

Geração de Dados Sintéticos em Bases Médicas com Amostras Reduzidas para a Predição de Câncer

Yuri Lanzini¹

¹ Universidade Federal da Fronteira Sul (UFFS), Chapecó, Brasil

yuri.lanzini@estudante.uffs.edu.br

RESUMO

Durante a aplicação de técnicas de Inteligência Artificial na área da saúde, a obtenção de grandes conjuntos de dados clínicos representa um desafio devido aos custos de coleta, à disponibilidade limitada de pacientes e às restrições éticas. Essa limitação é especialmente relevante em estudos baseados em exames laboratoriais, marcadores bioquímicos e testes enzimáticos, podendo comprometer o treinamento e a generalização de modelos de aprendizado de máquina. Como alternativa, a geração de dados sintéticos permite ampliar artificialmente conjuntos de dados médicos. Este trabalho apresenta uma análise experimental de métodos de geração de dados sintéticos aplicados à predição de câncer em bases de dados médicas com amostras reduzidas. Para isso, são comparados o Tabular Prior-Fitted Generator (TabPFGGen) e as Redes Generativas Adversariais (GANs). Os dados gerados são avaliados com os modelos Tabular Prior-Data Fitted Network (TabPFN) e Extreme Gradient Boosting (XGBoost). Os resultados demonstram que a utilização de dados sintéticos pode melhorar o desempenho dos classificadores, alcançando ganhos de até 6,6% no F1-score em relação ao cenário base, com melhorias estatisticamente significativas observadas em parte dos experimentos.

Palavras-chave: Inteligência Artificial; Dados Sintéticos; TabPFGGen; GAN; TabPFN.

ABSTRACT

During the application of Artificial Intelligence techniques in healthcare, obtaining large clinical datasets remains a challenge due to data collection costs, limited patient availability, and ethical constraints. This limitation is particularly relevant in studies based on laboratory tests, biochemical markers, and enzymatic assays, potentially affecting the training and generalization of machine learning models. As an alternative, synthetic data generation enables the artificial expansion of medical datasets. This work presents an experimental analysis of synthetic data generation methods applied to cancer prediction in medical datasets with limited sample sizes. For this purpose, the Tabular Prior-Fitted Generator (TabPFGGen) and Generative Adversarial Networks (GANs) are compared. The generated data are evaluated using the Tabular Prior-Data Fitted Network (TabPFN) and Extreme Gradient Boosting (XGBoost) models. The results demonstrate that the use of synthetic data can improve classifier performance, achieving gains of up to 6.6% in F1-score compared to the baseline scenario, with statistically significant improvements observed in part of the experiments.

Keywords: Artificial Intelligence; Synthetic Data; TabPFGGen; GAN; TabPFN.

1 INTRODUÇÃO

O avanço da Inteligência Artificial (IA) tem transformado diversas áreas do conhecimento, especialmente a saúde, na qual modelos computacionais são capazes de identificar padrões complexos em grandes volumes de dados biomédicos Russell and Norvig (2021); Mitchell (1997). Nesse contexto, a aplicação de IA na análise de dados clínicos relacionados ao câncer tem se mostrado uma estratégia promissora para compreender as relações entre variáveis fisiológicas, bioquímicas e demográficas, auxiliando na identificação de padrões associados à presença ou à progressão da doença Kourou et al. (2015).

Apesar do crescente interesse, a escassez de dados clínicos continua sendo um dos principais desafios para a construção de modelos robustos Gonzales et al. (2023). Questões éticas, legais e de privacidade dificultam o compartilhamento de informações médicas, limitando o desenvolvimento de métodos capazes de generalizar adequadamente Gonzales et al. (2023). Dessa forma, torna-se necessário buscar alternativas que permitam ampliar conjuntos de dados reais sem comprometer a confidencialidade dos pacientes.

Em Faria et al. (2025), por exemplo, investiga-se o impacto da sinalização purinérgica no perfil inflamatório de mulheres com lesão intraepitelial escamosa de alto grau (HSIL), condição precursora do câncer do colo do útero. No trabalho, um conjunto de dados é analisado, composto por pacientes com HSIL e controles saudáveis, apresenta um número reduzido de amostras e elevada heterogeneidade biológica, características frequentemente observadas em estudos biomédicos envolvendo marcadores clínicos e moleculares Pezoulas et al. (2024). Nesse contexto, a geração de dados sintéticos surge como uma alternativa promissora Kannan et al. (2025). A geração sintética permite criar amostras artificiais que reproduzem as propriedades estatísticas dos dados originais, preservando as relações entre variáveis e ampliando a capacidade de análise. Estudos recentes indicam que dados sintéticos podem contribuir para reduzir o impacto do desbalanceamento de classes, ampliar conjuntos de treinamento e melhorar o desempenho de classificadores em cenários médicos com baixa disponibilidade de amostras Kannan et al. (2025); Gonzales et al. (2023). Entre as abordagens mais utilizadas para a geração de dados sintéticos, destacam-se as Redes Generativas Adversariais (GANs) Goodfellow et al. (2014); Creswell et al. (2017). Mais recentemente, o *Tabular Prior-Fitted Generator* (TabPFGen) foi proposto como uma alternativa para geração de dados tabulares baseada em modelos fundacionais Ma et al. (2024). No entanto, ainda são escassos os estudos que avaliem seu impacto em tarefas de predição de câncer utilizando bases médicas com amostras reduzidas.

Diante desse cenário, este trabalho investiga o uso de técnicas de geração de dados sintéticos em bases médicas com amostras reduzidas para predição de câncer. Especificamente, são avaliadas as abordagens GAN e TabPFGen quanto à sua capacidade de ampliar conjuntos de treinamento e melhorar o desempenho de modelos de classificação. Para isso, foram utilizadas duas bases de dados reais: a Base HSIL, composta por dados clínicos e marcadores biológicos relacionados ao câncer do colo do útero, e uma base proveniente da plataforma Gene Expression Omnibus (GEO), formada por dados de expressão gênica associados à mesma tarefa de classificação. Os resultados sugerem que o uso de dados sintéticos pode ser uma alternativa eficaz para mitigar os efeitos da escassez de dados em aplicações médicas baseadas em aprendizado de máquina.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta a fundamentação teórica relacionada à Inteligência Artificial, aprendizado de máquina, TabPFN, TabPFGen e GANs. A Seção 3 descreve os trabalhos relacionados. A Seção 4 apresenta a metodologia utilizada nos experimentos. A Seção 5 discute os resultados obtidos. Por fim, a Seção 6 apresenta as conclusões e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

74 Nesta seção, são apresentados os conceitos e modelos que fundamentam este trabalho. Abordam-se os
75 princípios de Inteligência Artificial (IA) e de Aprendizado de Máquina (ML), o *Tabular Prior-Data Fitted*
76 *Network* (TabPFN), o *Tabular Prior-Fitted Generator* (TabPFGen), além das *Generative Adversarial*
77 *Networks* (GANs), amplamente utilizadas na geração de dados sintéticos.

78 A Inteligência Artificial (IA) busca desenvolver sistemas capazes de executar tarefas associadas à
79 inteligência humana, como o reconhecimento de padrões, o aprendizado e a tomada de decisão Russell and
80 Norvig (2021). Dentro desse campo, o Aprendizado de Máquina (*Machine Learning* – ML) concentra-se
81 em métodos que permitem que os sistemas aprendam a partir de dados, melhorando seu desempenho
82 com a experiência Mitchell (1997). Os algoritmos de ML são geralmente classificados em três categorias
83 principais: aprendizado supervisionado, quando o modelo é treinado com dados rotulados; aprendizado
84 não supervisionado, que identifica padrões em dados sem rótulos; e aprendizado por reforço, baseado na
85 interação com o ambiente e nas recompensas associadas às ações.

86 2.1 Conceitos Fundamentais

87 O *In-Context Learning* (ICL) é a capacidade de modelos baseados em *Transformers* de resolver novas
88 tarefas a partir de exemplos fornecidos no próprio contexto, sem ajuste adicional de parâmetros. Nesse
89 processo, o modelo utiliza o mecanismo de atenção para relacionar exemplos e identificar padrões que
90 orientam a predição de novas instâncias Dong et al. (2024). No presente trabalho, esse conceito é relevante
91 porque o TabPFN utiliza essa abordagem ao tratar os dados de treino e de teste como um único contexto.

