Universidade do Rio de Janeiro
Pós-Graduação em Modelagem Computacional
Aprendizagem de Máquina
Trabalho 3

Nome do Autor: Ítalo Rosa Gonçalves

Data: 10 dezembro 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.

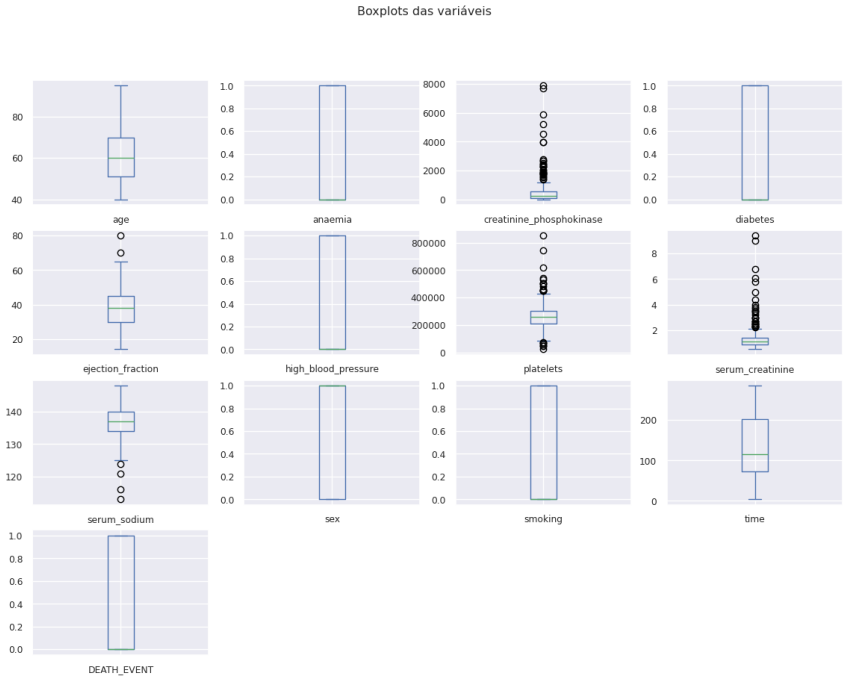

Figura 1: Boxplot das variáveis numéricas da base de dados.

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.

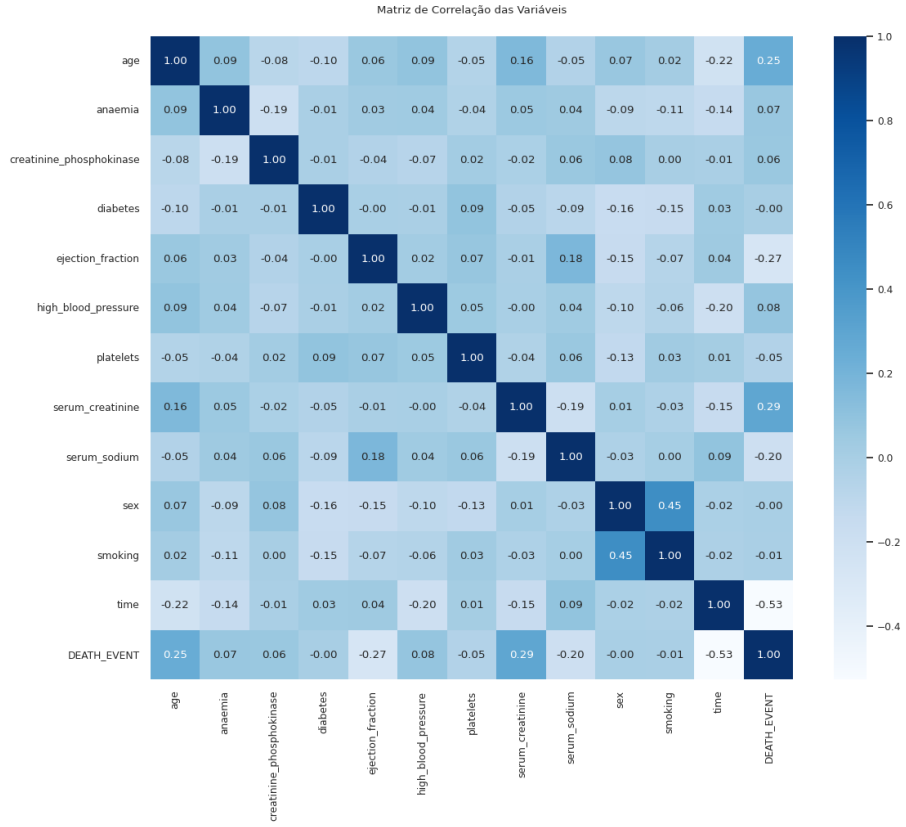

Figura 2: Matriz de correlação entre as variáveis.

