Classificação do Status Glicêmico a partir de Hemogramas: Avaliação de Estratégias de Aprendizado de Máquina em Dados Laboratoriais Reais do Brasil

Gabriel Eduardo Martini[1], **Mariana Recamonde-Mendoza**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

{gemartini,mrmendoza}@inf.ufrgs.br

Abstract. *Early diagnosis of diabetes is essential, yet glycated hemoglobin (A1c) testing is not always accessible. We investigated whether glycemic status (normal, pre-diabetic, and diabetic) can be classified exclusively from complete blood count (CBC) data using about 170000 real laboratory records. Binary, multiclass, binary decomposition, and ensemble strategies were evaluated. Neural network models achieved the best performance (F2 = 0.793 in the binary task and 0.551 in the multiclass setting), with no gains observed from the ensemble approach. Error analysis revealed higher misclassification rates near diagnostic A1c thresholds, indicating greater difficulty in transitional glycemic states. Age, leukocytes, and RDW were the most relevant predictors. These results suggest that CBC data contain signals associated with glycemic status, although with limitations for screening applications.*

Resumo. *O diagnóstico precoce do diabetes é essencial, porém a dosagem de hemoglobina glicada (A1c) nem sempre está acessível. Investigamos se o status glicêmico (normal, pré-diabético e diabético) pode ser classificado exclusivamente a partir de dados do hemograma completo utilizando cerca de 170 mil registros laboratoriais reais. Foram avaliadas estratégias de classificação binária, multiclasse, decomposição binária e ensemble. Redes neurais apresentaram o melhor desempenho (F2 = 0,793 na tarefa binária e 0,551 na multiclasse), e não foram observados ganhos no uso de ensemble. A análise de erros revelou maior taxa de classificações incorretas próxima aos limiares diagnósticos de A1c, indicando maior dificuldade em estados glicêmicos de transição. Idade, leucócitos e RDW foram os preditores mais relevantes. Os resultados indicam que dados do hemograma contêm sinais associados ao status glicêmico, embora com limitações para triagem.*

1. Introdução

O Diabetes Mellitus é uma doença metabólica caracterizada pelos altos níveis de glicose no sangue. Em recente relatório da revista *The Lancet*, estimativas apontam que em 2021 mais de 500 milhões de pessoas conviviam com diabetes e que em 2050 mais de 1,31 bilhão viverão com a doença [The Lancet 2023]. Por ser uma doença de sintomas silenciosos, muitos indivíduos permanecem sem diagnóstico por anos. O diagnóstico tardio está associado ao aumento do risco de complicações microvasculares e macrovasculares, como retinopatia, nefropatia, neuropatia e doenças cardiovasculares, além de maior impacto socioeconômico e sobrecarga aos sistemas de saúde [Lu et al. 2023].

Nesse contexto, a identificação precoce de alterações glicêmicas torna-se fundamental. O pré-diabetes representa um estado intermediário caracterizado por níveis glicêmicos acima da normalidade, porém abaixo dos critérios diagnósticos para diabetes. Evidências indicam que essa condição pode ser revertida por meio de intervenções no estilo de vida ou terapias farmacológicas, reduzindo significativamente o risco de progressão para diabetes tipo 2 [Galaviz et al. 2022]. Assim, a detecção correta do pré-diabetes representa uma oportunidade importante para prevenção e mitigação de complicações futuras.

A dosagem da hemoglobina glicada (A1c) é recomendada pela *American Diabetes Association* (ADA) como exame padrão para o diagnóstico de diabetes, refletindo a média glicêmica dos últimos 90 dias [World Health Organization 2021]. Entretanto, limitações de acesso, custo e infraestrutura em diversos contextos de saúde dificultam sua ampla utilização, contribuindo para subdiagnóstico e diagnóstico tardio [NCD Risk Factor Collaboration (NCD-RisC) 2023]. Nesse cenário, exames laboratoriais de rotina amplamente disponíveis e de baixo custo, como o hemograma, emergem como potenciais fontes de informação para triagem populacional.

O hemograma é um dos exames laboratoriais mais solicitados mundialmente [Le et al. 2022]. Alterações metabólicas e processos inflamatórios crônicos associados ao diabetes podem se refletir em parâmetros hematológicos, incluindo elevação na contagem de leucócitos e modificações em outras subpopulações celulares [Bambo et al. 2024]. Embora tais marcadores individualmente apresentem poder discriminativo limitado, seu conjunto pode conter padrões associados ao status glicêmico. Estudos recentes demonstram a viabilidade da aplicação de algoritmos de aprendizado de máquina (AM) na análise de dados hematológicos, evidenciando associações relevantes entre parâmetros do hemograma e o status glicêmico [Mansoori et al. 2023, Cardozo et al. 2022]. De forma mais ampla, abordagens baseadas em inteligência artificial têm sido empregadas no desenvolvimento de ferramentas de apoio ao diagnóstico de diabetes.

Entretanto, a maioria das abordagens reportadas concentra-se em classificação binária (normal *vs.* diabético) ou utiliza combinações de variáveis bioquímicas e demográficas, limitando a análise isolada do potencial preditivo do hemograma. Além disso, o desafio da classificação do pré-diabetes como classe intermediária, caracterizada por sobreposição fenotípica com indivíduos normais e diabéticos, pode comprometer o desempenho de modelos multiclasse convencionais. Aspectos relacionados à distribuição de erros entre subgrupos demográficos e à interpretabilidade dos modelos também permanecem pouco explorados nesse contexto.

