SHEILA NUNES COSTA SANTOS

Geração de Dados de Descomissionamento de Plataformas Offshore a Partir de Redes Adversariais Generativas

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientadora: Paula Medina Maçaira Louro
Coorientadora: Fernanda Araujo Baião

Departamento de Engenharia Industrial
Rio de Janeiro, 22 de novembro de 2024.

Dedico esse trabalho a todos que me amaram, apoiaram e confiaram em mim ao longo de minha jornada até aqui. Sem vocês, eu não teria conseguido. Muito obrigada por tudo.

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer a D'us, bendito seja, por ter me guiado e dado forças nos momentos em que mais precisei. Agradeço por todas as oportunidades e maravilhas que vivi até aqui.

Em seguida, agradeço à PUC-Rio, por ter me proporcionado experiências incríveis ao longo de minha graduação. Ter passado por essa instituição foi uma experiência transformadora que mudou a minha vida.

Também gostaria de agradecer a minha orientadora e coorientadora, pela sugestão do tema desse trabalho e por terem confiado na minha capacidade de desenvolver um estudo desse tipo.

À minha equipe de trabalho, gostaria de agradecer toda a compreensão e apoio ao longo desses últimos meses, vocês são incríveis.

Aos meus amigos, gostaria de agradecer a paciência. Vocês tornaram meus dias mais leves, e sem vocês, essa trajetória teria sido muito mais insustentável.

Aos meus pais, tios, tias e avós por todo amor e confiança, vocês me tornaram a pessoa que sou e me apoiaram ao longo de toda essa trajetória. Mas, em especial, preciso agradecer à minha irmã Shirley pelas palavras de conforto, suporte e direcionamento. Você é uma luz na minha vida, meu braço direito. Sem você, eu definitivamente não teria conseguido.

De forma geral, agradeço a todos que acreditaram em mim, especialmente nos momentos em que eu mesma não conseguia acreditar. Sou extremamente grata e sortuda por ter todos vocês na minha vida.

Por fim, gostaria de agradecer a mim mesma, por todas as vezes em que escolhi tentar mais uma vez, mesmo quando tudo parecia perdido.

RESUMO

Mesmo em um contexto em que a sustentabilidade e o uso de energias renováveis estejam cada vez mais em evidência, ainda é crescente a exploração de petróleo em bacias brasileiras. Nesse contexto, este trabalho apresenta uma aplicação de Redes Adversariais Generativas (GANs) na geração de dados sintéticos relacionados a plataformas offshore em descomissionamento. Levando-se em conta que o processo de descomissionamento é uma etapa crítica na exploração de petróleo, envolvendo desafios ambientais, econômicos e de segurança, é notória a escassez de dados públicos relacionados às plataformas brasileiras, o que dificulta o desenvolvimento de modelos preditivos robustos. Nesse contexto, as GANs surgem como uma alternativa promissora para ampliar a disponibilidade de dados estatisticamente semelhantes aos reais. Este estudo busca explorar a construção de uma GAN utilizando a biblioteca PyTorch, avaliando sua eficácia principalmente através de duas métricas, a comparação da correlação dos dados e a comparação da distribuição. Os resultados obtidos revelam que a rede geradora foi capaz de replicar características-chave dos dados reais, embora com limitações no caso de atributos com baixa variabilidade nos dados originais.

Palavras-chave: Redes Adversariais Generativas, GANs, dados sintéticos, descomissionamento de plataformas, aprendizado de máquina

ABSTRACT

Even in a context where sustainability and the use of renewable energy are increasingly in focus, the exploration of oil in Brazilian basins continues to grow. In this scenario, this work presents an application of Generative Adversarial Networks (GANs) for generating synthetic data related to decommissioning offshore platforms. Considering that the decommissioning process is a critical stage in oil exploration, involving environmental, economic, and safety challenges, there is a notable scarcity of public data regarding Brazilian platforms, which hinders the development of robust predictive models. In this scenario, GANs emerge as a promising alternative to increase the availability of data that are statistically like real ones. This study aims to explore the construction of a GAN using the PyTorch library, evaluating its effectiveness mainly through two metrics: correlation comparison and distribution comparison. The results show that the generator network was able to replicate key characteristics of the real data, although with limitations in features with low variability in the original dataset.

Key-words: Generative Adversarial Networks, GANs, synthetic data, platform decommissioning, machine learning

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GAN – Rede Adversarial Generativa

EI – *Energy Institute*

IEA – *International Energy Agency*

ANP - Agência Nacional de Petróleo, Gás Natural e Biocombustíveis

FPSO – Unidade Flutuante de Produção, Armazenamento e Transferência

SS – Plataforma Semissubmersível

RNA – Rede Neural Artificial

MSE - *Mean Squared Error*

VAE - Autocodificador Variacional

CGAN - *Conditional Generative Adversarial Network*

DCGAN - *Deep Convolutional Generative Adversarial Network*

LAPGAN - *Laplacian Pyramid Generative Adversarial Network*

PDI - Programa de Descomissionamento de Instalações

ReLU - *Rectified Linear Unit*

LeakyReLU - *Leaky Rectified Linear Unit*

Tanh - Tangente Hiperbólica

BCELoss - *Binary Cross Entropy Loss*

Adam - *Adaptive Moment Estimation*

AdaGrad - *Adaptive Gradient Algorithm*

RMSProp - *Root Mean Square Propagation*

KS - Kolmogorov-Smirnov

SUMÁRIO

1 INTRODUÇÃO

Em 2022 o Brasil foi o 9º maior produtor de petróleo e 30º maior produtor de gás natural no ranking mundial (EI, 2023). Já em relação às perspectivas futuras, a Agência Econômica Internacional estima que, até 2040, o país atinja a sétima posição entre os produtores mundiais de petróleo (IEA, 2020). A maior exploração dos campos levanta diversas questões em relação ao destino das estruturas no final da vida útil das bacias exploratórias. Nesse ponto, destaca-se o problema relacionado ao descarte das estruturas de exploração, que até pouco tempo eram simplesmente abandonadas, sem um fim bem determinado. Para contornar isso, surge o processo descomissionamento da estrutura, onde a escolha do método da remoção é crucial para o sucesso da operação, considerando questões como impactos ambientais, segurança e custos (SOUZA, 2022).

Estudos recentes têm explorado o uso de tecnologias avançadas no processo de decisão para descomissionamento, construindo modelos capazes de apoiar estratégias decisórias tanto no âmbito estratégico quanto operacional. Essas modelagens visam reduzir custos e riscos ao prever cenários e minimizar a geração de resíduos durante o descomissionamento (KRISHNAMOORTHY et. al, 2023). No entanto, observa-se que, no campo da Ciência de Dados, há uma escassez significativa de dados disponíveis relacionados às bacias em descomissionamento, o que dificulta o desenvolvimento de modelos robustos através de técnicas de aprendizado de máquina (VUTTIPITTAYAMONGKOL et. al, 2021).

Nesse contexto, o uso de Redes Adversariais Generativas (GANs), que têm como propósito gerar dados sintéticos estatisticamente semelhantes aos reais, surge como uma alternativa promissora para ampliar a disponibilidade de dados. O objetivo deste trabalho é avaliar se, no caso de plataformas brasileiras offshore de petróleo em descomissionamento, as GANs são capazes de gerar dados sintéticos com qualidade comparável aos dados reais. Além disso, busca-se determinar se esses dados possuem qualidade suficiente para serem utilizados em estudos futuros. Espera-se, portanto, realizar uma análise comparativa entre os dados gerados e os dados reais, verificando a similaridade em termos de distribuição das amostras e correlação, de modo a validar a utilidade dos dados sintéticos.

Destaca-se que a escolha de plataformas de petróleo offshore deu-se por conta do alto grau de complexidade dessas operações, que precisam de um longo

período de avaliação antes da tomada de decisão do descomissionamento. Além disso, os riscos envolvidos no processo, assim como o alto potencial de impactos ambientais, decorrente do abandono dessas estruturas, sugerem a necessidade de estudos mais abrangentes e técnicas mais robustas no auxílio da tomada de decisão (MARTINS, 2015).

Para atingir os objetivos desse estudo, escolheu-se uma arquitetura base para treinar a GAN. Em seguida, foi construído e aplicado um plano de experimentação, que permitiu encontrar a melhor combinação de hiperparâmetros para a Rede Geradora. Destaca-se que, para o melhor cenário, foi possível observar para alguns atributos, taxas de semelhança superiores a 70%. Ademais, a taxa de semelhança geral do melhor modelo ficou em 65,7%. Entende-se que esse valor não foi tão alto quanto o esperado, porém essa condição foi atrelada ao baixo desempenho do modelo em relação aos atributos com baixa variabilidade nos dados. Em suma, compreendeu-se que a GAN foi capaz de gerar dados condizentes para os atributos com maior variabilidade e mostrou-se promissora para aplicação em estudos futuros.

Este trabalho foi estruturado da seguinte forma: o Capítulo 2 apresenta uma revisão da literatura relacionada ao tema da pesquisa, abordando o descomissionamento de plataformas, a definição de dados sintéticos e a arquitetura da GAN. Por ser composta por duas redes neurais, também se aborda, de forma resumida, a estrutura de redes neurais artificiais. Em seguida, o Capítulo 3 descreve toda a metodologia aplicada ao longo do estudo, enquanto o Capítulo 4 apresenta os resultados observados, que serão analisados e discutidos, de forma sucinta, no Capítulo 5. Por fim, o Capítulo 6 traz as conclusões obtidas no estudo. Ao final do documento, em apêndice, encontram-se, em maior resolução, as imagens geradas pelo modelo que apresentou o melhor desempenho.

2 REFERENCIAL TEÓRICO

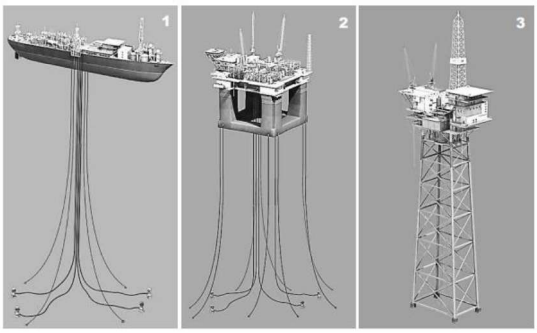
2.1 DESCOMISSIONAMENTO DE PLATAFORMAS DE PETRÓLEO

O Brasil é globalmente conhecido por seu potencial petrolífero, com destaque para a exploração das reservas do pré-sal. Conforme o último Boletim da Produção de Petróleo e Gás Natural (2024), da Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP), o país conta com 256 áreas de concessão de exploração, operadas por 54 empresas, sendo 67 em áreas marítimas e 205 em áreas terrestres. Em relação à produção, o último balanço totalizou 151.227 mil m^3/d de petróleo produzidos, equivalendo a 3,23 MMbbl/d (milhões de barris por dia).

Ademais, torna-se necessário compreender que todo campo explorado possui um ciclo de vida composto de três macro etapas: (1) Desenvolvimento, (2) Produção e (3) Desativação. Destaca-se que nas fases de desenvolvimento ocorre a escolha do tipo de estrutura utilizada na exploração da bacia, nos casos de projetos offshore brasileiros, principal foco desse estudo, ocorre o uso principal de três tipos de estrutura: a FPSO (Flutuante), a SS e a Fixa, conforme apresentado na Figura 1 (PROENÇA, 2023).

Conforme apresentado por André Proença et. al (2023), de forma resumida, uma FPSO consiste em um navio-tanque, que se ancora ao solo e explora os poços geralmente através de dutos flexíveis, seu conteúdo é desabastecido por meio de navios terceiros, os petroleiros. Já uma SS é uma instalação flutuante estabilizada por colunas, ela possui sistema de posicionamento dinâmico similar às sondas FPSO, sua conexão com o poço pode se dar por meio de dutos rígidos ou flexíveis e o escoamento da produção pode ser feito através de oleodutos ou através de petroleiros. Por fim, as Plataformas Fixas são estruturas rígidas, com profundidade de até 300m, cravadas no fundo do mar através de estacas, são estruturas mais antigas que não costumam ser empregadas em novos projetos, e sua conexão com a costa costuma se dar através de oleoduto, uma vez que costumam estar instaladas próximo a ela.

Figura 1 - Principais tipos de plataformas offshore utilizadas em bacias brasileiras



Fonte: MARTINS, 2015

Em relação à macro etapa de Desativação, compreende-se que ela se inicia no momento que ocorre um declínio na produção do poço e sua exploração passa a não ser mais vantajosa. A partir do momento de declínio, antes do processo de descomissionamento, ocorre o Abandono do Campo e do Poço, de acordo com a Resolução nº 27, de 18 de outubro de 2006, da ANP, o primeiro se trata do processo de alienação ou reversão de todas as instalações de produção de todos os poços que compõe o campo, já o segundo trata-se de uma série de operações que buscam isolar o local de extração, podendo ser permanente – quando não se pretende retomar a exploração – ou temporário – quando existe algum nível de interesse em relação ao retorno.

Apenas após o abandono o descomissionamento torna-se possível, de acordo com a definição apresentada por Karen Souza (2022), o descomissionamento é o conjunto de atividades que se iniciam com o arrasamento dos poços, englobam a remoção das instalações, a destinação adequada dos rejeitos, resíduos e materiais removidos e a recuperação ambiental da área. Até 2020, não existia a obrigatoriedade na remoção das estruturas ou do planejamento do processo de descomissionamento, porém através da resolução nº 817, de abril de 2020, da ANP, houve a formalização da obrigatoriedade por parte de todos os agentes responsáveis por instalações de exploração e produção de petróleo e gás natural no fim de sua vida útil.

O processo de descomissionamento envolve diversos setores diferentes, e processos extremamente complexos. Conforme as bacias brasileiras vão atingindo maturidade, a expectativa é que cada vez mais campos necessitem ser descomissionados, conforme apresentado no Painel Dinâmico da ANP (2024), a expectativa é que, até 2028, 3.883 campos necessitem de descomissionamento,

movimentando um total de R$ 64,39 bilhões em investimentos. Somente a região de Sergipe, região cujas plataformas desse estudo estão localizadas, tem um potencial de movimentar R$ 9 bilhões de reais.

Ainda assim, no presente momento, o acesso à informação de campos já descomissionados ou em descomissionamento são escassos, e no contexto da ciência de dados, o volume de informação é algo essencial para a produção de projeções e análises robustas. E é justamente nesse contexto que o uso de Redes Adversariais Generativas se faz necessário. Aumentar a disponibilidade de dados em relação a plataformas descomissionadas através de dados sintéticos pode vir a ajudar diversos outros estudos que englobem o tema do descomissionamento de plataformas.

2.2 DADOS SINTÉTICOS

Antes de compreender as Redes Adversárias Generativas (GANs), é necessário contextualizar o que são dados sintéticos. Em suma, tratam-se de dados criados artificialmente, por algum tipo de modelagem não humana, com o intuito de representarem, da forma mais semelhante possível, o mundo real. Nesse caso, espera-se que o conjunto de dados sintético tenha as mesmas propriedades matemáticas, como por exemplo características estatística dos dados originais, mas não contenham informações exatamente iguais às de entrada (RODRIGUES, 2021).

Tanto no meio acadêmico quanto fora dele, em diferentes momentos encontrar dados robustos o suficiente para o treinamento de modelos pode ser um desafio. Sendo assim, é principalmente nesse contexto que os dados sintéticos surgem, é possível gerar novos dados, com baixo custo computacional, respeitando as políticas de privacidade de dados de usuários. Em áreas de saúde, jurídicas e financeiras, que possuem dados sensíveis, é possível manter a fidelidade dos dados, multiplicando-os seguindo as mesmas informações estatisticamente relevantes sem expor dados confidenciais ou privados. Por fim, outro uso significativo seria na redução de viés de amostras de treinamento, onde por exemplo poderia ser possível aumentar os dados de treinamento para reduzir desequilíbrios na população da amostra (PORTELA, 2022).

Em geral, os dados sintéticos podem ser parciais ou completos, onde nos parciais apenas uma parte dos dados são complementados com dados gerados sinteticamente, enquanto nos completos todos os dados são novos (gerados por algum modelo). A grande diferença nesses casos é o objetivo da análise: no caso do uso de dados parcialmente sintético, um uso comum é na manutenção do anonimato de indivíduos, onde os dados originais são camuflados com dados gerados para evitar vazamentos de informações, como em estudos clínicos (AWS, 2024).

Em relação aos tipos de dados que podem ser gerados sinteticamente, eles podem ser de diferentes tipos incluindo imagens, vídeos, textos, números, tabelas, e outros tipos, gerados a partir de simulações e modelos computacionais, que ganham muito espaço com o avanço das diferentes técnicas de *machine learning* e inteligência artificial.

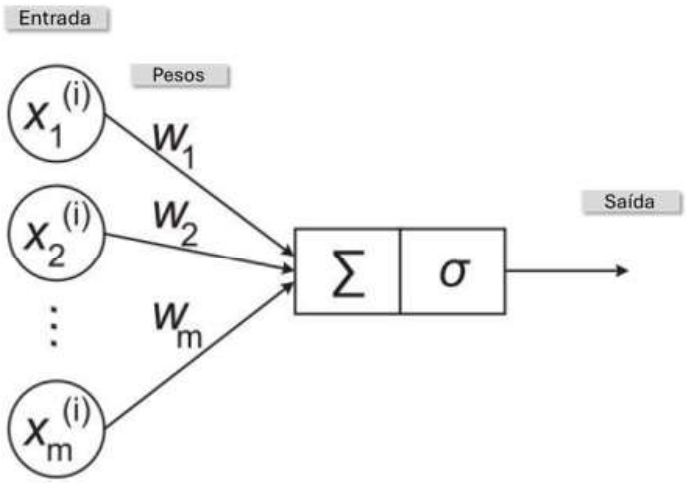
Por fim, vale destacar que as três formas mais comuns de geração desses dados são a distribuição estatística, que consiste na criação de um subconjunto de dados estatisticamente semelhante ao original através da aplicação dessas mesmas características, a modelagem de machine learning, onde um modelo é treinado para replicar as características dos dados reais, geralmente também replicando as características estatísticas, e por último, o aprendizado profundo, método escolhido para ser aplicado nesse estudo, que consiste na aplicação de técnicas mais avançadas, como Redes Adversárias Generativas (PORTELA,2022).

