ESCOLA SUPERIOR DE CIÊNCIAS DA SANTA CASA DE MISERICÓRDIA DE VITÓRIA – EMESCAM
GRADUAÇÃO EM MEDICINA

FABIANO NOVAES BARCELLOS FILHO

SELEÇÃO DE VARIÁVEIS COM MACHINE LEARNING PARA IDENTIFICAR FATORES DE RISCO ASSOCIADOS À MORTALIDADE INFANTIL

Vitória
2022

FABIANO NOVAES BARCELLOS FILHO

SELEÇÃO DE VARIÁVEIS COM MACHINE LEARNING PARA IDENTIFICAR FATORES DE RISCO ASSOCIADOS À MORTALIDADE INFANTIL

Trabalho de Conclusão de Curso apresentado à coordenação do curso de graduação em Medicina da Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM como requisito parcial para obtenção do grau de Bacharel em Medicina.

Orientador(a): Prof.ª Dr.ª Patrícia Casagrande Dias de Almeida

Coorientador: Prof. Dr. Gustavo Carreiro Pinasco

Vitória

2022

FABIANO NOVAES BARCELLOS FILHO

SELEÇÃO DE VARIÁVEIS COM MACHINE LEARNING PARA IDENTIFICAR FATORES DE RISCO ASSOCIADOS À MORTALIDADE INFANTIL

Trabalho de Conclusão de Curso apresentado à coordenação do curso de graduação em Medicina da Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM como requisito parcial para obtenção do grau de Bacharel emMedicina.

Aprovado em 13 de Junho de 2023.

BANCA EXAMINADORA

DocuSigned by:
FB9B4A389B4F4DE...

Prof.ª Dr.ª Patrícia Casagrande Dias de Almeida
Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM
Orientador(a) e avaliador(a)

DocuSigned by:
Prof. Dr. Gustavo Carreiro Pinasco
00D32E6B57F648D...

Prof. Dr. Gustavo Carreiro Pinasco
Universidade Federal do Espírito Santo - UFESCoorientador e avaliador

DocuSigned by:
Prof.ª Dr.ª Rachel Mocelin Dias Coelho
F47782BA7C30420...

Prof.ª Dr.ª Rachel Mocelin Dias Coelho
Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM
Avaliador(a)

DocuSigned by:
Prof.ª Dr.ª Andrea Lube Pereira
DF0AFEBE1EA0403...

Prof.ª Dr.ª Andrea Lube Pereira
Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória – EMESCAM
Avaliador(a)

DS
FNBF

Gostaria de expressar minha profunda gratidão a Deus, aos meus pais e às instituições parceiras que contribuíram para o meu estudo. Agradeço também à Prefeitura de Vitória pela generosa contribuição dos dados. A todos os professores, amigos e familiares que me apoiaram, meu sincero agradecimento por seu encorajamento e crença em meu potencial.

RESUMO

Introdução: A taxa de mortalidade infantil é uma medida utilizada para avaliar diversas características de uma população, como por exemplo a qualidade de vida e os cuidados de saúde de sua população infantil. Foi verificada a associação entre fatores de risco biológicos e óbito nos primeiros 28 dias de vida, descrita por vários autores, em especial no período perinatal. Sabe-se que, mesmo ao notificar dados de informações de nascimento e mortalidade infantil, a integração e o estudo dos dados gerados ainda são escassos. Nesse contexto, esse estudo espera fornecer uma alternativa viável e efetiva para a identificação e compreensão dos fatores de risco associados à mortalidade infantil, contribuindo para o desenvolvimento de políticas públicas mais abrangentes e contextualizadas. **Objetivo**: Desenvolver um algoritmo de predição de mortalidade infantil interpretável com base em dados do Sistema de Informação de Nascidos Vivos e de Sistema de Informação de Mortalidade. **Métodos**: Estudo tipo coorte retrospectiva em que serão avaliados todos os casos de óbito infantil em ambos os sexos, conforme dados do Sistema de Informação de Mortalidade e do Sistema de Informação de Nascidos Vivos da Secretaria Municipal de Saúde de Vitória (SEMUS), entre os anos 2000-2019. Os bancos de dados serão integrados por *linkage* determinístico. Foi feita análise exploratória e pré-processamento dos dados faltantes e categóricos. Em seguida, foi realizado o treinamento dos modelos de Machine Learning para criação do modelo preditivo de mortalidade infantil e, assim, obteve-se a interpretabilidade com os fatores mais importantes para a predição com o método SHAP. **Resultados:** Pode-se observar que os resultados da interpretabilidade do modelo de ML coincidem com os resultados da meta-análise que utilizou as mesmas bases de dados de forma nacional, na qual os fatores que mais influenciaram o desfecho final da mortalidade foram Peso, APGAR, Idade Gestacional e Presença de Anomalias. O uso de técnicas de interpretabilidade, como o SHAP, é bastante promissor para a seleção e identificação de fatores de risco populacionais relacionados à mortalidade infantil, utilizando bancos de dados existentes sem a necessidade de novos estudos populacionais. Além disso, esse conhecimento pode ser usado para auxiliar na tomada de decisões em saúde pública.

