# Quando os Erros Informam: Apoio ao Diagnóstico de Diabetes em Cenários de Alta Incerteza

**Samuel Norberto Alves[1], Celso França[1], Regina T. I. Bernal[1], Crizian S. Gomes[1], Oluwatoyin Joy Omole[1], Deborah Malta[1], Marcos André Gonçalves[1], Jussara M. Almeida[1]**

[1]Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

{samuelnorbertoalves,criziansaar17,oluwatoyinoj,dcmalta}@ufmg.br

{celsofranca,jussara,mgoncalv}@dcc.ufmg.br

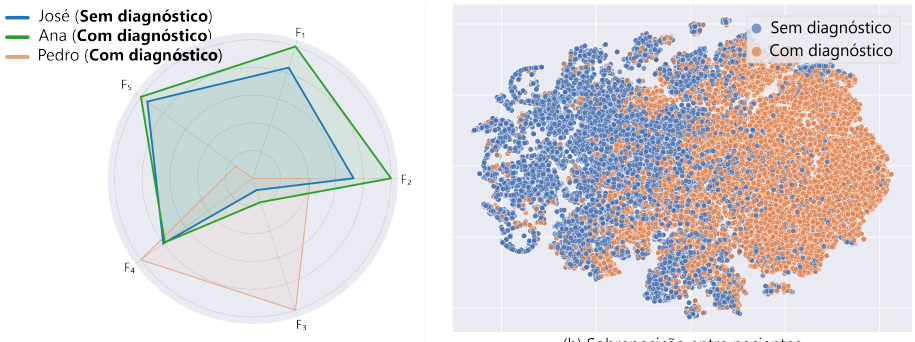
***Resumo.*** *Investigamos a efetividade de métodos supervisionados de aprendizado de máquina na identificação de indivíduos possivelmente não diagnosticados ou com alto risco de desenvolver Diabetes Mellitus (DM) no contexto de operadoras de saúde suplementar. O cenário é desafiador: há apenas dados administrativos indiretos (tipo e frequência de exames), sem resultados clínicos, além de baixa separabilidade entre classes e incerteza nos rótulos. Avaliamos três classificadores (XGBoost, Random Forest e Regressão Logística), obtendo desempenho robusto (Macro-F1 de 90,1%). A análise de erros sugere que falsos positivos podem indicar casos ainda não diagnosticados, enquanto falsos negativos podem refletir controle clínico inadequado.*

***Abstract.*** *We investigated the effectiveness of supervised machine learning methods in identifying individuals who may be undiagnosed or at high risk of developing Diabetes Mellitus (DM) in the context of private health insurance providers. The scenario is challenging: only indirect administrative data are available (such as the type and frequency of exams), without access to clinical results, along with low class separability and label uncertainty. We evaluated three classifiers (XGBoost, Random Forest, and Logistic Regression), achieving robust performance (Macro-F1 of 90.1%). Error analysis suggests that false positives may indicate undiagnosed cases, while false negatives may reflect inadequate clinical management.*

## 1. Introdução

A Agência Nacional de Saúde Suplementar (ANS) tem incentivado operadoras de planos de saúde a implementar programas voltados à promoção da saúde e à prevenção de doenças crônicas não transmissíveis (DCNT), com o objetivo de melhorar a qualidade de vida dos beneficiários [ANS 2021]. Tais ações são especialmente relevantes considerando que muitas dessas condições — como diabetes tipo 2, doenças cardiovasculares, respiratórias, renais e transtornos mentais — podem ser evitadas ou controladas com acompanhamento contínuo e mudanças sustentáveis no estilo de vida [Glechner et al. 2018].

Apesar da importância dessas diretrizes, identificar com precisão indivíduos em risco permanece sendo um desafio para as operadoras. Devido ao modelo de prestação de serviços, baseado em prestadores terceirizados, as operadoras frequentemente têm acesso apenas a registros administrativos — como o tipo e a frequência de exames




(a) Perfil de exames de pacientes.

