Universidade do Rio de Janeiro
Pós-Graduação em Modelagem Computacional
Aprendizagem de Máquina
Trabalho 2

Nome do Autor: Ítalo Rosa Gonçalves

Data: 18 de Novembro de 2025

Sumário

1 Descrição dos Dados

A base de dados utilizada neste trabalho é a ”Heart Failure Clinical Records”, que contém informações clínicas de 299 pacientes com insuficiência cardíaca. O conjunto de dados possui 13 atributos, sendo 12 características preditoras e uma variável alvo binária, `DEATH_EVENT`, que indica se o paciente faleceu (1) ou sobreviveu (0) durante o período de acompanhamento.

Uma análise inicial revelou que a base de dados é comportada, não contendo valores nulos ou amostras duplicadas. A variável alvo é desbalanceada, com 203 amostras da classe 0 e 96 da classe 1.

A Figura 1 apresenta os boxplots de todas as variáveis, permitindo uma análise da distribuição e a identificação de outliers. Nota-se a presença de outliers em atributos como `creatinine_phosphokinase` e `serum_creatinine`, o que reforça a importância da padronização dos dados para algoritmos sensíveis à escala.

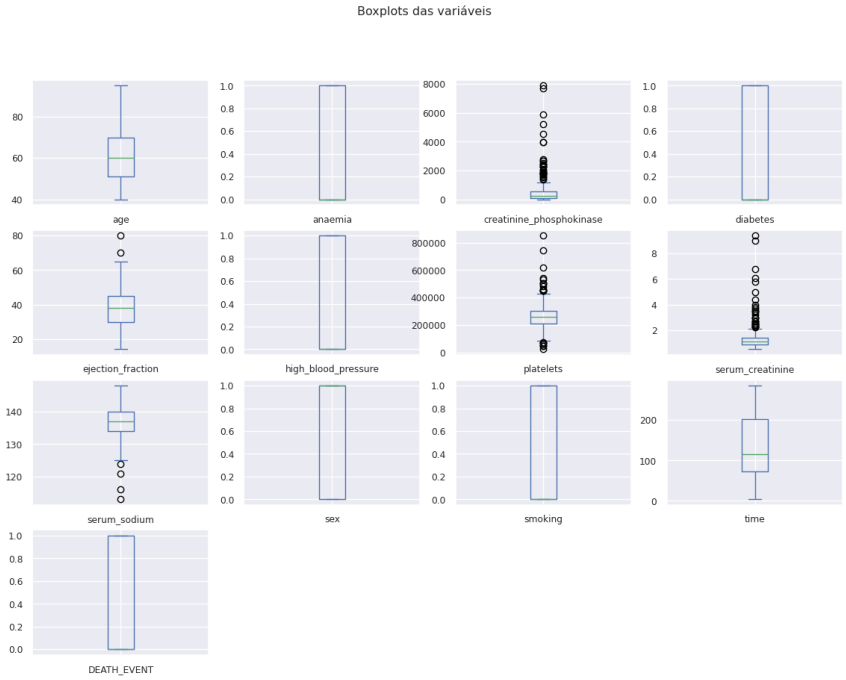

Figura 1: Boxplot das variáveis numéricas da base de dados.

Tabela 1: Resumo das estatísticas descritivas (Parte 1 de 2).

	age	anaemia	creatinine _phosphokinase	diabetes	ejection _fraction	high_blood _pressure
count	299.00	299.00	299.00	299.00	299.00	299.00
mean	60.83	0.43	581.84	0.42	38.08	0.35
std	11.89	0.50	970.29	0.49	11.83	0.48
min	40.00	0.00	23.00	0.00	14.00	0.00
25%	51.00	0.00	116.50	0.00	30.00	0.00
50%	60.00	0.00	250.00	0.00	38.00	0.00
75%	70.00	1.00	582.00	1.00	45.00	1.00
max	95.00	1.00	7861.00	1.00	80.00	1.00

Tabela 2: Resumo das estatísticas descritivas (Parte 2 de 2).

