Predição de Óbito Neonatal usando Dados dos Sistemas de Informação do SUS e de Censo Demográfico

Jorge R. H. Moreira[1], **Heder S. Bernardino**[1], **Alex B. Vieira**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF)

{jorge.moreira,heder}@ice.ufjf.br, alex.borges@ufjf.br

Abstract. *Infant mortality rate is one of the most important indicators of a society. This rate has improved in recent years, but Brazil still faces challenges to reduce this rate. In this context, neonatal deaths represent the majority of cases, requiring more attention from the government. Thus, predicting the risks of a baby dying death in its first days of life can generate positive impacts on the public health system and, consequently, on Brazilian society. This work used variables present in the SUS Information Systems and the Demographic Census to generate a classifier that makes it possible to issue an alert to the health system in case of neonatal risks, directing attention to maternal and newborn monitoring. The results show an accuracy and sensitivity in the prediction that exceed 89%, showing feasibility in the use of techniques and methodological approaches proposed in the prediction of neonatal deaths and that can be extended for application and improvement of the health system.*

Resumo. *A taxa de mortalidade infantil é um dos indicadores mais importantes dentro de uma sociedade. Apesar da melhora nos últimos anos, o Brasil ainda enfrenta desafios para reduzir esse índice. Nesse contexto, os óbitos neonatais representam a maior parcela dos casos, exigindo mais atenção do poder público. Assim, predizer os riscos de um bebê morrer nos seus primeiros dias de vida pode gerar impactos positivos ao sistema de saúde público e, consequentemente, à sociedade brasileira. Este trabalho utilizou variáveis presentes nos Sistemas de Informação do SUS e de Censo Demográfico para gerar um classificador que possibilita emitir um alerta ao sistema de saúde em caso de riscos neonatais, direcionando a atenção ao acompanhamento materno e ao recém-nascido. Os resultados revelam uma acurácia e sensibilidade na predição que ultrapassam 89%, mostrando viabilidade no emprego das técnicas e abordagens metodológicas propostas na predição de óbitos neonatais e que podem ser estendidas para aplicação e aprimoramento do sistema de saúde.*

1. Introdução

A mortalidade infantil é um grave problema de saúde pública mundial, sendo evidenciado principalmente nos países mais pobres ou emergentes [MDH 2020]. Ela é medida por um indicador conhecido como "*taxa de mortalidade infantil*". Essa taxa corresponde à razão entre o número de óbitos de crianças menores de um ano de idade e a quantidade de nascidos vivos durante o ano (em determinado limite geográfico), multiplicados por mil [Silva et al. 2019]. De acordo com Kropiwiec et al. (2017), a taxa de

mortalidade infantil permite analisar a disponibilidade, a utilização e a eficácia dos cuidados de saúde, em especial, da atenção ao pré-natal, ao parto, ao recém-nascido e à criança no primeiro ano de vida. Mais ainda, o índice é frequentemente utilizado para definir políticas públicas direcionadas à saúde materno-infantil [Kropiwiec et al. 2017].

Apesar da taxa de mortalidade infantil ter diminuído nos últimos anos, ainda é considerada um desafio para os indicadores do sistema de saúde brasileiro [Ramos et al. 2017]. Para Monteiro et al. (2021), os valores desse índice ainda permanecem aquém da realidade de países desenvolvidos, necessitando que políticas públicas direcionadas para as gestantes sejam mais eficazes.

As reduções dos índices de mortalidade neonatal e materna estão listados nos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030 [ONU 2022]. No Brasil, o objetivo é reduzir as mortes evitáveis de recém-nascidos e crianças menores de 5 anos, visando reduzir a mortalidade neonatal para, no máximo, 5 por mil nascidos vivos; e a mortalidade de crianças menores de 5 anos para, no máximo, 8 por mil nascidos vivos [IPEA 2020]. Assim, o monitoramento da ocorrência dos óbitos infantis e neonatais, além do desenvolvimento de medidas direcionadas à redução do risco de morte nesse grupo de indivíduos, representa uma estratégia fundamental para a sociedade.

Lansky et al. (2014) acreditam que as causas de óbito no primeiro ano de vida estão geralmente associadas à prematuridade, às anomalias congênitas, à asfixia no parto, às infecções perinatais e aos fatores maternos. Alves et al. (2020) afirmam que os fatores associados ao óbito neonatal são influenciados pelas características biológicas maternas, suas condições sociais e os cuidados prestados às gestantes pelos serviços de saúde. Ocorrências de óbitos com essas características são encontradas em países com muita desigualdade social e sistemas de saúde fraco, onde as mulheres e crianças não têm acesso a cuidados essenciais e vitais [Soares et al. 2021].

Há diversos modelos propostos para relacionar mortes infantis e fatores sociais. Por exemplo, um modelo muito difundido foi proposto por Mosley and Chen (1984). Este modelo teórico divide os fatores relacionados ao óbito infantil em uma estrutura hierarquizada e, segundo os autores, os fatores distais como os socioeconômicos, determinam comportamentos, os quais, por sua vez, impactam um conjunto de fatores proximais, como os biológicos, agindo de maneira direta sobre o desfecho. Através de uma estrutura hierarquizada, é possível considerar e modelar fatores distintos de acordo com a precedência no tempo e relevância para a determinação do desfecho [Lima et al. 2008]. Conforme Garcia et al. (2019), esse tipo de modelo trouxe um grande avanço para o desenvolvimento de políticas públicas. Além de modelos complexos, há também abordagens que utilizam métodos tradicionais de regressão para identificação das determinantes do óbito infantil e neonatal [Lima et al. 2008, Zanini et al. 2011, Morais and Pereira 2020].

