Journal of Health Informatics

J Health Inform. 2026;18:e1783
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1783

ARTIGO COMPLETO
CBIS 2026

Estabilidade temporal de modelo preditivo de mortalidade infantil em prematuros: validação walk-forward anual com dados nacionais (2014–2022)

Temporal stability of an infant mortality prediction model in preterm infants: annual walk-forward validation with national data (2014–2022)

Sayonara C. de O. Magalhães[1], Karolayne Santos de Azevedo[1], Luísa Christina de Souza[1], Matheus Dalmolin[1], Marcelo Augusto Costa Fernandes[2] ✉

ORCID: 0009-0009-5845-4711 · 0000-0001-5062-5920 · 0000-0002-3707-2097 · 0000-0002-5400-0909 · 0000-0001-7536-2506

[1] InovAI Lab, nPITI/IMD, Universidade Federal do Rio Grande do Norte (UFRN), Natal (RN), Brasil.

[2] InovAI Lab, nPITI/IMD, Universidade Federal do Rio Grande do Norte (UFRN); Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal (RN), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1783

DOI
10.59681/2175-4411.v18.iEspecial.2026.1783

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Marcelo Augusto Costa Fernandes
mfernandes@dca.ufrn.br

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Objetivo: Avaliar a estabilidade temporal de um modelo de aprendizado de máquina (XGBoost) para predição de mortalidade infantil em prematuros ao longo de nove anos de dados nacionais (2014–2022), quantificando degradação de desempenho e drift das variáveis preditoras, com ênfase no período pandêmico (2020–2021). **Método:** Validação walk-forward anual com seis janelas de treino acumulativas a partir de 2014, aplicando o modelo a cada ano de teste consecutivo (2017–2022). Calcularam-se AUC-ROC, AUPRC e Brier Score para performance drift, e teste de Kolmogorov-Smirnov (KS) e Population Stability Index (PSI) para data drift, utilizando registros vinculados do SINASC e SIM. **Resultados:** O modelo manteve AUC-ROC entre 0,922 e 0,933 em todos os anos, sem atingir nenhum limiar de alerta. O PSI máximo observado foi 0,070 (2022), bem abaixo do limiar de alerta (0,10). Nenhuma variável preditora atingiu status de alerta em qualquer ano, incluindo os anos pandêmicos (2020–2021). **Conclusão:** O modelo demonstrou estabilidade temporal robusta, com drift epidemiológico gradual e benigno, sustentando política de retreinamento anual com monitoramento contínuo de drift.

Descritores: mortalidade infantil; aprendizado de máquina; drift de modelo.

ABSTRACT

Objective: To evaluate the temporal stability of a machine learning model (XGBoost) for predicting infant mortality in preterm infants across nine years of national data (2014–2022), quantifying performance degradation and predictor variable drift, with emphasis on the pandemic period (2020–2021). **Method:** Annual walk-forward validation with six expanding training windows starting in 2014, applying the model to each consecutive test year (2017–2022). AUC-ROC, AUPRC, and Brier Score were computed for performance drift; Kolmogorov-Smirnov (KS) test and Population Stability Index (PSI) for data drift, using linked SINASC and SIM records. **Results:** The model maintained AUC-ROC between 0.922 and 0.933 across all years, never reaching any alert threshold. The maximum observed PSI was 0.070 (2022), well below the alert threshold (0.10). No predictor variable reached alert status in any year, including pandemic years (2020–2021). **Conclusion:** The model demonstrated robust temporal stability with gradual and benign epidemiological drift, supporting an annual retraining policy with continuous drift monitoring.

Keywords: infant mortality; machine learning; model drift.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Magalhães SCO, Azevedo KS, Souza LC, Dalmolin M, Fernandes MAC. Estabilidade temporal de modelo preditivo de mortalidade infantil em prematuros: validação walk-forward anual com dados nacionais (2014–2022). J Health Inform 2026;18:e1783. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1783

1 INTRODUÇÃO

A mortalidade infantil em prematuros constitui um dos principais desafios da saúde pública brasileira. Apesar da redução progressiva da mortalidade infantil observada nas últimas décadas, os óbitos neonatais, em especial aqueles associados à prematuridade, permanecem como componente relevante dessa mortalidade, refletindo desigualdades regionais, condições socioeconômicas e heterogeneidade no acesso à assistência pré-natal e perinatal [1]. A identificação precoce de recém-nascidos prematuros com maior risco de óbito constitui estratégia relevante para orientar ações de monitoramento intensivo e alocação racional de recursos no Sistema Único de Saúde [2].

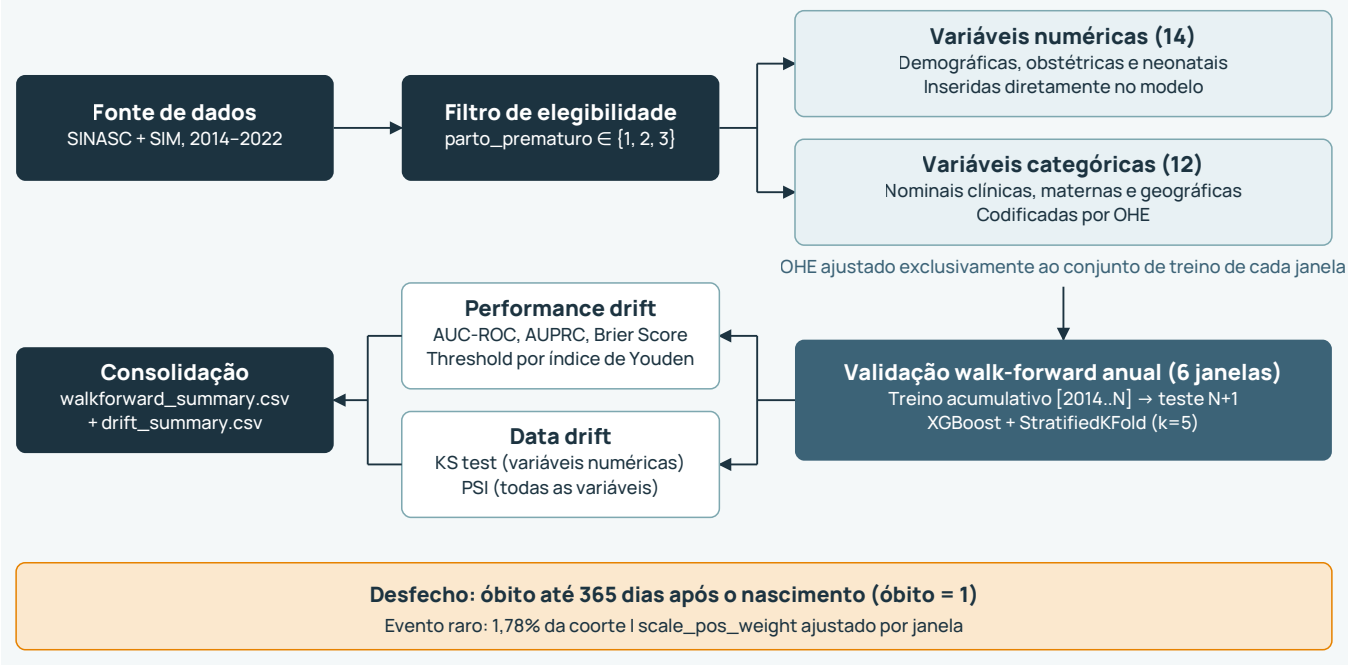
Modelos de aprendizado de máquina têm sido aplicados de forma crescente na estimativa de risco em contextos clínicos e epidemiológicos perinatais. Em estudos anteriores, técnicas de aprendizado não supervisionado foram utilizadas para estratificação de risco de prematuridade com dados socioeconômicos brasileiros [3-5], e modelos baseados em *gradient boosting* foram explorados para identificação de padrões de mortalidade em prematuros com dados do Sistema de Informações sobre Nascidos Vivos (SINASC) e do Sistema de Informações sobre Mortalidade (SIM) [6]. Revisões sistemáticas recentes indicam que modelos baseados em *Extreme Gradient Boosting* (XGBoost) e similares alcançam área sob a curva ROC (AUC-ROC) superiores a 0,90 em coortes de países de baixa e média renda, com destaque para o peso ao nascer, a idade gestacional e o escore de Apgar como variáveis de maior poder discriminativo [7]. No contexto nacional, estudos demonstraram que a integração de variáveis maternas, obstétricas, neonatais e geográficas em modelos preditivos populacionais melhora substancialmente a capacidade de concentração de risco em comparação a abordagens com variáveis isoladas [1]. Um desafio crítico e sistematicamente negligenciado na literatura é a avaliação da estabilidade temporal desses modelos após o período de treinamento. A maioria dos estudos reporta métricas de validação cruzada ou validação em um único ponto no tempo, sem investigar a degradação longitudinal do desempenho, fenômeno conhecido como *concept drift* ou *model drift* [8]. Alterações em protocolos clínicos, mudanças no perfil demográfico das parturientes, variações no acesso ao pré-natal e, mais recentemente, os impactos sistêmicos da pandemia de COVID-19 são fatores que podem deteriorar modelos treinados em períodos anteriores [9,10]. A ausência de monitoramento contínuo do desempenho representa risco real de uso clínico de modelos silenciosamente degradados, comprometendo a segurança e a utilidade das predições [11].

Neste contexto, o presente estudo tem como objetivo avaliar a estabilidade temporal de um modelo de aprendizado de máquina (*machine learning*, ML) do tipo XGBoost para predição de mortalidade infantil em prematuros, treinado com dados nacionais vinculados do SINASC e SIM no período de 2014 a 2022, por meio de validação *walk-forward* anual com janelas de treino crescentes. Adicionalmente, propõe-se uma análise sistemática de *data drift* das variáveis preditoras ao longo do período, com ênfase no impacto da pandemia de COVID-19. Os resultados demonstram que o modelo manteve AUC-ROC entre 0,922 e 0,933 em todos os anos avaliados (incluindo os anos pandêmicos) e que o *drift* epidemiológico observado foi gradual e benigno, sem atingir limiares de alerta em nenhuma das 26 variáveis do modelo ao longo de seis anos de validação prospectiva.

2 MÉTODOS

O *pipeline* metodológico adotado neste estudo está representado na Figura 1 e compreende cinco etapas principais: (1) carregamento e filtragem da coorte a partir da vinculação probabilística entre registros do SINASC e do SIM, disponibilizados pelo Ministério da Saúde, referentes ao período de 2014 a 2022, totalizando 3.064.338 registros; (2) preparação das variáveis preditoras, com 14 variáveis numéricas e 12 categóricas, sendo estas últimas codificadas por One-*hot encoding* (OHE) ajustado ao conjunto de treino de cada janela; (3) validação *walk-forward* anual com seis janelas de treino acumulativo e avaliação no ano subsequente; (4) cálculo paralelo de métricas de *performance drift* (AUC-ROC, AUPRC [*Area Under the Precision-Recall Curve*] e Brier Score) e de *data drift* usando o teste de Kolmogorov-Smirnov (KS) e a métrica de *Population Stability Index* (PSI); e (5) consolidação dos resultados em arquivos estruturados para análise de estabilidade temporal.

Figura 1. Pipeline metodológico do estudo. Etapas de construção da coorte, preparação de variáveis, validação walk-forward, avaliação de drift e consolidação dos resultados.



Fonte: Elaborado pelos autores (2026).

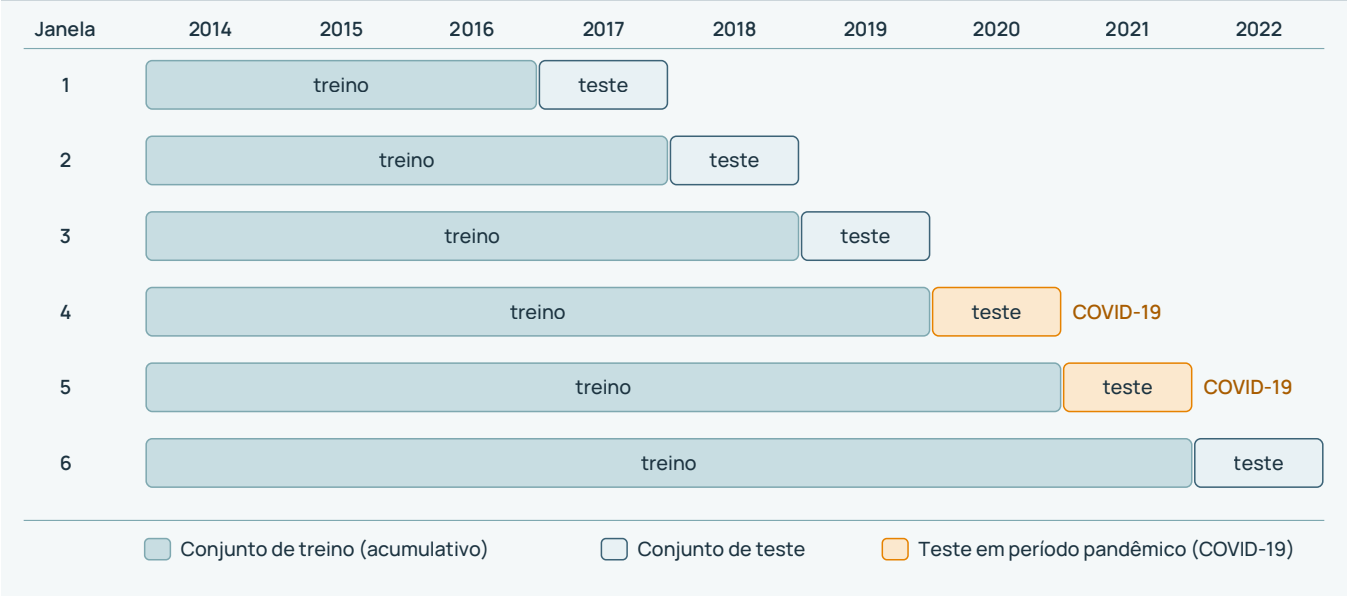
2.1 Fonte de Dados e Coorte

A população elegível foi composta por nascidos vivos classificados como prematuros nas categorias: prematuro confirmado por idade gestacional, inconclusivo por idade gestacional e inconclusivo por peso ao nascer inferior a 2.500 g (variável TIPO_PREMATURO ∈ {1, 2, 3}). O desfecho de interesse foi o óbito até 365 dias após o nascimento, identificado pela vinculação com o SIM e codificado de forma binária (1 = óbito; 0 = sobrevivente). O evento raro correspondeu a 1,78% da coorte (54.693 óbitos). O estudo utilizou exclusivamente dados secundários de acesso público, dispensando submissão ao Comitê de Ética em Pesquisa, conforme Resolução CNS 510/2016. O tratamento dos dados também observou os princípios da Lei Geral de Proteção de Dados (LGPD, Lei nº 13.709/2018), tratando-se de dados abertos e anonimizados disponibilizados pelo DATASUS. Foram utilizadas 26 variáveis de entrada, divididas em dois grupos conforme o tratamento aplicado no pré-processamento. O primeiro grupo compreende 14 variáveis numéricas, inseridas diretamente no modelo sem transformação: idade materna (IDADEMAE), peso ao nascer (PESO), semanas de gestação (SEMAGESTAC), Apgar no 1º e 5º minuto (APGAR1, APGAR5), classificação de prematuridade (TIPO_PREMATURO), tipo de gravidez (GRAVIDEZ), número de consultas pré-natais (CONSULTAS), índice de Kotelchuck (KOTELCHUCK) e histórico reprodutivo (QTDFILVIVO, QTDFILMORT, QTDGESTANT, QTDPARTNOR, QTDPARTCES). Embora GRAVIDEZ, CONSULTAS e KOTELCHUCK possuam categorias ordinais, foram mantidas como numéricas por preservarem relação de ordem cardinal informativa para o modelo. O segundo grupo compreende 12 variáveis categóricas nominais, codificadas por *OHE*: sexo do recém-nascido (SEXO), raça/cor materna (RACACORMAE), tipo de parto (PARTO), local de nascimento (LOCNASC), estado civil materno (ESTCIVMAE), escolaridade materna (ESCMAE), presença de anomalia congênita (IDANOMAL), nascimento em capital (NASC_CAPITAL), código e região da Unidade da Federação (UF) de nascimento (NASC_CODIGO_UF, NASC_REGIAO) e código e região da UF de residência (RES_CODIGO_UF, RES_REGIAO). O mapeamento de categorias do OHE foi ajustado exclusivamente a partir do conjunto de treino de cada janela, garantindo ausência de vazamento de informação entre treino e teste.

2.2 Estratégia de Validação Walk-Forward

A estratégia de validação *walk-forward* anual, ilustrada na Figura 2, consiste em treinar o modelo com todos os anos disponíveis até o N-ésimo ano e avaliá-lo exclusivamente nos dados do ano *N*+1, sem qualquer sobreposição entre treino e teste. Foram geradas seis janelas de validação, com janela mínima de treino de três anos (2014–2016) e janela máxima de oito anos (2014–2021), conforme detalhado na Figura 2. Cada ano de teste nunca integrou o conjunto de treino da mesma janela, garantindo avaliação prospectiva sem vazamento de informação. As janelas 4 e 5 correspondem aos anos pandêmicos (COVID-19).

Figura 2. Validação *walk-forward* anual (2014–2022). Cada janela utiliza treino acumulativo até o *N*-ésimo ano e avalia o modelo no ano *N*+1.



Fonte: Elaborado pelos autores (2026).

Tabela 1. Resultados da validação *walk-forward* anual (2017–2022).

Ano	Quant. de amostras		AUC-ROC	AUPRC	Brier	Recall	Youden	Threshold
	treino	teste						
2017	995.139	335.363	0,922	0,202	0,105	0,900	0,685	0,38
2018	1.330.502	352.268	0,927	0,208	0,104	0,903	0,699	0,41
2019	1.682.770	353.172	0,928	0,206	0,104	0,909	0,704	0,40
2020	2.035.942	341.012	0,929	0,200	0,105	0,916	0,705	0,39
2021	2.376.954	338.970	0,930	0,204	0,104	0,904	0,706	0,42
2022	2.715.924	348.414	0,933	0,195	0,103	0,907	0,719	0,45

Fonte: Elaborado pelos autores (2026).

Em cada janela, o modelo passou por validação cruzada estratificada interna de cinco *folds* (StratifiedKFold, *scikit-learn* 1.3) para monitoramento do desempenho no treino. O modelo final foi retreinado com a totalidade dos dados de treino da janela e avaliado no conjunto de teste do ano subsequente. O *threshold* de classificação foi otimizado pelo índice de Youden (J = Sensibilidade + Especificidade − 1) calculado sobre o conjunto de teste de cada janela. Adotou-se o algoritmo XGBoost com os seguintes hiperparâmetros, mantidos fixos em todas as janelas para isolar o efeito temporal: 800 estimadores, profundidade máxima de 6, taxa de aprendizado de 0,05, *subsample* de 0,9 e *colsample_bytree* de 0,9. O desbalanceamento de classes (1,78% de positivos) foi tratado pelo parâmetro scale_pos_weight, recalculado a cada janela como a razão entre o número de negativos e positivos do respectivo conjunto de treino.

2.3 Avaliação de Drift

Para cada ano de teste, calcularam-se: (i) AUC-ROC, que quantifica a capacidade discriminativa; (ii) AUPRC, adequada para desfechos raros; e (iii) Brier Score, medida de calibração-discriminação combinada. Adotou-se limiar de alerta em AUC-ROC < 0,90 e limiar crítico em AUC-ROC < 0,85. O *data drift* foi avaliado comparando a distribuição de cada variável preditora em cada ano de teste com a distribuição de referência (2014–2016). Para variáveis numéricas, aplicou-se o teste de KS bilateral, reportando a estatística D e o p-valor. Para todas as variáveis, calculou-se o PSI, categorizado em: estável (PSI < 0,10), alerta (0,10 ≤ PSI < 0,25) e crítico (PSI ≥ 0,25). Todas as análises foram conduzidas em Python 3.11 com as bibliotecas *pandas* 2.0, *scikit-learn* 1.3, *xgboost* 2.0 e *scipy* 1.11.

3 RESULTADOS

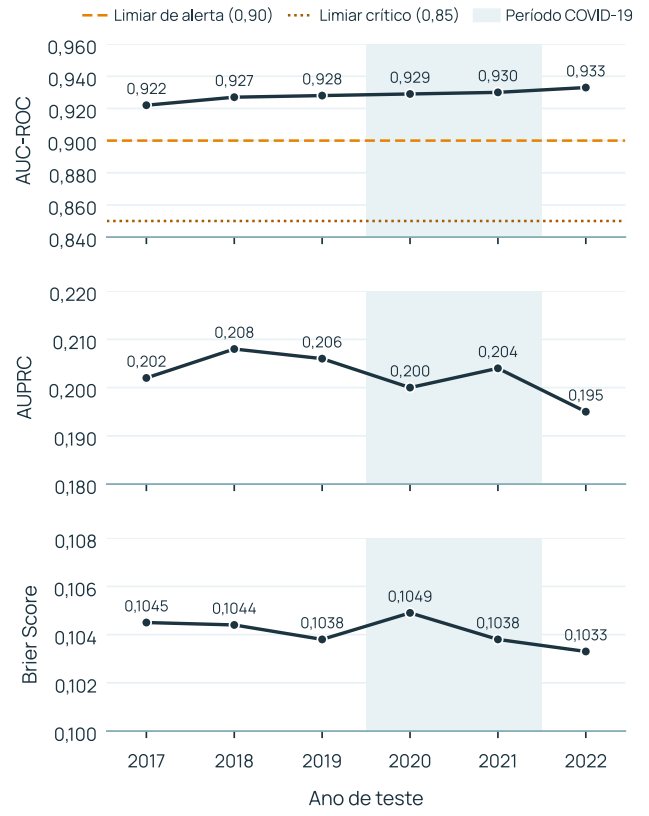
3.1 Características da Coorte

A distribuição anual manteve-se estável ao longo do período, com entre 326.195 (2016) e 353.172 (2019) nascimentos por ano. A taxa de óbito nos conjuntos de teste variou de 1,65% (2022) a 1,90% (2017), evidenciando tendência de redução gradual ao longo do período avaliado. O número de registros por janela de treino cresceu de 995.139 (janela 1, anos 2014–2016) a 2.715.924 (janela 6, anos 2014–2021), refletindo a estratégia de treino acumulativo adotada.

3.2 Performance Drift: Estabilidade do Modelo

Os resultados da validação *walk-forward* são apresentados na Tabela 1. O modelo manteve AUC-ROC entre 0,922 (2017) e 0,933 (2022) em todos os anos de teste, permanecendo consistentemente acima do limiar de alerta estabelecido em 0,90 e do limiar crítico de 0,85. Observou-se tendência de crescimento progressivo do AUC-ROC ao longo do período, acompanhando o aumento do volume de dados de treino, o que sugere que o modelo se beneficia do acúmulo histórico de informação. O AUPRC variou entre 0,195 (2022) e 0,208 (2018), e o Brier Score oscilou entre 0,103 (2022) e 0,105 (2020), ambos refletindo estabilidade na calibração ao longo do tempo. O índice de Youden, utilizado como critério de otimização do *threshold* de classificação, manteve-se entre 0,685 (2017) e 0,719 (2022), com *thresholds* otimizados variando entre 0,38 e 0,45. O Recall médio foi de 0,906, indicando que o modelo identifica corretamente 90% ou mais dos óbitos em todos os anos avaliados. Nenhum dos anos avaliados apresentou degradação clinicamente relevante do desempenho, incluindo os anos pandêmicos de 2020 e 2021. A evolução temporal das métricas de *performance drift* é apresentada na Figura 3.

Figura 3. Evolução das métricas de *performance drift* (2017–2022). AUC-ROC, AUPRC e Brier Score por ano de teste. Linha tracejada indica o limiar de alerta (AUC-ROC = 0,90). Anos pandêmicos destacados em cinza.



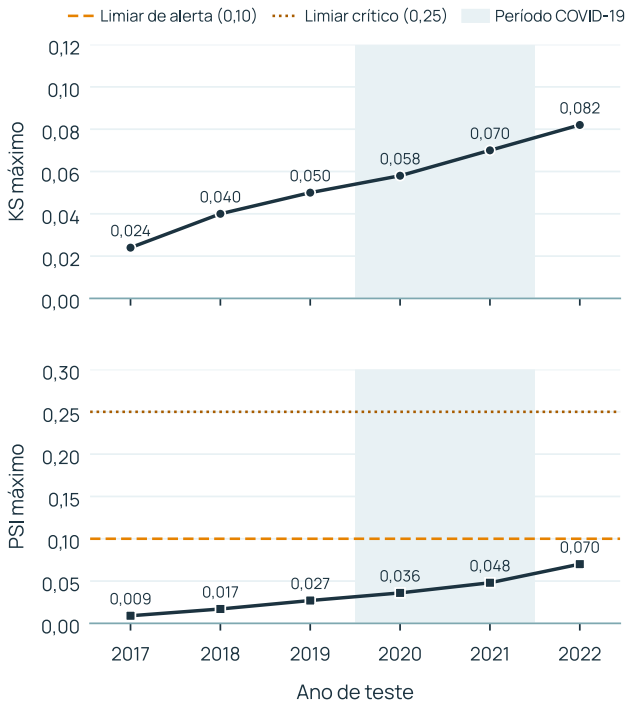
Fonte: Elaborado pelos autores (2026).

3.3 Data Drift: Estabilidade das Variáveis Preditoras

Os resultados da análise de *data drift* são apresentados na Tabela 2 e na Figura 4. Todas as 26 variáveis preditoras permaneceram na categoria *estável* (PSI < 0,10) em todos os anos avaliados. O PSI global médio cresceu de forma gradual e monotônica de 0,0013 (2017) para 0,0130 (2022), refletindo deriva epidemiológica lenta e esperada ao longo de nove anos. Nenhuma variável atingiu os limiares de alerta ou crítico em nenhum dos anos avaliados.

A distribuição individual do *drift* por variável e ano é apresentada nas Figuras 5 e 6. Entre as variáveis numéricas, CONSULTAS apresentou o maior KS em 2022 (0,082), seguida por KOTELCHUCK (0,079) e IDADEMAE (0,077), indicando que o perfil de acesso ao pré-natal e a idade materna foram as dimensões com maior variação temporal. A variável SEMAGESTAC apresentou KS máximo de 0,014, indicando alta estabilidade na distribuição da idade gestacional. Entre as variáveis categóricas, ESCMAE foi a que apresentou maior PSI ao longo do período (0,070 em 2022), refletindo mudanças graduais no perfil de escolaridade materna na população brasileira. O padrão diagonal visível nos *heatmaps* (intensidade crescente da esquerda para a direita) confirma que o *drift* é gradual e acumulativo, sem saltos abruptos associados ao período pandêmico.

Figura 4. Evolução do *data drift* por ano de teste (2017–2022). Superior: estatística KS máxima entre as variáveis numéricas. Inferior: PSI máximo entre todas as variáveis. Linhas tracejadas indicam os limiares de alerta (PSI = 0,10) e crítico (PSI = 0,25).



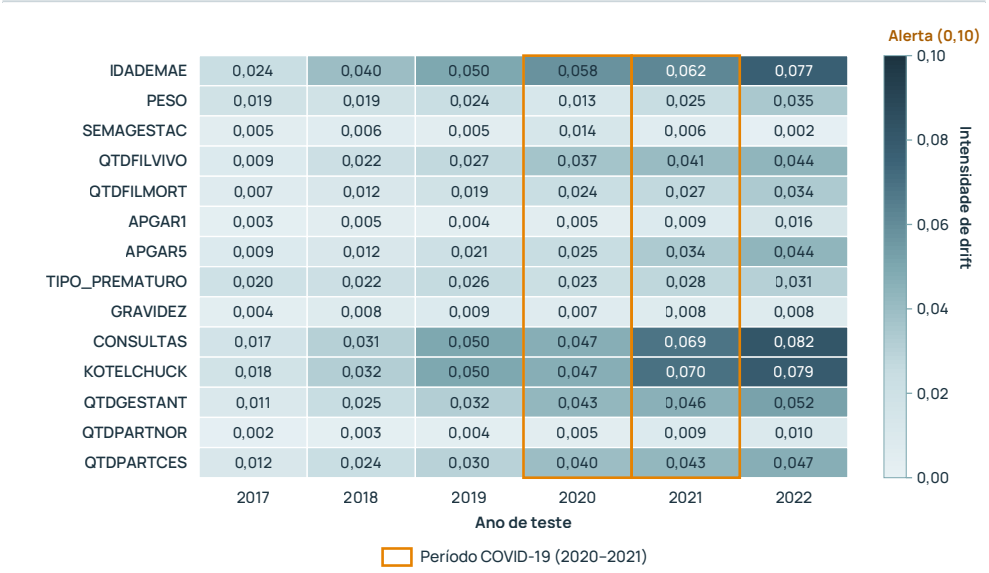
Fonte: Elaborado pelos autores (2026).

Tabela 2. Resumo do *data drift* por ano de teste.

Ano	Vars. KS sig. (p < 0,01)	KS máx.	PSI médio	PSI máx.	Status
2017	12/14	0,024	0,0013	0,009	Estável
2018	13/14	0,040	0,0033	0,017	Estável
2019	14/14	0,050	0,0056	0,027	Estável
2020	14/14	0,058	0,0073	0,036	Estável
2021	14/14	0,070	0,0099	0,048	Estável
2022	13/14	0,082	0,0130	0,070	Estável

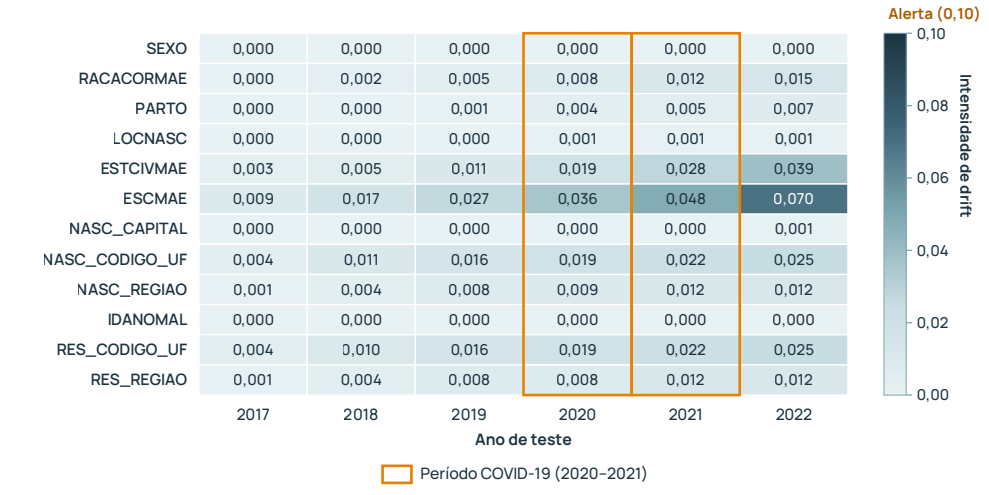
Fonte: Elaborado pelos autores (2026).

Figura 5. Heatmap de data drift: variáveis numéricas (2017–2022). Estatística KS por variável e ano de teste. Intensidade de azul proporcional ao valor de KS. Bordas âmbar destacam os anos pandêmicos (2020–2021). Nenhuma variável ultrapassou o limiar de alerta em qualquer ano.



Fonte: Elaborado pelos autores (2026).

Figura 6. Heatmap de data drift: variáveis categóricas (2017–2022). PSI por variável e ano de teste. Intensidade de azul proporcional ao valor de PSI. Bordas âmbar destacam os anos pandêmicos (2020–2021). Nenhuma variável ultrapassou o limiar de alerta (PSI = 0,10) em qualquer ano.



Fonte: Elaborado pelos autores (2026).

4 DISCUSSÃO

Este estudo demonstrou que um modelo XGBoost para predição de mortalidade infantil em prematuros, treinado com dados nacionais do SINASC e SIM, manteve estabilidade temporal clinicamente robusta ao longo de seis anos de validação prospectiva (2017–2022), incluindo o período da pandemia de COVID-19. O AUC-ROC variou de 0,922 a 0,933, permanecendo consistentemente acima do limiar de alerta preestabelecido (0,90), e nenhuma variável preditora

atingiu o limiar de alerta de *data drift* (PSI = 0,10) em qualquer ano avaliado.

A tendência de crescimento progressivo do AUC-ROC ao longo do período reflete a capacidade do XGBoost de extrair padrões mais robustos à medida que o volume e a diversidade dos dados de treino aumentam com a estratégia de janela acumulativa [12], resultado consistente com estudos de validação temporal de modelos preditivos em saúde perinatal [7]. O Brier Score oscilou em faixa estreita (0,103–0,105), evidenciando estabilidade da calibração, aspecto crítico em contextos clínicos, onde probabilidades mal calibradas comprometem a utilidade do modelo para triagem de risco e alocação de recursos em unidades neonatais [13]. Um dos achados mais relevantes deste estudo é a ausência de degradação clinicamente significativa do desempenho durante os anos pandêmicos (2020–2021), resultado que contrasta com relatos de degradação abrupta de modelos de ML em outros contextos clínicos durante a pandemia, como em modelos de predição de sepse e de readmissão hospitalar [9,11]. A explicação mais plausível reside na natureza das variáveis preditoras: peso ao nascer, idade gestacional e Apgar refletem condições biológicas do recém-nascido relativamente estáveis frente a perturbações sistêmicas de curto prazo. As variáveis que apresentaram maior *data drift* no período pandêmico, CONSULTAS e KOTELCHUCK (KS = 0,047 em 2020), refletem o impacto da COVID-19 sobre o acesso ao pré-natal, fenômeno documentado na literatura brasileira [10]. Contudo, apesar do *drift* nas variáveis de entrada, a relação entre as variáveis preditoras e o desfecho de óbito infantil não foi alterada de forma suficiente para comprometer o desempenho discriminativo do modelo.

O padrão de *data drift* observado (gradual, monotônico e sempre abaixo dos limiares de alerta) é consistente com mudanças epidemiológicas e demográficas de longo prazo na população brasileira. O PSI de ESCMAE (0,070 em 2022) reflete a melhoria progressiva nos níveis de escolaridade materna no Brasil [14], enquanto o *drift* em IDADEMAE (KS = 0,077 em 2022) é consistente com o adiamento da maternidade observado no país. A distinção entre *drift* estatisticamente significativo e *drift* clinicamente relevante é fundamental neste contexto: em uma coorte de mais de 300.000 nascimentos por ano, o teste KS detecta diferenças distribucionais muito pequenas, como evidenciado pelo fato de 12 das 14 variáveis numéricas apresentarem KS significativo já em 2017, com estatística D inferior a 0,025. A métrica PSI, por ser menos sensível ao tamanho amostral e orientada por limiares operacionais estabelecidos [15], oferece avaliação mais adequada da relevância prática do *drift* observado. Os resultados sugerem que a política de retreinamento anual combinada com limiar de alerta em AUC-ROC < 0,90 é estratégia conservadora e suficiente para o contexto avaliado. Contudo, o crescimento monotônico do PSI ao longo do período reforça a necessidade de monitoramento sistemático, pois a extrapolação da tendência sugere que o limiar de alerta poderia ser atingido em horizonte de cinco a dez anos sem retreinamento. A automação deste monitoramento a cada novo ciclo de dados do SINASC e SIM, conforme implementado no sistema Premonitor, permite detecção precoce de degradação com base em evidências objetivas [8]. Como limitações, destacam-se: o uso de dados administrativos sujeitos a variações de qualidade no preenchimento; a validação em escala nacional, que pode não refletir o desempenho em regiões com perfis epidemiológicos heterogêneos; a otimização do *threshold* de classificação (índice de Youden) sobre o próprio conjunto de teste de cada janela, o que pode introduzir viés otimista nas métricas dependentes de *threshold* (Recall, Sensibilidade, Especificidade e F1), embora as métricas independentes de *threshold* (AUC-ROC, AUPRC e Brier Score) permaneçam válidas; e a necessidade de continuidade do monitoramento com dados posteriores a 2022 para confirmar a estabilidade projetada.

