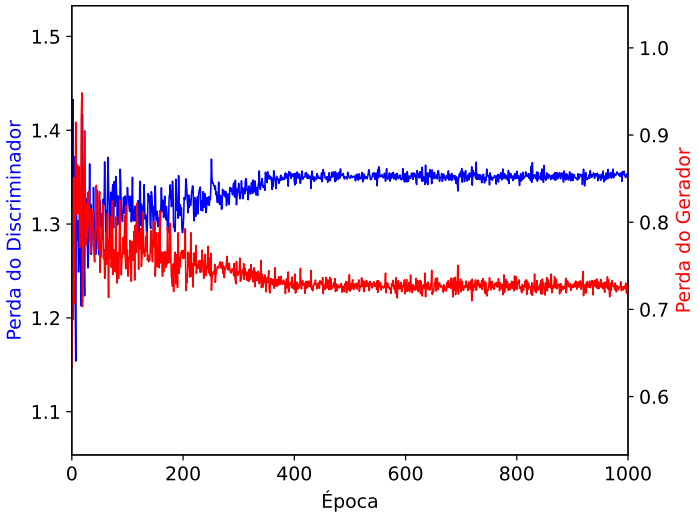
5.1 MODELO *BASELINE*

No início do treinamento para uma tarefa de aprendizado de máquina, é essencial estabelecer um modelo *baseline* que atue como referência durante o desenvolvimento subsequente. Esta fase inicial fornece entendimentos sobre a capacidade do modelo de gerar amostras convincentes e destaca áreas para identificação e correção de falhas, bem como oportunidades para aprimoramentos em iterações futuras. Para o modelo *baseline*, adotou-se um método de treinamento convencional com a arquitetura de GAN descrita no capítulo anterior, limitando-se ao uso exclusivo da entropia cruzada binária como função de perda, sem incorporar funções adicionais. O treinamento foi programado para 1000 épocas. Os dados foram formatados como representações visuais de tamanho 5×5, com cada célula representando um atributo da base de dados de doenças da tireoide. Dado que esta base contém 21 atributos, quatro células em cada uma das representações são vazias, apresentando valores nulos, conforme ilustrado na Figura 6.

A Figura 8 mostra o gráfico de perda por época para o modelo *baseline*. Nota-se que, a partir da época 400, as perdas do gerador e do discriminador começam a se equilibrar, indicando uma convergência e estabilidade do modelo no processo de otimização. Esse comportamento é o esperado para redes do tipo GAN, onde inicialmente há uma alta variabilidade na perda, visto que o gerador e o discriminador estão em uma fase de adversidade e aprendizado. No final, há o comportamento de equilíbrio observado, sugerindo que a GAN está gerando amostras de alta qualidade e o discriminador não consegue mais distinguir entre elas de forma significativa.

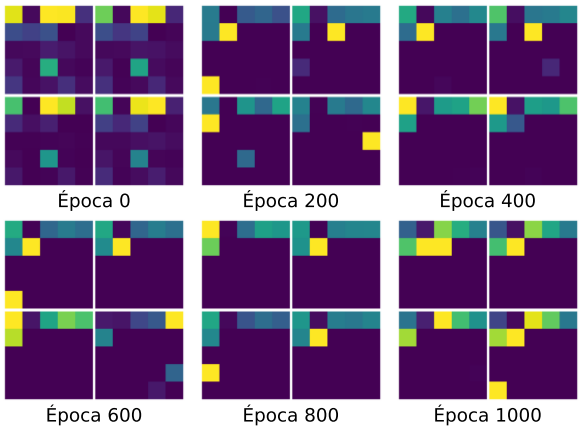
A cada época, a rede geradora produz novas amostras, denominadas amostras sintéticas. A Figura 9 exibe conjuntos de 4 dessas, cada um gerado em diferentes etapas do treinamento, partindo da primeira, em que as representações geradas são praticamente arbitrárias, até a última época, revelando uma organização mais refinadas e próximas das representações dos dados reais. É possível notar a progressão dessas representações até que o modelo alcance um padrão de valores e organização que se assemelha cada vez mais às amostras reais.

Figura 8 – Perda do gerador (vermelho) e discriminador (azul) do modelo *baseline*.



Fonte: Elaborada pela autora (2024).

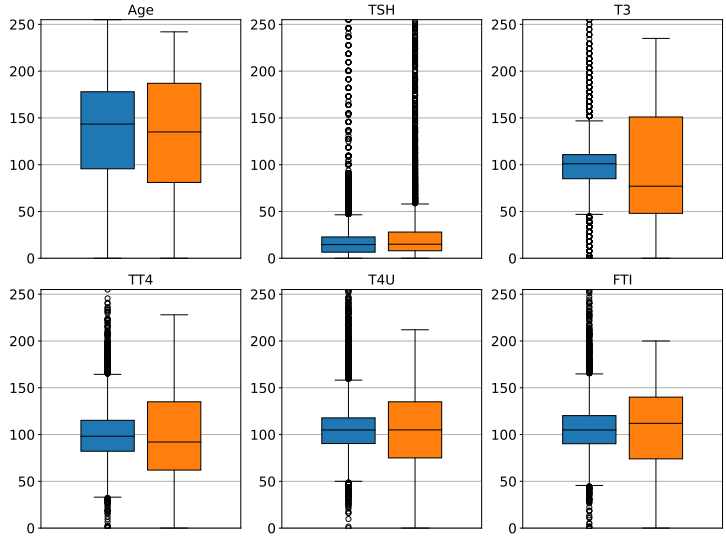
Figura 9 – Amostras geradas pelo modelo *baseline*.



Fonte: Elaborada pela autora (2024).

Além do gráfico de perda por época e das amostras geradas durante o treinamento, um aspecto crucial na avaliação dos resultados é a comparação da distribuição do conjunto de amostras sintéticas com o conjunto de amostras reais. A Figura 10 mostra os *boxplots* de cada atributo contínuo das bases de dados reais e sintéticas, facilitando a comparação entre ambos. É perceptível que as distribuições dos atributos T3, TT4, T4U e FTI mostram diferenças significativas entre as duas bases. Nas iterações subsequentes, explorou-se formas de refinamento do treinamento por meio de ajustes nos hiperparâmetros e através do desenvolvimento de um protocolo de treinamento visando gerar dados cujas distribuições dos atributos fossem mais semelhantes às dos dados reais, conforme será discutido nas próximas seções.

Figura 10 – *Boxplots* com os atributos contínuos do conjunto de dados reais (azul) e sintéticos (laranja) gerados pelo modelo *baseline*.



Fonte: Elaborada pela autora (2024).

Considerando os métodos de avaliação dos modelos apresentados anteriormente, a saber, a avaliação qualitativa por meio do progresso das amostras durante o treinamento e a perda de cada rede ao longo das épocas, pode-se concluir que o modelo *baseline* alcança a convergência em um número reduzido de épocas de treinamento, e as amostras sintéticas se aproximam das reais sob a perspectiva qualitativa. Contudo, ao examinar os *boxplots* dos atributos contínuos, observa-se que algumas distribuições dos atributos nos dados sintéticos divergem significativamente quando comparadas com as distribuições dos mesmos atributos nos dados reais. A melhoria almejada nos métodos subsequentes foca em alcançar distribuições mais similares para esses atributos.

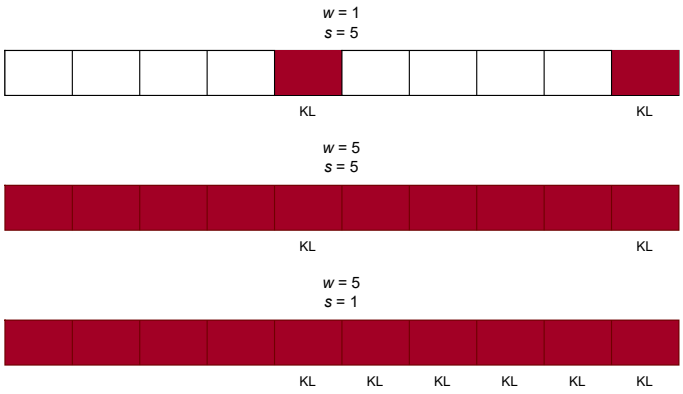
5.2 PARÂMETROS DE JANELA E PASSO

Na implementação do treinamento com a divergência KL, desenvolveu-se uma estratégia para determinar o melhor intervalo entre cada aplicação dessa função e o número ideal de amostras a ser considerado na sua aplicação. O custo computacional elevado exigido por essa função impôs a necessidade de planejar um treinamento mais rápido e eficiente. Duas variáveis foram essenciais nesse contexto: o passo (*stride*) s, que indica o número de épocas entre cada aplicação da função de perda customizada; e a janela (*window*) w, que define quantas épocas anteriores contribuem com a geração dos dados sintéticos para o cálculo da divergência KL. Um critério inicial importante foi que a aplicação da função customizada só deveria iniciar a partir da w-ésima época. A partir dessa premissa, identificaram-se três abordagens distintas para o treinamento.