Neste trabalho, utilizamos técnicas e abordagens de aprendizado de máquina para compreender as interações entre os diferentes fatores e dimensões da mortalidade neonatal. Mais precisamente, nós prevemos o risco de óbito neonatal com técnicas de aprendizado de máquina aplicadas aos dados epidemiológicos do Sistema Único de Saúde (SUS) e do censo brasileiro entre 2012 e 2014. Em suma, são gerados modelos de classificação usando os métodos *Naive Bayes*, Regressão Logística, k-vizinhos mais próximos (KNN), Árvore de decisão, além dos métodos de comitês (*Ensemble*) florestas aleatórias (*Random Forests*) e estímulo adaptativo (*AdaBoost*) usando esses dados públicos de governo.

Além de fatores biológicos maternos e de nascimento, o conjunto de dados foi enriquecido adicionando variáveis que tratam questões interurbanas e socioeconômicas de cada região brasileira. Assim, diferente dos trabalhos que usam apenas os atributos presentes nos conjuntos de dados públicos dos sistemas de informação de saúde do SUS, como em Alves et al. (2020), propusemos uma abordagem mais ampla, com novas variáveis que carecem de investigação na influência do desfecho do óbito neonatal. Cabe destacar que o trabalho também foi pautado por etapas e técnicas de mineração de dados com enfoque preditivo. A avaliação e aplicação de teste de significância relevaram que os classificadores gerados baseados no *Random Forest* e Regressão Logística apresentaram os melhores resultados, acima de 89% em termos de acurácia, sensibilidade e *f1-score* na classificação do desfecho neonatal.

2. Trabalhos Relacionados

Há interesse em estudos demográficos e epidemiológicos que aplicam o processo de mineração de dados para classificação do desfecho neonatal. Por exemplo, Alves et al. (2020) investigam as características relacionadas ao risco de mortalidade neonatal no Brasil usando aprendizado de máquina. Os autores usam os conjuntos de dados públicos do SIM[1] e SINASC[2] entre os anos de 2006 e 2016. Por meio de uma amostra de 30 milhões de registros, o trabalho gerou um classificador para prever o risco de óbito neonatal com desempenho superior a 90% de área sob a curva ROC[3] (AUC[4]). O classificador foi aplicado na investigação da importância das variáveis usadas no estudo. Os resultados apontaram as variáveis peso do recém-nascido, Apgar[5] de primeiro e quinto minuto, malformações congênitas, semanas gestacionais e número de consultas de pré-natal as seis características mais expressivas para a classificação do risco de óbito neonatal.

Ramos et al. (2017) desenvolveram um sistema de análise de saúde que utiliza mineração de dados para gerar alertas de risco de morte de recém-nascidos. Os autores usam dados de óbitos e nascimentos do estado do Ceará registrados no SIM e SINASC entre 2013 e 2014. Dentre os métodos de aprendizado de máquina empregados no estudo, o método *Naive Bayes* se destacou com uma AUC de 92,1% e desempenho geral de 98,2% na classificação de risco de óbito de recém-nascidos.

Batista et al. (2021) usaram dados de nascimento e óbitos do município de São Paulo no período de 2012 a 2017. Os pesquisadores geraram um modelo usando o algoritmo *Extreme Gradient Boosting Trees (XGBoost)*, que obteve uma AUC de 97% e F1-score de 55% para prever a mortalidade neonatal.

Silva et al. (2020) também usaram dados dos Sistemas de Informação do SUS (SIM e SINASC) para descoberta de conhecimento no contexto da mortalidade infantil. Os pesquisadores aplicaram o algoritmo baseado em árvore C5.0 para gerar um classificador de desfecho de óbito infantil. Para tratar o desbalanceamento das classes usaram

[1] Sistema de Informações sobre Mortalidade (SIM): sistema de vigilância epidemiológica nacional criado pelo Datasus para obter de forma regular dados sobre mortalidade em todo território nacional.

[2] Sistema de Informações sobre Nascidos Vivos (SINASC): sistema mantido pelo DATASUS com o objetivo de reunir informações epidemiológicas referentes aos nascidos vivos em todo território nacional.

[3] *Receiver Operating Characteristic* (ROC): métrica de desempenho para problemas de classificação. Representa a relação da taxa de verdadeiro positivo (sensibilidade) e da taxa de falso positivo (especificidade), variando o ponto de corte na probabilidade estimada (*threshold*).

[4] *Area Under the Curve* (AUC): área sob a curva ROC. Representa o grau de separabilidade do modelo. Quanto maior a AUC, melhor será o modelo em distinguir entre casos de óbitos e não-óbitos neonatais.

[5] Teste de Apgar: utilizado para avaliar o estado geral e a vitalidade do recém-nascido, ajudando a identificar se é necessário qualquer tipo de tratamento ou cuidado médico especial após o nascimento.

o *Synthetic Minority Oversampling Technique (SMOTE)* como técnica de sobreamostragem (*oversampling*), que consiste na criação de instâncias sintéticas da classe minoritária. Usaram a divisão de treinamento e teste estratificado com validação cruzada *k-fold* com k=10. Os experimentos se basearam na média da execução de 30 vezes do C5.0, chegando a uma acurácia de 98,58% e AUC de 72,6%. Conforme o estudo, concluíram ainda que as três principais características relacionadas à morte de uma criança antes de um ano de idade são o peso ao nascer, o escore de Apgar de cinco minutos e as semanas de gestação.

