UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

MATHEUS RODRIGUES BUENO GODINHO

Avaliação do Uso de Dados Sintéticos em Algoritmos de Radiolocalização Baseados em Fingerprinting e Aprendizagem de Máquina

Recife

2025

MATHEUS RODRIGUES BUENO GODINHO

Avaliação do Uso de Dados Sintéticos em Algoritmos de Radiolocalização Baseados em Fingerprinting e Aprendizagem de Máquina

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Área de Concentração: Inteligência Artificial Aplicada a Sistemas de Localização e Redes Sem Fio

Orientador (a): Daniel Carvalho da Cunha

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Godinho, Matheus Rodrigues Bueno.
Avaliação do Uso de Dados Sintéticos em Algoritmos de Radiolocalização Baseados em Fingerprinting e Aprendizagem de Máquina / Matheus Rodrigues Bueno Godinho. - Recife, 2025.
43 p. : il., tab.

Orientador(a): Daniel Carvalho da Cunha
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Engenharia da Computação - Bacharelado, 2025.
Inclui referências.

1. Aprendizado de Máquina. 2. Radiolocalização. 3. Geração de dados Sintéticos. 4. Redes Neurais Adversariais Generativas. 5. Fingerprinting. I. Cunha, Daniel Carvalho da. (Orientação). II. Título.

000 CDD (22.ed.)

MATHEUS RODRIGUES BUENO GODINHO

AVALIAÇÃO DO USO DE DADOS SINTÉTICOS EM ALGORITMOS DE RADIOLOCALIZAÇÃO BASEADOS EM FINGERPRINTING E APRENDIZAGEM DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de bacharel em Engenharia da Computação.

Aprovado em: 10 / 04 / 2025

BANCA EXAMINADORA

_____

Prof. Dr. Daniel Carvalho da Cunha (Orientador)

Universidade Federal de Pernambuco

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto (Examinador Interno)

Universidade Federal de Pernambuco

Dedico esse trabalho à minha falecida avó Nilza e ao meu avô Benedicto, que por tempos fizeram do seu, o meu lar. Com uma saudosa lembrança dos nossos almoços de domingo.

AGRADECIMENTOS

Gostaria de expressar minha sincera gratidão a todos que estiveram ao meu lado ao longo dessa jornada acadêmica.

Agradeço, primeiramente, aos meus familiares, meu Pai Leonardo, minha mãe Fátima, meu irmão Lucas, minhas avós Nilza e Arnete e meu avô Benedicto, pelo apoio incondicional e por sempre acreditar no meu potencial. À minha namorada Júlia, agradeço por todo seu amor, suporte e incentivo. Agradeço também aos meus mais próximos amigos e colegas de curso, Caio, Júlio e João pelo companheirismo e trocas de aprendizado. Por fim, agradeço a todos os professores que fizeram parte da minha graduação, em especial ao meu orientador, Daniel Cunha, por sua paciência, dedicação e por compartilhar seus conhecimentos de forma tão generosa.

RESUMO

Uma localização precisa é fundamental para diversas aplicações, desde a navegação e a mobilidade urbana até a segurança e os serviços personalizados. Para tanto, o sistema de posicionamento global (GPS) é amplamente utilizado, mas apresenta limitações em ambientes internos e áreas urbanas densas. Para superar essas limitações, técnicas de radiolocalização baseadas em *fingerprinting* e aprendizado de máquina têm sido exploradas. Essas técnicas utilizam parâmetros de sinais de rádio frequência, como o nível do sinal recebido, o tempo de ida e volta e a informação do estado do canal, para treinar modelos capazes de estimar a posição de um usuário. No entanto, a coleta de *fingerprints* para treinamento desses modelos é um processo custoso e pode gerar preocupações com a privacidade. Uma alternativa promissora para tratar esses problemas é a geração de dados sintéticos, que permite reduzir a necessidade de grandes volumes de dados reais, diminuindo os custos de coleta e preservando informações pessoais. Este trabalho tem como objetivo investigar e comparar o impacto de diferentes técnicas de geração de dados sintéticos no desempenho de modelos de localização. Para isso, foram feitos experimentos de geração de dados sintéticos, em duas bases de dados distintas, considerando duas formas de geração condicional, geração não condicional, geração seletiva e não seletiva. Foram utilizados três modelos localizadores baseados em aprendizagem de máquina, com o intuito de avaliar o impacto das diferentes formas de geração de dados sintéticos na precisão dos modelos. No total foram treinados 90 modelos, dentre os modelos base, que receberam apenas dados reais e os modelos dos experimentos, que receberam dados reais e sintéticos. Os resultados obtidos desses experimentos indicam que a geração de dados sintéticos pode melhorar o desempenho de modelos de predição baseados em aprendizagem de máquina.

Palavras-chaves: Radiolocalização. *Fingerprinting*. Aprendizado de Máquina. Redes Neurais Adversariais. Geração de dados Sintéticos.

ABSTRACT

Precise localization is essential to a variety of applications, from navigation to urban mobility, to security and custom services. For such, the global positioning system (GPS) is widely used, nevertheless, it has limitations to its precision in indoor environments and dense urban areas. To overcome these limitations, machine learning and fingerprint based radiolocation techniques have been explored. These methods use radio frequency signal parameters, such as received signal stregth, round-trip time and channel state information, to train models capable of estimating the user's location. However, collecting fingerprints in order to train such models is a costly task that may implicate issues of privacy. A promissing alternative to mitigate these challenges is to generate synthetic data, which reduces the need for large ammounts of data, therefore diminishing the cost of data collection while also preserving personal information. This work is concerned with investigating and comparing the impact of different synthetic data generation techniques in the performance of positioning models. To do so, synthetic data generation experiments were made in two different databases, considering two distinct forms of conditional generation, non-conditional generation, selective and non-selective generation. Three machine learning based positioning models were trained in order to evaluate the impact of the different synthetic data generation methods in the models' precision. In total, 90 models were trained, among baseline models, which were trained solely on real data, and the experiment models, which were trained using real and synthetic data. The results gathered with these experiments indicate that the generation of synthetic data may improve the performance of machine learning based positioning models.

Keywords: Radiolocation. Fingerprinting. Machine Learning. Generative Adversarial Networks. Syntetic Data Generation.

SUMÁRIO

1 INTRODUÇÃO

Os serviços de localização e navegação, como o sistema de posicionamento global (do inglês, *global positioning system* – GPS), são amplamente utilizados e se tornaram indispensáveis no dia-a-dia de muitos. Porém, o GPS tem suas limitações quando se trata da localização em ambientes internos e áreas de grande densidade urbana (AL-TAHMEESSCHI et al., 2022), (FENG; NGUYEN; LUO, 2024a). Por isso, outras soluções de localização são propostas (CHONG; YEO; LIM, 2022), (KARANAM; KORANY; MOSTOFI, 2018).

Uma dessas soluções é a técnica de radiolocalização baseada em *fingerprinting* e aprendizagem de máquina. Nessa técnica, são utilizadas informações, como o nível do sinal recebido (do inglês, *received strength signal* – RSS), o tempo de ida e volta do sinal (do inglês, *round-trip time* – RTT) ou a informação do estado do canal (do inglês, *channel state information* - CSI), parâmetros obtidos a partir dos sinais de rádio frequência (RF), como *Wi-Fi*, *Bluetooth* ou redes de telefonia celular, chamados de *fingerprints*. Os *fingerprints* são utilizados para treinar modelos de aprendizagem de máquina a associar os parâmetros dos sinais de RF a uma posição geográfica. Ao final do treinamento, esses modelos podem ser utilizados para fazer a localização de um usuário, dado o conjunto de sinais recebidos pelo dispositivo móvel.

Apesar do bom desempenho alcançado por esse tipo de técnica, a coleta de *fingerprints* é um processo custoso, já que os modelos precisam de uma grande quantidade de dados para seu treinamento e otimização (NJIMA et al., 2021) e que por vezes necessita ser feito repetidamente, em virtude de mudanças drásticas no ambiente, especialmente para implantações de larga escala e longa duração (LIU et al., 2019). Além do custo elevado associado a grande quantidade de dados necessária, surge também a preocupação com a privacidade (NABATI et al., 2020).

Dito isso, a geração de dados sintéticos é uma possível solução para resolver o problema do elevado custo de coleta de dados de treinamento dos sistemas de radiolocalização baseado em fingerprinting e que empregam modelos de aprendizagem de máquina (GRIRA; MSADAA; GRAYAA, 2023), (YEAN et al., 2021),(NJIMA et al., 2021), (NABATI et al., 2020). O princípio dessa solução é adicionar dados sintéticos, gerados a partir dos dados reais coletados, à base de dados para possibilitar que o treinamento dos algoritmos de localização seja feito com uma menor quantidade de dados coletados, diminuindo os custos associados ao processo de coleta de medições.

1.1 OBJETIVOS

O principal objetivo desse trabalho é analisar se a adição de dados de *fingerprint* sintéticos aos conjuntos de treinamento afeta o desempenho de modelos de localização baseados em *fingerprinting* e aprendizagem de máquina.

Atrelado a esse objetivo, busca-se avaliar também o impacto do aumento da proporção de dados sintéticos no desempenho dos modelos, entender qual técnica de geração de dados sintéticos proposta promove o melhor resultado e investigar os efeitos da aplicação de um filtro seletivo nesses dados.

O primeiro passo para se atingir esses objetivos foi escolher o conjunto de dados a ser utilizado. Nesse caso, foram escolhidos dois conjuntos, o primeiro foi UJIIndoorLoc (TORRES-SOSPEDRA et al., 2014), amplamente utilizada em trabalhos de localização e disponível publicamente no site da UCI Machine Learning Repository[1]. O segundo conjunto foi uma base de dados colhida na Universidade Federal de Pernambuco (UFPE) com a ajuda da operadora de telefonia móvel celular Claro, a partir de três grupos de antenas de redes móveis situadas no entorno da UFPE.

O segundo passo foi utilizar uma arquitetura de rede adversarial generativa denominada CTGAN (do inglês, *Conditional Tabular Generative Adversarial Networks*), que tem a capacidade de aprender a distribuição de dados do conjunto de *fingerprints* coletados em campo e gerar dados sintéticos semelhantes aos originais, com a possibilidade de fazer amostras de dados sintéticos de forma condicional, em conjunção com o agrupamento das localizações em zonas, feito a partir do método de análise hierárquica de clusters (do inglês, *Hierarchical Cluster Analysis* – HCA). Com o agrupamento das localizações por zonas, foi possível fazer a geração condicional de dados sintéticos limitado apenas às zonas desejadas.

Por fim, foram escolhidos, devido a sua ampla utilização no contexto de localização por *fingerprinting*, três modelos de aprendizagem de máquina: regressão por vetor de suporte (do inglês, *support vector regression* – SVR) (JONDHALE et al., 2022), aumento de gradiente extremo (do inglês, *extreme gradient boosting* – XGBoost) (SINGH et al., 2022) e perceptron de múltiplas camadas (do inglês, *multi-layer perceptron* – MLP) (NABATI et al., 2020), com o objetivo de avaliar seus desempenhos comparando os impactos de diferentes métodos de geração de dados sintéticos.

[1] Disponível em https://archive.ics.uci.edu/dataset/310/ujiindoorloc

1.2 ESTRUTURA DO DOCUMENTO

Neste documento, o Capítulo 2 aborda a fundamentação teórica necessária para o entendimento da pesquisa. Em seguida, no Capítulo 3, será apresentada a metodologia aplicada no trabalho, além dos resultados obtidos e discussão. Por fim, será apresentada a conclusão do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, é feito o embasamento conceitual das técnicas utilizadas durante a pesquisa. São descritas a técnica de localização, modelos de predição e modelo de geração de dados sintéticos, método de clusterização aplicados no desenvolvimento do trabalho.

