Predição de Mortalidade em Prematuros no Brasil com Otimização de Limiar Orientada à Sensibilidade Clínica e Validação Temporal

Sayonara C. de O. Magalhães[1,2], **Karolayne S. Azevedo**[1,2], **Luísa C. de Souza**[1,2]
Matheus Dalmolin[1,2] **e Marcelo A. C. Fernandes**[1,2,3]

[1]InovAI Lab, nPITI/IMD, UFRN, 59.078-900, Natal, RN, Brasil

[2]Leading Advanced Technologies Center of Excellence (LANCE)
nPITI/IMD, UFRN, 59.078-900, Natal, RN, Brasil

[3]Departamento de Engenharia da Computação e Automação (DCA), UFRN
59.078-900, Natal, RN, Brasil

sayonara.magalhaes.068@ufrn.edu.br, karolayneazevsantos@gmail.com

luisa.souza.103@ufrn.edu.br, matheusdalmolinrs@gmail.com

mfernandes@dca.ufrn.br

Resumo. *A mortalidade em prematuros constitui relevante desafio para a saúde pública brasileira. Este estudo propõe um modelo baseado em XGBoost para predição de óbito até 365 dias, utilizando coorte nacional vinculada dos sistemas SINASC e SIM (2014–2022), com 3.064.338 registros e prevalência de 1,78%. O modelo alcançou AUC-ROC de 0,9296. O critério com precisão mínima de 20% identificou 54,8% dos óbitos sinalizando 5% da população, com enriquecimento de 11,2 vezes. A validação temporal (2021–2022) confirmou robustez interanual (AUC-ROC = 0,9312). Os resultados indicam potencial para suporte decisório na alocação de recursos no SUS.*

Abstract. *Mortality among preterm infants remains a major public health challenge in Brazil. This study proposes an XGBoost-based model for predicting death within 365 days, using a nationally linked cohort from the SINASC and SIM databases (2014–2022), comprising 3,064,338 records with a 1.78% prevalence. The model achieved an AUC-ROC of 0.9296. The threshold criterion with minimum precision of 20% identified 54.8% of deaths by flagging 5% of the population, yielding an 11.2-fold enrichment. Temporal validation (2021–2022) confirmed inter-annual robustness (AUC-ROC = 0.9312). The results indicate potential for decision support in resource allocation within Brazil's public health system (SUS).*

1. Introdução

A mortalidade em prematuros permanece como um dos principais desafios da saúde pública, especialmente no contexto da prematuridade, reconhecida como fator determinante para desfechos adversos no primeiro ano de vida. Apesar da redução progressiva da mortalidade infantil no Brasil nas últimas décadas, os óbitos neonatais

ainda representam parcela significativa desses eventos, refletindo desigualdades regionais, condições socioeconômicas, acesso aos serviços de saúde e qualidade da assistência pré-natal e perinatal. A identificação precoce de recém-nascidos prematuros com maior probabilidade de óbito constitui estratégia relevante para orientar ações de monitoramento intensivo e alocação racional de recursos no Sistema Único de Saúde [Modell et al. 2012, Butler et al. 2007, World Health Organization 2012]

Modelos de aprendizado de máquina têm sido amplamente aplicados na estimativa de risco em contextos clínicos e epidemiológicos [Sullivan et al. 2024, Lopes et al. 2022, Motlagh et al. 2020, Lee et al. 2021]. Estudos recentes têm explorado diferentes abordagens para modelagem de desfechos perinatais em bases brasileiras. Em [Silva et al. 2025], técnicas de balanceamento de dados aplicadas a modelos baseados em árvores demonstraram que estratégias híbridas superam abordagens isoladas na predição de parto prematuro, alcançando acurácia de 70%, recall de 64% e precisão de 74%. De forma complementar, [Victor et al. 2025] utilizaram algoritmos de boosting com SMOTE para predição de baixo peso ao nascer na coorte Araraquara, obtendo AUROC de 0,94 com o XGBoost (XGB) e destacando idade gestacional e fatores sociodemográficos como principais preditores. Já [Rocha et al. 2022], analisando 3,5 milhões de nascimentos do CIDACS, evidenciaram que os fatores de risco para parto prematuro incidente e recorrente diferem conforme o histórico obstétrico, ressaltando o papel do intervalo interpartal, idade materna e número de consultas pré-natais. Esses achados reforçam a relevância de estratégias analíticas adaptadas ao contexto epidemiológico e ao perfil de risco materno.

De forma complementar, [Tietzmann et al. 2020] analisaram fatores associados à mortalidade em prematuros com idade gestacional $\leq$32 semanas no Sul do Brasil, utilizando modelos de regressão de Cox aplicados a dados vinculados do Sistema de Informações sobre Nascidos Vivos (SINASC) e Sistema de Informações sobre Mortalidade (SIM). Os autores identificaram maior risco de óbito associado a baixo peso ao nascer, sexo masculino, menor número de consultas pré-natais e nascimento em hospitais públicos, além de interação entre peso e via de parto, com efeito protetor da cesariana em recém-nascidos de muito baixo peso. Já em [Oliveira et al. 2023], foi explorada abordagem não supervisionada baseada em t-SNE para estratificação de risco em prematuros, evidenciando separação consistente entre sobreviventes e não sobreviventes e destacando como variáveis mais discriminativas o peso ao nascer, idade gestacional, índices de Apgar e qualidade do pré-natal. Esses estudos reforçam a complexidade da mortalidade em prematuros e a relevância de estratégias analíticas orientadas ao perfil clínico e epidemiológico.