O trabalho atual adotada uma nova abordagem, adicionando além de características presentes nos sistema de informação em saúde do SUS e que são comumente usadas nos estudos demográficos e epidemiológicos com foco em mortalidade infantil/neonatal, outros fatores influenciadores como o índice de desenvolvimento humano municipal, presença e capacidade de centros obstétricos e pediátricos dos estabelecimentos de saúde e distância no deslocamento entre residência e local de nascimento.

3. Apresentação, Preprocessamento e Análise dos Dados

Para compreender melhor o cenário histórico de nascimentos e óbitos neonatais no Brasil, os Órgãos e Entidades Nacionais responsáveis por esse acompanhamento, recorrem aos dados registrados nos Sistemas de Informação em Saúde do SUS. Esses sistemas agregam bases de dados de abrangência nacional com a função de registrar dados ligados à saúde pública. Podemos citar dois grandes conjuntos de dados largamente utilizadas na epidemiologia, que são o Sistema de Informação sobre Nascidos Vivos (SINASC) e o Sistema de Informação sobre Mortalidade (SIM). O SINASC tem por propósito quantificar nascidos vivos e fornecer informações sobre a gravidez, o parto e as condições da criança ao nascer. Esse sistema armazena informações que propiciam a construção de diagnóstico das condições de nascimento, o que possibilita a realização de ações de promoção, de prevenção e de planejamento em saúde. Por outro lado, o SIM tem a finalidade de reunir dados sobre óbitos ocorridos no Brasil, sendo considerado uma importante ferramenta de gestão na área da saúde que subsidia a tomada de decisão em diversas áreas da vigilância e assistência à saúde.

Neste trabalho foram utilizados os conjuntos de dados disponibilizados na Plataforma de Ciência de Dados da Fiocruz (PCDas)[6] que reúne, dentre outros, os microdados dos sistemas SINASC e SIM do DATASUS[7]. Foram usados os registros de nascidos vivos no período de 2012 a 2014, contabilizando um total de 8.789.075 de nascidos vivos de mães residentes nas 27 unidades federativas brasileiras. A escolha dessa janela temporal visou buscar um cenário de menos inconsistências e mais correspondências para vinculação dos registros entre os atributos presentes nas bases. São as seguintes motivações existentes para essa escolha:

• a publicação da portaria[8] Nº 116 de 11 de fevereiro de 2009, que regulamenta a coleta de dados, fluxo e periodicidade de envio das informações sobre óbitos e nascidos vivos para os Sistemas de Informação em Saúde, trazendo diretrizes quanto obrigatoriedade da emissão de Declaração de Nascido Vivo (DN) para todo nascido vivo, independente da duração da gestação, peso e estatura do recém-nascido;

[6] https://pcdas.icict.fiocruz.br/conjunto-de-dados/

[7] Departamento de Informática do Sistema Único de Saúde (DATASUS): responsável por prover os órgãos do SUS de sistemas de informação e suporte de informática, necessários ao processo de planejamento, operação e controle do SUS. Fonte: https://datasus.saude.gov.br/sobre-o-datasus/

[8] https://bvsms.saude.gov.br/bvs/saudelegis/svs/2009/prt0116_11_02_2009.html

• a vigoração da lei Nº 12.662, de 5 de junho de 2012, que assegura validade nacional à Declaração de Nascido Vivo - DNV;
• as mudanças em variáveis e na forma de coleta da Declaração de Nascido Vivo a partir de 2011, prevendo aumento na cobertura em todas as regiões, conferindo uma maior representatividade as informações geradas a partir do SINASC [Ministério da Saúde 2011]; e
• a presença do atributo com o número da declaração de nascido vivo para vinculação de registros entre os conjuntos de dados dos sistemas SINASC e SIM do DATASUS.

O estudo atual considera a divisão regional brasileira em vigor conforme IBGE (2022), como mostra a Tabela 1. As informações sobre a quantidade de nascidos vivos, óbitos neonatais e a taxa de mortalidade neonatal extraídas dos conjuntos de dados estão disponíveis nesta tabela. Os dados apresentados corroboram a hipótese da existência de possíveis diferenças entre as regiões do país em relação aos fatores relacionados ao óbito infantil/neonatal. Esse quadro motiva avançar para uma maior investigação das questões ligadas às variáveis demográficas, socioeconômicas e de assistência à saúde na gestação e parto associados sob a visão regional brasileira.

Tabela 1. Total de nascidos vivos (#NV), quantidade de óbitos neonatais (#ON) e taxa de mortalidade neonatal (#TMN), por região, entre 2012-2014.

		2012	2013	2014
NORTE	**#NV**	308375	313272	321682
	#ON	3338	321682	3320
	TMN(%)	10,82	10,70	10,32
NORDESTE	**#NV**	832631	821458	833090
	#ON	8970	8999	8594
	TMN(%)	10,77	10,95	10,32
CENTRO OESTE	**#NV**	230279	234687	245076
	#ON	2203	2185	2199
	TMN(%)	9,57	9,31	8,97
SUDESTE	**#NV**	1152846	1147627	1182949
	#ON	9659	9385	9634
	TMN(%)	8,38	8,18	8,14
SUL	**#NV**	381658	386983	396462
	#ON	2943	2808	2999
	TMN(%)	7,71	7,26	7,56

A Figura 1 mostra as quantidades de óbitos neonatais e infantis entre 2012 e 2015. Pode-se observar uma quantidade maior de óbitos neonatais em relação aos infantis. Além disso, essas quantidades tendem a permanecer constantes aos longo do período estudado.