	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
count	299.00	299.00	299.00	299.00	299.00	299.00	299.00
mean	263358.03	1.39	136.63	0.65	0.32	130.26	0.32
std	97804.24	1.03	4.41	0.48	0.47	77.61	0.47
min	25100.00	0.50	113.00	0.00	0.00	4.00	0.00
25%	212500.00	0.90	134.00	0.00	0.00	73.00	0.00
50%	262000.00	1.10	137.00	1.00	0.00	115.00	0.00
75%	303500.00	1.40	140.00	1.00	1.00	203.00	1.00
max	850000.00	9.40	148.00	1.00	1.00	285.00	1.00

As Tabelas 1 e 2 apresentam um resumo estatístico detalhado de todos os atributos da base de dados. Geradas a partir do método describe() da biblioteca Pandas, elas consolidam informações essenciais como média, desvio padrão e quartis, fornecendo uma visão geral da distribuição e escala de cada variável.

Uma inspeção preliminar do conjunto de dados revelou que este não possui valores faltantes ou duplicados, e todas as suas variáveis já se encontram em formato numérico, o que simplifica as etapas de pré-processamento. A análise dos boxplots, apresentada anteriormente na Figura 1, complementa esta descrição ao evidenciar a presença de outliers em atributos como creatinine_phosphokinase e serum_creatinine.

2 Análise dos Dados

As principais medidas estatísticas dos atributos, como média, desvio padrão e quartis, estão resumidas na Tabela 1 e 2. Esta tabela fornece uma visão geral da escala e distribuição de cada variável. Assim, pode-se partir para a matriz de correlação que irá nos dizer como uma característica influencia na outra. Para ficar mais fácil a visualização, foi gerado a partir da matriz de correlação um heatmap.

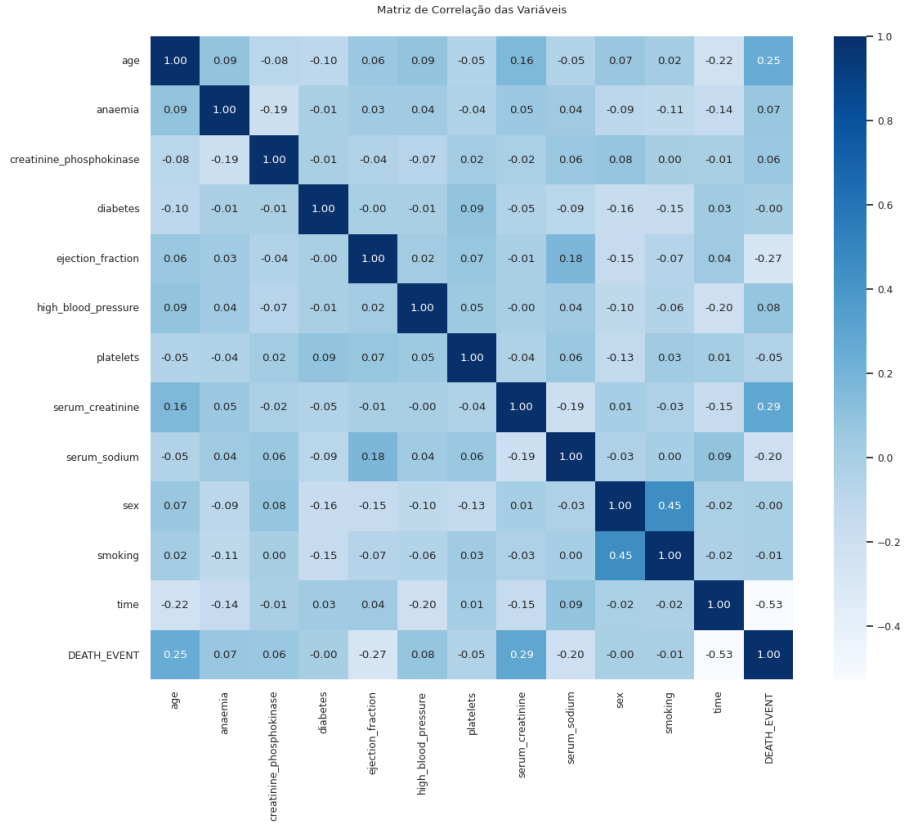

Figura 2: Matriz de correlação entre as variáveis.