Palavras-chave: 1. Mortalidade Infantil; 2. Fatores de risco; 3. Mineração de dados; 4. Seleção de recursos; 5. Inteligência Artificial Explicável.

ABSTRACT

Introduction: The infant mortality rate is a measure used to assess various characteristics of a population, such as the quality of life and healthcare for its infant population. The association between biological risk factors and death in the first 28 days of life has been observed by several authors, especially during the perinatal period. It is known that despite the reporting of data on birth information and infant mortality, the integration and study of the generated data are still scarce. In this context, this study aims to provide a viable and effective alternative for identifying and understanding the risk factors associated with infant mortality, contributing to the development of more comprehensive and contextualized public policies. **Objective:** To develop an interpretable machine learning algorithm for predicting infant mortality based on data from the Live Birth Information System and the Mortality Information System. **Methods:** This retrospective cohort study will evaluate all cases of infant death in both sexes, based on data from the Mortality Information System and the Live Birth Information System of the Municipal Health Department of Vitória (SEMUS), between the years 2000-2019. The databases will be integrated through deterministic linkage. Exploratory analysis and preprocessing of missing and categorical data will be performed. Machine learning models were trained to create the predictive model of infant mortality, followed by interpretability analysis using the SHAP method to identify the most important factors for prediction. **Results:** It was observed that the interpretability results of the ML model coincide with the results of the meta-analysis that used the same national databases, in which the factors that most influenced the outcome of mortality were: Weight, APGAR score, Gestational Age, and Presence of Anomalies. The use of interpretability techniques, such as SHAP, is very promising for selecting and identifying population risk factors related to infant mortality, using existing databases without the need for new population studies. Moreover, this knowledge can be used to assist in making decisions in public health.

Keywords: 1. Infant Mortality; 2. Risk Factors; 3. Data Mining; 4. Feature Selection; 5. Explainable Artificial Intelligence.

LISTA DE SIGLAS

IA	Inteligência Artificial
ML	Machine Learning
RN	Recém-Nascido
SIM	Sistema de Informação de Mortalidade
SINASC	Sistema de Informação de Nascidos Vivos
OMS	Organização Mundial de Saúde
NA	Not Available Values (Valores Nulos)
Fimp	Feature Importance (Importância da Variável)
SEMUS	Secretaria Municipal de Saúde de Vitória
SHAP	SHapley Additive exPlanations

SUMÁRIO

1 INTRODUÇÃO

O avanço nas condições ambientais, sociais e nos serviços de saúde identificados nos últimos anos têm contribuído para uma diminuição significativa da mortalidade infantil em todo o mundo, sendo a melhoria dos cuidados com o componente pós-neonatal o principal contribuinte para este fato. Contudo, mortes neonatais não responderam similarmente às melhorias citadas e ainda se apresentam como um grande desafio para o Brasil e outros países em desenvolvimento.

Desde a promulgação da Constituição Federal de 1988, os municípios assumiram uma importante responsabilidade no enfrentamento da mortalidade neonatal, tornando-se protagonistas na implementação de políticas públicas de saúde (VICTORA, C., et al.,2011; MATIJASEVICH, A., *et al.,* 2016). Portanto, é essencial realizar a análise dos fatores de risco para o óbito neonatal no nível local, a fim de que os municípios brasileiros possam desenvolver, em parceria com os estados e o governo federal, políticas públicas adaptadas a cada realidade (2016, MATIJASEVICH, A., *et al*.)

O objetivo deste estudo é integrar e analisar os bancos de dados de informações materno-infantis. A cidade de Vitória, capital do estado do Espírito Santo, localizada no litoral capixaba, possui uma população estimada de 358.267 habitantes em 2018. Em 2015, a taxa de mortalidade infantil era de 9,17 por 1000 nascidos vivos (2018, INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE).

Atualmente, a taxa de mortalidade infantil é utilizada para avaliar a qualidade de vida de uma população e os cuidados de saúde de sua população infantil. Os óbitos relacionados à esta população estão associados a uma grande variedade de fatores de risco socioeconômicos, comportamentais e biológicos.

Com o objetivo de avaliar tais fatores de forma integrada, Mosley e Chen propuseram, em 1984, um modelo hierárquico baseado na hipótese de que fatores socioeconômicos determinam comportamentos, que, por sua vez, têm impacto sobre fatores biológicos. (2003, MOSLEY, W. H. & CHEN, L. C.). Esse modelo hierárquico traz um grande avanço para o desenvolvimento de políticas públicas, uma vez que as informações provenientes de estudos que não se limitam a apenas um subconjunto desses fatores de risco, resultam em recomendações úteis e adequadas para avaliar as mortes entre as crianças, pois apresentam uma visão ampla e contextualizada da realidade.