(b) Sobreposição entre pacientes com diagnóstico e sem diagnóstico

**Figura 1. (a) Padrões de exames entre indivíduos com e sem diagnóstico de diabetes. (b) Projeção t-SNE de indivíduos com e sem diagnóstico de diabetes.**

realizados —, sem acesso aos respectivos laudos laboratoriais. Isso limita a obtenção de informações clínicas mais detalhadas e restringe a implementação de estratégias preventivas personalizadas. Nesse cenário, coexistem dois perfis: (i) indivíduos com diagnóstico formal de DCNT, documentado nos sistemas da operadora; e (ii) indivíduos que já apresentam a condição, mas ainda não foram identificados — seja por ausência de diagnóstico clínico ou por falhas no fluxo de informações entre prestadores e operadoras.

Soluções baseadas em aprendizado de máquina supervisionado têm se mostrado promissoras na identificação automatizada de padrões compatíveis com doenças crônicas [Ferreira et al. 2021, Kiran et al. 2025]. No entanto, a aplicação prática desses métodos em bases administrativas de operadoras de saúde suplementar apresenta desafios computacionais relevantes, dentre os quais destacamos: (i) o uso de atributos indiretos, como registros de exames sem acesso aos resultados; (ii) a baixa separabilidade entre classes, já que indivíduos com e sem diagnóstico podem apresentar padrões de exames semelhantes; e (iii) a incerteza nos rótulos, dado que nem todos os casos de DCNT estão formalmente identificados no sistema.

O primeiro desafio refere-se à presença de **atributos indiretos**. As operadoras conhecem os tipos e a frequência dos exames realizados, mas não dispõem dos laudos ou resultados clínicos associados, limitando a capacidade de inferência direta sobre a condição de saúde do paciente. O segundo desafio está relacionado à **baixa separabilidade entre classes**. Indivíduos diagnosticados e não diagnosticados frequentemente compartilham padrões semelhantes de exames, enquanto pacientes de uma mesma classe podem exibir perfis bastante distintos. A Figura 1(a) exemplifica esse cenário com três pacientes extraídos diretamente de um conjunto de dados reais de uma grande operadora de saúde na região Centro-Sul do Brasil: José (sem diagnóstico), Ana e Pedro (ambos diagnosticados com diabetes tipo 2). José e Ana, apesar de pertencerem a classes distintas, apresentam perfis de exames similares, enquanto Pedro e Ana, da mesma classe, têm perfis bastante diferentes. A Figura 1(b), utilizando t-SNE, reforça a sobreposição entre as classes, evidenciando a dificuldade de separação com base apenas na frequência dos exames.

O terceiro desafio refere-se à **incerteza nos rótulos**. Embora os registros de pacientes diagnosticados sejam confiáveis, ausência de diagnóstico não garante inexistência da condição. Indivíduos rotulados como não diabéticos podem apresentar padrões compatíveis com a doença, indicando possíveis casos subdiagnosticados, em consonância com evidências de que muitas ocorrências de diabetes tipo 2 permanecem

sem diagnóstico formal [Banday et al. 2020].

Diante desses desafios, este trabalho propõe uma abordagem supervisionada baseada em aprendizado de máquina, utilizando exclusivamente dados administrativos (tipos e frequência de exames) para identificação de indivíduos com padrões consistentes com diabetes tipo 2, mesmo que ainda não tenham sido formalmente diagnosticados. Além disso, buscamos também identificar pacientes diagnosticados que apresentam baixo nível de acompanhamento médico, o que pode indicar risco aumentado para complicações como retinopatia, nefropatia, neuropatia e doença cardiovascular aterosclerótica.