2.1 LOCALIZAÇÃO

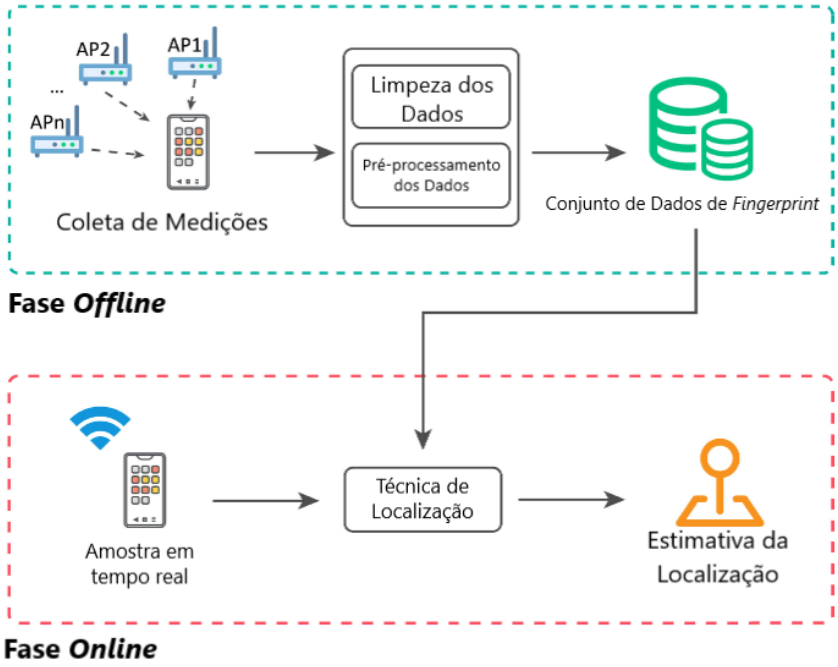
Como apontado na introdução, o GPS tem suas limitações quando se trata da localização em ambientes internos e áreas de grande densidade urbana (FENG; NGUYEN; LUO, 2024a), (AL-TAHMEESSCHI et al., 2022). Isso se dá devido a fenômenos físicos que afetam o sinal, como o efeito multipercurso, reflexões, bloqueios e absorção do sinal, que podem tornar instáveis as medições de sinal de GPS, o que acarreta na performance não confiável nesses cenários (AL-TAHMEESSCHI et al., 2022). Por isso, técnicas baseadas em angulo de chegada (do inglês, *Angle of Arrival* - AoA), tempo de chegada (do inglês, *Time of Arrival* - ToA) ou *fingerprinting* tem sido escolhidas para contornar as limitações do GPS. Nesse trabalho, foi explorada a técnica de localização baseada em *fingerprinting*.

2.1.1 Técnicas de Localização baseada em *fingerprinting*

A técnica de localização baseada em *fingerprinting* consiste no princípio de que, dado um ambiente com um conjunto de sinais de redes sem fio, sejam eles Wi-Fi, Bluetooth, de redes móveis ou outro tipo de comunicação sem fio, cada posição tem uma espécie de impressão digital desse conjunto de sinais. Devido à complexa propagação, dentre refrações, reflexões, atenuações e interferências de multipercurso, dos sinais eletromagnéticos e os diferentes caminhos que cada sinal percorreu para chegar naquela posição, cada localização recebe um conjunto único de medições de nível de sinais, chamado de *fingerprint* (FENG; NGUYEN; LUO, 2024a).

O *fingerprint* de uma localização pode ser construído a partir de diversas informações recebidas sobre o sinal, sendo RSS, RTT e CSI as mais utilizadas. Ao se coletar meticulosamente *fingerprints* de sinal e suas respectivas coordenadas em diferentes localizações, é possível criar um conjunto de dados de *fingerprints* (FENG; NGUYEN; LUO, 2024a).

Figura 1 – Fases da técnica de localização baseada em *fingerprinting*.



Fonte: FENG; NGUYEN; LUO (2024a).

A partir do conjunto de dados coletado, que pode utilizar qualquer uma das informações sobre o sinal ou uma combinação delas (FENG; NGUYEN; LUO, 2024b), são utilizados algoritmos de localização para determinar a posição atual de um usuário ao comparar o *fingerprint* da sua localização atual com àqueles coletados anteriormente (FENG; NGUYEN; LUO, 2024a).

O método de localização por *fingerprint* é dividido em duas fases. A primeira, a fase *offline*, se trata da coleta pré-processamento e limpeza dos dados de *fingerprint*, assim como os treinamentos dos modelos de aprendizagem de máquina. Vale ressaltar que a coleta desses dados de *fingerprint* é custosa em função do tempo e esforço necessário para coletar uma quantidade de dados que possa ser utilizada efetivamente para o treinamento e avaliação de desempenho dos modelos de aprendizagem (FENG; NGUYEN; LUO, 2024a). Já na segunda, a fase *online*, os modelos já treinados recebem o conjunto de sinais e estimam a localização do usuário em tempo real. A Figura 1 exemplifica esse processo.

2.2 MODELOS DE PREDIÇÃO

Algoritmos baseados aprendizagem de máquina e aprendizagem profunda vem sendo amplamente aplicados no contexto das técnicas de localização por *fingerprinting*, graças à precisão

e à robustez desses modelos, principalmente quando comparados com abordagens determinísticas (BAHL; PADMANABHAN, 2000) ou probabilísticas (YOUSSEF; AGRAWALA, 2005).

Foram utilizados no trabalho, para a investigação da mudança de precisão com diferentes formas de adição de dados sintéticos no conjunto de dados de treinamento, três modelos de regressão. Desses modelos, dois são baseados em aprendizagem de máquina: SVR e XGBoost, e um em aprendizagem profunda: MLP.

2.2.1 Perceptron de Múltiplas Camadas

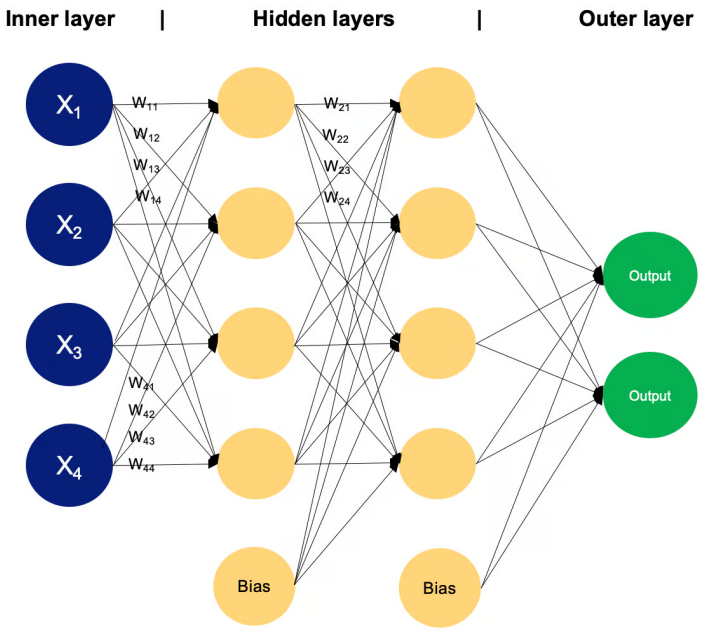
O MLP é um tipo de rede neural artificial que consiste em várias camadas de neurônios, que são a unidade fundamental de processamento de informação de uma rede neural artificial e recebem uma série de pesos. Os neurônios do MLP normalmente usam funções de ativação não lineares, permitindo que a rede aprenda padrões complexos nos dados. Os MLPs são importantes no aprendizado de máquina porque podem aprender relações não lineares nos dados, o que os torna modelos poderosos para tarefas como classificação, regressão e reconhecimento de padrões (JAISWAL, 2024). A Figura 2 apresenta um exemplo de MLP com cinco camadas de neurônios, uma camada de entrada, representada em azul, duas camadas ocultas em amarelo e uma camada de saída em verde. A camada de entrada consiste em nós ou neurônios que recebem os dados de entrada iniciais. Cada neurônio representa um recurso ou uma dimensão dos dados de entrada. Entre as camadas de entrada e saída, existem as camadas ocultas, nas quais cada neurônio recebe entradas de todos os neurônios da camada anterior (seja a camada de entrada ou outra camada oculta) e produz uma saída que é passada para a próxima camada. A camada de saída consiste em neurônios que produzem a saída final da rede. O número de neurônios nessa camada depende da natureza da tarefa. No caso da regressão, haverá um neurônio para cada variável a ser prevista.

Para o caso desse trabalho, o MLP foi utilizado para a tarefa de regressão de encontrar a latitude e longitude, dado um *fingerprint* como entrada.

2.2.2 Regressão por Vetor de Suporte

O SVR é uma técnica de regressão baseada no modelo de aprendizagem de máquina chamado de máquina de vetor de suporte (do inglês, *support vector machine* - SVM). Entretanto, diferente da SVM, que é utilizada para tarefas de classificação, o objetivo da SVR é resolver

Figura 2 – Representação de um MLP com duas camadas ocultas.



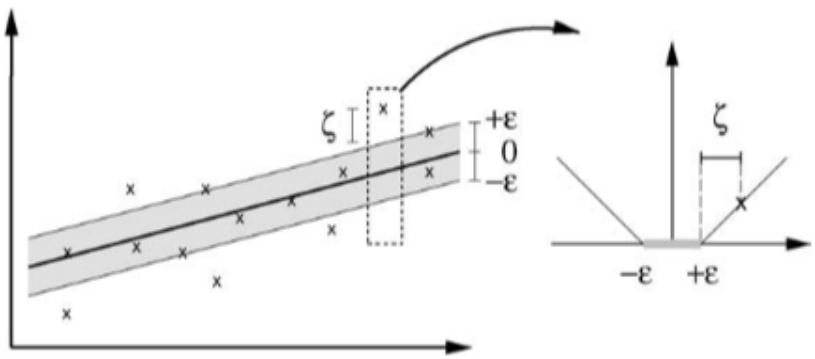
Fonte: JAISWAL (2024).

problemas de regressão, prevendo valores contínuos em vez de classes discretas.

Para isso, o SVR busca encontrar uma função $f(x)$ que tenha um desvio de no máximo ϵ dos valores reais y_i para todos os dados de treinamento (como na Figura 3) (SMOLA; SCHÖLKOPF, 2004). Depois do treinamento, o modelo pode então utilizar a aproximação aprendida da função $f(x)$ para estimar um valor y, a partir de um valor de entrada x.

Para o caso desse trabalho, o SVR foi utilizado para a tarefa de regressão de encontrar a latitude e longitude, dado um *fingerprint* como entrada.

Figura 3 – A configuração de margem de perda suave para um SVR linear.



Fonte: SMOLA; SCHÖLKOPF (2004).

2.2.3 Aumento de Gradiente Extremo

XGBoost é uma biblioteca que implementa algoritmos de aprendizagem de máquina baseados na arquitetura de *Gradient Boosting* (CHEN; GUESTRIN, 2016), que combina vários modelos de árvore de regressão fracos para criar um modelo mais robusto, chamado de *tree ensemble*.