2.3 REDES NEURAIS ARTIFICIAIS

Visando replicar computacionalmente o funcionamento dos neurônios humanos, Redes Neurais Artificiais (RNA) são algoritmos aplicados para previsões e análise de grandes volumes de dados, onde redes matemáticas, no formato similar a neurônios, são formadas a fim de resolver problemas complexos de forma mais ágil e eficaz. Essas soluções ocorrem através do treinamento da rede, por meio de pesos atribuídos a cada atributo do conjunto de dados, de forma resumida, esses atributos são relacionados a um modelo matemático que busca se adaptar aos dados de treinamento, replicando o comportamento aprendido a novos dados que passará a receber (WEBER, 2024).

De forma mais específica, conforme Kraus et al. (2020), uma RNA é uma conexão de unidades denominadas neurônios artificiais, sendo que cada um desses neurônios, também conhecidos como *perceptrons*, recebe entradas externas à rede, ou de outro neurônio. A RNA mais simples é aquela que possui apenas uma camada de neurônios, conforme a Figura 2, onde o modelo computa a soma ponderada entre o vetor de pesos (w) e as entradas, aplicando a função de ativação (σ) de forma livre de ciclos, ou seja, o neurônio passando os dados de entrada apenas para frente, porém, essas redes tem poder apenas para resolver problemas linearmente separáveis, ou seja enquadrados em soluções que estejam no contexto de uma reta, um plano ou um hiperplano.

Figura 2 - Ilustração de Neurônio em uma RNA simples



Fonte: Kraus, Feuerrigel e Oztekin (2020) – adaptado

Acrescentando complexidade à rede, surgem as redes neurais multicamadas, que tem como característica possuir uma ou mais camadas ocultas, entre a camada de entrada e a camada de saída. Cada camada oculta pode possuir um número independente de neurônios e ter uma função de ativação independente das demais camadas presentes na rede, destaca-se que individualmente esses neurônios realizam cálculos semelhantes aos apresentados anteriormente. A presença de camadas ocultas permite que a rede aprenda relações mais complexas e passe a representar dados com maior hierarquia, agora a rede é capaz de gerar qualquer polinômio, nesse caso, quanto maior o número de camadas ocultas e o número de neurônios dentro delas, maior seu poder de representar a complexidade das funções (BATISTA et.al, 2022)

Conforme apresentado por Márcio Batista et. al (2022), para construir e treinar uma rede neural é necessário definir alguns conceitos e parâmetros. As funções de ativação presentes nas camadas tem como objetivo fornecer capacidade de processamento às redes através da aplicação de transformações não lineares, a escolha da função utilizada impacta diretamente a eficiência e capacidade de aprendizado da rede.

De forma complementar, ainda segundo Batista, toda rede é avaliada em relação à sua perda, através da chamada Função de Perda. Essas funções medem de alguma forma a diferença entre o valor gerado pela rede e o valor verdadeiro, ao longo do treinamento, uma função muito popular é a *Mean Squared Error* (MSE), que calcula a média dos erros quadráticos das predições. No processo de estimar os melhores valores possíveis para os pesos (W_i) da rede, claramente, espera-se minimizar essas perdas ao longo do treinamento do problema e com isso surge outro participante das redes neurais, o otimizador.

No caso do otimizador, de modo geral, dentre as diversas opções disponíveis, aplica-se o conceito de máximos e mínimos do cálculo diferencial através do gradiente descendente. De forma resumida e simplificada, o otimizador através de algum método matemático estima o gradiente da função de perda para cada rodada do treinamento, com isso é possível obter uma visão parcial da direção em que o erro mínimo possa estar e através da taxa de aprendizado o modelo se otimiza e direciona gradualmente ao ponto.

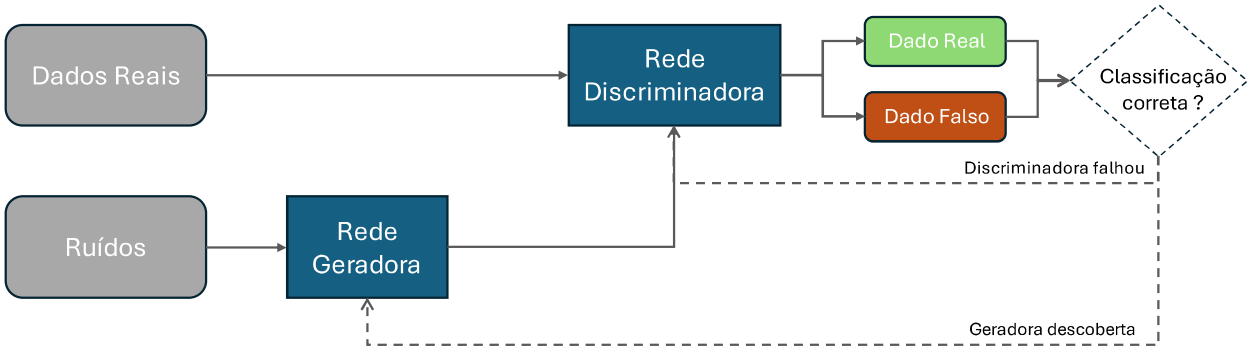
Nesse caso, devem ser considerados dois conceitos, o de épocas e o da taxa de aprendizado, onde a época é a quantidade de vezes que o modelo fará essa estimação e balanceamento do peso (W_i) e a taxa de aprendizado, a magnitude do direcionamento, até essa direção. Taxas de aprendizados muito elevadas podem fazer com que a rede dê saltos ao longo do treinamento nunca encontrando o mínimo, enquanto a época desbalanceada pode causar *overfitting* - situação em que o modelo memoriza o conjunto de treino e ao ser aplicado em dados reais não tem a desempenho adequado (BATISTA et. al, 2022).

2.4 REDES ADVERSARIAIS GENERATIVAS

Segundo Alankrita Aggarwal et al. (2021), uma Rede Adversarial Generativa (GAN) consiste em uma categoria de aprendizado de máquina, mais

especificamente voltado para o aprendizado não supervisionado, onde duas redes neurais competem entre si. Conceitualizada em 2014, por Ian Goodfellow, as GANs basicamente são formadas por uma parte geradora, que cria dados de forma realista, aproximando-os dos dados de entrada do modelo, e outra discriminadora, que avalia se os dados recebidos por ela são reais ou falsos. Em suma, o objetivo é que enquanto a rede geradora produza cada vez dados mais convincentes, a rede discriminadora aprenda a diferenciar cada vez melhor esses dados. A grande vantagem desse processo adversarial, são modelos muito robustos, capazes e produzir dados sintéticos muito próximos dos reais, de forma muito eficiente (CAMPOS, 2022).

Figura 3 - Exemplo genérico de GANs



Fonte: Autoria Própria

Em relação à sua aplicação, as GANs têm sido muito aplicadas em campos como detecção facial, processamento de imagens, processamento e geração de textos, mas ainda em campos como controle de tráfego, medicina e geração de imagens 3D e até mesmo criação de músicas. Em suma, essa modelagem vem transformando o campo do aprendizado profundo (conhecido como *deep learning*), permitindo que modelos gerem dados sintéticos realistas a partir de grandes conjuntos de dados não rotulados. Ainda assim, existem muitos desafios atrelados às GANs, Aggarwal e outros (2021) citam *mode colapse*, ou no português colapso de modo, que seria uma modelagem viesada, quando apenas um subconjunto dos dados é considerado no momento de gerar os dados sintéticos, o que limita em alguns casos, o poder de diversificar os exemplos realistas nas saídas do modelo. Outros problemas citados foram o poder de generalização, pois em alguns casos as GANs não conseguem generalizar conjuntos de dados complexos e variados, e por fim a falsificação de imagens, onde o uso dessa tecnologia pode ser atrelado a

fraudes e falsificações, como *deepfakes*, tem levantado preocupações éticas (AGGARWAL et al., 2021).

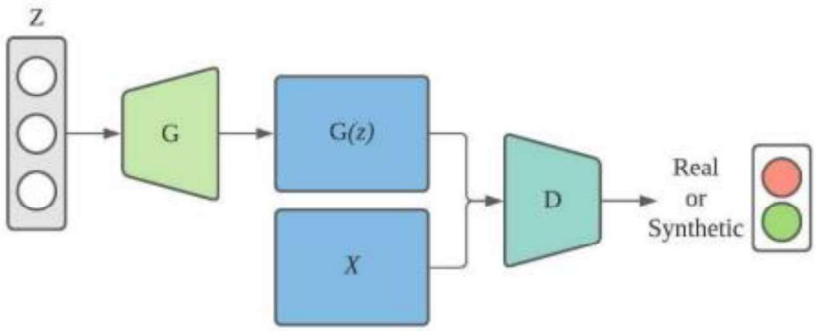
De fato, a função mais popular das Redes Adversariais Generativas é a criação de imagens, principalmente de pessoas, ainda que a maior capacidade dessa modelagem seja de gerar dados sintéticos que parecem indistinguíveis de dados reais. Como vantagem, destaca-se o fato de que esse tipo de modelo funciona muito bem com bases de dados menores, ou seja, quando inicialmente existem poucos dados disponíveis, as GANs realmente são uma saída alternativa. Além disso, em relação a criação de imagens, esse tipo de modelagem se sai muito bem se comparado a outros modelos como autocodificadores variacionais (VAEs) (Ruiz-Gándara et al. 2024). Ainda assim, o uso das GANs pode vir a ser explorado em outros tipos de dados, como nos casos de geração de dados tabulares e numéricos (CAMPOS, 2022).

Segundo os conceitos apresentados por Julio Gomes et al. (2023), abordando de forma mais simplificada o princípio no qual os dados sintéticos são gerados no contexto desse modelo, uma GAN tenta probabilisticamente determinar se um dado pertence ou não a um conjunto, em outras palavras, ela pode ser considerada um classificador. A grande diferença no caso dos modelos generativos é que eles avaliam, dada uma instancia, qual a probabilidade dele pertencer a uma classe y, algo no formato p(x | y), onde x seria o elemento ou dado e y o conjunto a que ele pertence.

Abordando de forma mais detalhada, Ian Goodfellow (2014) apresenta em seu estudo que uma GAN aprende a gerar uma distribuição sintética (p_g), através do mapeamento de uma variável de entrada, que se trada de um ruído, sem significado prático, chamada por ele de z. Esses dados de entrada estão relacionados ao espaço de dados G(z , θ_g), onde G é a função geradora, que na verdade representa uma RNA, de parâmetros θ_g. Além disso, é definida uma segunda RNA, discriminadora, a D(x , θ_d). Essa segunda regressão será responsável por inferir uma probabilidade absoluta a observação x ser real e não p_g (sintética). A fim de maximizar a assertividade na classificação de observações como reais ou sintéticas, a rede neural D deve ser treinada tanto com dados reais, quanto com dados sintéticos gerados pela rede neural G.

Por outro lado, os autores destacam que o treinamento de G é feito de modo a minimizar log(1 – D(G(z, , θ_g), θ_d), uma vez que o objetivo geral de uma GAN é minimizar a probabilidade de D classificar dados sintéticos como falso. Observando a minimização da função logarítmica apresentada, nota-se quanto maiores as chances de a informação ser verdadeira (não sintética), maior será a probabilidade calculada por D, e com isso o valor obtido no logaritmo ficará mais próximo de zero, respeitando a lógica de que o algoritmo se retroalimenta, melhorando a qualidade dos dados gerados através da disputa entre gerador e discriminador (GOODFELLOW, 2014).

Figura 4 - Funcionamento de uma GANs na visão probabilística



Fonte: CAMPOS, Afonso (2022)

Em relação aos tipos de GANs disponíveis, conforme apresentado por Africa Ruiz-Gándara et al. (2024), conforme o avançar dos estudos a respeito desse tipo de modelagem algumas variantes surgiram ao longo do tempo. Dentre as variações disponíveis o autor destacou quatro tipos sendo eles o Vanilla GAN, que é o tipo mais básico, sendo uma das principais características a simplicidade na arquitetura, ainda que não performe tão bem quando encontra dados mais complexos. Em seguida o autor também cita o *Conditional GAN*, conhecido pela sigla CGAN, que funciona como uma espécie de extensão da forma simplificada do modelo, onde dados condicionados a rótulos específicos podem ser considerados na geração de novos dados. Em seguida apresenta o Deep *Convolutional GAN* (DCGAN), que ficou conhecido por melhorar a arquitetura do GAN clássico, substituindo esses *perceptrons* multicamadas por redes convolucionais profundas, entendendo-se essas redes convulsionais como operações matemáticas que destacam características importantes dos dados. Por último, foi destacado também o *Laplacian Pyramid GAN* (LAPGAN), esse modelo trata-se de uma combinação do

CGAN com uma pirâmide laplaciana – técnica utilizada para processar imagens representando-as em múltiplas resoluções – sua aplicação permite gerar imagens de resolução mais detalhada e com maior precisão.

Os autores ainda destacaram outros tipos de GANs que que a literatura tem abordado como: SRGAN, StackGAN, CycleGAN, PassGAN, WGAN, GAN espaço-temporal, GAN restrita, H-GAN, pix2pix, Android-GAN, UNIT, RGAN, RaGAN, AnoGAN, GANomaly, RCGAN, EGAN e TimeGAN. A maior diferença entre essas variações no modelo original são a forma como a rede neural generativa e discriminativa são construídas, assim como o uso individual desses modelos varia de acordo com a essência dos dados trabalhados e os objetivos em relação a aplicação da GAN. Como dito anteriormente, o maior uso desse tipo de modelagem tem sido para a transformação e o tratamento de imagens, mas ainda assim diversos estudos aplicam as GANs em outros contextos e em muitos casos da literatura pode-se observar sucesso nas pesquisas avaliadas (RUIZ-GÁNDARA et al., 2024).

Os estudos recentes destacam aplicações inovadoras de Redes Adversariais Generativas (GANs) no setor de petróleo, especialmente no monitoramento e mitigação de impactos ambientais. Gou et al. (2024) apresentaram o uso de GANs para detecção de manchas de óleo offshore, utilizando uma rede generativa atenta às características espectrais de imagens hiperespectrais, o que contribui para uma identificação mais precisa em ambientes marítimos. Já Chemudupati (2024) propôs um modelo baseado em GANs para simular cenários de derramamento de óleo, possibilitando a criação de dados sintéticos para treinar redes neurais convolucionais (CNNs) com o objetivo de recomendar contramedidas de remediação adequadas, como boias de contenção e biorremediação. Complementarmente, Bui et al. (2023) aplicaram GANs condicionais (Pix2Pix) para gerar imagens sintéticas de derramamentos de óleo, aumentando a robustez e precisão de modelos de classificação e detecção com dados limitados.

Apesar desses avanços, o uso de GANs no setor de petróleo ainda é restrito a etapas específicas, como monitoramento ambiental e manutenção, evidenciando uma lacuna na aplicação dessas redes durante outras fases do ciclo de vida das plataformas, como o descomissionamento. Tal escassez representa uma lacuna relevante que pode ser explorado em estudos futuros, ampliando a utilização de

GANs para a geração de dados sintéticos em cenários onde a disponibilidade de informações é limitada.

Adentrando no cenário de pesquisas nacionais, o estudo de Araujo et al. (2024) explorou a aplicação de GANs na geração de dados sintéticos referentes a plataformas britânicas em descomissionamento, utilizando dados disponibilizados pelo governo do Reino Unido e avaliando de forma sucinta uma única arquitetura de rede criada no estudo. Em contraste, este trabalho se diferencia ao aplicar as GANs em dados brasileiros, além de testar diferentes configurações de arquitetura e realizar uma comparação minuciosa das métricas de avaliação dos dados gerados. Essa abordagem mais detalhada amplia a análise da qualidade dos dados sintéticos e valida sua aplicabilidade em contextos reais. Ainda assim, vale ressaltar a escassez de estudos nacionais que aplicam redes adversariais generativas nessa área, evidenciando o potencial e a relevância de pesquisas futuras nesse campo.

Por todos esses pontos apresentados, fez-se muito necessária e condizente a abordagem proposta nesse trabalho, de gerar artificialmente dados de plataformas através da aplicação de GANs. Espera-se que as GANs aprendam as distribuições dos dados das plataformas e representem seu comportamento de forma coerente.

3 METODOLOGIA

Neste trabalho, optou-se por desenvolver uma Rede Adversarial Generativa (GAN) usando a linguagem *Python 3* e a biblioteca *PyTorch*. Amplamente conhecida e difundida na ciência de aprendizado profundo de máquinas, a *PyTorch* possui uma vasta documentação e diversas implementações públicas, especialmente para geração de imagens e criação de modelos preditivos. Dessa forma, mostrou-se interessante utilizá-la para explorar a geração de dados sintéticos voltados para o cenário tabular, mais especificamente de plataformas em descomissionamento.

Além disso, a biblioteca *Pandas* foi utilizada para manipulação de dados tabulares, enquanto a *Scikit-learn* foi aplicada no processo de normalização dos dados durante o pré-processamento e pós-processamento. De forma complementar, a biblioteca SDMetrics foi empregada para avaliar a qualidade dos dados sintéticos gerados e a *Matplotlib* foi utilizada para visualização de diferentes dados ao longo do trabalho.

O código foi desenvolvido em *notebooks Jupyter* utilizando o *Google Colab,* com o objetivo de otimizar a execução do modelo por meio das redes de processamento da *Google*. O código completo está disponível em: https://abrir.link/ktcrc.