Um componente central da proposta está na análise dos **casos de erro** do modelo, que se concentram frequentemente em regiões de fronteira entre classes, onde os padrões são ambíguos. Essas regiões, embora desafiadoras para o classificador, podem revelar informações clínicas relevantes. Por exemplo, pacientes classificados como diabéticos pelo modelo, mas sem diagnóstico formal, podem representar casos de subdiagnóstico, favorecendo ações de triagem precoce. Por outro lado, pacientes diagnosticados, mas classificados como não diabéticos, podem indicar baixa adesão ao acompanhamento clínico — um sinal de alerta para potenciais complicações. A interpretação desses erros pode apoiar as operadoras de saúde na priorização de ações preventivas, em consonância com as diretrizes da ANS. Neste contexto, este trabalho busca responder às seguintes questões de pesquisa:

**QP1:** Qual é a efetividade da abordagem proposta na classificação de pacientes com e sem diabetes, considerando os desafios descritos?

**QP2:** Em um cenário com alta incerteza nos atributos e nos rótulos, qual é a confiabilidade do modelo?

**QP3:** Como interpretar os casos de erro gerados pelo modelo e quais são suas implicações clínicas?

Os experimentos demonstram que a abordagem proposta atinge alta efetividade, mesmo sob forte desbalanceamento de classes. O modelo com melhor desempenho (XGBoost) alcançou uma Macro-F1 de 90.1 (com intervalo de confiança de 95%). Em termos de confiabilidade, observa-se uma forte correlação entre a confiança das predições e sua acurácia, sendo o modelo bem calibrado — com *Brier Score Loss* de $0.054$ —, o que indica que as probabilidades estimadas estão bem alinhadas com os rótulos reais. Por fim, análise dos erros dos erros de classificação revelaram informações relevantes: falsos positivos podem indicar casos ainda não diagnosticados, enquanto falsos negativos sugerem baixa adesão ao tratamento. Tais evidências destacam o potencial do modelo tanto para fins preditivos quanto para subsidiar decisões clínicas em contextos de saúde pública.

## 2. Trabalhos Relacionados

Diversos estudos investigam o uso de aprendizado de máquina na predição de doenças crônicas, geralmente a partir de atributos extraídos diretamente de laudos clínicos — como os níveis de glicose para diabetes. [Alnowaiser 2024] propôs um modelo Tri-Ensemble, combinando XGBoost, Random Forest e Extra Trees via soft voting, com imputação de dados ausentes por KNN. [Tuppad and Devi Patil 2024] classificou diabetes e pré-diabetes utilizando modelos como Random Forest e Gradient Boosting e uma abordagem híbrida de seleção de atributos (*Feature Interaction-based Greedy Sequential Selection* e *Aglomerative Clustering*), com análise explicativa via SHAP.

[Dinh et al. 2019] explorou modelos com e sem dados laboratoriais, combinando classificadores (como Regressão Logística e Random Forest) por média ponderada das probabilidades. Os resultados indicam que a inclusão de exames laboratoriais melhora significativamente o desempenho na predição de diabetes.

Nosso estudo tem como objetivo identificar casos de diabetes tipo 2 usando apenas atributos indiretos, como o tipo e a frequência de exames realizados. Essa abordagem reflete um cenário mais desafiador e mais próximo da realidade das operadoras de planos de saúde, que frequentemente não têm acesso aos resultados clínicos desses exames.

## 3. Metodologia

O conjunto de dados usado neste estudo foi disponibilizado por uma grande operadora de saúde suplementar da região Centro-Sul do Brasil. Ele é composto por $116.145$ amostras, das quais $27.089$ correspondem a indivíduos com diagnóstico formal de diabetes tipo 2, enquanto os $89.056$ restantes não possuem esse diagnóstico registrado. A partir dessa distinção, definimos duas classes: com diagnóstico e sem diagnóstico de diabetes. Para garantir a qualidade dos dados, foram considerados apenas diagnósticos registrados manualmente, oriundos de acompanhamento clínico, com alto grau de confiabilidade.

Cada amostra do conjunto de dados corresponde a um único indivíduo, de modo que todas as amostras são independentes entre si. As variáveis preditoras representam a quantidade e frequência de realização de exames laboratoriais relevantes para o monitoramento da condição, como hemoglobina glicada, glicemia em jejum e perfil lipídico, além de indicadores indiretos de acompanhamento clínico. Entretanto, não há nenhum dado relacionado aos resultados de exames laboratoriais. A seleção dessas variáveis foi orientada pelo protocolo de acompanhamento do diabetes da Secretaria de Estado de Saúde de Minas Gerais (SES-MG) [da Cunha Paula 2014], assegurando relevância clínica e alinhamento com diretrizes do sistema público de saúde. A variável-alvo indica a presença ou ausência de diagnóstico de diabetes, totalizando 25 variáveis por amostra.