Esse modelo é importante para a concepção de políticas públicas realizadas pelos próprios municípios no contexto de descentralização administrativa e tributária em que o país está inserido (2000, HOSMER D. *et al.*), assim como pode fornecer dados para o resto do país e à políticas em regiões de outros países em desenvolvimento. Isso contribui para a diminuição de óbitos infantis neonatais.

A preocupação no estudo proposto se deve ao fato de que a associação entre fatores de risco biológicos e óbito nos primeiros 28 dias de vida já foi descrita por vários autores (2015, BORGES, T. S., & VAYEGO, S. A.; 2014, LANSKY, S. *et al.*; 2016, GAIVA, M.A. *et. Al*; 2011, ALMEIDA, M.F *et al.*), assim como o dado de que, aproximadamente um quarto (28%) de todas as crianças do mundo nascem com baixo peso ao nascer (2017, LAIS, S. *et al.*) e aproximadamente 60-80% de todos os óbitos neonatais estão associados a esse fator. Os fatores relacionados às mortes perinatais, portanto, tornaram-se um problema de saúde pública mundial.

A presente abordagem inovadora deste estudo fundamenta-se na integração sinérgica de bancos de dados previamente consolidados, promovendo a confluência de informações obtidas dos sistemas de informação sobre mortalidade (SIM) e nascidos vivos (SINASC) (2019, GARCIA *et al.*), os quais são extraídos de forma direta da plataforma oficial da Secretaria Municipal de Saúde de Vitória (SEMUS).

Além disso, a metodologia emprega técnicas de Inteligência Artificial (IA) (2017, HAMET, P., & TREMBLAY, J.) para realizar a predição de fatores de risco determinantes para a mortalidade infantil. Tal abordagem, pautada em recursos avançados de análise de dados e algoritmos inteligentes, visa aprimorar substancialmente o conhecimento e a compreensão acerca dos determinantes fundamentais para o desfecho fatal na população infantil.

Essa integração e visão de longo prazo com a predição permite o acesso a dados completos sobre nascimentos em todo o território municipal, incluindo informações maternas e infantis futuras relacionadas à natalidade. Além disso, são fornecidos insights sobre mortalidade, permitindo que as várias esferas de gestão na saúde pública tenham subsídios para avaliar os fatores de risco relacionados à morte neonatal e aprimorar sua capacidade de análise.

Embora haja uma redução significativa na taxa de mortalidade infantil no Brasil e no mundo nas últimas décadas, ainda há muito a ser feito para alcançar o objetivo global de zerar a taxa até 2030 (OMS-2020), conforme o plano da OMS.

Garcia *et al.* (2019), mostrou em seu estudo, alguns pontos positivos, incluindo o uso de bases nacionais já consolidadas e um modelo hierárquico, que permitiu análise integrada dos fatores socioeconômicos, comportamentais e de risco para a saúde, o índice de APGAR e os fatores de risco biológicos, que podem apoiar o desenvolvimento de políticas públicas mais abrangentes e contextualizadas.

A metanálise (2022, EL DIB, R.) é uma técnica estatística que tem sido amplamente utilizada na área da saúde para combinar os resultados de diferentes estudos sobre um mesmo tema e obter conclusões mais robustas. Esse tipo de análise permite identificar padrões, relações entre variáveis e avaliar a força das evidências disponíveis. No contexto da mortalidade infantil, a realização de uma metanálise pode ser muito útil para entender os diferentes fatores de risco associados a essa condição.

Ao agrupar dados de diferentes estudos, é possível ter uma visão mais ampla e detalhada sobre o tema em questão. Por exemplo, é possível identificar quais os fatores de risco mais consistentes para a mortalidade infantil, bem como quais os fatores que apresentam resultados controversos ou inconsistentes. Além disso, a metanálise pode ajudar a avaliar a qualidade dos estudos incluídos e identificar possíveis fontes de viés.

No caso da mortalidade infantil, a realização de uma metanálise pode ser particularmente útil para a elaboração de políticas públicas e programas de prevenção. A partir dos resultados obtidos, é possível identificar quais os fatores de risco mais importantes a serem abordados e quais as medidas mais eficazes para reduzir a taxa de mortalidade infantil em determinada população. (2019, VELOSO, F. *et al.*)

Algumas entidades, no entanto, como a Sociedade Australiana de Medicina considera Sistemas de Suporte a Decisão Clínica no topo das evidencias quando comparada a revisão sistemática ou Metanálise (2001, SIM, I. *et al*; 2023, GUIDES: ANSWERING CLINICAL QUESTIONS: HIERARCHY OF EVIDENCE).