Quanto a estrutura adotada, o desenvolvimento desse estudo procurou seguir as seguintes etapas (1) Pré-processamento e normalização dados, (2) Definição de hiperparâmetros, (3) Criação das Redes Generativas e Discriminadora, (4) Treinamento da Rede Adversarial Generativa, (5) Avaliação do modelo.

3.1 OBTENÇÃO DOS DADOS

A Resolução ANP Nº 817/2020 define o Programa de Descomissionamento de Instalações (PDI) como um documento textual que deve ser elaborado pela empresa responsável pela plataforma para planejar e executar o descomissionamento das instalações de produção e exploração de petróleo e gás. Cada PDI é um programa específico de uma bacia de exploração e contém detalhes

sobre as plataformas presentes, os riscos envolvidos no processo, além do cronograma de execução e das medidas de segurança ambiental.

Com base nesses documentos, o conjunto de dados reais foi criado. Esse conjunto, posteriormente, foi utilizado para treinar a rede discriminadora e avaliar os dados sintéticos gerados pelo modelo. Destaca-se que a formação do conjunto de dados a partir desses documentos só foi possível, dado que cada PDI possui um formato padronizado. Dentre os relatórios disponíveis, optou-se por focar nas bacias do estado de Sergipe, uma vez que, dentre os relatórios encontrados, os referentes às bacias sergipanas eram os mais completos e com o maior número de plataformas inclusas.

Dentre as bacias de exploração na região de Sergipe com plataformas em descomissionamento, apenas três possuíam relatórios completos: Caioba, Camorim e Guaricema. Todas as plataformas em descomissionamento, apresentadas nos respectivos relatórios, eram do tipo fixa, e a quantidade por bacia era a seguinte: na bacia de Caioba, quatro (4) plataformas; em Camorim, onze (11) plataformas; e em Guaricema, sete (7) plataformas (PETROBRAS, 2023a; PETROBRAS, 2023b; PETROBRAS, 2024).

A Tabela 1 apresenta cada atributo considerado na base de dados e sua respectiva descrição. Destaca-se que ao longo do PDI diversas informações ´são apresentadas, mas apenas as que constam na Tabela 1 foram consideradas relevantes para esse estudo.

Tabela 1 - Descrição dos atributos da base de dados

Característica	Descrição
Profundidade da água (m)	Profundidade do mar no local da plataforma.
Peso (t)	Peso total da estrutura a ser removida, medido em toneladas.
Idade da estrutura	Tempo em anos desde a instalação, afetando a condição e a complexidade da remoção.
Idade de parada definitiva de produção	Tempo em anos desde a instalação, até o momento em que a plataforma parou de produzir
Tipo de produção	Tipo de recursos anteriormente extraídos, ao qual é atribuído um valor: 1 para óleo e gás, 2 para somente óleo e 3 para somente gás.
Número de pernas	Quantidade de pernas que precisam ser desmontadas ou removidas.
Altura da jaqueta (m)	Altura da estrutura de suporte (jaqueta) a ser desmantelada, medida em metros.
Distância da costa (km)	Distância da plataforma até a costa.
Risco Operacional Tolerável	Número de casos identificados que podem ocorrer ao longo da operação de descomissionamento classificados como de Baixo Risco.

Característica	Descrição
Risco Operacional Moderado	Número de casos identificados que podem ocorrer ao longo da operação de descomissionamento classificados como de Risco Moderado.
Risco Operacional Não Tolerável	Número de casos identificados que podem ocorrer ao longo da operação de descomissionamento classificados como de Alto Risco.
Riscos Ambientais Efetivo - Pequeno	Número de casos identificados classificados como de Baixo Impacto ambiental, em relação ao Risco Efetivo: quando o impacto está associado a condições normais de operação.
Riscos Ambientais Efetivo - Média	Número de casos identificados classificados como de Médio Impacto ambiental, em relação ao Risco Efetivo: quando o impacto está associado a condições normais de operação.
Riscos Ambientais Efetivo - Grande	Número de casos identificados classificados como de Grande Impacto ambiental, em relação ao Risco Efetivo: quando o impacto está associado a condições normais de operação.
Riscos Ambientais Potencial - Pequeno	Número de casos identificados classificados como de **Baixo Impacto ambiental**, em relação ao **Risco Potencial**: quando se trata de um impacto associado a condições anormais do empreendimento.
Riscos Ambientais Potencial - Média	Número de casos identificados classificados como de **Médio Impacto ambiental**, em relação ao **Risco Potencial**: quando se trata de um impacto associado a condições anormais do empreendimento.
Riscos Ambientais Potencial - Grande	Número de casos identificados classificados como de **Grande Impacto ambiental**, em relação ao **Risco Potencial**: quando se trata de um impacto associado a condições anormais do empreendimento.
Impacto Socioeconômico Efetivo - Pequeno	Número de casos identificados classificados como de **Baixo Impacto socioeconômico**, em relação ao **Risco Efetivo:** quando o impacto está associado a condições normais de operação.
Impacto Socioeconômico Efetivo - Médio	Número de casos identificados classificados como de **Médio Impacto socioeconômico**, em relação ao **Risco Efetivo**: quando o impacto está associado a condições normais de operação.
Impacto Socioeconômico Efetivo - Grande	Número de casos identificados classificados como de **Grande Impacto socioeconômico**, em relação ao **Risco Efetivo**: quando o impacto está associado a condições normais de operação.
Impacto Socioeconômico Potencial - Pequeno	Número de casos identificados classificados como de **Baixo Impacto socioeconômico**, em relação ao **Risco Potencial:** quando se trata de um impacto associado a condições anormais do empreendimento.
Impacto Socioeconômico Potencial - Médio	Número de casos identificados classificados como de **Médio Impacto socioeconômico**, em relação ao **Risco Potencial**: quando se trata de um impacto associado a condições anormais do empreendimento.
Impacto Socioeconômico Potencial - Grande	Número de casos identificados classificados como de **Grande Impacto socioeconômico**, em relação ao **Risco Potencial:** quando se trata de um impacto associado a condições anormais do empreendimento.
Duração do descomissionamento	Intervalo de tempo que durarão todas as fases do descomissionamento, mensurado no planejamento, em meses.

Fonte: Elaboração Própria

3.1.1 Estrutura da Base de Dados

Conforme dito anteriormente, a partir de cada PDI foram extraídas as informações das bacias analisadas, para uma melhor visualização, a Tabela 2 exibe parcialmente a base de dados, considerando apenas as plataformas da bacia de Caioba. Ainda que parcialmente, é possível observar alguns pontos importantes que devem ser considerados para compreender a performance do modelo construído nesse estudo.

Tabela 2 - Base parcial com dados da bacia de Caioba

	CAIOBA			
	PCB-01	**PCB-02**	**PCB-03**	**PCB-04**
Profundidade da água (m)	26	27	27	27
Peso (t)	1407	1292	1016	1088
Idade da estrutura	53	50	46	40
Idade de parada definitiva de produção	49	44	30	36
Tipo de produção	1	1	1	1
Número de pernas	6	7	4	4
Altura da jaqueta (m)	34,9	36,1	35,1	31,8
Distância da costa (km)	11,63	12,29	12,32	12,34
Risco Operacional Tolerável	2	2	2	2
Risco Operacional Moderado	2	2	2	2
Risco Operacional Não Tolerável	0	0	0	0
Riscos Ambientais Efetivo - Pequeno	27	27	27	27
Riscos Ambientais Efetivo - Média	19	19	19	19
Riscos Ambientais Efetivo - Grande	0	0	0	0
Riscos Ambientais Potencial - Pequeno	28	28	28	28
Riscos Ambientais Potencial - Média	14	14	14	14
Riscos Ambientais Potencial - Grande	4	4	4	4
Impacto Socioeconômico Efetivo - Pequeno	1	1	1	1
Impacto Socioeconômico Efetivo - Médio	14	14	14	14
Impacto Socioeconômico Efetivo - Grande	5	5	5	5
Impacto Socioeconômico Potencial - Pequeno	0	0	0	0
Impacto Socioeconômico Potencial - Médio	12	12	12	12
Impacto Socioeconômico Potencial - Grande	3	3	3	3
Duração (meses)	150	150	150	150

Fonte: Elaboração Própria

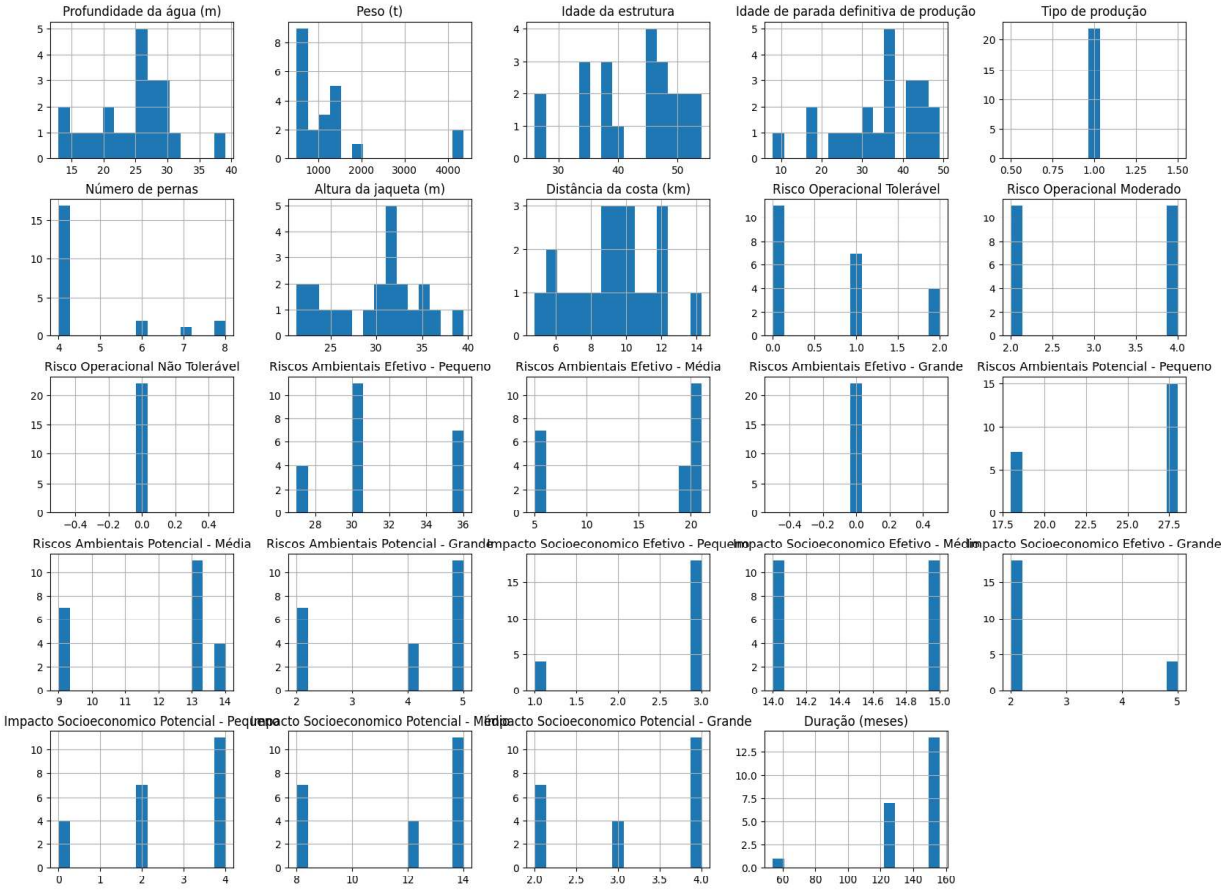
Em relação aos dados de riscos, todo Programa de Descomissionamento de Instalações (PDI) classifica os riscos de maneira similar. Observou-se, ainda, que

os cenários de risco são geralmente classificados por bacia, e os relatórios listam possíveis cenários conforme cada classificação. Com base nessas informações, a base de dados foi montada de acordo com os cenários mapeados para cada classificação de risco, conforme ilustrado na Tabela 1. Nesse caso, para todas as plataformas de um mesmo campo ocorre a repetição de valores referentes aos riscos, já que eles são mapeados de acordo com a bacia.

Em relação à base de dados como um todo, o conjunto final obteve um total de 22 colunas e 24 linhas, ainda que pequeno, para a geração de dados com GAN não se trata de um problema, mas a presença de dados duplicados deve ser um ponto de atenção. Além disso, foi possível observar que plataformas mais distantes da costa, possuem uma profundidade maior, assim como peso total da estrutura. Além disso, observa-se um número levemente maior de cenários mapeados de risco associados à essas plataformas.

Por fim, destaca-se que três colunas obtiveram todos os valores iguais, sendo elas: Tipo de produção, Riscos Ambientais Efetivo – Grande e Risco Operacional Não Tolerável. A Figura 5 apresenta a distribuição dos dados no formato de histograma, onde, da esquerda para a direita e de cima para baixo, cada subconjunto representa os dados da respectiva coluna do conjunto conforme a ordem apresentada na Tabela 1. Destaca-se a pouca variabilidade dos dados em relação ao risco e à duração dos projetos.

Figura 5 - Distribuição dos dados por Feature

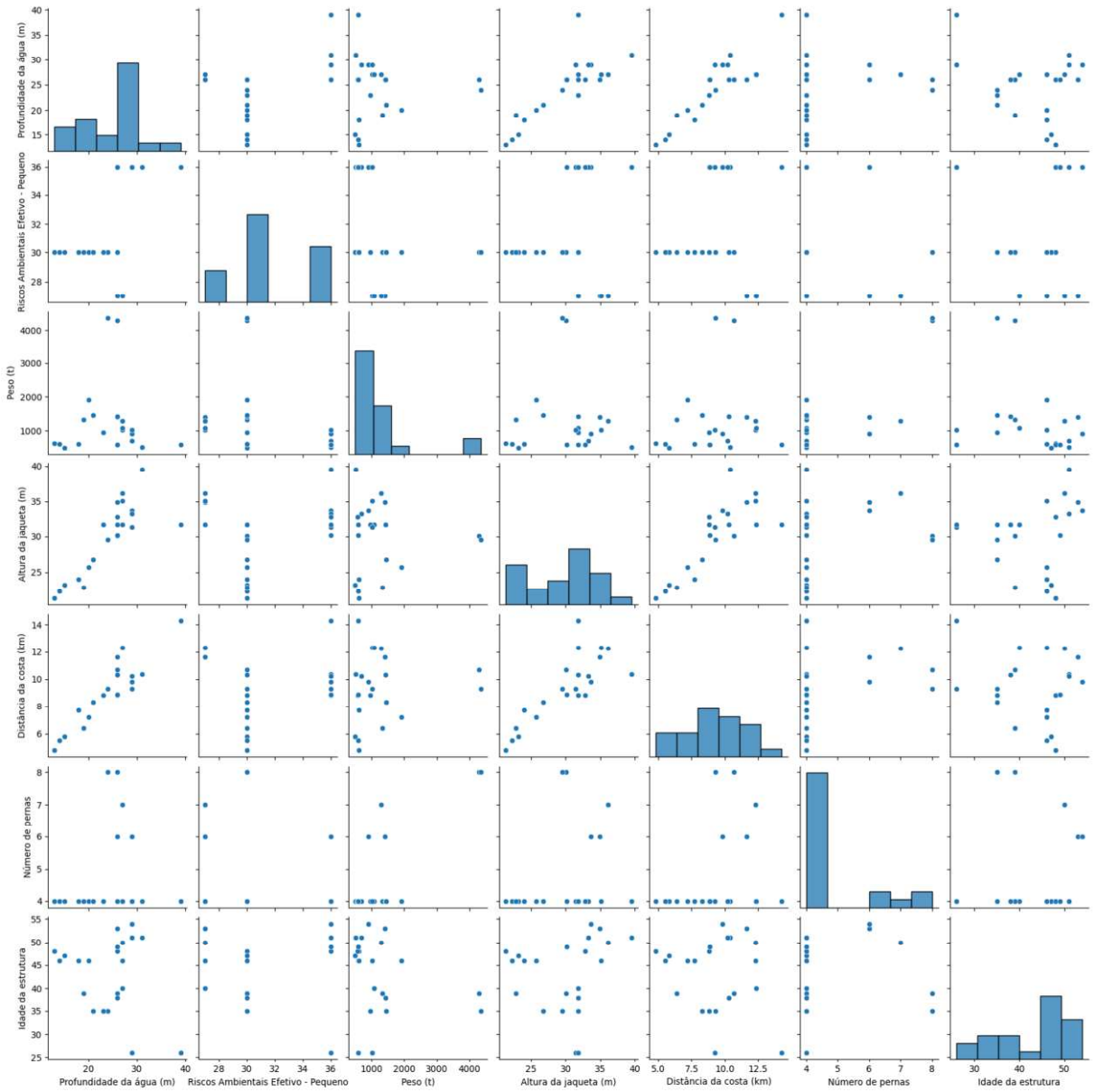


Fonte: Elaboração Própria

Para uma maior compreensão do relacionamento dos valores dessas colunas, gerou-se a Figura 6, que mostra apenas as colunas consideradas de maior relevância, em relação a distribuição dos dados e a característica da informação. As colunas selecionadas foram: Profundidade da água (m), Riscos Ambientais Efetivo – Pequeno, Peso (t), Altura da jaqueta (m), Distância da costa (km), Número de pernas e Idade da estrutura, de cima para baixo as linhas da imagem seguem essa mesma ordem de informação.

Na Figura 6 é possível observar os comportamentos citados anteriormente, onde a Profundidade da água (m) possui uma relação com a Altura da jaqueta (m) e a Distância da costa (km). Em relação ao Riscos Ambientais Efetivo – Pequeno, apesar de ele estar concentrado em um domínio contendo apendas três valores, ele parece aumentar conforme a Distância da costa (km) e a Idade da estrutura.