A Figura 2 apresenta a matriz de correlação entre todas as variáveis do conjunto de dados. A diagonal principal da matriz possui valor 1, pois representa a correlação de cada característica consigo mesma.

Observa-se que, de modo geral, as correlações lineares entre os atributos são fracas. As correlações mais expressivas com a variável alvo, `DEATH_EVENT`, são as do tempo de acompanhamento (`time`), da creatinina sérica (`serum_creatinine`) e da fração de ejeção (`ejection_fraction`). Ainda assim, por serem correlações de magnitude moderada a fraca, a análise do heatmap sugere que o desfecho do paciente não é determinado por uma única variável isoladamente, mas sim por uma combinação mais complexa de fatores.

2.1 Etapa de pré-processamento

O pré-processamento dos dados foi fundamental para garantir que os algoritmos de agrupamento e validação pudessem funcionar de forma eficaz.Assim após a fase de analise dos dados e verificar a ausência de valores nulos ou categóricos, aplicamos a padronização utilizando o StandardScaler do scikit-learn.Esta etapa é fundamental, pois na base de dados há atributos como o `creatinine_phosphokinase` com valores com grande diferença entre seu mínimo e máximo contidos na tabela 1. A padronização transforma os dados de modo que cada atributo tenha média zero e desvio padrão unitário, permitindo que todos os atributos contribuam de forma equilibrada para o cálculo das distâncias entre as amostras.

3 Metodologia

A metodologia adotada para a analise do trabalho seguiu algumas etapas, primeiramente a verificação de valores faltantes, dados duplicados e a presença de dados no formatado caractere, seguindo pela fase de balanceamento e redução de dimensionalidade sempre que for necessário a visualização, e usando o`GridSearchCV` com o intuito de descobrir os melhores parametros para o método de classificação `Multi-Layer Perceptron (MLPClassifier)` .

3.1 Tratamento de Desbalanceamento (ADASYN)

Para endereçar o desbalanceamento de classes (67.9% vs 32.1%), foi aplicada a técnica de superamostragem `ADASYN` (Adaptive Synthetic Sampling). Esta técnica foi aplicada **apenas** ao conjunto de treino, como ditam as boas práticas, para evitar vazamento de dados (*data leakage*) para o conjunto de teste. A Tabela 3 mostra a distribuição das classes no conjunto de treino antes e depois da aplicação do ADASYN.

Tabela 3: Distribuição de classes no conjunto de treino antes e depois do ADASYN.

Conjunto	**Classe 0 (Sobreviveu)**	**Classe 1 (Óbito)**
Treino (Original)	203	96
Treino (Pós-ADASYN)	203	194

3.2 Modelo de Classificação e Validação

O modelo de classificação utilizado foi o `Multi-Layer Perceptron (MLPClassifier)`, conforme designado. Para a validação e otimização de hiperparâmetros, foi empregado o `GridSearchCV` com uma estratégia de validação cruzada K-Fold ($k = 5$) conforme o padrão, podendo-se obter melhores resultados com maiores K's, ao custo de maior poder computacional . Para garantir robustez estatística, o processo foi repetido 30 vezes (`N_RUNS = 30`), e a média de acurácia foi usada para selecionar o melhor modelo. A grelha de parâmetros explorada incluiu:

- `hidden_layer_sizes`: [(5),(5,5),(5,5,5),(5,5,5,5),(10),(10,10),(10,10,10),(50)]
- `activation`: ['identity', 'logistic', 'tanh', 'relu']
- `max_iter`: [1000] (fixado, por conta do tempo de execução)

4 Experimentos Computacionais e Resultados

Foram conduzidos dois experimentos principais para avaliar o impacto do balanceamento de dados no desempenho do MLP.