A Figura 11 ilustra o treinamento utilizando diferentes configurações de tamanho de janela e passo. No caso de uma janela de tamanho 1 e um passo de tamanho 5, a

divergência KL irá calcular a diferença entre os dados reais e os dados sintéticos gerados pelo modelo até a época imediatamente anterior. O passo de tamanho 5 significa que a função de perda contendo o cálculo da divergência KL será aplicada a cada 5 épocas. A figura também demonstra outros dois métodos de treinamento: com janela de tamanho 5 e passo de tamanho 5, e com janela de tamanho 5 e passo de tamanho 1. Ambas as configurações apresentam uma janela de tamanho 5, indicando que a divergência KL avaliará a diferença entre os dados reais e os dados gerados pelos modelos das últimas 5 épocas. Portanto, nesses cenários, uma quantidade maior de dados é considerada, além de potencialmente incluir modelos menos precisos por serem "mais antigos". É importante observar que o treinamento com uma janela de tamanho 5 e um passo de tamanho 5 aplicará a função com divergência KL a cada 5 épocas, enquanto a configuração com um passo de tamanho 1 a utilizará em todas as épocas.

Figura 11 – Exemplo de treinamento com $w = 5$.



Fonte: Elaborada pela autora (2024).

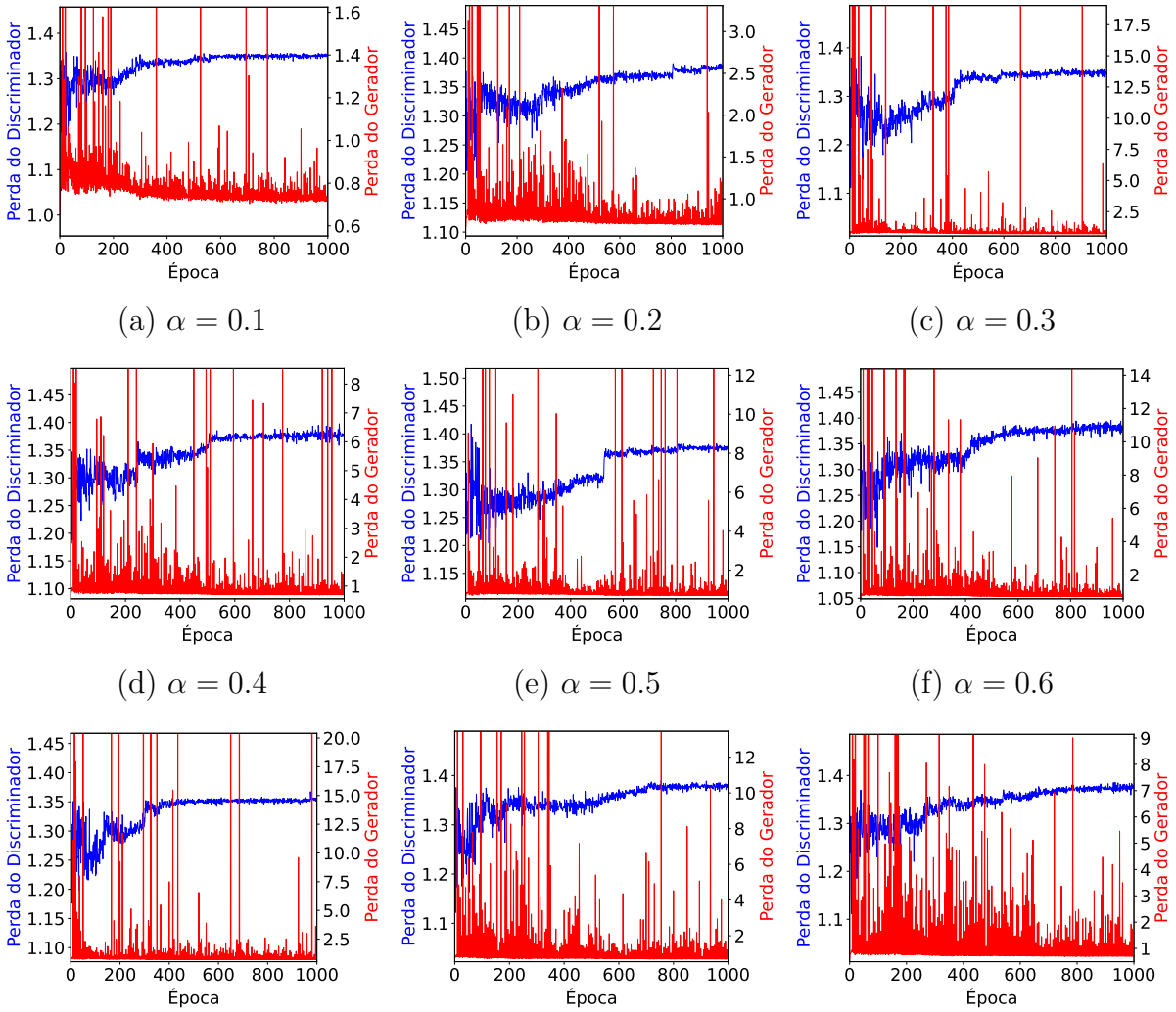
A função de perda customizada para esta configuração de treinamento é composta pela entropia cruzada binária somada à divergência KL, expressa por:

$$\text{Loss} = \text{BCE} + \alpha \cdot D_{KL}(p||q), \tag{5.1}$$

onde α é uma constante que varia entre 0 e 1. Durante o treinamento, foram testados valores de α de 0,1 a 0,9 para examinar o impacto e a evolução da perda com a atenuação da contribuição da divergência KL ao longo das épocas.

A Figura 12 mostra a perda por época do modelo treinado com janela de tamanho 1 e passo de tamanho 5 para cada valor de α. Os gráficos evidenciam a característica instável da perda do gerador para essa configuração de janela e passo, que atinge valores muito altos se comparado com a perda do discriminador. Nota-se também que quanto maior o valor de α, ou seja, quanto maior a contribuição da divergência KL, mais instável a perda do gerador. Nesse caso com a janela de tamanho 1, a divergência KL é aplicada entre a distribuição dos dados reais e a distribuição dos dados gerados na mesma época em que é aplicada. O passo, por sua vez, indica que essa aplicação da divergência KL ocorre a cada 5 épocas.

Figura 12 – Perda do gerador (vermelho) e discriminador (azul) do modelo com $w = 1$ e $s = 5$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