Figura 6 - Relação das colunas relevantes



Fonte: Elaboração Própria

3.2 TRATAMENTO DOS DADOS

Não houve a necessidade de pré-processamento muito complexo dos dados antes da rodada do modelo, além disso o funcionamento do treinamento da GAN é diferente de modelos de previsão, mais amplamente difundidos. No caso, a base não precisa ser dividida em treino e teste por exemplo, ela por completo é passada como dado de entrada ao discriminador. Fora isso, a base foi construída sem valores ausentes, logo não houve necessidade de tratamentos desse tipo.

O único tratamento necessário foi a normalização dos dados, que inicialmente foi feita para um intervalo entre zero e um utilizando o método *MinMaxScaler* da biblioteca *Scikit-learn.* A normalização é necessária uma vez que a rede neural consegue convergir mais rapidamente quando os dados estão normalizados, fora que as funções de ativação presentes nas camadas da rede, como a *Tanh* e a *Sigmoid* reduzem sua capacidade de aprendizado quando atuam muito fora dos intervalos esperados, respectivamente [-1,1] e (0,1). Por esses e outros fatores ao trabalhar com GANs deve-se sempre optar por normalizar os dados antes do processamento (RADFORD et. al, 2015).

3.3 ESCOLHA DOS HIPERPARÂMETROS INICIAIS

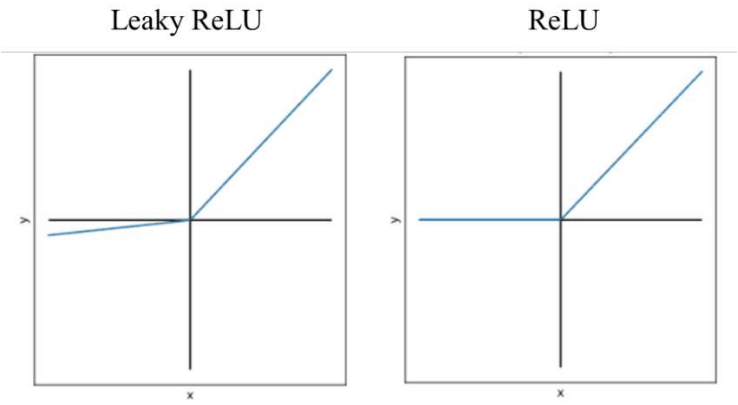
Nesse capítulo será explicado de forma mais detalhada o funcionamento de algumas funções utilizadas na metodologia. No referencial teórico foi abordado de forma geral que Redes Neurais são compostas por camadas que possuem neurônios que são modelados internamente por funções. Essas funções são chamadas de funções de ativação e existem uma infinidade disponível para uso.

Nesse estudo optou-se para as camadas intermediárias pelas opções *Rectified Linear Unit* (*ReLU*), no português Unidade Linear Retificada, com domínio de [0; ∞ [e pela versão modificada da *ReLU*, a *Leaky ReLU*, que ajusta os valores muito negativos para próximo de zero, através da aplicação de um fator de divisão, para todos os cenários foi utilizado um fator de 0,2 conforme o estudo aplicado por Alec Radford et. al (2015), que trata-se de uma escolha eficaz tanto para o gerador quanto para o discriminador. A Figura 7 ilustra o comportamento dessas funções, onde matematicamente se tem:

$$ReLU(x) = (x)^{+} = \text{Max}\,(0, x)$$

$$LeakyReLU(x) = \begin{cases} x, & x \geq 0 \\ fator\ de\ divisão * x, & x < 0 \end{cases}$$

Figura 7 - Funções de ativação camadas intermediárias



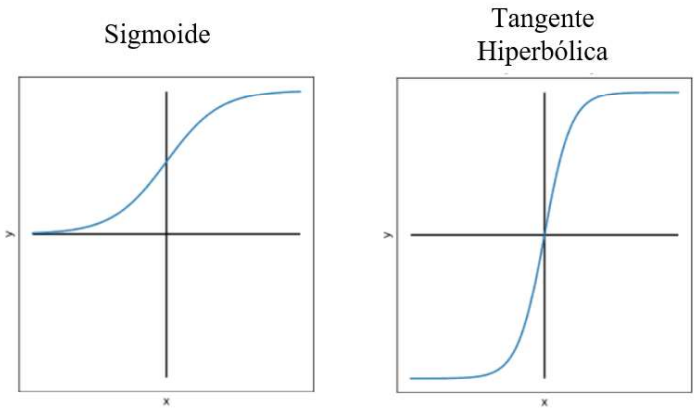
Fonte: PyTorch Documentation - Adaptado

Para as camadas de saída, foram escolhidas diferentes funções de ativação. Na Rede Geradora, optou-se pelas funções Tangente Hiperbólica e Sigmoide, enquanto na Rede Discriminadora foi utilizada apenas a função Sigmoide. Essa escolha para a Rede Discriminadora se deve ao fato de a Sigmoide ser uma função logística que direciona seus valores para os extremos, no intervalo [0,1], já em conformidade com o intervalo probabilístico da classificação de dado sintético ou real. Já na Rede Geradora, a escolha da função de ativação depende do intervalo de normalização esperado nos dados de entrada. Tanto a Tangente Hiperbólica quanto a Sigmoide apresentam mecanismos de funcionamento semelhantes, diferenciando-se apenas pelo intervalo em que atuam. Para essas funções tem-se matematicamente que:

$$\text{Sigmoid}(x) = \sigma(x) = \frac{1}{1+ e^{-x}}$$

$$\text{Tanh}(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Figura 8 - Funções de ativação camadas de saída



Fonte: PyTorch Documentation - Adaptado

Em relação à função de perda, optou-se por utilizar a função originalmente empregada por Goodfellow na primeira GAN desenvolvida, ainda que posteriormente problemas mais complexos tenham desenvolvido outros tipos de função, para uma geração tabular simples espera-se que ela atenda bem. No caso, trata-se da *Binary Cross Entropy Loss* (*BCELoss*), amplamente utilizada em tarefas de classificação binária. No caso da GAN ela calcula a diferença entre as probabilidades preditas pelo discriminador e os rótulos esperados (GOODFELLOW, 2014). Para essa função matematicamente tem-se que:

$$BCELoss = \frac{1}{N} \sum_{i=1}^{N} (y_i * \log(p_i) + (1 - y_i) * \log (1 - p_i))$$

Onde:

y_i = Rótulo real do dado, sendo 1 para verdadeiro e 0 para sintético.

p_i = Probabilidade predita pelo discriminador

N = Número de amostras

Já em relação ao otimizador escolhido, para ambas as redes, geradora e discriminadora, optou-se pelo Adaptive Moment Estimation (Adam). Esse otimizador combina as vantagens de dois outros otimizadores muito populares, o *AdaGrad* e o *RMSProp*. Entre suas vantagens destaca-se sua eficiência computacional, fácil implementação e boas propriedades de convergência. Em relação ao seu funcionamento, o otimizador utiliza estimativas dos momentos de primeira ordem (média) e segunda ordem (variância) dos gradientes para calcular taxas de aprendizado adaptativas, em outras palavras ele calcula dois valores para cada parâmetro ao longo do treinamento, sendo eles estimativas adaptativas da média e da variância dos gradientes (KINGMA et. al, 2015).

Em relação a taxa de aprendizado, a literatura aborda que ao otimizar a função de perda quanto maior a taxa de aprendizado maiores as chances de se cair em um mínimo local, em relação ao gradiente da função de perda. Nesse caso, optou-se por utilizar taxas muito pequenas para maximizar as chances de encontrar a menor perda possível.

Já em relação ao número de épocas, ao tamanho do lote percorrido em cada época e tamanho do vetor de ruído da Rede Geradora, não existe um valor consensual acerca de qual deve ser assumido. Nesse caso, para as épocas

buscou-se observar o comportamento da taxa de fidelidade dos dados ao longo de até 70.000 épocas a fim de determinar qual seria a mais adequada. Já para o vetor de ruído da Rede Geradora optou-se por não variar seu tamanho e com isso se escolheu um vetor de tamanho intermediário ao número de atributos, ou seja, de tamanho 15. Já os lotes percorridos optaram-se por variá-los de acordo com o plano de experimentação.

Por fim, em relação ao número de amostras geradas, ao longo do treinamento, buscou-se manter sempre a geração e avaliação do modelo em 100 amostras, já que a literatura também não aborda de forma clara a quantidade que deve ser gerada ao longo do treinamento e da avaliação.

3.4 PLANO DE EXPERIMENTAÇÃO

O plano de experimentação foi desenvolvido com o objetivo de identificar a combinação de hiperparâmetros que possibilitasse a Rede Adversarial Generativa produzir dados sintéticos com características mais próximas dos dados reais, tornando o modelo o mais eficiente possível, sem causar *overfitting*.

De acordo com a literatura disponível sobre GANs e Redes Neurais, levou-se em conta que dentre os hiperparâmetros disponíveis, seriam aplicados nesse estudo os mais comumente utilizados e de menor complexidade, uma vez que a base de dados reais conta com uma estrutura pouco complexa, quando comparada com gerações de imagens, ou vídeos. Além disso, outros estudos como o de Alec Radford et. al (2015), não aplicam variações nas redes discriminadoras, apenas na rede geradora, já que uma rede discriminadora muito forte pode interferir no processo de aprendizado da rede geradora, com isso optou-se por manter a arquitetura da rede discriminadora simples e estática.

Sendo assim, optou-se por variar a normalização dos dados de entrada entre [0,1] e [-1,1]. Consequentemente, a função de ativação da camada de saída da rede geradora deveria acompanhar a característica dos dados de entrada, nesse caso optou-se por variá-las usando as funções *Sigmoid* e *Tanh*. Em relação ao número de camadas, por mais que se tratasse de uma rede simples, entende-se que a quantidade de camadas poderia interferir significativamente na qualidade do modelo, por isso variou-se entre 3, 4 e 6 camadas funcionais. Vale destacar que o aumento do número de camadas fornece ao modelo capacidade de aprender

relações mais complexas, porém aumenta o risco de *overfitting*. Em relação as função de ativação dos neurônios, como entende-se que sua escolha também impactam significativamente o resultado final, optou-se por testar duas variações, *LeakyReLU(0.2)* e *ReLU()*. Por fim, sabe-se que a taxa de aprendizado pode interferir na localização do erro mínimo global, no caso, optou-se por variá-la em 0,0002 e 0,0001. A Tabela 3 apresenta cada cenário e sua variação de parâmetros.

Tabela 3 - Plano de Experimentação

Cenário	batch_ size	Normalizarion feature_range	Camada de Saída (Gerador)	Nº de Camadas	Função de Ativação (Gerador)	Taxa de Aprendizado
1	32	(0,1)	Sigmoid()	3	LeakyReLU(0.2)	0,0002
2	32	(-1,1)	Tanh()	3	LeakyReLU(0.2)	0,0002
3	24	(0,1)	Sigmoid()	3	LeakyReLU(0.2)	0,0002
4	24	(-1,1)	Tanh()	3	LeakyReLU(0.2)	0,0002
5	32	(0,1)	Sigmoid()	4	LeakyReLU(0.2)	0,0002
6	32	(-1,1)	Tanh()	4	LeakyReLU(0.2)	0,0002
7	24	(0,1)	Sigmoid()	4	LeakyReLU(0.2)	0,0002
8	24	(-1,1)	Tanh()	4	LeakyReLU(0.2)	0,0002
9	32	(0,1)	Sigmoid()	6	LeakyReLU(0.2)	0,0002
10	32	(-1,1)	Tanh()	6	LeakyReLU(0.2)	0,0002
11	24	(0,1)	Sigmoid()	6	LeakyReLU(0.2)	0,0002
12	24	(-1,1)	Tanh()	6	LeakyReLU(0.2)	0,0002
13	32	(0,1)	Sigmoid()	3	LeakyReLU(0.2)	0,0001
14	32	(-1,1)	Tanh()	3	LeakyReLU(0.2)	0,0001
15	24	(0,1)	Sigmoid()	3	LeakyReLU(0.2)	0,0001
16	24	(-1,1)	Tanh()	3	LeakyReLU(0.2)	0,0001
17	32	(0,1)	Sigmoid()	4	LeakyReLU(0.2)	0,0001
18	32	(-1,1)	Tanh()	4	LeakyReLU(0.2)	0,0001
19	24	(0,1)	Sigmoid()	4	LeakyReLU(0.2)	0,0001
20	24	(-1,1)	Tanh()	4	LeakyReLU(0.2)	0,0001
21	32	(0,1)	Sigmoid()	6	LeakyReLU(0.2)	0,0001
22	32	(-1,1)	Tanh()	6	LeakyReLU(0.2)	0,0001
23	24	(0,1)	Sigmoid()	6	LeakyReLU(0.2)	0,0001
24	24	(-1,1)	Tanh()	6	LeakyReLU(0.2)	0,0001
25	32	(0,1)	Sigmoid()	3	ReLU()	0,0002
26	32	(-1,1)	Tanh()	3	ReLU()	0,0002
27	24	(0,1)	Sigmoid()	3	ReLU()	0,0002
28	24	(-1,1)	Tanh()	3	ReLU()	0,0002
29	32	(0,1)	Sigmoid()	4	ReLU()	0,0002
30	32	(-1,1)	Tanh()	4	ReLU()	0,0002
31	24	(0,1)	Sigmoid()	4	ReLU()	0,0002
32	24	(-1,1)	Tanh()	4	ReLU()	0,0002
33	32	(0,1)	Sigmoid()	6	ReLU()	0,0002
34	32	(-1,1)	Tanh()	6	ReLU()	0,0002
35	24	(0,1)	Sigmoid()	6	ReLU()	0,0002

Cenário	batch_ size	Normalizarion feature_range	Camada de Saída (Gerador)	Nº de Camadas	Função de Ativação (Gerador)	Taxa de Aprendizado
36	24	(-1,1)	Tanh()	6	ReLU()	0,0002
37	32	(0,1)	Sigmoid()	3	ReLU()	0,0001
38	32	(-1,1)	Tanh()	3	ReLU()	0,0001
39	24	(0,1)	Sigmoid()	3	ReLU()	0,0001
40	24	(-1,1)	Tanh()	3	ReLU()	0,0001
41	32	(0,1)	Sigmoid()	4	ReLU()	0,0001
42	32	(-1,1)	Tanh()	4	ReLU()	0,0001
43	24	(0,1)	Sigmoid()	4	ReLU()	0,0001
44	24	(-1,1)	Tanh()	4	ReLU()	0,0001
45	32	(0,1)	Sigmoid()	6	ReLU()	0,0001
46	32	(-1,1)	Tanh()	6	ReLU()	0,0001
47	24	(0,1)	Sigmoid()	6	ReLU()	0,0001
48	24	(-1,1)	Tanh()	6	ReLU()	0,0001

Fonte: Elaboração Própria

3.5 EXECUÇÃO DA EXPERIMENTAÇÃO

Como não foi encontrada nenhuma biblioteca que rodasse de forma automática os cenários no formato de arquitetura de uma GAN, foi necessário aplicar manualmente as variações para posteriormente analisá-las. Dessa forma, os parâmetros de cada cenário foram alterados na arquitetura da GAN e avaliados de acordo com as métricas de avaliação da biblioteca *SDMetrics.*

3.6 MÉTRICAS DE AVALIAÇÃO

Como mencionado anteriormente, a biblioteca *SDMetrics* é usada para avaliar a qualidade da GAN, uma vez que possui ferramentas específicas para comparar o volume de dados gerados com os dados reais. Para tal utilizou-se o módulo *QualityReport* e o módulo *visualization.*

Para obter o *QualityReport* é necessário fornecer três parâmetros: os dados reais, os dados gerados e um metadado. O metadado consiste em um dicionário que deve listar para cada coluna do conjunto de dados, seu nome, se é numérica ou categórica e em caso de numérica se é inteiro ou ponto flutuante. Dentre os atributos e métricas disponíveis no relatório, optou-se apenas por utilizar a *Column Shapes*, a *Column Pair Trends* e *a get_score*.

A *Column Shapes*, através da estatística Kolmogorov-Smirnov (KS), na função *KSComplement* da biblioteca, calcula a taxa de semelhança entre os dados da coluna real e da coluna sintética, em relação ao formato, ou seja, apenas em relação a distribuição marginal. Após a comparação ser efetuada, um valor percentual é atribuindo, nesse caso, quanto mais próximo de 0 menor a similaridade.

Já para a *Column Pair Trends,* o grau de correlação para cada coluna da base de dados é calculado a partir da função CorrelationSimilarity, onde para cada possível par de colunas dois coeficientes são avaliados: o coeficiente de Pearson, que mede a correlação linear entre as variáveis, e o coeficiente de Spearman, que mede o relacionamento monotônico entre elas. Por fim, uma ponderação é feita para todas as medidas calculadas e um valor percentual absoluto é obtido para cada coluna, nesse caso quanto mais próximo de 1, mais a correlação dos dados sintéticos se aproxima com os dados reais.

Por fim a função *get_score* retorna uma métrica chamada de *Overall Score* que retrata a taxa e fidelidade absoluta, ou taxa de acerto, dos dados sintéticos em relação aos dados reais. Essa métrica é calculada através de uma ponderação entre a média geral das duas métricas apresentadas anteriormente.

Em relação ao módulo *visualization,* para utilizá-lo basta fornecer o conjunto de dados reais e sintético e o tipo de gráfico que deseja visualizar. O módulo será capaz de gerar excelentes imagens, com alto poder ilustrativo, que permite ao usuário compreender e visualizar a qualidade dos dados gerados.

3.7 ALGORÍTIMO ESCOLHIDO

Dentre todos os cenários gerados no plano de experimentação, comparou-se os resultados de forma absoluta através do *Overall Score.* A Tabela *4* apresenta para cada cenário a Taxa de fidelidade absoluta, ou taxa de acerto geral, por época para cada cenário testado, a fim de facilitar a comparação, uma escala de cor foi aplicada, onde tons próximos do branco representam taxas menores e próximos do verde escuro representam taxas maiores. Como é possível observar na Tabela *4*, de forma geral, para épocas maiores, obtiveram-se resultados melhores, em relação a taxa de aprendizado, cenários com taxa de aprendizado menor (0,0001) obtiveram resultado levemente melhor se comparado com os demais. Por fim,

destaca-se que cenários com a função de ativação *ReLU*, concentrados no final da tabela, também aparentam ter uma maior taxa de fidelidade se comparado com os demais.