Dado um conjunto de dados com entradas $D = \{(x_i, y_i)\}$, o modelo de *tree ensemble* é definido por

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F, \tag{2.1}$$

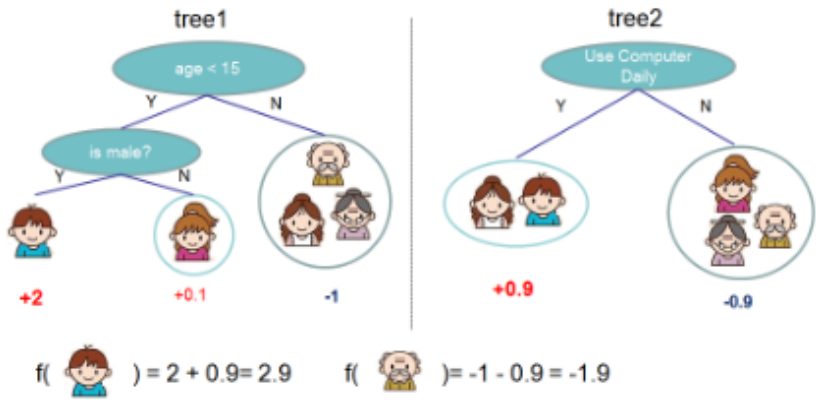
em que x_i são as entradas e $\hat{y}_i$ é a saída esperada. Para prever $\hat{y}_i$, são utilizadas K funções aditivas f_k, sendo K, a quantidade de árvores do conjunto. Cada função f_k corresponde a uma estrutura de árvore e seus pesos de folha independentes. A Figura 4 ilustra um modelo de *tree ensemble*, na qual a previsão final para uma dada entrada é a soma das predições de cada árvore. No caso da figura, caso a entrada seja um menino é somado o valor calculado na primeira árvore ao valor calculado na segunda árvore, o que resulta em $2 + 0.9 = 2.9$, que é o valor previsto pelo *tree ensemble*.

Para aprender o conjunto de funções utilizadas no modelo, deve-se minimizar o seguinte objetivo de aprendizagem regularizado

$$\mathcal{L}(\phi) = \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \ , \tag{2.2}$$

em que $\ell(\cdot)$ é uma função de perda convexa diferenciável que mede a diferença entre a predição

Figura 4 – Modelo de conjunto de árvores. A previsão final para um dado exemplo é a soma das predições de cada árvore.



Fonte: CHEN; GUESTRIN (2016).

$\hat{y}_i$ e o valor real y_i. O segundo termo $\Omega(\cdot)$ penaliza a complexidade do modelo, nesse caso, as funções de regressão f_k das árvores.

Para o caso desse trabalho, o XGBoost foi utilizado para a tarefa de regressão de encontrar a latitude e longitude, dado um *fingerprint* como entrada.

2.3 REDES ADVERSARIAIS GENERATIVAS

As redes adversariais generativas (do inglês, *Generative Adversarial Networks* – GANs) , introduzidas por Goodfellow em (GOODFELLOW et al., 2014), são redes neurais adversariais que tem a capacidade de aprender a distribuição de um conjunto de dados de treinamento e posteriormente gerar dados sintéticos semelhantes aos originais.

As GANs são modelos compostos internamente por duas redes neurais envolvidas em um processo adversarial: a rede Geradora G, que captura a distribuição dos dados e gera dados sintéticos, e a Discriminadora D, que estima a probabilidade de que um dado veio do conjunto de treinamento e não foi gerado por G. A rede neural Geradora tem a função durante o treinamento de maximizar a probabilidade de D cometer um erro.

Essa arquitetura corresponde a um jogo de minimax de dois jogadores, G e D, que pode ser descrito pela função valor $V(D, G)$ tal que (GOODFELLOW et al., 2014):

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[1 - log(D(G(z)))] \tag{2.3}$$

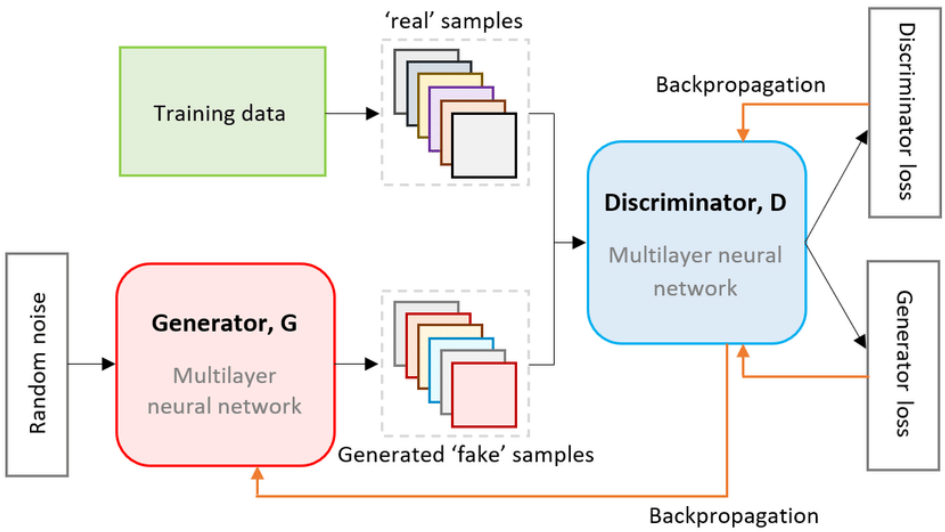
na qual $D(x)$ representa a probabilidade que x veio do dado de treinamento e não da distribuição p_g, gerada por G. Essa é a parte da função que o discriminador pretende maximizar. O gerador, por sua parte, deve maximizar a probabilidade de o discriminador identificar dados sintéticos como reais, maximizando $D(G(z))$, efetivamente minimizando a função $log(1 - D(G(z)))$, em que z é um vetor da distribuição aleatória p_z.

Ao fim do treinamento, depois do refinamento da geração de G e da detecção de D através da disputa, o gerador pode produzir dados sintéticos com uma distribuição próxima da original.

2.3.1 Redes Adversariais Generativas Condicionais Tabulares

A *Conditional Tabular GAN* é um método baseado em GAN para modelar a distribuição de dados tabulares. Na CTGAN, diferentemente das GANs tradicionais, possui um gerador

Figura 5 – Arquitetura da GAN.



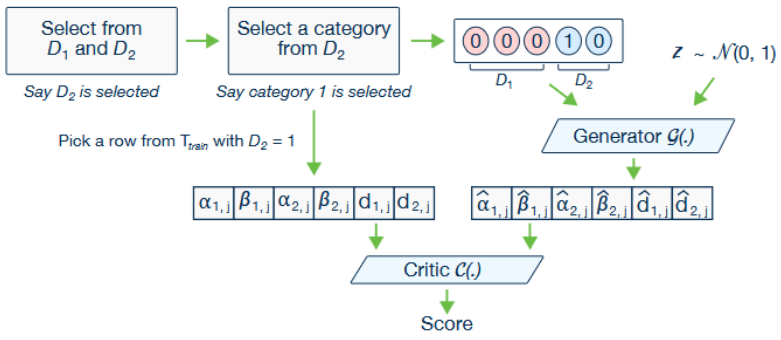
Fonte: LITTLE et al. (2021).

condicional, que foi treinado com um método chamado treinamento por amostra, para lidar com colunas discretas desbalanceadas (XU et al., 2019).

Como indicado na Figura 6, o gerador condicional pode gerar linhas sintéticas concionadas a uma, ou mais, das colunas discretas. Com o treinamento por amostra, os dados condicionais e de treinamento são amostrados de acordo com a frequência logarítmica de cada categoria, então a CTGAN pode explorar todos os valores discretos possíveis (XU et al., 2019).

Nesse trabalho, a CTGAN foi utilizada para para a geração de dados sintéticos de *fingerprint* tanto de forma não condicional, como de forma condicional, a partir de zonas definidas atráves do algoritmo de aglomeração HCA.

Figura 6 – Modelo CTGAN. O gerador condicional pode gerar linhas sintéticas condicionadas a uma das colunas discretas.



Fonte: XU et al. (2019).

2.4 ANÁLISE HIERÁRQUICA DE CLUSTERS

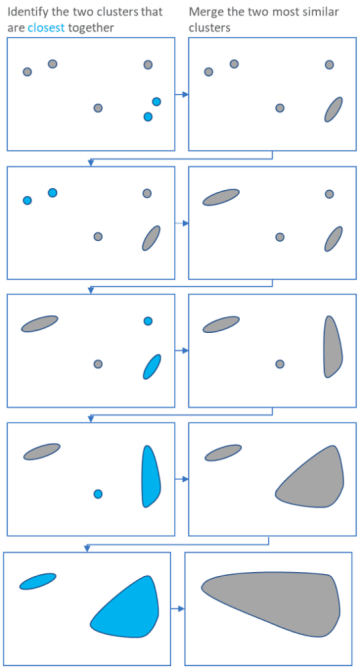
O HCA é um algoritmo que agrupa objetos similares de um conjunto em subgrupos chamados *clusters*. O objetivo final da é obter um conjunto de *clusters* que são distintos entre si, que contenham, internamente, elementos similares entre si.

O HCA funciona da seguinte maneira: ele começa tratando cada registro como um *cluster* de apenas um elemento e depois repete os seguintes dois passos até que todos os dados estejam contidos em um só *cluster*: primeiro, identificar os *clusters* mais similares e segundo juntar esses *clusters* em um só, a Figura 7 exemplifica esse comportamento.

A saída desse processo é um dendograma, como apresentado na Figura 8. A partir do dendograma é identificado quantos *clusters* existem em cada passo do processo. Com isso, é possível escolher um ponto de corte no processo de HCA e extrair os *clusters* computados naquele passo.

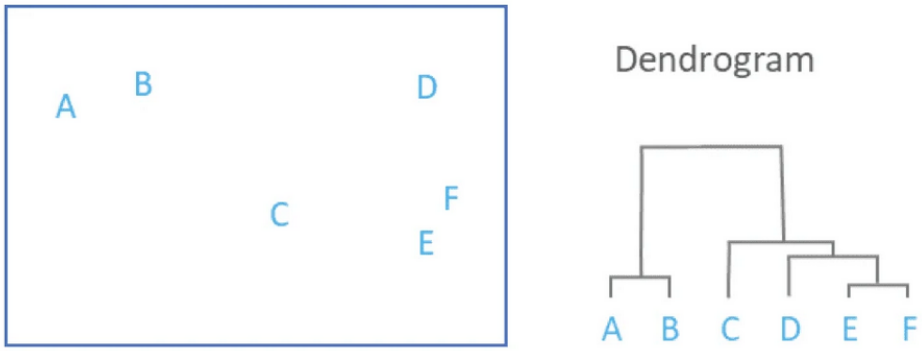
É com essa função que o trabalho utiliza o algoritmo de HCA para dividir o conjunto de localizações mapeadas no dataset de teste em zonas distintas, que serão utilizadas na geração condicional de dados sintéticos.

Figura 7 – Exemplo de passo a passo da utilização do algoritmo HCA.



Fonte: BOCK (2025).

Figura 8 – Exemplo de dendograma.



Fonte: BOCK (2025).

3 METODOLOGIA E RESULTADOS

São apresentados nesse capítulo a metodologia abordada para a aplicação dos experimentos e os resultados obtidos. Primeiramente, serão descritas as duas bases de dados exploradas no trabalho e os pré-processamentos realizados. Logo após, serão descritos os diferentes métodos utilizados para a geração de dados sintéticos e treinamento dos modelos de predição. Por fim, são apresentados os resultados e a discussão.

