Random Survival Forest for Neonatal Survival Prediction in Preterm Infants from a Brazilian National Cohort of 2014 to 2022

Edson M. de B. Junior[1,2] **e Marcelo A. C. Fernandes**[1,2,3]

[1] InovAI Lab – nPITI/IMD – Universidade Federal do Rio Grande do Norte (UFRN)
Natal – RN – Brazil

[2]Leading Advanced Technologies Center of Excellence (LANCE) – nPITI/IMD
UFRN – Natal – RN – Brazil

[3]Department of Computer Engineering and Automation (DCA) – UFRN
Natal – RN – Brazil

edson.maciel.017@ufrn.edu.br, mfernandes@dca.ufrn.br

Abstract. *Prematurity accounts for 60% to 70% of neonatal deaths in the first week of life in Brazil, yet the Cox proportional-hazards model, the standard survival tool, frequently violates its own assumptions in neonatal cohorts, where protective effects concentrate in the first days and dissipate thereafter. This work applies a Random Survival Forest (RSF) to a national cohort of preterm infants from linked SINASC–SIM records covering more than 3 million births between 2014 and 2022. Using an analytical sample of 40,970 observations, the RSF is benchmarked against the Cox model. Gestational age was the dominant predictor in both approaches, with permutation importance eight times higher than the second-ranked variable. The RSF captured non-linear contributions from birth weight and maternal age that did not reach significance in the Cox model. The proportional-hazards assumption was violated in four of five covariates (Schoenfeld test, $p < 0.05$), formally supporting the RSF as the primary method. Significant regional disparities, particularly between the North and South, point to inequalities in neonatal intensive care access beyond clinical birth factors.*

Resumo. *A prematuridade responde por 60% a 70% das mortes neonatais na primeira semana de vida no Brasil, mas o modelo de Cox frequentemente viola seu pressuposto de proporcionalidade em coortes neonatais, onde o efeito protetor da idade gestacional se concentra nos primeiros dias e se dilui progressivamente. Este trabalho aplica um Random Survival Forest (RSF) a uma coorte nacional de prematuros construída a partir de registros vinculados do SINASC e do SIM, cobrindo mais de 3 milhões de nascimentos entre 2014 e 2022. Em uma amostra analítica de 40.970 observações, o RSF é comparado ao modelo de Cox como baseline. A idade gestacional foi o preditor dominante em ambos os modelos, com importância por permutação oito vezes superior à da segunda variável. O RSF capturou efeitos não-lineares do peso ao nascer e da idade materna que não atingiram significância no Cox. O pressuposto de proporcionalidade foi violado em quatro das cinco covariáveis (teste de Schoenfeld, $p < 0{,}05$), justificando formalmente o RSF como abordagem*

principal. Diferenças regionais expressivas, especialmente entre Norte e Sul, evidenciam desigualdades no acesso a cuidados intensivos neonatais além dos fatores clínicos ao nascimento.

1. Introdução

A mortalidade neonatal constitui um dos principais indicadores de desenvolvimento em saúde pública, e a prematuridade é sua causa mais frequente tanto no Brasil quanto no cenário global [Blencowe et al. 2012]. Nascimentos antes de 37 semanas completas de gestação respondem por aproximadamente 11% dos partos no país e por 60% a 70% dos óbitos ocorridos na primeira semana de vida [Victora et al. 2011]. Além da magnitude epidemiológica, a prematuridade apresenta uma característica temporal que a distingue de outros problemas de saúde pública: o risco de morte não é estático, mas evolui de forma distinta ao longo do período neonatal. Prematuros extremos concentram seu risco nos primeiros dois a três dias de vida, ao passo que prematuros tardios exibem padrões distribuídos ao longo de todo o período de 28 dias. Essa heterogeneidade temporal impõe limites à análise baseada em desfechos binários (óbito sim/não), que ignora o *quando* do evento e subestima as diferenças de risco entre subgrupos clínicos e regionais.

A análise de sobrevivência fornece o arcabouço metodológico adequado para capturar essa dimensão temporal. O estimador não paramétrico de Kaplan–Meier [Kaplan and Meier 1958] permite descrever a função de sobrevivência sem impor pressupostos sobre sua forma, enquanto o modelo de riscos proporcionais de Cox [Cox 1972] possibilita estimar a contribuição independente de cada fator prognóstico por meio do Hazard Ratio (HR), que quantifica o risco instantâneo relativo de óbito. Essas ferramentas têm sido amplamente aplicadas à mortalidade neonatal em países de renda média e alta, com resultados consistentes quanto à centralidade da idade gestacional e do peso ao nascer como determinantes do prognóstico [World Health Organization 2023]. No contexto brasileiro, estudos com dados do SINASC e do SIM confirmaram a relevância desses fatores, mas raramente exploraram a dimensão não-linear das relações entre covariáveis e sobrevida ou questionaram formalmente a validade do pressuposto de proporcionalidade dos riscos em coortes de grande porte.

