Aplicação de Algoritmos de Aprendizado de Máquina na Classificação de Tipos de Anemia: Uma Abordagem Baseada em Dados Clínicos

Vinicius de Souza Santos

*Email do autor: viniciussouza742@gmail.com

Resumo: *A anemia é uma condição prevalente cujo diagnóstico diferencial pode ser apoiado por modelos de aprendizado de máquina a partir de hemogramas (CBC). Este estudo compara algoritmos para* multiclasse *usando um conjunto público (Kaggle) com 15 variáveis hematológicas; após limpeza (remoção de duplicatas e ausentes; sem remoção global de outliers) o conjunto consolidado contou com 596 instâncias. Cada modelo foi encapsulado em pipelines reprodutíveis com SMOTE aplicado* ***apenas*** *no treino e validação cruzada aninhada (nested CV). O critério primário foi o F1 ponderado, sendo F1 macro e acurácia métricas secundárias; também analisamos ROC/PR e matriz de confusão. Os* ensembles *tabulares foram superiores: o* ***XGBoost*** *obteve* $F1_{ponderado} = 0{,}978$ *e acurácia* $= 0{,}978$ *no teste, enquanto a* ***Random Forest*** *apresentou melhor equilíbrio entre classes (*$F1_{macro} = 0{,}896$ *e* AP_{macro} *maior). As principais confusões foram clinicamente plausíveis (e.g., ferropriva vs outras microcíticas; normocítica hipocrômica vs normocrômica; leucemia vs leucemia com trombocitopenia), e as curvas PR mostraram-se mais informativas que as ROC em cenário desbalanceado. Recomendamos, para implantação, calibração de probabilidades, limiares por classe (*one-vs-rest*) e abstenção assistida em casos de baixa margem. Conclui-se que* ensembles *baseados em árvores, com governança adequada, podem aprimorar a triagem e o apoio à decisão na classificação de anemias a partir de CBC, sem substituir o julgamento clínico.*

Palavras-chave: Hemograma + aprendizado de máquina na classificação de anemias: desempenho consistente, leitura clínica das confusões e boas práticas de adoção.

---

1 Introdução

A anemia é uma condição médica caracterizada pela diminuição da quantidade de glóbulos vermelhos ou hemoglobina no sangue, afetando a capacidade de transporte de oxigênio pelo organismo (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2025). Existem diversos tipos de anemia, como a ferropriva, a megaloblástica e a anemia de células falciformes,

com diferentes causas subjacentes, como a deficiência nutricional ou fatores genéticos (SILVA et al., 2016), (SCHOUERI JÚNIOR et al., 1990). A detecção precoce e precisa dessas condições é essencial para a implementação de tratamentos adequados, especialmente em populações de risco, como crianças e gestantes (ASARE; APPIAHENE; DONKOH, 2023). A caracterização e o diagnóstico das diversas formas de anemia podem ser aprimorados com o uso de estatísticas, como a análise de dados clínicos por meio de aprendizado de máquina, conforme mostrado por estudos recentes (ELSABAGH et al., 2023), (KARAGÜL YILDIZ; YURTAY; ÖNEÇ, 2021). A aplicação dessas técnicas tem se mostrado eficaz na classificação de tipos de anemia com base em exames laboratoriais, como a contagem de células sanguíneas e a análise da morfologia dos glóbulos vermelhos, contribuindo significativamente para um diagnóstico mais rápido e preciso (ASARE; APPIAHENE; DONKOH, 2023). No contexto da anemia, os modelos de aprendizado de máquina, especialmente as redes neurais artificiais, têm se mostrado altamente eficazes na classificação de diferentes tipos com base em exames laboratoriais, como a contagem de células sanguíneas e a morfologia dos glóbulos vermelhos (UVALIYEVA; BELGINOVA; BOROZENETS, 2023), (KOVAČEVIĆ et al., 2022). Diversos estudos, como o de (ELMALEEH, 2024), que comparou três modelos de redes neurais, e (KAZEMINIA et al., 2022), que propuseram um método eficaz para classificar distúrbios anêmicos raros, confirmam a eficácia dessas abordagens na análise de dados médicos complexos. Além disso, o uso de aprendizado supervisionado tem se expandido na área da saúde, permitindo a análise de grandes volumes de dados clínicos e a identificação de padrões complexos, como o reconhecimento morfológico e colorimétrico de eritrócitos (NANGO et al., 2023). Esses avanços não só melhoram a detecção precoce de anemias, mas também ampliam as perspectivas para o diagnóstico automatizado de doenças hematológicas. Modelos supervisionados, como as redes neurais artificiais, demonstram um grande potencial para identificar padrões complexos, facilitando diagnósticos mais rápidos e precisos, especialmente na classificação de condições como a anemia, que continua a ser um dos maiores desafios para a medicina moderna (GARCIA, 2020), (MOREIRA; SALERNO; TSUNODA et al., 2020). Em resumo, este estudo investiga como algoritmos de aprendizado de máquina podem melhorar a classificação de anemias a partir do CBC, utilizando um conjunto público e focando em métricas robustas ao desbalanceamento e em requisitos de implantação (calibração, limiares e abstenção).

2 Materiais e Métodos

2.1 Fonte e Estrutura do Conjunto de Dados

Este estudo utilizou o conjunto público *"Anemia Types Classification"*, disponibilizado na plataforma Kaggle por (ABOELNAGA) (ABOELNAGA, 2023).[1] O pacote possui tamanho aproximado de 125,59 KB e compreende 15 variáveis (numéricas e categóricas) derivadas

1 https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification (acesso em 21/08/2025). Licença: Apache-2.0.

do hemograma completo (CBC – *Complete Blood Count*). Após saneamento (remoção de duplicatas e valores ausentes), obteve-se um conjunto analítico com $N = 596$ instâncias. O manejo de valores extremos é descrito no protocolo de pré-processamento (ver §2.4).

Variáveis disponíveis. As colunas do CBC capturam contagens celulares e índices eritrocitários/plaquetários de uso rotineiro em hematologia: **WBC**, **LYMp**, **NEUTp**, **LYMn**, **NEUTn**, **RBC**, **HGB**, **HCT**, **MCV**, **MCH**, **MCHC**, **PLT**, **PDW**, **PCT** e a variável-alvo **Diagnosis**. Marcadores bioquímicos (ferritina, vitamina B12, folato) **não** estão disponíveis; portanto, a classificação neste trabalho é *fenotípica* com base *exclusivamente* no CBC.

Taxonomia fenotípica (nove classes). Os diagnósticos observáveis no CBC foram organizados em **nove classes fenotípicas**: *Healthy*, *Iron deficiency anemia* (IDA), *Other microcytic anemia* (OMA), *Normocytic hypochromic anemia* (NHA), *Normocytic normochromic anemia* (NNA), *Macrocytic anemia*, *Thrombocytopenia* (TP), *Leukemia* e *Leukemia with thrombocytopenia* (Leukemia+TP). A distribuição é desbalanceada, com menor prevalência para *Macrocytic anemia* e *Leukemia+TP*.

Mapeamento dos rótulos originais. O Quadro 1 mostra o mapeamento 1–1 entre os rótulos do arquivo bruto e as nove classes fenotípicas adotadas (apenas normalização ortográfica quando necessário).

Tabela 1 – Mapeamento entre os rótulos originais do dataset e as classes fenotípicas utilizadas.

Rótulo original (dataset)	**Classe fenotípica (este estudo)**
Healthy	Healthy
Iron deficiency anemia	Iron deficiency anemia (IDA)
Other microcytic anemia	Other microcytic anemia (OMA)
Normocytic hypochromic anemia	Normocytic hypochromic anemia (NHA)
Normocytic normochromic anemia	Normocytic normochromic anemia (NNA)
Macrocytic anemia	Macrocytic anemia
Thrombocytopenia	Thrombocytopenia (TP)
Leukemia	Leukemia
Leukemia with thrombocytopenia	Leukemia with thrombocytopenia (Leukemia+TP)

2.2 Objetivo e Abordagem de Modelagem

O objetivo deste trabalho foi desenvolver e *comparar sistematicamente* diferentes algoritmos de classificação multiclasse para identificar tipos de anemia a partir de parâmetros hematológicos (CBC), priorizando desempenho preditivo, performance dos modelos e reprodutibilidade. A tarefa foi formulada no paradigma de aprendizado supervisionado, em que se aprende um mapeamento de variáveis preditoras para rótulos previamente anotados por especialistas.

A abordagem adotada baseou-se em um *benchmark* entre famílias de modelos complementares, evitando a concentração em uma única classe de algoritmos. Foram avaliados: Regressão Logística (modelo linear probabilístico), SVM (margem máxima com núcleos), *k*-Nearest Neighbors (método baseado em instâncias), Random Forest (conjunto de árvores), XGBoost (gradient boosting para dados tabulares), Naive Bayes Gaussiano (probabilístico com suposição de normalidade) e uma Rede Neural Multicamadas (MLP) de baixa complexidade como representante de *deep learning* para dados tabulares. Essa seleção visa cobrir hipóteses lineares e não lineares, fronteiras de decisão locais e globais, além de modelos de *ensemble* com diferentes vieses e variâncias.