Tabela 4 - Taxa de acerto geral por cenário e por época

Cenário	Avaliação	500	1000	2000	5000	10000	15000	25000	30000	50000	60000	70000
1	**Overall Score %**	42,2%	37,4%	46,4%	58,5%	58,8%	61,5%	64,9%	64,2%	64,2%	64,1%	63,4%
2	**Overall Score %**	52,3%	54,4%	58,7%	58,4%	57,1%	57,7%	63,2%	61,7%	61,8%	61,1%	58,3%
3	**Overall Score %**	42,9%	51,8%	55,3%	59,0%	58,2%	58,3%	64,8%	64,0%	64,0%	64,0%	63,2%
4	**Overall Score %**	42,9%	51,8%	55,3%	59,0%	58,2%	58,3%	64,8%	64,0%	64,0%	64,0%	63,2%
5	**Overall Score %**	39,2%	37,0%	44,1%	57,3%	60,6%	61,6%	66,1%	66,1%	65,4%	65,9%	68,4%
6	**Overall Score %**	41,9%	43,5%	50,7%	53,8%	57,4%	57,3%	59,4%	60,8%	61,5%	62,6%	60,9%
7	**Overall Score %**	40,0%	37,6%	43,0%	53,5%	58,9%	59,2%	64,5%	64,0%	64,9%	64,8%	64,4%
8	**Overall Score %**	37,9%	40,2%	49,9%	60,3%	60,7%	60,5%	62,9%	63,5%	63,9%	65,6%	67,2%
9	**Overall Score %**	35,8%	39,6%	42,5%	53,3%	60,5%	58,2%	59,2%	60,8%	64,2%	64,0%	64,8%
10	**Overall Score %**	41,7%	39,6%	46,6%	44,8%	58,6%	61,2%	61,3%	60,6%	58,5%	57,6%	56,0%
11	**Overall Score %**	37,5%	43,1%	50,4%	59,4%	60,4%	57,5%	59,9%	61,9%	62,3%	64,1%	64,0%
12	**Overall Score %**	46,8%	47,4%	48,1%	55,5%	59,5%	59,9%	58,1%	59,0%	63,0%	61,1%	63,1%
13	**Overall Score %**	48,4%	48,6%	49,7%	56,6%	61,9%	64,3%	**67,2%**	67,5%	65,9%	66,5%	65,8%
14	**Overall Score %**	50,8%	55,4%	58,0%	59,2%	61,7%	63,8%	**67,1%**	65,3%	67,7%	67,4%	69,6%
15	**Overall Score %**	50,3%	51,1%	51,8%	57,9%	61,6%	60,7%	64,8%	66,0%	63,8%	63,5%	62,6%
16	**Overall Score %**	53,4%	56,7%	60,1%	58,7%	62,1%	62,6%	64,7%	62,9%	65,3%	63,7%	59,5%
17	**Overall Score %**	39,8%	40,4%	42,7%	60,2%	64,4%	**65,6%**	**67,3%**	65,1%	68,0%	68,0%	66,4%
18	**Overall Score %**	43,5%	49,1%	52,7%	58,4%	61,2%	61,2%	**68,9%**	**69,3%**	**70,8%**	**70,4%**	**69,8%**
19	**Overall Score %**	40,2%	37,5%	40,3%	46,9%	45,5%	54,6%	56,2%	55,7%	54,9%	56,5%	57,2%
20	**Overall Score %**	42,2%	49,6%	54,0%	61,0%	64,6%	**66,5%**	**68,3%**	66,0%	68,5%	68,1%	69,1%
21	**Overall Score %**	37,4%	36,3%	37,4%	49,6%	46,8%	53,4%	65,5%	66,6%	67,3%	66,0%	68,0%
22	**Overall Score %**	40,6%	47,3%	51,4%	58,3%	62,5%	63,5%	64,5%	63,9%	65,4%	66,4%	67,5%
23	**Overall Score %**	38,5%	40,0%	48,9%	49,9%	58,9%	56,8%	60,1%	63,5%	67,1%	66,3%	65,8%
24	**Overall Score %**	50,5%	45,5%	49,9%	51,6%	60,1%	59,5%	66,5%	66,3%	66,3%	66,3%	63,9%
25	**Overall Score %**	40,2%	43,8%	54,0%	60,1%	60,8%	62,2%	63,2%	63,9%	64,4%	64,1%	63,6%
26	**Overall Score %**	46,5%	51,8%	52,2%	60,5%	61,1%	61,5%	65,1%	65,7%	64,8%	64,1%	64,6%
27	**Overall Score %**	40,6%	37,0%	42,8%	54,6%	54,6%	54,8%	55,7%	56,3%	57,8%	56,7%	55,4%
28	**Overall Score %**	51,9%	51,6%	56,2%	58,2%	59,5%	60,4%	65,5%	64,7%	62,3%	61,7%	61,0%
29	**Overall Score %**	36,5%	39,3%	46,7%	59,8%	58,4%	63,3%	63,4%	63,8%	63,3%	60,6%	61,6%
30	**Overall Score %**	48,1%	48,9%	52,5%	55,2%	57,9%	60,1%	61,1%	64,3%	64,4%	63,2%	64,1%
31	**Overall Score %**	37,9%	40,2%	49,9%	60,3%	60,7%	60,5%	62,9%	63,5%	63,9%	65,6%	67,2%
32	**Overall Score %**	44,3%	44,1%	46,8%	47,1%	45,2%	52,2%	40,7%	42,9%	41,8%	42,2%	42,7%
33	**Overall Score %**	44,3%	44,1%	46,8%	47,1%	45,2%	52,2%	40,7%	42,9%	41,8%	42,2%	42,7%
34	**Overall Score %**	45,3%	49,9%	50,3%	50,3%	58,1%	54,1%	51,7%	56,1%	61,9%	65,0%	63,6%
35	**Overall Score %**	38,5%	45,3%	51,5%	51,9%	61,4%	56,6%	64,5%	66,3%	63,2%	64,9%	60,5%
36	**Overall Score %**	39,4%	39,9%	42,8%	43,1%	43,9%	46,7%	49,0%	55,8%	59,3%	59,1%	59,0%
37	**Overall Score %**	48,1%	39,9%	45,3%	56,4%	**65,0%**	64,9%	**67,8%**	**67,3%**	66,1%	66,0%	66,8%
38	**Overall Score %**	51,9%	50,5%	54,3%	59,2%	62,5%	**65,1%**	**67,7%**	66,7%	66,2%	67,8%	67,2%
39	**Overall Score %**	47,6%	46,5%	48,6%	56,2%	**65,8%**	**66,8%**	66,8%	66,7%	66,7%	68,1%	66,9%

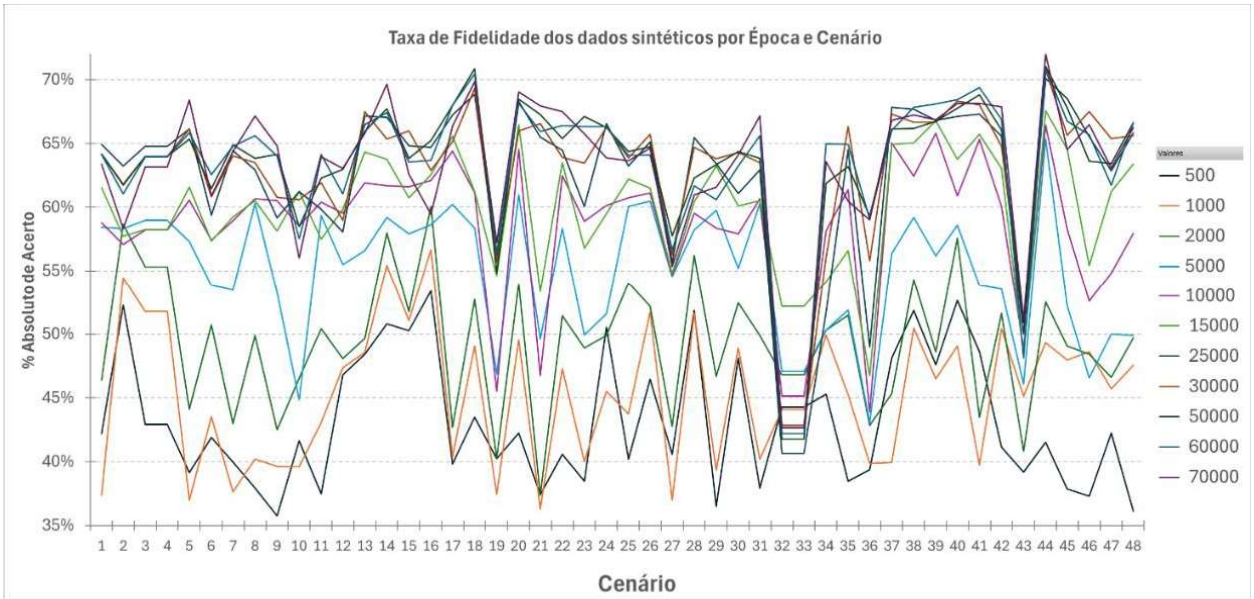
Cenário	Avaliação	500	1000	2000	5000	10000	15000	25000	30000	50000	60000	70000
40	**Overall Score %**	52,7%	49,1%	57,6%	58,6%	60,9%	63,8%	**67,1%**	**68,3%**	**67,7%**	**68,5%**	**68,1%**
41	**Overall Score %**	48,6%	39,7%	43,5%	53,9%	**65,3%**	**65,7%**	**67,3%**	**68,0%**	**68,8%**	**69,4%**	**68,2%**
42	**Overall Score %**	41,2%	50,4%	51,6%	53,6%	60,1%	63,1%	65,6%	64,8%	66,1%	66,9%	67,9%
43	**Overall Score %**	39,2%	45,1%	40,8%	46,1%	48,9%	48,4%	50,0%	51,2%	50,9%	48,1%	51,1%
44	**Overall Score %**	41,5%	49,3%	52,5%	65,4%	**66,4%**	**67,6%**	**70,1%**	**70,8%**	**71,0%**	**70,9%**	**72,0%**
45	**Overall Score %**	37,9%	48,0%	49,1%	52,2%	58,3%	64,6%	**68,5%**	65,7%	67,8%	66,7%	64,5%
46	**Overall Score %**	37,3%	48,6%	48,4%	46,6%	52,6%	55,5%	65,5%	67,5%	63,6%	65,7%	66,5%
47	**Overall Score %**	42,2%	45,7%	46,6%	50,0%	54,8%	61,2%	62,9%	65,4%	63,4%	61,7%	63,1%
48	**Overall Score %**	36,1%	47,6%	49,7%	49,9%	58,0%	63,4%	65,9%	65,6%	66,6%	66,7%	66,3%

Fonte: Elaboração Própria

Nesse contexto, os cenários que mais se destacaram foram os de número: 18, 40, 41, 44. Ainda assim, seria necessário definir o número ótimo de épocas a serem percorridas pelo modelo a fim de detalhar a qualidade dos dados gerados no fim desse percurso. Nesse caso, gerou-se a Figura 9, onde é possível observar que à medida que o úmero de épocas aumenta, a taxa de fidelidade também cresce. No entanto, após ultrapassar 10.000 épocas, a taxa passa a crescer em uma velocidade muito menos significativa. Ainda assim, não fica claro se a análise deveria ser feita considerando 10.000, 15.000 ou 25.000 épocas. Destaca-se também que os cenários não foram capazes de ultrapassar os 72% de fidelidade geral.

Ainda com a Figura 9 é possível identificar que, diferente dos demais, os cenários de 13 a 16, que compreendem aqueles com 3 camadas, função de ativação *LeakyReLU*(0.2) e taxa de aprendizado de 0,0001, tiveram um comportamento positivo em todo conjunto de épocas.

Figura 9 - Taxa de Fidelidade dos dados gerados por Época e Cenário



Fonte: Elaboração Própria

A fim de complementar a análise e definir o melhor cenário e melhor época, analisou-se as métricas *Column Pair Trends* e *Column Shapes* separadamente, para cada cenário de cada época com potencial levantado nas etapas anteriores. A Tabela 5 apresenta, de forma resumida, os melhores cenários encontrados para cada época analisada. Destaca-se que ao analisar de forma detalhada, o cenário 44 deixou de ser o de maior desempenho, ou seja, ao analisar o desempenho do modelo por atributo, outros cenários se destacaram mais.

Tabela 5 - Comparação dos melhores resultados de cada época

	10.000 Épocas		15.000 Épocas		25.000 Épocas	
	Cenário 41		Cenário 41		Cenário 13	
	Column Pair Trends	Column Shapes	Column Pair Trends	Column Shapes	Column Pair Trends	Column Shapes
Altura da jaqueta (m)	91,0%	46,8%	86,9%	41,4%	81,5%	36,8%
Distância da costa (km)	90,5%	55,5%	87,6%	46,2%	86,9%	32,8%
Idade da estrutura	72,2%	49,9%	72,2%	68,9%	83,0%	69,6%
Idade de parada definitiva de produção	69,7%	72,2%	69,9%	72,6%	81,2%	74,4%
Impacto Socioeconômico Efetivo - Grande	64,2%	62,2%	66,2%	69,2%		81,8%
Impacto Socioeconômico Efetivo - Médio	88,3%	50,0%	90,9%	50,0%	92,0%	50,0%
Impacto Socioeconômico Efetivo - Pequeno	80,2%	53,2%	78,3%	56,2%	78,1%	81,8%
Impacto Socioeconômico Potencial - Grande	91,0%	50,0%	93,3%	50,0%	77,0%	50,0%
Impacto Socioeconômico Potencial - Médio	95,3%	50,0%	98,8%	50,0%	85,8%	50,0%
Impacto Socioeconômico Potencial - Pequeno	77,7%	50,0%	79,0%	50,0%	87,0%	50,0%
Número de pernas	90,5%	57,7%	90,7%	67,7%	66,4%	77,3%
Peso (t)	83,0%	60,9%	84,3%	65,1%	71,7%	68,1%
Profundidade da água (m)	84,6%	57,9%	80,9%	40,9%	81,5%	34,8%
Risco Operacional Moderado	89,8%	50,0%	91,9%	50,0%	88,0%	50,0%
Risco Operacional Não Tolerável						
Risco Operacional Tolerável	84,6%	50,0%	87,0%	50,0%	84,1%	50,0%
Riscos Ambientais Efetivo - Grande						
Riscos Ambientais Efetivo - Média	81,4%	50,0%	76,4%	50,0%	85,0%	50,0%
Riscos Ambientais Efetivo - Pequeno	60,2%	27,2%	68,0%	41,8%	56,6%	68,2%
Riscos Ambientais Potencial - Grande	83,2%	50,0%	84,4%	50,0%	90,1%	50,0%
Riscos Ambientais Potencial - Média	79,7%	28,2%	71,6%	19,2%	74,9%	63,8%
Riscos Ambientais Potencial - Pequeno	78,1%	67,8%	75,0%	68,2%	86,5%	68,2%
Tipo de produção		86,0%		96,0%		100,0%

Fonte: Elaboração Própria

Observa-se na Tabela 5, que à medida que as épocas vão aumentando, ainda que o percentual de assertividade aumente, algumas métricas deixam de ser exibidas, isso ocorre uma vez que o modelo aprende o comportamento esperado pelo discriminador e passa a gerar apenas um valor para aquela coluna. Nesse caso, a rede geradora perde o poder generalista, que não é nosso interesse. Além

disso, houve colunas que para todas as três épocas observadas, os valores das métricas não puderam ser calculados, sendo esse outro fator que reforça o indício de que a baixa variabilidade de certas colunas atrapalhou o desempenho do modelo. Com isso, a fim de maximizar a taxa de fidelidade por atributo e minimizar as chances de *overfitting*, optou-se por utilizar 15.000 épocas, e consequentemente o cenário 41.