Confrontado com a hipótese de que as condições socioeconômicas, ambientais e de nascimento influem os resultados da mortalidade neonatal e infantil, modelos de Machine Learning podem trazer uma indicação mais absoluta da seleção de características-chave, incluindo características pré-natais, perinatais e socioeconômicas e fatores de risco associados à predição da mortalidade infantil. (2019, BATISTA *et al.*).

Esse estudo visa utilizar modelos de aprendizado de máquina focados na interpretabilidade como uma abordagem alternativa mais barata e rápida à meta-análise, o objetivo deste trabalho é avaliar a viabilidade de utilizar esses modelos para identificar e compreender os fatores de risco associados à mortalidade infantil.

Além disso, busca-se comparar os resultados obtidos com essa abordagem com aqueles obtidos por meio da meta-análise, avaliando os benefícios em termos de custo-benefício e rapidez que o uso de modelos de aprendizado de máquina pode oferecer em relação à meta-análise. A partir dos resultados obtidos, espera-se fornecer uma alternativa viável e efetiva para a identificação e compreensão dos fatores de risco associados à mortalidade infantil, contribuindo para o desenvolvimento de políticas públicas mais abrangentes e contextualizadas.

2 MÉTODOS

2.1 DESENHO DO ESTUDO E CASUÍSTICA

O estudo adotou um desenho de coorte retrospectivo para a análise dos dados de óbitos infantis em ambos os sexos, com idades de 0 a 1 ano, entre os anos de 2006 a 2019 na cidade de Vitória. (2004, MASSAD, E. *et al*; 2008, CUMMINGS, S. R., *et al.*). Para isso, foram utilizados os registros do Sistema de Informações sobre Mortalidade (SIM) e do Sistema de Informações sobre Nascidos Vivos (SINASC) disponibilizados pela Secretaria Municipal de Saúde de Vitória. Foram excluídas notificações com dados incompletos para análise das variáveis de interesse.

Este desenho de estudo é particularmente útil quando as variáveis preditoras são caras ou difíceis de serem medidas no momento da coleta de dados, pois elas já foram registradas e armazenadas pela Secretaria Municipal de Saúde de Vitória. Isso permite uma execução de baixo custo e rápida, sem a necessidade de coleta de novos dados.

Antes do acesso aos dados, foi obtida a aprovação do estudo pelo Comitê de Ética e Pesquisa (CEP) sob o parecer número 3.280.796 (ANEXO) da Escola Técnica do SUS (ETSUS), garantindo que a pesquisa fosse conduzida de acordo com os princípios éticos e regulamentações nacionais de coleta de dados e informações. Além disso, foram tomadas todas as medidas para preservar a privacidade e confidencialidade das informações dos participantes, conforme previsto na legislação vigente.

2.2 INTEGRAÇÃO DO BANCO DE DADOS

Para integração dos Bancos de Dados de Nascimento (SINASC) e Mortalidade (SIM), utilizou-se método de associação determinística baseado na comparação dos registros de nascimento, principalmente o Número de Identificação do Recém-Nascido (NUMERODN) nas duas bases de dados para integrar os dados de notificação de mortalidade e nascidos vivos e criar um banco para a predição de Mortalidade Infantil.

Essas informações são verificadas pelo centro de Vigilância Epidemiológica e identificados minuciosamente para vincular dados de nascimento e óbito, visto que depende de dados pessoais mais precisos. Depois de consolidar os bancos de dados dos dois sistemas, analisamos, e pré-processamos os dados conforme descreveremos.

2.3 ANÁLISE, LIMPEZA E PREPARAÇÃO DOS DADOS

Durante a leitura do banco de dados, alguns valores nulos foram substituídos pela representação de valor nulo (NA) apropriada, com base no dicionário de dados fornecido pelos sistemas dos bancos de dados de origem, pois algum código foi usado no banco de dados para representar que um determinado valor não era -existente. Ainda com base no dicionário de dados, os tipos de dados da coluna foram corrigidos.

No banco de dados integrado, que inicialmente tinha 248 colunas e 173.353 registros, foram excluídas as colunas do Sistema de Informações sobre Mortalidade, pois esse banco de dados servia apenas para identificar quais dos nascidos haviam falecido em até um ano de vida, assim o banco de dados foi reduzido em 88 colunas, agora com 160 colunas. Em seguida, foram excluídas outras 52 colunas do Sistema de Informações sobre Nascidos Vivos, restando 108 atributos. Por fim, foram excluídas 63 colunas que não possuem importância médica ou epidemiológica, como: nome, número do registro, conforme dicionário de dados, restando 45 colunas no banco de dados.

As colunas que apresentavam grande quantidade de valores nulos (NA) também foram excluídas. Como critério, foram excluídas todas as 18 colunas que tiveram mais de 20% de valores nulos, ao se considerar a classe de óbito, restando 27 colunas. Outras 11 colunas tiveram de ser excluídas, pois apresentavam valores duplicados com outras colunas, ou irrelevantes, com dados não relacionados à saúde, como códigos, ou que na pré-análise e pré-processamento se mostraram irrelevantes.

