1 Avaliação da Geração Sintética de Séries Temporais (TSG)

A Geração Sintética de Séries Temporais (TSG) é fundamental em diversas aplicações, incluindo aumento de dados, detecção de anomalias e preservação de privacidade. A avaliação das metodologias de TSG é crucial para garantir que os dados gerados sejam representativos e úteis para aplicações práticas. O objetivo do trabalho é avaliar a qualidade das séries temporais geradas, abordando questões fundamentais como: como determinar a qualidade de uma série sintética? O que caracteriza uma série sintetizada como "boa"? E quais métodos são mais eficazes para essa avaliação? Dentro das metodologias encontradas para a avaliar a qualidade dos dados temporais gerados estão:

TSGBench: Time Series Generation Benchmark: é um benchmark de geração de séries temporais composto por três módulos: (1) uma coleção de datasets reais adaptados para TSG e um pipeline padronizado de pré-processamento; (2) um conjunto completo de medidas de avaliação, incluindo avaliações baseadas em distância e ferramentas de visualização; (3) um teste de generalização com Adaptação de Domínio (DA). Foram realizados experimentos em dez datasets multivariados, abrangendo domínios como tráfego, finanças, energia e medicina. As medidas de avaliação incluem Euclidean Distance, DTW, Discriminative Score (DS), Predictive Score (PS), MDD, C-FID, ACD, SD, KD, e o tempo de treinamento, além de visualizações como t-SNE e Distribution Plot. A Tabela 1 apresenta as medidas avaliadas no TSGBench, organizadas por categorias específicas para avaliar diferentes aspectos das séries temporais geradas.

Tabela 1: Medidas avaliadas no TSBench

Categoria	**Medida de Avaliação**	**Descrição**
Medidas Baseadas em Distância	Euclidean Distance (ED)	Mede a distância euclidiana entre as séries originais e geradas, considerando cada ponto como uma dimensão.
Medidas Baseadas em Distância	Dynamic Time Warping (DTW)	Um método que mede a distância entre duas séries temporais, permitindo o alinhamento óptimo entre elas, independente do ritmo ou tempo.
Medidas de Discriminação	Discriminative Score (DS)	Utiliza um modelo de classificação de séries temporais para distinguir entre séries reais e geradas. Um classificador RNN é treinado com essas labels, e o erro de classificação no conjunto de teste mede a fidelidade do modelo.
Medidas de Previsão	Predictive Score (PS)	Avalia um modelo de previsão de séries temporais treinado com dados sintéticos. A performance é avaliada usando a média do erro absoluto (MAE) no conjunto original.

Medidas de Distribuição	Marginal Distribution Difference (MDD)	Compara as distribuições marginais das séries geradas e originais, calculando a diferença média absoluta entre os histogramas empíricos das duas séries para cada dimensão e passo de tempo.
Medidas de Contexto	Contextual-FID (C-FID)	Mede como as séries geradas se ajustam ao contexto local das séries originais, utilizando embeddings de séries temporais.
Medidas de Autocorrelação	AutoCorrelation Difference (ACD)	Compara as funções de autocorrelação das séries originais e geradas, avaliando como as dependências temporais são preservadas.
Medidas de Assimetria	Skewness Difference (SD)	Mede a assimetria das distribuições das séries originais e geradas. A diferença de skewness entre as duas séries é calculada para avaliar a fidelidade.
Medidas de Kurtosis	Kurtosis Difference (KD)	Avalia o comportamento das caudas das distribuições, calculando a diferença de kurtosis entre as séries originais e geradas.
Medidas de Eficiência	Training Time	Refere-se ao tempo total necessário para treinar o método de geração de séries temporais, importante para a viabilidade econômica e eficiência do método.
Medidas de Visualização	t-SNE	Uma técnica de visualização que ajuda a entender a distribuição das séries geradas em relação às originais, representando-as em um espaço bidimensional.
Medidas de Visualização	Distribution Plot	Um gráfico que mostra a diferença entre as séries de entrada e geradas em termos de densidade, dispersão e tendência central.

How to evaluate the quality of the synthetic data – measuring from the perspective of fidelity, utility, and privacy: (Como Avaliar a Qualidade dos Dados Sintéticos – Medindo a partir das Perspectivas de Fidelidade, Utilidade e Privacidade). O artigo aborda três componentes principais na análise de dados sintéticos: fidelidade, utilidade e privacidade. A fidelidade mede o quão bem os dados sintéticos replicam as propriedades estatísticas dos dados reais, garantindo a integridade estrutural. A utilidade avalia a aplicabilidade prática dos dados sintéticos, como seu desempenho em modelos de aprendizado de máquina. Por fim, a privacidade assegura que a criação dos dados sintéticos não comprometa informações sensíveis, utilizando técnicas como privacidade diferencial. A tabela 2 apresenta as medidas principais para cada componente, fornecendo uma visão geral do processo de avaliação.

Tabela 2 Medidas de Avaliação e suas Descrições para o artigo "Como Avaliar a Qualidade dos Dados Sintéticos – Medindo a partir das Perspectivas de Fidelidade, Utilidade e Privacidade"

Fidelidade	**Nome da Métrica**	**Descrição da Métrica/Medida**
Comparação Estatística	Comparações Estatísticas Exploratórias	Compara medidas estatísticas chave (média, mediana, desvio padrão, valores ausentes) entre os dados originais e sintéticos.
Similaridade do Histograma	Similaridade do Histograma	Avalia a sobreposição das distribuições marginais entre as features dos dados originais e sintéticos.
Informação Mútua	Pontuação de Informação Mútua	Mede a dependência mútua entre duas features, indicando preservação de relações no conjunto de dados sintéticos.
Correlação	Pontuação de Correlação	Verifica se as correlações entre variáveis no dataset original foram preservadas no dataset sintético.
Autocorrelação	Autocorrelação e Correlação Parcial	Mede a qualidade da preservação das autocorrelações em dados sequenciais no conjunto de dados sintéticos.

Privacidade	**Nome da Métrica**	**Descrição da Métrica/Medida**
Cópia Exata	Pontuação de Correspondência Exata	Avalia se algum registro do conjunto original foi copiado diretamente no conjunto de dados sintéticos.
Privacidade de Vizinhança	Pontuação de Privacidade de Vizinhança	Mede a semelhança entre registros sintéticos e reais, identificando potenciais riscos de vazamento de privacidade.