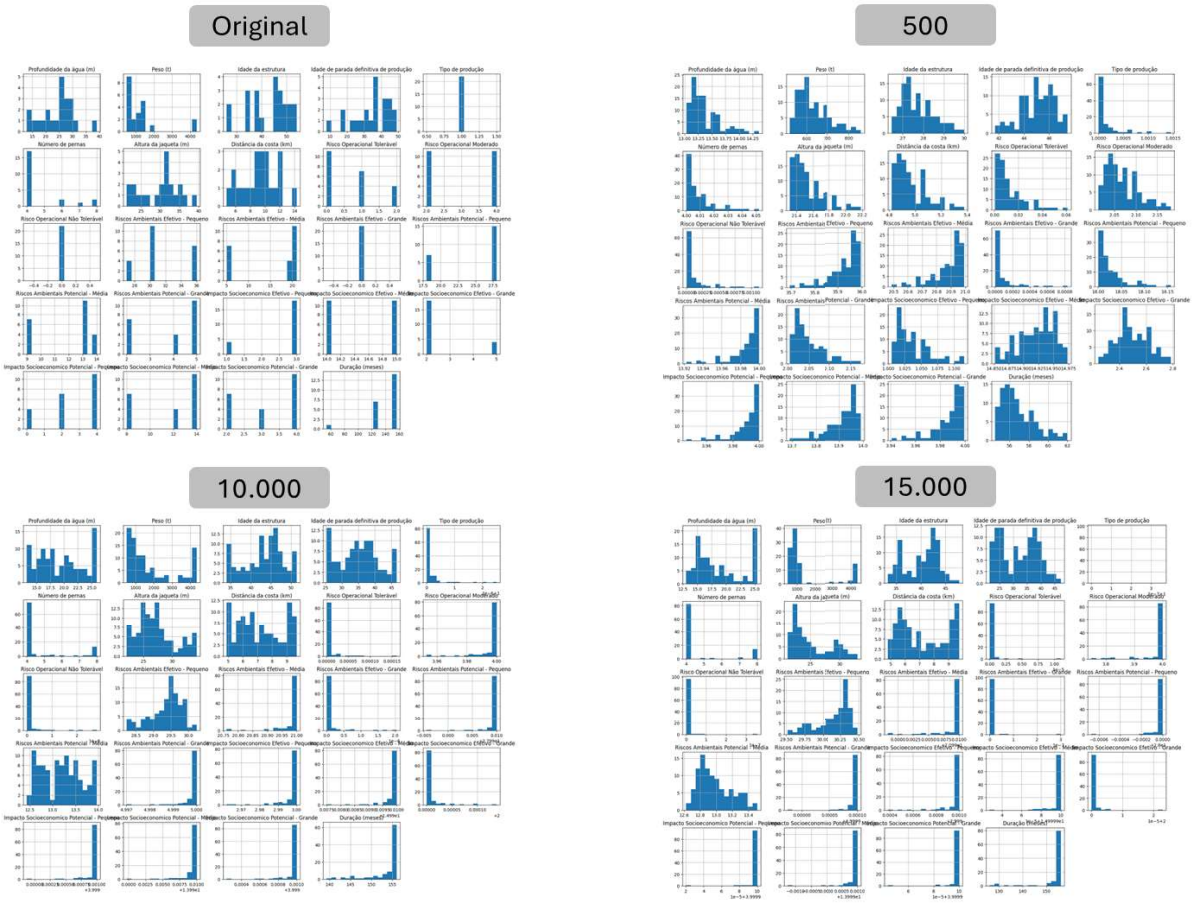
Sendo assim, a rede geradora utilizada nesse estudo foi criada conforme o cenário 41, com 4 camadas, sendo uma camada de entrada, duas ocultas e uma de saída, nas camadas de entrada e ocultas foi utilizada a função de ativação *ReLU*, com normalização dos dados de entrada no intervalo (0,1) e por isso na camada de saída foi utilizada a função *Sigmoid*. Por fim, a taxa de aprendizado foi de 0,0001.

4 RESULTADO

Ao rodar a Rede Adversarial Generativa utilizando a arquitetura da rede geradora selecionada anteriormente, observa-se, ao longo das épocas, uma progressiva melhoria na capacidade de geração de dados. Essa evolução reflete-se na aproximação, para a maioria das colunas, entre a distribuição dos dados gerados e dos dados reais, indicando um aumento na eficácia do modelo em replicar as características originais do conjunto de dados.

Ainda assim, como é possível observar na Figura 10, ao atingir 15.000 épocas, o modelo passa a escolher valores específicos a serem gerados, isso ocorre principalmente para colunas com pouca variabilidade nos dados, como as colunas de risco.

Figura 10 - Comparação dos dados gerados para algumas épocas

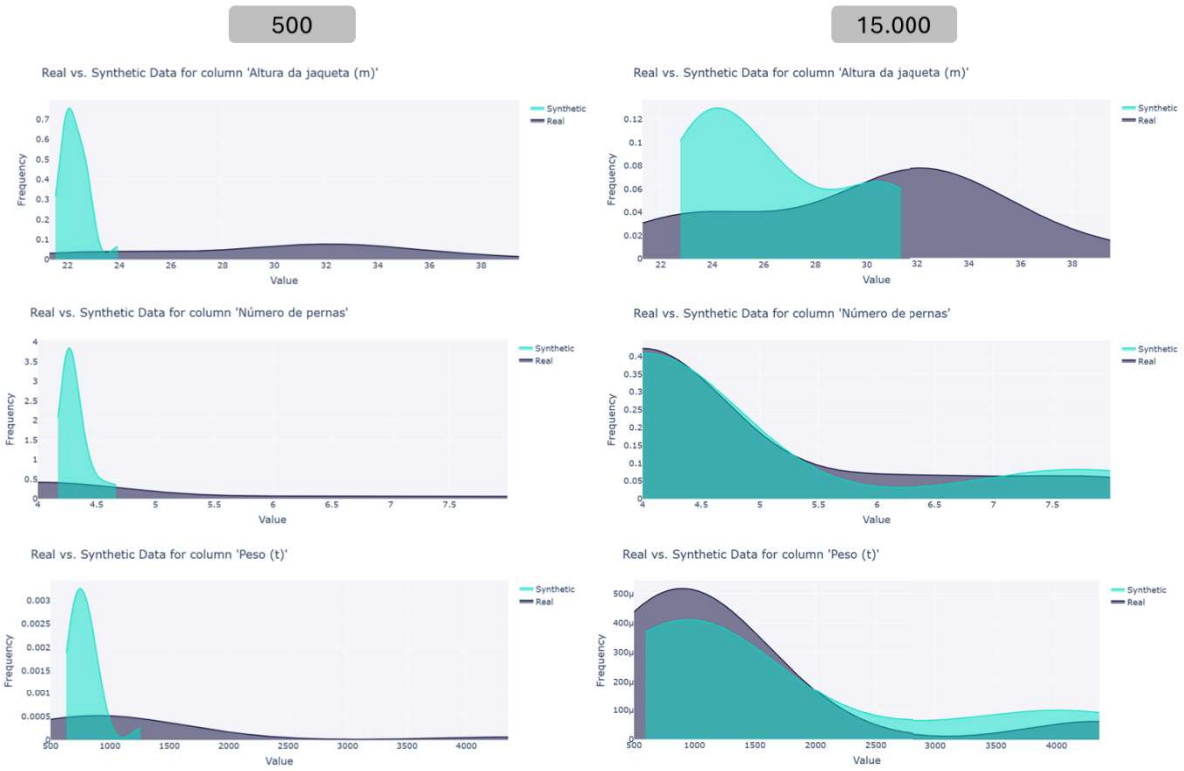


Fonte: Elaboração Própria

Em relação à distribuição marginal dos dados, a Figura 11 ilustra a evolução do modelo em relação ao intervalo de dados gerados e a similaridade das curvas para as colunas: Altura da Jaqueta (m), Número de pernas e Peso (t), por conta da

melhor qualidade dos dados de entrada dessas colunas. Destaca-se que nem em todos os casos o modelo foi capaz de gerar dados marginais com curvas sobrepostas o suficiente, ainda que ao longo do treinamento ele tenha sido capaz de melhorar sua eficácia. Ademais, observa-se aqui, para a coluna Altura da Jaqueta (m), o fenômeno citado no referencial teórico *mode colapse*, que é a limitação da geração a um subconjunto de dados.

Figura 11 - Comparação de algumas distribuições marginais



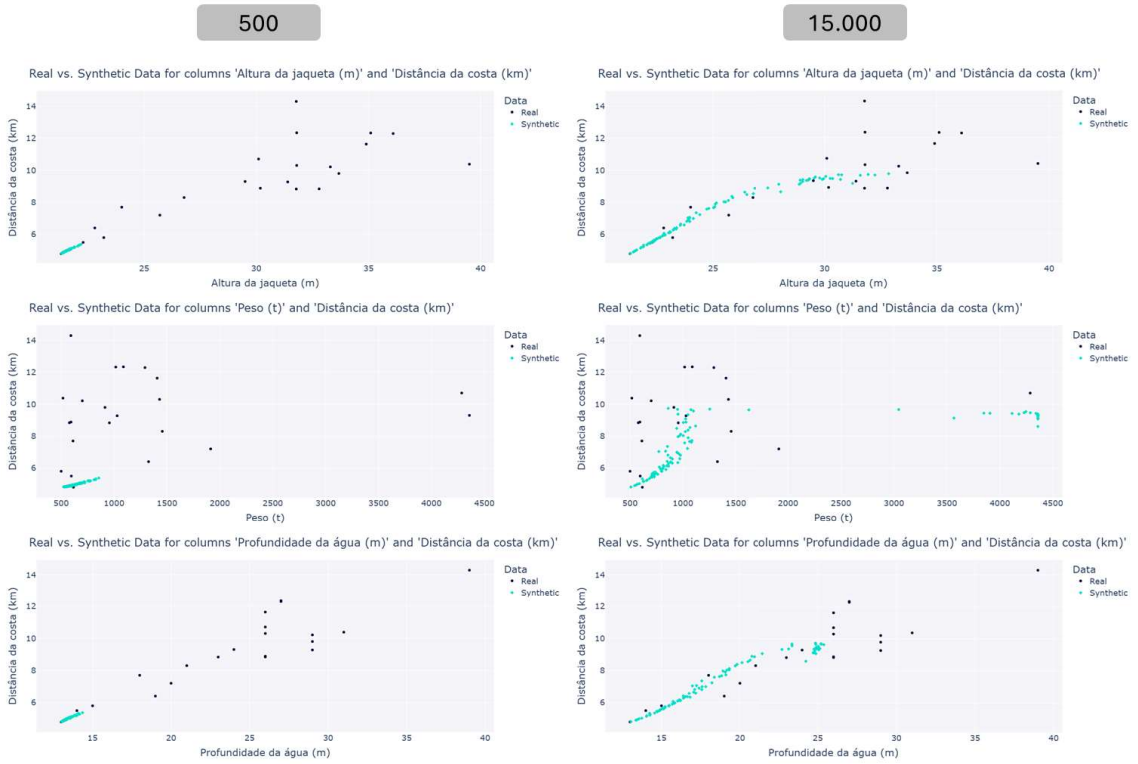
Fonte: Elaboração Própria

Em relação à correlação dos dados, a Figura 12 ilustra para alguns pares de colunas a evolução da correlação dos dados gerados. Observa-se que o poder da GAN em correlacionar os dados aumentou significativamente ao passar das épocas. A rede foi capaz de respeitar as relações existentes nos dados anteriores.

Por fim, a Figura 13 ilustra a evolução da função de perda, tanto para a rede geradora (curva azul) quanto para a rede discriminadora (curva laranja). Observa-se nesse tópico que a rede foi capaz de convergir, indicando a capacidade de aprendizado da rede. Além disso, destaca-se o comportamento atípico, onde a rede discriminadora inicia a disputa perdendo o combate, enquanto a rede geradora consegue enganá-la. Ao passar de poucas épocas, a rede discriminadora logo

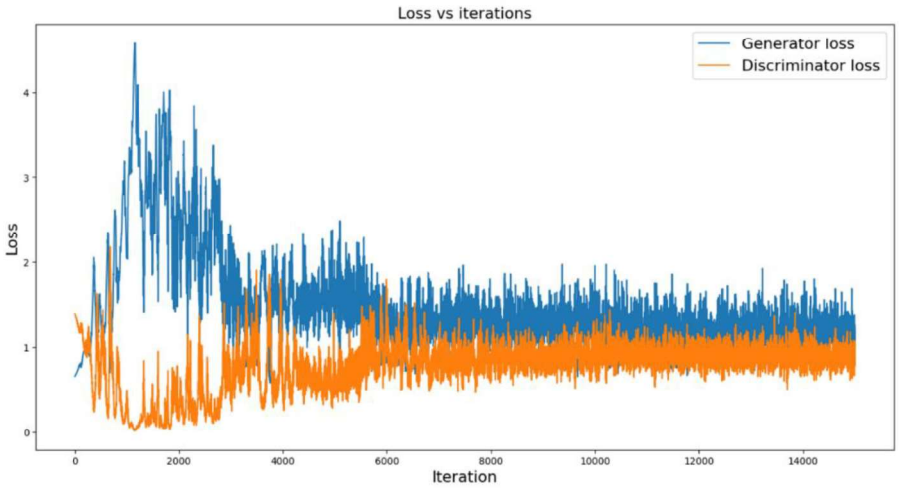
aprende a verdadeira distribuição dos dados e reinicia o embate. Aproximadamente com 3.000 épocas a rede geradora parece passar por um grande processo de aprendizado, atingindo o início da fase de convergência próximo às 6.000 épocas até finalmente a rede chegar na convergência final centrada em 1, conforme o esperado.

Figura 12 - Comparação da correlação dos dados gerados de algumas colunas



Fonte: Elaboração Própria

Figura 13 - Função de perda ao longo das épocas



Fonte: Elaboração Própria

5 ANÁLISE DOS RESULTADOS

Conforme observado na Tabela 5, Rede Adversarial Generativa do cenário 41 no intervalo de 15.000 épocas foi capaz de gerar dados robustos e suficientemente variados para a maioria das colunas. A rede foi capaz de convergir antes de atingir 15.000 interações no treinamento e tanto sua eficiência em relação à correlação, quanto em relação à distribuição marginal os dados foi relativamente aceitável, ficando próxima dos 70% ou ultrapassando em alguns casos.

Entende-se a necessidade de cautela ao trabalhar com uma base de dados pouco variada, como no caso das plataformas aqui presentes, visto que as chances de *overfitting* do modelo eram elevadas. Ademais, considerando intervalos de épocas mais elevadas o poder generativo da rede reduziu drasticamente, ainda que o percentual da taxa de fidelidade se mantivesse em crescimento.

Em relação às medidas de avaliação do cenário escolhido, a Tabela 6 apresenta uma medida geral, apenas do cenário vencedor, das duas métricas analisadas na Tabela 5. É possível observar que ainda que após a época 15.000 houve um leve declínio das taxas, seguidos de uma retomada no crescimento. Esse comportamento pode estar atrelado ao início da redução da diversidade dos dados, relacionado ao *overfitting*, que corrobora para a escolha da época 15.000 como ponto ótimo desse modelo.

Tabela 6 - Resumo do resultado por métrica e época

	500	1000	2000	5000	10000	15000	25000	30000	50000	60000	70000
Column Shapes %	27,1%	24,2%	24,2%	35,1%	49,2%	50,5%	55,6%	56,5%	56,8%	58,6%	55,9%
Column Pair Trends %	70,1%	55,3%	62,8%	72,6%	81,5%	80,9%	79,0%	79,5%	80,8%	80,2%	80,4%
Overall Score %	48,6%	39,7%	43,5%	53,9%	65,3%	65,7%	67,3%	68,0%	68,8%	69,4%	68,2%

Fonte: Elaboração Própria

6 CONCLUSÃO

O descomissionamento de plataformas é um processo crucial no cenário de exploração offshore, principalmente quando atrelado ao contexto de segurança e econômico. Ainda assim, seu planejamento envolvendo tecnologias mais robustas ocorre de maneira falha por conta da escassez de dados. Nesse contexto, esse trabalho apresentou o uso das Redes Adversariais Generativas como uma alternativa promissora.

Ao explorar a geração de dados sintéticos através do uso de GANs, o modelo mostrou-se capaz de manter uma taxa de fidelidade geral, no cenário e época escolhidos, de 65,7%. Apesar de não ser tão alto, notou-se que individualmente, para colunas com dados de entrada mais adequados, o poder generativo do modelo ultrapassou taxas de 80%. Nesse ponto, fica-se claro que em alguns momentos a baixa variabilidade em algumas colunas e o tamanho reduzido do conjunto de dados limitaram a capacidade de generalização do modelo.

Ainda assim, entende-se que, de maneira geral, esse estudo consegue contribuir de forma ativa na compreensão do uso de GANs em dados tabulares, demonstrando a viabilidade de sua aplicação no contexto de plataformas offshore. Dado que o modelo foi capaz de gerar dados condizentes para alguns atributos. De forma geral, a rede se mostrou capaz de aprender tanto os padrões de correlação dos dados, como os de distribuição ao longo do período de treinamento.

Para estudos futuros, sugere-se abordagens mais dinâmicas, especialmente ao lidar com bases de dados reduzidas, como a utilizada neste estudo. Uma alternativa seria a aplicação de técnicas de *data augmentation* no pré-processamento, permitindo a ampliação da diversidade dos dados e potencialmente melhorando o desempenho geral do modelo. Além disso seria interessante testar outras variações de vetores de ruído e amostras geradas, além de comparar a GAN com outras técnicas existentes. Por fim, seria relevante propor uma reformulação no formato dos documentos do Programa de Descomissionamento de Instalações (PDI), de modo que os riscos fossem discriminados por plataforma, em vez de apresentados apenas por bacia. Essa mudança enriqueceria a base de dados e possibilitaria análises mais detalhadas e específicas para cada instalação, podendo inclusive diminuir os problemas causados pela baixa variabilidade nos dados de risco.

7 REFERÊNCIAS

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS (ANP). ***Resolução nº 817,*** *de 24 de abril de 2020*. Diário Oficial da União: Seção 1, Brasília, DF, 27 abr. 2020.

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS (ANP). **Boletim da Produção de Petróleo e Gás Natural – Julho/2024**. Disponível em:https://app.powerbi.com/view?r=eyJrIjoiZjZhZDliMTYtOWIyZi00OGY5LWJkYzItOTQ1MzFjZGMzMDNkIiwidCI6IjQ0OTlmNGZmLTI0YTYtNGI0Mi1iN2VmLTEyNGFmY2FkYzkxMyJ9. Acesso em: 10 nov. 2024.

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS (ANP). **Painel Dinâmico da ANP**. Disponível em: https://app.powerbi.com/view?r=eyJrIjoiZjFlMWI0MDgtNWNiNC00OTZlLWI3NGQtOGM3MjQwODhjMTMwIiwidCI6IjQ0OTlmNGZmLTI0YTYtNGI0Mi1iN2VmLTEyNGFmY2FkYzkxMyJ9. Acesso em: 10 nov. 2024.

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS (ANP). **Resolução nº 27,** de 18 de outubro de 2006. Diário Oficial da União: seção 1, Brasília, DF, 20 out. 2006.

AGGARWAL, Alankrita; MITTAL, Mamta; BATTINENI, Gopi. **Generative adversarial network: An overview of theory and applications**. International Journal of Information Management Data Insights, v. 1, n. 1, p. 100004, 2021.

AMAZON WEB SERVICE (AWS) **O que são dados sintéticos?** Disponível em: <https://aws.amazon.com/pt/what-is/synthetic-data/>. Acesso em: 15 set. 2024.