Para assegurar comparação justa e prevenir vazamento de informação, cada algoritmo foi encapsulado em um *pipeline* independente, contemplando: (i) padronização (`StandardScaler`) quando o estimador é sensível à escala (Regressão Logística, SVM, KNN e MLP); (ii) balanceamento via SMOTE aplicado **exclusivamente** ao subconjunto de treino de cada partição; e (iii) o estimador final. A seleção de hiperparâmetros foi realizada por busca em grade (`GridSearchCV`) dentro de validação cruzada estratificada interna, enquanto o desempenho foi estimado em validação cruzada estratificada externa (*nested cross-validation*), mitigando o otimismo na seleção de modelos.

O critério primário de seleção foi o **F1-score ponderado** (ponderação pelo suporte das classes), adequado ao cenário originalmente desbalanceado; como critérios secundários, consideraram-se F1 macro e acurácia. Quando aplicável, probabilidades foram calibradas para avaliação adicional por métricas baseadas em probabilidade (e.g., Brier score, AUCs macro *one-vs-rest*). A reprodutibilidade foi assegurada por fixação de sementes pseudoaleatórias em todas as etapas (`random_state`) e por separação estratificada treino–teste, na qual o modelo selecionado ao final do *benchmark* foi readequado ao conjunto de treinamento completo e avaliado no conjunto de teste retido.

2.3 Procedimentos de Análise

A análise foi conduzida integralmente na linguagem de programação Python, utilizando as seguintes bibliotecas:

- `Pandas` e `NumPy` para manipulação e transformação dos dados;
- `Matplotlib` e `Seaborn` para análise exploratória e visualização gráfica;
- `Scikit-Learn` para construção, treinamento e avaliação dos modelos preditivos;
- `Imbalanced-learn` para tratamento do desbalanceamento de classes.

O conjunto de dados foi particionado em dois subconjuntos: 80% dos dados foram utilizados para treinamento e 20% para teste. Técnicas de normalização dos atributos numéricos

foram aplicadas utilizando `StandardScaler`, garantindo que todas as variáveis tivessem média zero e desvio padrão igual a um, o que é essencial para o desempenho adequado de algoritmos sensíveis à escala dos dados, como redes neurais e regressão logística.

A robustez dos modelos foi avaliada por meio de métricas clássicas de classificação: acurácia, precisão, *recall* e F1-Score. A avaliação foi conduzida tanto no conjunto de teste quanto por meio de validação cruzada estratificada com 5 *folds*, permitindo melhor estimativa da capacidade de generalização.

Detalhes adicionais sobre o pré-processamento dos dados, escolha dos algoritmos, ajuste de hiperparâmetros e resultados quantitativos da avaliação encontram-se nas seções subsequentes deste artigo.

2.4 Pré-processamento dos Dados

O conjunto de dados, originalmente composto por 15 variáveis provenientes de exames hematológicos (CBC), foi submetido a um protocolo de pré-processamento para garantir consistência, reprodutibilidade e adequação às etapas de modelagem. As ações contemplaram: (i) saneamento e integridade, (ii) tratamento de valores ausentes, (iii) detecção e manejo de *outliers*, (iv) padronização de escalas e (v) codificação de variáveis.

Saneamento e integridade. Inicialmente, procedeu-se à verificação de tipos (*type casting*) e à harmonização de nomes de colunas, assegurando correspondência semântica entre rótulos e grandezas laboratoriais. Instâncias duplicadas foram removidas por comparação exata de registros, evitando redundâncias e garantindo que cada linha representasse um único paciente. Também se verificou a presença de valores não numéricos em campos contínuos, os quais foram coerentemente convertidos ou invalidados, conforme apropriado.

Valores ausentes. Para lidar com dados faltantes, adotou-se exclusão por lista completa (*listwise deletion*) quando presente qualquer valor ausente em variáveis preditoras essenciais. Essa decisão visou evitar vieses de imputação em um conjunto de amostras relativamente enxuto e manter a comparabilidade entre observações usadas no treinamento e avaliação. Alternativas baseadas em imputação múltipla foram consideradas metodologicamente, mas não aplicadas neste estudo.

Detecção e manejo de *outliers*. Para evitar vazamento de informação, **nenhuma** exclusão de observações por *outliers* foi realizada globalmente. Em vez disso, adotou-se uma estratégia *in pipeline*: em cada *fold* de treino, calcularam-se os limiares baseados em IQR (ou, alternativamente, limites percentílicos) **apenas no treino** e aplicou-se *winsorização* (clipping) às variáveis contínuas via transformador dentro do `Pipeline`/`ImbPipeline`. Os parâmetros ajustados no treino foram então utilizados para transformar validação e teste no respectivo *fold*. Essa abordagem preserva o tamanho amostral e evita otimismo indevido por remoção

global de casos.

Padronização de escalas. Com o objetivo de tornar comparáveis as magnitudes das variáveis e favorecer algoritmos sensíveis à escala (e.g., modelos lineares), as características contínuas foram padronizadas via transformação z (média zero e desvio padrão unitário). Para prevenir vazamento de informação (*data leakage*), os parâmetros de padronização (médias e desvios) foram ajustados exclusivamente no conjunto de treinamento e posteriormente aplicados ao conjunto de teste.

Codificação da variável-alvo e diagnóstico de classes. A variável *Diagnóstico* (alvo) foi codificada por rótulos inteiros (*label encoding*) para viabilizar o treinamento. Antes do balanceamento, inspecionou-se a distribuição das classes a fim de caracterizar possíveis assimetrias de frequência entre tipos de anemia. O balanceamento sintético entre classes, quando aplicável, foi realizado apenas após a divisão treino–teste e restrito ao subconjunto de treinamento, de modo a preservar a validade da avaliação e evitar vazamento.

Reprodutibilidade. Sempre que pertinente, fixou-se a semente pseudoaleatória (`random_state`) nas rotinas de divisão e de treinamento, assegurando reprodutibilidade dos resultados.

O pipeline de pré-processamento priorizou a eliminação de inconsistências (duplicatas, ausentes e *outliers*), a normalização de escalas e a preparação apropriada da variável-alvo, estabelecendo as condições para treinamento justo e avaliação robusta dos modelos.

2.5 Desenvolvimento e Avaliação dos Modelos de Machine Learning

A tarefa de classificação multiclasse foi formulada sob o paradigma de aprendizado supervisionado. Para assegurar reprodutibilidade e evitar vazamento de informação (*data leakage*), adotou-se um *pipeline* único por modelo, composto por: (i) balanceamento sintético via SMOTE (aplicado **apenas** no conjunto de treino de cada *fold*); (ii) padronização (`StandardScaler`) quando o algoritmo for sensível à escala; e (iii) estimador final. Esse arranjo foi implementado com `Pipeline`/`ImbPipeline`, garantindo que todas as transformações fossem ajustadas exclusivamente no treino e posteriormente aplicadas ao teste em cada partição.

A avaliação seguiu *nested cross-validation*: uma validação cruzada estratificada externa com k=5 *folds* para estimar o desempenho generalizável e uma validação interna com k=3 *folds* para seleção de hiperparâmetros (`GridSearchCV`). A métrica primária adotada foi o **F1-score ponderado** (ponderado pelo suporte de cada classe), por refletir melhor o desempenho em cenários originalmente desbalanceados; a **acurácia** e o **F1 macro** foram utilizadas como métricas secundárias. Quando aplicável, estimativas de probabilidade foram calibradas (Platt/*isotonic*) no conjunto de validação interna.

2.5.1 Random Forest (RF)

A Random Forest foi empregada como modelo de referência baseado em conjuntos (*ensemble*) de árvores, capaz de modelar relações não lineares e interações entre variáveis. Por ser **invariante à escala**, não se realizou padronização neste *pipeline*. A busca contemplou:

- `n_estimators` $\in \{100, 200, 300, 500\}$,
- `max_depth` $\in \{$`None`$, 10, 20, 30\}$,
- `max_features` $\in \{$`sqrt`$,$ `log2`$\}$,
- `min_samples_split` $\in \{2, 5, 10\}$,
- `min_samples_leaf` $\in \{1, 2, 4\}$.

A importância de atributos foi posteriormente inspecionada por importância por permutação, mitigando o viés conhecido de importâncias baseadas em impureza.

2.5.2 Regressão Logística (RL)

A Regressão Logística, modelo linear probabilístico, foi utilizada como *baseline* pela interpretabilidade dos coeficientes e eficiência computacional. Requereu padronização dos preditores. A busca incluiu:

- `C` $\in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ (inverso da regularização),
- `penalty` $=$ `l2`,
- `solver` $\in \{$`lbfgs`$,$ `liblinear`$\}$,
- `class_weight` $\in \{$`None`$,$`balanced`$\}$ (sensibilidade exploratória a desbalanceamento).

