MODELO XGBOOST NA PREDIÇÃO DE OBSTRUÇÃO PULMONAR: UMA ABORDAGEM BASEADA EM DADOS DE ESPIROMETRIA DO NHANES

Ítalo Gabriel da Silva[1], Romário de Sousa Almeida[1,2], José Wanderson Ferreira de Freitas[1], Elizeu Silva de Medeiros[1], Maria Helena Ramos de Macedo[1], Morramed da Silva Souto[1]

[1]Universidade Estadual da Paraíba - UEPB, Campina Grande, Brasil (italo.gabriel@aluno.uepb.edu.br)
[2]Universidade Federal de Lavras - UFLA, Lavras, Brasil

Resumo: Este estudo teve como objetivo desenvolver e avaliar um modelo preditivo baseado no algoritmo XGBoost para identificação da obstrução do fluxo aéreo utilizando dados de espirometria do NHANES. Foram analisados dados de 16.596 indivíduos, considerando variáveis demográficas, antropométricas e espirométricas. O modelo foi treinado e testado por meio de divisão estratificada (70/30) e avaliado por métricas de classificação e análise de discriminabilidade. Os resultados evidenciaram desempenho elevado, com acurácia de 98,37% e AUC de 0,9992, indicando excelente capacidade preditiva. A análise de interpretabilidade baseada em SHAP revelou que a razão FEV1/FVC é o principal fator associado às predições do modelo. Além disso, a análise de decisão demonstrou potencial aplicabilidade clínica. Os achados indicam que o XGBoost é uma abordagem robusta, precisa e interpretável para aplicações em saúde respiratória.

Palavras-chave: Modelagem preditiva; Função pulmonar; Aprendizado de máquina; Saúde respiratória.

INTRODUÇÃO

A avaliação da função pulmonar tem um papel fundamental na compreensão do estado de saúde respiratória da população, sendo amplamente empregada tanto em contextos clínicos quanto em pesquisas epidemiológicas. Nesse sentido, a espirometria é largamente reconhecida como o principal método para avaliação da função ventilatória, por possibilitar o acompanhamento de forma mais objetiva de parâmetros fisiológicos essenciais para o funcionamento pulmonar, como o volume expiratório forçado no primeiro segundo (VEF1), a capacidade vital forçada (CVF) e a razão VEF1/CVF (JOHNSON *et al.*, 2023).

Os testes de função pulmonar apresentam uma longa trajetória histórica, com origem no século XIX, e continuam sendo amplamente aplicados na avaliação da saúde respiratória. Apesar dos avanços, ainda persistem desafios relacionados à sua realização, interpretação e aplicação na prática clínica. Nos últimos tempos, diversos aspectos desses testes têm sido reavaliados, especialmente no que se refere ao uso de valores de referência e à interpretação dos parâmetros espirométricos (HAYNES, 2025).

Esses parâmetros espirométricos constituem ferramentas importantes no diagnóstico, monitoramento e acompanhamento de doenças respiratórias, incluindo asma e doença pulmonar obstrutiva crônica (DPOC), sendo frequentemente empregados na prática clínica (FORNO *et al.*, 2024).

A interpretação dos resultados espirométricos baseia-se, de forma geral, na comparação entre os valores observados e valores de referência obtidos a partir de populações saudáveis, permitindo a identificação de possíveis alterações na capacidade pulmonar, uma vez que valores de referência são obtidos em indivíduos sem tabagismo e sem doenças cardiorrespiratórias e são essenciais para interpretar resultados de espirometria (ALBUQUERQUE *et al.*, 2024).

Nesse contexto, a construção de equações de referência adequadas torna-se fundamental para a interpretação correta das medições espirométricas. Essas equações consideram fatores individuais importantes, como idade, altura e sexo, além de outros aspectos que podem influenciar a função pulmonar, incluindo condições socioeconômicas e ambientais, contribuindo para estimativas mais precisas e representativas, uma vez que esses fatores

influenciam diretamente a capacidade respiratória (KANJ *et al.*, 2023).