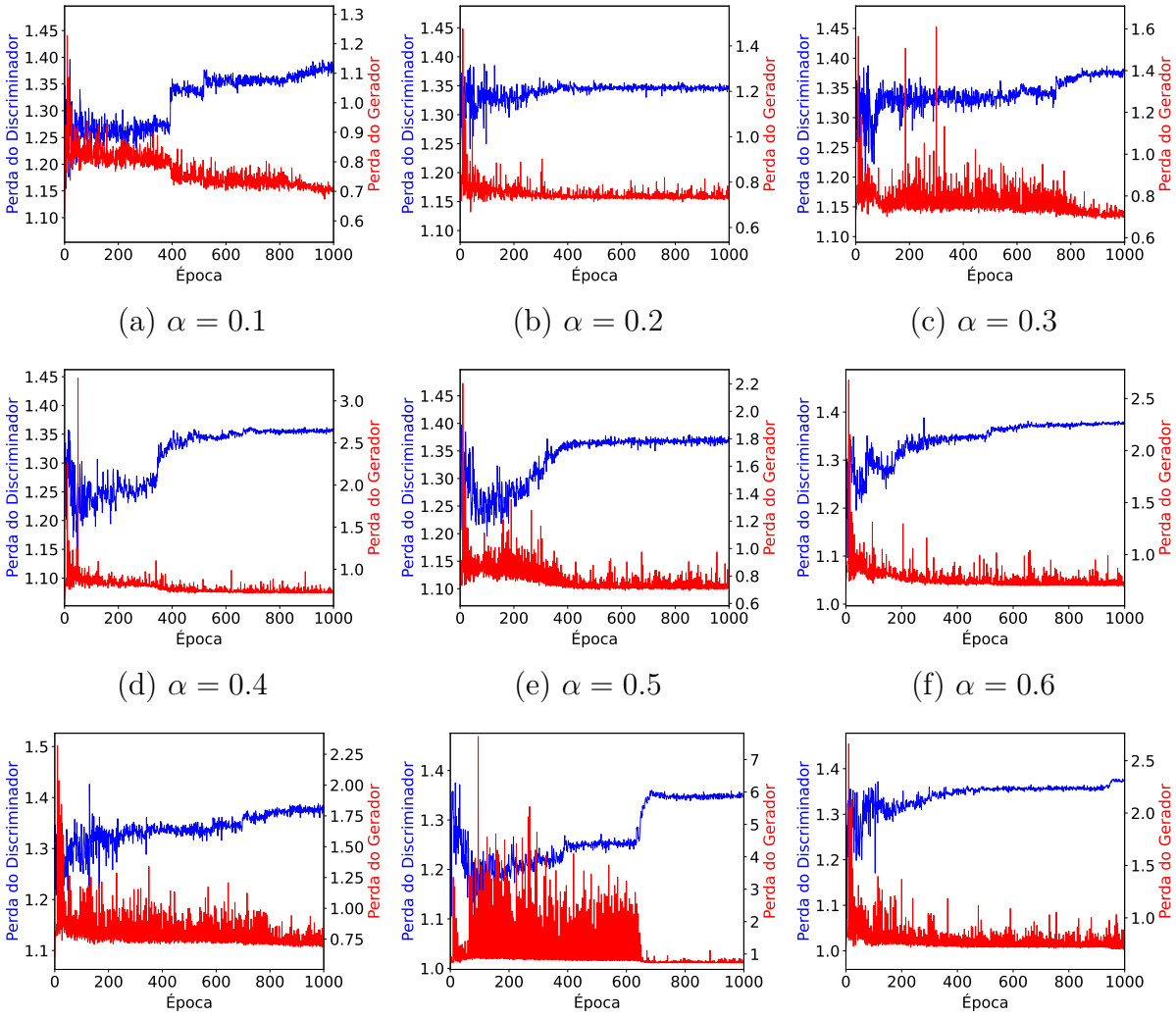
(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

Nota-se, também, que enquanto a perda do discriminador se comporta de maneira menos ruidosa, a do gerador, por sua vez, parece ser mais caótica. Valores maiores e menores de perda do gerador fazem com que a curva oscile em uma amplitude maior se comparada a do discriminador. Esse ruído observado pode ser devido ao tamanho da janela, que só considera as amostras geradas na época anterior para calcular a matriz de covariância dos dados gerados. Esse conjunto de dados, além de ser pequeno, muda a cada aplicação do KL, o que pode fazer com que as matrizes de covariância de cada conjunto sejam bem diferentes entre si, principalmente no começo do treinamento, quando o gerador ainda está aprendendo a representação dos dados. Devido a isso, os outros dois experimentos realizados fixam o tamanho de w em 5, visando uma maior estabilidade na perda do gerador além de proporcionar um maior número de amostras para se gerar a matriz de covariância.

As Figuras 13 e 14 mostram a perda por época dos modelos com $w = 5$ e $s = 5$ e $w = 5$ e $s = 1$, respectivamente, para cada valor de α. Nesses casos, a divergência KL é aplicada comparando a distribuição dos dados reais com a distribuição do conjunto de dados gerados nas 5 últimas épocas. Se comparado com os modelos treinados com $w = 1$ e $s = 5$, os modelos com $w = 5$ apresentam uma estabilidade maior, gerando poucos ou nenhum valor muito diferente, como era de se esperar ao aumentar o tamanho de w. Destaca-se que para $w = 5$ e $s = 5$, principalmente com valores de *alpha* iguais a $0,3$, $0,7$ e $0,8$, as perdas do gerador exibem uma oscilação de valores maior em comparação aos outros valores de *alpha*, mas ainda bem menos oscilantes se comparados aos gráficos do treinamento com $w = 1$.

Figura 13 – Perda do gerador (vermelho) e discriminador (azul) do modelo com $w = 5$ e $s = 5$ para cada valor de α.



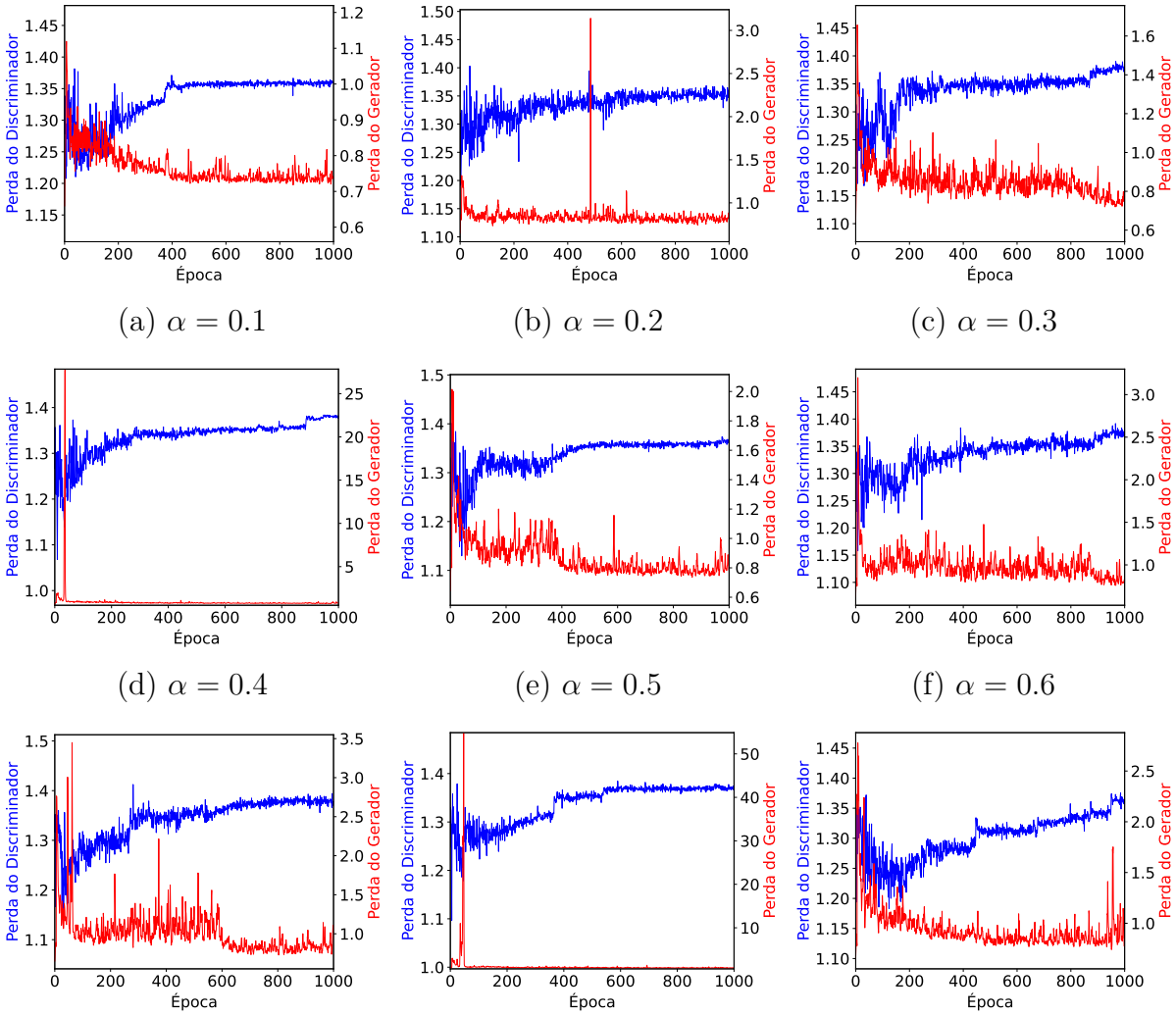
(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

Figura 14 – Perda do gerador (vermelho) e discriminador (azul) do modelo com $w = 5$ e $s = 1$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