3.1 BASES DE DADOS

Foram utilizadas duas bases de dados nas investigações desse trabalho e, em ambas, foram realizados os mesmos experimentos.

A primeira foi a UJIIndoorLoc (TORRES-SOSPEDRA et al., 2014), que se trata de uma base de dados amplamente utilizada em trabalhos de localização (YEAN et al., 2021), (NJIMA et al., 2021), (FENG; NGUYEN; LUO, 2024a). A base UJIIndoorLoc contém 19.937 registros para treinamento e 1.111 registros para teste e validação, obtidos em múltiplos prédios e andares da universidade espanhola *Universitat Jaume I*. A base contém os seguintes atributos: 520 colunas de indicadores de nível de sinal recebido (do inglês, *received signal strength indicator* - RSSI), relativos a 520 pontos de acesso sem fio (do inglês, *wireless access points* - WAPs) da universidade, duas colunas de coordenadas geográficas de latitude e longitude em metros, seguindo o Sistema Geodésico Mundial de 1984 (do inglês, *World Geodetic System* 1984 – WGS84), uma coluna que identifica o andar do prédio em que foi feita a medição, uma coluna que identifica o prédio, uma coluna que identifica o espaço em que foi feita a medição e uma coluna com a posição relativa a esse espaço, uma coluna que identifica o usuário que fez a medição e uma que identifica o aparelho celular utilizado para a medição e por fim uma coluna que registra a data e hora da medição.

Já a segunda base de dados utilizada foi colhida nas redondezas da Universidade Federal de Pernambuco (UFPE). Esse é um conjunto com 6.775 registros, que representam medições feitas em ambientes externos e internos da faculdade com relação a três grupos de antenas de redes móveis, no qual cada grupo tem 3 antenas. A base contém os seguintes atributos: 9 colunas de indicadores de força de sinal, 3 colunas de indicadores de atraso de propagação, duas colunas de coordenadas geográficas latitude e longitude, em graus decimais e uma coluna

que indica se o registro foi coletado em ambiente externo ou interno.

3.2 PRÉ-PROCESSAMENTO

Para o conjunto de dados UJIIndoorLoc, esse trabalho utiliza apenas as colunas de RSSI e as de coordenadas geográficas (latitude e longitude). Foi aplicado um filtro para se manter apenas os registros do térreo (andar zero), com a intenção de manter o problema de predição restrito a apenas duas dimensões, o que diminuiu o tamanho do conjunto para 4.305 registros de treino e 132 de teste. Para se alcançar a proporção de 90% de dados de treino e 10% de dados de teste, o conjunto de treino foi reduzido para 1.188 registros, amostrados de forma aleatória.

Além disso, para que fosse possível o treinamento e geração de dados sintéticos pela CTGAN, foi necessário diminuir a dimensionalidade do conjunto. Para isso, foi aplicado o algoritmo de analise de componentes principais (do inglês, *Principal Component Analysis* – PCA) nas colunas de RSSI. O objetivo foi extrair as informações importantes de um conjunto de dados e representá-lo como um conjunto de variáveis ortogonais chamadas de componentes principais, o que diminuiu a dimensão de 520 para 12 colunas.

Já para o conjunto de dados de redes móveis da UFPE, primeiramente foi feita a limpeza de registros duplicados com todas as informações repetidas. Em seguida, foi aplicado um filtro para utilizar nesse apenas os dados colhidos em ambientes externos, já que os registros internos foram coletados em grande quantidade e em poucas posições, o que poderia gerar um viés durante os treinamentos dos modelos. Após o filtro, sobraram 2.154 registros externos não duplicados. Desse conjunto, foi feita uma divisão de 90% de dados para treinamento e 10% de dados para teste, para que a CTGAN recebesse uma boa quantidade de dados para o treinamento.

Para ambos os conjuntos de dados, além da diminuição de dimensionalidade no conjunto da UJIIndoorLoc, não houve nenhum outro tipo de seleção de atributos.

3.2.1 Clusterização com HCA

O passo seguinte no pré-processamento foi a aplicação do algoritmo HCA com o propósito de dividir os conjuntos em zonas. Essas zonas agruparam os registros de localização, como feito em (GRIRA; MSADAA; GRAYAA, 2023). Em ambos os conjuntos de dados, o algoritmo foi

Figura 9 – Zonas definidas por meio da aplicação do algoritmo HCA na base de dados UJIIndoorLoc. A cor do ponto indica de qual zona aquele registro pertence.

Fonte: Elaborada pelo Autor (2025).

aplicado levando em consideração apenas as colunas de coordenadas geográficas (latitude e longitude) do conjunto de treinamento.

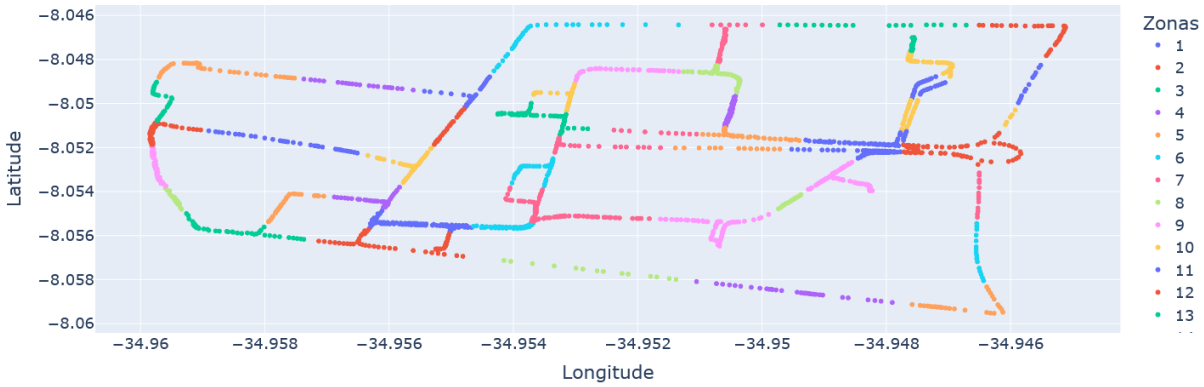
Primeiramente, foi feita a normalização desses registros e, em seguida, utilizando os métodos *linkage*[1], *dendogram*[2] e *fcluster*[3], da biblioteca *scipy.cluster.hierarchy*[4], foi possível gerar o dendograma e escolher um ponto de corte para a formação dos *clusters* (GRIRA; MSADAA; GRAYAA, 2023).

A distância Euclidiana de ponto de corte escolhida foi de 2 para o conjunto da UFPE, o que resultou na divisão em 42 zonas, equivalente fusão de aproximadamente $97,88\%$ dos *clusters*. Seguindo a mesma proporção de *clusters* fundidos, a distância euclidiana encontrada para o conjunto da UJIIndoorLoc foi de $1,27$, o que resultou em 27 zonas. O resultado da clusterização é apresentado nas Figuras 9 e 10, nas quais a zona de um registro é representada pela sua cor. Vale ressaltar que, devido a grande quantidade de zonas, algumas cores de zonas podem se repetir.

Ao fim do processo de clusterização, foi adicionada uma nova coluna categórica chamada *cluster*, que indica a qual zona cada registro pertence.

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html
[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html
[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html
[4] https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

Figura 10 – Zonas definidas por meio da aplicação do algoritmo na base de dados da UFPE. A cor do ponto indica de qual zona aquele registro pertence.



Fonte: Elaborada pelo Autor (2025).

3.3 TREINAMENTO E VALIDAÇÃO DA CTGAN

Nos experimentos, foram treinadas duas CTGANs baseadas na classe *CTGANSynthetizer* da biblioteca *sdv.single_table*[5], uma para cada conjunto de dados. O treinamento das CTGANs foi feito em 7.000 épocas e levou em consideração a coluna *cluster* como coluna categórica, de forma que fosse possível a geração de amostras sintéticas por zonas.

Para a validação da qualidade da geração de dados sintéticos dessas CTGANs, foram observadas três medidas: a divergência de Kullback-Leibler (KL), a distância de Wasserstein e a distância de início de Fréchet (do inglês, *Fréchet inception distance* – FID), que buscaram comparar a distribuição dos dados sintéticos com a dos dados reais.

A divergência KL é uma medida de diferença entre duas distribuições de probabilidade P e Q, interpretada como a quantidade de informação perdida ao aproximar uma distribuição pela outra. Essa medida é calculada pela seguinte fórmula:

$$D_{KL}(P||Q) = \sum_i P(i) log(\frac{P(i)}{Q(i)}) \ . \tag{3.1}$$

Na divergência KL, quanto mais alto for o valor da divergência, maior é a discrepância entre as distribuições, indicando que Q não é uma boa representação de P. Essa medida não é uma métrica propriamente dita, já que não é uma função simétrica, ou seja, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ (DHINAKARAN, 2023). Mesmo assim, em conjunção com a distância de Wasserstein e a FID, é possível julgar a qualidade da distribuição sintética e, por consequência, a qualidade do treinamento da CTGAN.

[5] https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/ctgansynthesizer

Tabela 1 – Medidas de diferença entre conjuntos avaliada entre dados reais e sintéticos.

Base de Dados	Divergência KL	Distância de Wasserstein	FID
UFPE	0,0064	0,0504	0,4571
UJIIndoorLoc	0,0072	0,0943	1,3408

Fonte: Elaborada pelo autor (2025).

A distância de Wasserstein é uma métrica que quantifica a diferença entre duas distribuições de probabilidade. Essa métrica é baseada no conceito de transporte ótimo, que busca encontrar a maneira mais eficiente de mover uma distribuição para se igualar a outra (CHILAMKURTHY, 2020). A distância de Wasserstein é amplamente utilizada no contexto de aprendizado de máquina, principalmente no treinamento de modelos generativos, como as Wassertein GANs (WGANs) (WENG, 2019). No caso desse trabalho, foi usada juntamente com a divergência de KL e a FID para avaliar a qualidade do treinamento da CTGAN a partir da qualidade dos dados sintéticos gerados. Quanto menor a distância de Wasserstein, melhor a aproximação das distribuições.

A FID também é uma métrica popularmente utilizada na avaliação da qualidade de treinamento e geração de dados sintéticos das GANs (THAKUR, 2022), que pode ser utilizada para medir a distância entre distribuições. Ela é mais comumente utilizada na avaliação de GANs que geram imagens e é calculada por

$$FID = ||\mu_X - \mu_Y||^2 - Tr(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}) \ , \tag{3.2}$$

em que X e Y são duas distribuições normais multivariadas, μ_X e μ_Y são os vetores de médias das distribuições, Σ_X e Σ_Y são as matrizes de covariância das distribuições e, por fim, $Tr(\cdot)$ é o traço da matriz. Quanto menor a FID, melhor a aproximação das distribuições.

Para o caso das CTGANs treinadas, as medidas encontradas estão apresentadas na Tabela 1.

Os baixos valores na Tabela 1 avaliados pelas medidas indicam que os treinamentos foram bem sucedidos.