Mais recentemente, métodos estatísticos mais avançados têm sido propostos com o objetivo de representar melhor o comportamento da função pulmonar, permitindo descrever de maneira mais completa a variabilidade e outras características presentes em informações fisiológicas. Essas abordagens possibilitam uma análise mais detalhada do comportamento das variáveis na população, contribuindo para maior precisão na construção de equações de referência (LI *et al.*, 2024; BANZE *et al.*, 2025).

Paralelamente, algoritmos de aprendizado de máquina, como o XGBoost, vêm sendo cada vez mais adotados na área da saúde e na análise de grandes bases de informações, principalmente por sua elevada capacidade preditiva e eficiência no processamento de grandes volumes de registros.

Pesquisas recentes indicam que esse tipo de modelo é bastante eficaz na identificação de padrões complexos em bases extensas, podendo apresentar desempenho superior a outros métodos em diferentes aplicações, como na previsão de desfechos clínicos e na classificação de riscos em pacientes (CHEN *et al.*, 2023; GUAN *et al.*, 2023).

Além disso, investigações recentes destacam que o uso do XGBoost, quando combinado com técnicas de interpretabilidade, permite não apenas gerar boas previsões, mas também compreender melhor quais variáveis exercem maior influência nos resultados. Isso amplia o potencial desse método em estudos epidemiológicos e clínicos, tornando-o uma ferramenta relevante para a análise de informações complexas e para o apoio à tomada de decisão em diferentes contextos científicos (LIANG *et al.*, 2024; ZHANG *et al.*, 2024).

Para a avaliação dessas abordagens, é fundamental a utilização de bases amplas e representativas. Nesse cenário, bases populacionais como a *National Health and Nutrition Examination Survey* (NHANES) destacam-se como uma importante fonte de informações em saúde pública para a análise da função pulmonar e para a construção de equações de referência, por reunirem informações detalhadas obtidas por meio de entrevistas, exames físicos e testes laboratoriais, contribuindo para pesquisas epidemiológicas mais robustas (JOHNSON *et al.*, 2023).

Dessa forma, o objetivo da presente pesquisa foi desenvolver e avaliar um modelo preditivo baseado no algoritmo XGBoost para identificação da obstrução do fluxo aéreo utilizando dados de espirometria do NHANES.

MATERIAL E MÉTODOS

Base de dados e pré-processamento

Os dados utilizados para esse estudo foram obtidos da base pública do National Health and Nutrition Examination Survey (NHANES), que é disponibilizado pelo *Centers for Disease Control and Prevention* (CDC), um programa promovido pelo Centro Nacional de Estatísticas em Saúde dos Estados Unidos que coleta e disponibiliza informações sobre a saúde e a nutrição da população por meio de entrevistas, exames clínicos e testes laboratoriais. Para esta pesquisa, foram considerados dados de espirometria coletados entre os anos de 2007 e 2012.

O conjunto de dados analisado nesse estudo inclui as informações de 16.596 participantes com idades entre 6 e 80 anos e foram incluídos apenas os participantes onde os exames de espirometria atendem aos critérios mínimos de qualidade técnica estabelecidos pela *American Thoracic Society*, considerando manobras classificadas como grau A ou grau B de aceitabilidade. Foram excluídos indivíduos com dados ausentes nas variáveis de interesse desse estudo, como aqueles com exames espirométricos que são incompletos nos dados.

As principais variáveis espirométricas consideradas neste estudo foram: Volume expiratório forçado no primeiro segundo (VEF1, em litros), Capacidade vital forçada (CVF, em litros) e Razão entre VEF1 e CVF (VEF1/CVF, em %). Além dessas medidas, também foram consideradas variáveis demográficas e antropométricas dos participantes, como idade, altura, peso corporal e índice de massa corporal (IMC), fatores que podem influenciar diretamente na função pulmonar

A variável resposta considerada no estudo foi a presença de obstrução do fluxo aéreo, tratada como variável categórica binária.