Embora estudos recentes evidenciem a importância do tratamento do desbalanceamento de classes e o potencial de algoritmos de *gradient boosting*, como o XGB, na modelagem de desfechos perinatais em dados tabulares do SUS, persistem desafios metodológicos relevantes em cenários de evento raro. Em particular, a elevada capacidade discriminativa global nem sempre se traduz em utilidade clínica, uma vez que a definição do limiar de decisão pode alterar substancialmente o perfil operacional do modelo. Além disso, a ausência de critérios explicitamente orientados à sensibilidade clínica e a limitada avaliação de robustez temporal reduzem a aplicabilidade prática dessas abordagens. Tais lacunas evidenciam a necessidade de estratégias que integrem controle do desbalanceamento, definição de thresholds alinhados à priorização assistencial e validação temporal,

visando maior estabilidade e utilidade em saúde pública.

Assim, este trabalho desenvolveu uma coorte nacional a partir da vinculação estruturada dos sistemas SINASC e SIM no período de 2014 a 2022, resultando em um dataset padronizado com variáveis maternas, obstétricas e neonatais, e desfecho definido como óbito até 365 dias. A prevalência de 1,78% caracterizou cenário de evento raro e forte desbalanceamento de classes, impondo desafios à modelagem preditiva. Diante disso, propõe-se um modelo baseado em XGB para predição de mortalidade em prematuros, incorporando estratégias de definição de limiar orientadas à priorização clínica e validação temporal prospectiva. A principal contribuição reside na integração entre coorte populacional em larga escala, controle do desbalanceamento e critérios operacionais alinhados à sensibilidade clínica, visando maior robustez e aplicabilidade em saúde pública.

2. Metodologia

2.1. Dados Utilizados

A coorte analisada foi composta por 3.064.338 registros de nascidos vivos no Brasil entre 2014 e 2022, extraídos do SINASC e vinculados ao SIM para identificação de óbitos atribuídos à prematuridade ocorridos até 365 dias após o nascimento. As variáveis utilizadas como entrada do XGB abrangem características demográficas maternas, variáveis obstétricas, condições neonatais e informações geográficas/contextuais.

As estatísticas descritivas das variáveis numéricas são apresentadas na Tabela 1. A idade materna média foi de 27,15 anos (DP = 7,14), refletindo o perfil reprodutivo típico da população brasileira. O peso médio ao nascer foi de 2.430,67 g (DP = 682,76 g), evidenciando a presença significativa de recém-nascidos prematuros ou com baixo peso na amostra. Os escores médios de Apgar no 1º e 5º minuto (7,87 e 9,01, respectivamente) indicam ampla variabilidade nas condições neonatais imediatas. As variáveis obstétricas relacionadas ao histórico reprodutivo (QTDFILVIVO, QTDFILMORT, QTDGESTANT, QTDPARTNOR e QTDPARTCES) apresentam distribuição assimétrica, com concentração em valores baixos, conforme esperado em coortes populacionais amplas.

Tabela 1. Estatísticas descritivas das variáveis numéricas.

Variável	Tipo	Média	DP	Mín	P95	Máx
IDADEMAE	Demográfica materna	27,15	7,14	14	39	43
PESO (g)	Neonatal	2430,67	682,76	532	3550	4000
SEMAGESTAC	Obstétrica	4,09	0,66	2	5	5
QTDFILVIVO	Obstétrica	0,97	1,26	0	4	6
QTDFILMORT	Obstétrica	0,27	0,59	0	1	3
QTDGESTANT	Obstétrica	1,23	1,48	0	4	7
QTDPARTNOR	Obstétrica	0,67	1,19	0	3	6
QTDPARTCES	Obstétrica	0,34	0,65	0	2	3
APGAR1	Neonatal	7,87	1,68	1	9	10
APGAR5	Neonatal	9,01	1,17	3	10	10

A caracterização das variáveis categóricas principais encontra-se na Tabela 2. Observa-se predominância de mães autodeclaradas pardas (54,97%) e solteiras (46,53%), além de elevada proporção com 8 a 11 anos de escolaridade (60,65%). No âmbito obstétrico, 88,83% das gestações foram únicas, e 57,06% das gestantes realizaram sete ou mais consultas pré-natais. O tipo de parto apresentou predominância de cesáreas (58,84%), padrão consistente com o cenário nacional. A presença de anomalias congênitas foi registrada em 1,99% dos casos, variável de elevada relevância clínica para o desfecho analisado.

Tabela 2. Distribuição das variáveis categóricas.

Variável	Tipo	Código	Categoria	N	%
RACACORMAE	Demográfica materna	1	Branco	1.135.819	37,07
		2	Preto	204.913	6,69
		3	Amarelo	13.794	0,45
		4	Pardo	1.684.613	54,97
		5	Indígena	25.199	0,82
SEXO	Neonatal	1	Masculino	1.541.175	50,29
		2	Feminino	1.523.163	49,71
GRAVIDEZ	Obstétrica	1	Única	2.721.910	88,83
		2	Dupla	332.028	10,84
		3	Tripla ou mais	10.400	0,34
CONSULTAS	Obstétrica	1	Nenhuma consulta	39.889	1,30
		2	1 a 3 consultas	306.165	9,99
		3	4 a 6 consultas	969.918	31,65
		4	7 ou mais consultas	1.748.366	57,06
PARTO	Obstétrica	1	Vaginal	1.261.307	41,16
		2	Cesáreo	1.803.031	58,84
LOCNASC	Obstétrica	1	Hospital	3.037.206	99,11
		2	Outro estabelecimento	13.185	0,43
		3	Domicílio	8.729	0,28
		4	Outros	5.218	0,17
ESTCIVMAE	Demográfica materna	1	Solteira	1.425.871	46,53
		2	Casada	970.991	31,69
		3	Separada/Divorciada	5.893	0,19
		4	Viúva	44.769	1,46
		5	União consensual	616.814	20,13
ESCMAE	Demográfica materna	1	Nenhuma	13.781	0,45
		2	1 a 3 anos	65.414	2,13
		3	4 a 7 anos	502.383	16,39
		4	8 a 11 anos	1.858.568	60,65
		5	12 anos ou mais	624.192	20,37
NASC_CAPITAL	Geográfica/Contextual	N	Não (interior)	2.024.416	66,06
		S	Sim (capital)	1.039.922	33,94
IDANOMAL	Neonatal	1	Sim	61.126	1,99
		2	Não	3.003.212	98,01
KOTELCHUCK	Obstétrica	1	Não tem	38.087	1,24
		2	Inadequado	644.928	21,05
		3	Intermediário	430.388	14,05
		4	Adequado	339.650	11,08
		5	Mais que adequado	1.611.285	52,58