92 O aprendizado bayesiano baseia-se na atualização de crenças a partir de novas evidências, formalizada
93 pela Regra de Bayes, que combina conhecimento prévio (*prior*) com os dados observados para obter
94 a distribuição posterior dos parâmetros Garcia and Lima Jr. (2021). Diferentemente de métodos que
95 produzem estimativas pontuais, a abordagem bayesiana permite modelar explicitamente a incerteza, sendo
96 particularmente útil em cenários com poucos dados, como em aplicações clínicas.

97 O *Stochastic Gradient Langevin Dynamics* (SGLD) é um método que combina gradiente estocástico
98 com ruído gaussiano para explorar a distribuição posterior dos parâmetros Welling and Teh (2011). Essa
99 abordagem permite aproximar métodos de amostragem bayesiana com menor custo computacional. Em
100 modelos baseados em energia, como o TabPFGen, o SGLD auxilia na geração de dados sintéticos ao
101 explorar, de forma estocástica, o espaço de possíveis amostras.

102 2.2 O Modelo TabPFN

103 O *Tabular Prior-Data Fitted Network* (TabPFN) é um modelo baseado em *Transformers* desenvolvido
104 para dados tabulares. Diferentemente de métodos tradicionais de aprendizado de máquina, que precisam
105 ser treinados para cada conjunto de dados, o TabPFN é pré-treinado com milhões de dados sintéticos.
106 Assim, consegue realizar predições diretamente em um único *forward pass* (ou seja, uma única passagem
107 da entrada pela rede para gerar a saída, sem necessidade de re-treinamento), sem necessidade de ajustes
108 adicionais de parâmetros, sendo particularmente adequado para cenários com poucos dados Hollmann et al.
109 (2024).

110 Seu funcionamento baseia-se no aprendizado em contexto (*in-context learning*), no qual exemplos de
111 treino e teste são processados conjuntamente como um único contexto. Dessa forma, o modelo identifica
112 padrões nos dados e realiza a predição sem etapas adicionais de otimização.

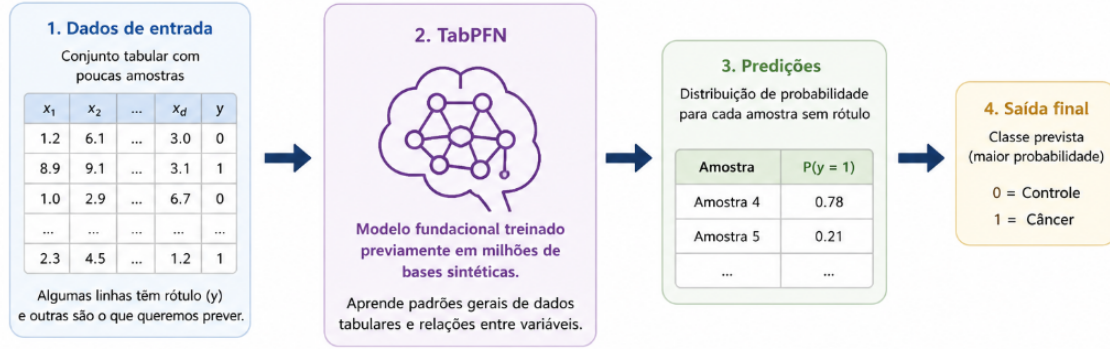


Figura 1. Fluxo simplificado de inferência do TabPFN em dados tabulares.

A Figura 1 apresenta uma visão geral do processo de inferência do TabPFN. Inicialmente, o modelo recebe um conjunto de dados tabulares com amostras rotuladas e instâncias para as quais se deseja realizar a predição. Em seguida, o conhecimento adquirido durante o pré-treinamento em milhões de bases sintéticas é utilizado para estimar a probabilidade de cada amostra pertencer a uma determinada classe. Por fim, a classe associada à maior probabilidade é retornada como resultado da predição. Essa abordagem permite que o modelo obtenha bom desempenho mesmo em bases com amostras reduzidas, um cenário comum em aplicações biomédicas.

2.3 O Modelo TabPFGen

O *Tabular Prior-Fitted Generator* (TabPFGen) é um modelo de geração de dados tabulares proposto por Ma et al. (2024), que utiliza o TabPFN pré-treinado para produzir amostras sintéticas realistas. Diferentemente de GANs e VAEs, o TabPFGen é formulado como um modelo baseado em energia (*Energy-Based Model*), reinterpretando o comportamento discriminativo do TabPFN como uma função de energia, sem exigir um novo treinamento.

O modelo atribui uma energia a uma amostra (x), indicando sua compatibilidade com os dados. Em outras palavras, a energia funciona como uma medida de quão bem uma amostra se ajusta aos padrões aprendidos pelo modelo. Amostras mais compatíveis tendem a apresentar menores valores de energia, enquanto amostras menos compatíveis recebem valores maiores. A energia condicional é definida a partir dos *logits*, conforme a Equação 1:

$$E(x | y) = -f(x)[y] \quad (1)$$

em que menores valores indicam maior plausibilidade.

A geração de dados é realizada por meio do *Stochastic Gradient Langevin Dynamics* (SGLD), que utiliza a função de energia definida na Equação 1 para guiar a geração de novas amostras:

$$x_{t+1} = x_t - \alpha \nabla_x E(x_t | y) + \sigma \cdot \mathcal{N}(0, I) \quad (2)$$

A Equação 2 utiliza a função de energia para ajustar gradualmente as amostras geradas, aproximando-as dos padrões aprendidos pelo modelo. Ao mesmo tempo, um termo de ruído é adicionado para aumentar a diversidade das amostras e evitar que todas sejam muito semelhantes entre si. Dessa forma, é possível gerar

137 dados sintéticos mais realistas e consistentes com os dados originais. Outra característica do método é a
 138 possibilidade de utilizar informações dos dados reais para orientar a geração, contribuindo para a produção
 139 de amostras sintéticas mais realistas Ma et al. (2024).

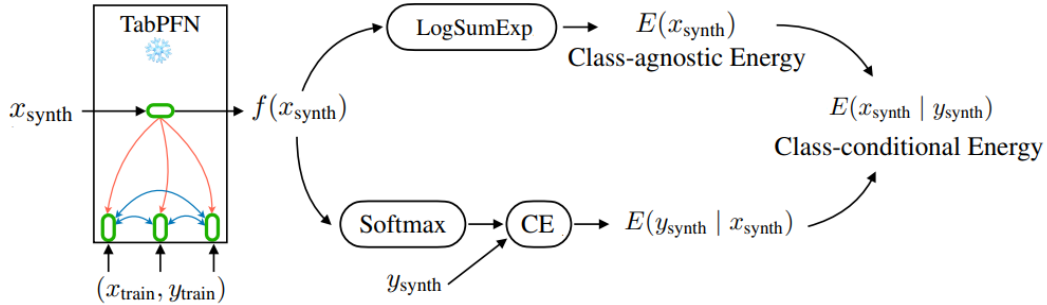


Figura 2. Fluxo do TabPFGen e definição das energias *class-agnostic* e *class-conditional* (adaptado de Ma et al. (2024)).

140 A Figura 2 ilustra o fluxo do modelo. A amostra x_{synth} é processada pelo TabPFN, gerando $f(x)$, a partir
 141 do qual são derivadas energias independentes de classe (*LogSumExp*) e energias condicionais baseadas em
 142 *Softmax* e entropia cruzada.

143 2.4 O Modelo GAN (Generative Adversarial Network)

144 As *Generative Adversarial Networks* (GANs), propostas por Goodfellow et al. (2014), são modelos
 145 generativos que aprendem a distribuição dos dados para produzir amostras sintéticas realistas. Uma GAN
 146 é composta por duas redes neurais: o gerador (G) e o discriminador (D), treinados de forma adversarial
 147 Creswell et al. (2017).

148 O gerador transforma um vetor de ruído z em uma amostra sintética $x' = G(z)$, enquanto o discriminador
 149 recebe amostras reais x e sintéticas x' , estimando a probabilidade de a entrada ser real.