Diante dessas lacunas, este trabalho investiga o uso de AM para classificar o status glicêmico (normal, pré-diabético e diabético) a partir de parâmetros do hemograma completo. O estudo utiliza uma base de 169.999 registros laboratoriais reais e compara diferentes estratégias de modelagem, incluindo classificação binária, multiclasse e decomposição do problema em classificadores binários especializados combinados por *ensemble*. As principais contribuições são: (i) a avaliação sistemática do potencial preditivo do hemograma completo em uma das maiores bases de dados brasileiras já usadas neste contexto; (ii) a comparação estruturada entre diferentes estratégias de modelagem, evidenciando os limites da classificação direta do pré-diabetes; (iii) a análise de padrões de erro e separabilidade entre classes, destacando a dificuldade intrínseca da classe pré-diabética; e (iv) uma análise de interpretabilidade baseada em importância por permutação

e SHAP, identificando parâmetros hematológicos relevantes para triagem glicêmica.

2. Trabalhos Relacionados

A aplicação de AM à predição do diabetes é um tema amplamente abordado por pesquisas no meio científico e acadêmico. A hiperglicemia crônica pode induzir alterações mensuráveis em parâmetros hematológicos [Bambo et al. 2024], motivando a investigação de modelos preditivos treinados com esses dados. [Mansoori et al. 2023] utilizaram dados hematológicos e bioquímicos de mais de 9 mil indivíduos iranianos para predição do diabetes, obtendo o melhor desempenho com *Random Forest* (Tabela 1). [Cardozo et al. 2022] utilizaram dados de mais de 62 mil pacientes brasileiros com exames laboratoriais de rotina para identificar diabéticos não diagnosticados, obtendo melhor desempenho com Redes Neurais Artificiais. Já [Tahir et al. 2024] analisaram mais de 200 mil registros laboratoriais combinando hemogramas e Proteína C Reativa (PCR), confirmando alterações hematológicas induzidas pela hiperglicemia crônica.

Outros estudos utilizaram dados clínico-laboratoriais diversos. Por exemplo, [Alhassan et al. 2021] investigaram elevações de A1c em 18 mil prontuários eletrônicos, enquanto [Al-hussein et al. 2025] analisaram 3.000 pacientes, utilizando 18 parâmetros clínicos, laboratoriais e demográficos. Por fim, [Cheng et al. 2023] identificaram IMC e atividade física como fatores de risco associados ao descontrole glicêmico utilizando AM.

A análise da literatura (Tabela 1) revela três lacunas principais. Primeiro, a maioria dos estudos concentra-se em classificação binária, negligenciando a complexidade da classe intermediária de pré-diabetes. Segundo, muitos trabalhos utilizam conjuntos de dados relativamente pequenos ou combinam múltiplas fontes de informação clínica, dificultando avaliar o potencial isolado do hemograma. Terceiro, poucos estudos comparam sistematicamente diferentes estratégias de modelagem para lidar com a sobreposição entre classes glicêmicas. Este trabalho busca contribuir nesses três aspectos.

Tabela 1. Resumo comparativo dos principais estudos correlatos.

Autor	Dados	Atributos (n)	Amostra	Algoritmos	Tarefa	Melhor Res.
Alhassan et al. (2021)	Clín. + Lab.	EHR + dados longit. (n.d.)	18.844	MLR, RF, SVM, LR, MLP	Bin. (A1c $\geq$5,7%)	MLP: AUC=0,83
Al-hussein et al. (2025)	Clín. + Lab. + Demo.	Clín., lab., demo. (18$\rightarrow$4)	3.000	LR, DT, RF, SVM, NB, ANN, KNN, AdaB., GB, XGB + híb.	Bin. (A1c $\geq$6,5%)	RF+LR: Acc=93%, AUC=0,88
Cardozo et al. (2022)	Hemat. + Lab.	HT, MCH, RDW, MPV, PLT, dif. leuc. (15)	62.496	KNN, SVM, NB, RF, ANN	Multi. (N/PD/DM) + Reg.	ANN: Sens.=78,1%, Prec.=78,7%
Cheng et al. (2023)	Clín. + Lab. + Demo.	IMC, HRV, lipídios, A1c (18)	647	SVM, Boost., DT, NN, KNN, RF	Bin. (A1c $\geq$6,5%)	RF: Acc=84%, AUC=0,95
Mansoori et al. (2023)	Hemat. + Lab.	HGB, HCT, PLT, RDW, WBC, etc. (16)	9.000	LR, DT, BF	Bin. (DM vs não-DM)	BF: Acc=83,3%, AUC=0,91
Tahir et al. (2024)	Hemat. + Inflam.	WBC, NLR, PLR, CRP, RBC, MCV, MCH (7)	208.137	LDA (4 modelos)	Multi. (N/Lim./Hiper.)	LDA: Acc=89,5%, AUC=0,87

3. Metodologia

3.1. Coleta e Preparação de Dados

O conjunto de dados utilizado nesta pesquisa foi obtido em parceria com um laboratório de análises clínicas privado localizado no sul do Brasil. Os dados consistem em registros laboratoriais secundários, previamente anonimizados de forma irreversível antes do

acesso pelos pesquisadores. O projeto foi submetido e aprovado pelo Comitê de Ética em Pesquisa institucional (CAAE 92847125.8.0000.5347).