Informações adicionais referentes à classificação da prematuridade e ao contexto geográfico são detalhadas nas Tabelas 3 e 4. Cerca de 35,87% dos registros foram classificados como prematuros confirmados, enquanto os demais apresentaram indícios baseados na idade gestacional ou no peso ao nascer. A distribuição regional evidencia maior concentração de nascimentos nas regiões Sudeste (42,39%) e Nordeste (24,88%), refletindo a distribuição populacional brasileira. A incorporação de variáveis regionais e de unidade federativa permite capturar diferenças estruturais e contextuais potencialmente associadas ao risco de mortalidade em prematuros.

A definição de prematuridade adotada neste estudo seguiu critério composto, combinando idade gestacional e peso ao nascer. Foram classificados como *prematuros confirmados* os recém-nascidos com idade gestacional inferior a 37 semanas [World Health Organization 2012]. Os demais registros foram incluídos por apresentarem indícios compatíveis: *Inconclusivo-IG* corresponde a casos com idade gestacional não confirmada, porém com características clínicas sugestivas, enquanto *Inconclusivo-Peso* refere-se a recém-nascidos com peso inferior a 2.500 g. Essa estratégia ampliada visa maximizar a cobertura da coorte, evitando a exclusão de casos potencialmente prematuros devido a inconsistências no registro. Quanto ao desfecho, este compreende qualquer óbito ocorrido até 365 dias após o nascimento entre os indivíduos da coorte, sem restrição a causas específicas codificadas na declaração de óbito. O estudo não adotou subcate-

Tabela 3. Distribuição de variáveis categóricas adicionais.

Variável	Tipo	Código	Categoria	N	%
TIPO_PREMATURO	Obstétrica	1	Inconclusivo-IG	1.254.689	40,94
		2	Inconclusivo-Peso ($<$ 2500 g)	710.556	23,19
		3	Prematuro	1.099.093	35,87
NASC_REGIAO	Geográfica/Contextual	–	Centro-Oeste	244.486	7,98
		–	Nordeste	762.355	24,88
		–	Norte	300.415	9,80
		–	Sudeste	1.299.023	42,39
		–	Sul	458.059	14,95
RES_REGIAO	Geográfica/Contextual	–	Centro-Oeste	244.244	7,97
		–	Nordeste	761.614	24,85
		–	Norte	302.207	9,86
		–	Sudeste	1.298.042	42,36
		–	Sul	458.231	14,95

gorias de prematuridade por gravidade (e.g., < 32 ou < 28 semanas), tratando a coorte de forma agregada para preservar representatividade populacional e poder estatístico em cenário de evento raro.

Tabela 4. Distribuição das UFs segundo nascimento e residência.

Código	UF	NASC_CODIGO_UF		RES_CODIGO_UF	
		N	%	N	%
11	RO	18.519	0,60	18.546	0,61
12	AC	11.908	0,39	11.252	0,37
13	AM	76.022	2,48	76.601	2,50
14	RR	15.951	0,52	16.000	0,52
15	PA	131.181	4,28	134.004	4,37
16	AP	20.013	0,65	18.527	0,60
17	TO	26.821	0,88	27.277	0,89
21	MA	101.499	3,31	103.357	3,37
22	PI	41.259	1,35	38.352	1,25
23	CE	112.072	3,66	111.826	3,65
24	RN	46.111	1,50	45.961	1,50
25	PB	56.770	1,85	56.666	1,85
26	PE	162.857	5,31	156.433	5,10
27	AL	47.158	1,54	47.539	1,55
28	SE	31.162	1,02	29.982	0,98
29	BA	163.467	5,33	171.498	5,60
31	MG	285.172	9,31	286.830	9,36
32	ES	58.188	1,90	57.960	1,89
33	RJ	233.996	7,64	234.027	7,64
35	SP	721.667	23,55	719.225	23,47
41	PR	177.029	5,78	176.707	5,77
42	SC	109.246	3,57	109.804	3,58
43	RS	171.784	5,61	171.720	5,60
50	MS	55.779	1,82	56.332	1,84
51	MT	65.258	2,13	65.314	2,13
52	GO	77.169	2,52	87.545	2,86
53	DF	46.280	1,51	35.053	1,14

Em conjunto, o grupo de variáveis selecionado contempla múltiplas dimensões (demográfica, obstétrica, neonatal e contextual) proporcionando ao modelo capacidade de capturar tanto fatores clínicos diretos quanto determinantes sociais e estruturais associados ao desfecho.