No intuito de abranger mais aspectos das dimensões da mortalidade neonatal, foram utilizados os dados do Cadastro Nacional de Estabelecimentos de Saúde (CNES), dados do Índice de Desenvolvimento Humano do Programa das Nações Unidas para o Desenvolvimento (PNUD)[9], e os códigos dos municípios publicados pelo Instituto Brasileiro de Geografia e Estatistica (IBGE)[10]. O conjunto de dados final foi obtido a partir do processo de vinculação de registros (*linkage*), como apresentado na Figura 2. A Tabela 2 resume o total de registros por região encontrados após o processo de *linkage*.

Etapa 1: Vinculação dos registros de nascimento (SINASC) e óbitos neonatais (SIM) por meio do atributo "*número de declaração de nascido vivo (NUMERNDN)*" presente em ambos conjuntos de dados. Nesta etapa também foi realizada a rotulagem das classes,

[9] https://www.br.undp.org/content/brazil/pt/home/idh0/rankings/idhm-municipios-2010.html
[10] https://www.ibge.gov.br/explica/codigos-dos-municipios.php

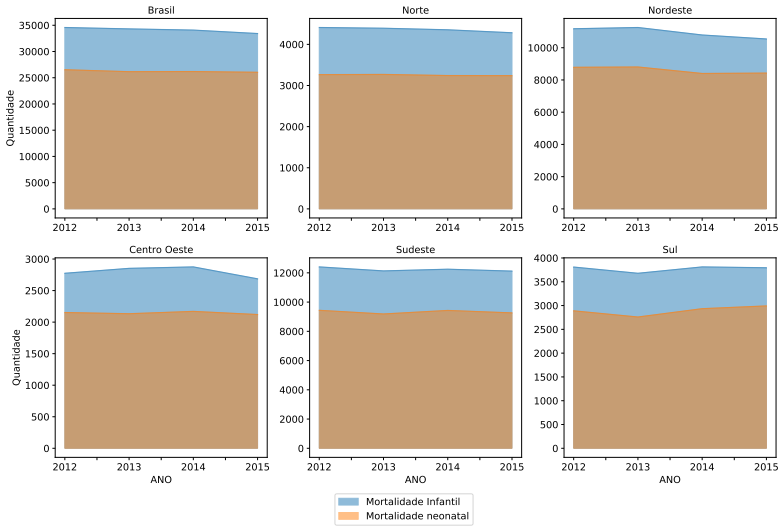

Figura 1. Quantidades de óbitos neonatais e infantis de 2012 a 2015.

Tabela 2. Total de registros após processo de linkage.

	NORTE		NORDESTE		CENTRO OESTE		SUDESTE		SUL	
	VIVOS	ÓBITOS	VIVOS	ÓBITOS	VIVOS	ÓBITOS	VIVOS	ÓBITOS	VIVOS	ÓBITOS
2012	285136	2040	711927	4793	225590	1419	1125583	6439	372439	2514
2013	293487	2192	699486	4757	229921	1545	1120922	6721	378826	2449
2014	302197	2176	709846	4776	236962	1611	1155839	6937	387863	2660
TOTAL	**880820**	**6408**	**2121259**	**14326**	**692473**	**4575**	**3402344**	**20097**	**1139128**	**7623**

onde os registros de óbitos até os 27 primeiros dias de vida foram separados e rotulados como casos de óbitos neonatais. Esses registros foram removidos do conjunto de dados de nascimentos, sendo os demais casos rotulados como classe de não-óbitos.
Etapa 2: Utilizando o "*código do estabelecimento de saude (CNES)*", bem como o mês e ano de competência do CNES, foi possível realizar a vinculação dos registros entre os conjuntos de dados. Com isso, formou-se uma base rotulada de nascidos vivos e óbitos neonatais com os respectivos dados do estabelecimento de saúde do nascimento.
Etapa 3: Vinculação dos registros do IDHM/PNUD com os municípios.
Etapa 4: Vinculação dos registros dos conjuntos de dados gerados nas etapa 3 e 4 por meio do "*código dos municípios*".
Etapa 5: Enriquecimento do conjunto de dados gerado adicionando um atributo que corresponde à distância entre a residência da gestante e o local de nascimento.

Antes de cada processo de vinculação de registros, na fase de pré-processamento, foi feita uma análise dos atributos chaves usados na vinculação, a exemplo do "*NUMERNDN*". Os registros que continham o atributo chave faltante ou inconsistente foram eliminados. Após o processo de vinculação, os dados foram separados por classe e região, aplicando os seguintes procedimentos nos demais atributos não chave que estava faltantes ou com preenchimento de "*valor ignorado*", ou seja, sem dados disponíveis:
Categórico nominal e numéricos discretos: preenchimento ou substituição pelo valor mais frequente do atributo;
Numéricos contínuos: preenchimento ou substituição pela média dos valores do atributo.