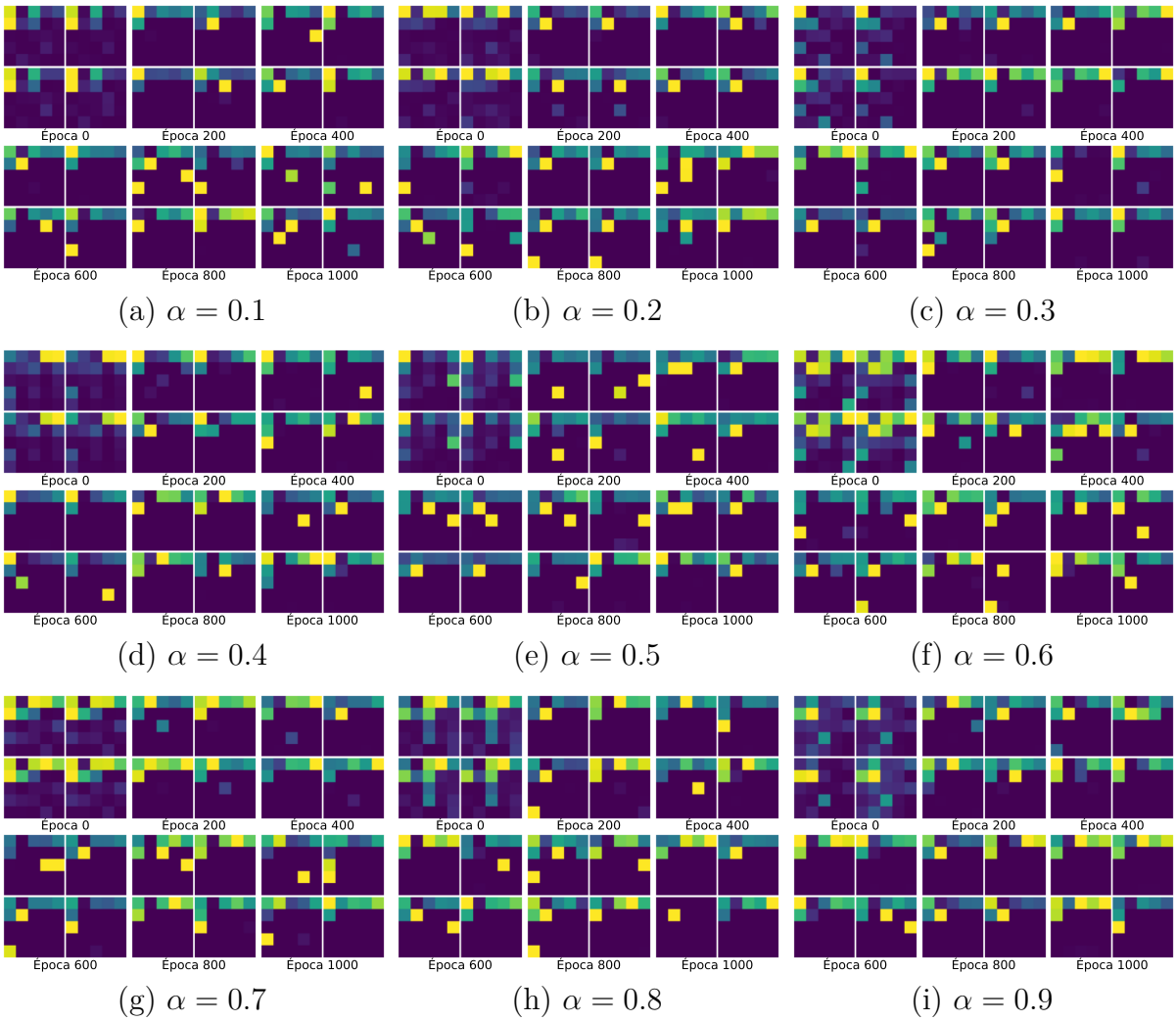
(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

Os gráficos de perda para o modelo com $w = 5$ e $w = 1$, por sua vez, apresentam maior estabilidade na curva da perda do gerador. Com $s = 1$, a divergência KL é aplicada a todas as épocas a partir da quinta, pegando sempre as amostras geradas nas cinco últimas épocas. Dessa maneira, a cada época, apenas uma parte dos dados é nova, a que se refere às amostras geradas na época em questão, enquanto as outras quatro são comuns à aplicação anterior. Devido a essa baixa variabilidade, é de se esperar que a função de perda do gerador não apresente valores muito discrepantes entre si a cada época, mantendo uma maior estabilidade durante o treinamento. Isso também é evidenciado observando que em todos os valores de *alpha* a curva de perda se comporta dessa maneira.

As Figuras 15, 16 e 17 ilustram algumas das amostras geradas ao longo das épocas para cada valor de α nos modelos com $w = 1$ e $s = 5$, $w = 5$ e $s = 5$ e $w = 5$ e $s = 1$, respectivamente.

Figura 15 – Amostras geradas pelo modelo com $w = 1$ e $s = 5$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

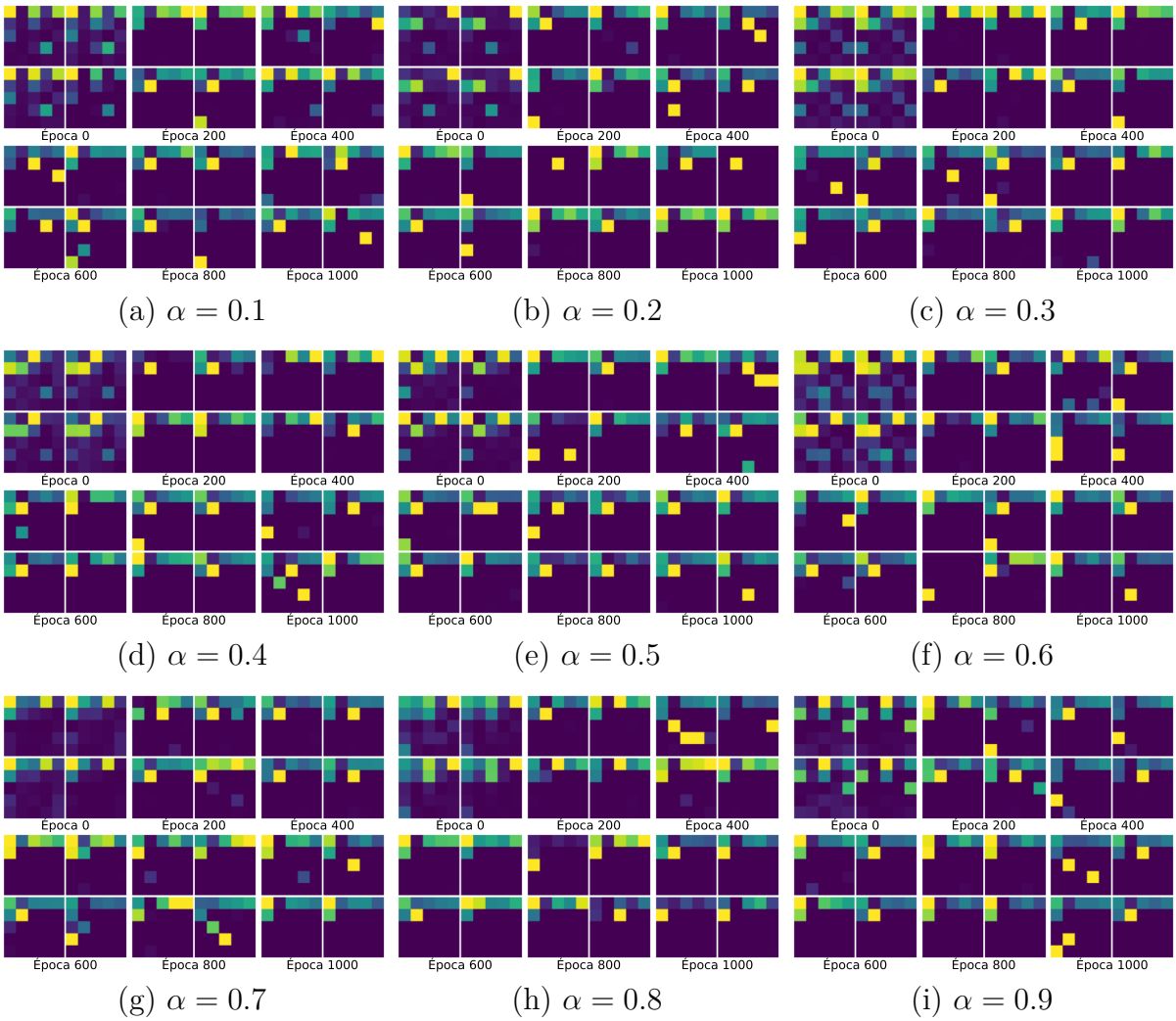
(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

As Figuras 18, 19 e 20 comparam as distribuições dos atributos contínuos dos dados da base de dados reais e dos dados sintéticos para cada valor de α nos modelos com $w = 1$ e $s = 5$, $w = 5$ e $s = 5$ e $w = 5$ e $s = 1$, respectivamente. É possível perceber pelas figuras que os atributos dos dados sintéticos, em geral, seguem uma distribuição próxima daqueles dos dados reais, incluindo os atributos que divergiam dos dados reais no modelo *baseline*.