4.1 Configuração Experimental e Reprodutibilidade

As análises foram conduzidas com o objetivo de avaliar a robustez do classificador frente ao desbalanceamento severo presente na base de insuficiência cardíaca. Para garantir reprodutibilidade, toda a experimentação seguiu uma configuração fixa de particionamento, normalização e validação cruzada.

A base foi dividida em conjuntos de treino e teste usando a função `train_test_split` do `scikit-learn`, na proporção de 80% para treino e 20% para teste, com `random_state = 42`.

As variáveis numéricas foram padronizadas por meio de `StandardScaler`, ajustado apenas no conjunto de treino, evitando vazamento de informação para o teste.

O balanceamento das classes foi realizado por ADASYN, aplicado **exclusivamente** ao conjunto de treino, após a divisão dos dados, conforme boas práticas. A busca pelos melhores hiperparâmetros foi feita com `GridSearchCV` utilizando validação cruzada em 5 folds, também restrita ao conjunto de treino.

Cada experimento foi repetido N = 30 vezes, com seeds distintos, permitindo estimar variabilidade estatística das métricas. Para cada repetição registrou-se: acurácia, precisão, recall, F1-score, AUC-ROC, AUC-PR e matriz de confusão. Os valores apresentados ao longo dos resultados correspondem à média e ao desvio-padrão dessas 30 execuções.

A Tabela 4 resume os principais parâmetros utilizados no processo experimental.

Tabela 4: Configuração experimental e parâmetros de reprodutibilidade.

Item	Valor
Conjunto de treino/teste	80% / 20%
Semente (random_state)	(ex.: 42)
Normalização	StandardScaler na base antes de ambos os grid
Oversampling	ADASYN (aplicado **somente** ao conjunto de treino)
Validação	GridSearchCV com K-fold = 5 (aplicada somente no treino)
Número de repetições	N_RUNS = 30 (rodadas independentes com seeds diferentes)

4.2 Métricas

o desempenho de um modelo não deve ser avaliado apenas por uma única execução, pois tanto a inicialização aleatória dos pesos quanto a divisão dos dados em treino e teste introduzem variabilidade nos resultados. Para obter uma avaliação mais estável e confiável, o procedimento experimental foi repetido 30 vezes, cada uma com uma semente aleatória distinta. Dessa forma, cada métrica (acurácia, precisão, sensibilidade, F1-score, AUC-ROC e AUC-PR) pôde ser analisada não apenas pelo seu valor médio, mas também por seu desvio-padrão.

O uso da média fornece uma estimativa central do desempenho típico do modelo, enquanto o desvio-padrão indica o quanto esse desempenho oscila entre diferentes execuções. Valores baixos de desvio-padrão sugerem que o modelo é estável e pouco sensível às variações na inicialização e no particionamento dos dados; em contrapartida, desvios elevados indicam que o comportamento do modelo é mais inconsistente e depende fortemente das condições de execução. Essa análise é particularmente importante em cenários com bases pequenas e desbalanceadas, como é o caso deste estudo, onde pequenas mudanças no conjunto de treino podem alterar de forma significativa o comportamento da rede neural.

Ao apresentar as métricas no formato média $\pm$ desvio-padrão, o relatório busca fornecer uma visão mais completa da robustez do modelo, permitindo comparar não apenas o nível de desempenho, mas também sua estabilidade. Esse procedimento segue práticas consolidadas em experimentos de Aprendizado de Máquina, garantindo maior rigor estatístico e transparência na interpretação dos resultados.