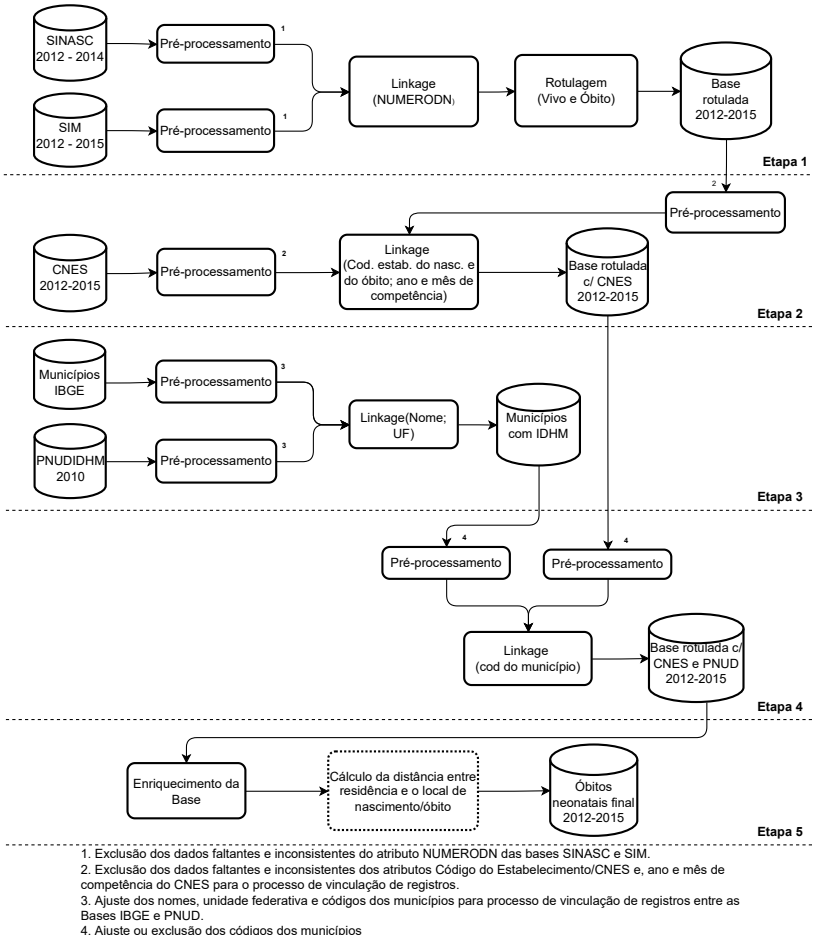

Figura 2. Etapas da vinculação de registros.

Para o atributo "*idade materna*", foi definido um intervalo entre 10 e 49 anos[11]. Não foi identificada idade materna significativamente discrepante e que se distanciasse significativamente do valor médio deste atributo (*outlier*). Todas as variáveis foram categorizadas para aplicação nos modelos de acordo com a Tabela 3. Na fase de transformação, as variáveis categóricas foram transformadas em variáveis *dummy*. Este processo gera N novos atributos, onde N é dado pelo número de valores únicos. A categoria do atributo de correspondência de posição é preenchida com 1 (um), enquanto o restante das posições é preenchida com o valor 0 (zero).

Foi feita uma análise exploratória para observar os atributos separados por região e por rótulo da classe em busca de diferenças na relação entre as classes de vivos e casos de óbitos. Nessa fase, destacamos:
(i) Em todas as regiões, o maior número de casos, sejam de vivos ou óbitos neonatais, ficaram entre mães na faixa de idade de 20 a 30 anos, representando mais de 22% de toda a amostra de dados desse atributo.
(ii) Em relação ao número de consultas durante o pré-natal, foi observado uma diferença de frequência entre as classes. Mulheres que perderam seus filhos nas primeiras semanas de vida fizeram menos que sete consultas pré-natal durante à gestação.

[11] https://conselho.saude.gov.br/ultimas_noticias/2007/politica_mulher.pdf

Tabela 3. Atributos e categorias usadas no estudo.