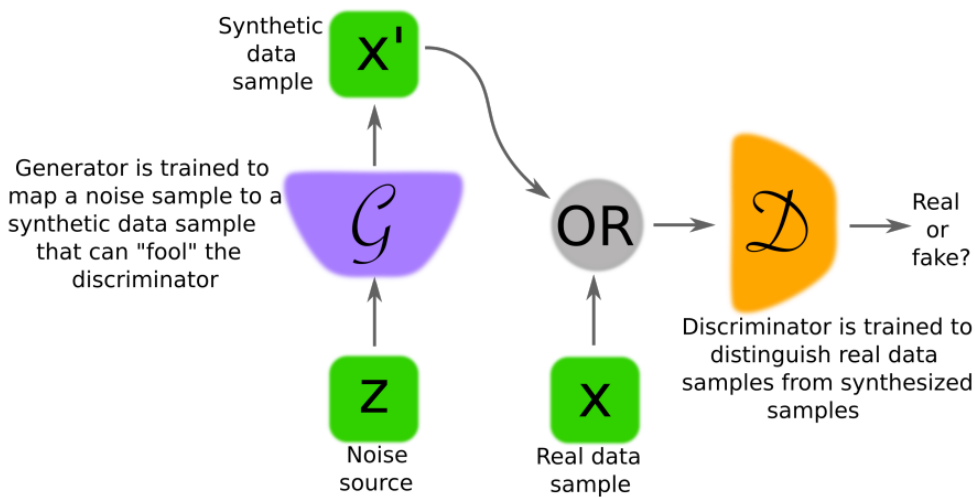


Figura 3. Fluxo de uma GAN, mostrando a interação entre o gerador (G) e o discriminador (D) (adaptado de Creswell et al. (2017)).

150 A Figura 3 ilustra o processo: o gerador recebe ruído z e produz x' , que é avaliado pelo discriminador
151 juntamente com dados reais x . O discriminador decide se a amostra é real ou sintética, enquanto o gerador
152 é otimizado para enganá-lo. O treinamento ocorre de forma alternada entre G e D , em um jogo competitivo
153 que leva o gerador a produzir amostras cada vez mais realistas.

3 TRABALHOS RELACIONADOS

154 Nesta seção, são apresentados trabalhos que utilizam modelos para dados tabulares, geração de dados
155 sintéticos e técnicas de aprendizado em cenários biomédicos. São descritos as abordagens adotadas e os
156 principais resultados obtidos.

157 **3.1 TACO: Um Framework Baseado em Inferência Causal e TabPFN para Detecção** 158 **Precoce da Long COVID**

159 O estudo de Piñero et al. (2025) propõe o *framework* TACO (*TabPFN-Augmented Causal Outcomes*),
160 voltado à detecção precoce da COVID longa ainda na fase presintomática. A abordagem combina inferência
161 causal com o TabPFN para aumentar a interpretabilidade e a eficiência em cenários clínicos.

162 Inicialmente, perfis de expressão gênica de pacientes com COVID aguda e Long COVID são analisados
163 por meio de vias do KEGG e da técnica *Differential Causal Effect* (DCE), que identifica os genes com
164 maior relevância causal para a progressão da doença. Esses genes são então utilizados para reduzir a
165 dimensionalidade dos dados, e o perfil de um novo paciente é fornecido ao TabPFN para realizar a predição
166 do risco de Long COVID.

167 A abordagem identificou 411 genes associados à doença, dos quais 97 já haviam sido descritos na
168 literatura, principalmente relacionados à resposta imune e inflamação. Utilizando esse conjunto reduzido de
169 atributos, o TabPFN foi aplicado sem ajuste fino e obteve desempenho superior a modelos tradicionais em
170 métricas como precisão e ROC-AUC. Em contraste, a presente pesquisa investiga o uso de dados sintéticos
171 para ampliar os conjuntos de treinamento e mitigar os efeitos da escassez de dados.

172 **3.2 Geração Híbrida de Dados Sintéticos para Predição de Doenças: Deep-CTGAN,** 173 **ResNet e TabNet em Bases Médicas Desbalanceadas**

174 O trabalho de Kannan et al. (2025) investiga o uso de dados sintéticos para lidar com o desbalanceamento
175 em bases médicas, comum em cenários com poucos dados e forte assimetria entre classes. Os autores
176 propõem um *framework* que integra pré-processamento, *oversampling*, geração de dados e classificação. O
177 modelo central é o TabNet Arik and Pfister (2021), que utiliza atenção sequencial para selecionar atributos
178 relevantes, oferecendo boa capacidade de modelagem e interpretabilidade. Diferentemente do TabPFN, que
179 é pré-treinado e não requer ajuste, o TabNet depende de treinamento supervisionado, sendo mais indicado
180 quando há maior volume de dados.

181 Inicialmente, são avaliadas técnicas clássicas de geração e balanceamento de dados sintéticos, que
182 apresentam limitações em relações não lineares complexas. Para superar esse problema, o estudo propõe o
183 *Deep-CTGAN*, uma extensão do CTGAN com blocos ResNet, capaz de gerar amostras mais diversas e
184 estáveis. A avaliação é realizada via protocolo *Train on Synthetic, Test on Real* (TSTR), com resultados
185 próximos de 99% de acurácia em bases de COVID-19, doença renal e dengue, indicando boa preservação
186 de padrões clínicos.

187 Em complemento, o uso de SHAP permite interpretar as decisões do modelo, revelando relações coerentes
188 com a literatura médica. Por fim, os autores destacam que o ganho de desempenho depende do equilíbrio

189 entre dados reais e sintéticos, exigindo calibração cuidadosa para evitar distorções e garantir consistência
190 clínica.

191 **3.3 Algoritmo Integrado com Seleção de Atributos, Geração de Dados Sintéticos e** 192 **XGBoost Aplicado ao Câncer de Ovário**

193 Cai et al. (2024) apresenta uma abordagem integrada para a classificação do câncer de ovário, baseada na
194 geração de dados sintéticos e no aprendizado de máquina. O método foi desenvolvido utilizando seis bases
195 de microarranjo, totalizando 502 amostras e mais de 11 mil atributos genéticos. A metodologia combina
196 seleção de atributos por Random Forest, geração de dados sintéticos com AC-GAN e classificação final por
197 meio do modelo XGBoost.

198 Os resultados demonstraram que a combinação entre geração sintética e XGBoost proporcionou
199 melhorias significativas no desempenho do modelo, alcançando acurácia de 99,01%, superando abordagens
200 tradicionais como SVM, regressão logística e o próprio XGBoost aplicado diretamente sobre os dados
201 originais.

202 Segundo os autores, a utilização de técnicas generativas contribuiu para reduzir problemas relacionados
203 ao desbalanceamento e à limitação do conjunto de treinamento, aumentando a capacidade de generalização
204 do classificador.

205 Diferentemente do presente trabalho, que investiga o uso de TabPFGen e GANs em conjunto com os
206 modelos TabPFN e XGBoost, Cai et al. (2024) utiliza uma abordagem baseada em AC-GAN aplicada
207 a dados genéticos de câncer de ovário. Ainda assim, o estudo reforça o potencial da geração de dados
208 sintéticos em cenários biomédicos com pequena disponibilidade de dados, aspecto também explorado nesta
209 pesquisa.

210 **3.4 Estratégia Híbrida com CTGAN para Diagnóstico de Câncer do Colo do Útero**

211 O trabalho de Tang et al. (2025) propõe uma estratégia híbrida para diagnóstico de câncer do colo do
212 útero utilizando seleção de atributos, balanceamento de classes e geração de dados sintéticos. A abordagem
213 combina os métodos mRMR e RFE baseado em XGBoost para seleção de atributos, SMOTE-ENN para
214 balanceamento e CTGAN para geração de dados tabulares sintéticos.

215 Os experimentos foram realizados na base RFCC (*Risk Factors of Cervical Cancer*), contendo 858
216 amostras e 36 atributos clínicos e comportamentais. Foram avaliados modelos como Logistic Regression,
217 KNN, Decision Tree e SVM. Os resultados demonstraram melhorias significativas no desempenho dos
218 classificadores, alcançando até 99,02% de F1-score utilizando Logistic Regression.

219 Os autores também realizaram análises estatísticas para avaliar a qualidade dos dados sintéticos gerados
220 pelo CTGAN, demonstrando que as distribuições sintéticas permaneceram consistentes em relação aos
221 dados reais. Dessa forma, o estudo evidencia o potencial da geração de dados sintéticos para aprimorar
222 modelos de aprendizado de máquina em cenários biomédicos com pequena disponibilidade de dados.