3.4 EXPERIMENTOS

Neste trabalho, foram feitos quatro tipos de experimentos para investigar o impacto que a forma de geração de dados sintéticos tem no treinamento dos modelos de predição. No primeiro experimento, foi considerada a geração indiscriminada dos dados sintéticos, que gerou

dados desconsiderando as zonas introduzidas pelo algoritmo HCA. No segundo, foi realizada a geração estratificada, que gerou dados sintéticos seguindo a mesma proporção por zona que a distribuição original. O terceiro experimento assumiu o uso de uma geração balanceada, que criou registros sintéticos de forma a balancear a quantidade de registros por zona. Por fim, o quarto tipo de experimento considerou um método de geração envolvendo todas as formas de geração anteriores. Neste último caso, os dados foram gerados seguindo um dos três tipos de geração (indiscriminada, estratificada ou balanceada) e, posteriormente, foi aplicado um filtro para selecionar apenas os dados mais semelhantes aos originais.

3.4.1 Geração Indiscriminada de Dados Sintéticos

A geração indiscriminada de dados sintéticos se trata da geração de dados sintéticos sem levar em consideração as zonas criadas após o treinamento da CTGAN. Foram colhidas as amostras sem nenhuma condição aplicada sobre as zonas em que seriam gerados os dados. Esse experimento teve como objetivo servir como resultado base para avaliar o impacto da geração de dados sintéticos na qualidade da predição dos modelos localizadores.

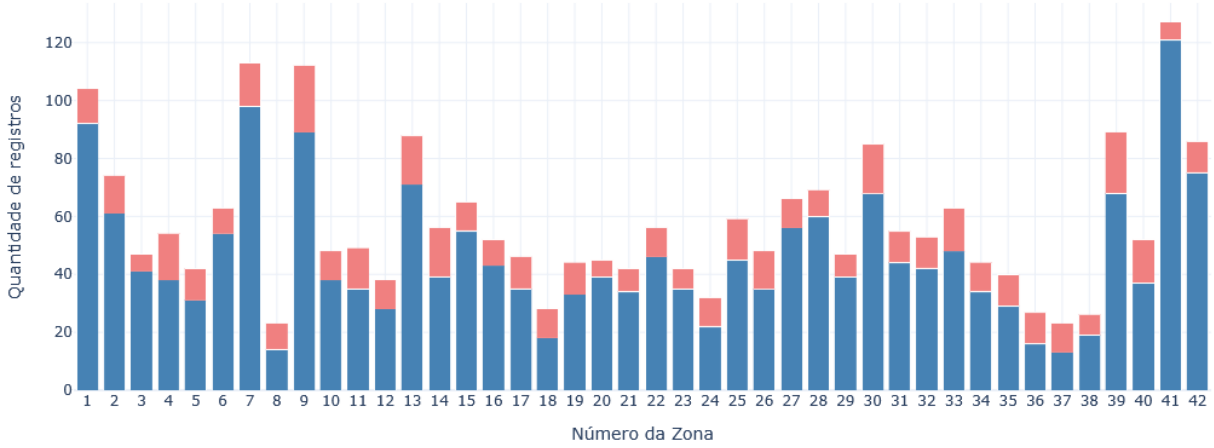
Para o caso desse experimento, foram feitas três gerações para cada conjunto de dados de treinamento. As gerações foram feitas de forma que, ao serem adicionados aos dados de treinamento originais, os dados sintéticos iriam compor uma certa porcentagem dos dados totais. Essas porcentagens foram: 20%, 40% e a mesma porcentagem alcançada pelo geração balanceada, que foi de aproximadamente 54% para o conjunto UJIIndoorLoc e 62% para o conjunto da UFPE, respectivamente. A Figura 11 ilustra uma distribuição de dados gerados indiscriminadamente pelas zonas.

Os conjuntos de dados gerados a partir desse experimento foram então utilizados no treinamento dos modelos de predição, para que seus resultados fossem comparados.

3.4.2 Geração Estratificada de Dados Sintéticos

A geração estratificada de dados sintéticos se trata da geração de dados sintéticos seguindo a mesma proporção de registros por zona da distribuição original, a partir da geração condicional da CTGAN. Esse experimento procurou avaliar se a geração de dados sintéticos por zona conforme a distribuição original poderia ser benéfico no treinamento dos modelos localizadores. Isso porque evitaria que grandes quantidades de dados sintéticos fossem adicio-

Figura 11 – Geração **indiscriminada** de dados sintéticos: Para cada zona, são indicadas as quantidades de amostras baseadas em dados reais (em azul) e em dados sintéticos (em vermelho).



Fonte: Elaborada pelo autor (2025).

nados a zonas pouco representadas, o que em teoria poderia dificultar a predição dos modelos para aquela zona, já que seriam treinados com dados majoritariamente sintéticos.

Para o caso desse experimento, assim como na geração indiscriminada, foram feitas três gerações para cada conjunto de dados de treinamento. As gerações foram feitas de forma que, ao serem adicionados aos dados de treinamento originais, os dados sintéticos iriam compor uma certa porcentagem dos dados totais.

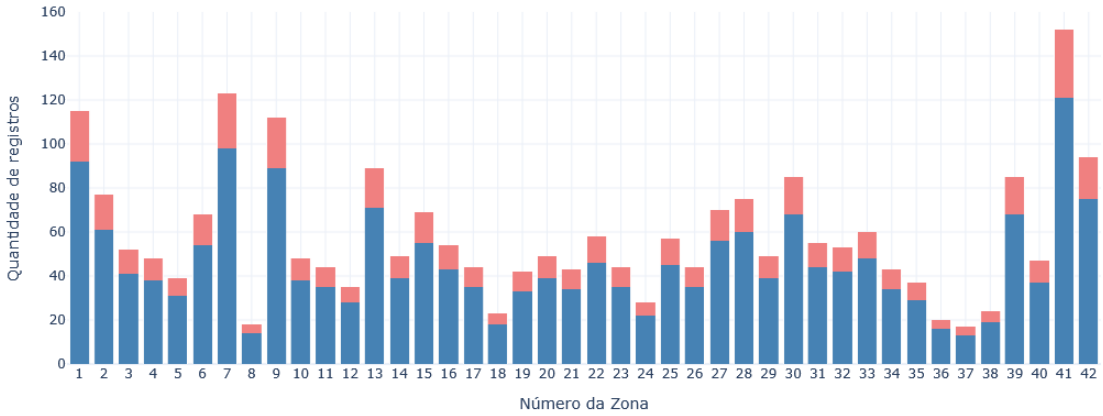
As porcentagens escolhidas foram as mesmas do experimento anterior: 20%, 40% e a mesma porcentagem alcançada pelo geração balanceada, que foi de aproximadamente 54% para o conjunto UJIIndoorLoc e 62% para o conjunto da UFPE, respectivamente. A Figura 12 mostra uma distribuição de dados gerados de maneira estratificada pelas zonas.

Os conjuntos de dados gerados a partir desse experimento foram então utilizados no treinamento dos modelos de predição, para que seus resultados fossem comparados.

3.4.3 Geração Balanceada de Dados Sintéticos

A geração balanceada de dados sintéticos se trata da geração de dados sintéticos de forma que, ao serem adicionados aos dados de treinamento originais, todas as zonas teriam a mesma quantidade de dados. Isso foi atingido ao se contar a quantidade de registros presentes em

Figura 12 – Geração **estratificada** de dados sintéticos: Para cada zona, são indicadas as quantidades de amostras baseadas em dados reais (em azul) e em dados sintéticos (em vermelho).



Fonte: Elaborada pelo autor (2025).

cada zona e gerar, a partir da geração condicional da CTGAN, para cada zona, exatamente a quantidade de registros necessária para aquela zona atinja o número de registros da zona com mais registros. Esse experimento procurava avaliar se, ao se balancear a quantidade de registros por zona, garantindo que todas as zonas tivessem a mesma representatividade durante o treinamento, o modelo treinado seria melhor preparado para generalizar e não ser tendencioso com relação a localizações em zonas majoritárias.

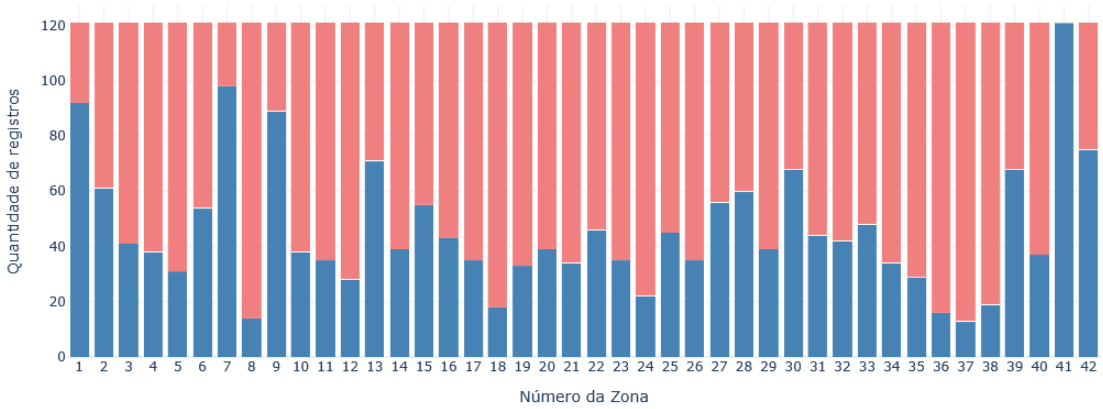
Para o caso desse experimento, foi feita uma geração para cada conjunto de dados de treinamento. Para o conjunto UJIIndoorLoc, isso significou que, ao final da geração, aproximadamente 54% dos dados eram sintéticos. Já para o conjunto da UFPE a proporção foi de aproximadamente 62% de dados sintéticos. A Figura 13 indica uma distribuição de dados gerados de maneira balanceada pelas zonas.

Os conjuntos de dados gerados a partir desse experimento foram então utilizados no treinamento dos modelos de predição, para que seus resultados fossem comparados.

3.4.4 Geração Seletiva

A geração seletiva foi uma extensão feita em conjunto com os outros métodos de geração apresentados. Na geração seletiva, a quantidade inicial a ser gerada por um determinado era multiplicada por 20 e sem seguida era aplicado um método que seleciona essas amostras

Figura 13 – Geração **balanceada** de dados sintéticos: Para cada zona, são indicadas as quantidades de amostras baseadas em dados reais (em azul) e em dados sintéticos (em vermelho).



Fonte: Elaborada pelo autor (2025).

para que apenas as melhores fossem utilizadas. O objetivo desse filtro aplicado pela geração seletiva é de procurar evitar que os dados sintéticos introduzidos no treinamento possam acabar confundindo o modelo e piorando a capacidade de predição, principio também aplicado em (NJIMA et al., 2021).

O método utilizado ordena o conjunto de dados gerados a partir da distância Euclidiana média de cada amostra com relação a um conjunto de dados reais. Para o caso dos experimentos de geração indiscriminada, esse conjunto de dados reais se trata do conjunto inteiro de treinamento. Já para as gerações balanceada e estratificada, o conjunto real a ser utilizado na comparação é o subconjunto da zona para qual foi gerado aquele conjunto de dados sintéticos.