Como variáveis explicativas, foram incluídas características demográficas, antropométricas e espirométricas dos indivíduos, a saber: idade, sexo, raça, índice de massa corporal (IMC), altura, peso corporal, volume expiratório forçado no primeiro segundo (VEF1), capacidade vital forçada (CVF), razão VEF1/CVF, pico de fluxo expiratório (PEF) e fluxo expiratório forçado entre 25% e 75% da CVF (FEF25-75).

Desenvolvimento do modelo XGBoost

Foi ajustado um modelo preditivo baseado no algoritmo XGBoost. pode ser representado como a soma de funções de árvores de decisão adicionadas sequencialmente:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \qquad (1)$$

onde $\hat{y}_i$ é o valor predito para a i-ésima observação, x_i representa o vetor de variáveis explicativas, K é o número total de árvores no modelo e f_k. corresponde à k-ésima árvore de decisão pertencente ao espaço de funções F.

A função objetivo otimizada pelo XGBoost é dada por:

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \qquad (2)$$

onde L representa a função objetivo, $l\left(y_i, \hat{yi}\right)$ é a função de perda que mede a discrepância entre os valores observados y_i e preditos $\hat{y}_i$, e $\Omega\left(f_k\right)$ é o termo de regularização que penaliza a complexidade do modelo, onde é dada por:

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad (3)$$

onde γ é o parâmetro que penaliza o número de folhas, T representa o número de folhas da árvore, λ é o parâmetro de regularização associado aos pesos das folhas, e w_j corresponde ao peso da j-ésima folha.

O modelo XGBoost foi ajustado utilizando a função objetivo logística binária ("*binary:logistic*"), adequada para problemas de classificação. Como métrica de avaliação durante o treinamento, foi adotada a área sob a curva ROC (AUC), permitindo monitorar a capacidade discriminatória do modelo.

Os principais hiperparâmetros foram definidos como: taxa de aprendizado (*learning rate*, η) igual a 0,1, profundidade máxima das árvores (*max_depth*) igual a 5 e número de iterações (*nrounds*) igual a 150. Esses parâmetros controlam o processo de aprendizado, sendo a taxa de aprendizado responsável por regular a contribuição de cada árvore adicionada ao modelo, enquanto a profundidade máxima limita a complexidade das árvores e o número de iterações define a quantidade total de árvores geradas.

A escolha desses hiperparâmetros foi realizada de forma empírica, buscando um equilíbrio entre desempenho preditivo e controle de sobreajuste (*overfitting*).

Estratégia de treinamento e teste

Os dados foram divididos em conjuntos de treinamento (70%) e teste (30%) por meio de amostragem estratificada, estratégia amplamente utilizada para preservar a proporção das classes e melhorar a avaliação do desempenho dos modelos (JOSEPH, 2022; SIVAKUMAR *et al.*, 2024).

Análise de desempenho

O desempenho do modelo foi avaliado no conjunto de teste utilizando as seguintes métricas: Acurácia (Accuracy), Precisão (Precision), Sensibilidade (Recall), F1-score e Área sob a curva ROC (AUC).
A *Acurácia* mede a proporção total de acertos do modelo:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

A Precisão avalia, entre os casos previstos como positivos, quantos são realmente positivos:

$$Precisão = \frac{TP}{(TP + FP)} \qquad (5)$$

O *Recall* (sensibilidade) mede a capacidade de identificar corretamente os positivos reais:

$$Recall = \frac{TP}{(TP + FN)} \qquad (6)$$

O F1-Score corresponde à média harmônica entre precisão e recall:

$$F1 - Score = 2 \cdot \frac{Precisão \cdot Recall}{Precisão + Recall} \qquad (7)$$

A AUC (Área sob a curva ROC) avalia a capacidade discriminatória do modelo:

$$TPR = \frac{TP}{(TP + FN)} \qquad (8)$$

e

$$FPR = \frac{FP}{(FP + TN)} \qquad (9)$$

onde TPR e FPR representam o desempenho do modelo para diferentes limiares de classificação.