4 METODOLOGIA

223 Neste trabalho, é proposta uma metodologia experimental para avaliar o impacto da geração de dados
224 sintéticos em bases médicas de pequeno porte, investigando sua influência no desempenho de modelos de
225 aprendizado de máquina aplicados à classificação de pacientes.

O fluxo ilustrado na Figura 4 apresenta a metodologia utilizada nesta pesquisa. Inicialmente, os dados passam por uma etapa de pré-processamento, que envolve limpeza, tratamento de valores ausentes, imputação de dados quando necessário e preparação dos atributos para uso nos modelos. Em seguida, aplica-se a técnica de validação cruzada estratificada (*Stratified K-Fold*), que permite dividir os dados em conjuntos de treinamento e teste. Após essa etapa, são avaliados dois cenários: um *baseline*, utilizando apenas dados reais, e outro baseado na expansão do conjunto de treinamento por meio da geração de dados sintéticos utilizando TabPFGen e Redes Generativas Adversariais (GANs).

Os conjuntos obtidos são utilizados no treinamento dos modelos TabPFN e XGBoost, sendo posteriormente avaliados por métricas de classificação. Por fim, é realizada uma comparação estatística utilizando o teste t (*t-test*) para analisar possíveis diferenças entre os cenários avaliados.

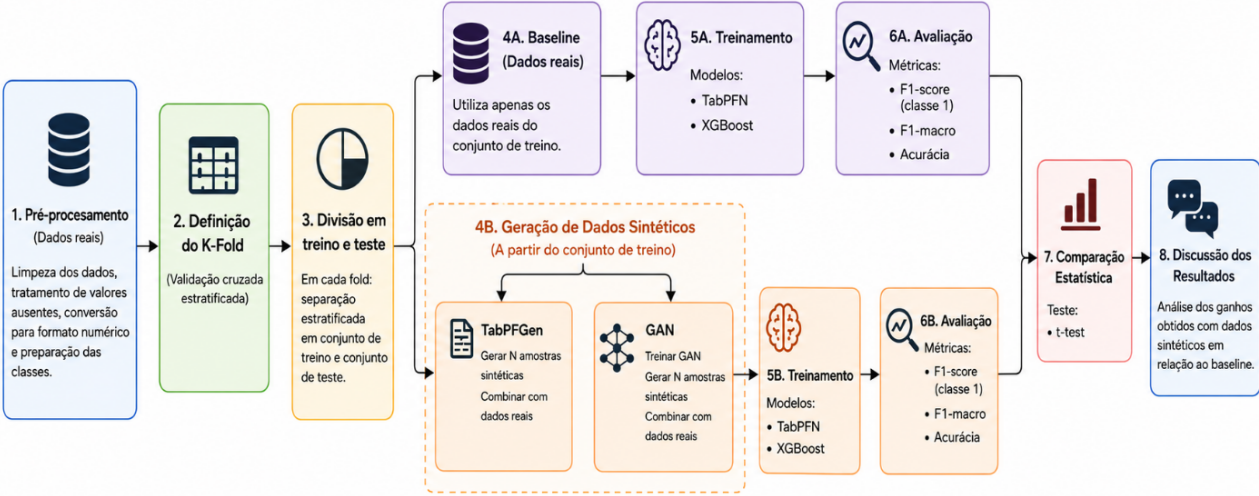


Figura 4. Fluxo metodológico utilizado no desenvolvimento deste trabalho.

5 EXPERIMENTOS

5.1 Base de dados

As bases de dados utilizadas neste estudo provêm de duas fontes. A primeira corresponde ao conjunto de dados utilizado por Faria et al. (2025), denominado neste trabalho Base HSIL, composto por dados clínicos e marcadores biológicos relacionados ao câncer do colo do útero. A segunda base foi obtida a partir da plataforma Gene Expression Omnibus (GEO), contendo dados de expressão gênica associados à mesma tarefa de classificação.

Para padronizar os experimentos, ambas as bases foram organizadas em uma estrutura binária de classificação, na qual os rótulos foram definidos da seguinte forma:

- 1 (positivo): pacientes diagnosticados com câncer;
- 0 (negativo): grupo controle, sem diagnóstico da doença.

A Tabela 1 apresenta um resumo das características das bases de dados utilizadas nos experimentos.

Observa-se que ambas as bases apresentam um número semelhante de amostras, porém diferem quanto à origem e ao tipo de informação disponível. Enquanto a base HSIL contém marcadores bioquímicos,

Tabela 1. Características das bases de dados utilizadas no estudo.

Base	Amostras	Atributos	Positivos	Negativos
HSIL	82	16	42	40
GEO	106	10	68	38

249 imunológicos e moleculares, a base GEO é composta por dados de expressão gênica, o que permite avaliar
250 o desempenho dos métodos em diferentes contextos biológicos.

251 5.1.1 Distribuição dos atributos

252 As Figuras 5 e 6 apresentam a projeção bidimensional das amostras das bases HSIL e GEO,
253 respectivamente, obtida por meio da Análise de Componentes Principais (*Principal Component Analysis*
254 – PCA) Jolliffe and Cadima (2016). Essa técnica permite representar, em um espaço de apenas duas
255 dimensões, a variação conjunta de todos os atributos originais, possibilitando observar visualmente a
256 relação entre as amostras das classes controle e câncer.

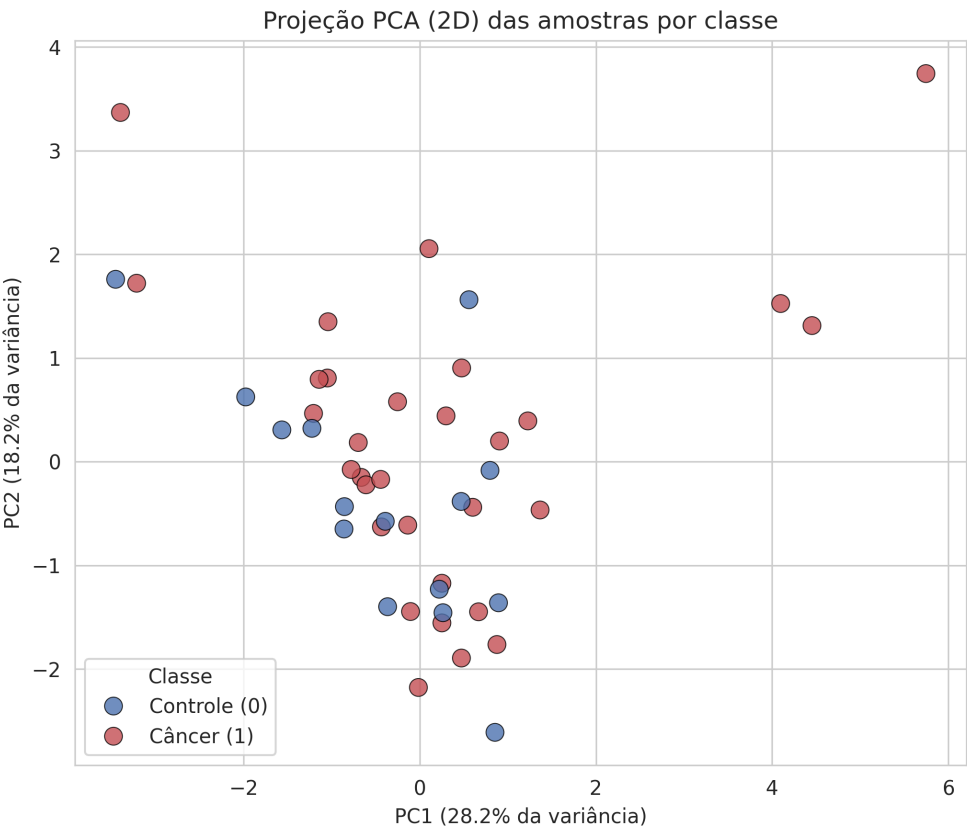


Figura 5. Projecção bidimensional (PCA) das amostras da base HSIL, coloridas por classe.

257 Na base HSIL, as duas primeiras componentes principais (PC1 e PC2) explicam, em conjunto,
258 aproximadamente 46,4% da variância total dos dados (28,2% e 18,2%, respectivamente). Observa-se
259 sobreposição considerável entre as amostras das classes controle e câncer na região central da projecção,

260 embora algumas amostras da classe câncer situem-se em regiões mais distantes do centro, indicando maior
261 heterogeneidade nesse grupo.