A base original continha 1.190.541 registros laboratoriais correspondentes a resultados de hemograma completo e dosagem de hemoglobina glicada (A1c). Inicialmente, foram removidas todas as instâncias em que pelo menos um desses exames estava ausente, o que ocorre quando hemograma e A1c não são realizados na mesma coleta, reduzindo o número de registros para 230.573. Em seguida, foram excluídas instâncias que apresentavam campos vazios em algum dos parâmetros hematológicos, geralmente decorrentes de falhas de leitura automática ou impossibilidade de dosagem, resultando em 170.021 registros. Adicionalmente, foram removidas 22 instâncias contendo valores extremos fisiologicamente implausíveis, identificados por inspeção visual das distribuições individuais dos parâmetros hematológicos. A base recebida já estava deduplicada por paciente, garantindo que exames de um mesmo indivíduo não se distribuíssem entre treino e teste. Após o pré-processamento, o conjunto final passou a conter 169.999 instâncias.

A base final inclui 23 atributos preditivos. Além das variáveis demográficas idade e sexo, foram considerados 21 parâmetros do hemograma completo, organizados em três grupos. A série branca compreende 13 atributos, incluindo leucócitos totais, neutrófilos segmentados, linfócitos, monócitos, eosinófilos, basófilos e suas respectivas proporções. A série vermelha inclui 6 atributos: eritrócitos (RBC), hemoglobina (HGB), hematócrito (HCT), hemoglobina corpuscular média (HCM), concentração de hemoglobina corpuscular média (CHCM) e amplitude de distribuição dos eritrócitos (RDW). Por fim, a série plaquetária abrange a contagem de plaquetas e volume plaquetário médio (VPM).

A variável-alvo é a Hemoglobina Glicada (A1c), que sintetiza a média glicêmica do indivíduo nos últimos três meses. Com base nos critérios da Organização Mundial da Saúde [World Health Organization 2021],os indivíduos foram classificados como normais (A1c $<$ 5,7%), pré-diabéticos (5,7% $\leq$ A1c $<$ 6,5%) ou diabéticos (A1c $\geq$ 6,5%). A distribuição entre classes resultante apresenta desbalanceamento moderado: Normal (N) = 85.170 registros (50,1%), Pré-diabetes (PD) = 49.157 registros (28,9%) e Diabetes (D) = 35.672 registros (21,0%).

3.2. Pré-processamento e Balanceamento de Classes

Após a preparação inicial descrita na Seção 3.1, todo o pré-processamento utilizado na modelagem foi implementado por meio de *pipelines* do *scikit-learn* [Pedregosa et al. 2011], garantindo reprodutibilidade, ausência de vazamento de dados e aplicação consistente das transformações. Os atributos numéricos foram escalonados por *StandardScaler*, *MinMaxScaler* ou *QuantileTransformer*, a depender da categoria experimental (detalhada na Seção 3.3). A variável categórica *Sexo* foi representada por *one-hot encoding* binário, descartando-se uma categoria para resultar em um atributo binário.

Para tratamento do desbalanceamento entre classes, foram avaliadas técnicas de sobreamostragem (SMOTE e *RandomOverSampler*/ROS), subamostragem (*TomekLinks* e *RandomUnderSampler*/RUS), abordagens híbridas (SMOTEENN e SMOTETomek) e penalização via *class_weight* [Chawla et al. 2002, Lemaître et al. 2017]. Todas as técnicas de reamostragem foram aplicadas exclusivamente nos dados usados para treinamento, evitando vazamento de informações para validação ou teste.

3.3. Estratégias de Classificação, Algoritmos e Categorias Experimentais

O problema foi investigado por meio de quatro estratégias de classificação (*i.e.*, tarefas): **(T1)** classificação binária entre indivíduos normais contra pré-diabéticos e diabéticos; **(T2)** classificação multiclasse com as três classes; **(T3)** decomposição do problema multiclasse em classificadores binários especializados, incluindo três fronteiras *one-vs-one* (N *vs.* D, N *vs.* PD, PD *vs.* D) e um classificador adicional (N+PD *vs.* D) que captura a separação entre indivíduos sem diabetes estabelecida e diabéticos; e **(T4)** *ensemble* composto por combinação dos melhores modelos binários por votação (*hard* ou *soft voting*).

Foram avaliados 12 algoritmos de classificação, incluindo modelos lineares (Regressão Logística, *LinearSVC*), modelos baseados em árvores (*Decision Tree*, *Random Forest*, *Bagging*), métodos de *boosting* (*Gradient Boosting*, *AdaBoost*, *HistGradientBoosting*, *XGBoost*, *LightGBM*, *CatBoost*) e Redes Neurais Artificiais (MLP). Cada configuração de modelo foi encapsulada em um *pipeline* composto por três etapas: escalonamento, reamostragem e classificação. Para permitir uma análise sistemática do impacto de cada etapa do *pipeline*, as configurações foram agrupadas em três categorias experimentais, cada uma isolando um fator de influência sobre o desempenho (Tabela 2): **(i)** transformação numérica, **(ii)** escolha do algoritmo, e **(iii)** método de balanceamento.

Tabela 2. Configurações de *pipeline* por categoria experimental. N representa o número de configurações executadas por estratégia de classificação.

Categoria	Escalonamento	Método de Balanceamento	Famílias de Algoritmos	N
C1: Escalonamento	*MinMaxScaler*, *StandardScaler* ou sequência (*MinMax*→*Standard*)	SMOTE ($k = 5$)	Decision Tree, Bagging, Logistic Regression, MLP, Gradient Boosting, AdaBoost	11
C2: Algoritmos	*QuantileTransformer* → *StandardScaler*[*]	SMOTE ($k = 5$)	Decision Tree, Random Forest, HistGB, XGBoost, LightGBM, CatBoost, Logistic Regression, LinearSVC, MLP	9
C3: Balanceamento	Diversos[†]	SMOTEENN, SMOTETomek, TomekLinks, ROS, RUS, *class_weight*	MLP, Logistic Regression, LinearSVC, Decision Tree, Bagging, Random Forest, XGBoost, CatBoost	8
Total executado por tarefa				**28**

[*] *HistGradientBoosting* e *LightGBM* dispensam escalonamento numérico devido ao *histogram binning* nativo.
[†] Cada algoritmo utiliza o escalonamento adequado ao seu tipo; modelos baseados em histogramas não recebem normalização explícita. ROS = *RandomOverSampler*; RUS = *RandomUnderSampler*.