Ataque de Inferência de Membro	Pontuação de Inferência de Membro	Verifica a probabilidade de um ataque de inferência de membros determinar que um registro fazia parte dos dados originais.
Utilidade	**Nome da Métrica**	**Descrição da Métrica/Medida**
Pontuação de Predição	Pontuação TSTR e TRTR	Compara a performance de modelos treinados com dados sintéticos e originais em tarefas de predição.
Importância da Feature	Pontuação de Importância da Feature	Compara a ordem de importância das features entre dados sintéticos e originais em tarefas de predição.
QScore	QScore	Avalia a similaridade de resultados de consultas de agregação aplicadas aos dados sintéticos e originais.

Evaluation is key: a survey on evaluation measures for synthetic time series(A Avaliação é Fundamental: uma Revisão sobre Medidas de Avaliação para Séries Temporais Sintéticas) Os autores realizaram uma revisão das medidas de avaliação para a geração de séries temporais, com o objetivo de auxiliar os usuários na escolha das mais adequadas. Organizam as medidas em uma taxonomia que facilita sua seleção e aplicabilidade, destacando a ausência de um padrão universal para esses procedimentos, o que dificulta o progresso na área. O estudo propõe uma revisão de 83 medidas extraídas de 56 trabalhos, oferecendo uma taxonomia estruturada e uma análise do uso dessas medidas na literatura. A figura 1 apresenta taxonomia da avaliação de sínteses de séries temporais. Ela inclui a qualidade como o objetivo geral, critérios que representam aspectos da qualidade, e, por fim, medidas que as quantificam. As medidas estão codificadas por cores, onde laranja representa qualitativas, ciano representa quantitativas, e cinza representa propósitos secundários.

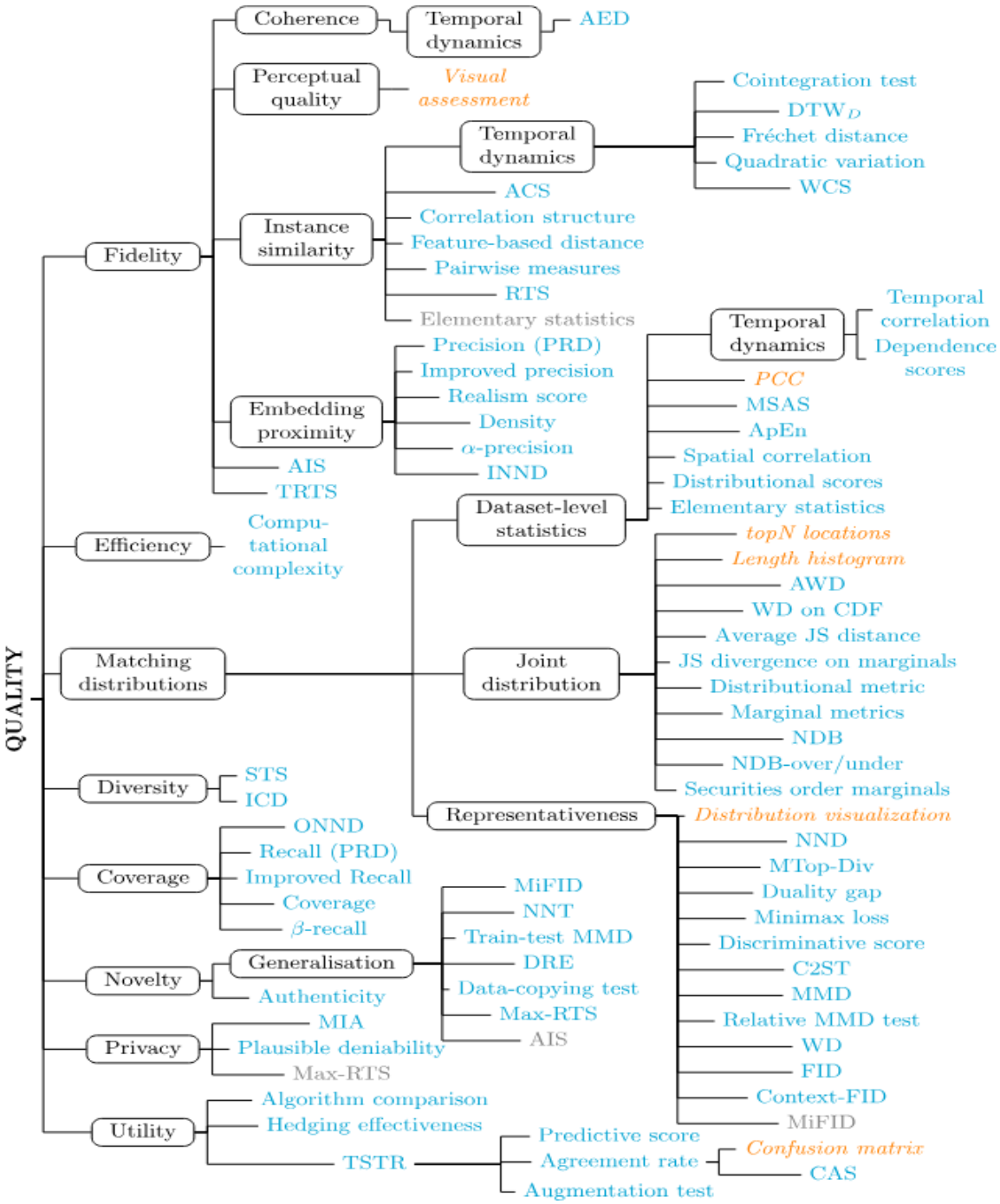


Fig1. Taxonomia da avaliação de sínteses de séries temporais proposto no artigo Evaluation is key: a survey on evaluation measures for synthetic time series

A Methodology for Validating Diversity in Synthetic Time Series Generation: (Uma Metodologia para Validar a Diversidade na Geração de Séries Temporais Sintéticas) Neste artigo, apresentam uma abordagem para gerar e avaliar muitos dados de séries temporais. O algoritmo de construção e a estrutura de validação são descritos em detalhes, juntamente com uma análise do nível de diversidade presente no conjunto de dados sintéticos.