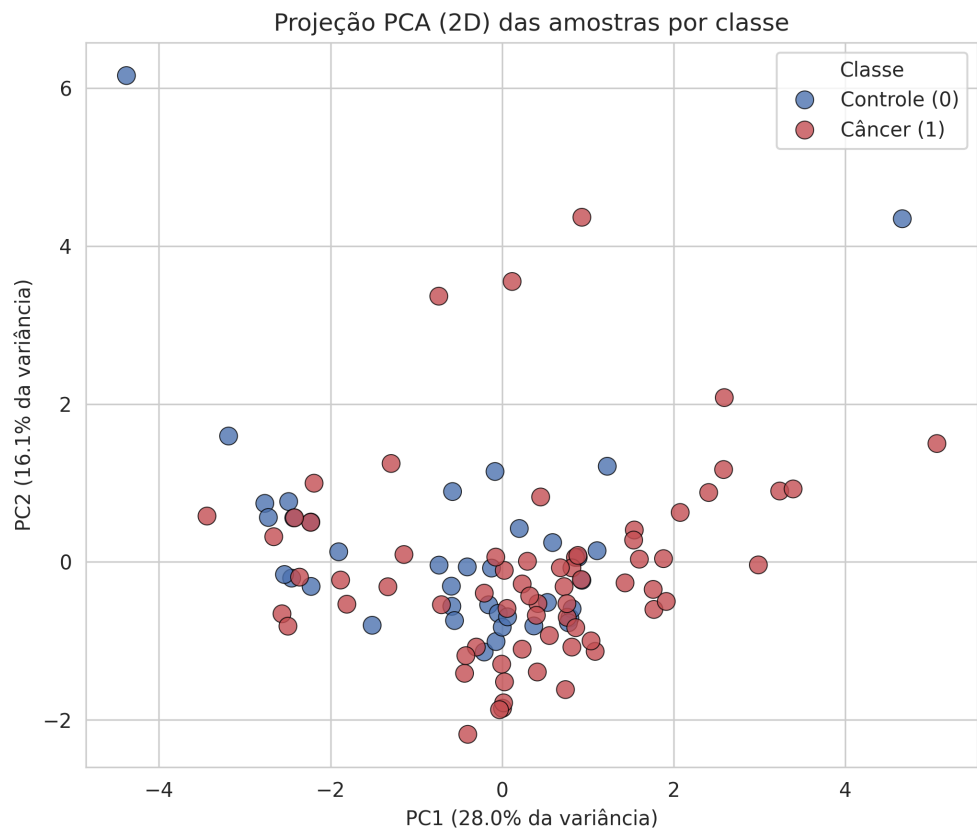


Figura 6. Projecção bidimensional (PCA) das amostras da base GEO, coloridas por classe.

262 Na base GEO, PC1 e PC2 explicam 28,0% e 16,1% da variância total, respectivamente (44,1%
263 combinados), padrão semelhante ao observado na base HSIL. Também há predominância de sobreposição
264 entre as classes, com alguns pontos da classe câncer mais dispersos em relação ao restante das amostras.

265 Em ambas as bases, a ausência de uma separação linear nítida entre as classes no espaço bidimensional
266 sugere que a distinção entre os grupos controle e câncer não pode ser obtida por meio de uma simples
267 combinação linear das variáveis originais. Esse resultado reforça a relevância de modelos capazes de
268 capturar relações não lineares entre os atributos, como o TabPFN e o XGBoost, empregados neste trabalho.

269 5.1.2 Importância das variáveis

270 As Figuras 7 e 8 apresentam a importância global das variáveis para as bases HSIL e GEO,
271 respectivamente, calculada por meio da média dos valores absolutos de SHAP Lundberg and Lee (2017)

272 Na base HSIL, observa-se que a variável *Atividade ADA em linfócitos* apresentou a maior contribuição
273 para as decisões do modelo, seguida por IL-10 e IFN- γ . As demais variáveis apresentaram impacto
274 relativamente reduzido, indicando que a classificação é influenciada principalmente por um pequeno
275 conjunto de marcadores.

276

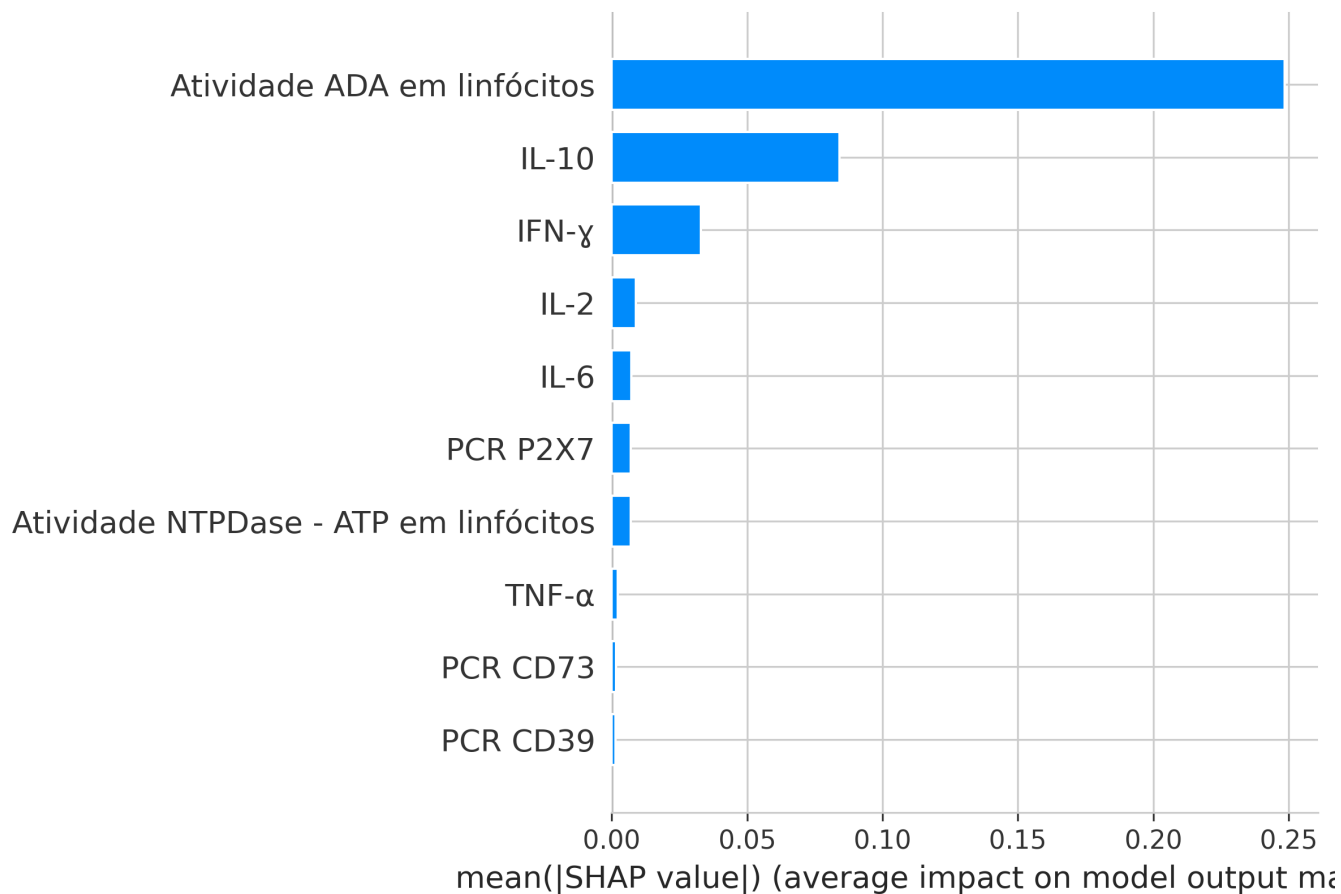


Figura 7. Importância das variáveis da base HSIL obtida por meio dos valores SHAP.