A categoria **C1: Escalonamento** investiga exclusivamente o efeito das transformações numéricas nos atributos. Nessa categoria, o método de balanceamento é mantido constante (SMOTE com $k = 5$) e variam-se os escalonadores: *MinMaxScaler*, *StandardScaler* ou a aplicação sequencial *MinMax*→*Standard*. Foram testadas 11 combinações, distribuídas entre os algoritmos incluídos. A categoria **C2: Algoritmos** fornece uma comparação entre diferentes famílias de modelos sob condições controladas. Todos os algoritmos recebem exatamente o mesmo pré-processamento (*QuantileTransformer*→*StandardScaler*) e o mesmo método de balanceamento (SMOTE). Essa categoria funciona como *baseline*, isolando o impacto da escolha do algoritmo. Por fim, a categoria **C3: Balanceamento** examina exclusivamente o efeito de diferentes estratégias de tratamento de desbalanceamento. Visando limitar os custos computacionais, um total de oito combinações foi definido de forma direcionada: cada método de balanceamento foi emparelhado com o algoritmo mais relevante segundo recomendações da literatura ou características estruturais. No total, para as tarefas T1, T2 e T3, executamos 28 configurações de *pipeline*, permitindo uma comparação ampla e sistemática.

3.4. Avaliação de Desempenho e Explicabilidade de Modelos

A base pré-processada foi dividida em 80% para treino e 20% para teste. A otimização de hiperparâmetros foi realizada exclusivamente nos dados de treino por meio de validação cruzada aninhada (*nested CV*), utilizando 10 *folds* externos para estimativa de variância e 5 *folds* internos com *HalvingRandomSearchCV* para seleção dos hiperparâmetros. Todas as divisões foram estratificadas e mantidas fixas entre algoritmos e estratégias de modelagem, garantindo comparabilidade. O desempenho final foi reportado no conjunto de teste, e a estabilidade entre os dez *folds* externos foi usada para verificar ausência de sobreajuste.

As principais métricas de avaliação foram F2-Score (β=2), Recall e AUC-PR, complementadas por Acurácia, Precisão e AUC-ROC. O F2-Score enfatiza *recall*, refletindo o maior custo clínico de falsos negativos. Na tarefa multiclasse, utilizou-se a média macro. Na T4 (ensemble), seis especialistas baseados em modelos binários, dois por classe alvo, foram selecionados a partir dos melhores desempenhos da T3, seguindo um ranking multicritério (F2, AUC-PR, Acurácia Balanceada e Recall). Por fim, a interpretabilidade dos modelos foi examinada por importância por permutação e valores SHAP (*SHapley Additive exPlanations*), permitindo identificar a contribuição individual dos parâmetros hematológicos na classificação do status glicêmico.

4. Resultados

Esta seção resume os principais achados do estudo, iniciando pela comparação entre as estratégias binária e multiclasse e enfatizando os desafios associados ao pré-diabetes, que motivam o uso de classificadores binários especializados. Na sequência, apresentamos os resultados dos modelos da T3, o desempenho do *ensemble* da T4, a análise dos atributos mais relevantes e os padrões de erro observados. Resultados completos, incluindo métricas adicionais e configurações complementares, arquitetura dos modelos e melhores hiperparâmetros encontram-se no repositório[1].

4.1. Desempenho Geral: Classificação Binária (T1) e Multiclasse (T2)

Nas primeiras análises, investigamos o desempenho dos modelos ao detectar indivíduos com alteração glicêmica, bem como diferenciar entre diabetes e pré-diabetes, a partir do hemograma. A Tabela 3 consolida o melhor desempenho por estratégia, evidenciando a superioridade recorrente dos modelos MLP. Os valores obtidos em treino, validação e teste são semelhantes, indicando boa estabilidade preditiva e ausência de *overfitting*.

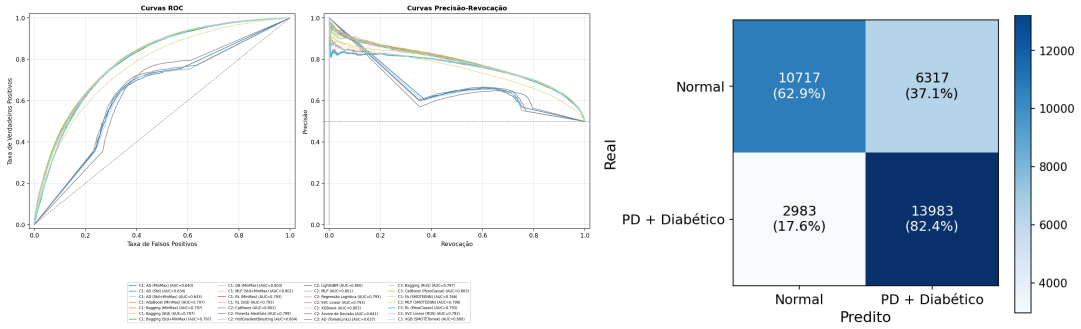
Tabela 3. Desempenho consolidado dos melhores modelos por tarefa, com métricas extraídas dos dados de teste e da validação cruzada aninhada.