ARAUJO, Carlos Vinicio D.; OLIVEIRA DA SILVA, Jailon William B.; SOARES, Pablo Luiz Braga; VIDAL, Priscila da Cunha Jacome. **Aplicação de Redes Adversariais Generativas na Geração de Dados para Plataformas Offshore a Serem Descomissionadas.** Anais do Simpósio Brasileiro de Pesquisa Operacional (SBPO), 2024

BATISTA, Márcio; MARTINS, A. **Redes Neurais no Ensino Básico**. Revista Eletrônica da Sociedade Brasileira de Matemática, v. 10, n. 4, p. 454-481, 2022.

BUI, N. A.; OH, Y. G.; LEE, I. P. Improving the accuracy of an oil spill detection and classification model with fake datasets. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, v. X-1/W1, p. 51-56, 2023.

CAMPOS, Afonso Teberga *et al*. **Redes adversárias generativas**: uma alternativa para modelagem de dados de entrada em projetos de simulação. 2022.

CHEMUDUPATI, P.; MARSHALL, H.; RAI, S. A Dynamic Machine Learning Model for Accelerated Oil Spill Remediation. Journal of Environmental Technology, v. 58, n. 2, p. 45-60, 2024.

ENERGY INSTITUTE (EI). **Statistical Review of World Energy 2023.** Disponível em https://www.energyinst.org/statistical-review/resources-and-data-downloads. Acesso em: 12 nov 2024.

GOMES, Julio Cesar; BRUNO, Diego Renan. **GANs–REDES ADVERSARIAS GENERATIVAS: definições e aplicações**. Revista Interface Tecnológica, v. 20, n. 2, p. 182-194, 2023.

GOODFELLOW, Ian J *et al*. **Generative Adversarial Nets**. arXiv, [s. l.], p. 1–9, 2014.

GOU, Y.; XU, Q.; LIU, D.; WANG, Y. **Detection of offshore oil slicks utilizing a novel joint feature extraction technique**. The Photogrammetric Record, v. 39, n. 3, p. 1-12, 2024.

INTERNATIONAL ENERGY AGENCY (IEA). ***World Energy Outlook 2020***. Paris: IEA, 2020. Disponível em: https://www.iea.org/reports/world-energy-outlook-2020. Acesso em: 12 nov. 2024

KINGMA, Diederik P.; BA, Jimmy**. Adam: a method for stochastic optimization**. In: 3RD INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 2015, San Diego. San Diego: ICLR, 2015.

KRAUS, Mathias; FEUERRIEGEL, Stefan; OZTEKIN, Asil. **Deep learning in business analytics and operations research: Models, applications and managerial implications R.** European Journal of Operational Research, [s. l.], v. 281, n. 3, p. 628–641, 2020. Disponível em: https://doi.org/10.1016/j.ejor.2019.09.018.

KRISHNAMOORTHY, Paramasivam; TREMBLAY, Charles; JACKSON, Paul. **Offshore underwater fixed structure integrity assessment using LRUT, PEC & ACFM advanced technologies with minimal marine growth removal**. In: *Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC)*, 2023, Abu Dhabi. 2023.

LINS, Anthony José da Cunha Carneiro. **Aplicação de aprendizagem de máquina no diagnóstico de declínio cognitivo e demência de Alzheimer baseado em testes cognitivos e marcadores genéticos.** Tese (Programa de Pós-Graduação em Biotecnologia (Renorbio)) – Universidade Federal Rural de Pernambuco, Recife, 2018.

MARTINS, C. F. **O descomissionamento de estruturas de produção offshore no Brasil**. Monografia-Curso de Pós-graduação em Engenharia Ambiental, Universidade Federal do Espírito Santo. Vitória, 2015.

PETROBRAS. **Programa de Descomissionamento de Instalação Marítima – PDI Executivo Parcial – Escopo Plataformas e Poços da Concessão de Guaricema.** Versão 0, março de 2024.

PETROBRAS. **Programa de Descomissionamento de Instalação Marítima – PDI Executivo Parcial – Escopo Plataformas e Poços da Concessão de Caioba.** Versão 0, outubro de 2023a.

PETROBRAS. **Programa de Descomissionamento de Instalação Marítima – PDI Executivo Parcial – Escopo Plataformas e Poços da Concessão de Camorim.** Versão 0, outubro de 2023b.

PORTELA, Marvio**. Dados sintéticos: a chave para a inovação sustentável**. MIT Technology Review Brasil, 13 maio 2022. Disponível em: <https://mittechreview.com.br/dados-sinteticos-a-chave-para-a-inovacao-sustentavel>. Acesso em: 15 set. 2024.

PROENÇA, André Luís Paes; SANTOS, Felipe Vilarin; MANZELA, M. Sc André Aleixo. **Descomissionamento de plataformas de petróleo offshore.** Revista de Engenharias da Faculdade Salesiana, n. 17, p. 7-21, 2023.

PYTORCH. ***PyTorch Documentation***. Disponível em: https://pytorch.org/docs/stable/. Acesso em: 9 nov. 2024.

RADFORD, Alec; METZ, Luke; CHINTALA, Soumith. **Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.** In: ARXIV, nov. 2015.

RODRIGUES, Mateus Franco. **Geração de dados sintéticos utilizando redes neurais artificiais**. 2021. 34 f. Trabalho de Conclusão de Curso (Graduação em Engenharia de Software) - Universidade Federal do Ceará, Campus de Russas, Russas, 2021.

RUIZ-GÁNDARA, Africa; GONZALEZ-ABRIL, Luis. **Generative Adversarial Networks in Business and Social Science**. Applied Sciences, v. 14, n. 17, p. 7438, 2024.

SDMETRICS. ***SDMetrics Documentation***. Disponível em: https://sdv.dev/SDMetrics. Acesso em: 9 nov. 2024

SOUZA, Karen Alves de**. O descomissionamento e desmantelamento como oportunidades de negócio**. In: **29º Congresso Internacional de Transporte Aquaviário, Construção Naval e Offshore - SOBENA**, Rio de Janeiro, 26 de

outubro de 2022. Disponível em:https://www.gov.br/anp/pt-br/centrais-de-conteudo/apresentacoes-palestras/2022/arquivos/apresentacao-sobena-descomissionamento-no-brasil-26-10-2022.pdf. Acesso em: 10 nov. 2024.

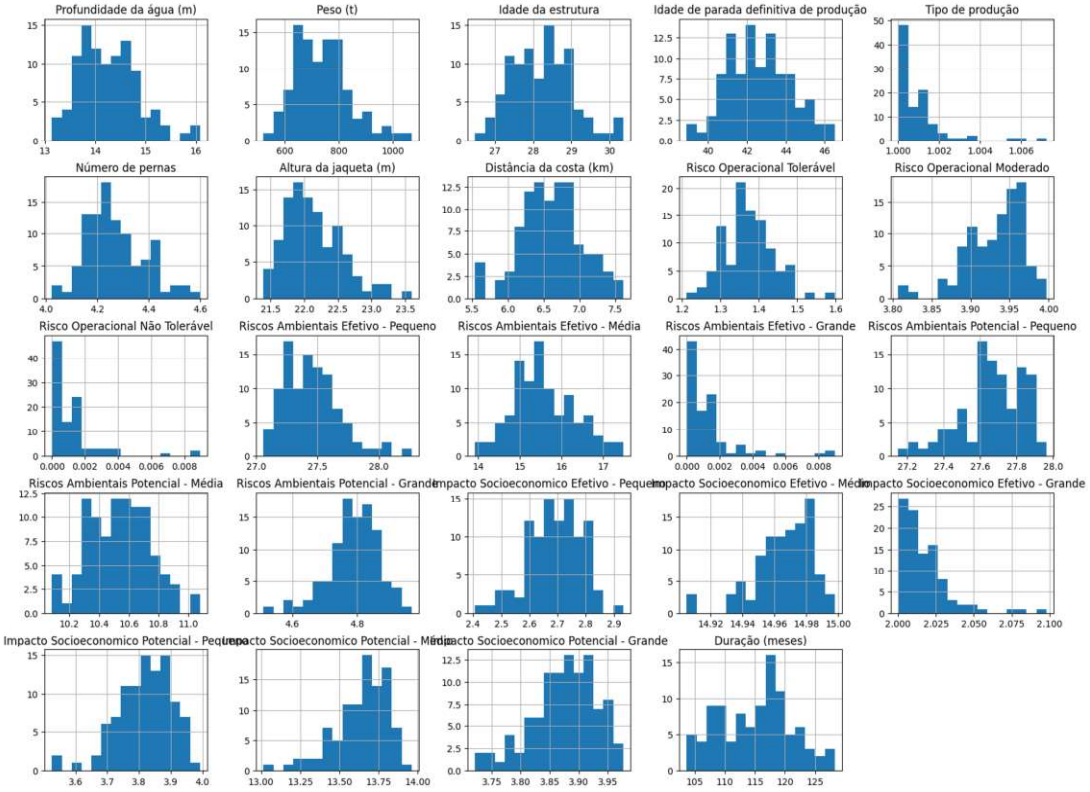
VUTTIPITTAYAMONGKOL, Pattaramon; TUNG, Aaron; ELYAN, Eyad. **Towards machine learning-driven practices for oil and gas decommissioning – introduction of a new offshore pipeline dataset.** *Energies*, v. 14, n. 21, p. 1-20, 2021. DOI: 10.3390/en14216994. Disponível em: https://www.mdpi.com/1996-1073/14/21/6994. Acesso em: 12 nov. 2024.

WEBER, Luciano. **Explorando redes neurais e engenharia do conhecimento uma revisão narrativa.** Cadernos do IME-Série Informática, v. 49, n. 1, p. 157-166, 2024.

APÊNDICIES

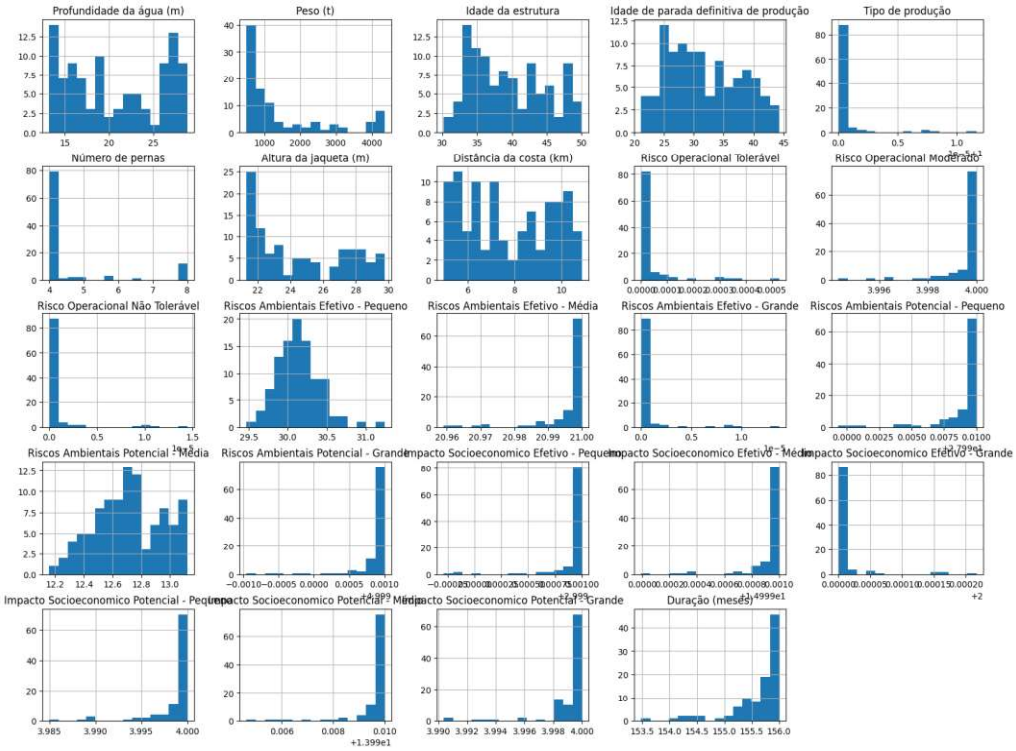
IMAGENS COMPARATIVAS GERADAS PELO MODELO DO CENÁRIO 41

Figura 14 - Distribuição dos dados gerados com 500 épocas



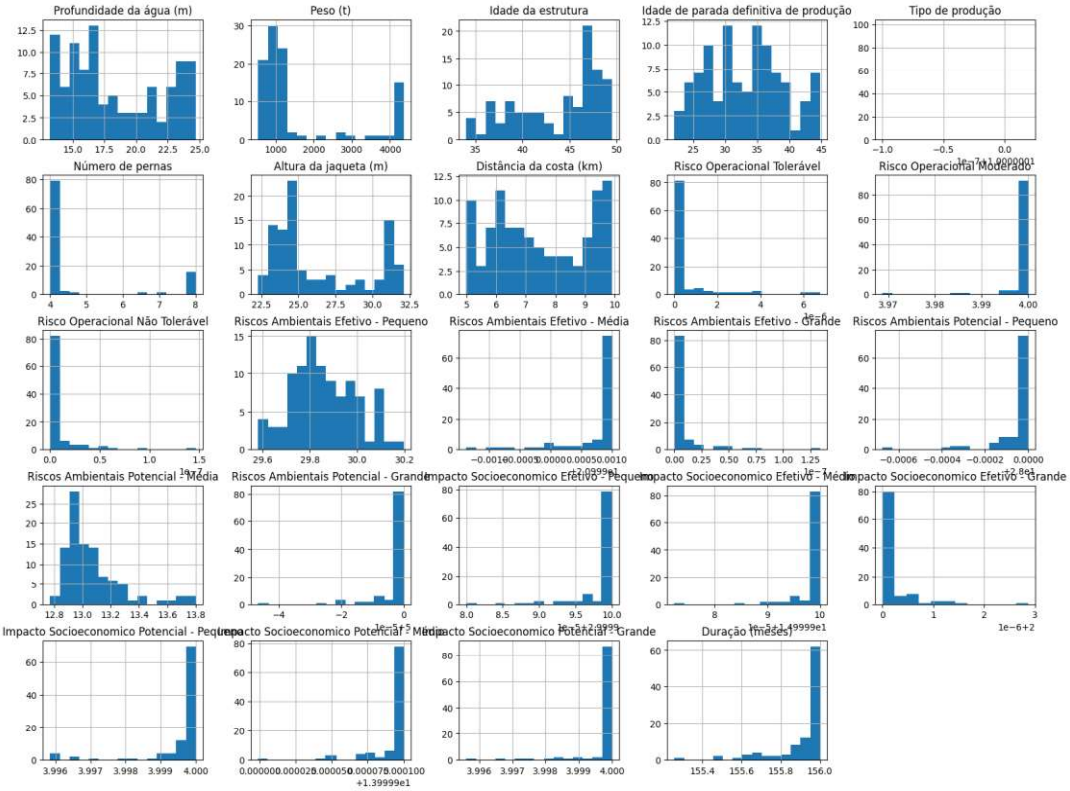
Fonte: Elaboração Própria

Figura 15 - Distribuição dos dados gerados com 10.000 épocas



Fonte: Elaboração Própria

Figura 16 - Distribuição dos dados gerados com 15.000 épocas



Fonte: Elaboração Própria

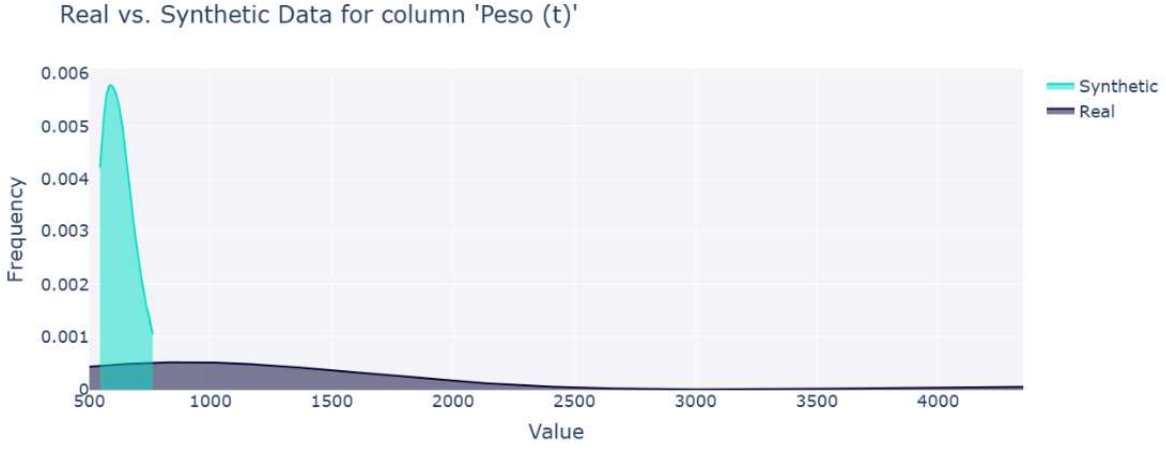
Figura 17 - Distribuição Marginal dos dados gerados com 500 épocas – Altura da Jaqueta (m)

Fonte: Elaboração Própria

Figura 18 - Distribuição Marginal dos dados gerados com 500 épocas – Número de Pernas