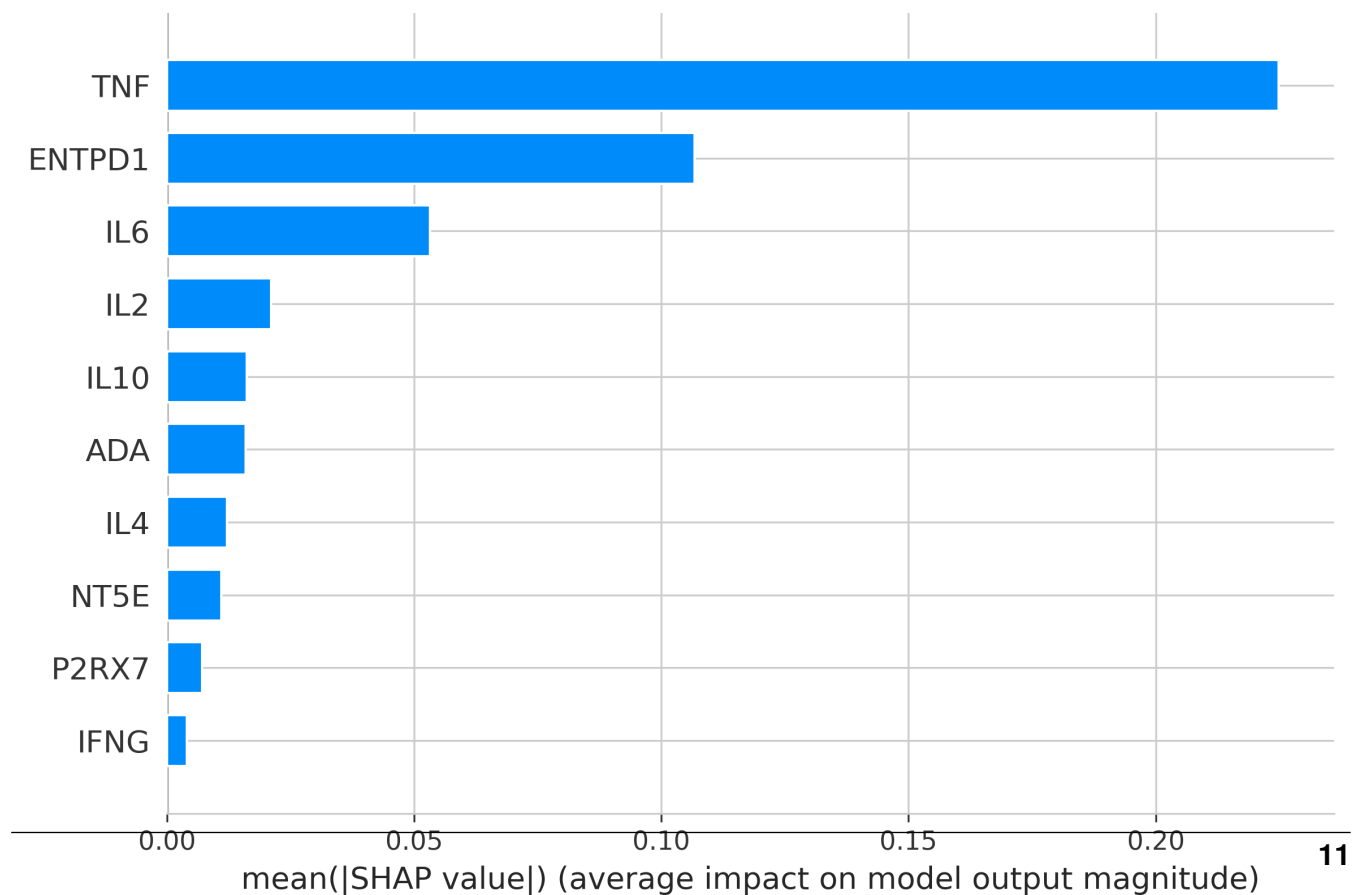


Figura 8. Importância das variáveis da base GEO obtida por meio dos valores SHAP.

277 Na base GEO, a variável TNF apresentou a maior importância para o modelo, seguida por ENTPD1 e
278 IL6. As demais variáveis contribuíram de forma significativamente menor para as predições, indicando que
279 o processo de classificação depende principalmente desses três atributos. A comparação entre as duas bases
280 mostra que, embora ambas apresentem um conjunto reduzido de variáveis mais relevantes, os atributos
281 que concentram maior importância diferem entre os conjuntos de dados, refletindo suas características
282 biológicas e experimentais.

283 **5.2 Pré-processamento**

284 Antes da execução dos experimentos, as bases de dados passaram por uma etapa de pré-processamento
285 para padronizar os atributos e garantir a consistência das informações utilizadas pelos modelos. Inicialmente,
286 foram removidos valores ausentes, campos vazios e registros contendo símbolos inválidos, como “-”. Em
287 seguida, os atributos foram convertidos para formato numérico, permitindo sua utilização pelos algoritmos
288 de aprendizado de máquina.

289 Para a base HSIL, devido à presença de valores ausentes e visando minimizar a perda de informação,
290 foi aplicada a técnica de imputação *Multiple Imputation by Chained Equations* (MICE) van Buuren and
291 Groothuis-Oudshoorn (2011). Essa abordagem realiza a estimativa iterativa dos valores faltantes com base
292 nas relações existentes entre os atributos da base de dados.

293 **5.3 Configuração experimental**

294 Foi utilizada validação cruzada estratificada (*Stratified K-Fold*) com cinco *folds*, preservando a proporção
295 entre as classes em cada divisão dos dados.

296 Para os métodos de geração de dados sintéticos, foram avaliadas diferentes configurações para o número
297 de amostras geradas, o número de etapas do SGLD (*n_sgl_steps*) e o número de épocas (*epochs*). Os
298 valores considerados em cada experimento são apresentados na Tabela 2.

Tabela 2. Valores avaliados para os hiperparâmetros dos geradores sintéticos

Parâmetro	Valores avaliados
Número de amostras sintéticas	0, 50, 100, 200, 500, 1000
<i>epochs</i> e <i>n_sgl_steps</i>	200, 500, 1000, 2000, 5000, 10000, 15000

299 Para o modelo XGBoost, foram avaliadas três configurações distintas. A primeira utilizou uma
300 configuração padrão, servindo como referência para comparação. A segunda empregou técnicas adicionais
301 de regularização, incluindo redução da taxa de aprendizado, amostragem parcial dos dados e penalização
302 dos nós de decisão, buscando minimizar possíveis efeitos de sobreajuste (*overfitting*). A terceira adotou uma
303 abordagem mais conservadora, com aprendizado mais lento, maior número de estimadores e penalizações
304 mais fortes, visando aumentar a capacidade de generalização do modelo. Os parâmetros utilizados em cada
305 configuração são apresentados na Tabela 3.

Tabela 3. Configurações avaliadas para o modelo XGBoost

Parâmetro	Padrão	Regularizado	Conservador
<i>n_estimators</i>	100	300	500
<i>max_depth</i>	3	4	3
<i>learning_rate</i>	0,10	0,05	0,03
<i>subsample</i>	1,0	0,8	0,8
<i>colsample_bytree</i>	1,0	0,8	0,8
<i>gamma</i>	0	1	1
<i>reg_lambda</i>	1	1	2
<i>reg_alpha</i>	0	0,5	1

5.4 Resultados

Os resultados obtidos são apresentados separadamente para cada base de dados avaliada. Para cada cenário, são analisados os efeitos da geração de dados sintéticos sobre o desempenho dos classificadores TabPFN e XGBoost, considerando diferentes configurações dos métodos de geração. Além da comparação com os respectivos cenários *baseline*, foi realizada uma análise estatística para avaliar a significância dos ganhos observados.

5.4.1 Base HSIL

A Figura 9 apresenta o mapa de calor dos resultados obtidos para a base HSIL utilizando o método TabPFGem em conjunto com o modelo TabPFN. Os resultados representam diferentes combinações entre a quantidade de amostras sintéticas geradas e o número de etapas do SGLD utilizadas no processo de geração.

Observa-se que a utilização de dados sintéticos apresentou melhorias em relação ao cenário *baseline*. Os melhores resultados foram obtidos com 50 amostras sintéticas associadas a 2000 e 5000 etapas de SGLD, atingindo um F1-score de 0,890 para a classe positiva. Em comparação ao *baseline* (F1-score de 0,838), esse resultado representa um ganho relativo de aproximadamente 6,1%. Entretanto, o teste t não indicou diferença estatisticamente significativa entre os resultados.