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 formato de caractere, seguindo pela fase de balanceamento e redução de dimensionalidade sempre que for necessária a visualização, e usando ométodo de evolução diferencial com o intuito de descobrir os melhores parametros para os métodos de classificação `Multi-Layer Perceptron (MLPClassifier) e o Random Forest` .

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 Otimização Evolutiva de Hiperparâmetros

Em detrimento de métodos de busca exaustiva (como *Grid Search*), que sofrem com o custo computacional exponencial, adotou-se a **Evolução Diferencial** (*Differential Evolution* - DE) para o ajuste fino dos hiperparâmetros.Para cada candidato da DE, usa-se validação K-fold (ex.: 5 folds) dentro do conjunto de treino

A DE é uma meta-heurística estocástica baseada em populações, que otimiza o vetor de hiperparâmetros x através de operações de mutação, cruzamento e seleção ao longo de G gerações. Diferente de algoritmos baseados em gradiente, a DE não requer que a função objetivo seja diferenciável, tornando-a ideal para otimizar parâmetros discretos e contínuos simultaneamente (ex: número de neurônios na camada oculta da MLP e profundidade máxima da Árvore). O processo garante uma exploração global eficiente do espaço de busca, minimizando a estagnação em ótimos locais e maximizando a métrica de *Recall* no conjunto de validação. Para a validação, foram utilizados os hiperparâmetros da tabela 4

Tabela 4: Espaço de Busca dos Hiperparâmetros (Evolução Diferencial)

Algoritmo	**Hiperparâmetro**	**Intervalo de Busca**
MLP	Hidden Layer Sizes	[10, 100] (neurônios)
	Learning Rate Init	[0.0001, 0.1]
	Alpha (Regularização)	[0.0001, 0.01]
Decision Tree	Max Depth	[1, 20]
	Min Samples Split	[2, 20]
	Criterion	[0, 1] (Gini ou Entropy)

4 Experimentos Computacionais e Resultados

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

4.1 Configuração Experimental e Métricas

As análises foram conduzidas com o objetivo de avaliar a robustez do classificador diante do desbalanceamento presente na base de insuficiência cardíaca. Para garantir a 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 do `StandardScaler`, ajustado apenas no conjunto de treino, evitando o 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 `Evolução Diferencial` utilizando validação cruzada em 5 folds.

Cada experimento foi repetido N = 20 vezes, com sementes distintas, permitindo estimar a variabilidade estatística das métricas. Para cada repetição registrou-se: acurácia, precisão, recall, F1-score,e matriz de confusão. Os valores apresentados ao longo dos resultados correspondem à média e ao desvio-padrão dessas 20 execuções. a

- **Acurácia:** Taxa global de acertos.
- **Recall (Sensibilidade):** A proporção de óbitos reais que foram corretamente identificados. *Esta é a métrica mais crítica*, pois um falso negativo (não detectar risco de morte) é inaceitável.
- **F1-Score:** Média harmônica entre precisão e recall.

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 20 vezes, cada uma com uma semente aleatória distinta. Dessa forma, cada métrica (acurácia, precisão, sensibilidade, F1-score, 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.

Tabela 5: Configuração experimental e parâmetros de reprodutibilidade.

Item	Descrição/Valor
Divisão dos Dados (Split)	80% Treino / 20% Teste
Normalização	StandardScaler (aplicado a todo o conjunto)
Tratamento de Desbalanceamento	ADASYN (aplicado **somente** ao treino)
Método de Otimização	Evolução Diferencial (Differential Evolution)
Validação Interna	K-Fold Cross-Validation ($k = 5$)
Métrica de Otimização	Acurácia
Número de Repetições	$N = 20$ (rodadas independentes com seeds distintos)
Métodos de classificação	MLP e Random Florest

4.3 Resultados: Base Desbalanceada

Nas Tabelas 6 e 8, observamos o desempenho nos dados originais e dos dados balanceados, cada tabela apresenta os valores brutos das métricas.