Tabela 3 Medidas de Avaliação e suas Descrições para o artigo “uma Metodologia para Validar a Diversidade na Geração de Séries Temporais Sintéticas”

Medida	Descrição
Agreement rate	Taxa de concordância entre a distribuição gerada e a real, utilizada para avaliar a precisão de modelos de geração.
Algorithm comparison	Comparação de diferentes algoritmos em termos de desempenho na geração de dados.
AIS	Medida de Similaridade de Distribuição que utiliza modelos de decodificação.
ApEn (TS)	Entropia aproximada, utilizada para avaliar a complexidade e a aleatoriedade de séries temporais.
Augmentation test	Teste que avalia a robustez de um modelo em uma tarefa específica após a aplicação de técnicas de aumento de dados.
ACS	Medida que avalia a complexidade de séries temporais em relação a uma função de transformação.
Average JS distance	Distância média de Jensen-Shannon entre distribuições, utilizada para medir a similaridade entre séries temporais.
AED	Distância Euclidiana Absoluta, que mede a diferença entre séries temporais em um espaço de duas dimensões.
AWD seasonal	Avaliação de séries temporais sazonais, que considera a variação ao longo do tempo.
CAS	Medida que avalia a complexidade de um sistema por meio de suas distribuições.
C2ST	Teste de consistência entre duas distribuições que avalia a sua similaridade.
Computational complexity	Avaliação da complexidade computacional de um gerador de dados.
Confusion matrix	Matriz que representa o desempenho de um modelo de classificação, mostrando as previsões corretas e incorretas.
Context-FID	Medida de distância que considera o contexto das amostras em uma distribuição.
Correlation structure	Avaliação da estrutura de correlação entre múltiplas séries temporais.
Coverage	Avaliação da cobertura das distribuições geradas em relação às reais.
Data-copying test	Teste que avalia a capacidade de um modelo em copiar dados dentro de uma distância específica.
Density	Avaliação da densidade das distribuições geradas em comparação com as reais.
Dependence scores	Medidas que avaliam a dependência estatística em séries temporais.
Discriminative score	Medida que avalia a capacidade de um modelo em diferenciar entre classes em uma distribuição.
DRE	Avaliação de Redes Geradoras de Dados que mede a qualidade das amostras geradas.
Distribution visualization	Representação visual da distribuição de dados gerados em comparação com dados reais.
Distributional metric	Medidas que avaliam a similaridade entre distribuições.
Distributional scores	Medidas que avaliam características específicas de distribuições em relação a uma série temporal.
Duality gap	Medida que avalia a diferença entre a distribuição real e a gerada em modelos GAN.

Feature-based correlation	Análise de correlação baseada em características de séries temporais multivariadas.
FID images	Distância de Fréchet aplicada a imagens, utilizada para comparar distribuições de imagens geradas e reais.
Hedging effectiveness	Avaliação da eficácia de estratégias de hedge em finanças.
Improved precision	Medida que avalia o aumento da precisão em modelos de classificação ou geração.
Improved recall	Medida que avalia o aumento da capacidade de recuperação em modelos de classificação ou geração.
ICD	Medida que avalia a capacidade de um modelo de gerar dados que seguem uma distribuição específica.
JS divergence on marginals	Divergência de Jensen-Shannon aplicada a marginais, que mede a similaridade entre distribuições.
Length histogram	Avaliação da distribuição do comprimento de séries temporais.
Marginal metrics	Medidas que avaliam características específicas de marginais em séries temporais.
MTop-Div	Medida que avalia a diversidade em um conjunto de dados gerados.
MMD	Distância máxima média, utilizada para comparar distribuições.
Max-RTS	Avaliação da robustez em séries temporais geradas.
MIA	Medida que avalia a independência das amostras geradas em relação à distribuição real.
MiFID	Medida que avalia a conformidade em mercados financeiros.
Minimax loss	Função de perda que minimiza o pior caso em modelos de geração.
MSAS	Avaliação de séries temporais em múltiplas dimensões, considerando diferentes estatísticas.
NND	Distância do vizinho mais próximo, utilizada para medir a similaridade entre amostras geradas e reais.
NDB	Teste de dependência baseado em distância, que avalia a capacidade de um modelo em manter a estrutura de dependência.
ONND	Medida de dependência que utiliza Análise de Componentes Principais (PCA) para avaliar a similaridade.
PPC	Medida que avalia a capacidade de um modelo em gerar dados realísticos em contextos reais.
PRD	Distância de representação probabilística, que mede a diferença entre distribuições geradas e reais.
Predictive score	Medida que avalia a capacidade preditiva de um modelo em gerar dados futuros.
Relative MMD test	Teste de comparação de múltiplas distribuições usando a distância máxima média.
Spatial correlation	Avaliação da correlação espacial entre diferentes séries temporais.
STS	Medida que avalia a similaridade temporal entre séries.
Temporal correlation	Avaliação da correlação ao longo do tempo em séries temporais.
topN locations	Medida que avalia a eficácia de um modelo em prever as localizações mais prováveis em um conjunto de dados.

TRTS	Avaliação de séries temporais em tarefas específicas, considerando sua adequação para a geração de dados.
TSTR	Medida que avalia a robustez em tarefas de séries temporais específicas.
WD	Distância Wasserstein, utilizada para medir a diferença entre distribuições.
WD on CDF	Medida que avalia a distância entre distribuições com base na Função de Distribuição Acumulada (CDF).
WCS	Medida que avalia a similaridade em séries temporais com base na sua estrutura.
β-recall	Medida que avalia a capacidade de um modelo em recuperar informações relevantes em séries temporais.

2. Resumo das metodologias existentes para avaliação de geração de dados sinteticos

A tabela 4 resume as medidas de avaliação para séries temporais sintéticas, organizadas em critérios que testam aspectos como fidelidade, semelhança de instâncias, proximidade de embedding, eficiência, representatividade, cobertura, novidade e generalização.

Tabela 4: resumo de medidas para a avaliação na geração ode dados

Critério de Avaliação	Descrição	Medidas
Fidelidade	Refere-se à preservação das características e padrões da série temporal original.	AIS, TRTS, Coerência, Dinâmica Temporal (AED), Qualidade Perceptual (Avaliação Visual)
Semelhança de Instâncias	Compara séries temporais reais e sintéticas em nível de amostra.	ACS, Estrutura de Correlação, Distância Baseada em Recursos, Medidas Parciais, RTS, Estatísticas Elementares, Dinâmica Temporal (Teste de Cointegração, DTWD, Distância de Fréchet, Variação Quadrática, Coerência de Wavelet)
Proximidade de Embedding	Mede a proximidade de uma amostra sintética em relação aos seus vizinhos reais em um espaço de embedding.	Precisão, Precisão Aprimorada, Pontuação de Realismo, Densidade, α-Precisão, INND
Eficiência	Avalia a qualidade dos dados sintéticos em relação ao esforço necessário para gerá-los.	Complexidade Computacional
Distribuições Correspondentes	Compara as distribuições de dados reais e sintéticos.	Estatísticas de Nível de Conjunto, Estatísticas Conjuntas, Estatísticas Elementares (Média, Correlação Espacial)
Representatividade	Avalia a similaridade geral das distribuições dos dados sintéticos em relação aos dados reais.	MMD, WD, Visualizações de Distribuição (t-SNE), C2ST, Pontos de Minimax, FID, MiFID
Cobertura	Avalia a capacidade dos dados sintéticos de cobrir a região definida pela distribuição dos dados reais.	Recall, Recall Aprimorado, Cobertura, β-recall, ONND, STS, ICD
Novidade	Refere-se à capacidade de gerar amostras que sejam significativamente diferentes das amostras de treinamento.	Autenticidade

Generalização	Avalia a capacidade de gerar amostras que sejam variações das amostras de treinamento.	Distância de Memorização (MiFID), NNT, MMD

2 Seleção e detalhamento das metodologias estudadas para avaliação de representações em cenários de múltiplas tarefas

A avaliação da geração de dados sintéticos para o estudo foi dividida em seis partes: medidas baseadas em modelos, medidas baseadas em características, visualização, algoritmos de aprendizado não supervisionado e privacidade.