Quanto aos algoritmos de classificação, adotamos três abordagens supervisionadas: (i) **XGBoost**, pela robustez frente a variáveis heterogêneas e capacidade de lidar com desbalanceamento; (ii) **Random Forest**, um modelo de ensemble não paramétrico eficaz em cenários com dados ruidosos e correlacionados; e (iii) **Regressão Logística**, pela simplicidade, calibração[Cunha et al. 2025], interpretabilidade e ampla aplicação.

O protocolo experimental consistiu em validação cruzada com 5 partições (5-fold cross-validation). A efetividade dos modelos foi avaliada por meio das métricas de precisão, revocação e F1-score por classe, e F1 médio (Macro-F1), que considera o desbalanceamento das classes. Para comparação estatística entre os modelos, aplicou-se o teste t pareado com nível de confiança de 95% [Cunha et al. 2023, França et al. 2024].

Devido à distribuição desbalanceada de classes, aplicamos subamostragem aleatória da classe majoritária no conjunto de treinamento, ajustando a proporção para 2:1 em relação à classe minoritária. O conjunto de teste foi mantido inalterado.

## 4. Resultados Experimentais

**QP1.** A Tabela 1 apresenta os resultados de efetividade dos três modelos avaliados. O **XGBoost** obteve desempenho estatisticamente superior, com Macro-F1 de 90.1, destacando-se como o classificador mais eficaz no cenário proposto. Os demais algoritmos — Random Forest e Regressão Logística — apresentaram desempenho competitivo,

**Tabela 1. Efetividade de cada classificador.**

| Modelo | Sem Diagnóstico | | | Com Diagnóstico | | | Macro-F1 |
|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 | |
| **XGBoost** | **96.0(0.2)** | **94.5(0.1)** | **95.3(0.1)** | **82.8(0.4)** | **87.2(0.8)** | **84.9(0.5)** | **90.1(0.3)** |
| **RF** | 95.5(0.2) | 93.8(0.3) | 94.6(0.2) | 80.7(0.8) | 85.4(0.8) | 83.0(0.5) | 88.8(0.4) |
| **LR** | 94.2(0.2) | 92.8(0.2) | 93.5(0.1) | 77.4(0.5) | 81.3(0.8) | 79.3(0.4) | 86.4(0.3) |

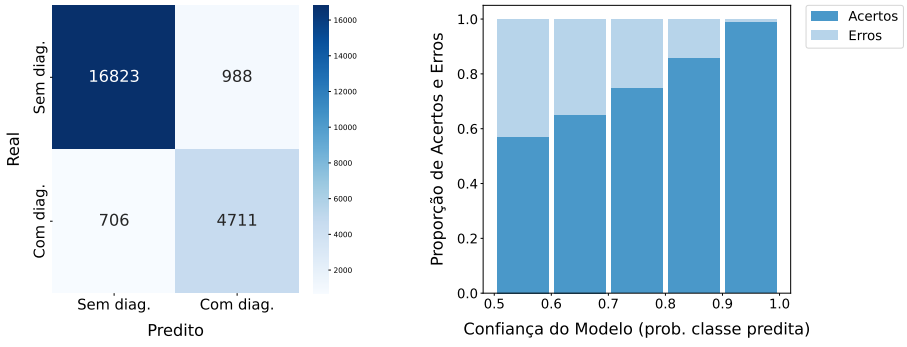



**Figura 2. (a) Matriz de confusão. (b) Proporção de acertos e erros por faixa de confiança do modelo.**

com métricas ligeiramente inferiores em ambas as classes. Esses resultados refletem o impacto positivo da subamostragem aleatória da classe majoritária na redução do viés e na melhoria da performance da classe minoritária.