A distribuição da variável usada como rótulo, ÓBITO, e da coorte por ano de nascimento é apresentada na Tabela 5. Observa-se que o desfecho óbito até 365 dias após o nascimento ocorreu em 54.693 casos (1,78%), enquanto 98,22% dos registros corresponderam a recém-nascidos vivos ao final do período de acompanhamento. Esse perfil evidencia cenário de evento raro e forte desbalanceamento de classes, condição típica em estudos populacionais de mortalidade associada à prematuridade. Tal desbalanceamento impõe desafios relevantes ao treinamento de modelos de aprendizado de máquina, especialmente no que se refere à otimização de métricas sensíveis à classe minoritária. No presente estudo, esse aspecto foi tratado por meio da utilização de métricas adequadas a eventos raros (AUPRC, MCC) e da definição de limiares operacionais orientados à sensi-

bilidade clínica.

Em relação à distribuição temporal, observa-se relativa estabilidade no número de nascimentos ao longo dos anos analisados, com proporções anuais variando entre aproximadamente 10,6% e 11,5% da coorte total. Essa homogeneidade anual reduz o risco de viés estrutural associado a anos específicos e possibilita a realização de validação temporal robusta, conforme explorado na Seção 3.2. Importante destacar que a variável ANO_NASCIMENTO foi utilizada exclusivamente para caracterização temporal da amostra e para a construção do experimento de validação temporal (treino: 2014–2020; teste: 2021–2022), não sendo incluída como variável de entrada no modelo do XGB. Essa decisão metodológica evita vazamento de informação temporal e assegura avaliação prospectiva realista do desempenho do modelo.

Tabela 5. Distribuição da variável rótulo (ÓBITO) e por ano de nascimento.

Variável	Tipo	Código	N	%
ÓBITO	Desfecho	0	3.009.645	98,22
		1	54.693	1,78
ANO_NASCIMENTO	Temporal	2014	333.959	10,90
		2015	334.985	10,93
		2016	326.195	10,64
		2017	335.363	10,94
		2018	352.268	11,50
		2019	353.172	11,53
		2020	341.012	11,13
		2021	338.970	11,06
		2022	348.414	11,37

2.2. Obtenção dos Resultados

A obtenção dos resultados foi conduzida a partir de dois experimentos complementares, com o objetivo de avaliar tanto a capacidade discriminativa quanto a robustez temporal do modelo proposto. O primeiro experimento, denominado E1 – Validação Aleatória Estratificada, consistiu na divisão aleatória da base completa referente ao período de 2014 a 2022 em conjuntos de treinamento e teste, preservando a proporção do desfecho (variável ÓBITO) por meio de estratificação. O conjunto de treinamento foi submetido a validação cruzada em cinco folds, permitindo estimar a estabilidade do modelo e calcular médias e desvios-padrão das métricas de desempenho. Após o treinamento final utilizando todo o conjunto de treino, o modelo foi aplicado ao conjunto de teste independente. O segundo experimento, denominado E2 – Validação Temporal Prospectiva, adotou validação temporal estrita. Os registros de 2014 a 2020 foram utilizados exclusivamente para treinamento e validação cruzada, enquanto os dados de 2021 e 2022 foram reservados como conjunto de teste futuro. Essa estratégia simula cenário prospectivo e permite avaliar a generalização do modelo frente a possíveis variações interanuais. Os limiares de decisão foram definidos apenas com dados do período de treinamento e posteriormente aplicados ao conjunto temporal, evitando qualquer vazamento de informação.

As variáveis utilizadas como entrada do modelo correspondem às descritas nas Tabelas 1, 2 e 3. As variáveis IDADEMAE, PESO, SEMAGESTAC, QTDFILVIVO, QTDFILMORT, APGAR1, APGAR5, parto_prematuro, GRAVIDEZ, CONSULTAS, KOTELCHUCK, QTDGESTANT, QTDPARTNOR e QTDPARTCES foram tratadas como numéricas. Embora algumas dessas variáveis possuam natureza originalmente categórica (por exemplo, GRAVIDEZ, CONSULTAS e KOTELCHUCK), sua codificação ordinal reflete progressão clínica ou intensidade de risco, de modo que o valor

numérico preserva relação monotônica plausível com o desfecho (variável ÓBITO). Essa abordagem permite ao modelo capturar gradientes de risco associados ao aumento ou redução desses valores. As demais variáveis categóricas como SEXO, RACACORMAE, PARTO, LOCNASC, ESTCIVMAE, ESCMAE, nasc_CAPITAL, NASC_CODIGO_UF, NASC_REGIAO, IDANOMAL, RES_CODIGO_UF e RES_REGIAO foram transformadas por meio de codificação one-hot encoding. Essa estratégia evita a introdução artificial de ordenação entre categorias nominais e permite que o modelo avalie independentemente o efeito de cada classe. A aplicação do one-hot encoding é particularmente importante para variáveis geográficas e demográficas, nas quais não há hierarquia intrínseca entre categorias.