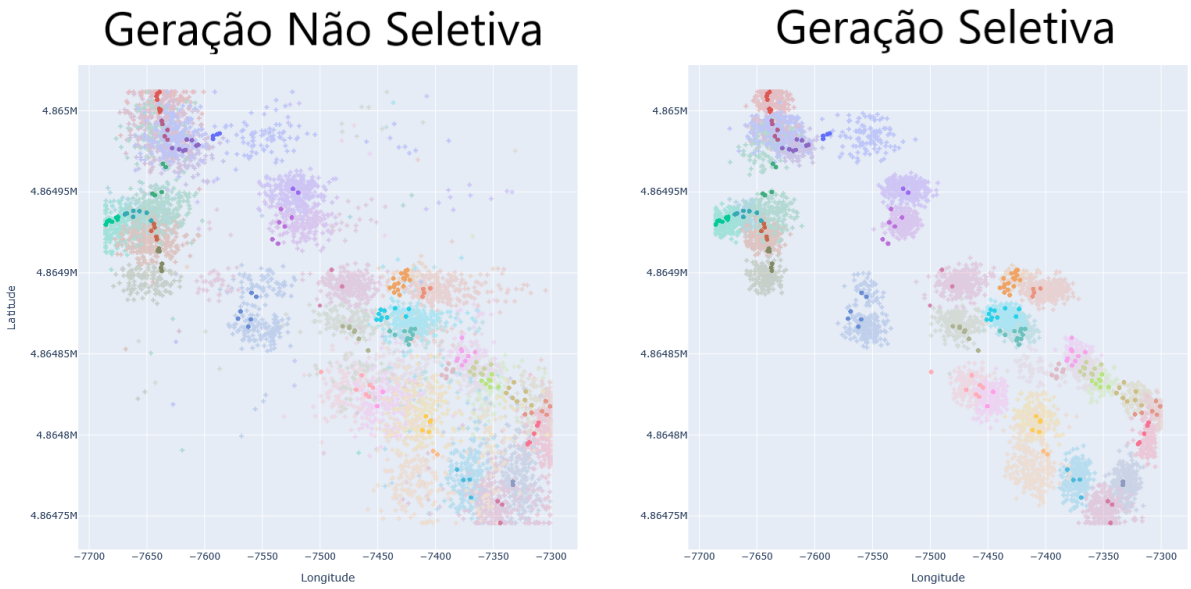
A partir da ordenação desses dados sintéticos, são filtrados apenas os n registros mais próximos, a depender da quantidade necessária de dados necessários para aquele experimento. A Figura 14 mostra uma comparação da geração de dados de forma seletiva e não seletiva.

Os conjuntos de dados gerados a partir desse experimento foram então utilizados no treinamento dos modelos de predição, para que seus resultados fossem comparados.

3.4.5 Treinamento de preditores

Para o treinamento dos modelos preditores (MLP, SVR e XGBoost), foi utilizada a biblioteca *Optuna* para a escolha do melhor conjunto de hiperpâmetros para cada treinamento. A

Figura 14 – Exemplo de distribuição de dados gerados de forma não seletiva versus seletiva. Os pontos mais escuros são os dados reais e os mais claros são dados sintéticos.



Fonte: Elaborada pelo autor (2025).

função objetivo a ser minimizada pela biblioteca avalia o modelo por *cross-validation* usando *k-fold* de 5 divisões e calcula o MSE negativo. A partir do MSE negativo das 5 divisões do *k-fold*, essa função retorna a média dos MSEs em valor absoluto, que será usada pelo Optuna para encontrar a melhor configuração de hiperparâmetros. Os hiperparâmetros estudados são apresentados na Tabela 2

Depois da validação e escolha dos melhores hiperparâmetros, os modelos foram treinados e testados com o conjunto de testes. No total, houve o treinamento de 90 modelos, sendo seis modelos de base treinados somente com dados reais, 36 modelos treinados com dados sintéticos de geração indiscriminada seletiva e não seletiva adicionados, 36 modelos treinados com dados sintéticos de geração estratificada seletiva e não seletiva adicionados e, por fim, 12 modelos treinados com dados sintéticos de geração balanceada seletiva e não seletiva adicionados. Os resultados obtidos desses experimentos serão descritos na próxima Seção.

3.4.6 Métricas

Para a avaliação dos experimentos, foram utilizadas três métricas: o erro quadrático médio (do inglês, *mean square error* – MSE) e o erro de predição de distância em metros, do qual foi avaliada a média $\bar{\epsilon}$ e o desvio padrão σ_ϵ.

Define-se o erro quadrático médio como a média da diferença entre os valores predito $\hat{y}$ e real y elevada ao quadrado, dado por

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ . \tag{3.3}$$

O fato dessa diferença ser elevada ao quadrado exacerba as diferenças e faz com que sejam mais fortemente penalizados valores previstos muito diferentes dos reais. Por isso, o MSE foi utilizado como métrica para a otimização dos hiperparâmetros dos modelos. Entretanto, um ponto negativo do MSE é que a unidade é distorcida por ser elevada ao quadrado, o que dificulta a interpretabilidade.

Por isso, foi avaliado também o erro de predição de distância, dado em metros, é a métrica mais utilizada no contexto da localização por sua fácil interpretação e clara ligação com situações reais. Para o caso do conjunto UJIIndoorLoc, o erro é calculado a partir da distância euclidiana da posição prevista $\hat{P} = (\hat{x}, \hat{y})$ e a posição real $P = (x, y)$, já que unidade dos atributos de coordenada a serem previstos está em metros no formato do WGS84. A equação que calcula distância Euclidiana é a seguinte:

$$\epsilon_1 = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2} \ . \tag{3.4}$$

A unidade dos atributos previstos conjunto da UFPE, por sua vez, é o grau decimal, que precisou ser transformado em metros para que fosse feita a comparação de forma mais

Tabela 2 – Hiperparâmetros sugeridos nos estudos do *Optuna*.

Modelo	**Hiperparâmetro**	**Estudo**
XGBoost	*learning_rate*	Número de ponto flutuante entre $0,001$ e $0,1$
	max_depth	Número inteiro entre 1 e 10
	subsample	Número de ponto flutuante entre $0,05$ e $1,0$
	colsample_bytree	Número de ponto flutuante entre $0,05$ e $1,0$
	min_child_weight	Número inteiro entre 1 e 20
MLP	*learning_rate_init*	Número de ponto flutuante entre $0,0001$ e $0,1$
	activation	Categórico entre *identity*, *tanh* e *relu*
	learning_rate	Categórico entre *constant*, *invscaling* e *adaptive*
SVR	*kernel*	Categórico entre *linear*, *poly*, *rbf* e *sigmoid*
	C	Número de ponto flutuante entre $0,01$ e $10,0$
	gamma	Categórico entre *scale* e *auto*
	epsilon	Número de ponto flutuante entre $0,01$ e $1,0$

Fonte: Elaborada pelo autor (2025).

interpretável e comparável. A aproximação foi feita utilizando a fórmula de Haversine, tal que

$$\epsilon_2 = 2R \arcsin(\sqrt{\sin^2(\frac{\phi_2 - \phi_1}{2}) + \cos(\phi_1)\cos(\phi_2)\sin^2(\frac{\lambda_2 - \lambda_1}{2})}) \ , \tag{3.5}$$

na qual ϕ_1 e λ_1 são, respectivamente, a latitude e a longitude verdadeiras; ϕ_2 e λ_2 são, respectivamente, a latitude e a longitude preditas e, por fim, R é o raio aproximado da terra, cujo valor é de 6.371.000 metros. Cabe ressaltar que todas as coordenadas geográficas aqui mencionadas são expressas em radianos.

Para os conjuntos de distâncias de erro de predição medidos, foi avaliado a média $\bar{\epsilon_1}$ e o desvio padrão σ_{ϵ_1}, para a base UJIIndoorLoc, e a média $\bar{\epsilon_2}$ e o desvio padrão σ_{ϵ_2}, para a base da UFPE.

3.5 RESULTADOS E DISCUSSÃO

O objetivo principal desse trabalho é avaliar se o uso de dados sintéticos melhora o desempenho de modelos de radiolocalização baseada em *fingerprinting* e aprendizado de máquina em datasets pequenos, visando a diminuição de custos de coleta de dados para treinamento. Ao longo dessa avaliação, surgiram quatro indagações:

1. Como se comporta o desempenho dos modelos treinados com dados sintéticos e reais quando comparado aos modelos treinados apenas com dados reais?

2. Como o aumento na proporção de dados sintéticos no conjunto de treinamento afeta os resultados?

3. Qual dos três tipos de geração de dados sintéticos (indiscriminada, estratificada ou balanceada) promove o melhor resultado?

4. Aplicar um método de seletividade aos dados sintéticos impacta os resultados?

As Tabelas 3 a 8 apresentam o MSE, além do valor médio e do desvio-padrão do erro de predição de distância, estes últimos em m, denotados, respectivamente, por $\bar{\epsilon_1}$ e σ_{ϵ_1} para a base da UJIIndoorLoc e $\bar{\epsilon_2}$ e σ_{ϵ_2} para a base da UFPE. Em todas as tabelas anteriormente mencionadas, a coluna *Ranking* indica o ranqueamento dos modelos, ordenando-os de forma crescente conforme o valor de $\bar{\epsilon}$. A partir do *Ranking*, observa-se, primeiramente, que o benefício do uso de dados sintéticos pode estar associado com o tipo de modelo de aprendizagem utilizado. O modelo SVR no conjunto da UFPE (vide Tabela 3) teve melhoria no desempenho

Tabela 3 – MSE e erro de predição de distância (ϵ) do modelo de predição SVR, considerando diferentes tipos de geração de dados sintéticos (com e sem seletividade) e percentuais distintos de dados sintéticos no conjunto de treinamento da base da UFPE. Foi acrescentado um ranking dos casos investigados.

Geração	**Seletividade**	**% Sintéticos**	**MSE**	$\bar{\epsilon_2}\ (m)$	$\sigma_2\ (m)$	**Ranking**
Indiscriminada	Não Seletivo	62%	6,03e-07	106,76	57,62	3
		40%	6,93e-07	116,19	58,17	9
		20%	6,72e-07	113,32	59,69	7
	Seletivo	62%	7,50e-07	119,39	63,92	11
		40%	8,12e-07	123,81	67,28	15
		20%	7,56e-07	123,43	57,0	14
Estratificada	Não Seletivo	62%	5,98e-07	107,62	54,8	4
		40%	6,26e-07	111,56	53,24	6
		20%	7,27e-07	116,77	64,06	10
	Seletivo	62%	**5,79e-07**	**103,28**	**58,91**	**1**
		40%	6,90e-07	113,84	62,29	8
		20%	6,56e-07	110,37	62,05	5
Balanceada	Não Seletivo	62%	7,45e-07	119,7	61,87	12
	Seletivo	62%	5,97e-07	104,26	60,8	2
	Modelo Base		7,35e-07	122,88	53,83	13

Fonte: Elaborada pelo autor (2025).

para a maioria dos experimentos utilizando dados sintéticos quando comparados ao modelo base, que ficou ranqueado entre os três piores modelos dos 15 avaliados. Já para os modelos de SVR, no conjunto UJIIndoorLoc, e MLP e XGBoost, em ambos os conjuntos, (vide Tabelas 4, 5, 6, 7 e 8), os modelos base estiveram sempre ranqueados entre os cinco melhores modelos, o que possivelmente indica que o uso de dados sintéticos com tais modelos tende a piorar o desempenho. Mesmo assim, nas Tabelas 4, 5, 6, e 7, houveram modelos com adição de dados sintéticos que foram capazes de superar o desempenho dos modelos base. Isso sugere que, ainda que os modelos tenham a tendência a piorar os resultados com o uso de dados sintéticos, é possível melhorar os seus desempenhos utilizando a técnica correta de geração, proporção de dados sintéticos utilizados no treinamento e algum critério de seletividade nos dados sintéticos adicionados ao conjunto de treinamento.