2.1 Qualidade Perceptual (Avaliação Visual)

A comparação de séries temporais pode ser realizada utilizando diferentes métodos de visualização, como gráficos de linhas, FFT e espectrogramas. Cada uma dessas técnicas oferece uma perspectiva única sobre os dados:

Gráficos de Linhas: Proporcionam uma visualização direta das flutuações nos dados ao longo do tempo, permitindo a identificação de tendências, padrões sazonais e anomalias. No entanto, podem ser limitados na representação de informações de frequência.

Transformada Rápida de Fourier (FFT): Essa técnica converte a série temporal do domínio do tempo para o domínio da frequência, permitindo a análise das componentes de frequência presentes nos dados. A FFT é particularmente útil para identificar padrões periódicos, mas pode não capturar bem as variações temporais que ocorrem ao longo do tempo.

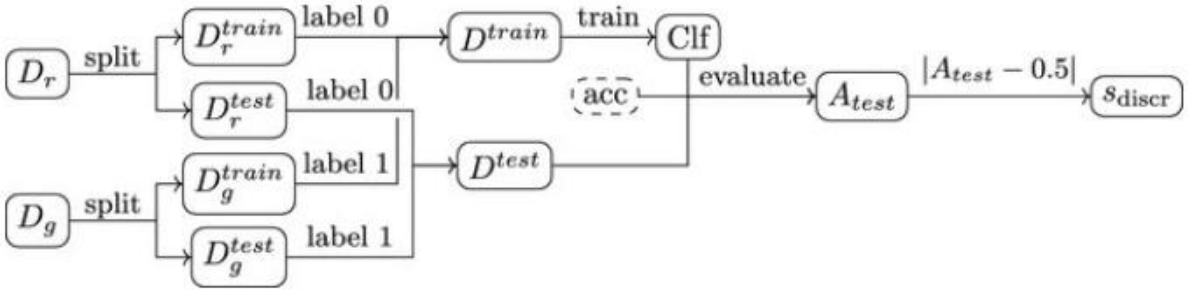
Espectrogramas: Oferecem uma representação visual da densidade espectral de potência de uma série temporal ao longo do tempo. Isso é feito através da divisão da série temporal em segmentos e da aplicação da FFT em cada segmento. Os espectrogramas são extremamente eficazes para visualizar como as frequências variam ao longo do tempo, revelando informações que podem ser invisíveis em gráficos de linha ou em análises de FFT isoladas.

2.2 Medidas Baseadas em Modelos

As medidas baseadas em modelos utilizam o esquema TSTR (Train on Synthetic, Test on Real). Neste processo, primeiro, séries temporais sintéticas são geradas para representar características que se esperam encontrar em dados reais. Em seguida, uma rede neural é treinada com essas séries sintéticas, permitindo que o modelo aprenda a identificar padrões e relações nos dados. Após o treinamento, o modelo é avaliado em sua capacidade de lidar com séries temporais originais. Essa abordagem é útil para verificar como um modelo, que foi treinado em dados simulados, se comporta quando aplicado a dados do mundo real, fornecendo insights sobre sua robustez e eficácia. Essa abordagem é crucial

em HAR, pois permite verificar como um modelo treinado em dados simulados se comporta ao ser aplicado a dados do mundo real.

Discriminative Score (DS): Essa medida utiliza um modelo de classificação de séries temporais pós-hoc com GRUs ou LSTMs de duas camadas para diferenciar entre séries originais e geradas. Cada série original recebe o rótulo "real", enquanto as geradas são rotuladas como "sintéticas". A partir dessas etiquetas, um classificador RNN é treinado, e o erro de classificação em um conjunto de teste quantifica a fidelidade do modelo gerador.



Representação do Score Discriminativo da Medida Fonte: Evaluation is key

Predictive Score (PS). Envolve treinar um modelo de predição de séries temporais pós-hoc nos dados sintéticos. Utilizando GRUs ou LSTMs, o modelo prevê os vetores temporais de cada série de entrada para os próximos passos ou o vetor inteiro. O desempenho do modelo é avaliado no conjunto de dados original com base no erro absoluto médio.

Interpretação da Pontuação: A pontuação preditiva resultante é uma métrica que indica a utilidade dos dados sintéticos. Quanto menor for a pontuação (ou seja, menor for o MAE), melhor será a qualidade do conjunto de dados sintético em replicar as características dos dados reais, permitindo que o modelo de previsão funcione de forma eficaz.

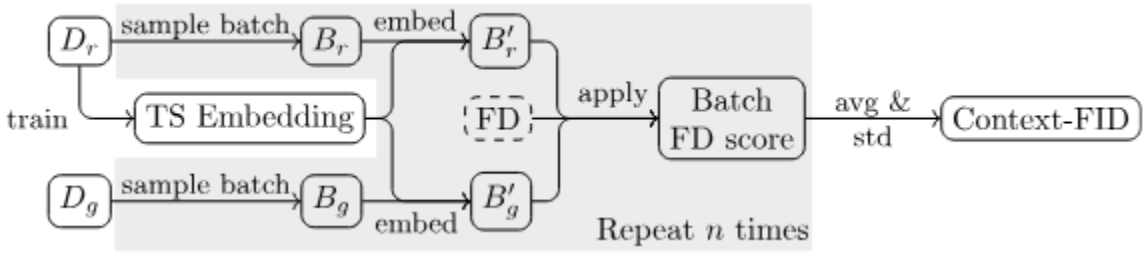
Contextual-FID (C-FID). Esta medida estende o conceito de Frechet Inception Distance (FID), usado em geração de imagens. Ela quantifica o quão bem as séries temporais sintéticas se ajustam ao contexto local das séries temporais originais, utilizando embeddings que se integram harmoniosamente com o contexto local.

Processo de Cálculo:

- Extração de Embeddings: Para calcular o C-FID, primeiro são geradas representações (embeddings) das séries temporais, tanto das originais quanto das sintéticas. Essa etapa geralmente envolve o uso de um modelo pré-treinado que pode capturar características temporais e contextuais relevantes, como uma rede neural convolucional ou recorrente, que extrai características temporais em diferentes níveis de abstração.