Mesmo com oscilações maiores na curva de perda do gerador do modelo com $w = 1$ e $s = 5$, observa-se que a distribuição dos dados gerados evidencia um resultado tão bom quanto para os outros dois casos. A aplicação da divergência KL em cada configuração de treinamento tem como resultado amostras sintéticas com distribuição mais próxima das amostras dos dados reais, como pode ser observado nas figuras. Esse resultado evidencia a contribuição desse erro, além de demonstrar o efeito que cada tamanho de w e s causa no treinamento, o que também demonstra a flexibilidade desses parâmetros, com certas configurações apresentando comportamentos na perda da rede bem distintos em comparação com outras.

Figura 16 – Amostras geradas pelo modelo com $w = 5$ e $s = 5$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

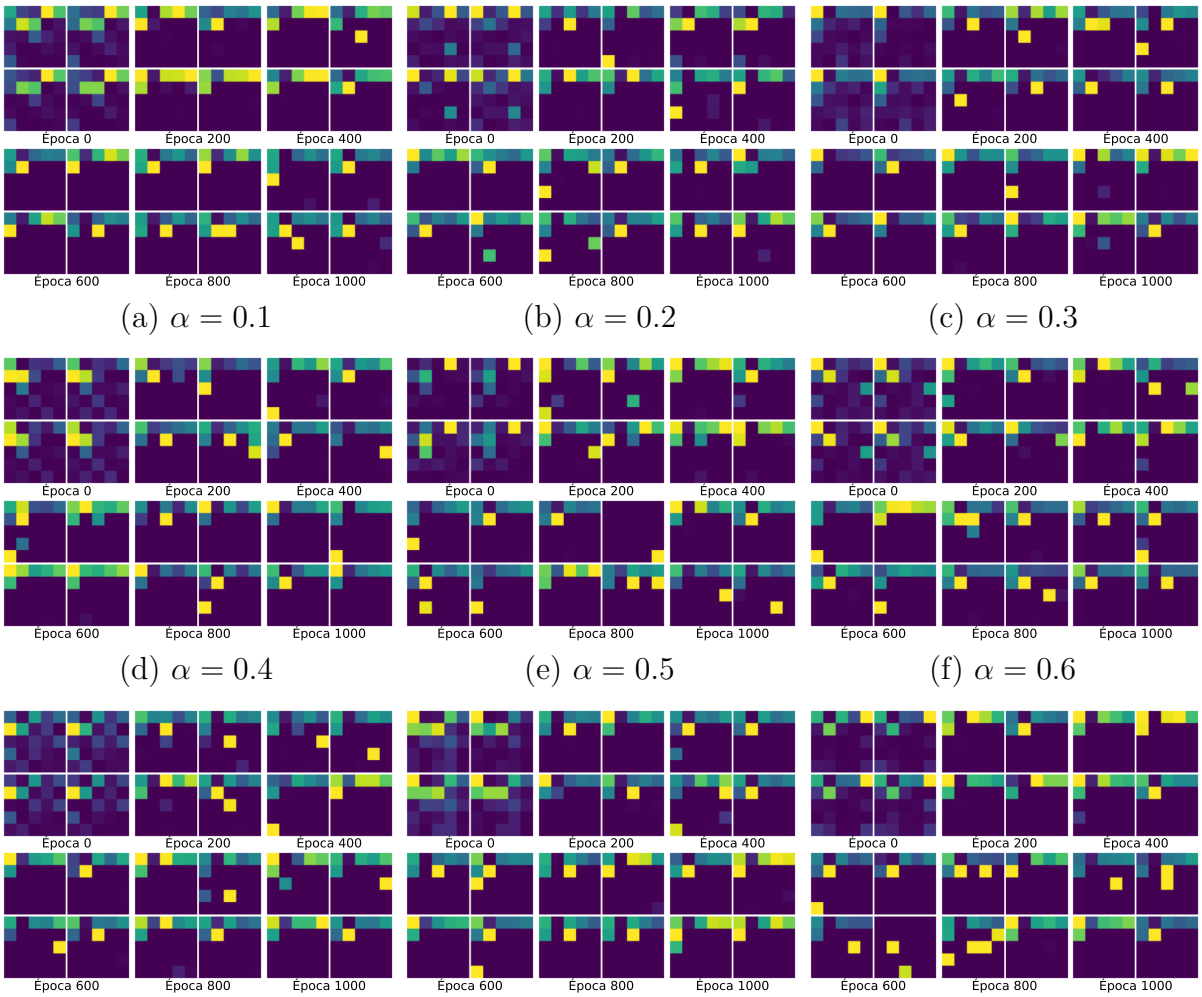
(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

A Tabela 2 apresenta a média, desvio padrão e mediana das perdas do gerador e do discriminador para comparação de cada uma das configurações de janela e passo. O menor valor dentre cada medida estatística está em destaque.

Para $alpha = 0,1$ os valores de média e desvio padrão apresentam uma discrepância acentuada se comparados a outros valores de *alpha*. Isso se evidencia ao se comparar a tabela com a Figura 12, que mostra a perda por época da configuração $w = 5$ e $s = 1$ para cada valor de *alpha*. Como discutido anteriormente, para essa configuração de w e s, a perda do gerador oscila bastante. Especificamente em $alpha = 0,1$, os valores de perda do gerador alcançam faixas muito maiores se comparados aos outros valores de *alpha*. Os experimentos foram realizados mais vezes, gerando o mesmo comportamento para esse valor específico de *alpha*.

Figura 17 – Amostras geradas pelo modelo com $w = 5$ e $s = 1$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

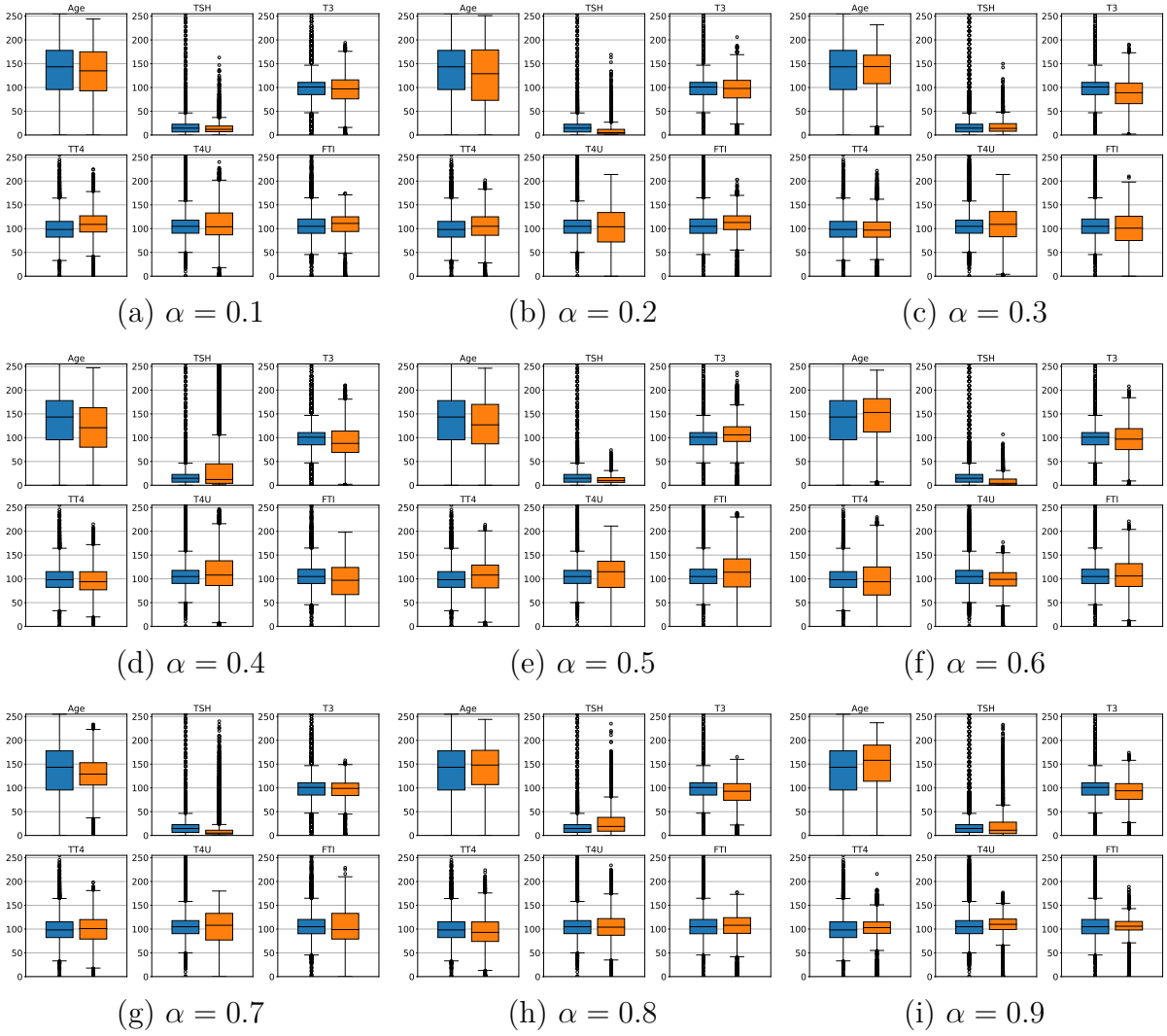
Fonte: Elaborada pela autora (2024).