Em seguida, foram excluídos todos os registros que apresentavam algum valor nulo, deixando a base de dados com 155.058 registros, dos quais 154.418 (99,59%) são da classe viva e 640 (0,41%) da classe morte.

As 14 colunas analisadas foram: presença de anomalia, APGAR 1, APGAR 5, número de consultas de pré-natal, estado civil da mãe, idade da mãe, tipo de parto, peso, número de filhos vivos e mortos, semanas de gestação, sexo, tipo de gravidez e resultado da morte. Detalhes são exemplificados na Tabela 1. Duas colunas auxiliares adicionais, data de nascimento e idade do óbito, também foram mantidas. Os registros foram divididos aleatoriamente, mas mantendo a proporção, em três novos bancos de dados: treinamento com 93.034 (60%) registros, teste com 31.012 (20%) e validação com 31.012 (20%) registros.

2.4 IDENTIFICAÇÃO DAS PRINCIPAIS VARIÁVEIS

Para identificar quais variáveis mais impactaram o modelo, e trazer maior confiabilidade e explicar melhor sua importância e potencial preditivo, podemos aplicar métodos como SHAP (SHapley Additive exPlanations) (2020, RODRÍGUEZ-PÉREZ, R., & BAJORATH, J.), *feature_importances* (do Scikit-Learn tree models) e selectKBest (também do Scikit-learn, baseado em métodos estatísticos como o X^2) (2011, PEDREGOSA *et al.*). Todas as ferramentas mencionadas acima são válidas para analisar as características mais importantes para a previsão. No entanto, no estudo, elas foram usadas para que pudéssemos combinar poder preditivo e explicabilidade dos modelos.

Nesse contexto, o método selectKBest foi o primeiro utilizado para selecionar as melhores características, com seu método estatístico, que pode ser avaliado antes do treinamento do modelo. Na mesma linha, ao treinar o primeiro modelo, foi utilizado o feature_importances dos algoritmos de árvore (*Random Forest* – 2013, KULKARNI, V.Y. *et al.*), em que o treinamento do modelo de referência (mais básico, antes do ajuste dos melhores parâmetros).

Este foi levado em consideração e comparado com as técnicas de selectKBest e posterior do SHAP, este último utilizado com o melhor modelo após treiná-lo e validá-lo em dados nunca vistos.

O SHAP (SHapley Additive exPlanations) é uma técnica que usa uma abordagem de teoria dos jogos para explicar a saída de modelos de aprendizado de máquina (6 - Documentação do SHAP).6 Por meio da teoria de jogos, ele conecta e agrupa as variáveis e calcula uma pontuação proporcional à contribuição desse fator para o alcance do objetivo, com base nos valores de Shapley e suas extensões relacionadas (2017, LUNDBERG *et al.*).

O principal modelo de ML utilizado como base para o SHAP foi o XGBoost (Extreme Gradient Boosting – 2016, CHEN, T. *et al.*) por ser o modelo com melhores resultados, sendo o mais indicado e usado hoje, quando se trata de dados estruturados ou tabulados. O XGBoost é um modelo supervisionado de algoritmo de aprendizado de máquina, baseado em árvore de decisão e usando uma estrutura de aumento de gradiente. Ele constrói várias árvores de decisão (que são basicamente como fluxogramas de decisão). Cada uma das árvores, individualmente, tem baixo poder preditivo, mas, com o poder do gradiente, ajustam os erros através dos resíduos anteriores e passam os resultados para as próximas árvores aumentando sua eficiência. Para que o resultado seja uma média dos resultados do conjunto de árvores.

A principal diferença entre o método de Feature Importance e o método de SHAP é que o primeiro fornece uma medida relativa da importância de cada recurso em relação aos outros, enquanto o último fornece uma medida absoluta da contribuição de cada recurso para a previsão do modelo em cada instância.

Para aumentar a capacidade preditiva do modelo, este passou por um ajuste de hiperparemetrização (“tuning”) para determinar quais parâmetros de configuração melhor se adequavam aos dados utilizados e ao objetivo do projeto. A seleção final das variáveis de interesse foi confrontada com as principais características encontradas na última revisão sistemática e metanálise sobre mortalidade infantil com dados dos Sistemas de Informações brasileiros (VELOSO *et al.,* 2019).

3 RESULTADOS

3.1 DESCRIÇÃO DOS DADOS

Inicialmente, o banco de dados contava com 173.353 registros de nascidos vivos, entre os anos de 2006 a 2019, mas após limpeza e preparo, passou a contar com 155.058 registros, dos quais 640 (0,41%) foram a óbito em até um ano de vida. Como podemos observar, os óbitos mais prevalentes foram, ao nascer, os de baixo peso (menos de 2500g), 413 (64,53%), que tiveram APGAR 5 menor que 7. 89 (13,91%), que nasceram em pré-termo (menos de 37 semanas de gestação), 391 (61,09%), e aquelas que fizeram menos de 6 consultas de pré-natal, 327 (50,47%). Dos que morreram, 404 (63,12%) estavam no período neonatal (até 28 dias).