- Contexto Local: O C-FID foca em como essas embeddings se integram ao contexto local das séries temporais. Isso significa que ele considera sequências de tempo mais curtas ou segmentos de dados que preservam as dinâmicas e padrões temporais locais, em vez de analisar as séries como um todo. Essa abordagem permite uma avaliação mais refinada da similaridade.

-Cálculo da Distância: Assim como no FID, uma vez que as embeddings são extraídas, o C-FID calcula a distância entre as distribuições dessas representações para as séries reais e sintéticas. Isso pode incluir a comparação da média e da covariância das distribuições dos embeddings, o que permite uma avaliação abrangente de quão bem as séries sintéticas se ajustam aos padrões observados nas séries reais.



Fonte: Evaluation is key

Interpretação: Um valor baixo do C-FID indica que as séries temporais sintéticas são muito semelhantes às reais em termos de comportamento local, sugerindo que o modelo gerador foi capaz de capturar eficazmente as dinâmicas temporais importantes. Um valor alto, por outro lado, sugere que há discrepâncias significativas entre as séries geradas e reais, indicando que o modelo pode não ter capturado com precisão os padrões temporais ou as relações contextuais.

2.3 Medidas Baseadas em Características (Feature-based Measures)

Essas medidas são projetadas para capturar correlações entre séries e dependências temporais, avaliando o quão bem as séries temporais geradas preservam as características originais. Uma vantagem importante dessas medidas é sua capacidade de fornecer resultados claros e determinísticos, oferecendo uma avaliação objetiva da qualidade das séries temporais geradas.

Diferença de Distribuição Marginal (Marginal Distribution Difference MDD). Essa medida calcula um histograma empírico para cada dimensão e passo temporal na série gerada, usando os centros e larguras dos bins da série original. Em seguida, calcula a diferença absoluta média entre esse histograma e o da série original, avaliando o quão próximas as distribuições das séries originais e geradas estão.

Processo:

Criação de Histogramas Empíricos: Para cada dimensão da série temporal (por exemplo, se a série contém dados de múltiplos sensores, cada sensor representa uma dimensão) e para cada passo temporal (os diferentes instantes da série), um histograma empírico é construído. Os histogramas são formados utilizando os mesmos centros e larguras dos bins (intervalos) que foram usados para gerar os histogramas das séries originais. Isso assegura que a comparação seja feita em uma base consistente.

Cálculo da Diferença: Após a criação dos histogramas para as séries originais e geradas, a diferença absoluta média entre os valores correspondentes dos histogramas é calculada. Isso envolve comparar os valores das contagens de cada bin entre os histogramas da série original e da série gerada. A diferença absoluta média é obtida somando todas as diferenças absolutas entre as contagens nos bins dos histogramas e dividindo pelo número total de bins.

Avaliação das Distribuições: O resultado desta análise é um número que representa o quanto as distribuições das séries originais e geradas se assemelham. Um valor menor indica que as distribuições são mais semelhantes, enquanto um valor maior indica uma diferença maior entre as distribuições.

Diferença de Autocorrelação (ACD). Esta medida computa a autocorrelação tanto das séries temporais originais quanto das geradas, determinando sua diferença. A comparação das autocorrelações permite avaliar o quanto as dependências temporais são mantidas nas séries geradas.

Diferença de Assimetria (AutoCorrelation Difference SD). Além da autocorrelação, considera-se também medidas estatísticas, como a assimetria, que é essencial para a distribuição marginal de uma série temporal. A assimetria quantifica a simetria da distribuição, e a diferença entre a assimetria da série gerada e da original é usada para avaliar sua fidelidade.

Diferença de Curtose (KD). Assim como a assimetria, a curtose avalia o comportamento das caudas de uma distribuição, revelando grandes desvios da média. A diferença de curtose entre a série original e a série gerada é calculada para medir a precisão das séries geradas em relação aos extremos da distribuição.

Tempo de Treinamento (Training Time): O tempo de treinamento se refere ao tempo total necessário para treinar um método de geração de séries temporais (TSG) em termos de relógio de parede. Este é um aspecto crucial para a avaliação e implementação desses métodos, uma vez que pode influenciar diretamente decisões econômicas e práticas no uso dessas técnicas. Métodos mais rápidos são frequentemente preferidos em contextos em que recursos computacionais são limitados ou quando há uma necessidade de geração em tempo real.

2.4 Agrupamento usando Gráficos de Visualização

Oferece uma perspectiva intuitiva e interpretativa que permite comparar diretamente as estruturas e padrões entre as séries temporais originais e as geradas.

t-SNE: O t-distributed Stochastic Neighbor Embedding (t-SNE) é uma técnica popular para visualização concisa da distribuição das séries temporais geradas em comparação com as originais, em um espaço bidimensional. Essa ferramenta ajuda a identificar padrões, similaridades e discrepâncias entre as séries geradas e as reais, facilitando a comparação visual.

Gráfico de Distribuição: O gráfico de distribuição revela as diferenças entre as séries temporais de entrada e as geradas, com foco em características como densidade, dispersão e tendência central. Ele ilustra como as séries geradas se aproximam das estatísticas das séries originais, oferecendo uma visão clara sobre a fidelidade dos dados gerados.

2.5 Agrupamento: usando medidas de algoritmos não supervisionados

As medidas de aprendizado não supervisionado desempenham um papel crucial na classificação do reconhecimento de atividades humanas (HAR) e em contextos similares, pois são fundamentais para a precisão e eficácia dos modelos desenvolvidos. No reconhecimento de atividades, a identificação de padrões complexos em grandes volumes de dados, como sequências de movimentos coletados por sensores, é vital. Medidas como o Coeficiente de Silhueta, Índice de Calinski-Harabasz, Índice de Davies-Bouldin, Confiabilidade (Trustworthiness), Variância Explicada Acumulada, e Sammon's Mapping são essenciais para avaliar e garantir a qualidade dos modelos. Essas medidas ajudam a garantir que diferentes atividades sejam bem separadas, que os dados semelhantes sejam agrupados corretamente e que a integridade das relações e padrões observados seja preservada, melhorando assim a precisão e a aplicabilidade dos modelos em monitoramento de saúde, segurança e automação de processos.

Silhouette Coefficient: é uma técnica de clustering utilizada para medir a similaridade dos dados dentro de um cluster em comparação com os outros clusters. O Coeficiente de Silhueta é uma representação numérica que varia de -1 a 1. O valor 1 significa que cada cluster está completamente distinto dos outros, enquanto o valor -1 indica que todos os dados foram atribuídos ao cluster errado. Um valor 0 indica que não há clusters significativos nos dados.