A violação desse pressuposto é, na verdade, esperada em dados neonatais: o efeito protetor do maior peso ao nascer, por exemplo, tende a ser intenso nos primeiros dias e se dilui após a primeira semana, à medida que os sobreviventes superam as complicações agudas. Quando o pressuposto é violado, o HR estimado pelo Cox representa uma média temporal do efeito real, potencialmente enganosa para fins de estratificação de risco. Métodos de aprendizado de máquina para sobrevivência, como o Random Survival Forest (RSF) proposto por [Ishwaran et al. 2008], superam essa limitação ao construir um ensemble de árvores de sobrevivência sem impor qualquer forma paramétrica à função de risco. O RSF acomoda naturalmente não-linearidades e interações de alta ordem entre covariáveis, e a importância das variáveis pode ser quantificada por permutação, oferecendo um critério preditivo direto que independe de testes de hipótese. Apesar dessas vantagens, sua aplicação a coortes neonatais brasileiras de escala nacional ainda é incipiente na literatura, o que representa uma lacuna metodológica relevante.

Este trabalho aplica o RSF a uma coorte nacional de prematuros brasileiros construída a partir de dados vinculados do SINASC e do SIM, cobrindo mais de 3 milhões

de nascimentos entre 2014 e 2022, com o objetivo de identificar os principais determinantes da sobrevida neonatal e avaliar o ganho do RSF em relação ao modelo de Cox em um cenário de dados administrativos. Os resultados mostram que a idade gestacional é o preditor dominante em ambos os modelos, com importância por permutação no RSF oito vezes superior à da segunda variável classificada, e que o RSF captura contribuições de peso ao nascer e idade materna que não alcançam significância no Cox, evidenciando efeitos não-lineares que o modelo paramétrico não consegue modelar. O pressuposto de proporcionalidade dos riscos foi violado em quatro das cinco covariáveis, validando formalmente o uso do RSF como abordagem principal. Diferenças regionais expressivas na sobrevida neonatal, em especial o contraste entre Norte e Sul do país, apontam para desigualdades no acesso a cuidados intensivos neonatais que transcendem os fatores clínicos ao nascimento.

2. Metodologia

2.1. Dados

O estudo utilizou dados vinculados do Sistema de Informações sobre Nascidos Vivos (SINASC) e do Sistema de Informações sobre Mortalidade (SIM), disponibilizados pelo DATASUS [Brasil. Ministério da Saúde. DATASUS 2022b, Brasil. Ministério da Saúde. DATASUS 2022a], cobrindo o período de 2014 a 2022. O SINASC registra variáveis clínicas e demográficas coletadas na Declaração de Nascido Vivo, enquanto o SIM registra os óbitos por meio da Declaração de Óbito. O vínculo entre as duas bases foi realizado por determinismo utilizando os campos de data de nascimento, município e características do recém-nascido.

A base original compreendeu 3.064.338 registros no período de 2014 a 2022. Para a construção da amostra analítica utilizada na análise de sobrevivência, foram selecionados os registros com `IDADE_OBITO` superior a zero, critério que resultou em 40.970 observações. Todos esses registros apresentaram `SEMAGESTAC < 37`, correspondendo à população de prematuros analisada. Não foi realizada subamostragem aleatória, balanceamento entre eventos e censuras ou seleção baseada na variável de desfecho além do critério temporal de inclusão. As cinco variáveis utilizadas nos modelos foram `SEMAGESTAC`, `PESO`, `APGAR1`, `APGAR5` e `IDADEMAE`.

2.2. Métodos

O problema de sobrevida em prematuros apresenta duas características que motivam o uso de aprendizado de máquina: o risco de óbito varia ao longo do tempo de forma não proporcional entre subgrupos clínicos, e as relações entre covariáveis e desfecho tendem a ser não-lineares e dependentes de interações de alta ordem. O modelo estatístico clássico de referência para esse tipo de dado é o modelo de riscos proporcionais de Cox [Cox 1972], ajustado com estimativa de Breslow para a função de risco basal. Sua adequação foi avaliada pelo teste de resíduos de Schoenfeld [Grambsch and Therneau 1994] e a discriminação pelo índice de concordância C-index de Harrell [Harrell et al. 1982]. O Cox serve aqui como *baseline* interpretável para comparação com a abordagem de ML.