O modelo preditivo foi implementado utilizando o XGB, amplamente empregado em tarefas de classificação com dados tabulares estruturados. Foi adotado o método de construção de árvores baseado em histogramas, que proporciona eficiência computacional em bases de grande escala. O treinamento foi realizado com até 800 árvores, profundidade máxima das árvores igual a 6 e taxa de aprendizado de 0,05. Para controle de variância e mitigação de sobreajuste, foram adotadas estratégias de amostragem estocástica tanto das instâncias quanto das variáveis a cada iteração de construção das árvores. A complexidade estrutural do modelo foi regulada por meio de restrições no crescimento dos nós, sem imposição de penalizações adicionais para divisões específicas. A regularização foi conduzida predominantemente via penalização L2, mantendo configuração conservadora quanto à penalização L1, de modo a preservar estabilidade numérica e capacidade de generalização. Em ambos os experimentos, a variável ANO_NASCIMENTO foi utilizada exclusivamente para construção do cenário de validação temporal e não foi incluída como variável preditora. Além das métricas tradicionais de desempenho (AUC-ROC, AUPRC, precision, recall, F1-score, índice de Youden e coeficiente de correlação de Matthews), foram conduzidas análises complementares de concentração de risco, incluindo curvas Precision–Recall (PR), Recall@k, Ganho Acumulativo e Lift, visando avaliar a eficiência de priorização em contexto de evento raro.

3. Resultados e Discussão

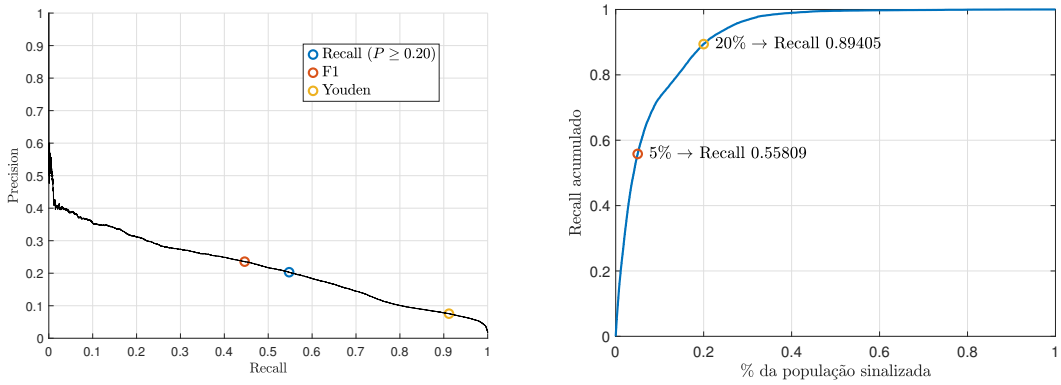
3.1. Experimento E1 – Validação Aleatória Estratificada

A Tabela 6 apresenta o desempenho do XGB no experimento E1 sob três critérios de definição de limiar: recall com precisão mínima de 20%, F1-score e índice de Youden. As métricas globais de discriminação permaneceram invariáveis entre os cenários (AUC-ROC = 0,9296; AUPRC = 0,2160 no teste), por serem independentes do threshold adotado, enquanto as diferenças concentraram-se nas métricas operacionais. O critério com precisão mínima de 20% apresentou o melhor equilíbrio global, alcançando recall de 54,8%, precisão de 20,3% e o maior coeficiente de correlação de Matthews (MCC = 0,315), identificando 5.894 dos 10.762 óbitos com sinalização de aproximadamente 5% da população. A otimização pelo F1-score elevou a precisão para 23,6%, com redução do recall para 44,6%, caracterizando abordagem mais conservadora. Já o índice de Youden priorizou sensibilidade máxima (91,2%), capturando 9.813 óbitos, ao custo de baixa precisão (7,6%), configurando triagem ampliada. A baixa variabilidade entre os folds reforça a estabilidade do modelo, e, do ponto de vista operacional, o critério com precisão mínima de 20% mostrou-se o mais adequado ao conciliar impacto clínico e viabilidade assistencial.

Tabela 6. Desempenho do exp. E1 sob diferentes critérios de otimização.

	Recall (P≥0,20)		F1-score		Youden	
Métrica	CV (± DP)	Teste	CV (± DP)	Teste	CV (± DP)	Teste
AUC-ROC	0,9275 ± 0,0010	0,9296	0,9275 ± 0,0010	0,9296	0,9275 ± 0,0010	0,9296
AUPRC	0,2039 ± 0,0040	0,2160	0,2039 ± 0,0040	0,2160	0,2039 ± 0,0040	0,2160
Precision	0,0839 ± 0,0012	0,2031	0,0839 ± 0,0012	0,2356	0,0839 ± 0,0012	0,0756
Recall	0,8734 ± 0,0070	0,5477	0,8734 ± 0,0070	0,4463	0,8734 ± 0,0070	0,9119
F1-score	0,1519 ± 0,0010	0,2963	0,1519 ± 0,0010	0,3084	0,1519 ± 0,0010	0,1396
MCC	0,6934 ± 0,0050	0,3147	0,6934 ± 0,0050	0,3076	0,6934 ± 0,0050	0,2284

A Figura 1a apresenta a curva PR do XGB no conjunto de teste, juntamente com os três pontos operacionais avaliados: maximização do recall com restrição de precisão mínima de 20%, otimização do F1-score e maximização do índice de Youden. Já a Figura 1b demonstra forte concentração de risco nos indivíduos com maior probabilidade estimada. Observa-se que os 5% da população com maior risco concentram aproximadamente 56% dos óbitos, enquanto 20% concentram cerca de 89% dos eventos. Esse comportamento evidencia elevada capacidade de priorização em contexto de evento raro.