Índice de Calinski-Harabasz: ou Critério de Razão de Variância, é uma métrica usada para avaliar a qualidade dos clusters, medindo a razão entre a dispersão entre os clusters e a dispersão dentro dos clusters. Basicamente, mede-se a diferença entre a soma das distâncias quadradas dos dados entre os clusters e dos dados dentro do cluster. Quanto

maior o valor do Índice de Calinski-Harabasz, melhor, pois significa que os clusters estão bem separados. No entanto, como não há um limite superior para o índice, essa métrica é mais adequada para avaliar diferentes valores de k (número de clusters) do que para interpretar o resultado isoladamente.

O Índice de Davies-Bouldin é uma métrica de avaliação de clusterização que mede a similaridade média entre cada cluster e o mais similar a ele. A similaridade é calculada pela razão entre as distâncias dentro do cluster e as distâncias entre os clusters. Quanto mais distantes e menos dispersos os clusters, melhor será o resultado. Diferente das medidas anteriores, o Índice de Davies-Bouldin busca um valor o mais baixo possível.

Variância Explicada Acumulada: (Cumulative Explained Variance - CEV) é uma métrica chave na avaliação da redução de dimensionalidade, que indica o percentual de variabilidade original dos dados retida à medida que as dimensões são reduzidas. A redução de dimensionalidade é uma técnica essencial para simplificar conjuntos de dados complexos, mantendo a maior parte da informação original. O objetivo principal é diminuir o número de variáveis (features) sem comprometer a integridade dos dados, facilitando a análise e visualização, além de melhorar a eficiência de algoritmos de aprendizado de máquina. A Análise de Componentes Principais (PCA) é um dos métodos mais usados para isso, pois transforma os dados em novas dimensões chamadas de componentes principais, que capturam a maior variabilidade possível. Geralmente, a meta é reter entre 90-95% da variância, o que pode resultar em uma redução significativa do número de features, sem perda substancial de informação.

Confiabilidade (Trustworthiness) é uma métrica que avalia a qualidade da técnica de redução de dimensionalidade, medindo quão bem a estrutura local dos dados originais foi preservada na nova dimensão. Ela verifica se os pontos que eram vizinhos no espaço original permanecem vizinhos após a redução de dimensões. A métrica varia de 0 a 1, sendo que valores próximos de 1 indicam uma preservação alta da estrutura dos dados, ou seja, os vizinhos mais próximos nas dimensões reduzidas continuam sendo vizinhos nas dimensões originais. Isso é crucial para garantir que as relações locais importantes entre os dados não sejam perdidas ao reduzir o número de variáveis.

Sammon's Mapping é uma técnica de redução de dimensionalidade não linear que visa preservar as distâncias entre pares de dados em um espaço de alta dimensionalidade ao ser reduzido para um espaço de menor dimensão. O objetivo é minimizar o erro de preservação dessas distâncias, medido pela Função de Stress de Sammon. Quanto menor o valor dessa função, melhor é a preservação das distâncias entre os pontos no espaço original e o espaço reduzido. A técnica é particularmente útil quando se busca manter a estrutura geométrica dos dados intacta durante a redução de dimensionalidade, algo importante em tarefas de visualização e análise de dados complexos. A Função de Stress de Sammon calcula a diferença entre as distâncias no espaço original e no espaço reduzido, e o valor resultante indica a qualidade da preservação.

2.6 privacidade

Com a evolução das regulamentações de privacidade, garantir a proteção de informações sensíveis tornou-se uma obrigação ética é um requisito legal para as organizações. Isso é especialmente relevante na geração de dados sintéticos a partir de dispositivos móveis para o reconhecimento de atividades, onde dados de sensores podem conter informações pessoais e sensíveis sobre os usuários. Antes que os dados sintéticos possam ser compartilhados e utilizados em diversas aplicações, é crucial avaliar medidas de privacidade que analisam o quão bem os dados sintéticos gerados mantêm a privacidade em comparação com os conjuntos de dados originais. Essa avaliação é fundamental, pois ajuda a informar as partes interessadas sobre possíveis vazamentos de informações e orienta decisões críticas sobre o compartilhamento e a aplicação de dados sintéticos.

M_16 Exact Match Score: serve como uma métrica direta e intuitiva para avaliar a privacidade. Ela quantifica o número de registros reais do conjunto de dados original que podem ser encontrados verbatim no conjunto de dados sintético. Idealmente, essa pontuação deve ser zero, indicando que nenhuma informação real foi retida em sua forma original. Essa métrica atua como uma ferramenta de triagem preliminar, permitindo que as organizações avaliem o nível de privacidade antes de se aprofundarem em avaliações mais complexas.

M_17 Neighbors' Privacy Score: avalia a proporção de registros sintéticos que são muito semelhantes aos registros originais. Embora esses registros possam não ser cópias exatas, sua proximidade em termos de semelhança representa riscos potenciais de vazamento de privacidade e pode facilitar ataques de inferência. Para calcular essa pontuação, realiza-se uma busca de vizinhos mais próximos em alta dimensão, comparando os dados sintéticos com o conjunto de dados original para identificar registros que se assemelham estreitamente.

M_18 Membership Inference Score: Mede a probabilidade de que um atacante consiga executar com sucesso um ataque de inferência de membros, que visa determinar se registros específicos fizeram parte do conjunto de dados de treinamento utilizado para gerar dados sintéticos. Uma vez que um modelo de sintetizador é treinado, ele pode produzir amostras sintéticas sem acesso posterior aos dados originais. No entanto, se a pontuação de inferência de membros for alta, isso indica um risco significativo de comprometimento da privacidade, pois os atacantes podem inferir detalhes sobre os dados originais.

3. Avaliação dos pontos fortes e limitações

A **avaliação visual da Qualidade Perceptual** oferece uma abordagem intuitiva e acessível para a interpretação de dados, permitindo a identificação rápida de padrões, tendências e anomalias em séries temporais, especialmente em contextos como o reconhecimento de atividades humanas. Essa técnica facilita a comparação entre diferentes conjuntos de dados, como séries temporais originais e sintetizadas, destacando a fidelidade dos dados gerados em relação aos reais. No entanto, a avaliação visual possui limitações, como a subjetividade, que pode influenciar a interpretação, e a simplicidade excessiva, que pode levar à omissão de informações críticas. Além disso, em conjuntos de dados muito grandes ou complexos, as visualizações podem se tornar confusas, dificultando a análise. A dependência de ferramentas e técnicas adequadas é essencial, uma vez que visualizações mal projetadas podem obscurecer informações importantes. Portanto, embora a avaliação visual seja valiosa, é fundamental complementá-la com medições quantitativas rigorosas para garantir uma análise abrangente e informada. A integração de ambas as abordagens potencializa a interpretação dos dados, resultando em decisões mais fundamentadas.