O método principal deste trabalho é o Random Survival Forest (RSF) [Ishwaran et al. 2008], algoritmo de aprendizado de máquina que estende as florestas aleatórias ao contexto de dados censurados. O RSF constrói um ensemble de árvores

de sobrevivência, em que cada árvore é treinada em uma amostra bootstrap e cada nó avalia apenas um subconjunto aleatório de $\sqrt{p}$ covariáveis, reduzindo a correlação entre as árvores e aumentando a generalização. A função de risco acumulado de Nelson-Aalen é estimada em cada folha terminal e agregada por média entre todas as árvores, produzindo uma estimativa de sobrevivência individualizada para cada observação sem pressupor qualquer forma paramétrica para a função de risco. O modelo foi configurado com 100 árvores, divisão mínima de 10 observações por nó interno e mínimo de 5 por folha terminal. A importância de cada covariável foi calculada por permutação com 5 repetições, estratégia que mensura a queda no C-index ao embaralhar aleatoriamente os valores de cada variável no conjunto de teste, quantificando sua contribuição marginal para o poder preditivo do ensemble. O desempenho final de ambos os modelos foi avaliado pelo C-index no conjunto de teste retido, permitindo comparação direta em condições idênticas.

Todas as análises foram conduzidas em Python 3.11, utilizando as bibliotecas *lifelines* (versão 0.27) para as curvas de Kaplan–Meier e o modelo de Cox, *scikit-survival* (versão 0.22) para o RSF e *scikit-learn* (versão 1.4) para o pré-processamento.

3. Resultados

3.1. Caracterização da amostra e análise descritiva

A taxa geral de óbito neonatal na base completa foi de 1,78%. Na amostra analítica (n = 40.970), 28.986 registros (70,7%) corresponderam a eventos de óbito e 11.984 (29,3%) a observações censuradas, distribuição que reflete os critérios de elegibilidade e completude aplicados, e não constitui estimativa populacional da mortalidade. A Tabela 1 apresenta a distribuição por faixa de idade gestacional, peso ao nascer e região geográfica.

A curva geral de Kaplan–Meier (Figura 1) mostra que a maior concentração de óbitos ocorre nos primeiros dias de vida, com mediana de sobrevida de aproximadamente 1,43 semanas, seguida de desaceleração progressiva ao longo do período neonatal.

As curvas estratificadas por faixa de idade gestacional e por peso ao nascer (Figura 2) revelam separação crescente entre os grupos à medida que a prematuridade aumenta ou o peso ao nascer diminui. Os testes log-rank pareados, visualizados como matrizes de $-\log_{10}(p)$ (Figura 3), quantificam a magnitude dessas diferenças. Para a idade gestacional, a comparação entre prematuros extremos (<28 sem) e tardios (32–36 sem) atingiu $p = 1{,}83 \times 10^{-72}$, e a comparação entre as faixas 28–31 sem e 32–36 sem foi ainda mais expressiva ($p = 9{,}23 \times 10^{-120}$). Para o peso ao nascer, os menores p-valores envolveram a faixa <1.000 g, tanto em relação a 1.500–2.499 g ($p = 1{,}38 \times 10^{-217}$) quanto a ≥ 2.500 g ($p = 2{,}68 \times 10^{-204}$); o único par sem diferença significativa foi ≥ 2.500 g vs. 1.500–2.499 g ($p = 0{,}71$). Para a análise regional (Figura 4), Norte vs. Sul ($p = 2{,}32 \times 10^{-13}$) e Sudeste vs. Sul ($p = 6{,}30 \times 10^{-12}$) foram as comparações mais expressivas, enquanto Centro-Oeste vs. Sul não atingiu significância ($p = 0{,}59$), provavelmente em função do tamanho amostral reduzido do Centro-Oeste (n = 951).

3.2. Modelo de Cox como baseline

Os VIFs foram todos inferiores a 5 (máximo: APGAR1 = 3,69), indicando ausência de colinearidade problemática. A Tabela 2 apresenta os resultados do modelo de Cox ajustado

Tabela 1. Distribuição da amostra por faixa de idade gestacional, peso ao nascer e região geográfica. SINASC-SIM, Brasil, 2014–2022.

Categoria	n	%
Idade Gestacional		
$<$ 28 semanas	11.312	27,6
28–31 semanas	11.853	28,9
32–36 semanas	16.519	40,3
$\geq$37 semanas*	1.286	3,1
Peso ao Nascer		
$<$ 1.000 g	11.388	27,8
1.000–1.499 g	7.012	17,1
1.500–2.499 g	11.647	28,4
$\geq$2.500 g	10.923	26,6
Região Geográfica		
Sudeste	23.496	57,3
Nordeste	6.867	16,7
Norte	5.094	12,4
Sul	4.562	11,1
Centro-Oeste	951	2,3

*Divergência de codificação no SINASC; mantidos como categoria de referência.

no conjunto de treinamento.

Tabela 2. Modelo de riscos proporcionais de Cox. Treinamento: n = 32.776; eventos: 23.189. Avaliado no conjunto de teste: n = 8.194. Covariáveis padronizadas. SINASC-SIM, Brasil, 2014–2022.

Covariável	β	HR	IC inf.	IC sup.	z	p
SEMAGESTAC	$-$0,24	0,78	0,77	0,80	$-$26,34	$<$0,005
APGAR5	$-$0,08	0,93	0,91	0,95	$-$6,46	$<$0,005
APGAR1	$+$0,02	1,02	0,99	1,04	$+$1,42	0,15
IDADEMAE	$-$0,00	1,00	0,98	1,01	$-$0,59	0,55
PESO	$-$0,00	1,00	0,98	1,02	$-$0,15	0,88

C-index (teste): 0,5765 AIC parcial: 457.649,47 LR test: 1.767,16 (df = 5; $p < 10^{-300}$)

Apenas SEMAGESTAC (HR = 0,78; IC 95%: 0,77–0,80) e APGAR5 (HR = 0,93; IC 95%: 0,91–0,95) alcançaram significância estatística ($p < 0{,}005$), indicando que cada desvio padrão adicional nessas variáveis reduz o risco instantâneo de óbito em 22% e 7%, respectivamente. As demais covariáveis não foram significativas ($p > 0{,}10$). A Figura 5 ilustra os Hazard Ratios com seus intervalos de confiança.

O teste de Schoenfeld identificou violação do pressuposto de proporcionalidade em quatro das cinco covariáveis: PESO, SEMAGESTAC e APGAR5 com $p < 5 \times 10^{-5}$, e IDADEMAE com $p = 0{,}03$. Apenas APGAR1 não violou o pressuposto ($p = 0{,}66$). Esse resultado indica que o efeito dessas variáveis sobre o risco de óbito varia ao longo do tempo, o que invalida a interpretação de um HR único e constante e motiva diretamente

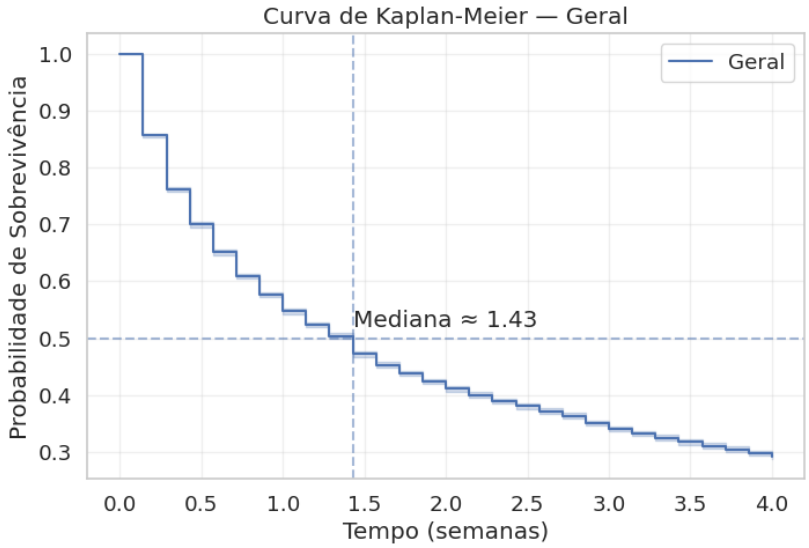
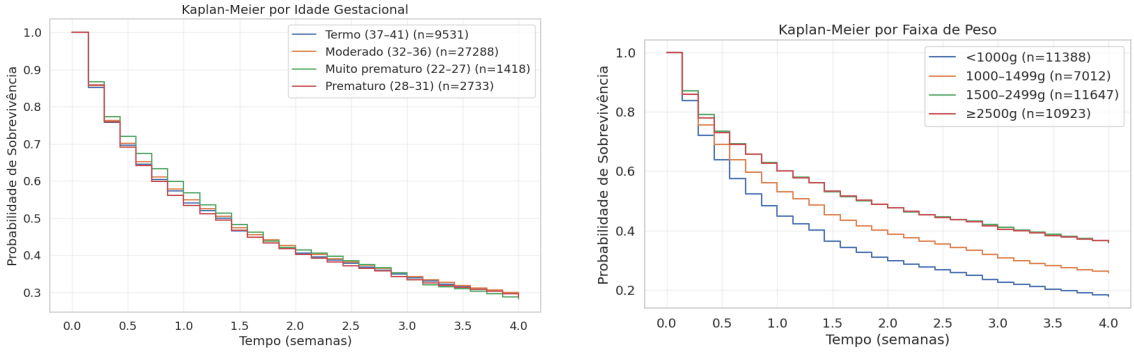


Figura 1. Curva geral de Kaplan–Meier para sobrevida neonatal em prematuros. Período neonatal (0–28 dias). Fonte: SINASC-SIM, Brasil, 2014–2022.



(a) Por faixa de idade gestacional. **(b) Por faixa de peso ao nascer.**

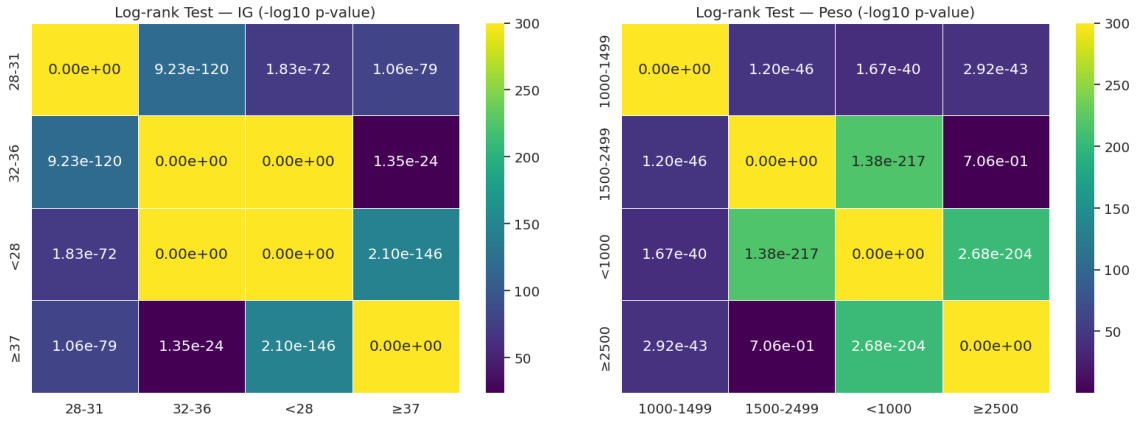
Figura 2. Curvas de Kaplan–Meier estratificadas. Período neonatal (0–28 dias). Fonte: SINASC-SIM, Brasil, 2014–2022.

o uso do RSF como abordagem principal. O C-index do Cox no conjunto de teste foi de 0,5765, servindo de referência para a comparação com o modelo de ML.

3.3. Random Survival Forest: importância das variáveis e desempenho

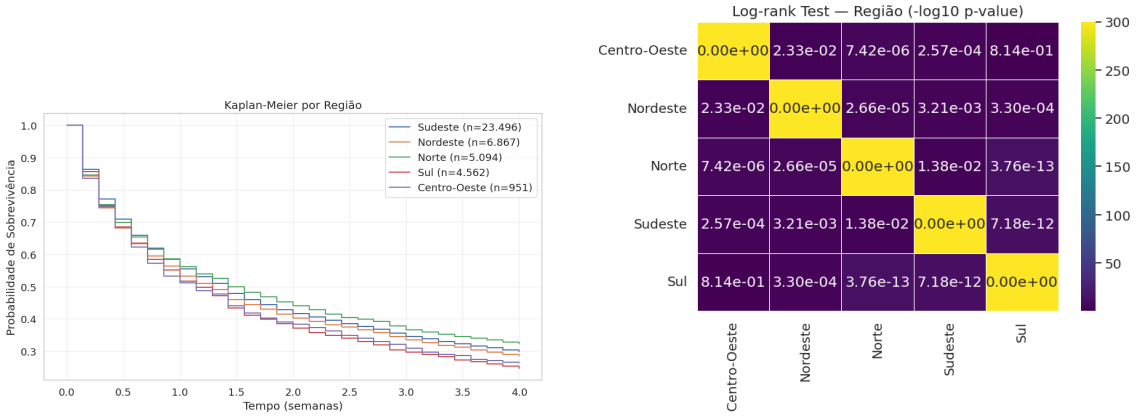
O RSF foi treinado no mesmo conjunto de dados e avaliado no mesmo conjunto de teste do Cox. A Tabela 3 apresenta a importância por permutação de cada covariável e compara o C-index dos dois modelos.

A importância por permutação confirmou a dominância de SEMAGESTAC (0,049), com distância expressiva em relação ao segundo colocado PESO (0,006). Esse resultado é consistente com o do Cox e reforça a centralidade da idade gestacional como preditor prognóstico, independentemente do método utilizado. A contribuição de PESO e IDADEMAE, embora pequena em magnitude absoluta, foi capturada pelo RSF mesmo sem ter atingido significância estatística no Cox, o que sugere que essas variáveis exercem efeitos não-lineares ou condicionais a interações que o modelo paramétrico não consegue



(a) Por faixa de idade gestacional. **(b) Por faixa de peso ao nascer.**

Figura 3. Matrizes de p-valores do teste log-rank pareado ($-\log_{10} p$). Amarelo indica maior significância. Fonte: SINASC-SIM, Brasil, 2014–2022.



(a) Curvas de Kaplan–Meier por região. **(b) Matriz de p-valores log-rank por região.**

Figura 4. Análise regional da sobrevida neonatal. Fonte: SINASC-SIM, Brasil, 2014–2022.

modelar. APGAR1 apresentou importância negativa ($-0{,}002$), evidenciando que, dado que APGAR5 já está no modelo, essa variável não agrega poder discriminativo ao ensemble e é candidata à exclusão em versões futuras.

O C-index do RSF no conjunto de teste foi de 0,5600, ligeiramente inferior ao do Cox (0,5765). Esse resultado é esperado em cenários com poucos preditores de baixa resolução: o ganho do RSF tende a ser mais pronunciado quando há muitas variáveis com relações complexas a explorar. A Figura 6 apresenta a importância das variáveis e as curvas de sobrevivência individualizadas estimadas pelo RSF para cinco pacientes amostrados aleatoriamente do conjunto de teste, ilustrando a capacidade do modelo de produzir prognósticos personalizados sem pressupor proporcionalidade dos riscos.

4. Discussão

Os resultados deste trabalho devem ser interpretados em dois planos complementares: o plano clínico-epidemiológico, que diz respeito ao que os modelos revelam sobre os determinantes da sobrevida neonatal, e o plano metodológico, que diz respeito ao que

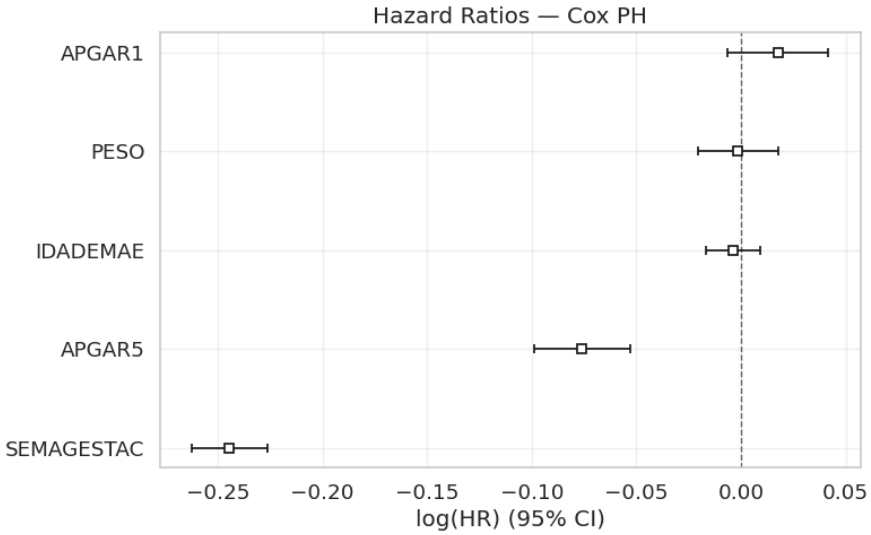


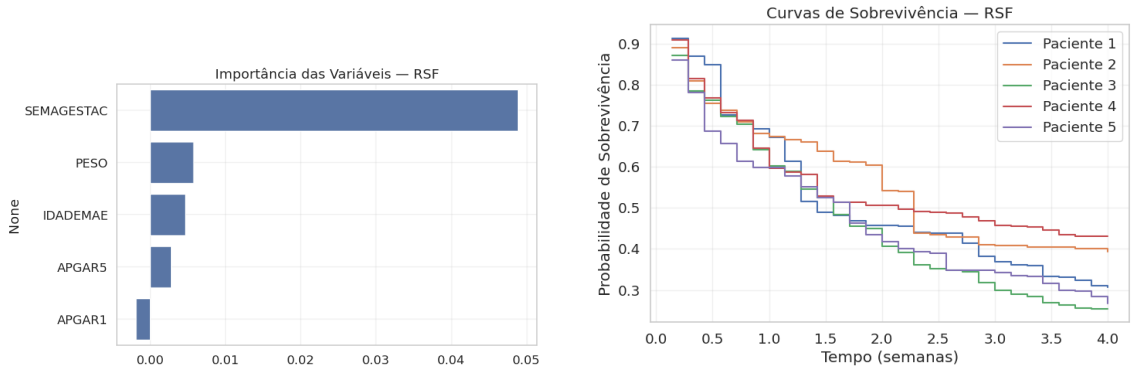
Figura 5. Forest plot dos Hazard Ratios do modelo de Cox com IC 95%. Linha tracejada em HR = 1 (ausência de efeito). Fonte: SINASC-SIM, Brasil, 2014–2022.

Tabela 3. Importância das variáveis no RSF por permutação (média de 5 repetições) e comparação de C-index entre modelos. SINASC-SIM, Brasil, 2014–2022.

Covariável	Importância (permutação)	Ranking
SEMAGESTAC	0,04886	1º
PESO	0,00575	2º
IDADEMAE	0,00463	3º
APGAR5	0,00275	4º
APGAR1	−0,00188	5º
C-index Cox (teste): 0,5765	C-index RSF (teste): 0,5600	

a comparação entre o Cox e o RSF ensina sobre a escolha de modelos para dados de sobrevivência com pressuposto violado.

No plano clínico-epidemiológico, a centralidade da idade gestacional como preditor prognóstico foi confirmada por ambos os modelos e por métricas distintas: o Cox atribuiu a SEMAGESTAC o único coeficiente com magnitude expressiva (HR = 0,78), e o RSF atribuiu a essa variável importância por permutação de 0,049, oito vezes maior do que a segunda colocada. Esse resultado é consistente com a literatura, que posiciona a idade gestacional como o principal determinante isolado da mortalidade neonatal em prematuros [Blencowe et al. 2012, Victora et al. 2011]. O Apgar no 5º minuto foi o segundo preditor significativo no Cox (HR = 0,93), o que reflete sua natureza de marcador composto da condição clínica imediata do recém-nascido e de sua resposta às primeiras intervenções pós-parto. A ausência de significância do APGAR1 no modelo multivariado, por sua vez, é esperada: como APGAR5 é calculado após APGAR1 e incorpora informação adicional sobre a resposta à reanimação, os dois escores compartilham substancial conteúdo prognóstico, e a inclusão de ambos no mesmo modelo produz supressão



(a) Importância por permutação das variáveis.

(b) Curvas de sobrevivência individualizadas (5 pacientes do conjunto de teste).

Figura 6. Random Survival Forest: importância das variáveis e predições individualizadas. Fonte: SINASC-SIM, Brasil, 2014–2022.

do efeito de APGAR1. A importância negativa de APGAR1 no RSF ($-0{,}002$) confirma essa interpretação de forma independente do pressuposto de linearidade.

O peso ao nascer merece atenção especial. No modelo de Cox, PESO não atingiu significância estatística após ajuste para SEMAGESTAC, o que poderia ser interpretado precipitadamente como ausência de efeito prognóstico. O RSF, contudo, atribuiu a PESO a segunda maior importância por permutação (0,006), sugerindo que essa variável exerce influência sobre a sobrevida por meio de relações não-lineares ou de interações com outras covariáveis que o modelo linear-aditivo do Cox não consegue capturar. Essa discrepância entre os dois modelos é metodologicamente informativa: a não-significância no Cox não implica irrelevância preditiva, mas sim que o efeito de PESO não se expressa de forma proporcional e linear ao longo de todo o período de acompanhamento, o que é biologicamente plausível dado que o benefício do maior peso tende a se concentrar nos primeiros dias de vida, quando as complicações agudas da prematuridade dominam o risco.

No plano metodológico, o achado mais relevante deste trabalho é a violação generalizada do pressuposto de riscos proporcionais em quatro das cinco covariáveis, detectada pelo teste de Schoenfeld. Esse resultado não é uma falha dos dados, mas sim uma característica estrutural do problema: em coortes neonatais, o perfil de risco muda qualitativamente ao longo do período de acompanhamento, pois os prematuros que sobrevivem à primeira semana enfrentam um conjunto de complicações distinto daquele que domina os primeiros dias. O modelo de Cox, ao impor um HR único e constante para cada covariável, produz uma média ponderada do efeito ao longo do tempo que pode ser enganosa quando o efeito real é concentrado em uma janela temporal específica. O RSF não impõe nenhum pressuposto sobre a forma da função de risco e, portanto, acomoda naturalmente essa variação temporal, tornando-o mais apropriado para a modelagem do problema em questão.

O C-index do RSF (0,5600) ficou ligeiramente abaixo do Cox (0,5765) no conjunto de teste, o que pode parecer contraditório dado o argumento anterior. Essa aparente contradição se dissolve quando se considera a natureza dos dados disponíveis: com apenas cinco covariáveis provenientes de um sistema administrativo, com apenas

cinco covariáveis provenientes de um sistema administrativo e limitada diversidade de informações clínicas, o RSF tem pouca estrutura não-linear para explorar. O ganho do RSF em relação ao Cox tende a ser substancial quando o conjunto de preditores é amplo, heterogêneo e inclui variáveis contínuas de alta resolução como dados de monitoramento clínico contínuo, resultados laboratoriais e características hospitalares. A ausência dessas variáveis nas bases do SINASC e do SIM é, portanto, a principal limitação da capacidade discriminativa de ambos os modelos, e não uma limitação intrínseca do RSF como método.

As desigualdades regionais identificadas pelos testes log-rank reforçam que o desfecho neonatal não é determinado apenas por fatores clínicos ao nascimento. O contraste entre Norte e Sul ($p = 2{,}32 \times 10^{-13}$) e entre Sudeste e Sul ($p = 6{,}30 \times 10^{-12}$) sugere que o acesso diferenciado a UTIs neonatais, a disponibilidade de surfactante exógeno e de suporte ventilatório especializado, e as condições socioeconômicas das mães operam como modificadores do prognóstico de forma independente das características clínicas ao nascimento. Esses fatores territoriais não estão disponíveis nas bases administrativas utilizadas, o que implica que os modelos ajustados neste trabalho subestimam sistematicamente a variância explicável do desfecho. A incorporação de variáveis de acesso a serviços de saúde em versões futuras do modelo representa, portanto, a principal via de melhora do poder discriminativo.

Em síntese, a contribuição metodológica central deste trabalho é demonstrar que, mesmo em um cenário de dados administrativos com poucos preditores, o RSF oferece vantagens diagnósticas relevantes: ele dispensa o pressuposto de proporcionalidade, produz estimativas de sobrevivência individualizadas sem forma paramétrica imposta, e quantifica a importância das variáveis por um critério preditivo direto que independe de testes de significância. A combinação do Cox como referência interpretável e do RSF como modelo preditivo flexível representa uma estratégia robusta para a análise de sobrevivência em coortes neonatais de grande porte, e pode ser diretamente estendida a bases de dados com maior riqueza clínica em trabalhos futuros.

5. Conclusões

Este trabalho demonstrou que o Random Survival Forest é uma abordagem metodologicamente mais adequada do que o modelo de Cox para a análise de sobrevida neonatal em prematuros, dado que o pressuposto de proporcionalidade dos riscos foi violado em quatro das cinco covariáveis avaliadas. A idade gestacional confirmou-se como o preditor dominante em ambos os modelos, com importância por permutação no RSF oito vezes superior à da segunda variável classificada, resultado convergente com a literatura e obtido por dois critérios independentes. O RSF revelou ainda contribuições não-lineares do peso ao nascer e da idade materna que o Cox não foi capaz de capturar, evidenciando que a não-significância estatística em modelos paramétricos não implica ausência de relevância preditiva. As desigualdades regionais expressivas na sobrevida neonatal, especialmente o contraste entre Norte e Sul do país, indicam que fatores relacionados ao acesso a cuidados intensivos neonatais exercem influência sobre o desfecho de forma independente das características clínicas ao nascimento, e que as bases administrativas disponíveis subestimam sistematicamente a variância explicável do problema. Como trabalho futuro, a incorporação de variáveis clínicas de maior resolução, como uso de surfactante, corticoterapia antenatal e suporte ventilatório, representa a principal via para aumentar o poder

discriminativo dos modelos além do C-index de 0,58 obtido neste estudo. A extensão da abordagem para análise de equidade racial e regional, bem como a estimação do impacto da pandemia de COVID-19 sobre a mortalidade neonatal por meio de séries temporais interrompidas, constitui a agenda imediata do módulo PreEpi do qual este trabalho faz parte.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio e financiamento. Este trabalho foi apoiado pela Rede SOFIA – Soluções em Observação, Formação, Inovação e Aplicação em Saúde Materno-Infantil, no âmbito do Programa Redes de Colaboração em Saúde Digital, coordenado pela Rede Nacional de Ensino e Pesquisa (RNP).

Referências

Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., Adler, A., Garcia, C. V., Rohde, S., Say, L., and Lawn, J. E. (2012). National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*, 379(9832):2162–2172.

Brasil. Ministério da Saúde. DATASUS (2022a). Sistema de informações sobre mortalidade (SIM). Disponível em: `http://www.datasus.gov.br`. Acesso em: maio 2026.

Brasil. Ministério da Saúde. DATASUS (2022b). Sistema de informações sobre nascidos vivos (SINASC). Disponível em: `http://www.datasus.gov.br`. Acesso em: maio 2026.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Victora, C. G., Aquino, E. M. L., do Carmo Leal, M., Monteiro, C. A., Barros, F. C., and Szwarcwald, C. L. (2011). Maternal and child health in Brazil: progress and challenges. *The Lancet*, 377(9780):1863–1876.

World Health Organization (2023). *Born too soon: decade of action on preterm birth*. World Health Organization, Geneva.