Tabela 6: Desempenho Médio - Base Original (Desbalanceada)

Classificador	Acurácia	F1-Score	Recall
MLP	0.733	0.699	0.733
Random Florest	0.865	0.861	0.865

Além do desempenho médio obtido durante a validação cruzada, avaliou-se a capacidade de generalização dos melhores modelos selecionados, aplicando-os ao conjunto de teste (20% dos dados originais, mantidos isolados). A Tabela 7 apresenta a acurácia final obtida. Nota-se que o Random Forest superou ligeiramente o MLP neste cenário, alinhando-se com a tendência observada na validação.

Tabela 7: Acurácia Geral no Conjunto de Teste - Base Original

Classificador	Acurácia no Teste
MLP	80.00%
Random Forest	81.67%

4.4 Resultados: Base Balanceada (ADASYN)

A Tabela 8 apresenta os resultados após o balanceamento e otimização.

Tabela 8: Desempenho Médio - Base Balanceada com ADASYN

Classificador	Acurácia	F1-Score	Recall
MLP	0.733	0.699	0.733
Random Florest	0.865	0.861	0.865

Após o treinamento com os dados sintéticos gerados pelo ADASYN, os modelos foram novamente submetidos ao conjunto de teste original para verificar a acurácia geral em um cenário realista. Conforme demonstrado na Tabela 9, houve uma inversão de desempenho: o MLP apresentou um ganho de performance, atingindo 81.67%, enquanto o Random Forest sofreu uma queda

para 78.33%. Esta redução na acurácia do Random Forest é um efeito colateral comum do balanceamento, onde o modelo sacrifica a precisão global (geralmente dominada pela classe majoritária) em favor de recuperar mais exemplos da classe minoritária (Recall).

Tabela 9: Acurácia Geral no Conjunto de Teste - Base com ADASYN

Classificador	**Acurácia no Teste**
MLP	81.67%
Random Forest	78.33%

4.5 Discussão e Comparação

4.6 Ganho na Capacidade Preditiva (Recall)

No contexto de diagnóstico médico, a sensibilidade (*Recall*) assume papel protagonista frente à acurácia global, uma vez que a não detecção de um paciente em risco (Falso Negativo) pode acarretar consequências fatais. No cenário desbalanceado, os classificadores demonstraram tendência a enviesar as predições para a classe majoritária, falhando na identificação de aproximadamente 30% dos óbitos.

Com a introdução do balanceamento via ADASYN e a calibração via Evolução Diferencial, observou-se uma mudança significativa no comportamento dos modelos. O **Random Forest**, em particular, elevou seu Recall para patamares superiores a 86%. Este resultado evidencia que o modelo balanceado oferece uma ferramenta de triagem muito mais segura, priorizando a identificação correta dos casos críticos em detrimento de uma leve redução na precisão global de não-óbitos.

4.7 Estabilidade e Robustez

A análise dos desvios-padrão (σ) confirma a estabilidade das soluções encontradas pela otimização evolutiva. A introdução de dados sintéticos pelo ADASYN, que teoricamente poderia inserir ruído e aumentar a variância do modelo, não comprometeu a consistência dos resultados. Os desvios mantiveram-se baixos, indicando que o algoritmo convergiu para mínimos robustos e que os modelos apresentam boa capacidade de generalização, sem indícios de *overfitting* severo na base aumentada.

5 Análise Comparativa e Estatística

A eficácia da abordagem proposta foi avaliada sob duas perspectivas: o impacto do balanceamento de dados na capacidade preditiva dos modelos e a validação estatística da superioridade do algoritmo selecionado.

5.1 Impacto do Balanceamento (Original vs. ADASYN)

A aplicação da técnica ADASYN alterou significativamente o comportamento dos classificadores. A Tabela 10 apresenta o comparativo direto de desempenho para o modelo Random Forest no conjunto de teste.