Medidas Baseadas em Modelos: As medidas baseadas em modelos, como Discriminative Score, Predictive Score, Contextual-FID e Diferença de Distribuição Marginal, são ferramentas valiosas para a avaliação quantitativa da qualidade das séries temporais geradas. Seus pontos fortes incluem a objetividade e quantificabilidade, permitindo comparações entre modelos, além da capacidade de capturar padrões complexos e relações temporais, sendo flexíveis para diversas aplicações em áreas como finanças e saúde. Contudo, essas medidas também apresentam limitações, como a sensibilidade a dados específicos, a falta de consideração de dependências temporais, a complexidade de implementação e a possibilidade de não capturarem todas as dimensões de qualidade desejadas, como diversidade e coerência. Além disso, a qualidade das avaliações pode depender das configurações do modelo utilizado, introduzindo viés nas interpretações. Portanto, embora essas métricas ofereçam uma abordagem poderosa para a análise, é essencial complementá-las com outras avaliações qualitativas e quantitativas para uma compreensão mais abrangente da eficácia dos modelos geradores.

Medidas Baseadas em Características: oferecem uma abordagem eficaz para avaliar a qualidade das séries temporais geradas, destacando suas correlações e dependências temporais. Entre suas vantagens, destaca-se a capacidade de fornecer resultados claros e determinísticos, permitindo uma avaliação objetiva das características das séries geradas em comparação com as originais. Medidas como a Diferença de Distribuição Marginal (MDD), Diferença de Autocorrelação (ACD), Diferença de Assimetria (SD) e Diferença de Curtose (KD) são valiosas para identificar o quão bem os padrões estatísticos e as propriedades das séries temporais foram preservados. O Tempo de Treinamento também é um critério relevante, uma vez que a eficiência do processo de geração pode influenciar a viabilidade prática dos modelos. Contudo, essas medidas têm limitações. A dependência

de histogramas e estatísticas simples pode não capturar nuances complexas nas relações temporais, levando a uma subavaliação da qualidade das séries geradas. Além disso, o foco em características específicas pode resultar em uma visão parcial, negligenciando outros aspectos relevantes, como a diversidade ou a coerência das séries temporais. Também é importante considerar que a interpretação das medidas pode ser influenciada por configurações de parâmetros, o que pode introduzir viés nas comparações. Portanto, enquanto as Medidas Baseadas em Características são ferramentas valiosas para avaliação, é crucial integrá-las com outras métricas e análises qualitativas para uma compreensão mais completa da eficácia dos modelos geradores.

Agrupamento usando graficos de visualização: As técnicas de visualização como t-SNE, PCA e UMAP são cruciais para a avaliação de séries temporais geradas, pois oferecem uma perspectiva intuitiva, facilitando comparações entre estruturas e padrões das séries originais e geradas. O t-SNE, em particular, destaca-se por projetar séries temporais em um espaço bidimensional, permitindo a identificação de padrões e discrepâncias, tornando as análises mais acessíveis.

Por outro lado, os gráficos de distribuição são eficazes para ilustrar diferenças nas características fundamentais das séries, como densidade, dispersão e tendência central. Eles fornecem uma visão clara da fidelidade dos dados gerados, permitindo a avaliação de quão bem as estatísticas das séries originais foram preservadas.

Entretanto, essas técnicas de visualização também apresentam limitações. A interpretação dos resultados pode ser subjetiva e dependente da experiência do analista, o que pode levar a conclusões imprecisas. Além disso, a redução de dimensionalidade do t-SNE pode ocultar informações importantes ao comprimir dados complexos em um espaço bidimensional. Gráficos de distribuição, embora informativos, podem ser enganosos se não forem acompanhados de análises estatísticas robustas que contextualizem as diferenças observadas. Portanto, embora as técnicas de visualização sejam ferramentas poderosas, é essencial utilizá-las em conjunto com outras métricas e análises quantitativas para uma avaliação mais abrangente da qualidade das séries temporais geradas.

Agrupamento usando medidas de algoritmos não supervisionados: A utilização de algoritmos não supervisionados para agrupamento oferece diversas vantagens, como a identificação e quantificação de padrões intrínsecos nos dados sem depender de rótulos. Isso possibilita uma análise exploratória robusta, revelando relações e similaridades entre atividades que poderiam passar despercebidas em análises supervisionadas. Métricas como o Coeficiente de Silhueta e o Índice de Davies-Bouldin fornecem insights valiosos sobre a qualidade da clusterização, permitindo que os pesquisadores avaliem a separação e coesão dos dados, fatores essenciais para a precisão no reconhecimento de atividades.

No entanto, essas métricas apresentam limitações significativas. A escolha de parâmetros pode resultar em inconsistências e interpretações subjetivas. Além disso, a falta de rótulos

dificulta a captura de características relevantes dos dados sintéticos, e a ausência de uma referência padrão complica a validação da qualidade dos dados gerados. A falta de robustez em relação a diferentes distribuições de dados e a dificuldade de generalizar os resultados para cenários do mundo real também são desafios notáveis.

Privacidade: A avaliação das métricas de privacidade na geração de dados sintéticos para reconhecimento de atividades humanas (HAR) é crucial para garantir a proteção de informações sensíveis, especialmente à luz das regulamentações em constante evolução. Entre os pontos fortes dessas métricas, destaca-se a capacidade de fornecer uma análise direta e intuitiva da privacidade dos dados gerados. Métricas como o Exact Match Score permitem uma triagem inicial, identificando rapidamente a retenção de informações originais nos dados sintéticos. Isso facilita a tomada de decisões informadas sobre o compartilhamento e uso desses dados, contribuindo para a conformidade legal e a ética organizacional.

Outro ponto positivo é a habilidade das métricas, como o Neighbors' Privacy Score e o Membership Inference Score, de quantificar riscos potenciais de vazamento de privacidade. Essas avaliações ajudam as organizações a entender a semelhança entre dados sintéticos e originais, bem como a probabilidade de ataques de inferência. Ao identificar registros sintéticos que se assemelham a dados reais, as organizações podem implementar medidas de mitigação mais eficazes e adaptadas a suas necessidades específicas.

No entanto, essas métricas também apresentam limitações. O Exact Match Score, por exemplo, pode não capturar nuances de privacidade em dados mais complexos, onde a simples ausência de correspondências exatas não garante a proteção contra vazamentos de informação. Da mesma forma, o Neighbors' Privacy Score pode ser influenciado por variáveis que não são capturadas na análise, como a variabilidade nos dados originais. Por fim, o Membership Inference Score, embora valioso, pode não refletir a realidade de situações onde a estrutura dos dados sintéticos é suficientemente diferente da original, resultando em uma avaliação imprecisa dos riscos.