O resumo da descrição dos dados do registro de nascimento pode ser visualizado na TABELA 1. Na TABELA 2, é possível observar a descrição dos dados antes do processo de limpeza e preparo.

3.2 SELEÇÃO DAS MELHORES VARIÁVEIS

A interpretabilidade do modelo foi realizada utilizando três metodologias e comparando a identificação dos fatores de risco. Os métodos SelectKBest, Feature Importance e SHAP (SHapley Additive exPlanations) identificaram vários fatores de risco semelhantes para morte precoce, conforme demonstrado no Quadro 1, o que reforça a importância dessas características para predizer o óbito infantil.

As variáveis epidemiológicas com maior força na predição do óbito neonatal e infantil até 1 (um) ano encontradas foram: peso ao nascer, APGAR 5, número de semanas gestacionais, presença de anomalias congênitas e número de consultas de pré-natal, o que é corroborado pelos resultados obtidos por outros estudos (GAIVA *et al*., 2016; VELOSO *et al.*, 2019).4,9 As principais variáveis encontradas em nossos resultados também são corroboradas pelos achados de outros pesquisadores, que as acharam com maior relevância o seguintes variáveis: baixo peso ao nascer e prematuridade (BORGESA *et al.*, 2015; GAIVA *et al.*, 2016).

Alguns fatores significativamente associados à mortalidade também foram encontrados: idade da mãe acima de 35 anos, sexo masculino, gravidez múltipla, ausência de companheiro, complicações da gravidez e do parto de Cesário (VELOSO *et al.,* 2019).

Observou-se que a maior concentração de óbitos ocorreu nos primeiros 28 dias de vida, com um total de 484 casos, correspondendo a 65% dos óbitos registrados. As mortes neonatais nos primeiros seis dias são causadas principalmente por fatores maternos, complicações da gravidez e do parto (KASSAR *et al.*, 2013). Pesquisadores realizaram um estudo dos primeiros 28 dias de vida da população de São Paulo e constataram as seguintes variáveis como as mais importantes para a predição: APGAR5, peso ao nascer, APGAR1, presença de anomalia congênita e idade gestacional - que também são resultados que corroboram os achados do modelo SHAP empregado nesse estudo. (BATISTA *et al.,* 2021).

Tabela 1 – Descrição das Variáveis Numéricos após pré-processamento.

	Total			**RN vivo**			**RN morte**		
	155.058 (100%)			**154.418 (99.59%)**			**640 (0.41%)**		
Variáveis	**Média (DP)**	**IC 95,0%**	**Mediana (P25 – P75)**	**Média (DP)**	**IC 95,0%**	**Mediana (P25 – P75)**	**Média (DP)**	**IC 95,0%**	**Mediana (P25 – P75)**
APGAR 1	8.13 (0.99)	8.13 – 8.14	8.0 (8.0–9.0)	8.14 (0.97)	8.14–8.15	8.0 (8.0–9.0)	6.07 (2.29)	5.89–6.25	7.0 (5.0–8.0)
APGAR 5	9.05 (0.7)	9.04 – 9.05	9.0 (9.0–9.0)	9.05 (0.69)	9.051–9.058	9.0 (9.0–9.0)	7.80 (1.68)	7.67–7.93	8.0 (7.0–9.0)
Idade Materna	27.54 (6.50)	27.51–27.57	28.0 (22.0–32.0)	27.54 (6.49)	(27.51–27.57)	28.0 (22.0–32.0)	27.49 (7.02)	26.96–28.04	27.0 (22.0–33.0)
Peso ao nascer	3216 (547.78)	3213 –3218	3250 (2940–3555)	3221 (538.29)	3218–3223	3250 (2940–3555)	1955 (1094)	1870–2040)	1916 (885–2942)
Número de filhos mortos	0.18 (0.51)	0.18–0.18	0 (0–0)	0.18 (0.51)	(0.1883–0.1888)	0 (0–0)	0.25 (0.63)	0.20-0.29	0 (0-0)
Número de filhos vivos	0.79 (1.06)	0.78–0.79	0 (0–1)	0.79 (1.06)	0.786–0.796	1.0 (0–1)	0.94 (1.28)	0.84-1.04	0 (0-1)

Fonte: Elaborada pelo autor.

DP: desvio padrão; IC: intervalo de confiança; RN: recém-nascido.

Tabela 2 – Descrição das Variáveis Categóricos após pré-processamento.