Variáveis	Categorias
Demográficas e socioeconômicas maternas	
Idade materna	(10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49)
Raça/cor da pele	(Branca, Preta, Amarela, Parda, Indígena)
Estado civil	(Solteiro, Casado, Viúvo, Separado judicialmente/divorciado, União estável)
Escolaridade	(Nenhuma, de 1 a 3 anos, de 4 a 7 anos, 8 a 11 anos, 12 anos e mais)
Município de residência faz parte da Amazônia Legal	(Sim/Não)
Município de residência faz parte da faixa de fronteira	(Sim/Não)
Município de residência é capital de UF	(Sim/Não)
Índice de Desenvolvimento Humano Municipal 2010	(Alto, Baixo, Muito baixo, Médio)
Referentes à atenção/assistência à saúde na gestação e parto	
Local de nascimento	(Hospital, Outros estabelecimentos de saúde, Domicílio, Via pública, Outros)
Número de consultas durante o pré-natal	(Nenhuma, de 1 a 3, de 4 a 6, 7 e mais)
Tipo de parto	(Vaginal, Cesáreo)
Número de semanas de gestação	(<22, 22-27, 28-31, 32-36, 37-41, 42 ou mais)
Mês de gestação em que iniciou o pré-natal	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Trabalho de parto induzido	(Sim, Não, Não se aplica)
Cesárea ocorreu antes do trabalho de parto iniciar	(Sim, Não, Não se aplica)
Assistência do nascimento	(Médico, Enfermeira/obstetriz, Parteira, Outros)
Município de nascimento faz parte da Amazônia Legal	(Sim/Não)
Município de nascimento é capital de UF	(Sim/Não)
Distância entre residência e local de nascimento (km)	(<50, 50-100, 100-150, 150-200, 200 ou mais)
Qtd. de salas de pré parto em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de parto normal em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de curetagem em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de cirurgias em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de leitos de pré parto em centro obstétrico	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Referentes às condições de nascimento do recém-nascido	
Semana de gestação	(Menos de 22 semanas, 22 a 27 semanas, 28 a 31 semanas, 32 a 36 semanas, 37 a 41 semanas, 42 semanas ou mais)
Tipo de gravidez	(Única, Dupla, Tripla e mais)
Tipo de apresentação do RN	(Cefálico, Pélvica ou podálica, Transversa)
Número de gestações anteriores	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Número de partos vaginais	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Número de partos cesáreos	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Número de filhos vivos	(0-2, 3-5, 6-8, 9-11, 12 ou mais)
Número de filhos mortos	(0-2, 3-5, 6-8, 9-11, 12 ou mais)
Sexo	(Masculino, Feminino)
Apgar no primeiro minuto	(0-3: grave, 4-6: moderado, 7: leve, 8-10: ótimo)
Apgar no quinto minuto	(0-3: grave, 4-6: moderado, 7: leve, 8-10: ótimo)
Peso ao nascer em gramas	(<2500, 2500-2999, 3000-3999, 4000 ou mais)
Anomalia congênita	(Ignorado; Sim; Não)
Prematuridade do nascimento	(Termo, Inconclusivo-IG, Inconclusivo-Peso, Prematuro)
Qtd. salas/consultórios atend. pediátrico urg./emerg.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. salas de repouso/observação pediátrico urg./emerg.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de salas de repouso ou obs. pediátrico de atend. ambul.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de leitos repouso ou obs. pediátrico de urg./emerg.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Qtd. de leitos de repouso ou obs. pediátrico de atend. ambul.	(0-3, 4-7, 8-11, 12-15, 16-19, 20 ou mais)
Existência instalação física atend. hosp. em centro obstétrico	(Sim/Não)

(iii) Observou-se também, em todas as regiões, que os óbitos neonatais eram mais evidentes quando a gestação ocorreu em menos de 22 semanas.
(iv) A frequência relativa do tipo de apresentação pélvica ou podálica do recém-nascido na classe óbito superou em aproximadamente 10% os casos da classe vivo.
(v) Outro atributo que incorreu em diferenças significativas entre as classes foram os índices de vitalidade Apgar de um e cinco minutos (Apgar1 e Apgar5). O nível grave de ambos os índices marcou presença nos casos de óbito com um diferença de aproximadamente 30% em relação à mesma categoria da classe dos bebês que não foram a óbito.

(vi) O atributo peso ao nascer também apresentou diferenças marcantes entre as classes. Os casos de óbitos em que o peso ao nascer eram menores que 2500 gramas, mostrou aproximadamente 60% a mais de casos em relação a mesma categoria da classe vivo.
(vii) Em todas as regiões, o fator prematuridade se mostrou expressivo em valores absolutos na classe óbito. Foi observada uma diferença de aproximadamente 50% dos casos de óbitos neonatais em relação a classe vivo.

4. Métodos de Classificação

Neste estudo, é utilizada a abordagem supervisionada de classificação binária. Por esta abordagem, há a presença das classes rotuladas de bebês que sobreviveram ou que foram ao óbito nos seus 27 primeiros dias de vida. Os seguintes métodos de classificação foram usados aqui para a geração de modelos: *Naive Bayes*, Regressão Logística, k-vizinhos mais próximos (KNN), Árvore de decisão, além dos métodos de comitês (*Ensemble*) florestas aleatórias (*Random Forests*) e estímulo adaptativo (*AdaBoost*).

Para todos os métodos, foi utilizada a estratégia de validação cruzada *k-fold* com k=10. A validação cruzada consiste em dividir os dados em k subconjuntos chamados folds e, a cada iteração, um desses conjuntos é usado para testar os classificadores gerados, sendo os demais, usados no treinamento. A variante estratificada mantém a proporção original das classes nos conjuntos (*folds*).

Utilizamos também uma abordagem de subamostragem: método de balanceamento de classes que reduz o número de observações da classe majoritária para balanceamento das classes. Para isso, extraímos aleatoriamente da classe majoritária (casos de não óbito neonatal) um conjunto de mesma proporção da classe minoritária durante o processo de treinamento dos modelos. Como limitação do uso dessa abordagem, citamos a possível perda de informação causada pela eliminação de registros representativos da classe majoritária [Castro and Braga 2011]. Os modelos foram avaliados por meio da matriz de confusão com a frequências de classificação para cada classe do modelo, e também, o resultado das métricas de precisão, sensibilidade (*recall*), *f1-score*, acurácia e AUC (*Area Under The Curve*) para análise da relação entre a sensibilidade e a especificidade dos classificadores gerados.

5. Experimentos, Resultados e Discussões

Todos os experimentos foram realizados em uma máquina com 10 núcleos de CPU e 64 GB de RAM, rodando Ubuntu 20.04 (64 bits). Foram utilizados também a linguagem de programação Python 3.6 com as bibliotecas Numpy (1.20.3), Pandas (1.3.2), Scikit-Learn (0.24.2) e MatplotLib (3.3.4). Os resultados foram determinados pela média e desvio padrão de 30 execuções dos algoritmos.