Também é possível constatar que o aumento contínuo do número de amostras sintéticas não resultou necessariamente em ganhos adicionais de desempenho. Em determinadas configurações, maiores quantidades de dados gerados resultaram em reduções nas métricas avaliadas, indicando a possível introdução de ruído ou alterações na distribuição original dos dados.

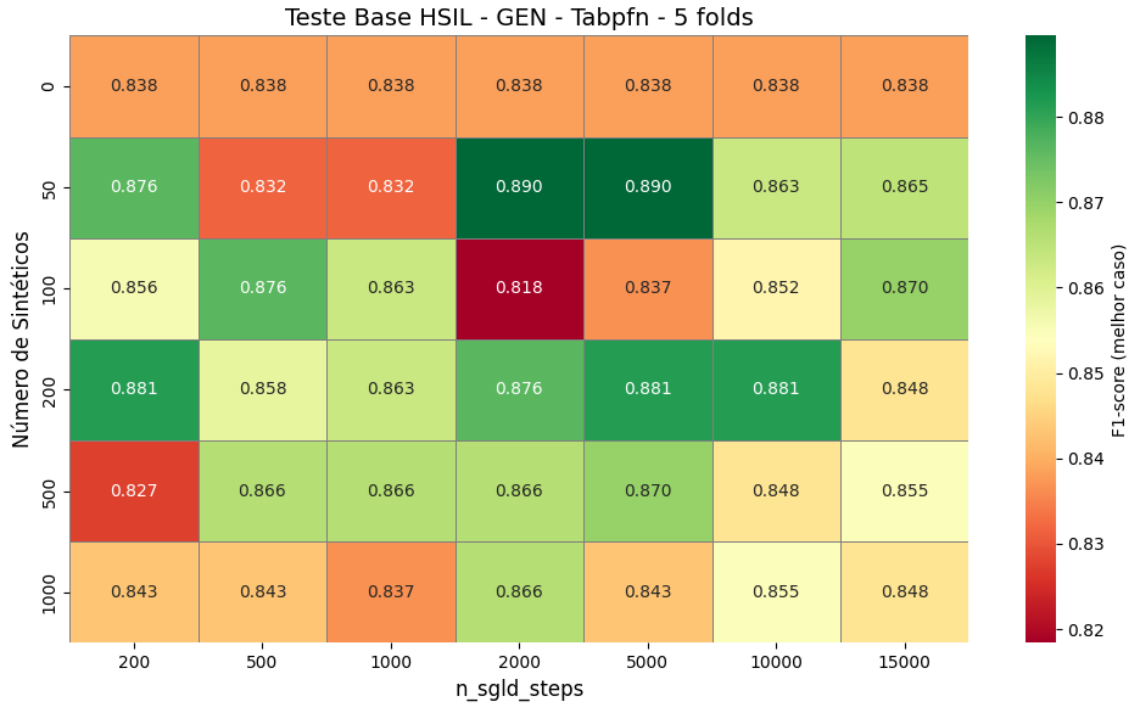


Figura 9. Mapa de calor do F1-score da classe positiva (classe 1) obtido na validação cruzada com 5 folds para diferentes quantidades de dados sintéticos e valores de n_sgld_steps , utilizando TabPFGen e TabPFN.

Tabela 4. Melhores resultados obtidos para a base HSIL

Métrica	Método	TabPFN	XGBoost
F1 Classe 1	Baseline	0,838	0,858 (XGB padrão) / 0,803 (XGB conservador)
	GEN	0,890 (+0,051) 50 sintéticos + 2000 SGLD	0,894 (+0,036) XGB (padrão) + 1000 sintéticos + 200 SGLD
	GAN	0,856 (+0,018) 100 sintéticos + 200 épocas	0,890 (+0,087) XGB (conservador) + 100 sintéticos + 500 épocas
F1 Macro	Baseline	0,755	0,625 (XGB conservador)
	GEN	0,828 (+0,073) 50 sintéticos + 2000 SGLD	0,806 (+0,181) XGB (conservador) + 100 sintéticos + 2000 SGLD
	GAN	0,775 (+0,020) 100 sintéticos + 200 épocas	0,829 (+0,204) XGB (conservador) + 100 sintéticos + 500 épocas

Na Tabela 4, são apresentados os melhores resultados obtidos para cada combinação entre o modelo de classificação e o método de geração de dados sintéticos. Para cada caso, são mostrados o F1-score(classe 1) *baseline*, o melhor F1-score(classe 1) obtido com dados sintéticos e o ganho correspondente. No XGBoost, os valores de *baseline* diferem entre GEN e GAN porque os melhores resultados foram obtidos com

330 configurações distintas do modelo. Embora tenham sido observadas melhorias em todas as combinações
331 avaliadas, nenhuma delas apresentou diferença estatisticamente significativa em relação ao respectivo
332 *baseline*.

333 5.4.2 Base GEO

334 Na Figura 10 são apresentados os resultados obtidos para a base GEO utilizando o método TabPFGem em
335 conjunto com o modelo TabPFN.

336 Na base GEO, a melhor configuração utilizou 200 amostras sintéticas e 200 etapas de SGLD, alcançando
337 um F1-score médio de 0,904 para a classe positiva. Em relação ao *baseline* (0,848), esse resultado
338 representa um ganho de aproximadamente 6,6%. No entanto, os testes estatísticos não indicaram diferença
339 significativa nesse cenário. Resultados semelhantes também foram observados em outras configurações,
340 indicando que ganhos de desempenho podem ser obtidos mesmo com diferentes quantidades de dados
341 sintéticos gerados.

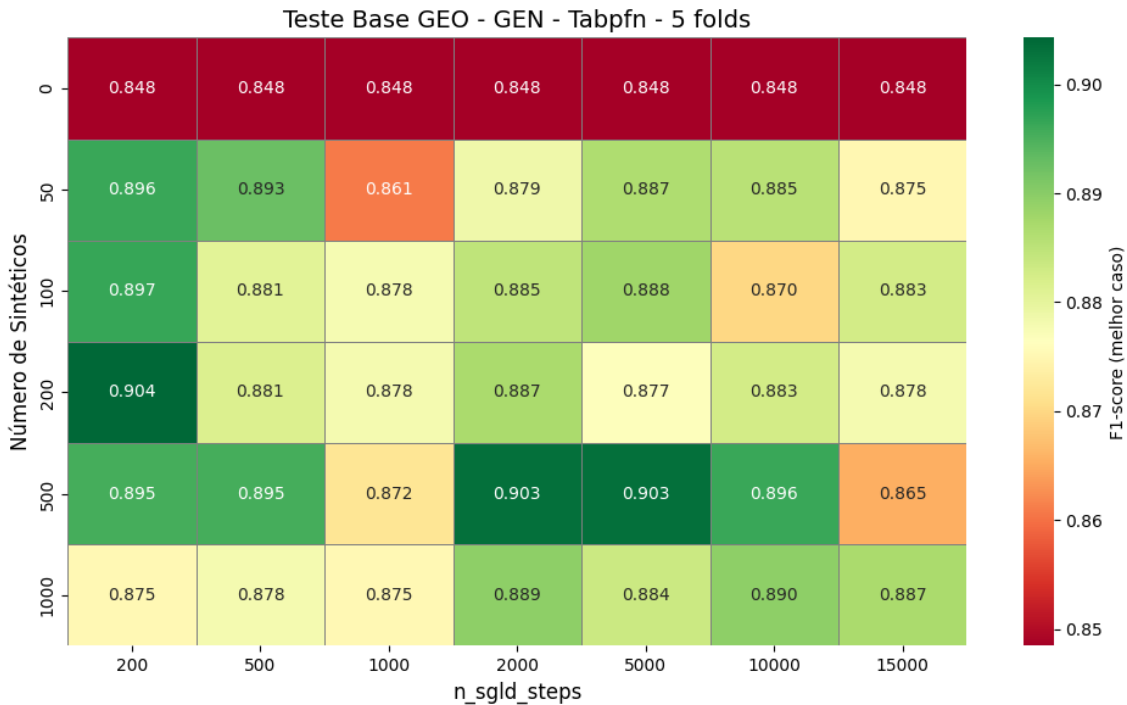


Figura 10. Mapa de calor dos resultados de F1-score da classe positiva (classe 1) para diferentes configurações de geração sintética com TabPFGem e TabPFN.