Tabela 5: Métricas de Desempenho Global (Média e Desvio Padrão em 30 Execuções)

Métrica	Média	Desvio Padrão
Acurácia	0.8196	0.0369
F1-Score (Weighted)	0.8191	0.0372
Recall (Weighted)	0.8196	0.0369

Nota: Resultados obtidos via validação cruzada com 30 repetições.

4.3 Experimento 1: Base de Dados Desbalanceada

No primeiro experimento, o `GridSearchCV` foi aplicado diretamente ao conjunto de treino original (desbalanceado). A Tabela 6 apresenta os hiperparâmetros ótimos encontrados.

Tabela 6: Experimento 1: Parâmetros Ótimos do MLP (Dados Desbalanceados).
Esta tabela foi gerada a partir da saída da célula 1202 do notebook

Hiperparâmetros	Valor
`activation`	'identity'
`hidden_layer_sizes`	[5, 5]
`max_iter`	1000
Acurácia Média (Validação Cruzada)	**0.8196**

Este modelo foi então avaliado no conjunto de teste (que também é desbalanceado). O resultado final de acurácia é apresentado na Tabela 6.

4.4 Experimento 2: Base de Dados Balanceada (ADASYN)

No segundo experimento, o `GridSearchCV` foi aplicado ao conjunto de treino balanceado pelo ADASYN. A Tabela 7 apresenta os hiperparâmetros ótimos encontrados.

Tabela 7: Experimento 2: Parâmetros Ótimos do MLP (Dados Balanceados com ADASYN).

Hiperparâmetros	Valor
`activation`	'identity'
`hidden_layer_sizes`	[50]
`max_iter`	1000
Acurácia Média (Validação Cruzada)	**0.8196**

Curiosamente, a acurácia média na validação cruzada foi idêntica à do Experimento 1, embora a arquitetura de rede ótima encontrada tenha sido diferente (uma única camada oculta com 50 neurónios, contra duas camadas de 5). Este modelo foi avaliado no mesmo conjunto de teste do Experimento 1.

4.5 Discussão e Comparação dos Resultados

A Tabela 8 e 9 apresenta uma comparação detalhada do desempenho dos dois modelos no conjunto de teste, utilizando métricas mais robustas (Recall, Precision e F1-Score) além da Acurácia. O foco é o desempenho na **Classe 1 (`DEATH_EVENT`)**, que é a classe de interesse primário em um diagnóstico médico.

Tabela 8: Comparação das Métricas de Desempenho no Conjunto de Teste (Parte 1 de 2).

Abordagem Experimental	**Acurácia Geral**	**Recall (Sensibilidade)**
Exp. 1 (Dados Desbalanceados)	0.7833	0.5789
Exp. 2 (Dados Balanceados c/ ADASYN)	**0.8000**	**0.7895**

Tabela 9: Comparação das Métricas de Desempenho no Conjunto de Teste (Parte 2 de 2).

Abordagem Experimental	**Precision (Precisão)**	**F1-Score**
Exp. 1 (Dados Desbalanceados)	**0.6875**	0.6300
Exp. 2 (Dados Balanceados c/ ADASYN)	0.6522	**0.7130**

A análise dos resultados demonstra um ganho de desempenho ao utilizar o tratamento de desbalanceamento.

4.5.1 Ganho na Capacidade Preditiva

O modelo treinado com dados balanceados pelo ADASYN (Exp. 2) alcançou uma acurácia de 80.00% no conjunto de teste, superando os 78.33% do modelo treinado nos dados originais (Exp. 1). O ganho mais crucial, no entanto, reside nas métricas específicas para a classe minoritária:

- **Recall (Sensibilidade):** Houve um aumento de 57.89% para **78.95**% no Recall para a classe 1 (Óbito). Em um contexto médico, o Recall é a capacidade de identificar corretamente pacientes que falecerão, minimizando Falsos negativos. O ganho de mais de 21 pontos percentuais é de grande importância clínica.
- **F1-Score:** O F1-Score, que é a média harmônica entre Precision e Recall, subiu de 0.6300 para **0.7130**, confirmando a melhoria geral na qualidade das predições da classe minoritária.