Para cada método, foi gerada a matriz de confusão conforme mostrado na Figura 3. Os classificadores obtiveram taxas de acerto superiores a 80% na previsão das classes. Para melhor visualização e compreensão dos resultados foi gerada a Tabela 4 com o resultado das métricas: acurácia, precisão, sensibilidade e *F1-score*.

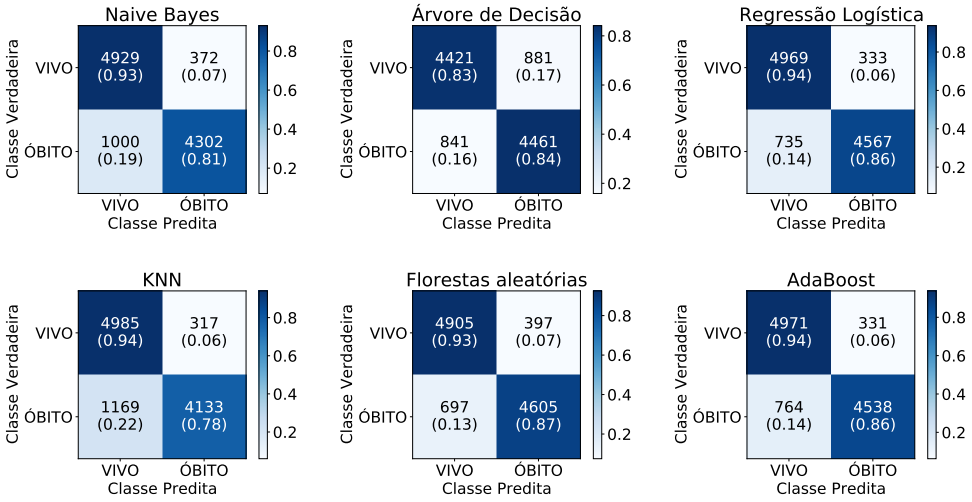

Figura 3. Matrizes de confusão com resultados médios.

Tabela 4. Médias e Desvios Padrões (DP) dos resultados alcançados.

	Acurácia		**Precisão**		**Sensibilidade**		**F1 score**	
	Média	**DP**	**Média**	**DP**	**Média**	**DP**	**Média**	**DP**
Naive Bayes	0.8704	0.0001	0.8345	0.0003	0.924	0.0003	0.877	0.0001
Regressão Logística	0.8998	0.0001	0.8721	0.0001	0.9372	0.0001	0.9034	0.0001
Árvore de Decisão	0.8341	0.0009	0.8396	0.001	0.8259	0.0014	0.8327	0.001
K-vizinhos mais próximos	0.8608	0.0005	0.8112	0.0005	0.9405	0.0006	0.871	0.0005
Florestas aleatórias	0.8971	0.0003	0.8773	0.0003	0.9232	0.0005	0.8997	0.0003
AdaBoost	0.8972	0.0002	0.8676	0.0003	0.9375	0.0003	0.9012	0.0002

Os classificadores avaliados apresentam desempenho médio semelhante e desvio padrão baixo. Em termos de desempenho geral dos modelos, os melhores resultados foram alcançados pelos métodos de Regressão Logística, *Random Forest* e *AdaBoost* com sensibilidades acima de 90%.

Ainda no processo de avaliação dos modelos foram projetadas as curvas ROC e extraídas suas respectivas áreas (AUC) visando observar a variação da sensibilidade e especificidade, para diferentes valores de corte. Por limitação de espaço, a figura com as curvas não são apresentadas no artigo, mas estão disponíveis em material suplementar[12]. Os classificadores Regressão Logística, *Random Forest* e *AdaBoost* apresentam curvas ROC quase sobrepostas e áreas de aproximadamente 89%. Os resultados mostram uma relação sensibilidade e especificação adequada, representando um bom indicador de desempenho dos modelos na classificação de óbitos neonatais.

Visando investigar as diferenças estatísticas entre o resultado dos algoritmos propostos foi conduzido os testes estatísticos de Friedman e *post-hoc* de Nemenyi[13] com um nível de confiança de 95%. O teste de Friedman resultou em zero, representando a existência de diferença entre o resultado da execução dos métodos propostos. Como a distância crítica do teste de Nemenyi foi de 1,377, podemos considerar que em relação a métrica acurácia dentre os algoritmos com maiores valores a Regressão Logística não possui diferença estatística significativa em relação ao *Random Forest*. Para a métrica

[12] http://netlab.ice.ufjf.br/index.php/neonataldeathsprediction/

[13] https://www.rdocumentation.org/packages/tsutils/versions/0.9.2/topics/nemenyi

precisão, o destaque fica para os métodos de Regressão Logística e *AdaBoost* que não possuem diferença estatística significativa entre seus resultados. Para as métricas sensibilidade e *F1-score*, a Regressão Logística e o *Random Forest* apresentam os melhores resultados sem diferença estatisticamente significantes entre os dois. Todos os testes são apresentados em figuras no material suplementar.