Tabela 2 – Média, desvio padrão e mediana das perdas do discriminador e do gerador para cada valor de α.

	loss \ α	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
	Média	**1,287**	1,342	1,368	1,378	1,380	1,380	1,381	1,381	1,381
D	Desvio padrão	0,093	0,010	0,006	0,004	0,003	0,002	0,002	**0,002**	0,002
	Mediana	**1,230**	1,345	1,369	1,377	1,381	1,380	1,381	1,381	1,381
	Média	211,240	0,773	0,738	0,740	0,736	**0,732**	0,736	0,746	0,743
G	Desvio padrão	5504,878	0,070	**0,067**	0,086	0,095	0,083	0,095	0,148	0,101
	Mediana	0,802	0,749	0,719	0,707	0,702	0,703	0,701	**0,701**	0,702

Fonte: Elaborada pela autora (2024).

Figura 18 – *Boxplots* com atributos contínuos do conjunto de dados reais (azul) e sintéticos (laranja) gerados pelo modelo com $w = 1$ e $s = 5$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

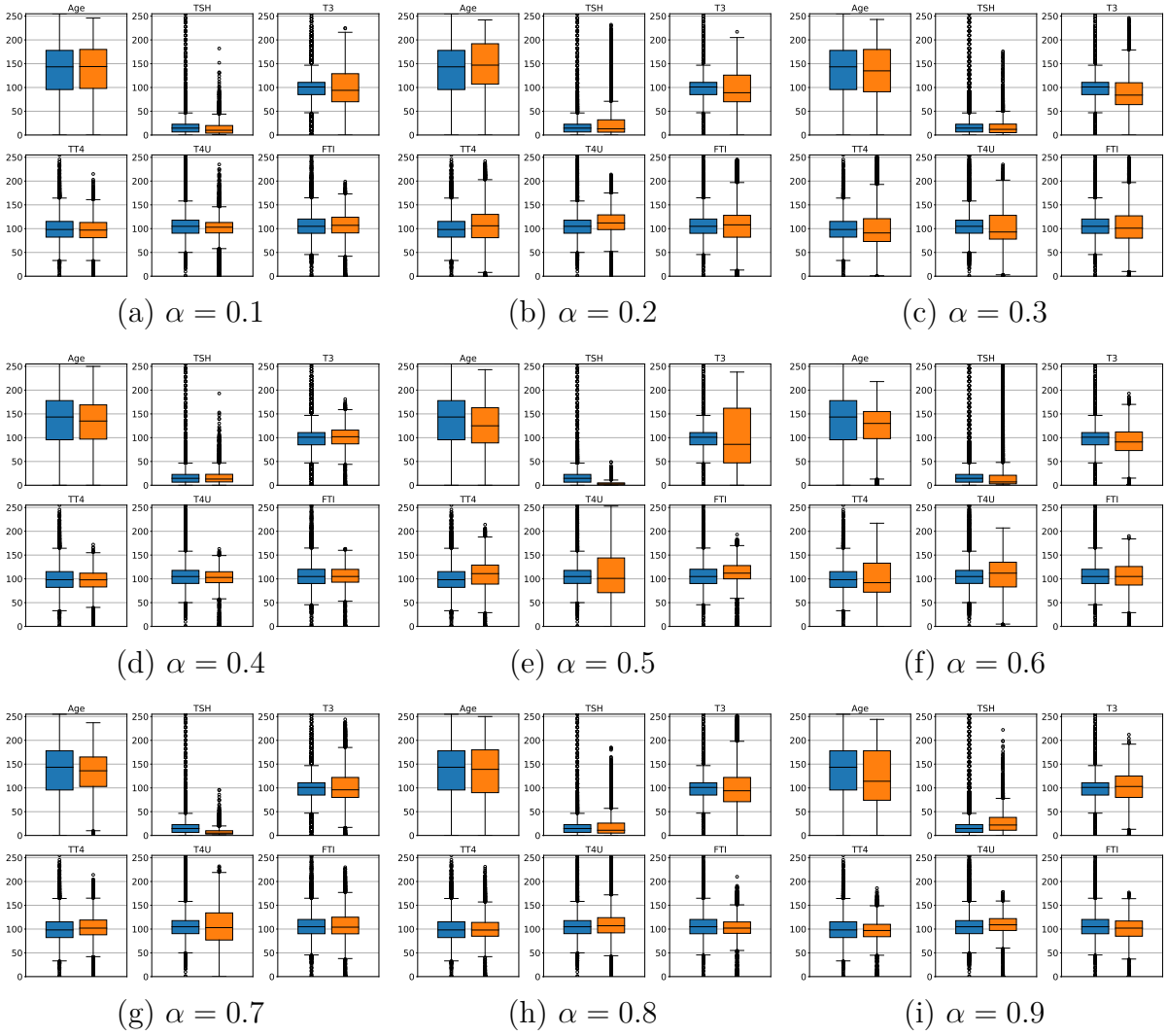
5.2.1 Investigação com MSE

Com o intuito de explorar mais as capacidades de uma função de perda com múltiplos erros, uma terceira função foi adicionada ao treinamento. Essa função é a já apresentada MSE e, assim como a divergência KL, foi acoplada à perda do gerador na forma:

$$\text{Loss} = \text{BCE} + \alpha \cdot D_{KL}(p||q) + \beta \cdot \text{MSE}. \tag{5.2}$$

Para esse experimento, optou-se por usar um vetor de 1×21 como entrada, diferente da matriz 5×5 utilizada anteriormente. A divergência KL foi calculada conforme a Equação 2.4, apresentada anteriormente, o que faz com que valores negativos possam ser gerados para o caso em que a distribuição dos dados sintéticos esteja em um intervalo inferior a dos dados reais. Por uma questão de coerência com as perdas anteriores, os valores absolutos foram tomados e utilizados para calcular a média, mediana e desvio padrão.