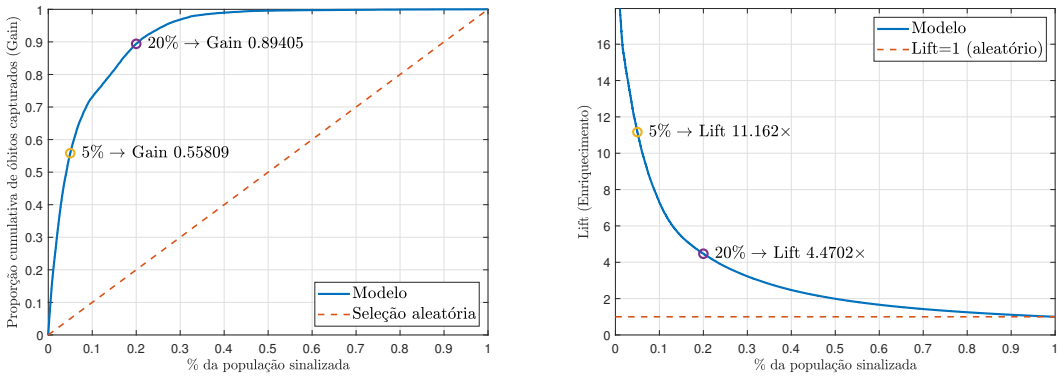


(a) Curva Precision–Recall. **(b) Curva Recall@k.**

Figura 1. Curvas Precision–Recall e Recall@k do XGB para o experimento E1.

O modelo mantém elevada precisão em faixas intermediárias de recall, evidenciando forte capacidade de concentração de risco. O ponto com precisão mínima de 20% situa-se nessa região, alcançando recall superior a 50% com controle da taxa de falsos positivos. A otimização pelo F1-score desloca o ponto operacional para maior precisão e menor sensibilidade, enquanto o critério de Youden posiciona-se em alta sensibilidade, caracterizando triagem ampliada com baixa precisão. A forma monotônica da curva confirma a estabilidade discriminativa em diferentes níveis de decisão. A análise de concentração indica que os 5% indivíduos com maior probabilidade estimada concentram 56% dos óbitos, com risco interno de aproximadamente 20%, cerca de 11,2 vezes superior à prevalência basal (1,78%). Ao considerar os 20% de maior risco, concentram-se 89% dos óbitos, correspondendo a enriquecimento de 4,4 vezes em relação ao risco médio populacional. A Figura 2a apresenta a curva de ganho cumulativo, enquanto a Figura 2b apresenta a curva de Lift, ambas derivadas da ordenação das probabilidades previstas pelo modelo no conjunto de teste. Essas curvas permitem avaliar a capacidade de priorização do modelo em diferentes frações da população.

A curva de ganho cumulativo (Figura 2a) demonstra forte concentração de risco nos indivíduos com maior probabilidade estimada. Observa-se que os 5% da população



(a) Curva de ganho cumulativo. **(b) Curva de Lift (enriquecimento).**

Figura 2. Curvas de ganho cumulativo e Lift do XGB para o experimento E1.

com maior risco concentram aproximadamente 55,8% dos óbitos, enquanto os 20% concentram cerca de 89,4% dos eventos. Em contraste, a linha de seleção aleatória apresenta crescimento linear, indicando ausência de capacidade discriminativa. A curva de Lift (Figura 2b) quantifica o enriquecimento relativo em comparação com a seleção aleatória. No ponto correspondente a 5% da população, o modelo apresenta Lift de aproximadamente 11,16×, indicando que a incidência de óbito nesse subconjunto é mais de onze vezes superior à prevalência basal. Para 20% da população, o Lift é de aproximadamente 4,47×, ainda substancialmente superior ao desempenho aleatório (Lift = 1). Esses resultados evidenciam elevada eficiência de priorização, permitindo identificar a maioria dos eventos com acompanhamento de fração reduzida da população, característica particularmente relevante para estratégias de alocação racional de recursos no contexto do Sistema Único de Saúde.

3.2. Experimento E2 – Validação Temporal Prospectiva

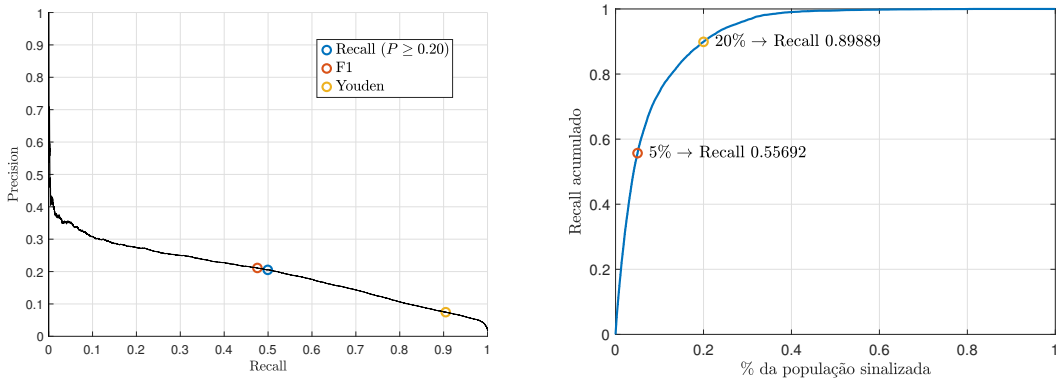
A Tabela 7 apresenta o desempenho do modelo no experimento E2 – Validação Temporal Prospectiva, no qual o treinamento foi realizado com dados de 2014–2020 e o teste com registros de 2021–2022. Observa-se que as métricas globais de discriminação permaneceram elevadas no conjunto temporal (AUC-ROC = 0,9312; AUPRC = 0,1987), indicando manutenção da capacidade preditiva em dados futuros. No critério com precisão mínima de 20%, o modelo atingiu recall de 49,9% e precisão de 20,5%, com MCC = 0,3025, valores próximos aos obtidos na divisão aleatória, evidenciando estabilidade operacional. A otimização pelo F1-score resultou em leve aumento de precisão (21,1%) e redução do recall para 47,6%, mantendo desempenho global semelhante. Já o critério baseado no índice de Youden priorizou alta sensibilidade (90,5%), ao custo de menor precisão (7,5%), caracterizando cenário de triagem ampliada. A pequena redução da AUPRC em comparação com o experimento aleatório é compatível com variações interanuais naturais da base e não indica degradação relevante do modelo, reforçando sua robustez temporal.