Variáveis		Total	RN vivo	RN morte
		155.058 (100%)	**154.418 (99.59%)**	**640 (0.41%)**
Anomalia Congênita	Presença	871 (0.56%)	778 (0.5%)	93 (14.53%)
	Ausência	154187 (99.44%)	153640 (99.5%)	547 (85.47%)
Estado civil da mãe	Solteira	73612 (47.47%)	73251 (47.44%)	361 (56.41%)
	Casada	74167 (47.83%)	73918 (47.87%)	249 (38.91%)
	Viúva	374 (0.25%)	372 (0.24%)	2 (0.31%)
	Divorciada	3162 (2.04%)	3146 (2.04%)	16 (2.5%)
	União Estável	3743 (2.41%)	3731 (2.42%)	12 (1.88%)
Idade Gestacional	Termo (37-42s)	140579 (90.6%)	140332 (90.88%)	247 (38.59%)
	Pré-termo (<37s)	12810 (8.26%)	12419 (8.04%)	391 (61.09%)
	Pós-termo (>42s)	1669 (1.08%)	1667 (1.08%)	2 (0.31%)
Tipo de parto	Vaginal	46296 (29.92%)	46159 (29.89%)	237 (37.03%)
	Cesáreo	108662 (70.08%)	108259 (70.11%)	403 (62.97%)
Consultas Pré-natal	Nenhuma	1160 (0.75%)	1135 (0.74%)	25 (3.91%)
	1 a 3	4450 (2.93%)	4460 (2.89%)	90 (14.06%)
	4 a 6	119926 (77.34%)	119609 (77.46%)	208 (32.5%)
	7 ou mais	73612 (47.47%)	73251 (47.44%)	317 (49.53%)
Sexo do RN	Masculino	79497 (51.27%)	79148 (51.26%)	349 (54.53%)
	Feminino	75561 (48.73%)	75270 (48.74%)	291 (45.47%)
Tipo de gravidez	Única	151461 (97.68%)	150895 (97.72%)	566 (88.44%)
	Gêmeos	3499 (2.26%)	3428 (2.21%)	71 (11.09%)
	Trigêmeos ou mais	98 (0.063%)	95 (0.061%)	3 (0.47%)
Ocorrência de morte	Período neonatal			404 (63.12%)
	Período pós-neonatal			236 (36.88%)

Fonte: Elaborada pelo autor.

DP: desvio padrão; IC: intervalo de confiança; RN: recém-nascido.

Tabela 3 – Ordem dos Fatores de Risco de Diferentes Técnicas em comparação com a Literatura.

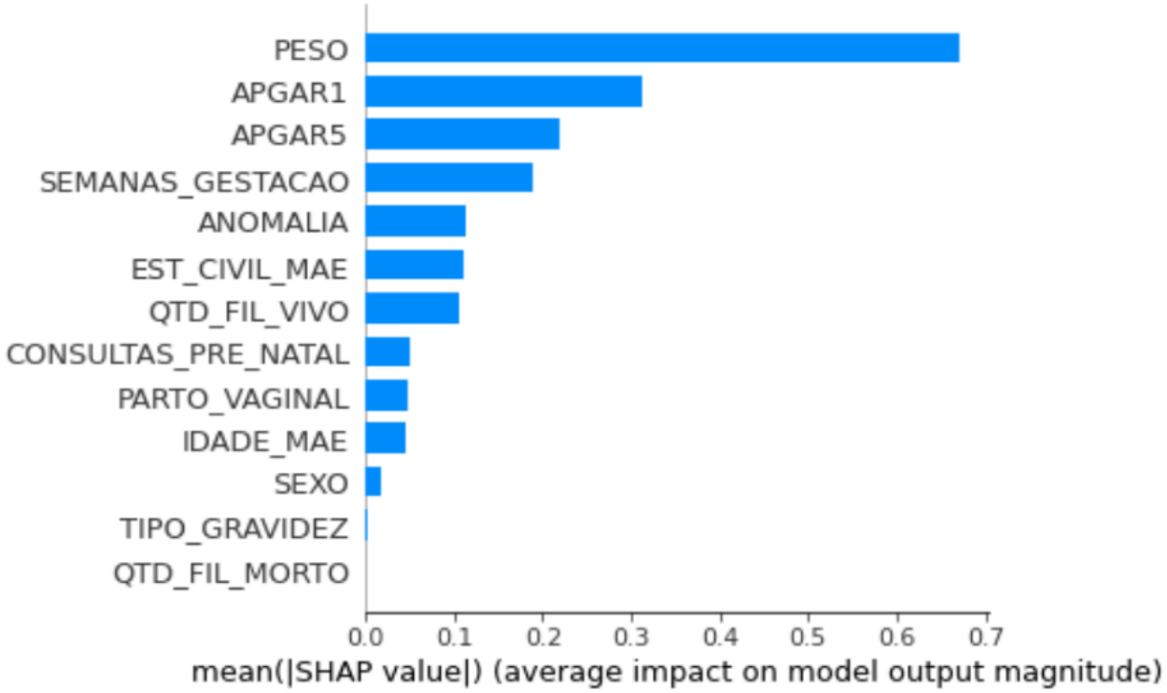
LITERATURE REVIEW		PRESENTE ESTUDO		
Metanálise [Garcia et. al]	**SHAP [Batista et. al]**	**Select K-Best**	**Feature Importance**	**SHAP**
1. Peso <1500g	1. APGAR5	1. Anomalias	1. Peso	1. Peso
2. Presença de anomalia	2. Peso	2. APGAR1	2. Idade materna	2. APGAR1
3. APGAR 5 (<7)	3. APGAR1	3. APGAR5	3. APGAR1	3. APGAR5
4. IG (<37s)	4. Anomalias	4. Consultas pré-natal	4. APGAR5	4. Idade Gestacional
5. Intercorrências na gestação	5. IG	5. Estado civil da mãe	5. Número de filhos vivos	5. Presença de anomalias
6. Pré-natal (Ausência)		6. Tipo de gravidez	6. Idade gestacional	6. Estado civil da mãe
7. Múltiplos RNs		7. Peso	7. Número de consultas pré-natal	Número de filhos vivos
8. Mãe sem parceiro		8. Número de filhos mortos	8. Estado civil da mãe	8. Número de consultas pré-natal
9. Histórico de RN morto		9. Número de filhos vivos	9. Sexo do RN	9. Tipo de gravidez
10. Educação		10. Idade gestacional	10. Tipo de gravidez	10. Idade materna