Tarefa	Problema	Modelo	F2	Recall	AUC-PR	AUC-ROC	F2 Treino CV ($\pm$dp)	F2 Valid CV ($\pm$dp)
T1	N vs PD+D	MLP (Std+MinMax)	**0,793**	0,824	**0,780**	0,801	$0{,}763 \pm 0{,}016$	$0{,}763 \pm 0{,}018$
T2	N vs PD vs D	MLP (Std+MinMax)	0,551	0,556	0,571	0,756	$0{,}548 \pm 0{,}003$	$0{,}545 \pm 0{,}005$
T3	N vs D	MLP (SMOTEENN)	0,770	**0,877**	0,686	**0,852**	$0{,}782 \pm 0{,}003$	$0{,}769 \pm 0{,}007$
T3	N vs PD	MLP (SMOTEENN)	0,750	0,847	0,616	0,764	$0{,}753 \pm 0{,}003$	$0{,}747 \pm 0{,}009$
T3	PD vs D	MLP (SMOTEENN)	0,697	0,762	0,592	0,681	$0{,}700 \pm 0{,}007$	$0{,}688 \pm 0{,}011$
T3	N+PD vs D	MLP (SMOTEENN)	0,650	0,851	0,456	0,780	$0{,}666 \pm 0{,}002$	$0{,}647 \pm 0{,}005$
T4*	N vs PD vs D	*Hard Voting*	0,499	0,504	—	—	—	—
T4*	N vs PD vs D	*Soft Voting*	0,497	0,510	0,562	0,758	—	—

**Hard Voting* não produz probabilidades contínuas, assim, AUC-PR e AUC-ROC não são calculáveis.

[1] https://github.com/MartiniGabriel/SBCAS2026_MaterialComplementar/

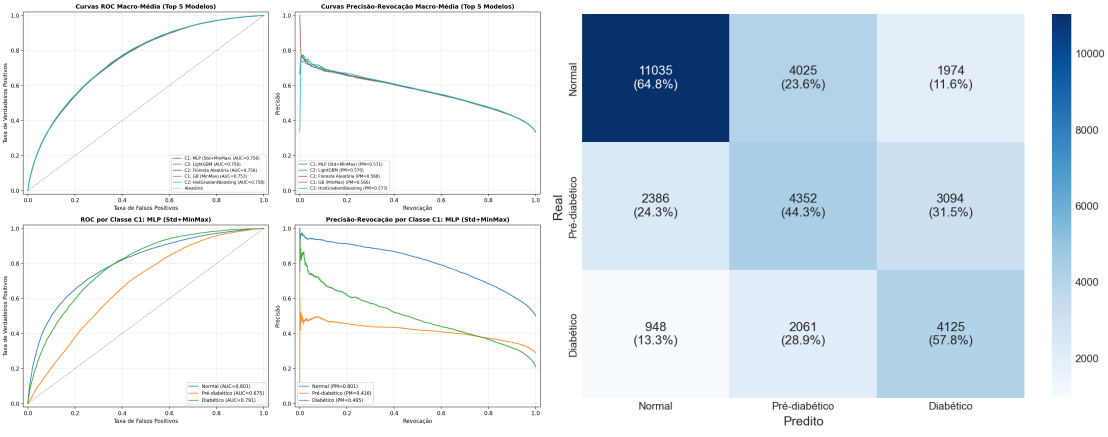
O modelo para T1 (binário) apresentou desempenho elevado (F2 = 0,793; AUC-ROC = 0,801), indicando que o hemograma contém informações suficientes para distinguir, com bom poder preditivo, indivíduos normais daqueles com alteração glicêmica inicial ou estabelecida (pré-diabetes ou diabetes). Esse comportamento se manteve consistente entre diferentes modelos e categorias de *pipeline* avaliadas para T1, conforme ilustrado na Figura 1(a). A matriz de confusão do melhor modelo, baseado em MLP, nos dados de teste (Figura 1(b)) mostra que o modelo consegue identificar 82,4% dos indivíduos com alteração glicêmica apenas a partir dos parâmetros hematológicos.



(a) Curvas ROC e PR para T1 (N x PD+D). (b) Matriz de confusão para o melhor modelo.

Figura 1. Desempenho de modelos binários (T1) nos dados de teste.

O melhor modelo para T2 (multiclasse) apresentou desempenho substancialmente inferior ao observado em T1, com F2 = 0,551 e Recall = 0,556. As curvas ROC e Precision-Recall (PR) (Figura 2) reforçam esse achado, evidenciando que o desempenho ternário é consistentemente menor do que o binário. A principal queda ocorre na classe Pré-diabético, cujo comportamento limitado aparece de forma clara nas curvas ROC e PR do melhor modelo (Figura 2(a)) e na matriz de confusão correspondente (Figura 2(b)). Esta análise evidencia a forte sobreposição hematológica entre classes adjacentes, dado que os erros mais frequentes (proporcionalmente) ocorrem nas células (Real/Predito) PD/D, D/PD, PD/N e N/PD. Esses resultados motivam a exploração da estratégia T3, em que os classificadores binários especializados visam capturar fronteiras específicas entre classes, especialmente as mais desafiadoras, como Normal *vs.* Pré-diabético.



(a) Curvas ROC e PR para os Top 5 e melhor modelo em T2. (b) Matriz de confusão para o melhor modelo.

Figura 2. Desempenho de modelos multiclasse (T2) nos dados de teste.