Além disso, foram construídas curvas ROC e estimados intervalos de confiança de 95% para a AUC por meio de reamostragem bootstrap com 2000 repetições.

Análise de interpretabilidade

Também foram realizadas análises adicionais, incluindo: Decision Curve Analysis (DCA), para avaliação da utilidade clínica, Importância das variáveis (*Feature Importance*), Análise de interpretabilidade baseada em valores SHAP.

Para melhor compreensão do comportamento do modelo XGBoost, foram utilizadas técnicas de interpretabilidade. A importância das variáveis foi avaliada por meio do próprio modelo XGBoost, enquanto os valores SHAP (*SHapley Additive exPlanations*) foram empregados para quantificar a contribuição de cada variável nas predições individuais (análise local) e no modelo como um todo (análise global).

Adicionalmente, foram construídos gráficos de dependência para avaliar o efeito de variáveis específicas, como a razão VEF1/CVF, sobre as predições do modelo.

Implementação computacional

As análises estatísticas foram realizadas na linguagem de programação R (versão 4.3.2), utilizando o ambiente RStudio (R CORE TEAM, 2023). Foram empregados diversos pacotes para modelagem estatística e aprendizado de máquina, incluindo caret, xgboost, pROC, vip, iml, fastshap, rms e rmda.

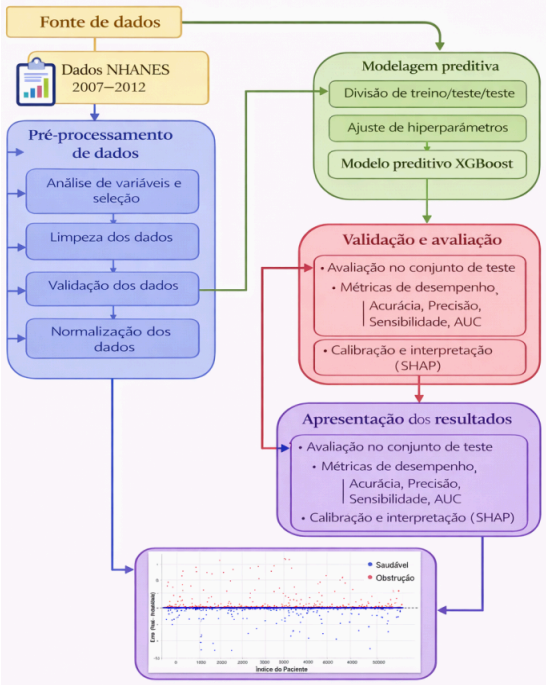

Figura 1. Fluxograma geral da metodologia empregada no estudo.

RESULTADOS E DISCUSSÃO

Inicialmente, a capacidade discriminatória do modelo foi avaliada por meio da curva ROC, apresentada na Figura 2. Observa-se que a curva se concentra próxima ao canto superior esquerdo do gráfico, indicando elevada sensibilidade e especificidade simultaneamente.

Esse comportamento evidencia forte capacidade do modelo em distinguir corretamente as classes, com desempenho próximo do ideal, uma vez que a curva ROC expressa a relação entre sensibilidade e taxa de falsos positivos e é amplamente utilizada para avaliar o poder discriminatório de modelos de classificação (ROBIN *et al.*, 2022).

Além disso, curvas mais próximas do canto superior esquerdo indicam maior acurácia e melhor desempenho diagnóstico do modelo (HAJIAN-TILAKI, 2023).

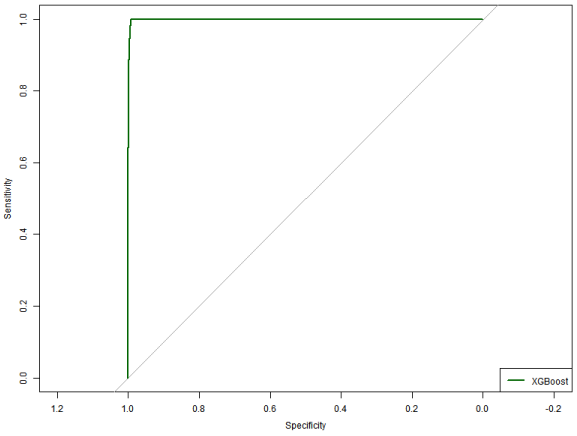

Figura 2. Curva ROC comparando o desempenho dos modelo de classificação.