Figura 19 – *Boxplots* com atributos contínuos do conjunto de dados reais (azul) e sintéticos (laranja) gerados pelo modelo com $w = 5$ e $s = 5$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

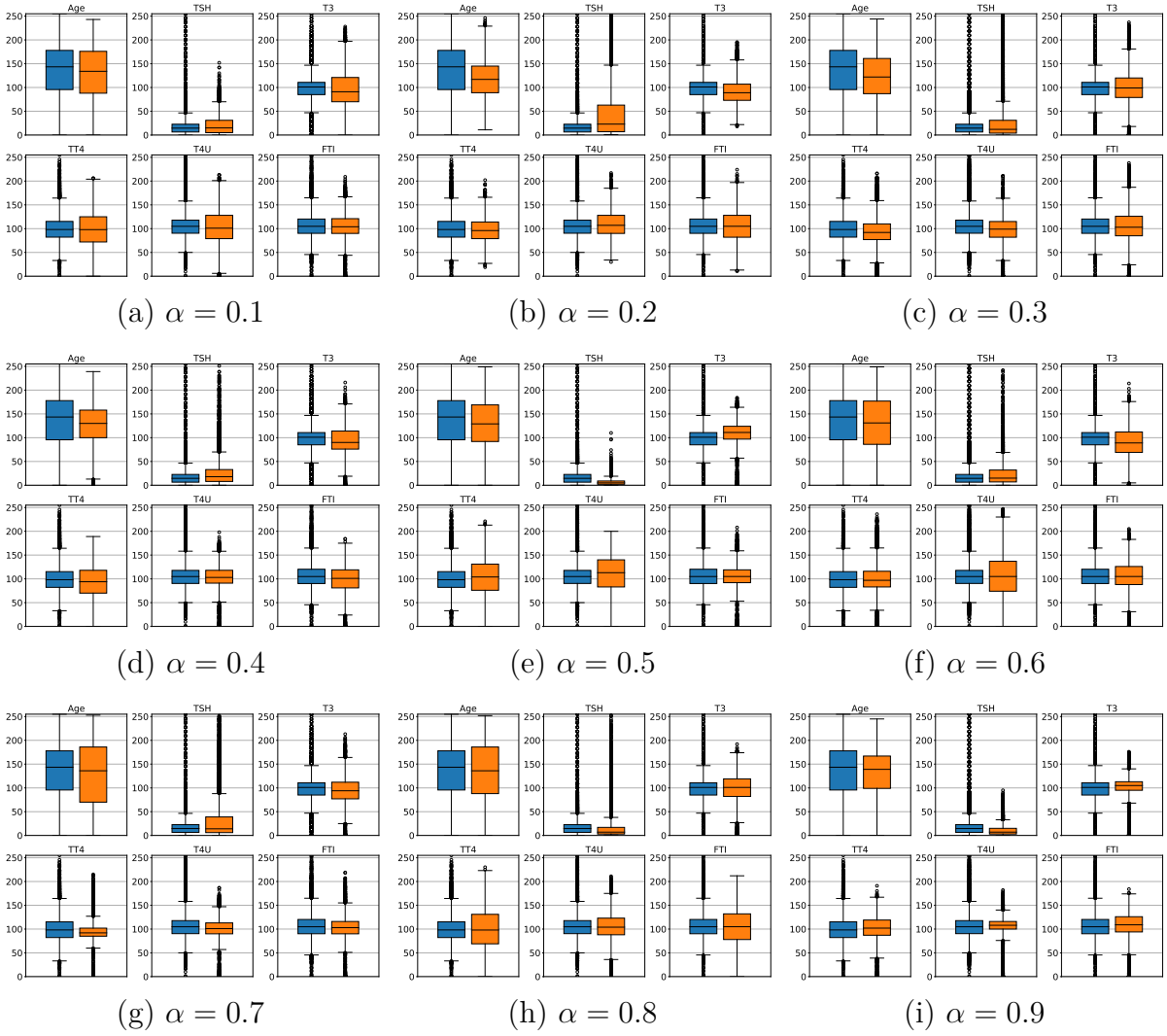
(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

As Tabelas 3 e 4 apresentam a média, desvio padrão e mediana das perdas do discriminador e do gerador para cada valor de α e β. Adicionalmente, a Tabela 5 mostra a distância euclidiana entre as perdas do discriminador e do gerador.

É possível observar pelas tabelas que os menores valores de média e mediana são do modelo com α e β iguais a 0. Quando se trata do desvio padrão, o menor valor está em $\alpha = 0,6$ e $\beta = 0$. Pelo que foi observado durante o treinamento dos modelos e pelas tabelas resultantes, a função de perda com MSE ponderado pelo parâmetro β não apresenta melhoria para ser incluída no modelo final. Por essa razão, resolveu-se excluir a função de perda MSE da função de perda.

Figura 20 – *Boxplots* com atributos contínuos do conjunto de dados reais (azul) e sintéticos (laranja) gerados pelo modelo com $w = 5$ e $s = 1$ para cada valor de α.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Fonte: Elaborada pela autora (2024).

Tabela 3 – Perda do discriminador (média) para cada combinação de α e β.

β \ α	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	**1,3109**	1,3837	1,3839	1,3845	1,3844	1,3845	1,3840	1,3838	1,3842	1,3842
1	1,3708	1,3838	1,3840	1,3844	1,3845	1,3845	1,3833	1,3839	1,3841	1,3844
2	1,3820	1,3838	1,3839	1,3846	1,3844	1,3845	1,3828	1,3839	1,3841	1,3843
3	1,3834	1,3840	1,3838	1,3845	1,3843	1,3846	1,3831	1,3837	1,3843	1,3842
4	1,3836	1,3840	1,3840	1,3847	1,3843	1,3221	1,3837	1,3840	1,3843	1,3844
5	1,3839	1,3841	1,3841	1,3847	1,3842	1,3685	1,3838	1,3839	1,3843	1,3843
6	1,3835	1,3840	1,3841	1,3846	1,3843	1,3719	1,3837	1,3840	1,3843	1,3843
7	1,3833	1,3840	1,3842	1,3845	1,3843	1,3814	1,3838	1,3841	1,3843	1,3844
8	1,3837	1,3838	1,3841	1,3845	1,3843	1,3835	1,3838	1,3842	1,3843	1,3844
9	1,3836	1,3838	1,3842	1,3845	1,3844	1,3843	1,3835	1,3840	1,3845	1,3843

Fonte: Elaborada pela autora (2024).

Tabela 4 – Perda do gerador (média) para cada combinação de α e β.

β \ α	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,7674	0,6975	0,6990	0,6988	0,6983	0,6990	**0,6890**	0,7025	0,7033	0,6994
1	0,9720	0,9509	0,9524	0,9518	0,9497	0,9518	0,9460	0,9562	0,9554	0,9563
2	1,2125	1,2034	1,2049	1,2025	1,2026	1,2051	1,2007	1,2074	1,2066	1,2039
3	1,4610	1,4558	1,4593	1,4559	1,4555	1,4542	1,4563	1,4586	1,4583	1,4570
4	1,7111	1,7095	1,7116	1,7070	1,7091	1,8962	1,7089	1,7138	1,7106	1,7071
5	1,9624	1,9623	1,9641	1,9590	1,9613	2,0273	1,9597	1,9632	1,9614	1,9633
6	2,2195	2,2162	2,2167	2,2111	2,2130	2,2685	2,2154	2,2166	2,2128	2,2155
7	2,4729	2,4680	2,4676	2,4650	2,4650	2,4687	2,4697	2,4680	2,4671	2,4654
8	2,7221	2,7242	2,7213	2,7178	2,7182	2,7130	2,7211	2,7149	2,7197	2,7194
9	2,9747	2,9764	2,9728	2,9683	2,9704	2,9600	2,9760	2,9719	2,9733	2,9749

Fonte: Elaborada pela autora (2024).

Tabela 5 – Distância $\sqrt{D^2 + G^2}$ para cada combinação de α e β.

β \ α	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	**1,5217**	1,5495	1,5504	1,5509	1,5506	1,5510	1,5462	1,5520	1,5527	1,5511
1	1,6803	1,6790	1,6800	1,6801	1,6789	1,6802	1,6760	1,6822	1,6820	1,6827
2	1,8384	1,8338	1,8350	1,8339	1,8338	1,8355	1,8314	1,8367	1,8363	1,8347
3	2,0120	2,0086	2,0111	2,0091	2,0087	2,0079	2,0085	2,0106	2,0108	2,0098
4	2,2006	2,1995	2,2011	2,1980	2,1994	2,3149	2,1989	2,2029	2,2006	2,1980
5	2,4013	2,4014	2,4028	2,3989	2,4006	2,4460	2,3991	2,4020	2,4008	2,4024
6	2,6154	2,6128	2,6133	2,6089	2,6103	2,6511	2,6120	2,6132	2,6102	2,6125
7	2,8335	2,8296	2,8293	2,8273	2,8271	2,8290	2,8310	2,8296	2,8290	2,8275
8	3,0536	3,0556	3,0531	3,0502	3,0504	3,0454	3,0527	3,0475	3,0518	3,0515
9	3,2807	3,2824	3,2793	3,2753	3,2772	3,2677	3,2819	3,2784	3,2798	3,2812

Fonte: Elaborada pela autora (2024).