4.2. Desempenho dos Classificadores Binários Especializados (T3) e Ensemble (T4)

A estratégia implementada em T3 avalia quatro fronteiras de decisão e, como esperado, as tarefas envolvendo classes mais distintas (N *vs.* D; N *vs.* PD) apresentaram os maiores F2-Scores, enquanto as fronteiras mais próximas (PD *vs.* D; N+PD *vs.* D) apresentaram os menores. Modelos com SMOTEENN foram os melhores nos modelos da T3 (Tabela 3), superando configurações com outras técnicas de balanceamento (*class_weight*, RUS), com ganhos de 1,2 a 4,4 p.p. em F2 e 6,2 a 8,9 p.p. em *recall*.

A T4 combina seis modelos binários enumerados (Tabela 4). Verificou-se que o *ensemble* por *soft voting* foi ligeiramente melhor que o *hard voting* em termos de desempenho geral. A Tabela 5 compara o resultado do *ensemble* via *soft voting* com o melhor modelo ternário direto. O *ensemble* melhora o desempenho para Normal e Pré-diabético, mas apresenta queda acentuada no diagnóstico de Diabético (*recall* de 0,213). Em média macro, o modelo ternário direto supera o *ensemble* (0,551 *vs.* 0,499), indicando que a combinação de especialistas não produziu ganhos expressivos nesta configuração.

Tabela 4. Especialistas binários do *ensemble* (T4).

Classe	Algoritmo	Dec.	F2	Rec.	Pr.
Normal	MLP	N vs PD+D	0,748	0,757	0,714
Normal	CatBoost	N vs PD+D	0,754	0,767	0,708
Pré-d.	MLP	N vs PD	0,628	0,638	0,589
Pré-d.	RF	PD vs D	0,424	0,399	0,568
Diab.	HistGB	N+PD vs D	0,230	0,199	0,586
Diab.	CatBoost	N+PD vs D	0,230	0,200	0,585

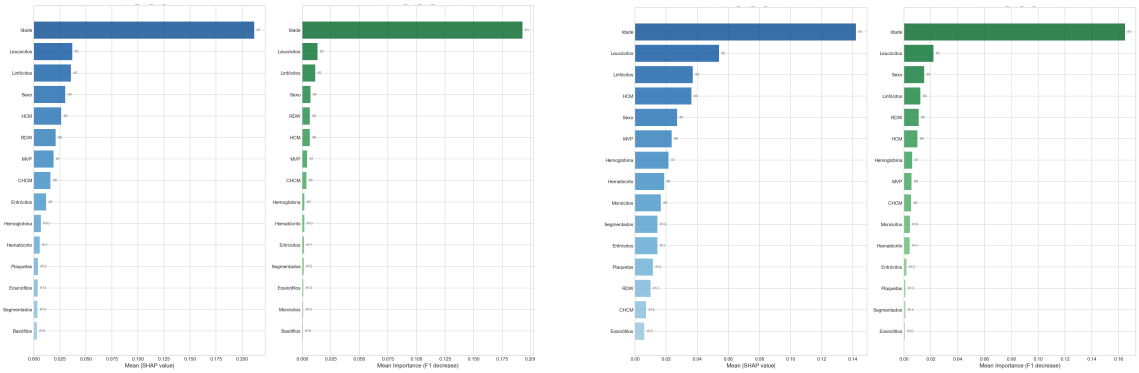
Tabela 5. Desempenho por classe: T2 *vs.* T4 (*soft voting*).

	T2			T4		
Classe	**F2**	**Rec.**	**Pr.**	**F2**	**Rec.**	**Pr.**
Normal	0,669	0,648	0,768	0,743	0,755	0,699
Pré-diab.	0,437	0,443	0,417	0,511	0,544	0,412
Diabético	0,547	0,578	0,449	0,244	0,213	0,581
Macro	**0,551**	**0,556**	**0,545**	**0,499**	**0,504**	**0,564**

4.3. Importância dos Atributos Preditivos

A importância de atributos por permutação analisada sobre os melhores modelos nas tarefas T1, T2 e T3 (Tabela 3), identificou Idade como o atributo mais relevante (1º em 4/6 tarefas, magnitude 3 a 18 vezes superior ao segundo), seguido por Leucócitos (1º em N+PD vs D e PD vs D, *top*-5 em 5/6 tarefas) e Sexo (*top*-5 em 5/6 tarefas). A série branca de células contribuiu com Leucócitos e Linfócitos, enquanto a série vermelha de células foi representada por RDW e Hemoglobina (*top*-5 em 3/6 tarefas cada). A contagem leucocitária elevada em pacientes diabéticos [Bambo et al. 2024] reforça o papel discriminativo desse parâmetro.

A análise de valores SHAP (Figura 3) confirmou Idade e Leucócitos como os atributos de maior contribuição nas predições em ambas as tarefas, em concordância com a importância por permutação. No problema ternário (T2), observa-se mudança no perfil de relevância: RDW, importante na T1 (6º), perde destaque (13º), enquanto atributos da série vermelha, como Hemoglobina e Hematócrito, ganham maior contribuição. A concordância entre os dois métodos reforça a robustez e interpretabilidade dos parâmetros hematológicos identificados. A correlação entre os rankings de importância foi maior em T1 do que em T2, indicando maior consistência na interpretação do modelo binário. Em análise complementar, a T1 foi reexecutada sem as variáveis demográficas, reduzindo o melhor F2 de 0,793 para 0,656 e elevando o RDW à posição de atributo mais relevante segundo SHAP, indicando que o hemograma preserva sinal preditivo autônomo, ainda que amplificado por Idade e Sexo.



(a) T1 - binário (N vs PD+D). Spearman $\rho = 0{,}975$.

(b) T2 - ternário (N vs PD vs D). Spearman $\rho = 0{,}889$.