2.5.3 k-Nearest Neighbors (KNN)

O KNN foi avaliado por capturar estruturas locais no espaço de características. Por depender de distâncias, aplicou-se padronização. Hiperparâmetros:

- `n_neighbors` $\in \{3, 5, 7, 11, 15, 21\}$,
- `weights` $\in \{$`uniform`$,$ `distance`$\}$,
- `metric` $=$ `minkowski`, `p` $\in \{1, 2\}$.

2.5.4 Máquinas de Vetores de Suporte (SVM)

O SVM foi considerado pela capacidade de maximização de margens e uso de *kernels*. Foi utilizada padronização. A busca contemplou:

- `kernel` $\in \{$`linear`, `rbf`$\}$,
- `C` $\in \{0.1, 1, 10, 100\}$,
- `gamma` $\in \{$`scale`, `auto`$, 10^{-3}, 10^{-2}\}$ (quando `rbf`),
- `decision_function_shape` $=$ `ovr` (multiclasse).

Para métricas baseadas em probabilidade, ativou-se `probability=True` com calibragem posterior quando necessário.

2.5.5 XGBoost

Modelos de *gradient boosting* foram incluídos pela alta acurácia em dados tabulares e controle refinado de *bias-variance*. Não foi empregada padronização. Grade principal:

- `n_estimators` $\in \{200, 400, 600\}$,
- `learning_rate` $\in \{0.05, 0.1, 0.2\}$,
- `max_depth` $\in \{3, 5, 8\}$,
- `subsample` $\in \{0.7, 1.0\}$, `colsample_bytree` $\in \{0.7, 1.0\}$,
- `reg_lambda` $\in \{1, 5, 10\}$, `reg_alpha` $\in \{0, 0.1\}$.

2.5.6 Naive Bayes Gaussiano (GNB)

Como classificador probabilístico simples e eficiente para atributos contínuos, avaliou-se o GNB, que assume normalidade condicional por classe. Não requer padronização estritamente, porém a padronização foi mantida por consistência do *pipeline*. Foi explorado:

- `var_smoothing` $\in \{10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}\}$.

2.5.7 Rede Neural Multicamadas (MLP)

Como representante de *deep learning* leve para dados tabulares, empregou-se um MLP com função de ativação ReLU. Foram aplicadas padronização e *early stopping* (paciência ≥ 10 épocas) para regularização. Como o `MLPClassifier` do `scikit-learn` não implementa *dropout*, utilizou-se regularização L2 (`alpha`) e *early stopping*. A busca incluiu:

- `hidden_layer_sizes` $\in \{(64,), (128,), (64, 32), (128, 64)\}$,
- `alpha` (L2) $\in \{10^{-5}, 10^{-4}, 10^{-3}\}$,
- `learning_rate_init` $\in \{10^{-3}, 3 \cdot 10^{-3}\}$,
- `batch_size` $\in \{32, 64\}$,
- `max_iter` $\in \{200, 300, 500\}$, `early_stopping` $=$ `True`.

Para estimação probabilística estável, utilizou-se `softmax` na camada de saída.

2.5.8 Esquema de Seleção, Métricas e Interpretação

O desempenho foi sumarizado por média $\pm$ desvio-padrão ao longo dos *folds* externos. O modelo de produção foi escolhido pelo maior F1 ponderado na validação externa; em caso de empate, privilegiou-se o F1 macro e, subsequentemente, a acurácia. Para interpretação, foram consideradas: (i) importância por permutação (RF/XGBoost); (ii) coeficientes padronizados (RL/SVM linear); e (iii) *SHAP values* (quando pertinente) para análise de contribuições locais e globais das variáveis.

Observação sobre balanceamento. O SMOTE foi aplicado exclusivamente no subconjunto de treino de cada *fold* interno, preservando a distribuição real no teste e prevenindo vazamento. Para algoritmos que suportam `class_weight`, realizou-se análise de sensibilidade comparando SMOTE versus ponderação de classes. Além do SMOTE padrão, avaliou-se, em análise de sensibilidade, o uso de `class_weight` (quando suportado), bem como variantes *Borderline-SMOTE* e SMOTE-ENN, ajustando k conforme a raridade de cada classe. Para mitigar alta variância em classes raras, adotou-se *Repeated Stratified* CV na análise complementar.

2.6 Avaliação com Validação Cruzada

A capacidade de generalização dos modelos foi estimada por validação cruzada estratificada. Empregou-se um arranjo de validação cruzada aninhada (*nested cross-validation*), no qual uma validação interna é utilizada para seleção de hiperparâmetros e uma validação externa para estimação imparcial do desempenho. Especificamente, adotou-se validação cruzada externa com k=5 *folds* estratificados e, em cada partição externa, uma validação cruzada interna com k=3 *folds* para a busca em grade (`GridSearchCV`). A estratificação preservou as proporções das classes em todas as divisões.A validação aninhada foi executada exclusivamente no subconjunto de treinamento (80%); o conjunto de teste (20%) permaneceu completamente fora do processo de seleção/ajuste e foi usado apenas para a avaliação final.

Para prevenir vazamento de informação (*data leakage*), todas as transformações do *pipeline* (padronização e balanceamento por SMOTE) foram ajustadas exclusivamente sobre os dados de treinamento de cada *fold* e, em seguida, aplicadas aos dados de validação correspondentes. O SMOTE foi utilizado apenas no subconjunto de treino de cada partição.

A métrica primária de seleção foi o **F1-score ponderado** (ponderação pelo suporte de cada classe), apropriada ao cenário originalmente desbalanceado. Como métricas secundárias, foram consideradas **F1 macro**, **acurácia** e a **matriz de confusão** por classe. Quando aplicável, foram avaliadas também métricas baseadas em probabilidade (e.g., *Brier score*, ROC-AUC e PR-AUC em regime *one-vs-rest*). As estimativas de desempenho foram reportadas como média e desvio padrão ($\mu \pm \sigma$) ao longo dos *folds* externos; intervalos de confiança (95%) foram obtidos por *bootstrap* estratificado quando indicado.

O modelo selecionado em cada comparação foi aquele que apresentou maior F1 ponderado na validação externa; em caso de empate, utilizou-se como critérios de desempate, sequencialmente, F1 macro e acurácia. Após a seleção, o estimador escolhido foi readequado ao conjunto completo de treinamento e avaliado no conjunto de teste retido, cuja análise quantitativa é apresentada na Seção *Resultados*.

2.7 Métricas de Desempenho

A avaliação dos modelos foi conduzida em regime de validação cruzada estratificada e conjunto de teste retido, sempre no *hold-out* de cada partição, evitando vazamento de informação. Em cada *fold*, calcularam-se métricas no conjunto de validação; ao final, reportaram-se médias e desvios padrão ($\mu \pm \sigma$) ao longo dos *folds* externos. A métrica primária adotada foi o **F1 ponderado** (ponderação pelo suporte de cada classe); como métricas secundárias, consideraram-se **acurácia**, **F1 macro**, **matriz de confusão** e, quando aplicável, **ROC-AUC** e **PR-AUC** macro-ovr, **kappa de Cohen**, **MCC** e **Brier score**.

Notação e matriz de confusão. No cenário multiclasse com K classes, define-se a matriz de confusão $\mathbf{C} \in \mathbb{N}^{K\times K}$, em que C_{ij} é o número de instâncias da classe verdadeira i previstas como classe j. Para a classe k:

$$\mathrm{TP}_k = C_{kk}, \quad \mathrm{FP}_k = \sum_{i\neq k} C_{ik}, \quad \mathrm{FN}_k = \sum_{j\neq k} C_{kj}, \quad \mathrm{TN}_k = \sum_{i\neq k}\sum_{j\neq k} C_{ij}.$$

Matrizes normalizadas por linha (percentuais por classe verdadeira) foram empregadas para comparação visual entre modelos.

Acurácia. A acurácia global mede a proporção de acertos:

$$\mathrm{Acc} = \frac{1}{N}\sum_{k=1}^{K} C_{kk},$$

onde N é o número total de instâncias no conjunto avaliado.

Precisão, *recall* e F1 por classe. Para cada classe k:

$$\mathrm{Precision}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \qquad \mathrm{Recall}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \qquad \mathrm{F1}_k = \frac{2\cdot \mathrm{Precision}_k \cdot \mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k}.$$

Tais métricas são apropriadas quando os custos de falsos positivos/negativos não são simétricos.