4.5.2 Análise das Matrizes de Confusão

A melhoria no Recall e a pequena queda na Precision no Exp. 2 podem ser visualizadas nas matrizes de confusão (Figuras 3 e 4). O ADASYN permitiu que o modelo cometesse menos erros do Tipo II (Falsos Negativos - prever sobrevida quando o paciente faleceu), em troca de um ligeiro aumento nos erros do Tipo I (Falsos Positivos - prever Óbito quando o paciente sobreviveu).

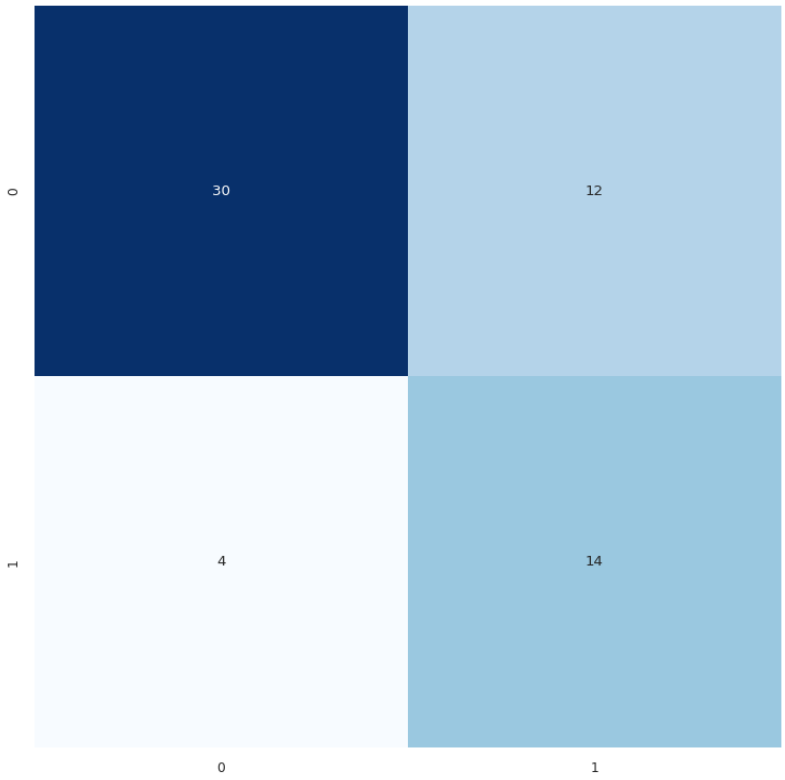

Figura 3: Matriz de Confusão do Experimento 1

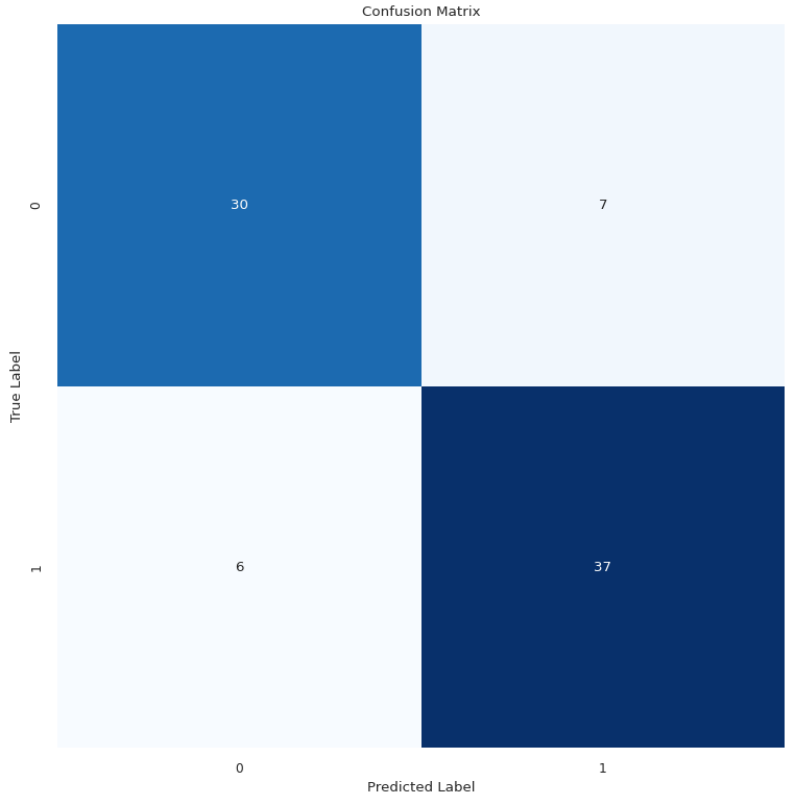

Figura 4: Matriz de Confusão do Experimento 2

Diferença de Arquitetura

A diferença mais curiosa foi a mudança na arquitetura ótima encontrada:

- **Exp. 1 (Desbalanceado):** (5, 5) - Duas camadas de 5 neurônios.
- **Exp. 2 (Balanceado):** (50) - Uma única camada com 50 neurônios.

5 Conclusão

Este estudo investigou a eficácia de Redes Neurais Artificiais como ferramenta de suporte ao prognóstico de mortalidade em pacientes com insuficiência cardíaca, com foco central na mitigação do desbalanceamento de classes. A análise comparativa entre o modelo treinado com dados originais e aquele submetido ao balanceamento sintético via ADASYN revelou implicações assistenciais significativas.

A aplicação do ADASYN promoveu um aumento substancial na sensibilidade (Recall) do classificador, mitigando o risco de falsos negativos — falha crítica em triagens médicas que poderia levar à omissão de pacientes graves. Embora esse ganho tenha ocorrido em detrimento da precisão, configurando o clássico trade-off de classificação, tal comportamento é clinicamente defensável em cenários de triagem, onde a prioridade é maximizar a detecção de casos de alto risco.