6 CONCLUSÃO

Este trabalho apresentou métodos de geração de dados sintéticos utilizando a arquitetura GAN em dados de saúde com o intuito de gerar uma base completamente sintética que tenha uma distribuição de atributos próxima da distribuição dos dados reais para que possa ser usada em tarefas de aprendizado de máquina. Essa abordagem tem como principal objetivo aumentar a eficácia e robustez dos modelos sem que uma base com informações potencialmente sensíveis precise ser utilizada. Para isso, explorou-se a utilização de métodos de treinamento específicos que demonstraram melhorias significativas na distribuição dos atributos dos dados sintéticos gerados.

O método utilizado aqui envolveu a aplicação de uma função de perda baseada na divergência de *Kullback-Leibler* acoplada à rede geradora, juntamente com um mecanismo de treinamento que varia o tamanho de janela e passo para aplicar essa função. Além disso, utiliza-se um parâmetro α com o intuito de atenuar a divergência KL e avaliar qual seria o fator de atenuação ótimo. Também foram realizados experimentos com um fator de atenuação β vinculado ao erro médio quadrático (MSE), além do já mencionado componente α vinculado à divergência KL.

Os resultados obtidos através de *boxplots* que comparam a distribuição dos dados reais com a dos dados sintéticos sugere que o método de janela e passo juntamente com o componente da divergência KL melhora a representatividade e a qualidade desses dados, tornando-os mais próximos da dos dados reais quando comparada com o método *baseline*. Isso é crucial para a construção de modelos de aprendizado de máquina confiáveis e generalizáveis em contextos de saúde, onde a fidedignidade dos resultados é de grande importância. Além da utilização dos dados como substituição de bases com informações sensíveis, essa abordagem oferece uma perspectiva adicional para lidar com desafios de escassez de dados e desequilíbrio de rótulos, um problema recorrente em aprendizado de máquina.

Como trabalhos futuros, pretende-se introduzir outros tipos de erro no treinamento para que se possa avaliar o seu impacto na distribuição dos dados gerados. A divergência KL se mostrou eficaz para aproximar a distribuição dos atributos contínuos dos dados sintéticos em relação aos dos dados reais, e novos experimentos com tamanhos de janela e passo diferentes, além de critérios adicionais para a aplicação desse erro são exemplos de treinamentos possíveis. Por fim, também deve-se considerar a adoção de outras métricas de avaliação de resultados, como o FID. A utilização dessa métrica requisita a separação da base de dados em bases de treinamento, validação e teste, algo que não foi seguido neste trabalho, pois optou-se por utilizar apenas um conjunto de treinamento.

REFERÊNCIAS

BORJI, Ali. Pros and cons of gan evaluation measures: New developments. **Computer Vision and Image Understanding**, Elsevier, v. 215, p. 103329, 2022.

EMAM, Khaled El; ARBUCKLE, Luk. **Anonymizing health data: case studies and methods to get you started**. [S.l.]: "O'Reilly Media, Inc.", 2013.

FEDER, Amir; KEITH, Katherine A; MANZOOR, Emaad; PRYZANT, Reid; SRIDHAR, Dhanya; WOOD-DOUGHTY, Zach; EISENSTEIN, Jacob; GRIMMER, Justin; REICHART, Roi; ROBERTS, Margaret E et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. **Transactions of the Association for Computational Linguistics**, MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA ..., v. 10, p. 1138–1158, 2022.

FRID-ADAR, Maayan; DIAMANT, Idit; KLANG, Eyal; AMITAI, Michal; GOLDBERGER, Jacob; GREENSPAN, Hayit. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. **Neurocomputing**, Elsevier, v. 321, p. 321–331, 2018.

GOODFELLOW, Ian; POUGET-ABADIE, Jean; MIRZA, Mehdi; XU, Bing; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron; BENGIO, Yoshua. Generative adversarial nets. **Advances in neural information processing systems**, v. 27, 2014.

HASHEMI, Atiye Sadat; SOLIMAN, Amira; LUNDSTRÖM, Jens; ETMINANI, Kobra. Domain knowledge-driven generation of synthetic healthcare data. In: IOS PRESS. **The 33rd Medical Informatics Europe Conference, MIE2023, Gothenburg, Sweden, 22-25 May, 2023**. [S.l.], 2023. v. 302, p. 352–353.

HEUSEL, Martin; RAMSAUER, Hubert; UNTERTHINER, Thomas; NESSLER, Bernhard; HOCHREITER, Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. **Advances in neural information processing systems**, v. 30, 2017.

IBRAHIM, Ibrahim; ABDULAZEEZ, Adnan. The role of machine learning algorithms for diagnosing diseases. **Journal of Applied Science and Technology Trends**, v. 2, n. 01, p. 10–19, 2021.

KULLBACK, Solomon; LEIBLER, Richard A. On information and sufficiency. **The annals of mathematical statistics**, JSTOR, v. 22, n. 1, p. 79–86, 1951.

LAKHEY, M. **Generative Adversarial Networks Desmystified**. 2019. . https://medium.com/datadriveninvestor/gans-demystified-f057f5 e32fc9.

LANILLOS, Pablo; MEO, Cristian; PEZZATO, Corrado; MEERA, Ajith Anil; BAIOUMY, Mohamed; OHATA, Wataru; TSCHANTZ, Alexander; MILLIDGE, Beren; WISSE, Martijn; BUCKLEY, Christopher L et al. Active inference in robotics and artificial agents: Survey and challenges. **arXiv preprint arXiv:2112.01871**, 2021.

LEDIG, Christian; THEIS, Lucas; HUSZÁR, Ferenc; CABALLERO, Jose; CUNNINGHAM, Andrew; ACOSTA, Alejandro; AITKEN, Andrew; TEJANI, Alykhan;

TOTZ, Johannes; WANG, Zehan et al. Photo-realistic single image super-resolution using a generative adversarial network. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 4681–4690.

LIU, Katherine; OK, Kyel; VEGA-BROWN, William; ROY, Nicholas. Deep inference for covariance estimation: Learning gaussian noise models for state estimation. In: IEEE. **2018 IEEE International Conference on Robotics and Automation (ICRA)**. [S.l.], 2018. p. 1436–1443.

PARK, Noseong; MOHAMMADI, Mahmoud; GORDE, Kshitij; JAJODIA, Sushil; PARK, Hongkyu; KIM, Youngmin. Data synthesis based on generative adversarial networks. **arXiv preprint arXiv:1806.03384**, 2018.

PIACENTINO, Esteban; ANGULO, Cecilio. Generating fake data using gans for anonymizing healthcare data. In: SPRINGER. **Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 6–8, 2020, Proceedings**. [S.l.], 2020. p. 406–417.

PIACENTINO, Esteban; GUARNER, Alvaro; ANGULO, Cecilio. Generating synthetic ecgs using gans for anonymizing healthcare data. **Electronics**, MDPI, v. 10, n. 4, p. 389, 2021.

QUINLAN, Ross. **Thyroid Disease**. 1987. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5D010.

RAJABI, Amirarsalan; GARIBAY, Ozlem Ozmen. Tabfairgan: Fair tabular data generation with generative adversarial networks. **Machine Learning and Knowledge Extraction**, MDPI, v. 4, n. 2, p. 488–501, 2022.

SHORTEN, Connor; KHOSHGOFTAAR, Taghi M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019.

SUN, Ke; XIAO, Bin; LIU, Dong; WANG, Jingdong. Deep high-resolution representation learning for human pose estimation. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2019. p. 5693–5703.

TRIASTCYN, Aleksei; FALTINGS, Boi. Generating artificial data for private deep learning. **arXiv preprint arXiv:1803.03148**, 2018.

WANG, Qi; MA, Yue; ZHAO, Kun; TIAN, Yingjie. A comprehensive survey of loss functions in machine learning. **Annals of Data Science**, Springer, p. 1–26, 2020.

WASSERMAN, Larry. **All of statistics: a concise course in statistical inference**. [S.l.]: Springer, 2004. v. 26.

YOON, Jinsung; DRUMRIGHT, Lydia N; SCHAAR, Mihaela Van Der. Anonymization through data synthesis using generative adversarial networks (ads-gan). **IEEE journal of biomedical and health informatics**, IEEE, v. 24, n. 8, p. 2378–2388, 2020.

YU, Lantao; ZHANG, Weinan; WANG, Jun; YU, Yong. Seqgan: Sequence generative adversarial nets with policy gradient. In: **Proceedings of the AAAI conference on artificial intelligence**. [S.l.: s.n.], 2017. v. 31, n. 1.