Agregações no multiclasse. Para sintetizar desempenho global, empregaram-se três agregações usuais:

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{F1}_k, \qquad \text{Weighted-F1} = \sum_{k=1}^{K} w_k \mathrm{F1}_k, \quad w_k = \frac{n_k}{N},$$

$$\text{Micro-F1} = \frac{2 \cdot \sum_k \text{TP}_k}{2 \cdot \sum_k \text{TP}_k + \sum_k \text{FP}_k + \sum_k \text{FN}_k}.$$

O *F1 macro* confere o mesmo peso a todas as classes (sensível a classes raras), enquanto o *F1 ponderado* pondera pelo suporte n_k e é menos volátil quando a distribuição é assimétrica. O *F1 micro* equivale, no multiclasse, à precisão/recall globais.

Curvas ROC e PR; AUC. Quando os modelos fornecem escores contínuos/probabilidades, calcularam-se: (i) ROC-AUC *one-vs-rest* (macro), que integra a curva taxa de verdadeiros positivos vs. taxa de falsos positivos para cada classe, e (ii) PR-AUC (área sob a curva Precisão–*Recall*) *macro-ovr*, mais informativa em cenários desbalanceados por enfatizar o equilíbrio entre precisão e cobertura nas classes minoritárias.

Acordo além do acaso e correlação. Para complementar a acurácia, reportou-se:

$$\kappa \text{ de Cohen} \quad (\text{acordo além do acaso}) \quad \text{e} \quad \text{MCC} \in [-1, 1],$$

onde MCC (coeficiente de correlação de Matthews) resume o desempenho considerando todas as células da matriz de confusão; seu uso é recomendado em presença de desbalanceamento. (Para multiclasse, utilizaram-se as extensões disponíveis em bibliotecas padrão.)

Calibração probabilística. Para modelos probabilísticos, avaliou-se a calibração por meio do *Brier score*:

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} (1 - \hat{p}_{i,y_i})^2,$$

em que $\hat{p}_{i,y_i}$ é a probabilidade prevista para a classe verdadeira y_i do exemplo i. Quando necessário, aplicou-se calibração (Platt ou isotônica) em dados de validação.

Incerteza e comparação entre modelos. Para quantificar incerteza, reportaram-se intervalos de confiança (95%) por *bootstrap* estratificado ($B \geq 1000$ reamostragens) para acurácia e *F1*. Para comparação pareada de modelos, utilizou-se *bootstrap* pareado no mesmo *hold-out*; para AUC, quando pertinente, empregou-se o teste de DeLong em configuração *one-vs-rest*.

Critério de seleção. O modelo de produção foi escolhido pelo maior *F1 ponderado* na validação externa. Em caso de empate, priorizou-se, sequencialmente, *F1 macro* e acurácia. Todas as métricas finais foram ainda reportadas no conjunto de teste retido, acompanhadas da matriz de confusão normalizada por classe.

3 Resultados e Discussão

3.1 Visão geral do desempenho

Após a seleção por *nested cross-validation*, os modelos foram readequados ao conjunto de treino completo e avaliados em um *hold-out* estratificado (20%). A Figura 1 sintetiza três

métricas no teste: Acurácia, F1 macro e F1 ponderado.

Ranking pelo critério primário. Pelo **F1 ponderado**—mais adequado ao desbalanceamento original—o **XGBoost** foi o melhor ($F1_w = 0{,}978$), seguido de **Random Forest** (0,969), **SVM** (0,862), **MLP** (0,831), **Regressão Logística** (0,794), **KNN** (0,717) e **GaussianNB** (0,496). A **acurácia** reproduziu essencialmente a mesma hierarquia (*XGBoost* 0,978; *Random Forest* 0,970), indicando baixa variância de erro entre classes majoritárias.

Equilíbrio entre classes. No **F1 macro**—que confere peso igual a todas as classes—houve uma inversão sutil entre os dois líderes: **Random Forest** (0,896) superou ligeiramente o **XGBoost** (0,877), sugerindo melhor *recall* em classes raras. SVM (0,807) e MLP (0,763) formam um segundo pelotão, enquanto Logística (0,669), KNN (0,601) e GaussianNB (0,494) ficam atrás.

O que os "gaps" revelam. A diferença entre $F1_w$ e $F1_{macro}$ quantifica o quanto o desempenho está concentrado nas classes frequentes. Esse *gap* é pequeno para **Random Forest** ($\Delta \approx 0{,}073$) e moderado para **XGBoost** ($\Delta \approx 0{,}101$), o que explica a troca de posições no F1 macro. Em contraste, a **Regressão Logística** apresenta um *gap* maior ($\Delta \approx 0{,}125$), evidenciando viés para classes dominantes; já o **GaussianNB** tem $F1_w \approx F1_{macro}$ (diferença $\approx 0{,}002$), refletindo baixo desempenho relativamente uniforme.

Famílias de modelos. Os **ensembles** para dados tabulares (XGBoost/Random Forest) dominam nas três métricas, com ganhos absolutos expressivos sobre os **modelos lineares** e o **KNN** (e.g., XGBoost vs. Logística: $+0{,}184$ em $F1_w$; XGBoost vs. KNN: $+0{,}261$). O **SVM** e a **MLP** alcançam resultados competitivos, porém ainda atrás dos ensembles — um padrão consistente em cenários com fronteiras não lineares e atributos tabulares heterogêneos. O **GaussianNB** foi sensível às correlações entre variáveis hematológicas (hipótese de independência violada), justificando o desempenho inferior.

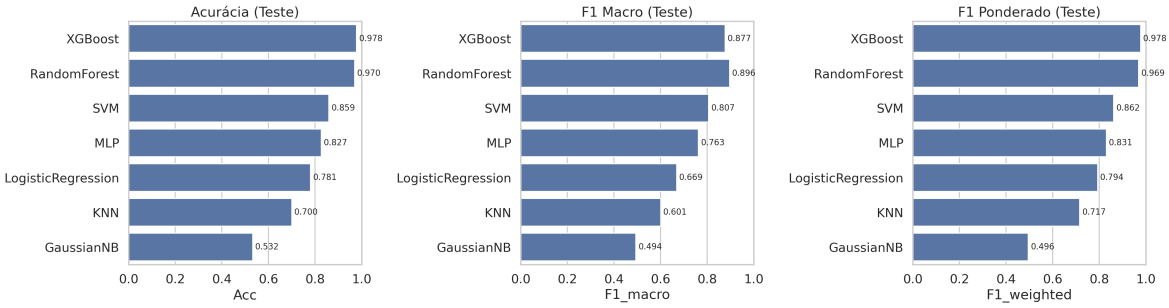

Figura 1 – Desempenho no conjunto de teste: Acurácia, F1 macro e F1 ponderado.

3.2 Métricas baseadas em probabilidade

As Figuras 2 e 3 apresentam as curvas ROC e Precisão–Recall (PR) com médias *micro* e *macro*. Recordando: a média *micro* pondera cada instância (favorecendo classes prevalentes),

enquanto a *macro* dá peso igual a todas as classes (mais sensível às raras).

ROC-AUC. Os modelos de *ensemble* apresentaram AUCs elevadas, indicando separabilidade consistente entre as classes no conjunto de teste. Em cenários desbalanceados e com baixo suporte em algumas categorias, contudo, as curvas ROC tendem a saturar próximo do teto e discriminam pouco entre os líderes. Por isso, priorizamos a PR-AUC (*macro*, *one-vs-rest*) em conjunto com $F1_{macro}$ e $F1_{ponderado}$, e reportamos intervalos de confiança de 95% obtidos via *bootstrap* estratificado; quando aplicável, comparamos AUCs pelo teste de DeLong.

Precisão–Recall (AP). Nas curvas PR, a linha de base de *cada classe* coincide com a sua *prevalência*; por isso, a PR-AUC *macro* é mais sensível ao desempenho nas classes raras. **Random Forest** e **XGBoost** tiveram $AP_{micro} \approx 0{,}997$ ambos, porém no $\mathbf{AP_{macro}}$ o Random Forest manteve vantagem (0,967 vs. 0,923), sugerindo maior *precisão sob altos níveis de recall* nas classes menos frequentes. **MLP** também foi consistente ($AP_{micro} = 0{,}918$; $AP_{macro} = 0{,}867$), seguido por **SVM** (0,842; 0,827). **KNN** (0,772; 0,662) e **Logística** (0,760; 0,712) mostraram quedas mais acentuadas quando a média ignora a prevalência. O **GaussianNB** ficou próximo à linha de base no início da curva e rapidamente perdeu precisão ($AP_{micro} = 0{,}530$), embora o $AP_{macro} = 0{,}584$ tenha superado o *micro*, indicando que seus erros concentram-se nas classes majoritárias.

Leituras práticas. (i) Com **ROC** saturada, a **PR** distingue melhor os modelos de topo; por isso, adotamos AP_{macro} como verificação adicional de equilíbrio entre classes. (ii) A vantagem do Random Forest em AP_{macro} corrobora o seu melhor $F1_{macro}$ na Seção anterior. (iii) Para implantação, recomenda-se **calibração** de probabilidades e **ajuste de limiar por classe** (otimizando $F1_{macro}$ ou custo clínico específico), uma vez que pequenas variações de limiar na faixa de *recall* alto impactam significativamente a precisão em classes raras.