Tendo em vista que os modelos gerados com os métodos *Random Forest* e Regressão Logística obtiveram os melhores resultados em termos de acurácia e sensibilidade, foi extraído destes modelos a relevância dos atributos quanto ao desfecho para a predição de óbito neonatal. No *Random Forest*, as vinte variáveis mais importantes para o modelo de classificação estão relacionadas ao peso ao nascer, teste de Apgar de um e cinco minutos, prematuridade, semanas de gestação, anomalia congênita, número de consultas, número de semanas de gestação e idade materna. Variáveis relacionadas ao IDHM do município de nascimento e capacidade dos centros de saúde apareceram entre as 50 variáveis mais relevantes para o modelo. Já fatores como a distância entre residência e nascimento ficaram entre as 100 variáveis mais importantes. No classificador gerado usando o método Regressão Logística, as variáveis prematuridade, semanas gestacionais e teste de Apgar de um e cinco minutos ainda se destacam na sinalização do risco de óbito neonatal.

6. Conclusões e Trabalhos Futuros

Neste artigo, utilizamos variáveis presentes nos sistemas de informação do SUS e do Censo Demográfico brasileiro para gerar classificadores para predizer o desfecho neonatal. Os classificadores consideram dados biológicos maternos, fatores socioeconômicos, dados da criança ao nascer, além de questões assistenciais e de atenção à saúde da gestante e do bebê. Além disso, consideramos características e fatores influenciadores do desfecho neonatal sob uma visão regional e de um conjunto de dados enriquecido com novas variáveis.

Os classificadores desenvolvidos são capazes de expor o risco de morte neonatal com alta acurácia e sensibilidade acima de 89%. Os classificadores Regressão Logística e o baseados em árvore (*Random Forest* e *AdaBoost*) se mostraram os mais eficientes na predição do óbito neonatal. O estudo mostrou também que as variáveis peso ao nascer, score de Apgar de um e cinco minutos, prematuridade, semana de gestação, anomalia congênita, número de consultas, número de semanas de gestação e idade materna são as que mais influenciam na predição do desfecho neonatal de acordo com os classificadores usando os métodos *Random Forest* e Regressão Logística. Apesar do emprego das técnicas de aprendizado máquina e abordagens metodológicas propostas neste estudo serem atemporais, enxergamos que a janela temporal do conjunto de dados utilizado no trabalho apresenta-se como um fator de limitação do estudo, carecendo de investigação com dados mais atuais e que podem revelar novos indícios do desfecho da mortalidade neonatal. Outra limitação do trabalho é a utilização direta de tratamento do desbalanceamento dos dados para geração dos modelos sem abordar e comparar resultados com outras técnicas de amostragem, o que pode gerar viés e alterar a realidade no tratamento do problema.

Para pesquisas futuras, pretende-se investigar outras abordagens e métodos de aprendizado de máquina, principalmente com combinações de classificadores *bagging* e *boosting* baseados em árvores, comparar e discutir o desempenho dos classificadores

com outras abordagens de amostragem, seleção de atributos e dados mais recentes.

Referências

Castro, C. L. and Braga, A. P. (2011). Mortalidade neonatal precoce em um hospital terciário do nordeste brasileiro. *Controle & Automação Sociedade Brasileira de Automatica*, 22:441–466.

IPEA (2020). Objetivos de desenvolvimento sustentável da onu. https://www.ipea.gov.br/ods/ods3.html. Acessado em: 03 de janeiro de 2022.

Kropiwiec, M. V., Franco, S. C., and do Amaral, A. R. (2017). Fatores associados à mortalidade infantil em município com índice de desenvolvimento humano elevado. *Revista Paulista de Pediatria*, 35.

Lima et al. (2008). Proposta de modelo hierarquizado aplicado à investigação de fatores de risco de óbito infantil neonatal. *Cadernos de Saúde Pública*, 24:1910–1916.

MDH (2020). Ministério da mulher, da família e dos direitos humanos. mortalidade e saúde infantil. `https://www.gov.br/mdh/pt-br/navegue-por-temas/crianca-e-adolescente/dados-e-indicadores/mortalidade-e-saude-infantil`. Acessado em: 03 de janeiro de 2022.

Ministério da Saúde (2011). Consolidação do sistema de informações sobre nascidos vivos 2011. `http://tabnet.datasus.gov.br/cgi/sinasc/Consolida_Sinasc_2011.pdf`. Acessado em: 10 de janeiro de 2022.

Morais, A. and Pereira, A. (2020). Mortalidade neonatal precoce em um hospital terciário do nordeste brasileiro. *Revista da Sociedade Brasileira de Enfermeiros Pediatras*, 19:89–96.

ONU (2022). Plataforma agenda 2030. disponível em: http://www.agenda2030.org.br/. http://www.agenda2030.org.br. Acessado em: 03 de janeiro de 2022.

Ramos et al. (2017). Using predictive classifiers to prevent infant mortality in the brazilian northeast. In *IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6.

Silva et al. (2019). Desigualdades socioeconômicas: uma análise sobre os determinantes da taxa de mortalidade infantil nos municípios brasileiros. *Revista Brasileira de Estudos Regionais e Urbanos*, 13(1):73–97.

Soares et al. (2021). Caracterizando a mortalidade infantil utilizando técnicas de machine learning: um estudo de caso em dois estados brasileiros - Santa Catarina e Amapá. *Brazilian Journal of Development*, 7.

Zanini et al. (2011). Determinantes contextuais da mortalidade neonatal no Rio Grande do Sul por dois modelos de análise. *Revista de Saúde Pública*, 45:79–89.