Sob a ótica da modelagem, observou-se que a alteração na distribuição dos dados influenciou a topologia da rede: o uso do ADASYN demandou uma arquitetura com maior capacidade (camada oculta mais larga) para acomodar a complexidade adicional da fronteira de decisão gerada pelas

amostras sintéticas. Isso evidencia que técnicas de pré-processamento alteram não apenas as métricas de desempenho, mas também os requisitos estruturais do modelo.

As conclusões, contudo, devem ser interpretadas considerando o tamanho restrito da amostra (n=299). A robustez dos achados requer validação em coortes independentes e a confirmação da significância estatística das diferenças observadas por meio de testes de hipótese (como Wilcoxon pareado). Como trabalhos futuros, sugere-se: a avaliação de métodos de ensemble (Random Forest, XGBoost) e estratégias de ponderação (class weights), a aplicação de técnicas de explicabilidade (XAI), como SHAP, para identificar biomarcadores críticos, e a realização de estudos prospectivos para mensurar o impacto operacional da ferramenta na prática clínica..

Referências

[1] O QUE é a análise de componentes principais (PCA)?. *IBM*, [s.d.]. Disponível em: https://www.ibm.com/br-pt/think/topics/principal-component-analysis. Acesso em: 10 out. 2025.

[2] SCIKIT-LEARN. *scikit-learn: machine learning in Python*. [s.l.]: scikit-learn, [s.d.]. Disponível em: https://scikit-learn.org/stable/. Acesso em: 10 out. 2025.

[3] AMARAL, Fernando. *Inteligência Artificial e Machine Learning*. Udemy, [s.d.]. Disponível em: https://www.udemy.com/course/inteligencia-artificial-e-machine-learning/?couponCode=MT251015G4. Acesso em: 10 out. 2025.