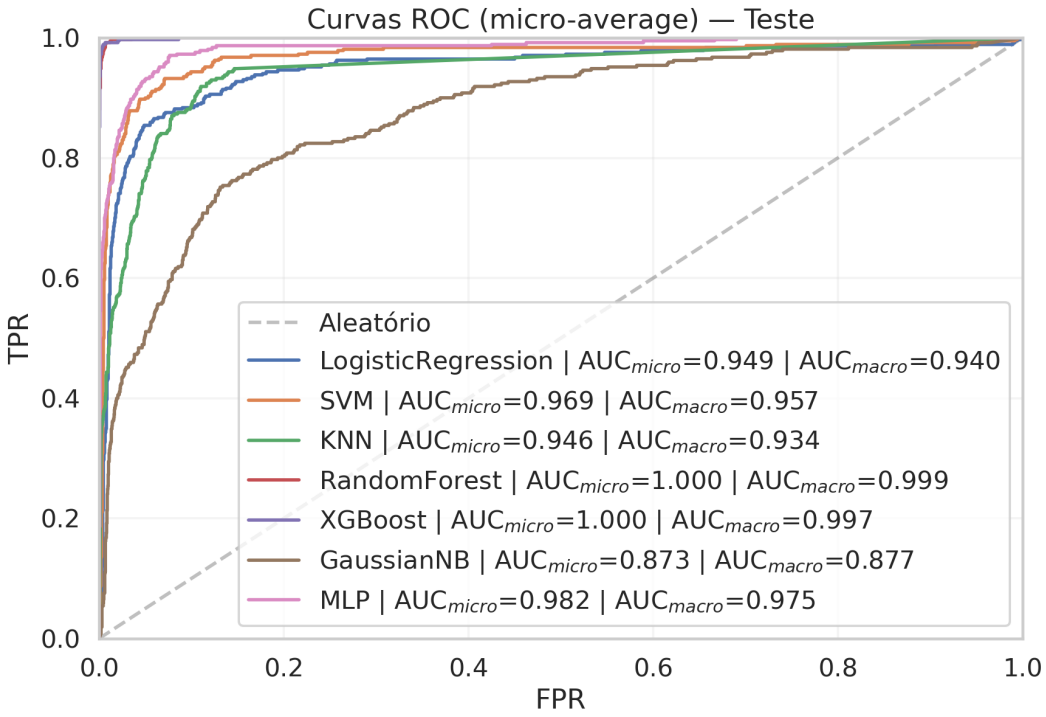

Figura 2 – Curvas ROC (médias *micro*/*macro*) no teste.

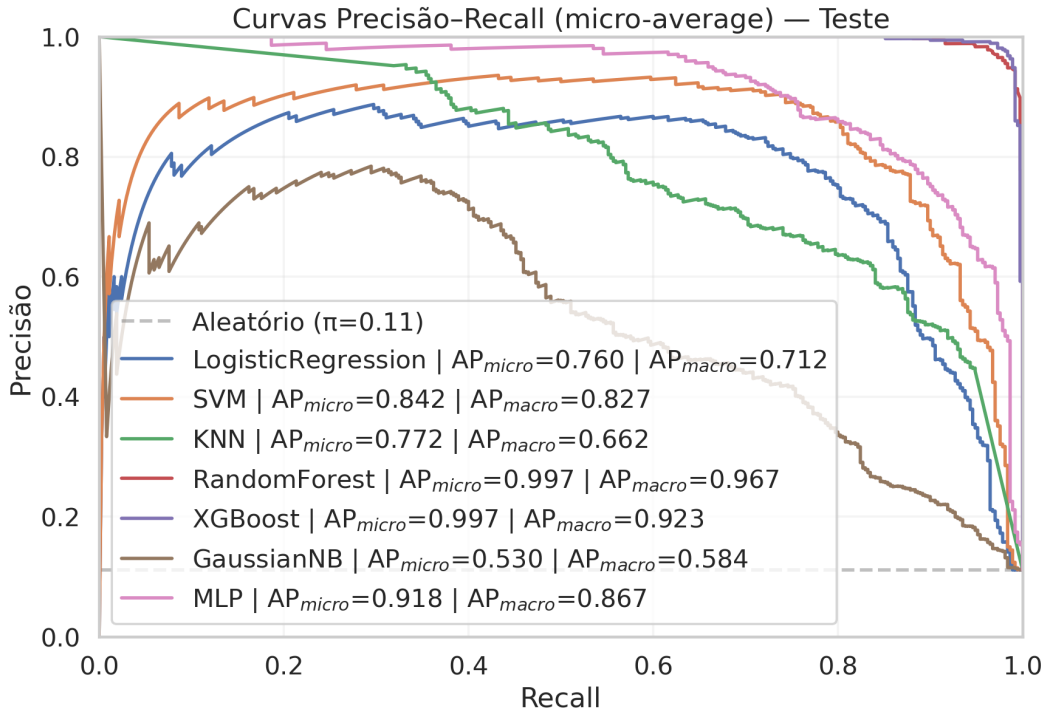

Figura 3 – Curvas Precisão–Recall (médias *micro*/*macro*) no teste. A linha pontilhada indica a linha de base (π = prevalência positiva).

3.3 Análise por classe e padrões de confusão

As matrizes de confusão foram disponibilizadas em formato textual (CSV; material suplementar) e analisadas na normalização por linha (*recall* por classe). No *hold-out* de teste, os suportes por classe foram: *Healthy* (97), *Normocytic hypochromic anemia* – NHA (81), *Normocytic normochromic anemia* – NNA (77), *Iron deficiency anemia* – IDA (55), *Thrombocytopenia* – TP (22), *Other microcytic anemia* – OMA (17), *Leukemia* (13), *Macrocytic anemia* (5) e *Leukemia with thrombocytopenia* – Leukemia+TP (3).

XGBoost (campeão por F1 ponderado). **Recall** por classe: Healthy 99,0%, IDA 98,2%, Leukemia 84,6%, Leukemia+TP 33,3%, Macrocytic 80,0%, NHA 100,0%, NNA 100,0%, OMA 100,0%, TP 95,5%. **Erros dominantes**: *Leukemia+TP* $\rightarrow$ *Leukemia* (66,7%, efeito do suporte muito baixo) e *Macrocytic* $\rightarrow$ *NHA* (20%). As demais confusões foram esparsas ($\leq$ 7,7%). Em **precisão**, as classes NHA, NNA, TP e IDA apresentaram pureza $\geq$ 98%.

Random Forest (melhor F1 macro). **Recall** por classe: Healthy 99,0%, IDA 100,0%, Leukemia 92,3%, Leukemia+TP 100,0%, NHA 98,8%, NNA 98,7%, OMA 82,4%, TP 95,5% e Macrocytic 40,0%. Mostrou-se *mais estável* nas classes raras de leucemia (acertos completos em Leukemia+TP), mas perdeu em *Macrocytic*, predominantemente confundida com NHA/NNA. Quanto à **precisão**, várias classes permaneceram muito altas (e.g., Healthy, Leukemia, OMA, TP $\approx$ 100%), ao passo que *Macrocytic* ficou com baixa pureza (50%).

Modelos basais. O **SVM** atingiu **recall** de 88,7% (Healthy), 90,9% (IDA), 96,1% (NNA) e 90,9% (TP), porém foi inferior aos ensembles nas classes hematológicas mais desafiadoras (e.g., Leukemia+TP = 66,7%; OMA com **precisão** $\approx$ 50%). **KNN** e **GaussianNB** apresentaram degradação geral: o KNN é sensível à escala/dimensionalidade e o NB sofre com correlações entre índices hematológicos (violação da independência condicional), o que explica tanto *recall* menor quanto maior confusão entre microcíticas (IDA/OMA) e entre NHA/NNA.

Leituras cruzadas dos padrões. (i) **IDA vs. OMA** (microcíticas): partilham MCV/MCH baixos; sem ferritina, a distinção depende de sutilezas (p. ex., RDW), justificando erros residuais. (ii) **NHA vs. NNA**: a fronteira *hipo/normocrômica* é sutil; trocas pontuais ocorrem sobretudo quando a hemoglobina corpuscular média oscila próximo ao limiar. (iii) **Leukemia vs. Leukemia+TP**: a classe composta tende a ser absorvida por *Leukemia* quando a plaquetopenia não domina o sinal; com suporte n=3, pequenas trocas inflacionam as porcentagens. (iv) **Macrocytic anemia**: suporte muito baixo (n=5) gera estimativas instáveis; ainda assim, o XGBoost manteve 80,0% de *recall*, enquanto o Random Forest caiu para 40,0%.

Implicação prática. Para cenários em que *todas* as classes importam (triagem/encaminhamento), o **Random Forest** oferece equilíbrio superior entre classes ($F1_{macro}$ e AP_{macro} maiores).

Quando o foco é **desempenho global** ponderado pela prevalência e alta pureza nas classes mais frequentes, o **XGBoost** é preferível. Em ambos os casos, recomenda-se calibração e ajuste de limiares *one-vs-rest* por classe, sobretudo para *Macrocytic* e *Leukemia+TP*.

3.4 Interpretação clínica dos erros

Os padrões de confusão observados são compatíveis com a fisiopatologia e com as limitações do hemograma isolado (ausência de marcadores como ferritina, B12 e folato). A seguir detalhamos os pares mais críticos e ações mitigadoras.

(i) **IDA vs. outras microcíticas (OMA).** *Racional:* ambas apresentam MCV/MCH baixos; sem ferritina, a separação depende de sutilezas como RDW (tende a ser maior na IDA) e contagem de eritrócitos (RBC)—frequentemente menor na IDA e normal/alta em traços talassêmicos, que compõem o grupo OMA. *Mitigação:* engenharia de índices derivados (p. ex., *Mentzer index* $=$ MCV$/$RBC, RDWI), calibração e ajuste de limiar *one-vs-rest* para IDA, e abordagem hierárquica "microcítica $\rightarrow$ IDA vs. OMA".

(ii) **NHA vs. NNA.** *Racional:* a fronteira entre hipocromia leve e normocromia é estreita; pequenas oscilações em MCH/MCHC e hemoglobina corpuscular podem inverter a classe. Isso se reflete em trocas pontuais, mais visíveis quando a amostra está próxima ao limiar. *Mitigação:* pós-processamento por margem (reclassificação conservadora quando $|p_{\text{NHA}} - p_{\text{NNA}}| < \delta$), normalização robusta por lote e *feature scaling* consistente; opcionalmente, um nó hierárquico "normocítica $\rightarrow$ hipo vs. normocrômica".

(iii) **Leukemia $\times$ Leukemia+TP.** *Racional:* quando a trombocitopenia ocorre no contexto leucêmico, a classe composta tende a ser absorvida por "Leukemia" se o sinal plaquetário não dominar o *split* (ou estiver em faixa limítrofe). Com apenas $n = 3$ no teste, pequenas trocas inflacionam as porcentagens. *Mitigação:* regra pós-treinamento baseada em plaquetas (se predito "Leukemia" e PLT $< \tau$, reclassificar como "Leukemia+TP", com τ escolhido em validação para maximizar F1 da classe), ou cabeça adicional *one-vs-rest* para trombocitopenia (modelo multi-rótulo leve).

(iv) **Macrocytic anemia.** *Racional:* suporte muito baixo ($n = 5$) torna as estimativas instáveis; casos com MCV próximo ao limite superior da normalidade podem migrar para NHA/NNA. Ainda assim, o XGBoost manteve *recall* de 80%, enquanto o Random Forest caiu para 40%. *Mitigação:* aumento dirigido da classe (p. ex., *SMOTE-ENN*/*Borderline-SMOTE* ou *class weights* sem *oversampling*), ajuste de limiar específico e validação estratificada com *repeated CV* para reduzir variância.

As confusões refletem sobreposição fisiológica esperada de índices eritrocitários/plaquetários. Medidas de **calibração**, **limiares por classe** e um **esquema hierárquico** (microcíticas $\rightarrow$ subtipos; normocíticas $\rightarrow$ hipo vs. normocrômica; neoplasias $\rightarrow$ com/sem trombocitopenia) tendem a reduzir erros clinicamente sensíveis sem sacrificar o desempenho global.

3.5 Implicações práticas e escolha do modelo

Cenários de uso e critério de escolha.

- **Alto volume, custo de erro proporcional à prevalência** (rotina de laboratório, *screening* amplo): priorize **XGBoost** — apresentou $F1_w$ e acurácia máximos e alta pureza em NHA/NNA/TP/IDA, mantendo baixo erro nas classes mais frequentes.
- **Equidade entre classes / foco nas raras** (triagem onde falsos-negativos de neoplasias são inaceitáveis): prefira **Random Forest**, que obteve $F1_{macro}$ superior e *recall* mais estável em *Leukemia* e *Leukemia+TP*, ainda que com queda pontual em *Macrocytic anemia*.
- **Recursos limitados ou necessidade de baseline interpretável**: **Regressão Logística** e **SVM** oferecem bom compromisso de simplicidade, devendo ser acompanhadas de limiares por classe para reduzir vieses em classes raras.

Política de decisão recomendada (produção).

1. **Calibração de probabilidades** (*Platt* ou isotônica) no conjunto de treino e validação, garantindo probabilidades bem-calibradas para uso clínico.
2. **Limiar por classe** τ_k (estratégia *one-vs-rest*): escolher τ_k para maximizar $F1_{macro}$ ou minimizar o risco esperado sob uma matriz de custos (e.g., penalizar mais falsos-negativos em *Leukemia*/*Leukemia+TP*).
3. **Abstenção assistida** (classificador seletivo): se a margem $|p_{(1)} - p_{(2)}| < \delta$ ou $p_{(1)} < \tau_{(1)}$, emitir *"necessita revisão humana"* e sugerir o *top-2* de classes. Isso reduz erros clínicos sem afetar muito o throughput.
4. **Estratégia hierárquica**: (i) microcítica $\rightarrow$ {IDA, OMA}; (ii) normocítica $\rightarrow$ {NHA, NNA}; (iii) neoplasias $\rightarrow$ {Leukemia, Leukemia+TP}. Essa decomposição reduz confusões estruturais observadas nas matrizes.

Custo computacional e manutenção.

- **XGBoost** (árvores *hist*) tende a ser compacto e rápido em inferência; bom para *batch* e *online*.
- **Random Forest** tem boa performance na maioria dos domínios e pouco sensível a hiperparâmetros, mas pode demandar mais memória em inferência quando há muitas árvores.
- **Monitoramento**: acompanhar, mensalmente, $F1_{macro}$, *recall* por classe e a taxa de abstenção; readequar limiares e recalibrar probabilidades diante de *drift*.

Interpretabilidade e governança. Aplicar **SHAP** global/local para evidenciar a contribuição de índices eritrocitários/plaquetários nas decisões (p. ex., MCV/MCH para microcíticas; PLT para Leukemia+TP), documentando regras de limiar e critérios de abstenção para auditoria clínica.

Recomendação. **Padrão**: XGBoost para maior desempenho global ($F1_w$ e acurácia). **Cenários sensíveis às raras**: Random Forest com calibração e limiares por classe para maximizar $F1_{macro}$ e *recall* em neoplasias. Quando a operação permitir, um **ensemble simples** (voto/empate resolvido por probabilidade calibrada) entre XGBoost e Random Forest pode oferecer robustez adicional sem custo expressivo.

3.6 Ameaças à validade e limitações

Validade interna (rigor da comparação).

- **Risco residual de vazamento de informação.** O uso de *pipelines* e *nested CV* mitigou vazamentos; ainda assim, qualquer pré-processamento executado fora do *pipeline* (p. ex., remoção de outliers globais, normalizações manuais, *feature selection* externa) pode introduzir otimismo. *Mitigação:* garantir que *todas* as transformações dependentes dos dados estejam dentro do *pipeline* e da CV.
- **SMOTE com k=1.** Escolha segura para classes raras, porém induz fronteiras muito locais e pode amplificar ruído quando n é muito baixo (e.g., n=3 em Leukemia+TP). *Mitigação:* testar *class weights* sem *oversampling*, variantes como *Borderline-SMOTE*/SMOTE-ENN, ou ajuste fino de k por classe.
- **Busca de hiperparâmetros limitada.** Grades compactas reduzirem custo computacional, mas podem subexplorar a hipótese do modelo. *Mitigação: search* híbrida (largura por *random search* + refino por *grid*) e *early stopping* onde aplicável.

Validade externa (generalização).

- **Mudança de domínio (*dataset shift*).** O *hold-out* de 20% provém da mesma base; desempenho pode se degradar em outros laboratórios, equipamentos, faixas etárias e prevalências. *Mitigação:* validação externa por sítio/tempo (*temporal split*, *site split*), recalibração e ajuste de limiares por prevalência local.
- **Dependência de faixas de referência.** Variações interlaboratoriais em MCV/MCH/MCHC e contagem plaquetária alteram a fronteira entre NHA/NNA, microcíticas e trombocitopenias. *Mitigação:* normalização por lote/equipamento e checagem de *drift* com monitoramento contínuo.

Validade de construto (o que foi medido).

- **Conjunto de atributos restrito ao hemograma.** A ausência de ferritina, B12 e folato limita a discriminação clínica de microcitoses e macrocitoses, favorecendo confusões plausíveis (IDA vs. OMA; Macrocytic vs. NHA/NNA). *Mitigação:* incluir variáveis bioquímicas quando disponíveis ou índices derivados (p. ex., Mentzer, RDWI) como *features*.
- **Qualidade/ruído de rótulos.** Bases públicas podem conter rótulos ruidosos ou heterogêneos quanto ao critério diagnóstico. *Mitigação:* auditoria amostral dos rótulos e, se possível, *noise-robust loss* ou *co-teaching*.

Validade estatística (precisão das estimativas).

- **Baixo suporte em classes raras.** Macrocytic (n=5) e Leukemia+TP (n=3) geram estimativas de *recall*/*precision* com alta variância; pequenas trocas alteram percentuais drasticamente. *Mitigação: repeated nested CV*, *bootstrap* para intervalos de confiança e reporte explícito de n por classe.
- **Métricas dependentes de limiar.** $\mathrm{F1}_{\mathrm{macro}}$/$\mathrm{F1}_{\mathrm{w}}$ dependem do *argmax* (ou de τ_k por classe); sem calibração, mudanças pequenas de limiar alteram decisões, sobretudo em raras. *Mitigação:* calibração (Platt/isotônica) e seleção de τ_k orientada a custo clínico.

Limitações operacionais.

- **Limpeza de dados.** A remoção de *NaN*/duplicatas pressupõe mecanismo de ausência ao acaso; se a ausência for informativa (p. ex., exames incompletos em casos graves), pode haver viés. *Mitigação:* análise de padrões de *missing* e imputação compatível com o *pipeline*.
- **Interpretabilidade consolidada.** Sem análise *post hoc* (SHAP), a explicabilidade por classe ainda é limitada para uso clínico amplo. *Mitigação:* adicionar explicações locais/globais e regras de decisão derivadas.

A combinação de **validação externa**, **calibração e limiares por classe**, **engenharia de índices hematológicos** e **aumento prudente** das classes raras deve reduzir o risco de superestimar o desempenho e melhorar a transportabilidade clínica dos modelos.

3.7 Comparação com pesquisas relacionadas

Diversos grupos de pesquisa aplicaram algoritmos de aprendizado de máquina para classificar ou prever tipos de anemia com base em dados clínicos, principalmente hemogramas (CBC) ou dados demográficos. A Tabela 2 resume estudos recentes (2022–2025) que utilizaram tecnologias similares àquelas empregadas neste trabalho.

Tabela 2 – Trabalhos que empregaram algoritmos de machine learning na classificação de anemias.

Autor/ano	**Conjunto de dados e algoritmos**	**Principais resultados**
(YIMER et al.) (2025)(YIMER et al., 2025)	Dados da pesquisa demográfica e de saúde da Etiópia para crianças menores de 5 anos; compararam regressão logística, árvore de decisão, *Random Forest*, SVM, Naive Bayes e KNN.	A Random Forest alcançou acurácia de 81,16 % (sensibilidade 83,07 %, especificidade 79,26 %), superando a regressão logística (54,79 %) e a SVM (59,94 %)(YIMER et al., 2025).
(VOHRA et al.) (2022)(VOHRA et al., 2022)	364 pacientes ambulatoriais com hemograma; avaliaram árvore de decisão, regressão logística, MLP, Naive Bayes, *Random Forest* e SVM, com seleção de características e SMOTE.	Para a classe *moderada*, a Random Forest obteve recall de 98,2 % e AUC de 99,8 %, superando outros algoritmos; MLP e regressão logística mostraram bom desempenho em classes leves(VOHRA et al., 2022).
(AWAAD et al.) (2025)(AWAAD et al., 2025)	Conjunto de 15 300 registros de CBC do Kaggle; testaram KNN, SVM, árvore de decisão, *Random Forest*, CNN, CNN+SVM, CNN+RF e XGBoost, enriquecendo as entradas com ontologias clínicas.	O modelo Onto-CNN+SVM alcançou F1 de 1,00 com oversampling e o XGBoost foi o classificador mais robusto; a ontologia melhorou em mais de 20 % o F1 para anemias raras como deficiência de folato e B12(AWAAD et al., 2025).
(SAMTANI) (2025)(SAMTANI, 2025)	Dados de hemograma de alta qualidade; comparou *Random Forest*, SVM e regressão logística com `GridSearchCV` para hiperparâmetros.	A Random Forest obteve 99,48 % de acurácia, enquanto SVM alcançou apenas 23,81 % e a regressão logística 57,23 %; empregou SHAP para selecionar as características mais relevantes(SAMTANI, 2025).
(DARSHAN et al.) (2025)(DARSHAN et al., 2025)	Atributos de hemograma para diferenciar anemia ferropriva e aplástica; compararam regressão logística, árvore de decisão, KNN, *Random Forest*, AdaBoost, CatBoost, LightGBM, XGBoost, modelos empilhados e ANN.	LightGBM e o modelo empilhado atingiram acurácia de 96 %; Random Forest e AdaBoost obtiveram 92%; CatBoost e XGBoost, 94 %; o trabalho também utilizou SHAP, LIME e outras técnicas de interpretabilidade(DARSHAN et al., 2025).
(GÓMEZ GÓMEZ et al.) (2025)(GÓMEZ GÓMEZ et al., 2025)	Conjunto de dados de anemia de 2022 (estudo de Sabatini); treinaram Análise Discriminante Linear, árvores de decisão e *Random Forest*.	A Random Forest alcançou acurácia de 99,82 % e permitiu sub-classificar anemias microcíticas, normocíticas e macrocíticas, superando abordagens binárias anteriores(GÓMEZ GÓMEZ et al., 2025).

Observa-se que métodos baseados em conjuntos de árvores (Random Forest, XGBoost, LightGBM) tendem a obter desempenho superior, muitas vezes com acurácia acima de 90 %, sobretudo após balanceamento de classes e otimização de hiperparâmetros. Esses resultados corroboram a escolha de algoritmos ensemble no presente estudo e reforçam a importância de técnicas de explicabilidade (SHAP, LIME) para a interpretação clínica dos modelos.

3.8 Discussão

Os resultados indicam, de forma consistente, a superioridade dos *ensembles* para dados tabulares (XGBoost e Random Forest) no cenário em estudo. Essa vantagem apareceu tanto nas métricas de ponto ($F1_w$, acurácia) quanto nas métricas baseadas em probabilidade, com separabilidade próxima do teto nas curvas ROC e melhor discriminação entre os dois líderes quando analisada a PR-AUC (*macro*) — onde a Random Forest manteve vantagem em classes menos prevalentes. Esses achados são coerentes com a literatura em dados clínicos tabulares e com a hipótese de fronteiras não lineares entre subtipos de anemia capturadas por combinações de índices eritrocitários e plaquetários.

Do ponto de vista clínico, a análise por classe mostrou que as confusões residuais seguem a fisiopatologia esperada: (i) *IDA* versus *OMA* pela sobreposição em MCV/MCH baixos; (ii) *NHA* versus *NNA* pela fronteira cromática sutil; (iii) *Leukemia* absorvendo *Leukemia+TP* quando o sinal plaquetário não domina; e (iv) maior instabilidade em *Macrocytic* pelo baixo suporte. Ainda assim, os líderes sustentaram altos *recalls* nas classes mais frequentes e desempenhos competitivos nas raras — com a Random Forest mais estável em *Leukemia/Leukemia+TP* e o XGBoost com *recall* superior em *Macrocytic* no *hold-out*.

Em termos operacionais, a PR-AUC (*macro*) serviu como *tie-break* importante quando a ROC estava saturada, reforçando a adoção de limiares específicos por classe (*one-vs-rest*) e calibração de probabilidades na implantação — medidas que reduzem o custo de erro em classes raras sem sacrificar o desempenho global. Na prática, o ajuste de τ_k orientado a $F1_{\text{macro}}$ (ou a uma matriz de custos clínicos) e políticas de abstenção (*selective classification*) quando a margem é pequena são estratégias pragmáticas para aumentar segurança clínica.

O desenho metodológico adotado (pipelines completos, SMOTE restrito ao treino, e *nested CV* com estratificação) mitigou vieses de seleção e vazamento de informação, elevando a confiança nas comparações. Ainda assim, permanecem ameaças clássicas: (i) grades de hiperparâmetros limitadas; (ii) fronteiras locais induzidas por SMOTE com k=1 em classes *muito* raras; e (iii) *dataset shift* em ambientes externos (equipamentos, faixas etárias, prevalências). Essas fragilidades foram explicitadas e guiam o plano de validação externa e robustificação (p.,ex., *random search* + *grid*, variações de SMOTE/ponderação de classe, e normalização por lote).

Por fim, a escolha entre os dois líderes depende do *use-case*. Se o objetivo é maximizar o acerto global ponderado pela prevalência em fluxos de alto volume, o XGBoost é adequado.