A Figura 3a apresenta a curva PR obtida na validação temporal, utilizando dados de 2021–2022 como conjunto de teste após treinamento em registros de 2014–2020. Já a Figura 3b apresenta a curva Recall@k obtida na validação temporal, utilizando dados de 2021–2022 como conjunto de teste. Observa-se que a concentração de risco permanece praticamente inalterada na avaliação temporal. Os 5% da população com maior risco

Tabela 7. Desempenho no exp. E2 sob diferentes critérios de otimização.

	Recall (P≥0,20)		F1-score		Youden	
Métrica	CV (± DP)	Teste	CV (± DP)	Teste	CV (± DP)	Teste
AUC-ROC	0,9275 ± 0,0010	0,9312	0,9275 ± 0,0010	0,9312	0,9275 ± 0,0010	0,9312
AUPRC	0,2039 ± 0,0040	0,1987	0,2039 ± 0,0040	0,1987	0,2039 ± 0,0040	0,1987
Precision	0,0839 ± 0,0012	0,2052	0,0839 ± 0,0012	0,2112	0,0839 ± 0,0012	0,0746
Recall	0,8734 ± 0,0070	0,4993	0,8734 ± 0,0070	0,4756	0,8734 ± 0,0070	0,9048
F1-score	0,1519 ± 0,0010	0,2909	0,1519 ± 0,0010	0,2925	0,1519 ± 0,0010	0,1378
MCC	0,6934 ± 0,0050	0,3025	0,6934 ± 0,0050	0,2998	0,6934 ± 0,0050	0,2273

estimado concentram aproximadamente 55,7% dos óbitos, enquanto os 20% concentram cerca de 89,9% dos eventos.



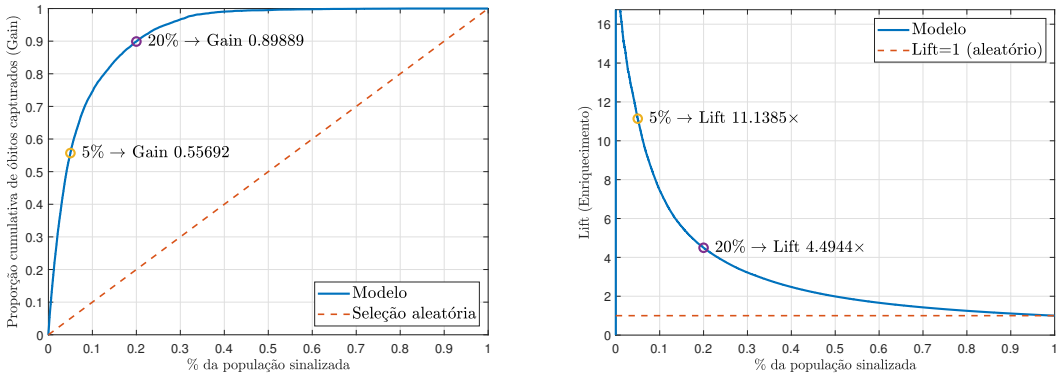
(a) Curva Precision–Recall. **(b) Curva Recall@k.**

Figura 3. Curvas Precision–Recall e Recall@k do XGB para o experimento E2.

Observa-se que a curva mantém formato semelhante ao obtido na divisão aleatória, indicando estabilidade da capacidade discriminativa ao longo do tempo. A área sob a curva PR (AUPRC = 0,1987) permanece substancialmente superior à prevalência basal do evento (1,78%), confirmando capacidade consistente de concentração de risco em dados futuros. O ponto operacional com restrição de precisão mínima de 20% (Recall = 0,499; Precision = 0,205) situa-se em região intermediária da curva, demonstrando manutenção do equilíbrio entre sensibilidade e especificidade mesmo sob separação temporal estrita. O ponto otimizado pelo F1-score apresenta maior precisão (0,211), porém menor recall (0,476), caracterizando estratégia mais conservadora. Já o critério baseado no índice de Youden posiciona-se em região de alta sensibilidade (Recall = 0,905), refletindo cenário de triagem ampliada. A semelhança estrutural entre esta curva e aquela obtida na validação do experimento E1 sugere ausência de degradação significativa do modelo ao longo do tempo, indicando robustez frente a possíveis variações anuais nos dados. A manutenção desses valores em comparação com a validação aleatória estratificada (experimento E1) indica estabilidade do ranking produzido pelo modelo ao longo do tempo, sugerindo baixa sensibilidade a possíveis variações anuais na distribuição dos dados. Esse resultado reforça a robustez da abordagem proposta, evidenciando que a capacidade de priorização não depende de sobreajuste específico ao período de treinamento.

No caso das Figuras 4a e 4b apresentam as curvas de ganho cumulativo e de Lift obtidas na validação temporal, utilizando dados de 2021–2022 como conjunto de teste. Observa-se que a concentração de risco permanece praticamente inalterada em comparação com a divisão aleatória. Na validação temporal, os 5% da população com

maior risco estimado concentram aproximadamente 55,7% dos óbitos, enquanto os 20% concentram cerca de 89,9% dos eventos.