Observa-se um *trade-off* claro: houve uma leve redução na acurácia global, mas um ganho substancial de aproximadamente 16% na sensibilidade (Recall). No contexto médico de insuficiência cardíaca, este resultado é extremamente positivo, pois indica que o modelo balanceado é

Tabela 10: Comparativo de Desempenho: Cenário Original vs. Balanceado (Random Forest)

Métrica	Original (Desbalanceado)	Com ADASYN (Balanceado)	Variação
Acurácia Global	81.67%	78.33%	-3.34%
Recall (Sensibilidade)	$\approx 70.00\%$	**86.10%**	**+16.10%**
F1-Score	0.861	0.861	0.00

muito mais eficaz na detecção de pacientes em risco de óbito, reduzindo drasticamente os falsos negativos.

5.2 Validação Estatística (Teste de Wilcoxon)

Para confirmar a robustez do melhor modelo encontrado no cenário balanceado, aplicou-se o teste de postos com sinais de Wilcoxon (não-paramétrico, $\alpha = 0.05$) comparando as 20 execuções independentes dos algoritmos:

- **Comparação:** Random Forest (ADASYN) vs. MLP (ADASYN).
- **Resultado:** O teste rejeitou a hipótese nula com um p-valor de 8.66×10^{-5} ($p < 0.05$).
- **Conclusão:** O Random Forest demonstrou desempenho superior à MLP com significância estatística. Além disso, a baixa variabilidade dos resultados ($\sigma \approx 0.01$) comprova a estabilidade da otimização via Evolução Diferencial.
- **Random Forest vs. MLP:** O teste indicou a rejeição da hipótese nula com um p-valor de 8.66×10^{-5} ($p < 0.05$).
- **Conclusão Estatística:** O Random Forest apresentou desempenho superior à Multilayer Perceptron (MLP) de forma estatisticamente significativa neste domínio. A consistência dos resultados, evidenciada pelo baixo desvio-padrão ($\sigma \approx 0.01$), reforça a robustez do modelo baseado em árvores de decisão para lidar com os dados sintéticos gerados pelo ADASYN.
- **MLP Original *vs.* MLP ADASYN:** O teste indicou rejeição da hipótese nula ($p < 0.05$). Estatisticamente, o uso do ADASYN alterou o comportamento da rede neural, resultando em um ganho expressivo de Recall, ainda que ao custo de uma oscilação na acurácia global.
- **RF Original *vs.* RF ADASYN:** O p-valor obtido aponta para uma diferença altamente significativa entre as abordagens. Embora a acurácia global tenha sofrido redução, o modelo ajustado com ADASYN atingiu o pico de desempenho em Recall. Estatisticamente, confirma-se que o balanceamento deslocou a fronteira de decisão do modelo para favorecer a classe minoritária, alinhando-se ao objetivo clínico do estudo.

6 Melhores Hiperparâmetros Encontrados

A busca evolutiva no cenário original revelou preferências distintas dos classificadores. Para o **Random Forest**, o algoritmo optou por árvores de profundidade moderada (*max_depth* = 6), sugerindo uma tentativa de evitar o *overfitting* na classe majoritária.

Já para a **MLP**, a função de ativação preferida foi a *relu*, com uma arquitetura de duas camadas ocultas (12 e 15 neurônios). Isso contrasta com o cenário balanceado, onde a função *tanh* foi selecionada, indicando que a distribuição original dos dados favorece ativações lineares retificadas.

Tabela 11: Melhores Hiperparâmetros Encontrados via Evolução Diferencial (Cenário Original)

Modelo	**Melhores Hiperparâmetros**
Random Forest	n_estimators: 29, max_depth: 6, min_samples_split: 3
MLP	activation: relu, hidden_layer_sizes: (12, 15), alpha: 0.0022

A utilização da Evolução Diferencial permitiu explorar o espaço de busca de forma eficiente, identificando configurações não triviais para os classificadores no cenário balanceado.

Para o **Random Forest**, o algoritmo convergiu para árvores mais profundas (*max_depth* = 9) e um número de estimadores próximo a 30, sugerindo que o modelo precisou de maior complexidade para capturar as nuances das fronteiras de decisão geradas pelos dados sintéticos do ADASYN.

Já para a **Multilayer Perceptron (MLP)**, a função de ativação *tanh* demonstrou melhor convergência que a *relu* neste espaço de busca, com uma topologia de duas camadas ocultas (11 e 14 neurônios), indicando uma preferência por uma rede mais profunda e estreita para este problema específico.

A Tabela 12 resume os melhores hiperparâmetros encontrados durante a validação cruzada.