342 A Tabela 5 apresenta os melhores resultados obtidos com os métodos GEN e GAN nos classificadores
343 avaliados. São apresentados os valores de F1-score(classe 1) obtidos no cenário *baseline*, os melhores
344 resultados com a inclusão de dados sintéticos e os respectivos ganhos observados. No caso do XGBoost,
345 os diferentes valores de *baseline* correspondem às configurações do modelo associadas aos melhores
346 resultados, conforme já discutido.

347 Também foi realizada uma análise estatística para verificar se as melhorias obtidas em relação ao
348 cenário *baseline* representavam diferenças significativas ou apenas variações ocasionais entre os *folds*.
349 Diferentemente do que ocorreu na base HSIL, a base GEO apresentou resultados estatisticamente
350 significativos em algumas das configurações avaliadas. Conforme apresentado na Tabela 6, os casos

Tabela 5. Melhores resultados de F1-score da classe positiva (classe 1) obtidos para a base GEO

Método	TabPFN	XGBoost
Baseline	0,848	0,868 (XGB conservador) / 0,876 (XGB regularizado)
GEN	0,904 (+0,056) 200 sintéticos + 200 SGLD	0,886 (+0,019) XGB (conservador) + 500 sintéticos + 1000 SGLD
GAN	0,865 (+0,016) 100 sintéticos + 1000 épocas	0,890 (+0,014) XGB (regularizado) + 50 sintéticos + 200 épocas

significativos foram identificados por meio do modelo TabPFN em conjunto com o método GEN. Para o F1-score da classe positiva, as configurações com 50 amostras sintéticas e 200 e 500 etapas de SGLD apresentaram valores de p inferiores ao nível de significância adotado. Resultados semelhantes também foram observados para a métrica F1-score macro, indicando que os ganhos não ocorreram apenas na classe de interesse, mas também se refletiram no desempenho geral do modelo. Uma possível explicação para esse comportamento é que a base GEO apresenta padrões mais consistentes e mais facilmente reproduzíveis pelos geradores sintéticos. Em contrapartida, a base HSIL possui um número menor de amostras e maior variabilidade dos atributos clínicos, o que pode dificultar a geração de dados sintéticos capazes de produzir melhorias estatisticamente significativas. Dessa forma, os resultados sugerem que a efetividade da geração de dados sintéticos está relacionada às características da base de dados utilizada.

Tabela 6. Resultados estatisticamente significativos obtidos para a base GEO

Métrica	TabPFN (GEN)
F1 classe 1	Baseline: 0,848 → 0,896 ($p = 0,041$) 50 sintéticos + 200 SGLD
	Baseline: 0,848 → 0,893 ($p = 0,041$) 50 sintéticos + 500 SGLD
F1 macro	Baseline: 0,796 → 0,857 ($p = 0,026$) 50 sintéticos + 200 SGLD
	Baseline: 0,796 → 0,858 ($p = 0,027$) 50 sintéticos + 500 SGLD
	Baseline: 0,796 → 0,848 ($p = 0,027$) 50 sintéticos + 5000 SGLD

6 CONCLUSÃO

361 Este trabalho apresentou uma análise experimental do uso de técnicas de geração de dados sintéticos
362 aplicadas a bases médicas de pequeno porte para tarefas de classificação de pacientes. Foram avaliados
363 os métodos TabPFGen e Redes Generativas Adversariais (GANs), em conjunto com os modelos TabPFN
364 e XGBoost, com o objetivo de investigar o impacto da expansão artificial de dados sobre o desempenho
365 preditivo.

366 Os resultados demonstraram que a utilização de dados sintéticos pode contribuir para a melhoria do
367 desempenho dos classificadores em diferentes cenários. Os melhores resultados foram obtidos com o
368 método GEN, alcançando F1-score de 0,894 na base HSIL e 0,904 na base GEO. Embora esses valores
369 representem os maiores desempenhos observados, os ganhos obtidos nem sempre corresponderam aos
370 cenários estatisticamente significativos.

371 Os testes estatísticos mostraram que parte das melhorias observadas pode estar associada à variabilidade
372 dos dados, não representando necessariamente diferenças relevantes em relação aos cenários baseline. Ainda
373 assim, algumas configurações avaliadas na base GEO apresentaram ganhos estatisticamente significativos,
374 sugerindo que a geração de dados sintéticos pode contribuir de forma consistente para o desempenho dos
375 modelos quando parâmetros adequados são empregados durante o processo de geração.

376 Como trabalhos futuros, pretende-se investigar outros métodos de geração de dados sintéticos, explorar
377 diferentes estratégias de processamento dos dados e avaliar os modelos em bases médicas maiores,
378 permitindo ampliar a análise da capacidade de generalização das abordagens estudadas.

REFERÊNCIAS

- 379 Arik, S. O. and Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In *Proceedings of the*
380 *AAAI Conference on Artificial Intelligence*. 6679–6687
- 381 Cai, J., Lee, Z.-J., Lin, Z., Hsu, C.-H., and Lin, Y. (2024). An integrated algorithm with feature selection,
382 data augmentation, and xgboost for ovarian cancer. *Mathematics* 12, 4041. doi:10.3390/math12244041
- 383 Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2017).
384 Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*
- 385 Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., et al. (2024). A survey on in-context learning. *arXiv*
386 *preprint arXiv:2301.00234*
- 387 Faria, P. C. L., Palú, G., Franciosi, M. L. M., Lima, M. D. M., Maciel, S. F. V. O., and Cardoso, A. M.
388 (2025). O papel do atp e do p2x7 no perfil inflamatório de mulheres com lesão intraepitelial de alto grau
389 do colo de útero Relatório técnico / manuscrito interno fornecido pela autora
- 390 Garcia, G. D. and Lima Jr., R. M. (2021). Introdução à estatística bayesiana aplicada à linguística. *Revista*
391 *da ABRALIN* 20, 1–24. doi:10.25189/rabralin.v20i2.1914
- 392 Gonzales, K. D., Ellis, M. T., Prince, J. A., et al. (2023). Synthetic data in health care: A narrative review.
393 *BMC Medical Research Methodology* 23, 32–47. doi:10.1186/s12874-023-01931-4
- 394 Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative
395 adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2672–2680
- 396 Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., et al. (2024). Accurate
397 predictions on small data with a tabular foundation model. *Nature* doi:10.1038/s41586-024-08328-6
- 398 Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments.
399 *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*
400 374, 20150202. doi:10.1098/rsta.2015.0202

401 Kannan, M., Umamaheswari, D., Manimekala, B., Stella Mary, I. P., Savitha, P. M., and Rozario, J. (2025).
402 An enhancement of machine learning model performance in disease prediction with synthetic data
403 generation. *Scientific Reports* 15, 33482. doi:10.1038/s41598-025-15019-3

404 Kourou, K. et al. (2015). Machine learning applications in cancer prognosis and prediction. *Computational*
405 *and Structural Biotechnology Journal*

406 Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances*
407 *in Neural Information Processing Systems (NeurIPS)*. 4765–4774

408 Ma, J., Dankar, A., Stein, G., Yu, G., and Caterini, A. (2024). Tabpfgn: Tabular data generation with
409 tabpfn. *arXiv preprint arXiv:2406.05216* Table Representation Learning Workshop at NeurIPS 2023

410 Mitchell, T. M. (1997). *Machine Learning* (New York, NY, USA: McGraw-Hill)

411 Pezoulas, V. C. et al. (2024). Synthetic data generation methods in healthcare: A review on open-
412 source tools and methods. *Computational and Structural Biotechnology Journal* 23, 2892–2910.
413 doi:10.1016/j.csbj.2024.07.005

414 Piñero, S., Li, X., Liu, L., Li, J., Lee, S. H., Winter, M., et al. (2025). Taco: Tabpfn augmented causal
415 outcomes for early detection of long covid. *medRxiv* doi:10.1101/2025.10.02.25337138. Preprint, not
416 peer-reviewed

417 Russell, S. J. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (London, UK: Pearson
418 Education), 4 edn.

419 Tang, M., Chen, H., Lv, Z., and Cai, G. (2025). Diagnosis of cervical cancer based on a hybrid strategy
420 with ctgan. *Electronics* 14, 1140. doi:10.3390/electronics14061140

421 van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations
422 in r. *Journal of Statistical Software* 45, 1–67. doi:10.18637/jss.v045.i03

423 Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In
424 *Proceedings of the 28th International Conference on Machine Learning (ICML)*. 681–688