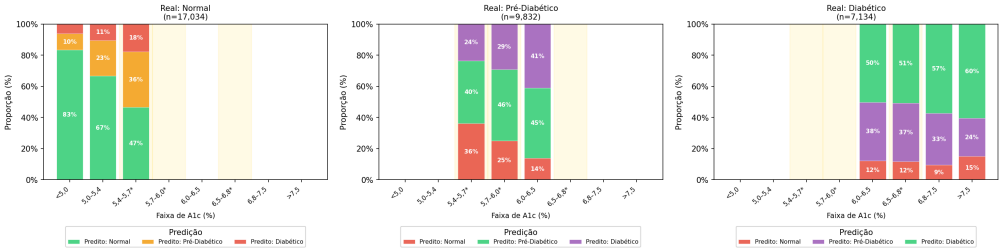
Figura 3. Comparação entre SHAP e importância por permutação do MLP nas tarefas binária (T1) e ternária (T2).

4.4. Análise dos Erros de Classificação

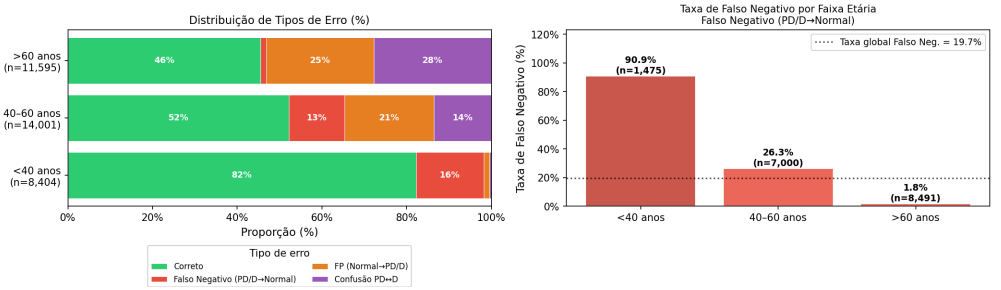
Por fim, realizamos uma análise dos padrões de erro do melhor modelo ternário (T2) obtido. A distribuição dos erros por faixa de A1c (Figura 4a) evidencia o comportamento do modelo nas regiões limítrofes dos critérios de diagnóstico da OMS. O pico de falsos positivos (36,6%) está concentrado na faixa 5,4 a 5,7%, imediatamente abaixo do limiar Normal/Pré-diabético (5,7%), enquanto o maior índice de falsos negativos (25,1%) ocorre na faixa 5,7 a 6,0%, logo acima desse mesmo limiar. Esse padrão confirma que a principal dificuldade preditiva está na divisão entre as classes Normal e Pré-diabético, onde os perfis hematológicos estão em maior sobreposição, corroborando a hipótese de que exames hematológicos apresentam limitação diagnóstica nas faixas glicêmicas limítrofes (ou transicionais).

A classe Pré-diabético (Figura 4a) apresentou taxa de erro de 55,7%, com dispersão bidirecional para as classes vizinhas: 56,5% dos erros para Diabético e 43,5% para Normal. Observa-se também que 23,6% dos indivíduos normais e 28,9% dos diabéticos foram classificados como Pré-diabético, confirmando a sobreposição dos perfis hematológicos nessa classe. Os 3.334 falsos negativos (PD→N: 2.386; D→N: 948), sendo 19,6% dos indivíduos com alteração glicêmica, representam o cenário de maior risco clínico por resultarem em subdiagnóstico da doença.

Também foi perceptível a variação no desempenho do modelo ternário por subgrupo demográfico (Figura 4b). Analisando os erros por faixa etária, notamos que 90,9% dos falsos negativos do modelo concentram-se em indivíduos com menos de 40 anos. Esse comportamento é consistente com a menor prevalência de alterações glicêmicas nessa faixa etária, o que reduz a capacidade discriminativa do modelo. Em contrapartida, os erros em indivíduos com mais de 60 anos ocorrem predominantemente como falsos positivos e confusões entre pré-diabetes e diabetes. Esses resultados sugerem que o modelo apresenta maior sensibilidade para detectar alterações glicêmicas em faixas etárias mais avançadas, enquanto tende a subestimar casos positivos em populações mais jovens. Em termos de diferenças de desempenho preditivo por sexo (disponível no repositório suplementar), verificamos que o modelo tende a ter menos falsos positivos ou cometer menos erros entre pré-diabéticos e diabéticos para mulheres, embora em termos de F2 macro a diferença seja de 0,562 (feminino) *vs.* 0,512 (masculino).



(a) Distribuição de erros por faixa de A1c.



(b) Distribuição de erros por faixa etária.

Figura 4. Análise de erros do classificador MLP ternário (T2).

5. Considerações Finais

Os resultados demonstram que o hemograma completo contém sinais relevantes para a classificação do status glicêmico, apesar de algumas limitações importantes. Modelos MLP apresentaram o melhor desempenho em todas as tarefas avaliadas, e, de modo geral, tarefas binárias envolvendo classes mais distintas (por exemplo, Normal *vs.* Diabético) apresentaram melhor desempenho do que aquelas entre classes adjacentes (Normal *vs.* Pré-diabético ou Pré-diabético *vs.* Diabético), refletindo a sobreposição progressiva dos perfis hematológicos ao longo do espectro glicêmico. Salienta-se que a modelagem foi deliberadamente restrita a parâmetros hematológicos, excluindo variáveis antropométricas ou bioquímicas que poderiam aumentar o desempenho preditivo, mas cuja inclusão reduziria a aplicabilidade em cenários de triagem baseados em exames de rotina.