4 Avaliação dos modelos HAR

O modelo treinado é empregado para gerar um conjunto de dados sintéticos, que, juntamente com os dados de treino e teste, é submetido a diversas técnicas de avaliação conforme definido em um arquivo de configuração. Para assegurar a eficácia do processo, foram implementadas várias medidas. A **Fig. 1** ilustra este fluxo, destacando as etapas de geração e avaliação dos dados, bem como as técnicas específicas utilizadas para mensurar a qualidade e a representatividade dos conjuntos de dados gerados. Essa abordagem sistemática permite uma análise abrangente do desempenho dos modelos e a validação dos dados sintéticos gerados.

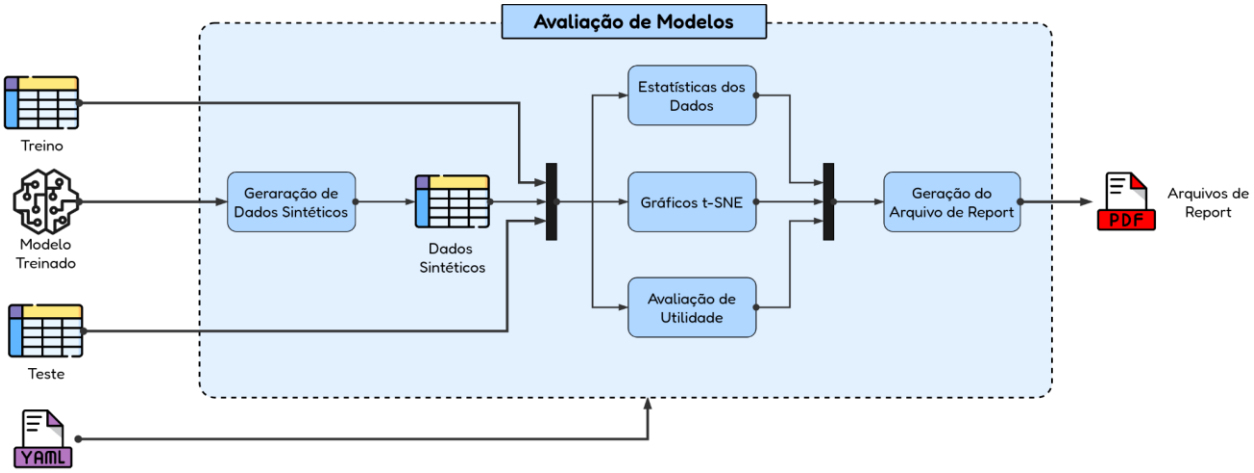


Fig 8 Fluxo para o treinamento de modelo

A classe **Evaluator** foi desenvolvida com o objetivo de estruturar e facilitar a avaliação de dados sintéticos e reais de forma sistemática. Essa classe permite a avaliação individual de conjuntos de dados, tanto reais quanto sintéticos, oferecendo uma abordagem estruturada para analisar e comparar as características de diferentes datasets.

Um dos principais recursos da classe é o treinamento de modelos de classificação. Dois classificadores são treinados com dados reais e sintéticos, possibilitando a medição de diversas métricas de desempenho. A **acurácia** é utilizada para medir a proporção de previsões corretas em relação ao total de previsões realizadas. A **precisão** avalia a proporção de verdadeiros positivos em relação ao total de positivos previstos, refletindo a qualidade das previsões positivas. O **recall** mede a capacidade do modelo de capturar todos os casos positivos, enquanto a **pontuação F1** fornece uma única métrica que reflete o equilíbrio entre precisão e recall, calculando a média harmônica entre ambas.

Além disso, a classe inclui funcionalidades para gerar gráficos que facilitam a visualização da distribuição dos dados por classe, ajudando na identificação de desbalanceamentos. As visualizações implementadas incluem gráficos de amostras, que mostram a distribuição de diferentes classes, e o método **t-SNE**, que projeta os dados em duas dimensões, permitindo a identificação de agrupamentos e padrões entre as amostras.

Outro aspecto importante da classe Evaluator é a capacidade de facilitar a comparação entre as distribuições dos dados reais e sintéticos. Isso é realizado através da **divergência de Kullback-Leibler**, que quantifica a diferença entre duas distribuições de probabilidade, e por meio de testes estatísticos que avaliam a semelhança das distribuições, assegurando que os dados sintéticos sejam representativos dos dados reais.

A classe também implementou diversas medidas de avaliação, incluindo o **Silhouette Score**, que analisa a qualidade da separação entre diferentes classes, o **Davies-Bouldin Score**, que mede a razão entre a soma das distâncias intra-cluster e as distâncias inter-cluster, e o **Calinski-Harabasz Score**, que avalia a proporção das distâncias entre clusters em relação às distâncias dentro de cada cluster. Além dessas, a classe possui métodos para calcular e visualizar as métricas, utilizando gráficos de distribuição que ilustram as diferenças nas características das séries temporais geradas e originais, focando em aspectos como densidade, dispersão e tendência central.

A tabela 13 apresenta as medidas que foram implementadas, com base nas métricas selecionadas no item 2. O objetivo é implementar todas as métricas listadas; no entanto, devido a restrições de tempo, algumas ainda não foram implementadas. A seção a seguir delineará os passos futuros necessários para completar essa implementação.

Tabela 13 Medidas de avaliação que foram implementadas

Categoria	Métrica	Implementada
Qualidade perceptual	Gráficos de Linhas	Sim
	Transformada Rápida de Fourier (FFT)	Sim
	Espectrogramas	Sim
Medidas Baseadas em Modelos	M1: Discriminative Score (DS)	Não
	M2: Predictive Score (PS)	Não
	M3: Contextual-FID (C-FID)	Não
Medidas Baseadas em Características	M4: Diferença de Distribuição Marginal (MDD)	Sim
	M5: Diferença de Autocorrelação (ACD)	Sim
	M6: Diferença de Assimetria (SD)	Sim
	M7: Diferença de Curtose (KD)	Não
	M8: Tempo de Treinamento	Não
Agrupamento usando graficos de visualização	M9: t-SNE	Sim
	M10: Gráfico de Distribuição	Sim
Medidas agrupamento usando algoritmos Não Supervisionados	M11: Coeficiente de Silhueta	Sim
	M12: Índice de Calinski-Harabasz	Sim
	M13: Índice de Davies-Bouldin	Sim
	M14: Variância Explicada Acumulada (CEV)	Não
	M15: Confiabilidade (Trustworthiness)	Não
	M16: Sammon's Mapping	Não

Privacidade	M17: Exact Match Score	Não
	M18: Neighbors' Privacy Score	Não
	M19: Membership Inference Score	Não