A Figura 2(a) apresenta a matriz de confusão de um dos cinco folds de validação do **XGBoost**. O modelo acertou 16.823 instâncias da classe sem diagnóstico e 4.711 da classe com diagnóstico, totalizando 21.534 classificações corretas nesse fold. Apesar do desbalanceamento entre as classes, o modelo manteve consistência ao atingir aproximadamente $0{,}83$ de precisão e $0{,}87$ de revocação na classe com diagnóstico.

O elevado valor de revocação indica que o modelo é capaz de identificar a maioria dos indivíduos com diagnóstico formal — uma característica relevante em cenários reais, nos quais muitos pacientes com padrões clínicos compatíveis podem ainda não ter sido diagnosticados. Já a alta precisão demonstra que, entre os indivíduos classificados como diabéticos, a maioria pertence de fato à classe positiva, reduzindo o risco de falsos positivos em estratégias de triagem.

O poder discriminativo dos modelos viabiliza sua aplicação prática em diferentes frentes no contexto das operadoras de saúde suplementar, especialmente no atendimento às diretrizes da ANS. As predições podem apoiar a verificação de registros clínicos, a auditoria de prestadores e o monitoramento da adesão a protocolos, fornecendo suporte analítico para ações preventivas e melhoria da qualidade assistencial.

**QP2.** Após avaliada a efetividade do modelo, torna-se essencial analisar sua confiabilidade, sobretudo em um contexto marcado por possíveis imprecisões nos atributos e rótulos dos dados. Apesar dessas limitações, o desempenho observado foi satisfatório. A Figura 2(b) ilustra a relação entre a confiança das previsões, representada pela probabilidade atribuída às classes preditas, e a ocorrência de acertos ou erros. Observa-se uma boa calibração: a proporção de acertos aumenta progressivamente com o grau de confiança, indicando que o modelo identifica padrões discriminativos relevantes. Ademais, as faixas de maior confiança apresentam taxas de acerto significativamente elevadas, o que reforça a confiabilidade das predições em zonas de maior certeza. Essa confiabilidade é corrobo-

rada pelo valor do *Brier Score Loss* [Sledzik and Zabihimayvan 2022], que foi de 0,054. Essa métrica penaliza discrepâncias entre as probabilidades atribuídas e os resultados reais, sendo que valores mais próximos de zero indicam melhor calibração. No presente caso, o baixo valor obtido evidencia que as probabilidades emitidas pelo modelo são coerentes com os desfechos observados. Assim, ao considerar conjuntamente o *Brier Score* e a análise da calibração, conclui-se que o modelo não apenas apresenta elevado desempenho preditivo, mas também gera estimativas probabilisticamente consistentes e confiáveis.

**QP3** Dadas a robustez e calibração do modelo, os erros de classificação revelam-se informativamente relevantes, ampliando sua aplicação além da predição tradicional. Os falsos positivos (*i.e.*, indivíduos classificados como "com diagnóstico", porém rotulados como "sem diagnóstico") podem apresentar características compatíveis com estágios iniciais ou ainda não diagnosticados da doença, já que seus perfis assemelham-se aos de pacientes com diagnóstico confirmado. Essa capacidade é crucial para a detecção precoce do diabetes, favorecendo intervenções oportunas e a redução de complicações a longo prazo.

Em relação aos falsos negativos, uma hipótese plausível é que esses pacientes tenham baixa adesão ao tratamento ou realizem exames com frequência insuficiente, resultando em perfis clínicos semelhantes aos de indivíduos sem a doença. Embora isso possa induzir erro no modelo, evidencia sua sensibilidade a padrões comportamentais e de acompanhamento clínico. Assim, a análise dos erros evidencia o potencial do modelo como ferramenta exploratória, capaz de identificar casos atípicos ou de manejo inadequado, contribuindo para estratégias de monitoramento mais eficazes.