A comparação direta com a literatura é limitada pela heterogeneidade de métricas, definições de classe e variáveis utilizadas. Ainda assim, a capacidade preditiva de modelos baseados em MLP neste estudo é consistente com achados prévios. Além disso, a base utilizada neste trabalho (169.999 registros) supera substancialmente o tamanho das bases reportadas em estudos correlatos com populações brasileiras, permitindo uma avaliação mais robusta do potencial preditivo do hemograma. A relevância de atributos como Idade, Leucócitos e RDW também está alinhada com evidências prévias que associam alterações hematológicas a processos inflamatórios e metabólicos relacionados ao diabetes [Bambo et al. 2024].

A elevada taxa de erro observada na classe Pré-diabético (55,7%), com confusão bidirecional para as classes Normal e Diabético, confirma que essa condição intermediária representa o principal desafio do problema. Esse resultado é consistente com a natureza fisiopatológica do pré-diabetes, caracterizada por alterações metabólicas ainda discretas e frequentemente sobrepostas aos perfis de indivíduos normais e diabéticos. Ainda assim, a tarefa binária Normal *vs.* Pré-diabético (F2=0,750) indica que o hemograma apresenta

algum grau de separabilidade para essa fronteira. Do ponto de vista clínico, essa distinção é particularmente relevante, uma vez que intervenções no estágio pré-diabético podem reduzir significativamente a progressão para diabetes tipo 2 [Galaviz et al. 2022].

Apesar dos resultados interessantes, este estudo apresenta algumas limitações. Primeiramente, os dados foram obtidos de um único laboratório clínico, o que pode introduzir vieses relacionados à população atendida ou aos equipamentos laboratoriais utilizados, limitando a generalização dos resultados. Em segundo lugar, a definição de classe baseia-se exclusivamente em valores de A1c, sem considerar outros critérios diagnósticos ou histórico clínico dos pacientes. Por fim, o estudo possui natureza transversal, utilizando exames coletados no mesmo momento clínico, o que limita inferências sobre capacidade preditiva longitudinal.

Em conjunto, os resultados indicam que modelos de AM aplicados ao hemograma completo podem capturar alguns padrões associados ao status glicêmico, apresentando desempenho particularmente promissor na distinção entre indivíduos normais e diabéticos. Entretanto, a identificação precisa do pré-diabetes permanece um desafio significativo devido à sobreposição entre perfis hematológicos. A principal contribuição deste trabalho reside na avaliação sistemática de diferentes estratégias de modelagem em uma base extensa de dados laboratoriais reais, evidenciando tanto o potencial quanto as limitações do hemograma como fonte de informação para triagem glicêmica. Como trabalhos futuros, destacam-se a validação externa com dados de outros centros, a incorporação de dados longitudinais e a integração com exames laboratoriais complementares para potencializar o apoio ao diagnóstico precoce.

Agradecimentos

Este estudo foi parcialmente financiado pela CAPES - Código de Financiamento 001, pela FAPERGS [Proj. nº 21/2551-0002052-0 e Proj. nº 22/2551-0000390-7] e pelo CNPq [Proj. nº 308075/2021-8].

Referências

Al-hussein, F., Tafakori, L., Abdollahian, M., Al-Shali, K., and Al-Hejin, A. (2025). A hybrid approach to enhance HbA1c prediction accuracy while minimizing the number of associated predictors: A case-control study in Saudi Arabia. *PLoS One*, 20(6):e0326315.

Alhassan, Z., Watson, M., Budgen, D., Alshammari, R., Alessa, A., and Moubayed, N. A. (2021). Improving current glycated hemoglobin prediction in adults: Use of machine learning algorithms with electronic health records. *JMIR Medical Informatics*, 9(5):e25237.

Bambo, G. M., Asmelash, D., Alemayehu, E., Gedefie, A., Duguma, T., and Kebede, S. S. (2024). Changes in selected hematological parameters in patients with type 1 and type 2 diabetes: A systematic review and meta-analysis. *Frontiers in Medicine*, 11:1294290.

Cardozo, G. et al. (2022). Use of machine learning and routine laboratory tests for diabetes mellitus screening. *BioMed Research International*, 2022:8114049.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Cheng, Y.-L., Wu, Y.-R., Lin, K.-D., Lin, C.-H. R., and Lin, I.-M. (2023). Using machine learning for the risk factors classification of glycemic control in type 2 diabetes mellitus. *Healthcare*, 11(8):1141.

Galaviz, K. I., Weber, M. B., Suvada, K., Gujral, U. P., Wei, J., Merchant, R., Dharanendra, S., Haw, J. S., Narayan, K. M. V., and Ali, M. K. (2022). Interventions for reversing prediabetes: A systematic review and meta-analysis. *American Journal of Preventive Medicine*, 62(4):614–625.

Le, V. O. H. et al. (2022). Formation and evaluation of complete blood count proficiency testing program. *Hematology Reports*, 14(2):73–84.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Lu, Y., Wang, W., Liu, J., Xie, M., Liu, Q., and Li, S. (2023). Vascular complications of diabetes: A narrative review. *Medicine*, 102(40):e35285.

Mansoori, A. et al. (2023). Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: A cohort study analysis. *Scientific Reports*, 13:663.

NCD Risk Factor Collaboration (NCD-RisC) (2023). Global variation in diabetes diagnosis and prevalence based on fasting glucose and hemoglobin A1c. *Nature Medicine*, 29(11):2885–2901.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tahir, A., Asghar, K., Shafiq, W., et al. (2024). Fingerprinting hyperglycemia using predictive modelling approach based on low-cost routine CBC and CRP diagnostics. *Scientific Reports*, 14(1):1090.

The Lancet (2023). Diabetes: a defining disease of the 21st century. *The Lancet*, 401(10394):2087.

World Health Organization (2021). Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus.