Journal of Health Informatics

J Health Inform. 2026;18:e1831
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1831

ARTIGO COMPLETO
CBIS 2026

Inteligência artificial para predição de mortalidade infantil em prematuros: recuperação de casos similares com dados nacionais (2014–2022)

Artificial intelligence for predicting infant mortality in preterm infants: retrieval of similar cases with national data (2014–2022)

Karolayne Santos de Azevedo[1], Luísa Christina de Souza[1], Matheus Dalmolin[1], Marcelo Augusto Costa Fernandes[2] ✉

ORCID: 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), Natal (RN), Brasil; Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal (RN), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1831

DOI
10.59681/2175-4411.v18.iEspecial.2026.1831

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: propor uma estratégia de predição de mortalidade infantil em prematuros que combina aprendizado de máquina com recuperação de casos históricos similares, visando maior explicabilidade clínica. **Método:** registros vinculados dos sistemas SIM e SINASC (2014–2022, n=3.064.338) foram utilizados para treinar um modelo de transformer com aprendizado contrastivo, capaz de representar cada recém-nascido em um espaço vetorial orientado ao desfecho. Para cada novo caso, os registros históricos mais *similares são recuperados e suas características clínicas agregadas enriquecem o modelo preditivo XGBoost com validação cruzada de cinco folds.* ***Resultados:*** *o modelo enriquecido alcançou AUC-ROC de 0,9267, com os 5% de maior risco concentrando 54,2% dos óbitos (Lift = 10,83×). As características dos casos recuperados contribuíram com 43% da importância preditiva total.* ***Conclusão:*** *a estratégia preserva o desempenho do modelo base e permite compreender o risco por analogia com casos reais anteriores, operando exclusivamente* sobre dados já disponíveis no SUS.

Descritores: Mortalidade Infantil; Recém-Nascido Prematuro; Inteligência Artificial; recuperação de casos.

ABSTRACT

Objective: to propose a mortality prediction strategy for preterm infants that combines machine learning with historical similar-case retrieval, aiming at greater clinical explainability. **Method:** linked SIM-SINASC records (2014–2022, n=3,064,338) were used to train a transformer model with contrastive learning, representing each newborn in an outcome-oriented vector space. For each new case, the most similar historical records are retrieved and their clinical features are aggregated to enrich an XGBoost predictive model trained *with five-fold cross-validation.* ***Results:*** *the enriched model achieved AUC-ROC of 0.9267, with the top 5% highest-risk individuals concentrating 54.2% of deaths (Lift = 10.83×). Retrieved case features contributed 43% of total predictive importance.* ***Conclusion:*** *the proposed strategy preserves baseline predictive performance and enables health professionals to understand risk by analogy with real previous cases, operating exclusively on data already available in the Brazilian Unified Health System (SUS).*

Keywords: Infant Mortality; Infant; Premature; Artificial Intelligence; case retrieval.

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: de Azevedo KS, de Souza LC, Dalmolin M, Fernandes MAC. Inteligência artificial para predição de mortalidade infantil em prematuros: recuperação de casos similares com dados nacionais (2014–2022). J Health Inform 2026;18:e1831. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1831

1 INTRODUÇÃO

A mortalidade em prematuros representa um dos principais desafios da saúde pública brasileira. Entre 2014 e 2022, mais de 3 milhões de nascimentos prematuros foram registrados no país, com 54.693 óbitos infantis associados, correspondendo a uma prevalência de 1,78% [1]. Desse total, 74% dos óbitos são atribuídos a causas evitáveis, o que evidencia o potencial de impacto de intervenções preventivas orientadas por informação [2]. A desigualdade regional no acesso a cuidados neonatais intensivos agrava o cenário, com taxas de mortalidade significativamente superiores nas regiões Norte e Nordeste em comparação ao Sul e Sudeste [3]. O Sistema Único de Saúde (SUS) coleta, de forma rotineira, dados clínicos e demográficos detalhados por meio do Sistema de Informações sobre Nascidos Vivos (SINASC) e do Sistema de Informações sobre Mortalidade (SIM), constituindo uma base de dados de escala nacional ainda subutilizada para fins preditivos e de vigilância ativa.

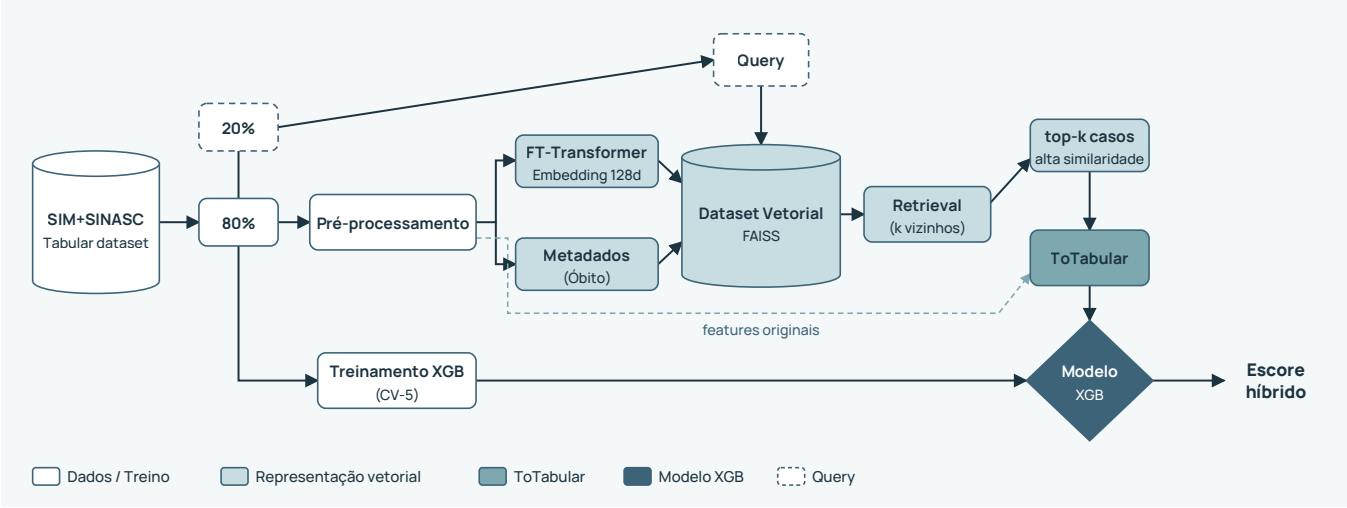
Nas últimas décadas, modelos de aprendizado de máquina têm sido crescentemente aplicados à predição de desfechos perinatais adversos. Estudos recentes demonstraram que algoritmos de *gradient boosting*, em particular o XGBoost, alcançam desempenho discriminativo superior em dados tabulares estruturados do SUS, com área sob a curva ROC (AUC-ROC) acima de 0,90 em cenários de evento raro [4,5]. A vinculação dos registros SINASC e SIM possibilita a construção de coortes populacionais de grande escala com desfecho definido, superando limitações de estudos baseados em amostras hospitalares [6]. Abordagens de otimização de limiar orientadas à sensibilidade clínica constituem estratégia relevante para reduzir o número de casos a monitorar sem perda expressiva de sensibilidade, aspecto crítico para a viabilidade operacional no contexto do SUS [4,5].

Paralelamente, avanços em representação de dados tabulares por redes neurais têm aberto novas possibilidades para a modelagem de desfechos clínicos. O Feature Tokenizer Transformer (FT-Transformer) projeta cada variável em um espaço latente de alta dimensão e aplica mecanismos de autoatenção para capturar interações não lineares entre *features*, superando modelos baseados em árvores em *benchmarks* de classificação tabular [7]. Estratégias de aprendizado contrastivo, como a *Triplet Loss*, permitem orientar o espaço de representação ao desfecho clínico, aproximando casos com mesmo resultado e afastando casos com resultados distintos [8]. A recuperação de casos por similaridade vetorial, viabilizada por ferramentas como o FAISS [9], tornou escalável a busca em bases de milhões de registros com latência de milissegundos. Esses elementos, combinados, aproximam-se do paradigma de recuperação aumentada aplicado à predição tabular, no qual informações recuperadas de casos históricos similares enriquecem o modelo preditivo. Diferentemente do paradigma de *Retrieval-Augmented Generation* (RAG) clássico, baseado em geração textual, a abordagem proposta utiliza a recuperação vetorial como mecanismo de enriquecimento preditivo e explicabilidade por analogia.

Apesar dos avanços descritos, persiste uma lacuna relevante na literatura de predição de mortalidade em prematuros: os modelos existentes produzem escores de probabilidade pontual sem oferecer ao profissional de saúde uma justificativa compreensível para a decisão. A explicabilidade baseada em valores SHAP, embora tecnicamente robusta, exige familiaridade com conceitos de aprendizado de máquina e apresenta limitações para comunicação em contextos assistenciais [10]. Adicionalmente, nenhum trabalho anterior sobre mortalidade em prematuros com dados SIM-SINASC explorou a recuperação de casos históricos similares como mecanismo de enriquecimento preditivo e explicativo, nem utilizou *transformers* com aprendizado contrastivo para geração de *embeddings* tabulares orientados ao desfecho neonatal. Essa lacuna é clinicamente relevante: profissionais de saúde raciocinam naturalmente por analogia com casos anteriores, e sistemas que apoiem esse raciocínio têm maior potencial de adoção na prática clínica [11].

Este trabalho propõe uma estratégia de predição de mortalidade em prematuros que combina um *transformer* com aprendizado contrastivo para representação vetorial dos recém-nascidos, busca por similaridade sobre 2,4 milhões de registros históricos e enriquecimento do modelo XGBoost com características clínicas agregadas dos casos mais similares recuperados. A abordagem é avaliada sobre a coorte nacional SIM-SINASC (2014–2022, n=3.064.338) e comparada ao modelo XGBoost base. Os resultados demonstram que o modelo enriquecido preserva o desempenho discriminativo do *baseline* (AUC-ROC = 0,9267 versus 0,9280), com as características dos casos recuperados contribuindo com 43% da importância preditiva total, e adiciona uma dimensão de explicabilidade clínica concreta: para cada recém-nascido de alto risco, o sistema identifica os casos históricos mais similares e informa quantos foram a óbito, fundamentando a predição em evidência histórica análoga diretamente interpretável pela equipe de saúde.

Figura 1. Pipeline da estratégia proposta. Os registros SIM-SINASC são divididos em 80% para indexação e treinamento e 20% para avaliação. O FT-Transformer gera *embeddings* orientados ao desfecho via *Triplet Loss*. O FAISS recupera os *k* casos mais similares, cujas *features* originais são agregadas (ToTabular) e concatenadas às *features* do caso consultado para treinar o XGBoost enriquecido.



Fonte: Elaborado pelos autores (2026).

2 PROPOSTA

A estratégia proposta fundamenta-se na ideia de que a predição de mortalidade em prematuros pode ser enriquecida com o raciocínio por analogia: em vez de produzir apenas uma probabilidade pontual para um novo caso, o sistema recupera registros históricos de bebês prematuros com perfil clínico semelhante e utiliza o desfecho e as características desses casos como evidência adicional para o modelo preditivo. A Figura 1 ilustra o *pipeline* completo da abordagem, composto por quatro etapas principais: representação vetorial dos registros, indexação e recuperação por similaridade, enriquecimento tabular e predição com XGBoost.

A primeira etapa consiste na construção de uma representação vetorial de cada recém-nascido prematuro a partir das variáveis clínicas, obstétricas e demográficas disponíveis nos sistemas SIM e SINASC. Para isso, utiliza-se um FT-Transformer, arquitetura que projeta cada variável em um vetor de dimensão fixa e aplica mecanismos de autoatenção para capturar interações entre as variáveis clínicas. Diferentemente de abordagens tradicionais de codificação tabular, o FT-Transformer permite que o peso ao nascer, por exemplo, influencie a representação das semanas de gestação e do índice de Apgar de forma não linear e adaptativa. O modelo é treinado com *Triplet Loss* contrastiva: para cada registro âncora, seleciona-se um caso positivo com o mesmo desfecho (óbito ou sobrevivência) e um caso negativo com desfecho oposto, de modo que o espaço vetorial aprendido agrupe casos de mesmo desfecho e separe casos de desfechos distintos. O resultado é um vetor de 128 dimensões por recém-nascido, normalizado e orientado ao desfecho clínico.

Na segunda etapa, os vetores dos 80% de registros destinados ao treinamento são indexados em uma estrutura de busca vetorial eficiente utilizando a biblioteca FAISS, que viabiliza a recuperação dos *k* casos mais similares entre milhões de registros com latência de milissegundos. Para cada novo caso do conjunto de avaliação, o vetor correspondente é gerado e consultado contra o índice, retornando os *k* vizinhos com maior similaridade pelo cosseno. Este processo é equivalente à pergunta clínica: “quais prematuros registrados no passado mais se assemelham a este bebê?”. Para evitar vazamento de informação durante o treinamento do modelo preditivo, cada registro do conjunto de treino busca seus vizinhos excluindo a si mesmo do índice.

A terceira etapa, denominada ToTabular, converte os *k* casos recuperados em novas variáveis preditivas. Para cada vizinho, as *features* tabulares originais são recuperadas diretamente do *dataset* histórico. Em seguida, estatísticas de agregação são calculadas sobre os *k* casos: média, desvio padrão, mínimo e máximo de cada variável clínica, além da taxa de óbito e da similaridade média entre o caso consultado e seus vizinhos. Essas estatísticas formam um vetor de contexto histórico que é concatenado às *features* originais do caso consultado, resultando em um conjunto de variáveis enriquecido que combina as características do próprio recém-nascido com informação derivada de casos análogos do passado.

Na quarta e última etapa, o conjunto de variáveis enriquecido alimenta um modelo XGBoost treinado com validação cruzada de cinco *folds*. A vantagem dessa abordagem em relação à simples combinação linear de escores é que o XGBoost aprende automaticamente o peso e as interações entre as *features* originais e as *features* de contexto histórico, sem necessidade de calibração manual de parâmetros de ponderação. O modelo resultante produz uma probabilidade de óbito que incorpora tanto as características individuais do recém-nascido quanto a experiência histórica de casos

similares. Para fins de explicabilidade, o sistema disponibiliza ao profissional de saúde não apenas o escore de risco, mas também a lista dos casos históricos mais similares com seus respectivos desfechos, permitindo fundamentar a predição em evidência concreta e compreensível.

3 MÉTODOS

3.1 Base de Dados e Divisão dos Dados

A coorte analisada foi construída a partir da vinculação estruturada dos registros do SINASC e do SIM referentes ao período de 2014 a 2022, resultando em 3.064.338 registros de nascidos vivos prematuros no Brasil. O desfecho de interesse, denominado Óbito, foi definido como a ocorrência de óbito no primeiro ano de vida, isto é, até 365 dias após o nascimento, entre recém-nascidos prematuros registrados no SINASC e vinculados ao SIM. A coorte apresentou proporção de óbitos de 1,78% (54.693 casos), caracterizando um cenário de evento raro e forte desbalanceamento de classes. As variáveis utilizadas abrangem quatro dimensões: demográfica materna (idade, raça/cor, estado civil, escolaridade), obstétrica (tipo de gravidez, número de consultas pré-natais, tipo de parto, histórico reprodutivo), neonatal (peso ao nascer, semanas de gestação, índices de Apgar, presença de anomalia congênita) e geográfico-contextual (região e unidade federativa de nascimento e residência). A variável Ano de Nascimento foi utilizada exclusivamente para caracterização temporal da coorte, não sendo incluída como variável preditora, a fim de evitar vazamento de informação. Os registros foram divididos em dois conjuntos por meio de amostragem aleatória estratificada pelo desfecho, com proporção de 80% para treinamento e indexação e 20% para avaliação. O conjunto de treinamento, composto por aproximadamente 2,4 milhões de registros, foi utilizado tanto para o treinamento do modelo de representação vetorial quanto para a construção do índice de busca por similaridade e o treinamento do modelo XGBoost enriquecido. O conjunto de avaliação, com aproximadamente 612 mil registros, foi utilizado exclusivamente para mensuração do desempenho dos modelos, sem qualquer participação nas etapas de ajuste ou indexação.

3.2 Representação Vetorial, Indexação e Recuperação por Similaridade

Cada recém-nascido prematuro foi representado por um vetor de 128 dimensões gerado por um FT-Transformer [7]. O modelo projeta cada variável de entrada em um vetor de dimensão $d = 128$ por meio de uma camada de tokenização, aplica dois blocos de *transformer* com quatro cabeças de autoatenção e dimensão interna de 256, e extrai o *token* de classe [CLS] como representação final do recém-nascido. As variáveis numéricas foram padronizadas por *z-score* e as variáveis categóricas foram codificadas por tabelas de *embedding* aprendidas. O modelo foi treinado com *Triplet Loss* contrastiva [8]: para cada registro âncora, selecionou-se aleatoriamente um caso positivo com o mesmo desfecho e um caso negativo com desfecho oposto, com margem $m = 0{,}5$. O treinamento foi conduzido por 20 épocas com otimizador AdamW, taxa de aprendizado de 10^{-4} e agendamento cosseno, utilizando *batches* de 512 triplas. Os vetores resultantes foram normalizados pela norma $L2$, de modo que a similaridade do cosseno entre dois vetores corresponde ao produto interno entre eles. Os vetores dos registros de treinamento foram então indexados em uma estrutura de busca exata por produto interno utilizando a biblioteca FAISS [9] com índice IndexFlatIP. Para cada caso do conjunto de avaliação, o vetor correspondente foi gerado pelo FT-Transformer e consultado contra o índice, retornando os k registros com maior similaridade do cosseno. Foram avaliados quatro valores de $k \in \{5, 10, 20, 30\}$. Para evitar vazamento de informação durante a fase de treinamento do XGBoost, cada registro do conjunto de treino buscou seus vizinhos com exclusão do próprio registro (*exclude-self*), garantindo que nenhum caso consultasse a si mesmo como vizinho.

3.3 Enriquecimento Tabular, Modelo Preditivo e Métricas de Avaliação

Para cada caso consultado, as *features* tabulares originais dos k vizinhos recuperados foram obtidas por busca direta no *dataset* de treinamento, utilizando o identificador único de cada registro como chave de acesso. Em seguida, quatro estatísticas de agregação foram calculadas sobre os k casos para cada variável clínica numérica: média, desvio padrão, mínimo e máximo. Adicionalmente, foram calculadas a taxa de óbito entre os vizinhos (viz_obito_mean), a contagem absoluta de óbitos (viz_obito_sum) e estatísticas de similaridade (média, máximo, mínimo e desvio padrão). O conjunto de variáveis agregadas, denominado vetor de contexto histórico, foi concatenado às *features* originais do caso consultado, resultando em um vetor enriquecido com 73 dimensões no total, 26 *features* originais codificadas e 47 *features* derivadas dos vizinhos. Esse vetor enriquecido alimentou um modelo XGBoost [12] configurado com até 800 árvores, profundidade máxima de seis, taxa de aprendizado de 0,05, subamostragem estocástica de instâncias e variáveis de 80% e regularização L2, treinado com validação cruzada estratificada de cinco *folds*. Para fins de comparação, um modelo XGBoost base foi treinado nas mesmas condições utilizando apenas as 26 *features*

Tabela 1. Comparação entre XGBoost base e modelo enriquecido (k = 10, configuração de referência por maior AUPRC). Critério operacional: precisão mínima de 20%.

Modelo	AUC-ROC	AUPRC	Recall@5%	Lift@5%	MCC	Recall (P≥20%)
XGBoost base	0,9280	0,2017	0,5445	10,89×	0,3034	0,5165
Enriquecido (*k* = 10)	0,9267	0,2010	0,5417	10,83×	0,3025	0,5138

Fonte: Elaborado pelos autores (2026).

Tabela 2. Métricas do modelo enriquecido por valor de k. A última linha apresenta o XGBoost base sem enriquecimento.

k	AUC-ROC	AUPRC	Recall@5%	Lift@5%
5	0,9268	0,1996	0,5426	10,85×
10	0,9267	0,2010	0,5417	10,83×
20	0,9267	0,1995	0,5394	10,79×
30	0,9269	0,1989	0,5423	10,85×
Base	0,9280	0,2017	0,5445	10,89×

Fonte: Elaborado pelos autores (2026).

originais codificadas, sem o enriquecimento pelos vizinhos. A análise de importância de *features* foi conduzida pelo ganho médio de cada variável nas divisões das árvores do modelo enriquecido. Dado o forte desbalanceamento de classes (prevalência de 1,78%), foram priorizadas métricas sensíveis a eventos raros: AUC-ROC para discriminação global, área sob a curva *Precision-Recall* (AUPRC) para equilíbrio entre precisão e sensibilidade, Recall@5% e Lift@5% para avaliação da capacidade de priorização, e coeficiente de correlação de Matthews (MCC) com limiar operacional definido pelo maior *recall* com precisão mínima de 20%, critério alinhado à viabilidade assistencial no contexto do SUS [4,13].

4 RESULTADOS

O modelo XGBoost base alcançou AUC-ROC de 0,9280 e AUPRC de 0,2017 no conjunto de teste, com variância entre os *folds* de validação cruzada de 0,9276 ± 0,0009. No critério operacional de precisão mínima de 20%, o modelo identificou 51,7% dos óbitos sinalizando 5% da população, com MCC de 0,303 e Lift de 10,89× em relação à prevalência basal de 1,78%. Esses resultados são consistentes com os reportados na literatura para modelos de predição de mortalidade em prematuros sobre dados SIM-SINASC [4,5].

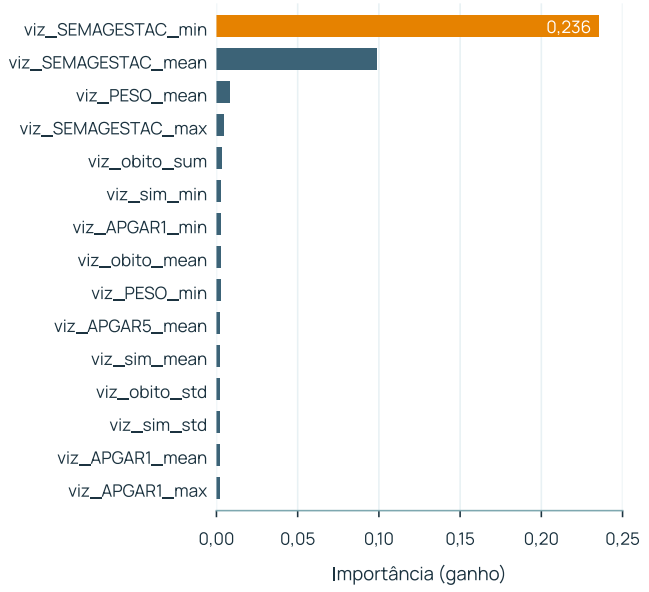
A Tabela 1 apresenta a comparação entre o XGBoost base e o modelo enriquecido com $k = 10$. O modelo enriquecido alcançou AUC-ROC de 0,9267 e AUPRC de 0,2010, com diferença absoluta de 0,0013 em AUC-ROC e 0,0007 em AUPRC em relação ao *baseline*. No critério operacional, o modelo enriquecido identificou 51,4% dos óbitos com precisão de 20,1% e MCC de 0,303, com Lift de 10,83× nos 5% de maior risco.

A Tabela 2 apresenta as métricas do modelo enriquecido para $k \in \{5, 10, 20, 30\}$. A AUC-ROC oscila entre 0,9267 e 0,9269, a AUPRC entre 0,1989 e 0,2010, e o Recall@5% entre 0,539 e 0,543. O valor $k = 10$ apresentou a maior AUPRC (0,2010) e foi adotado como configuração de referência operacional por equilibrar qualidade preditiva e custo computacional. As curvas de concentração de risco e Lift são apresentadas para $k = 30$, configuração de maior AUC-ROC (0,9269), cujos valores são representativos do comportamento geral do modelo dado que a variação entre valores de k é inferior a 0,001 em AUC-ROC, de 0,0021 em AUPRC e de 0,0032 em Recall@5%.

As 47 variáveis geradas pelo enriquecimento tabular (prefixo viz_) contribuíram com 43% da importância total do modelo enriquecido, enquanto as *features* originais codificadas contribuíram com os demais 57%. A Figura 2 apresenta as quinze *features* de maior ganho entre as variáveis viz_*. A *feature* viz_SEMAGESTAC_min, que representa a menor semana gestacional entre os k vizinhos recuperados, obteve ganho de 0,236, sendo superada apenas pela SEMAGESTAC do caso consultado (ganho = 0,327). As *features* viz_obito_sum e viz_obito_mean, que representam a contagem e a taxa de óbito entre os vizinhos, figuram entre as dez *features* de maior importância do modelo.

As curvas *Precision-Recall* dos dois modelos se sobrepõem (XGBoost base: AUPRC = 0,2017; enriquecido: AUPRC = 0,1989), conforme apresentado na Figura 3. A curva Recall@k (Figura 4) indica que os 5% da população com maior escore concentram 54,2% dos óbitos e os 20% concentram 88,7% dos eventos em ambos os modelos. A curva de Lift (Figura 5) registra enriquecimento de 10,85× nos 5% iniciais e de 4,44× nos 20% iniciais para o modelo enriquecido, valores próximos ao *baseline*.

Figura 2. Importância das 15 *features* de maior ganho geradas pelo enriquecimento tabular (viz_*). viz_SEMAGESTAC_min obteve ganho de 0,236, superada apenas pela SEMAGESTAC do próprio caso (0,327).

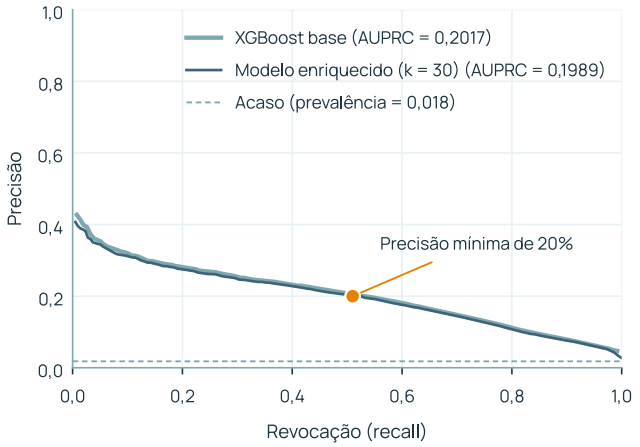


Fonte: Elaborado pelos autores (2026).

5 ANÁLISE DOS DADOS

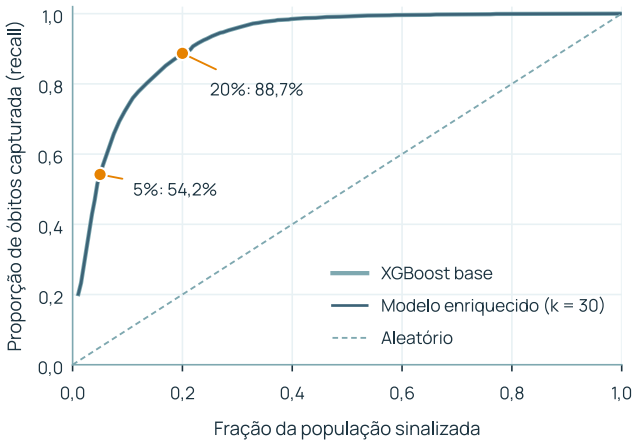
A diferença de 0,0013 em AUC-ROC entre o XGBoost base (0,9280) e o modelo enriquecido (0,9267) não representa inferioridade preditiva. Em amostras da magnitude utilizada (aproximadamente 612 mil casos de teste), o teste de DeLong [14] detecta diferenças estatisticamente significativas de AUC da ordem de 0,001, ainda que sem relevância clínica ou operacional [13]. A interpretação adequada é de equivalência prática: o enriquecimento com contexto histórico preserva a capacidade discriminativa do modelo base e adiciona uma dimensão de explicabilidade que o *baseline* não oferece.

Figura 3. Curvas *Precision-Recall* do XGBoost base e do modelo enriquecido (k = 30, configuração de maior AUC-ROC). O ponto marcado corresponde ao critério operacional de precisão mínima de 20%.



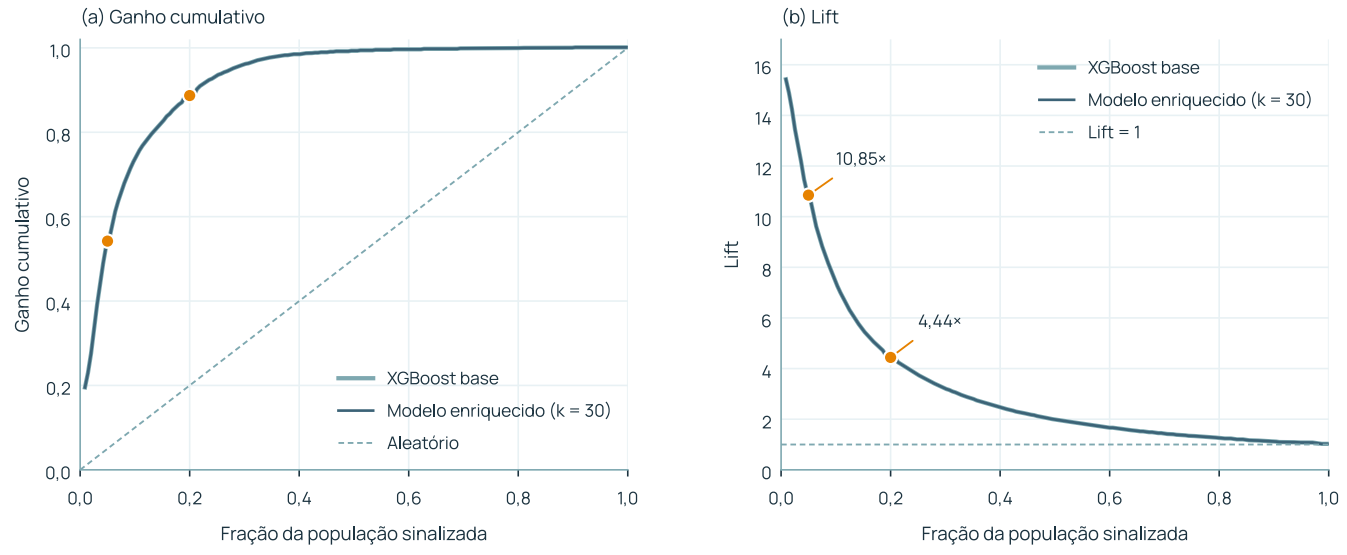
Fonte: Elaborado pelos autores (2026).

Figura 4. Curva Recall@k: proporção de óbitos capturada em função da fração da população sinalizada. Os 5% de maior risco concentram 54,2% dos óbitos; os 20% concentram 88,7%.



Fonte: Elaborado pelos autores (2026).

Figura 5. Curvas de Ganho Cumulativo e Lift. O enriquecimento de 10,85× nos 5% iniciais indica que a taxa de óbito nesse subgrupo é onze vezes superior à prevalência basal (1,78%).



Fonte: Elaborado pelos autores (2026).

As 47 variáveis geradas pelo enriquecimento tabular contribuíram com 43% da importância total do modelo. Se o contexto histórico fosse redundante com as *features* originais, o XGBoost atribuiria importância marginal a essas variáveis por sua natureza seletiva na escolha de divisões. A contribuição de 43% indica que as *features* dos vizinhos capturam informação não contida nas características individuais do recém-nascido, sustentando a premissa da abordagem de que casos históricos similares carregam sinal preditivo independente sobre o desfecho de novos casos.

A *feature* viz_SEMAGESTAC_min representa a menor semana gestacional entre os k vizinhos recuperados, ou seja, o grau máximo de prematuridade observado em casos similares ao caso consultado. Sua importância (ganho = 0,236) indica que a proximidade de um recém-nascido a casos de prematuridade extrema no espaço de representação vetorial constitui sinal de risco, mesmo que o próprio caso apresente semanas gestacionais maiores. Esse resultado aponta para a heterogeneidade dentro de faixas de prematuridade: combinações de variáveis que a semana gestacional isolada não discrimina são capturadas pela estrutura de vizinhança. A presença de viz_obito_mean e viz_obito_sum entre as dez *features* de maior importância valida a racionalidade do método: a taxa de óbito histórica entre os vizinhos constitui uma estimativa do risco baseada em casos reais, consistente com a literatura de raciocínio baseado em casos [11].

A variação máxima de 0,0002 em AUC-ROC e de 0,0032 em Recall@5% entre $k = 5$ e $k = 30$ indica que o método não depende da calibração de k dentro da faixa avaliada. Trabalhos anteriores sobre mortalidade em prematuros com dados SIM-SINASC produziram modelos com capacidade discriminativa global, mas restritos à produção de escores pontuais [4,5]. A abordagem proposta mantém esse desempenho e introduz um mecanismo de explicabilidade por analogia: a apresentação dos casos históricos mais similares e seus desfechos transforma o escore de risco em evidência fundamentada em registros reais, com potencial de aumentar a adoção do sistema na prática assistencial.

6 CONCLUSÕES

Este trabalho propôs uma estratégia de predição de mortalidade em prematuros que combina um *transformer* com aprendizado contrastivo, busca vetorial por similaridade e enriquecimento tabular com características de casos históricos recuperados. Avaliada sobre a coorte nacional SIM-SINASC (2014–2022, n=3.064.338), a abordagem alcançou AUC-ROC de 0,9267 com desempenho equivalente ao modelo base (0,9280), com os 5% de maior risco concentrando 54,2% dos óbitos (Lift = 10,83×). As variáveis derivadas dos casos recuperados contribuíram com 43% da importância preditiva total, com viz_SEMAGESTAC_min como segunda *feature* mais importante (ganho = 0,236), demonstrando que o contexto histórico de casos similares constitui sinal preditivo independente das características individuais do recém-nascido.

A contribuição central da abordagem é a explicabilidade que ela adiciona: para cada caso de alto risco, o sistema apresenta os registros históricos mais similares e seus desfechos, permitindo que o profissional de saúde compreenda a predição por analogia com casos reais, sem necessidade de familiaridade com aprendizado de máquina. O método opera exclusivamente sobre dados já coletados pelo SUS e integra-se ao ecossistema Premonitor como módulo de suporte à decisão em vigilância neonatal. Como trabalho futuro, propõe-se a validação da explicabilidade por analogia com profissionais de saúde em contexto assistencial real.

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. Ministério da Saúde (BR). Sistema de informações sobre nascidos vivos (SINASC) e sistema de informações sobre mortalidade (SIM) [bases de dados] [Internet]. Brasília: Ministério da Saúde; 2022 [citado 2026 out 18]. Disponível em: https://datasus.saude.gov.br

2. Comissão Nacional sobre Determinantes Sociais da Saúde. Análise da situação de saúde. In: As causas sociais das iniqüidades em saúde no Brasil. [s.l.]: Editora FIOCRUZ; 2008. p. 23-130. doi: https://doi.org/10.7476/9788575415917.0004

3. Maternal mortality rates.

4. Victor A, Almeida F, Xavier SP, Rondó PHC. Predicting low birth weight risks in pregnant women in Brazil using machine learning algorithms: data from the Araraquara cohort study. BMC Pregnancy Childbirth. 2025;25(1):320. doi: https://doi.org/10.1186/s12884-025-07351-3

5. Silva AB, Rocha EDS, Lorenzato JF, Endo PT. Evaluating how different balancing data techniques impact on prediction of premature birth using machine learning models. PLoS One. 2025;20(3):e0316574. doi: https://doi.org/10.1371/journal.pone.0316574

6. Rocha AS, de Cássia Ribeiro-Silva R, Fiaccone RL, Paixao ES, Falcão IR, Alves FJO, et al. Differences in risk factors for incident and recurrent preterm birth: a population-based linkage of 3.5 million births from the CIDACS birth cohort. BMC Med. 2022;20(1):111. doi: https://doi.org/10.1186/s12916-022-02313-4

7. Cappuzzo R. Deep learning models for tabular data curation. [s.l.]: Agence Bibliographique de l'Enseignement Supérieur; 2022. doi: https://doi.org/10.70675/d83bad15z52faz47ddz9b3czd25cd9ab0f08

8. Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [s.l.]: IEEE; 2015. p. 815-23. doi: https://doi.org/10.1109/cvpr.2015.7298682

9. Johnson J, Douze M, Jegou H. Billion-Scale Similarity Search with GPUs. IEEE Trans. Big Data. 2021;7(3):535-47. doi: https://doi.org/10.1109/tbdata.2019.2921572

10. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Red Hook, NY: Curran Associates; 2017. p. 4765-74. [citado 2026 out 31]. Disponível em: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

11. Aamodt A, Plaza E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications. 1994;7(1):39-59. doi: https://doi.org/10.3233/aic-1994-7104

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. New York, NY, USA: ACM; 2016. p. 785-94. doi: https://doi.org/10.1145/2939672.2939785

13. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157-72. doi: https://doi.org/10.1002/sim.2929

14. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45. doi: https://doi.org/10.2307/2531595

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