Além da avaliação de desempenho, foi analisada a utilidade prática do modelo por meio da curva de decisão (Figura 3). Essa abordagem permite avaliar o benefício clínico associado às previsões em diferentes limiares de probabilidade, por meio da quantificação do benefício líquido ao considerar o balanço entre verdadeiros positivos e falsos positivos (VICKERS; VAN CALSTER; STEYERBERG, 2022).

Os resultados indicam que o modelo apresenta ganho líquido superior em uma ampla faixa de limiares, reforçando sua aplicabilidade como ferramenta de apoio à decisão.

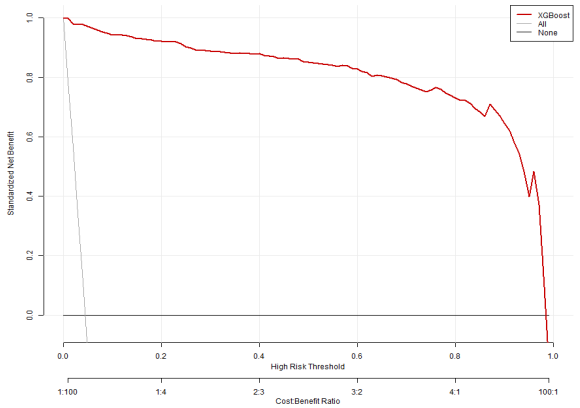

Figura 3. Análise de curva de decisão demonstrando o benefício líquido do modelo preditivo.

O desempenho do modelo também foi avaliado por meio de métricas derivadas da matriz de confusão, incluindo verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). A acurácia mede a proporção total de acertos, enquanto a precisão avalia a proporção de verdadeiros positivos entre os casos previstos como positivos. O recall (sensibilidade) mede a capacidade do modelo em identificar corretamente os casos positivos, e o F1-Score representa o equilíbrio entre precisão e recall.

A Tabela 1 apresenta os resultados dessas métricas. De modo geral, observa-se desempenho elevado em todas as medidas, destacando-se os altos valores de precisão (99,66%), recall (99,68%) e AUC (0,9992). O F1-Score também se manteve elevado (0,9674), indicando equilíbrio consistente entre as métricas. Esses resultados reforçam a robustez do modelo e sua elevada capacidade preditiva.

Tabela 1. Comparação das métricas de desempenho do modelo XGBoost.

Métricas	XGBoost
Acurácia	98,37%
Precisão	99,66%
Recall	99,68%
F1-Score	0,9674
AUC	0,9992

Após a avaliação do desempenho, buscou-se compreender quais variáveis mais influenciaram as previsões do modelo. De forma complementar, a análise global dos valores SHAP (Figura 4) mostra que as predições são predominantemente influenciadas por variáveis espirométricas, com destaque para a razão FEV1/FVC e, em menor grau, a idade.

Outras variáveis, como altura, sexo e FVC, apresentam influência moderada, enquanto IMC, peso e demais variáveis possuem impacto reduzido, e variáveis relacionadas à raça demonstraram contribuição praticamente nula no modelo.

A utilização dos valores SHAP permite quantificar a contribuição individual de cada variável para as previsões, fornecendo uma medida consistente de importância e interpretabilidade do modelo (ARRIETA *et al.*, 2022).