(a) Curva de ganho cumulativo. **(b) Curva de Lift (enriquecimento).**

Figura 4. Curvas de ganho cumulativo e Lift do XGB para o experimento E2.

A curva de Lift confirma esse comportamento, apresentando enriquecimento de aproximadamente 11,14× nos 5% iniciais e 4,49× nos 20% iniciais, valores muito próximos aos observados na avaliação aleatória. A manutenção desses níveis de enriquecimento ao longo do tempo indica estabilidade da ordenação de risco produzida pelo modelo, sugerindo baixa sensibilidade a variações interanuais e reforçando a robustez temporal da abordagem proposta. Os resultados indicam que o modelo mantém desempenho consistente mesmo quando avaliado em coortes temporais futuras não utilizadas no treinamento, sugerindo potencial aplicabilidade prospectiva em cenários reais de monitoramento populacional.

4. Conclusões

Este estudo desenvolveu e avaliou um modelo preditivo baseado em XGBoost para mortalidade em prematuros utilizando coorte nacional vinculada dos sistemas SINASC e SIM (2014–2022), composta por mais de 3 milhões de registros e prevalência de óbito de 1,78%. Mesmo em cenário de evento raro, o modelo apresentou elevada capacidade discriminativa (AUC-ROC superior a 0,93) e forte capacidade de concentração de risco. A principal contribuição metodológica consiste na definição de limiar orientada à sensibilidade clínica com restrição mínima de precisão, permitindo identificar mais da metade dos óbitos sinalizando aproximadamente 5% da população, com enriquecimento superior a 11 vezes em relação à prevalência basal. A validação temporal prospectiva confirmou a robustez interanual do modelo, com manutenção do desempenho discriminativo e da capacidade de priorização em dados futuros. Os resultados indicam que a integração de bases administrativas nacionais aliada a estratégias operacionais alinhadas à viabilidade assistencial pode oferecer suporte decisório relevante para alocação racional de recursos no Sistema Único de Saúde, contribuindo para estratégias mais eficientes de monitoramento e potencial redução da mortalidade associada à prematuridade.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CA-

PES) e ao projeto Redes de Colaboração em Saúde Digital (Rede SOFIA), financiado pela Rede Nacional de Ensino e Pesquisa (RNP) no âmbito do Programa Prioritário em Saúde Digital (PPI-SD), com apoio do Ministério da Ciência, Tecnologia e Inovação (MCTI).

Referências

Butler, A. S., Behrman, R. E., et al. (2007). Preterm birth: causes, consequences, and prevention.

Lee, J., Cai, J., Li, F., and Vesoulis, Z. A. (2021). Predicting mortality risk for preterm infants using random forest. *Scientific reports*, 11(1):1–9.

Lopes, M. L. B., Barbosa, R. d. M., and Fernandes, M. A. C. (2022). Unsupervised learning applied to the stratification of preterm birth risk in brazil with socioeconomic data. *International Journal of Environmental Research and Public Health*, 19(9).

Modell, B., Berry, R., Boyle, C. A., Christianson, A., Darlison, M., Dolk, H., Howson, C. P., Mastroiacovo, P., Mossey, P., and Rankin, J. (2012). Global regional and national causes of child mortality. *The Lancet*, 380(9853):1556.

Motlagh, A. J., Asgary, R., and Kabir, K. (2020). Evaluation of clinical risk index for babies to predict mortality and morbidity in neonates admitted to neonatal intensive care unit. *Electronic Journal of General Medicine*, 17(5).

Oliveira, Y. S. d. P., Freitas, G. R. d., Barros, W. K. P., Souza, L. C. d., Azevedo, K. S., Barbosa, R. d. M., and Fernandes, M. A. C. (2023). Aprendizagem de máquina aplicada à estratificação de risco de mortalidade de recém-nascidos associados a partos prematuros. In Simas, E., Ferreira, D. D., and Oliveira, L. R., editors, *Anais do XVI Congresso Brasileiro de Inteligência Computacional (CBIC'2023)*, pages 1–8, Salvador, BA. SBIC.

Rocha, A. S., de Cássia Ribeiro-Silva, R., Fiaccone, R. L., Paixao, E. S., Falcão, I. R., Alves, F. J. O., Silva, N. J., Ortelan, N., Rodrigues, L. C., Ichihara, M. Y., et al. (2022). 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 medicine*, 20(1):111.

Silva, A. B., Rocha, E. d. S., Lorenzato, J. F., and Endo, P. T. (2025). Evaluating how different balancing data techniques impact on prediction of premature birth using machine learning models. *PLOS ONE*, 20:1–21.

Sullivan, B. A., Beam, K., Vesoulis, Z. A., Aziz, K. B., Husain, A. N., Knake, L. A., Moreira, A. G., Hooven, T. A., Weiss, E. M., Carr, N. R., et al. (2024). Transforming neonatal care with artificial intelligence: challenges, ethical consideration, and opportunities. *Journal of perinatology*, 44(1):1–11.

Tietzmann, M. R., Teichmann, P. d. V., Vilanova, C. S., Goldani, M. Z., and Silva, C. H. d. (2020). Risk factors for neonatal mortality in preterm newborns in the extreme south of brazil. *Scientific Reports*, 10(1):7252.

Victor, A., Almeida, F., Xavier, S. P., and Rondó, P. H. (2025). Predicting low birth weight risks in pregnant women in brazil using machine learning algorithms: data from the araraquara cohort study. *BMC Pregnancy and Childbirth*, 25(1):320.

World Health Organization (2012). *Born Too Soon: The Global Action Report on Preterm Birth*. World Health Organization, Geneva, Switzerland.