Analisando os resultados conforme as porcentagens de dados sintéticos adicionados ao treinamento, percebe-se que a influência da porcentagem de dados sintéticos depende do modelo e do conjunto de dados em questão. Foram observadas melhorias nos resultados, como no caso dos modelos SVR aplicados ao conjunto de dados da UFPE (vide Tabela 3), o que

Tabela 4 – MSE e erro de predição de distância (ϵ) do modelo de predição SVR, considerando diferentes tipos de geração de dados sintéticos (com e sem seletividade) e percentuais distintos de dados sintéticos no conjunto de treinamento da base UJIIndoorLoc. Foi acrescentado um ranking dos casos investigados.

Geração	**Seletividade**	**% Sintéticos**	**MSE**	$\bar{\epsilon_1}$ (m)	σ_1 (m)	**Ranking**
Indiscriminada	Não Seletivo	54%	1.30e+03	41.29	29.78	9
		40%	1.29e+03	41.01	30.01	8
		20%	1.16e+03	40.02	26.78	6
	Seletivo	54%	1.57e+03	45.24	32.92	15
		40%	1.54e+03	44.84	32.54	14
		20%	1.36e+03	42.36	30.43	11
Estratificada	Não Seletivo	54%	1.43e+03	43.45	31.27	13
		40%	1.25e+03	40.23	29.55	7
		20%	1.32e+03	42.82	28.24	12
	Seletivo	54%	1.16e+03	39.26	27.9	5
		40%	1.13e+03	38.57	27.74	2
		20%	1.16e+03	38.91	28.23	3
Balanceada	Não Seletivo	54%	1.31e+03	41.98	29.25	10
	Seletivo	54%	**1.13e+03**	**38.5**	**27.85**	**1**
	Modelo Base		1.14e+03	39.11	27.31	4

Fonte: Elaborada pelo autor (2025).

sugere que os dados sintéticos enriqueceram o conjunto original, favorecendo o aprendizado de padrões mais robustos, assim como piora nos resultados, tal qual o caso dos modelos XGBoost, também aplicados ao conjunto de dados da UFPE (vide Tabela 8), o que indica que os dados sintéticos inseriram ruído ou distorceram a distribuição original, prejudicando a generalização dos modelos. Houve também casos em que os resultados parecem variar com o aumento da porcentagem, como na Tabela 4, na qual a porcentagem de 40% de dados sintéticos melhorou o desempenho dos modelos de geração estratificada em relação às outras porcentagens, o que leva a crer que pode haver um limite de adição de dados sintéticos, além do qual pode se prejudicar a generalização ao invés de enriquecê-la. Tal fato leva a crer que esse é um comportamento que vale ser analisado caso a caso.

A Tabela 9 apresenta as médias dos MSEs para cada modelo, base de dados e tipo de geração de dados sintéticos. Esta tabela tem o objetivo de facilitar a comparação entre os tipos de geração de dados sintéticos para cada combinação de modelo e base de dados. Observando as médias, percebeu-se que as gerações balanceadas e estratificadas tendem a ser melhores que as gerações indiscriminadas, que obtiveram o melhor resultado em apenas uma das seis

Tabela 5 – MSE e erro de predição de distância (ϵ) do modelo de predição MLP, considerando diferentes tipos de geração de dados sintéticos (com e sem seletividade) e percentuais distintos de dados sintéticos no conjunto de treinamento da base UJIIndoorLoc. Foi acrescentado um ranking dos casos investigados.

Geração	Seletividade	% Sintéticos	MSE	$\bar{\epsilon_1}$ (m)	σ_1 (m)	Ranking
Indiscriminada	Não Seletivo	54%	8.00e+02	32.54	23.26	9
		40%	8.01e+02	34.41	20.44	12
		20%	6.41e+02	29.2	20.72	8
	Seletivo	54%	6.39e+02	28.82	21.16	7
		40%	7.08e+02	28.56	24.51	6
		20%	**4.67e+02**	**24.99**	**17.59**	**1**
Estratificada	Não Seletivo	54%	5.25e+02	27.02	17.86	4
		40%	5.02e+02	27.58	15.64	5
		20%	9.01e+02	33.11	26.58	10
	Seletivo	54%	1.21e+03	36.37	33.07	13
		40%	5.06e+02	26.49	17.62	3
		20%	9.30e+02	34.02	26.51	11
Balanceada	Não Seletivo	54%	1.06e+03	41.39	20.0	15
	Seletivo	54%	9.45e+02	37.04	22.75	14
	Modelo Base		6.35e+02	26.33	24.03	2

Fonte: Elaborada pelo autor (2025).

comparações. Já entre as gerações balanceadas e estratificadas, a balanceada se dá melhor. Em três das seis comparações, as gerações balanceadas têm um melhor resultado e em duas das seis as gerações estratificadas obtiveram um melhor resultado. Isso indica que de modo geral a geração balanceada se dá melhor do que a estratificada, que por sua vez supera a indiscriminada. Ainda assim, os resultados mostram que a escolha do tipo de geração deve ser feita caso a caso, já que cada abordagem se destacou em contextos específicos.

Por fim, os desempenhos dos modelos com geração seletiva e não seletiva de dados sintéticos foram observados. A partir dessa observação, nota-se que o método com geração seletiva melhorou os resultados em 29 dos 42 pares de experimentos, o que equivale a aproximadamente 69% dos casos. Vale ressaltar que essa melhoria não foi uniforme por tipo de geração. Para o caso das gerações indiscriminadas, a utilização do filtro de seleção melhorou o resultado em 10 dos 18 pares de experimentos. Para a geração estratificada, esse valor aumenta para 13 dos 18 pares. Por fim, a geração balanceada foi a que mais se beneficiou da seleção dos dados sintéticos, com a melhoria em todas as seis comparações de experimentos. Uma hipótese levantada para explicar o motivo da melhoria dos resultados das gerações balance-

Tabela 6 – MSE e erro de predição de distância (ϵ) do modelo de predição MLP, considerando diferentes tipos de geração de dados sintéticos (com e sem seletividade) e percentuais distintos de dados sintéticos no conjunto de treinamento da base da UFPE. Foi acrescentado um ranking dos casos investigados.

Geração	Seletividade	% Sintéticos	MSE	$\bar{\epsilon_2}$ (m)	σ_2 (m)	Ranking
Indiscriminada	Não Seletivo	62%	6,40e-07	110,79	58,15	10
		40%	6,20e-07	110,8	53,61	11
		20%	5,30e-07	97,75	58,64	5
	Seletivo	62%	**5,01e-07**	**91,75**	**62,13**	**1**
		40%	7,09e-07	114,74	64,37	13
		20%	5,05e-07	97,69	52,99	3
Estratificada	Não Seletivo	62%	7,07e-07	112,15	68,62	12
		40%	5,65e-07	101,29	59,77	7
		20%	6,10e-07	108,26	56,79	9
	Seletivo	62%	7,00e-07	116,19	60,43	14
		40%	8,57e-07	128,33	67,25	15
		20%	5,22e-07	97,73	56,73	4
Balanceada	Não Seletivo	62%	6,14e-07	102,13	67,64	8
	Seletivo	62%	5,32e-07	100,99	53,24	6
Modelo Base			4,86e-07	95,38	53,0	2

Fonte: Elaborada pelo autor (2025).

adas e estratificadas com a seleção dos dados é que a seleção feita por zonas, como foi o caso das gerações balanceada e estratificada, garante que as amostras selecionadas vão estar distribuídas por todas as zonas e terão valores próximos aos reais, o que melhora a capacidade de generalização do modelo. Por outro lado, para o caso da seleção sem levar em consideração as zonas, como foi feito na geração indiscriminada, esse filtro pode levar a um conjunto de dados sintéticos não distribuídos, concentrados em regiões específicas. Isso pode introduzir nos modelos de predição um viés que prejudica sua capacidade de generalização e sua precisão. Portanto, o que se conclui sobre a utilização de métodos de seleção de dados sintéticos é que essa é uma técnica que melhora a precisão dos modelos e garante a qualidade dos dados sintéticos gerados, mas que deve ser aplicada com atenção, dado que a seleção pode enviesar os dados de treinamento e piorar os resultados.

Tabela 7 – MSE e erro de predição de distância (ϵ) do modelo de predição XGBoost, considerando diferentes tipos de geração de dados sintéticos (com e sem seletividade) e percentuais distintos de dados sintéticos no conjunto de treinamento da base UJIIndoorLoc. Foi acrescentado um ranking dos casos investigados.

Geração	**Seletividade**	**% Sintéticos**	**MSE**	$\bar{\epsilon_1}$ (m)	σ_1 (m)	**Ranking**
Indiscriminada	Não Seletivo	54%	2.20e+02	16.66	12.76	9
		40%	2.96e+02	18.72	15.53	13
		20%	2.03e+02	15.74	12.59	6
	Seletivo	54%	2.55e+02	16.59	15.3	8
		40%	2.46e+02	15.93	15.46	7
		20%	3.16e+02	16.79	18.69	10
Estratificada	Não Seletivo	54%	2.43e+02	17.38	13.55	11
		40%	2.86e+02	18.73	14.87	14
		20%	3.22e+02	19.31	16.49	15
	Seletivo	54%	2.08e+02	15.0	13.84	4
		40%	1.75e+02	14.94	11.25	3
		20%	1.86e+02	15.05	12.07	5
Balanceada	Não Seletivo	54%	2.91e+02	18.68	15.27	12
	Seletivo	54%	**1.53e+02**	**14.15**	**10.24**	**1**
	Modelo Base		1.84e+02	14.92	12.06	2

Fonte: Elaborada pelo autor (2025).

Tabela 8 – MSE e erro de predição de distância (ϵ) do modelo de predição XGBoost, considerando diferentes tipos de geração de dados sintéticos (com e sem seletividade) e percentuais distintos de dados sintéticos no conjunto de treinamento da base da UFPE. Foi acrescentado um ranking dos casos investigados.

Geração	**Seletividade**	**% Sintéticos**	**MSE**	$\bar{\epsilon_2}$ (m)	σ_2 (m)	**Ranking**
Indiscriminada	Não Seletivo	62%	9,08e-08	39,01	26,53	15
		40%	6,45e-08	32,78	22,53	11
		20%	5,45e-08	29,54	21,45	9
	Seletivo	62%	5,53e-08	29,05	22,61	7
		40%	5,19e-08	28,22	21,82	6
		20%	4,30e-08	25,59	19,94	3
Estratificada	Não Seletivo	62%	7,31e-08	35,06	23,71	13
		40%	6,99e-08	33,35	24,51	12
		20%	5,48e-08	29,4	21,86	8
	Seletivo	62%	5,90e-08	30,46	22,73	10
		40%	4,41e-08	26,04	20,06	4
		20%	4,14e-08	25,41	19,22	2
Balanceada	Não Seletivo	62%	8,72e-08	37,75	26,67	14
	Seletivo	62%	5,03e-08	28,07	21,08	5
	Modelo Base		**3,69e-08**	**23,59**	**18,64**	**1**

Fonte: Elaborada pelo autor (2025).

Tabela 9 – Médias de MSEs, agrupadas por modelo e tipo de geração, para ambas as bases de dados.

Base de Dados	**Modelo\Geração**	*Balanceada*	*Estratificada*	*Indiscriminada*
UFPE	MLP	**5,73e-07**	6,60e-07	5,84e-07
	SVR	6,71e-07	**6,46e-07**	7,14e-07
	XGBoost	6,88e-08	**5,70e-08**	6,00e-08
UJIIndoorLoc	MLP	1,00e+03	7,62e+02	**6,76e+02**
	SVR	**1,22e+03**	1,24e+03	1,37e+03
	XGBoost	**2,22e+02**	2,37e+02	2,56e+02

Fonte: Elaborada pelo autor (2025).