Tabela 12: Melhores Hiperparâmetros Encontrados (Cenário ADASYN)

Modelo	**Configuração Otimizada**
Random Forest	`n_estimators`: 29, `max_depth`: 9, `min_samples_split`: 4
MLP	`activation`: 'tanh', `hidden_layer_sizes`: (11, 14), `alpha`: 0.0327

A Tabela 12 detalha as configurações ótimas encontradas pela Evolução Diferencial para o cenário balanceado. Nota-se que o Random Forest exigiu árvores mais profundas (*max_depth* = 9) comparado ao cenário original, sugerindo a necessidade de modelos mais complexos para capturar as nuances das amostras sintéticas criadas pelo ADASYN. Já o MLP convergiu para uma função de ativação tangente hiperbólica (*tanh*) e uma arquitetura de duas camadas ocultas, ideal para modelar as fronteiras de decisão não-lineares induzidas pelo balanceamento.

6.1 Análise das Matrizes de Confusão

As matrizes de confusão (Figuras 3 e 4) ilustram visualmente a mudança de comportamento dos classificadores no conjunto de teste.

- **Cenário Original :** O modelo apresentou um comportamento conservador, privilegiando a classe majoritária. Embora tenha cometido apenas 1 erro de Falso Positivo, falhou em detectar 7 casos de óbito (Falsos Negativos).
- **Cenário Balanceado :** Com o uso do ADASYN, o modelo conseguiu recuperar mais casos da classe de interesse, elevando os Verdadeiros Positivos para 17 e reduzindo os Falsos Negativos para 5.

O aumento no número de Falsos Positivos (de 1 para 7) no cenário balanceado é uma consequência esperada e aceitável neste domínio. No contexto de insuficiência cardíaca, o **Erro do Tipo II (Falso Negativo)** — prever que um paciente irá sobreviver quando ele está em risco de morte — é muito mais custoso do que o Erro do Tipo I. Portanto, a matriz do cenário balanceado demonstra um modelo mais seguro para auxílio à decisão médica.