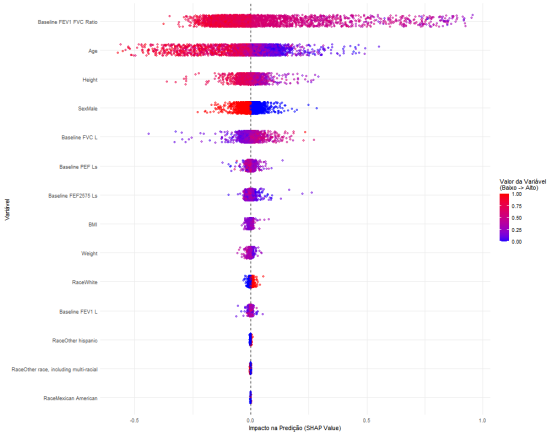

Figura 4. Importância e Impacto das Variáveis na Predição do Modelo XGBoost (Análise SHAP).

Para aprofundar essa análise, o gráfico de dependência SHAP (Figura 5) permite observar como diferentes valores dessa variável influenciam as previsões do modelo. Verifica-se que valores mais elevados da razão FEV1/FVC estão associados a maior contribuição para a predição, enquanto valores mais baixos reduzem essa probabilidade, evidenciando a relação entre o valor da variável e seu impacto nas saídas do modelo (ARRIETA *et al.*, 2022).

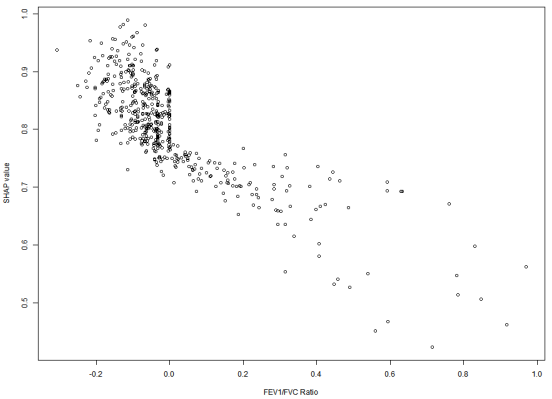

Figura 5. Análise de dependência SHAP do modelo XGBoost mostrando a relação entre o índice FEV1/FVC e sua contribuição para a predição do modelo.

A estrutura dos dados também foi analisada por meio do mapa de calor da matriz de correlação (Figura 6),

que permite identificar relações lineares entre as variáveis. Essa análise auxilia na compreensão de possíveis associações entre preditores que podem influenciar o comportamento do modelo, uma vez que mapas de calor da matriz de correlação são amplamente utilizados para explorar relações entre variáveis e compreender a estrutura dos dados em análises multivariadas (SINGH; NAGAHARA, 2024; Elsevier, 2024).

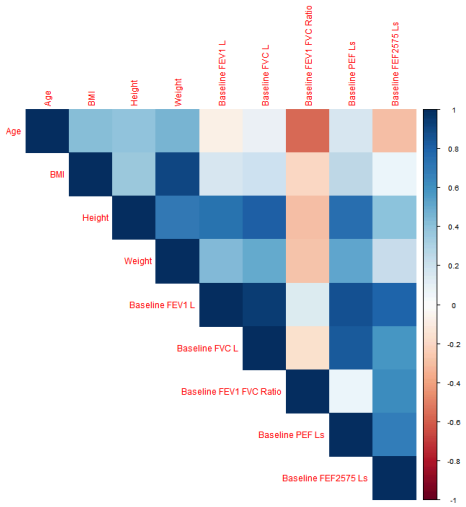

Figura 6. Matriz de correlação entre as variáveis do estudo.

Por fim, a análise dos resíduos do modelo (Figura 7) indica que os erros estão, em sua maioria, próximos de zero, sugerindo boa concordância entre valores previstos e observados.

A distribuição aleatória dos resíduos indica ausência de padrões sistemáticos, o que reforça o bom ajuste do modelo, uma vez que resíduos distribuídos aleatoriamente em torno de zero são indicativos de que o modelo capturou adequadamente a relação entre preditores e resposta e não apresenta falhas de especificação (VERMA, 2025). Embora existam alguns erros mais extremos, eles são pouco frequentes e não comprometem o desempenho geral.