Se a prioridade é equidade entre classes e maior *recall* nas raras (especialmente neoplasias), a Random Forest se mostrou mais favorável. No limite, um *ensemble* simples entre ambos, com probabilidades calibradas, tende a agregar robustez sem custo computacional proibitivo.

4 Conclusão

Este estudo comparou, de forma sistemática e reprodutível, diferentes algoritmos de classificação multiclasse para identificar tipos de anemia a partir de parâmetros hematológicos (CBC). O desenho metodológico — *pipelines* completos com padronização quando necessário, SMOTE aplicado **apenas** no treino e validação cruzada aninhada — reduziu riscos de vazamento de informação e otimismo na seleção de modelos, fornecendo estimativas mais fidedignas de desempenho.

Os resultados apontam a superioridade de *ensembles* tabulares. O **XGBoost** obteve o maior F1 ponderado e acurácia no teste ($F1_w = 0{,}978$; $Acc = 0{,}978$), sendo a melhor escolha quando o objetivo é maximizar acerto global ponderado pela prevalência. Já a **Random Forest** apresentou melhor equilíbrio entre classes ($F1_{macro} = 0{,}896$; AP_{macro} superior), favorecendo cenários em que falsos-negativos em classes raras são críticos (p. ex., neoplasias). As curvas PR foram mais informativas do que a ROC — frequentemente saturada — para discriminar os modelos de topo, reforçando a necessidade de avaliar métricas sensíveis ao desbalanceamento.

As confusões observadas (IDA vs. OMA; NHA vs. NNA; *Leukemia* vs. *Leukemia+TP*; *Macrocytic* com baixo suporte) são compatíveis com a fisiopatologia e com as limitações do hemograma isolado. Para implantação, recomendamos: (i) **calibração** de probabilidades; (ii) **limiares por classe** (*one-vs-rest*) orientados a $F1_{macro}$ ou custo clínico; e (iii) **abstenção assistida** em casos de baixa margem, estratégias que reduzem erros clinicamente relevantes sem sacrificar o throughput. Em ambientes de produção, um **ensemble simples** (voto/ desempate por probabilidade calibrada) entre XGBoost e Random Forest tende a agregar robustez com baixo custo adicional.

Como limitações, destacam-se o baixo suporte de algumas classes (p. ex., *Macrocytic* e *Leukemia+TP*), a ausência de marcadores bioquímicos (ferritina, B12, folato) e a possibilidade de *dataset shift* entre laboratórios. Trabalhos futuros devem priorizar: validação externa temporal/multissítio com recalibração e retuning de limiares; inclusão de índices derivados (Mentzer, RDWI) e marcadores bioquímicos; abordagens hierárquicas e/ou multi-rótulo para neoplasias com trombocitopenia; e explicabilidade sistemática (SHAP) com monitoramento de *drift* e auditoria de equidade entre classes.

Em síntese, os achados confirmam o potencial do aprendizado de máquina para apoiar a triagem e o encaminhamento de anemias a partir de CBC. Ressaltamos que o sistema proposto **não substitui** o julgamento clínico: seu uso recomendado é como ferramenta de apoio à decisão, com revisão humana nos casos de incerteza e em classes raras.

Referências

ABOELNAGA, Ehab. **Anemia Types Classification**. [S.l.: s.n.], 2023. Kaggle dataset. Licença: Apache-2.0. Acesso em: 21-08-2025. Disponível em: https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification.

ASARE, Justice Williams; APPIAHENE, Peter; DONKOH, Emmanuel Timmy. Detection of anaemia using medical images: A comparative study of machine learning algorithms – A systematic literature review. **Informatics in Medicine Unlocked**, v. 40, p. 101283, 2023. ISSN 2352-9148. DOI: https://doi.org/10.1016/j.imu.2023.101283. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352914823001272.

AWAAD, Amira S. et al. Exploring CBC Data for Anemia Diagnosis: A Machine Learning and Ontology Perspective. **BioMedInformatics**, v. 5, n. 3, p. 35, 2025. Ontology-enhanced models such as Onto-CNN+SVM achieved F1-score of 1.00; XGBoost consistently the most robust. DOI: 10.3390/biomedinformatics5030035.

DARSHAN, B. S. Dhruva et al. Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable Artificial Intelligence utilizing blood attributes. **Scientific Reports**, v. 15, p. 505, 2025. LightGBM and stacked models achieved 96% accuracy; Random Forest and AdaBoost 92%. DOI: 10.1038/s41598-024-84120-w.

ELMALEEH, Mohammed AA. The Identification and Categorization of Anemia Through Artificial Neural Networks: A Comparative Analysis of Three Models. **arXiv preprint arXiv:2404.04690**, 2024.

ELSABAGH, Ahmed Adel et al. Artificial intelligence in sickle disease. **Blood Reviews**, v. 61, p. 101102, 2023. ISSN 0268-960X. DOI: https://doi.org/10.1016/j.blre.2023.101102. Disponível em: https://www.sciencedirect.com/science/article/pii/S0268960X23000632.

GARCIA, Ana Cristina. Ética e inteligencia artificial. **Computação Brasil**, n. 43, p. 14–22, 2020.

GÓMEZ GÓMEZ, Jorge et al. Anemia Classification System Using Machine Learning. **Informatics**, v. 12, n. 1, p. 19, 2025. Random Forest accuracy 99.82% for sub-classifying microcytic, normocytic and macrocytic anemia. DOI: 10.3390/informatics12010019.

KARAGÜL YILDIZ, Tuba; YURTAY, Nilüfer; ÖNEÇ, Birgül. Classifying anemia types using artificial learning methods. **Engineering Science and Technology, an International Journal**, v. 24, n. 1, p. 50–70, 2021. ISSN 2215-0986. DOI: https://doi.org/10.1016/j.jestch.2020.12.003. Disponível em: https://www.sciencedirect.com/science/article/pii/S2215098620342646.

KAZEMINIA, Salome et al. Anomaly-aware multiple instance learning for rare anemia disorder classification. In: SPRINGER. INTERNATIONAL Conference on Medical Image Computing and Computer-Assisted Intervention. [S.l.: s.n.], 2022. P. 341–350.

KOVAČEVIĆ, Anđela et al. Application of Artificial Intelligence in Diagnosis and Classification of Anemia. In: 2022 11th Mediterranean Conference on Embedded Computing (MECO). [S.l.: s.n.], 2022. P. 1–4. DOI: 10.1109/MECO55406.2022.9797180.

MOREIRA, Paulo Sergio da Conceição; SALERNO, Byanca Neumann; TSUNODA, Denise Fukumi et al. Internet das coisas e aprendizado de máquina na área da saúde: uma análise bibliométrica da produção científica de 2009 a 2019. **Revista Eletronica De Comunicação, Informação and Inovação Em Saude**, Fundação Oswaldo Cruz. Instituto de Comunicação e Informação Científica e . . ., 2020.

NANGO, J. et al. A New Strategy for the Morphological and Colorimetric Recognition of Erythrocytes for the Diagnosis of Forms of Anemia based on Microscopic Color Images of Blood Smears. **arXiv preprint arXiv:2302.08214**, 2023.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Anemia — Biblioteca Virtual em Saúde MS**. [S.l.: s.n.], 2025. Acesso em: 10 fev. 2025. Disponível em: https://bvsms.saude.gov.br/anemia/.

SAMTANI, Prerna. Anemia Detection Using CBC Data: A Comparative Study. **International Journal for Research in Applied Science and Engineering Technology**, v. 13, n. 7, 2025. Random Forest accuracy 99.48%; SVM 23.81%; logistic regression 57.23%. DOI: 10.22214/ijraset.2025.73251.

SCHOUERI JÚNIOR, Roberto et al. Anemia megaloblástica em idosos. **Rev. paul. med**, p. 148–52, 1990.

SILVA, Geovane Carlos da et al. ANEMIA FERROPRIVA. **ANAIS DO FÓRUM DE INICIAÇÃO CIENTÍFICA DO UNIFUNEC**, v. 7, n. 7, 2016.

UVALIYEVA, Indira; BELGINOVA, Saule; BOROZENETS, David. Analysis of the Machine Learning Methods Effectiveness for Morphological Classification of Anemia. In: 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). [S.l.: s.n.], 2023. P. 1–5. DOI: 10.1109/ISMSIT58785.2023.10304939.

VOHRA, Rajan et al. Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting. **PLoS One**, v. 17, n. 7, e0269685, 2022. Random Forest recall of 98.2% for moderate anemia. DOI: 10.1371/journal.pone.0269685. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC9258850/.

YIMER, Ali et al. Optimizing machine learning models for predicting anemia among under-five children in Ethiopia: insights from Ethiopian demographic and health survey data. **BMC Pediatrics**, v. 25, n. 311, 2025. Random Forest achieved 81.16% accuracy while logistic regression and SVM performed worse. DOI: 10.1186/s12887-025-05659-9. Disponível em: https://pubmed.ncbi.nlm.nih.gov/.