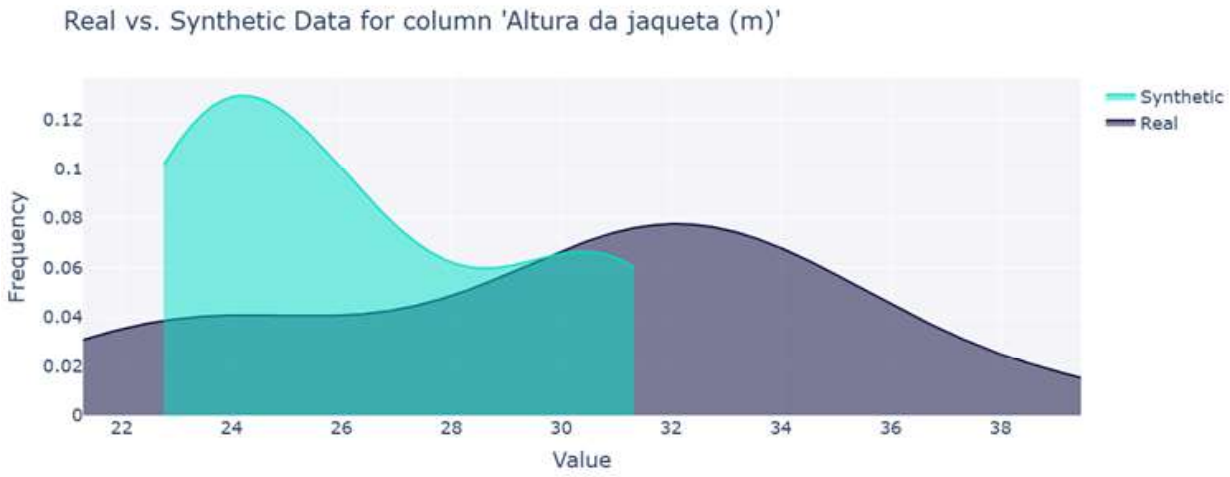
Fonte: Elaboração Própria

Figura 19 - Distribuição Marginal dos dados gerados com 500 épocas – Peso (t)



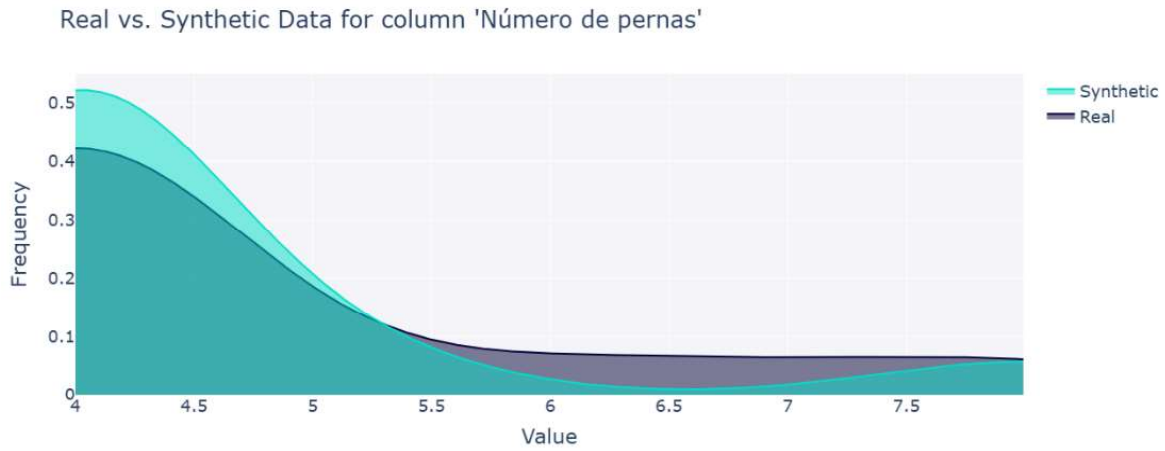
Fonte: Elaboração Própria

Figura 20 - Distribuição Marginal dos dados gerados com 15.000 épocas – Altura da Jaqueta (m)



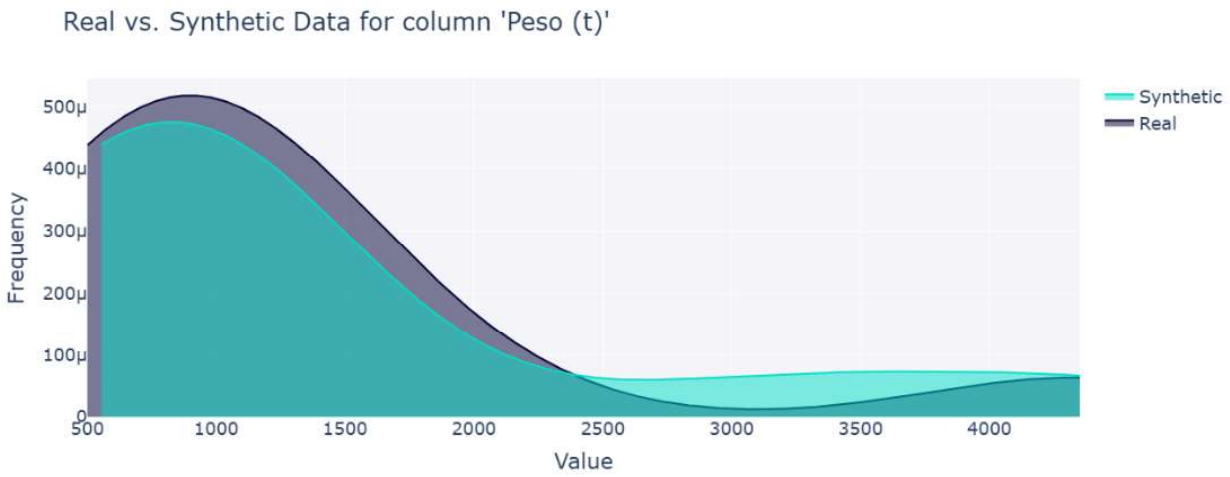
Fonte: Elaboração Própria

Figura 21 - Distribuição Marginal dos dados gerados com 15.000 épocas – Número de Pernas



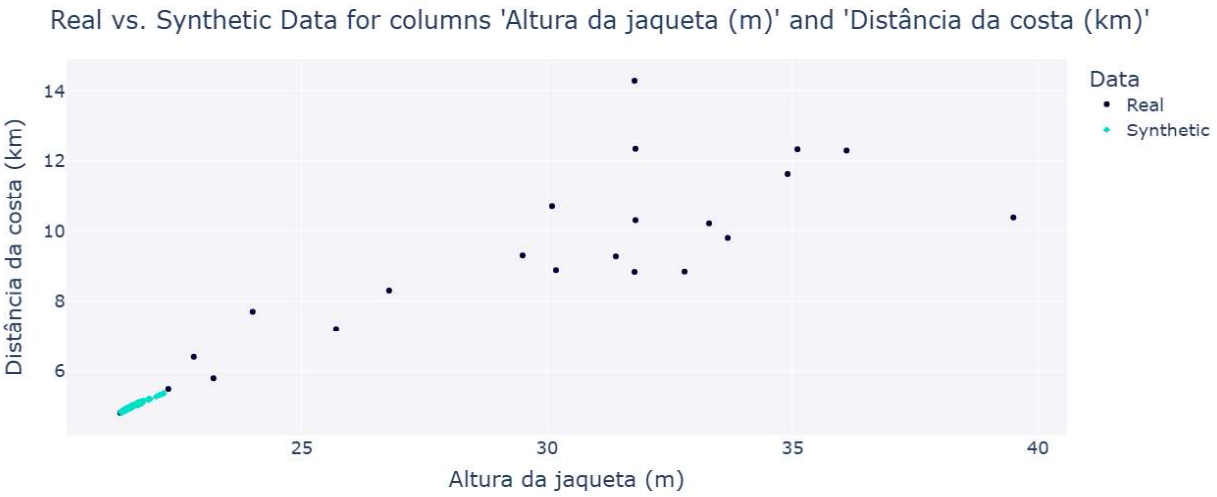
Fonte: Elaboração Própria

Figura 22 - Distribuição Marginal dos dados gerados com 15.000 épocas – Peso (t)



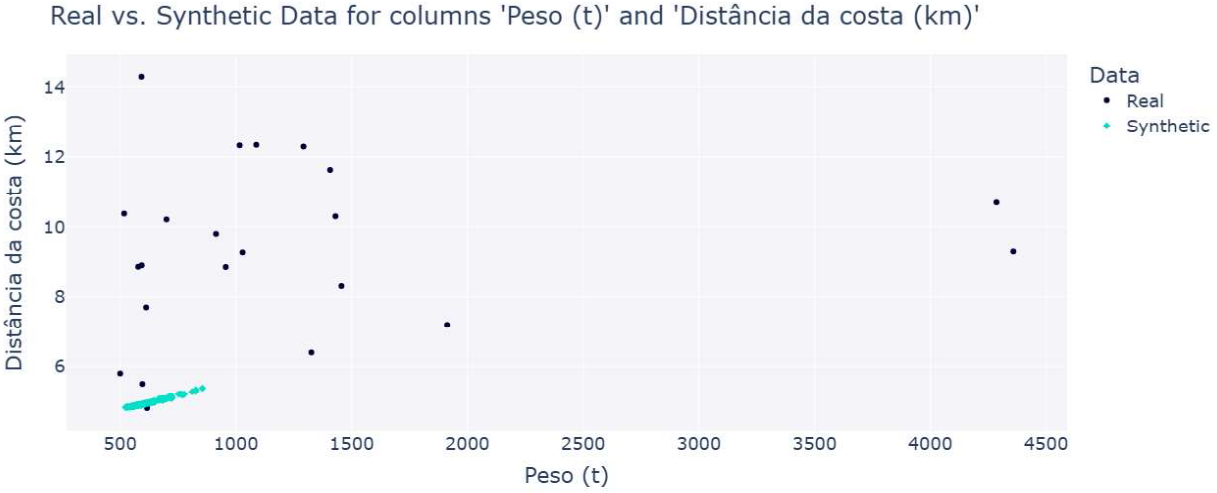
Fonte: Elaboração Própria

Figura 23 - Correlação dos dados gerados com 500 épocas – Altura da jaqueta (m) vs. Distância da costa(km)



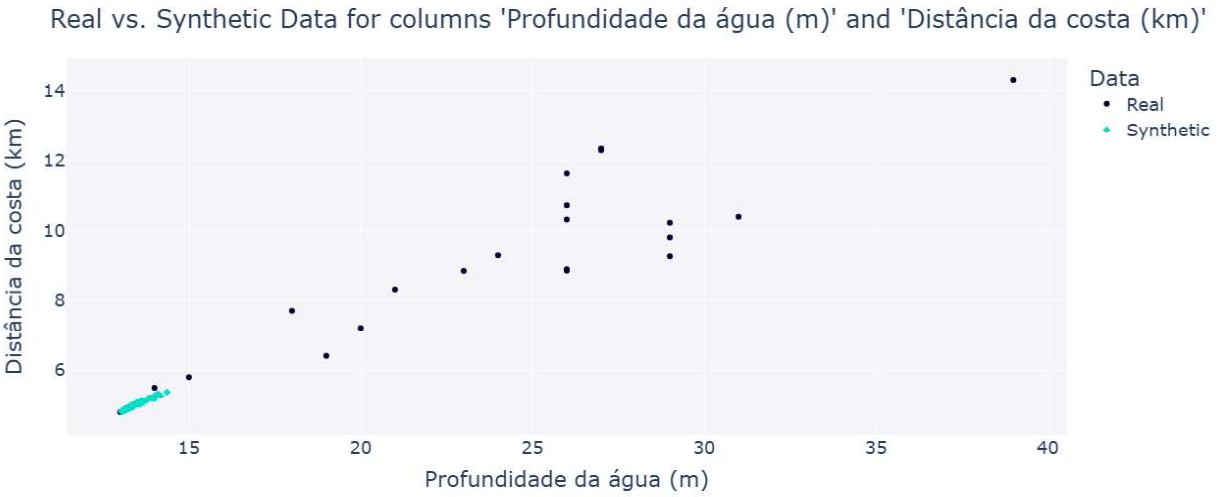
Fonte: Elaboração Própria

Figura 24 - Correlação dos dados gerados com 500 épocas – Peso (t) vs. Distância da costa (km)



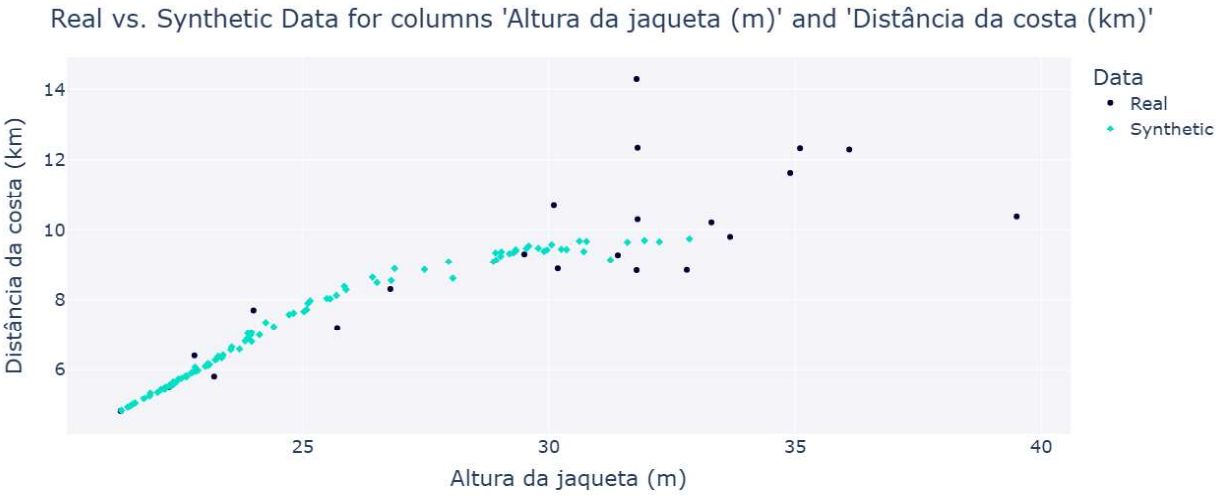
Fonte: Elaboração Própria

Figura 25 - Correlação dos dados gerados com 500 épocas – Profundidade da água (m) vs. Distância da costa(km)



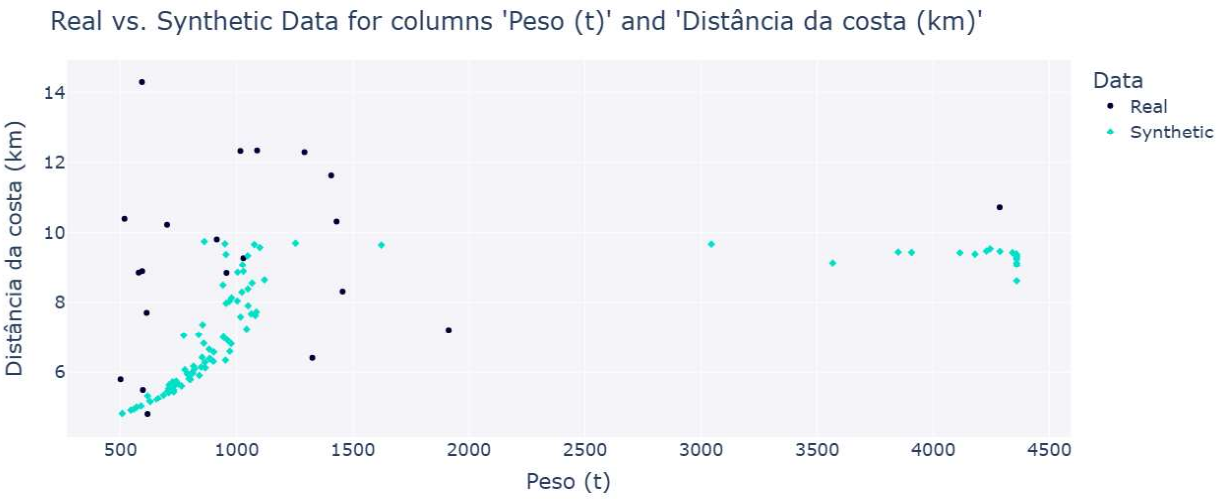
Fonte: Elaboração Própria

Figura 26 - Correlação dos dados gerados com 15.000 épocas – Altura da jaqueta (m) vs. Distância da costa(km)



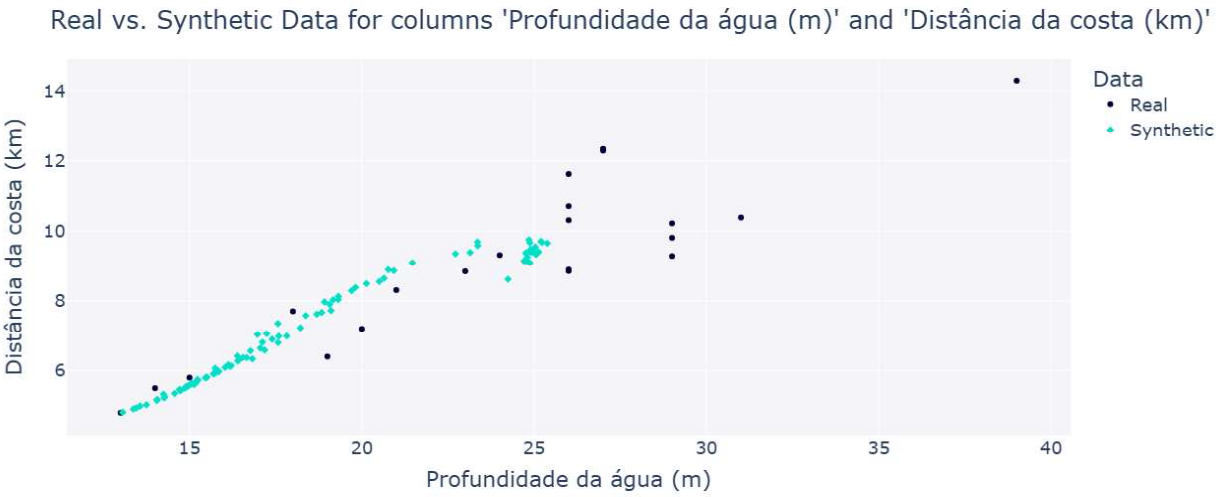
Fonte: Elaboração Própria

Figura 27 - Correlação dos dados gerados com 15.000 épocas – Peso (t) vs. Distância da costa (km)



Fonte: Elaboração Própria

Figura 28 - Correlação dos dados gerados com 15.000 épocas – Profundidade da água (m) vs. Distância da costa(km)



Fonte: Elaboração Própria