4 CONCLUSÃO

O custo de coleta dos dados para construir o mapa de rádio é um dos grandes problemas da radiolocalização baseada em *fingerprint*, devido ao esforço associado para a coleta, limpeza e constante atualização de grandes quantidades de dados (NJIMA et al., 2021) (LIU et al., 2019) . Por isso, a geração de dados sintéticos é uma solução frequentemente proposta (NJIMA et al., 2021), (YEAN et al., 2021), (GRIRA; MSADAA; GRAYAA, 2023), (NABATI et al., 2020). Essa solução busca possibilitar que o treinamento de modelos de predição seja feito com uma menor quantidade de dados coletados, uma vez que dados sintéticos parecidos com os reais serão adicionados para enriquecer o treinamento. Dito isso, esse trabalho procurou investigar o impacto de técnicas diferentes de geração de dados sintéticos baseadas em redes adversariais generativas condicionais tabulares (do inglês, Conditional Tabular Generative Adversarial Networks – CTGAN) no desempenho de algoritmos de localização baseados em aprendizagem de máquina. Para isso, foram feitos experimentos utilizando a CTGAN em dois conjuntos de dados, o UJIIndoorLoc, amplamente utilizado em trabalhos de localização (YEAN et al., 2021), (FENG; NGUYEN; LUO, 2024a), (NJIMA et al., 2021), e um banco de dados de *fingerprint* de sinal colhido na UFPE, e com três modelos de predição diferentes, o MLP, o SVR e o XGBoost. Além disso, foi utilizada uma técnica de agrupamento chamada HCA, para agrupar as localizações em zonas, que foram utilizadas para a geração condicional de dados sintéticos.

Nesses experimentos, procurou-se avaliar o impacto no desempenho dos modelos depois da adição de dados sintéticos, como a proporção de dados sintéticos adicionados afeta o resultado, qual técnica de geração promove o melhor resultado e se fazer uma seleção dos dados sintéticos a serem adicionados melhora os resultados dos experimentos. Foram comparados os desempenhos dos modelos base, sem adição de dados sintéticos no treinamento, com modelos que utilizaram dados sintéticos nos treinamentos. Para a geração de dados sintéticos, foram utilizadas quatro técnicas diferentes. A primeira técnica foi a de geração indiscriminada, que produziu dados sintéticos sem considerar as zonas definidas pelo método HCA. A segunda foi a de geração estratificada dos dados sintéticos, na qual foram gerados dados sintéticos para as zonas de localização na mesma proporção da densidade de amostras naquela zona. A ideia desse método foi a de evitar que grandes quantidades de dados sintéticos fossem gerados em zonas pouco representadas, uma vez que o treinamento com dados majoritariamente sintéticos poderia piorar a precisão dos modelos de localização. A terceira técnica investigada foi a de

geração balanceada de dados sintéticos, que igualou a quantidade de dados por zona, ao se gerar exatamente a quantidade necessária para que todas as zonas tivessem a mesma quantidade de amostras que a zona mais representada. O intuito desse terceiro experimento foi observar se o modelo treinado estaria menos tendencioso às zonas majoritárias e poderia generalizar melhor. Por último, foi considerada a geração seletiva, na qual foi a geração de dados foi realizada a partir de uma das três técnicas anteriores e depois aplicado um filtro para manter apenas dos dados sintéticos mais semelhantes aos reais, visando diminuir a confusão que poderia ser causada por dados sintéticos pouco realistas.

Os resultados obtidos indicaram que a geração de dados sintéticos é capaz de melhorar o desempenho de modelos de predição, principalmente aqueles baseados em SVR. Observou-se também que o impacto do aumento da proporção de dados sintéticos adicionados aos reais varia conforme o modelo preditor e a base utilizada, sem um padrão claro de melhora ou piora. Além disso, em geral a geração balanceada melhorou os resultados frente à geração estratificada de dados sintéticos, que por sua vez se saiu melhor que geração indiscriminada. Com relação à geração seletiva, notou-se que a aplicação dos filtros melhora a qualidade dos dados gerados de forma condicional (geração estratificada e balanceada) e piora a qualidade dos dados gerados de forma indiscriminada. Com isso, conclui-se que a aplicação de método de seleção resulta em conjuntos de dados sintéticos mais perecidos com os originais. Porém, se a seleção for feita sem o devido cuidado com a distribuição dos dados, ela é capaz de introduzir dados enviesados no treinamento e piorar os resultados.

Uma possível extensão desse trabalho consiste na investigação de conjuntos de dados ainda menores ou na mudança drástica da proporção de dados de treino e teste de forma que o conjunto de dados utilizados para o treinamento seja reduzido. Tal mudança permitiria entender se esses comportamentos se repetem com bases de dados menores. Podem ser explorados também novos modelos de predição mais avançados, como, por exemplo, *transformers* (NGUYEN; LE; HAVINGA, 2023) ou redes neurais convolucionais (QIN; ZUO; WANG, 2021).

REFERÊNCIAS

AL-TAHMEESSCHI, A.; TALVITIE, J.; LÓPEZ-BENÍTEZ, M.; RUOTSALAINEN, L. Deep learning-based fingerprinting for outdoor ue positioning utilising spatially correlated rsss of 5g networks. In: IEEE. *2022 International Conference on Localization and GNSS (ICL-GNSS)*. [S.l.], 2022. p. 1–7.

BAHL, P.; PADMANABHAN, V. N. Radar: An in-building rf-based user location and tracking system. In: IEEE. *Proceedings IEEE INFOCOM 2000. Conference on computer communications. Nineteenth annual joint conference of the IEEE computer and communications societies (Cat. No. 00CH37064)*. [S.l.], 2000. v. 2, p. 775–784.

BOCK, T. *What is Hierarchical Clustering?* 2025. Disponível em: <https://www.displayr.com/what-is-hierarchical-clustering/>. Acesso em: 28 fev. 2025.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.

CHILAMKURTHY, K. *O que é Wasserstein Distance (Distância de Wasserstein)?* 2020. Disponível em: <https://kowshikchilamkurthy.medium.com/wasserstein-distance-contraction-mapping-and-modern-rl-theory-93ef740ae867>. Acesso em: 07 mar. 2025.

CHONG, A.-M.-S.; YEO, B.-C.; LIM, W.-S. Integration of uwb rss to wi-fi rss fingerprinting-based indoor positioning system. *Cogent Engineering*, Taylor & Francis, v. 9, n. 1, p. 2087364, 2022.

DHINAKARAN, A. *Understanding KL Divergence*. 2023. Disponível em: <https://medium.com/towards-data-science/understanding-kl-divergence-f3ddc8dff254>. Acesso em: 28 fev. 2025.

FENG, X.; NGUYEN, K. A.; LUO, Z. A review of open access wifi fingerprinting datasets for indoor positioning. *IEEE Access*, IEEE, 2024.

FENG, X.; NGUYEN, K. A.; LUO, Z. A wifi rss-rtt indoor positioning system using dynamic model switching algorithm. *IEEE Journal of Indoor and Seamless Positioning and Navigation*, IEEE, 2024.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. *Advances in neural information processing systems*, v. 27, 2014.

GRIRA, H.; MSADAA, I. C.; GRAYAA, K. Enhancing fingerprinting indoor positioning systems through hierarchical clustering and gan-based cnn. In: IEEE. *2023 IEEE Symposium on Computers and Communications (ISCC)*. [S.l.], 2023. p. 1054–1057.

JAISWAL, S. *Perceptrons multicamadas em aprendizado de máquina: Um guia abrangente*. 2024. Disponível em: <https://www.datacamp.com/pt/tutorial/multilayer-perceptrons-in-machine-learning>. Acesso em: 28 fev. 2025.

JONDHALE, S. R.; MOHAN, V.; SHARMA, B. B.; LLORET, J.; ATHAWALE, S. V. Support vector regression for mobile target localization in indoor environments. *Sensors*, MDPI, v. 22, n. 1, p. 358, 2022.

KARANAM, C. R.; KORANY, B.; MOSTOFI, Y. Magnitude-based angle-of-arrival estimation, localization, and target tracking. In: IEEE. *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. [S.l.], 2018. p. 254–265.

LITTLE, C.; ELLIOT, M.; ALLMENDINGER, R.; SAMANI, S. Generative adversarial networks for synthetic data generation: A comparative study. *arXiv preprint arXiv.2112.01925*, 12 2021.

LIU, X.; CEN, J.; ZHAN, Y.; TANG, C. An adaptive fingerprint database updating method for room localization. *IEEE Access*, IEEE, v. 7, p. 42626–42638, 2019.

NABATI, M.; NAVIDAN, H.; SHAHBAZIAN, R.; GHORASHI, S. A.; WINDRIDGE, D. Using synthetic data to enhance the accuracy of fingerprint-based localization: A deep learning approach. *IEEE Sensors Letters*, IEEE, v. 4, n. 4, p. 1–4, 2020.

NGUYEN, S. M.; LE, D. V.; HAVINGA, P. J. Learning the world from its words: Anchor-agnostic transformers for fingerprint-based indoor localization. In: IEEE. *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. [S.l.], 2023. p. 150–159.

NJIMA, W.; CHAFII, M.; CHORTI, A.; SHUBAIR, R. M.; POOR, H. V. Indoor localization using data augmentation via selective generative adversarial networks. *IEEE access*, IEEE, v. 9, p. 98337–98347, 2021.

QIN, F.; ZUO, T.; WANG, X. Ccpos: Wifi fingerprint indoor positioning system based on cdae-cnn. *Sensors*, MDPI, v. 21, n. 4, p. 1114, 2021.

SINGH, N.; CHOE, S.; PUNMIYA, R.; KAUR, N. Xgbloc: Xgboost-based indoor localization in multi-building multi-floor environments. *Sensors*, MDPI, v. 22, n. 17, p. 6629, 2022.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and computing*, Springer, v. 14, p. 199–222, 2004.

THAKUR, A. *How to Evaluate GANs using Frechet Inception Distance (FID)*. 2022. Disponível em: <https://wandb.ai/ayush-thakur/gan-evaluation/reports/How-to-Evaluate-GANs-using-Frechet-Inception-Distance-FID---Vmlldzo0MTAxOTI>. Acesso em: 28 fev. 2025.

TORRES-SOSPEDRA, J.; MONTOLIU, R.; MARTÍNEZ-USÓ, A.; AVARIENTO, J. P.; ARNAU, T. J.; BENEDITO-BORDONAU, M.; HUERTA, J. Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems. In: IEEE. *2014 international conference on indoor positioning and indoor navigation (IPIN)*. [S.l.], 2014. p. 261–270.

WENG, L. From gan to wgan. *arXiv preprint arXiv:1904.08994*, 2019.

XU, L.; SKOULARIDOU, M.; CUESTA-INFANTE, A.; VEERAMACHANENI, K. Modeling tabular data using conditional gan. arxiv 2019. *arXiv preprint arXiv:1907.00503*, v. 1, 2019.

YEAN, S.; SOMANI, P.; LEE, B.-S.; OH, H. L. Gan+: Data augmentation method using generative adversarial networks and dirichlet for indoor localisation. In: *IPIN-WiP*. [S.l.: s.n.], 2021.

YOUSSEF, M.; AGRAWALA, A. The horus wlan location determination system. In: *Proceedings of the 3rd international conference on Mobile systems, applications, and services*. [S.l.: s.n.], 2005. p. 205–218.