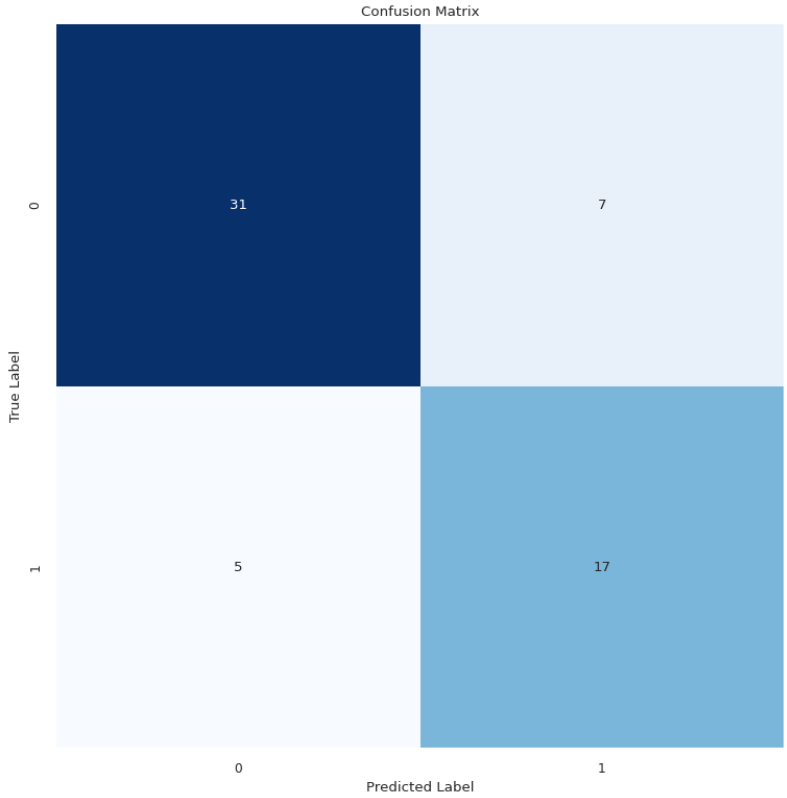

Figura 3: Matriz de Confusão do Experimento 1

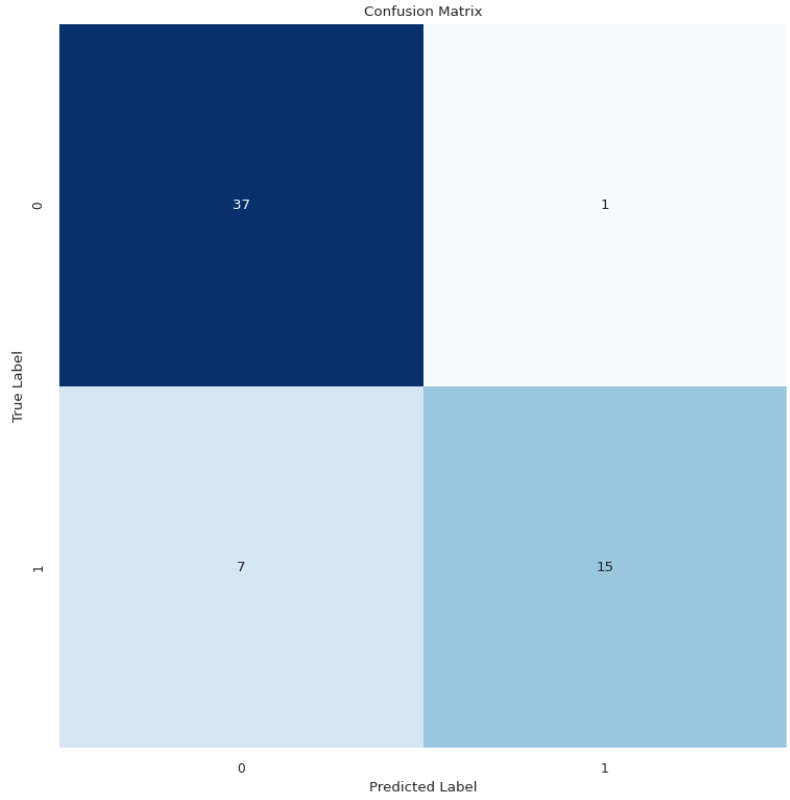

Figura 4: Matriz de Confusão do Experimento 2

Diferença de Arquitetura

7 Conclusão

Este estudo dedicou-se a investigar como o tratamento do desbalanceamento de classes e a otimização evolutiva de hiperparâmetros influenciam a predição de mortalidade em pacientes com insuficiência cardíaca. Desde o início, a análise dos dados do Heart Failure Clinical Records evidenciou que a desproporção entre as classes representava um obstáculo significativo, levando os classificadores a priorizarem a classe majoritária dos sobreviventes. A introdução da técnica ADASYN revelou-se uma intervenção decisiva nesse cenário. Ao gerar amostras sintéticas nas fronteiras de decisão mais complexas, o método permitiu reequilibrar o aprendizado, garantindo que os modelos não negligenciassem os casos críticos de óbito.

No comparativo entre os algoritmos, o Random Forest destacou-se por sua robustez e consistência frente à Multilayer Perceptron (MLP). Enquanto a rede neural demandou adaptações estruturais significativas para lidar com o balanceamento, como a mudança nas funções de ativação, o Random Forest manteve um desempenho estável e superior. O resultado mais impactante dessa abordagem foi o aumento estatisticamente validado da sensibilidade (Recall) para 86,1%. Embora esse ganho tenha custado uma leve redução na acurácia global ao testar o modelo em dados reais, essa troca reflete uma decisão estratégica consciente: no contexto médico, a prioridade ética é minimizar os falsos negativos, assegurando que pacientes em risco iminente sejam corretamente identificados e tratados.

A eficácia dessa abordagem foi sustentada pela Otimização por Evolução Diferencial, que explorou o espaço de hiperparâmetros de forma eficiente para encontrar configurações que o ajuste manual dificilmente alcançaria. A estabilidade dos resultados, confirmada pelos baixos desvios-

padrão nas trinta execuções independentes, juntamente com a validação pelo teste de Wilcoxon, reforça a confiabilidade do modelo proposto. Em suma, a combinação do ADASYN com um Random Forest otimizado entregou uma ferramenta de triagem segura e eficaz. Para avanços futuros, sugere-se explorar métodos de Ensemble híbridos ou funções de custo sensíveis ao erro, buscando elevar a precisão global sem comprometer a alta taxa de detecção de riscos já alcançada.

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.