5 CONCLUSÕES

Este estudo demonstrou que um modelo XGBoost para predição de mortalidade infantil em prematuros, treinado com dados nacionais do SINASC e SIM, manteve estabilidade temporal clinicamente robusta ao longo de seis anos de validação prospectiva *walk-forward* (2017–2022), com AUC-ROC variando de 0,922 a 0,933 e nenhuma variável preditora atingindo limiar de alerta de *data drift* (PSI < 0,10) em qualquer ano, incluindo os anos pandêmicos de 2020 e 2021. O *drift* epidemiológico observado foi gradual, monotônico e interpretável à luz de tendências demográficas documentadas na população brasileira, sem comprometer a capacidade discriminativa ou a calibração do modelo. Estes resultados indicam que a política de retreinamento anual combinada com monitoramento sistemático de métricas de *drift* é estratégia suficiente e tecnicamente fundamentada para garantir a validade contínua do modelo, com potencial de aplicação no suporte à decisão clínica e à alocação racional de recursos em saúde neonatal no Sistema Único de Saúde.

Contribuição dos autores. Os autores contribuíram igualmente em todas as etapas de construção do artigo: concepção e delineamento do estudo; coleta, análise e interpretação dos dados; redação e revisão crítica do manuscrito; e aprovação da versão final.

Agradecimentos. Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Financiamento. Este trabalho recebeu apoio da Rede SOFIA, financiada pela RNP/MCTI, no âmbito do Programa Redes de Colaboração em Saúde Digital (PRC-SD).

Disponibilidade de dados e materiais. Os autores não declararam a disponibilidade dos dados e materiais da pesquisa.

Conflitos de interesse. Os autores não declararam conflitos de interesse.

Declaração de uso de IA. Os autores declaram o uso de ferramentas de IA no apoio e desenvolvimento dos scripts de análise. A responsabilidade pelo conteúdo científico é exclusiva dos autores.

REFERÊNCIAS

1. Brasil. Ministério da Saúde. Sistema de Informações sobre Nascidos Vivos (SINASC) e Sistema de Informações sobre Mortalidade (SIM): dados de 2014 a 2022 [Internet]. Brasília: DATASUS; 2022 [citado 2026 out 15]. Disponível em: https://datasus.saude.gov.br
2. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577-9. doi: https://doi.org/10.1016/S0140-6736(19)30037-6
3. Sullivan BA, Beam K, Vesoulis ZA, Aziz KB, Husain AN, Knake LA, et al. Transforming neonatal care with artificial intelligence: challenges, ethical consideration, and opportunities. J Perinatol. 2024;44(1):1-11. doi: https://doi.org/10.1038/s41372-023-01848-5
4. Lopes MLB, Barbosa RM, Fernandes MAC. Unsupervised Learning Applied to the Stratification of Preterm Birth Risk in Brazil with Socioeconomic Data. Int J Environ Res Public Health. 2022;19(9):5596. doi: https://doi.org/10.3390/ijerph19095596
5. Lee J, Cai J, Li F, Vesoulis ZA. Predicting mortality risk for preterm infants using random forest. Sci Rep. 2021;11(1):7308. doi: https://doi.org/10.1038/s41598-021-86748-4
6. Oliveira YSP, Freitas GR, Barros WKP, Souza LC, Azevedo KS, Barbosa RM, et al. Aprendizagem de máquina aplicada à estratificação de risco de mortalidade de recém-nascidos associados a partos prematuros. In: Anais do XVI Congresso Brasileiro de Inteligência Computacional (CBIC 2023); Salvador, BA. Salvador: SBIC; 2023. p. 1-8. doi: https://doi.org/10.21528/CBIC2023-150
7. Mangold C, Zoretic S, Thallapureddy K, Moreira A, Chorath K, Moreira A. Machine Learning Models for Predicting Neonatal Mortality: A Systematic Review. Neonatology. 2021;118(4):394-405. doi: https://doi.org/10.1159/000516891
8. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv. 2014;46(4):1-37. doi: https://doi.org/10.1145/2523813
9. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021;181(8):1065-70. doi: https://doi.org/10.1001/jamainternmed.2021.2626
10. Carvalho-Sauer R, Costa MCN, Teixeira MG, Flores-Ortiz R, Leal JTFM, Saavedra R, et al. Maternal and perinatal health indicators in Brazil over a decade: assessing the impact of the COVID-19 pandemic and SARS-CoV-2 vaccination through interrupted time series analysis. Lancet Reg Health Am. 2024;35:100774. doi: https://doi.org/10.1016/j.lana.2024.100774
11. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In: Proceedings of the 4th Machine Learning for Healthcare Conference (PMLR 106); 2019 ago 9-10. [s.l.]: PMLR; 2019. p. 381-405. [citado 2026 out 15]. Disponível em: https://proceedings.mlr.press/v106/nestor19a.html
12. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16); San Francisco, CA, EUA. New York: ACM; 2016. p. 785-94. doi: https://doi.org/10.1145/2939672.2939785
13. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2. ed. Cham: Springer; 2019. doi: https://doi.org/10.1007/978-3-030-16399-0
14. Instituto Brasileiro de Geografia e Estatística (IBGE). Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua): divulgação trimestral [Internet]. Rio de Janeiro: IBGE; 2022 [citado 2026 out 15]. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html
15. Naeem Siddiqi. Intelligent credit scoring: Building and implementing better credit riskscore-cards. John Wiley & Sons, 2017.

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.

Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.

SBIS Sociedade Brasileira de Informática em Saúde