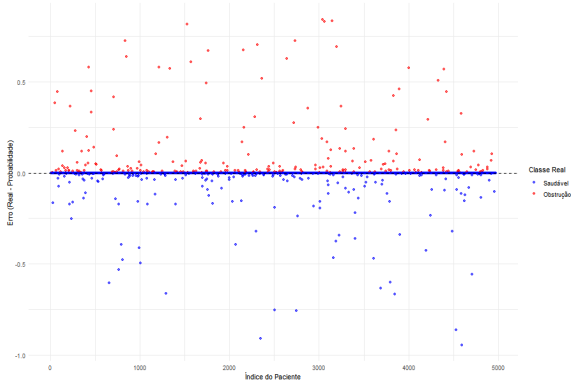

Figura 7. Resíduos do modelo XGBoost (erro entre valores observados e probabilidades previstas).

De maneira geral, os resultados indicam que o modelo XGBoost apresenta desempenho preditivo elevado, com forte capacidade discriminatória, boa calibração e previsões consistentes. A análise integrada das métricas e das visualizações do modelo evidencia sua habilidade em capturar padrões complexos nos dados, tornando-o uma ferramenta robusta e confiável para a tarefa de classificação proposta.

CONCLUSÃO

O modelo XGBoost apresentou desempenho preditivo elevado na identificação da obstrução do fluxo aéreo, com alta acurácia, excelente capacidade discriminatória e adequada calibração. A análise de interpretabilidade evidenciou que variáveis espirométricas, especialmente a razão FEV1/FVC, são os principais determinantes das predições, reforçando a consistência fisiológica dos resultados.

Além disso, a análise de decisão indicou potencial aplicabilidade clínica do modelo, com ganho líquido em diferentes limiares de probabilidade. De forma geral, os resultados demonstram que o XGBoost é uma abordagem robusta, precisa e interpretável, com elevado potencial para aplicações em estudos epidemiológicos e apoio à decisão em saúde respiratória.

Estudos futuros poderão validar o modelo em outras populações, explorar novas variáveis clínicas e ambientais e avaliar o impacto do uso dessa abordagem em contextos de apoio à decisão clínica, ampliando ainda mais sua aplicabilidade e relevância prática.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão de bolsas de pesquisa.

À Universidade Estadual da Paraíba (UEPB), pelo suporte institucional

REFERÊNCIAS

BANZE, S. et al. Global lung function initiative equations: recent advances and applications. *Respiratory Research*, 2025. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40975970/. Acesso em: 19 mar. 2026.

CHEN, C. et al. XGBoost-based machine learning test improves the accuracy of hemorrhage prediction

among geriatric patients. *BMC Geriatrics*, 2023. Disponível em: https://link.springer.com/article/10.1186/s12877-023-04049-z. Acesso em: 19 mar. 2026.

FORNO, E. et al. Spirometry interpretation after implementation of race-neutral reference equations in children. *JAMA Pediatrics*, v. 178, n. 7, p. 699–706, 2024. Disponível em: https://jamanetwork.com/journals/jamapediatrics/fullarticle/2819224#google_vignette. Acesso em: 19 mar. 2026.

GUAN, X. et al. Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. *BMC Medical Informatics and Decision Making*, 2023. Disponível em: https://link.springer.com/article/10.1186/s12911-023-02171-x. Acesso em: 19 mar. 2026.

HAYNES, J. M. Year in review: pulmonary function testing. *Respiratory Care*, v. 70, n. 8, p. 1033–1044, 2025. Disponível em: https://journals.sagepub.com/doi/10.1089/respcare.13066?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed. Acesso em: 19 mar. 2026.

JOHNSON, D. C.; JOHNSON, B. G. Spirometry reference equations including existing and novel parameters. *The Open Respiratory Medicine Journal*, v. 17, p. e187430642212260, 2023. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC10351349/. Acesso em: 18 mar. 2026.

KANJ, N. A. et al. Application of Global Lung Function Initiative global spirometry reference equations across a large, multicenter pulmonary function lab population. American Journal of Respiratory and Critical Care Medicine, v. 209, n. 1, p. 83–90, 2024. Disponível em: https://www.atsjournals.org/doi/10.1164/rccm.202303-0613OC?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed. Acesso em: 18 mar. 2026.