Dado esse contexto, o modelo pode ser muito útil para operadoras de saúde, especialmente aquelas que dispõem apenas de atributos indiretos sobre os beneficiários. Ele permite a identificação precoce de indivíduos que já apresentam ou estão em processo de desenvolver a doença, permitindo intervenções antecipadas para prevenir complicações que poderiam resultar em custos elevados. Além disso, o modelo pode detectar beneficiários cujo tratamento não está sendo realizado de forma adequada, permitindo à operadora agir proativamente, reduzindo riscos e gastos futuros.

## 5. Conclusão e Direções Futuras

Apesar dos desafios computacionais, como rótulos incertos, atributos indiretos, baixa separabilidade e forte desbalanceamento de classes, os modelos supervisonados propostos obtiveram desempenho bastante satisfatório na identificação de casos com diagnóstico de diabetes. Os bons resultados de efetividade, inclusive na classe minoritária, associadas à calibração adequada (baixo *Brier Score*), indicaram robustez frente à incerteza dos dados. Além disso, os erros de classificação mostraram-se bastante informativos: falsos positivos podem sinalizar casos subdiagnosticados, e falsos negativos, baixa adesão ao tratamento. Esses achados reforçam o potencial dos modelos tanto para predição quanto para apoio a decisões clínicas em saúde pública.

## Agradecimentos

Este trabalho foi apoiado por CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde e Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR; 408490/2024-1).

## Referências

Alnowaiser, K. (2024). Improving healthcare prediction of diabetic patients using knn imputed features and tri-ensemble model. *IEEE Access*, 12:16783–16793.

ANS (2021). Promoção da saúde e prevenção de doenças - PROMOPREV - https://www.gov.br/ans/pt-br/assuntos/operadoras/compromissos-e-interacoes-com-a-ans-1/programas-ans-1/promoprev. Atualizado em 06/06/2025.

Banday, M. Z., Sameer, A. S., and Nissar, S. (2020). Pathophysiology of diabetes: An overview. *Avicenna journal of medicine*, 10(04):174–188.

Cunha, W. et al. (2023). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *SIGIR*, page 665–674.

Cunha, W., Moreo Fernández, A., Esuli, A., Sebastiani, F., Rocha, L., and Gonçalves, M. A. (2025). A noise-oriented and redundancy-aware instance selection framework. *ACM Trans. Inf. Syst.*, 43(2).

da Cunha Paula, D. J. (2014). Análise de custo e efetividade do tratamento de diabéticos adultos atendidos no centro hiperdia de juiz de fora, minas gerais. Dissertação de mestrado, Universidade Federal de Juiz de Fora, Juiz de Fora, MG, Brasil. Aprovado em 17 de fevereiro de 2014.

Dinh, A., Miertschin, S., Young, A., and Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(1):211.

Ferreira, T., França, C., A. Gonçalves, M., Pagano, A., et al. (2021). Evaluating recognizing question entailment methods for a Portuguese community question-answering system about diabetes mellitus. In *Proc. Int'l Conf. on Recent Advances in Natural Language Processing*.

França, C., Lima, R. C., Andrade, C., Cunha, W., de Melo, P. O. V., Ribeiro-Neto, B., Rocha, L., Santos, R. L., Pagano, A. S., and Gonçalves, M. A. (2024). On representation learning-based methods for effective, efficient, and scalable code retrieval. *Neurocomputing*, 600:128172.

Glechner, A., Keuchel, L., Affengruber, L., Titscher, V., Sommer, I., Matyas, N., Wagner, G., Kien, C., Klerings, I., and Gartlehner, G. (2018). Effects of lifestyle changes on adults with prediabetes: A systematic review and meta-analysis. *Primary care diabetes*, 12(5):393–408.

Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscionek, B., and Russell, D. (2025). Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. *Frontiers in Digital Health*, 7:1557467.

Sledzik, R. and Zabihimayvan, M. (2022). Focal loss improves performance of high-sensitivity c-reactive protein imbalanced classification. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 114–118.

Tuppad, A. and Devi Patil, S. (2024). An efficient classification framework for type 2 diabetes incorporating feature interactions. *Expert Systems with Applications*, 239:122138.