LI, A. et al. Implications of the 2022 lung function update and GLI global reference equations among patients with interstitial lung disease. *Thorax*, v. 79, n. 11, p. 1024–1032, 2024. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC11503192/ Acesso em: 18 mar. 2026.

LIANG, Y. et al. Predicting higher risk factors for COVID-19 reinfection based on XGBoost algorithm. *Journal of Translational Medicine*, 2024. Disponível em: https://link.springer.com/article/10.1186/s12967-024-05982-2. Acesso em: 18 mar. 2026.

ZHANG, X. et al. Explainable machine learning with XGBoost for health prediction. *Bioengineering*, 2024. Disponível em: https://www.researchgate.net/publication/390965086_Explainable_AI_in_Maternal_Health_Utilizing_XGBoost_and_SHAP_Values_for_Enhanced_Risk_Prediction_and_Interpretation. Acesso em: 19 mar. 2026.

ROBIN, X.; TURCK, N.; HAINARD, A.; et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, Londres, v. 23, 2022. Disponível em: https://bmcbioinformatics.biomedcentral.com/. Acesso em: 20 mar. 2026.

HAJIAN-TILAKI, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine, v. 14, n. 2, 2023. Disponível em: https://caspjim.com/. Acesso em: 20 mar. 2026.

VICKERS, A. J.; VAN CALSTER, B.; STEYERBERG, E. W. Decision curve analysis: a technical note. Diagnostic and Prognostic Research, v. 6, n. 1, 2022. Disponível em: https://diagnprognres.biomedcentral.com/. Acesso em: 20 mar. 2026.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; DEL SER, J.; et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, v. 58, p. 82–115, 2022. Disponível em: https://www.sciencedirect.com/. Acesso em: 20 mar. 2026.

SINGH, Nitin Kumar; NAGAHARA, Masaaki. LightGBM-, SHAP-, and correlation-matrix-heatmap-based framework for data analysis. Energies, [S. l.], v. 17, 2024. Disponível em: https://www.mdpi.com/. Acesso em: 20 mar. 2026.

ELSEVIER. Using a 3D heat map to explore correlations in data. Data Science Journal, 2024. Disponível em: https://www.sciencedirect.com/. Acesso em: 20 mar. 2026.

VERMA, Vibhu. *A comprehensive framework for residual analysis in regression and machine learning.* Journal of Information Systems Engineering and Management, v. 10, n. 31s, p. 34–46, 2025. DOI: 10.52783/jisem.v10i31s.4958. Disponível em: https://jisem-journal.com/index.php/journal/article/view/4958. Acesso em: 21 mar. 2026.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2023. Disponível em: https://www.R-project.org/. Acesso em: 20 mar. 2026.

R STUDIO TEAM. RStudio: Integrated Development Environment for R. Boston, MA:

RStudio, PBC, 2023. Disponível em: https://www.rstudio.com/. Acesso em: 20 mar. 2026.

ALBUQUERQUE, A. L. P. et al. New spirometry recommendations from the Brazilian Thoracic Association – 2024 update: considerations on the interpretation of spirometry and reference values. Jornal Brasileiro de Pneumologia, 2024. Disponível em: https://www.scielo.br/j/jbpneu/a/83VrJwhFpKGNVnPgjpgJxvK/?lang=en. Acesso em: 21 mar. 2026.

SIVAKUMAR, M.; PARTHASARATHY, S.; PADMAPRIYA, T. Trade-off between training and testing ratio in machine learning for medical image processing. PeerJ Computer Science, 2024. Disponível em: https://peerj.com/articles/cs-2245/. Acesso em: 21 mar. 2026.

JOSEPH, V. Roshan. Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal, v. 15, n. 4, p. 531–538, 2022. Disponível em: https://doi.org/10.1002/sam.11583. Acesso em: 21 mar. 2026.