Fonte: Elaborada pelo autor.

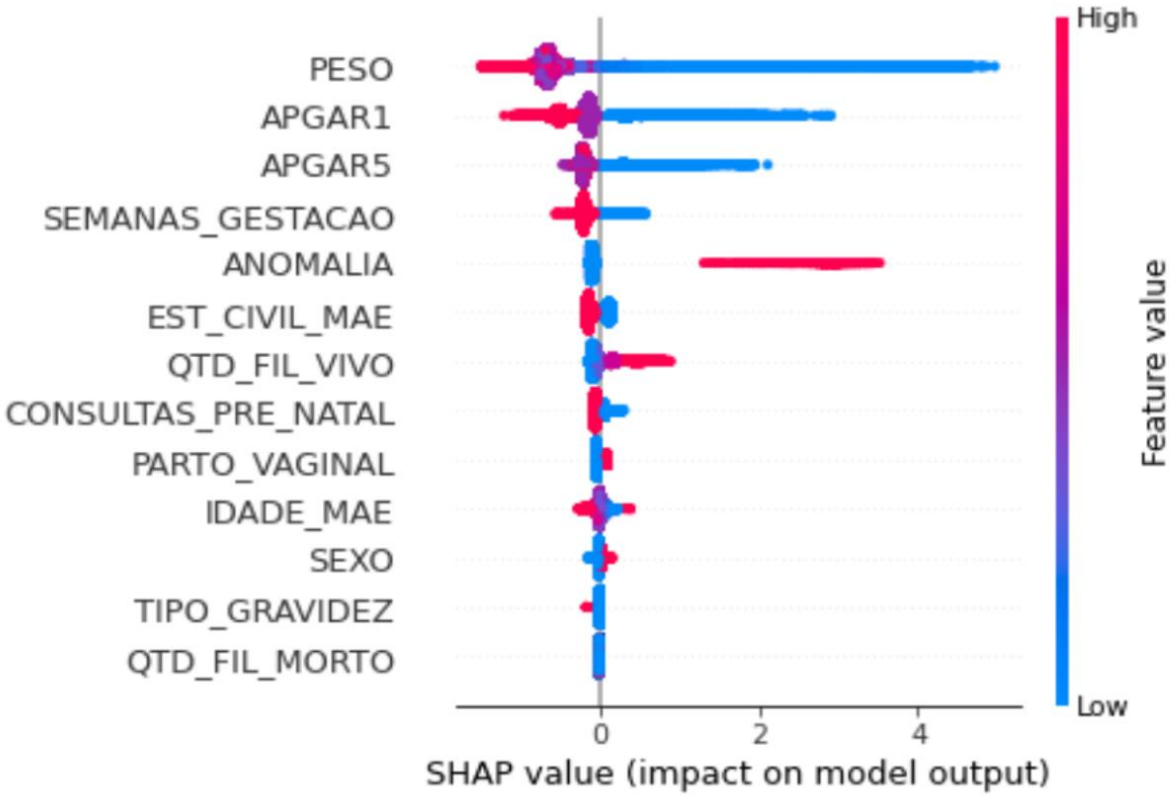
SHAP: Shappley Values; RN: recém-nascido; IG: Idade Gestacional; Tipo de gravidez (única ou mais de um filho); s: semanas (após o número de semanas de gestação).

Figura 1 - Resultados Gerados pelo Algoritmo na Seleção das Principais Variáveis.

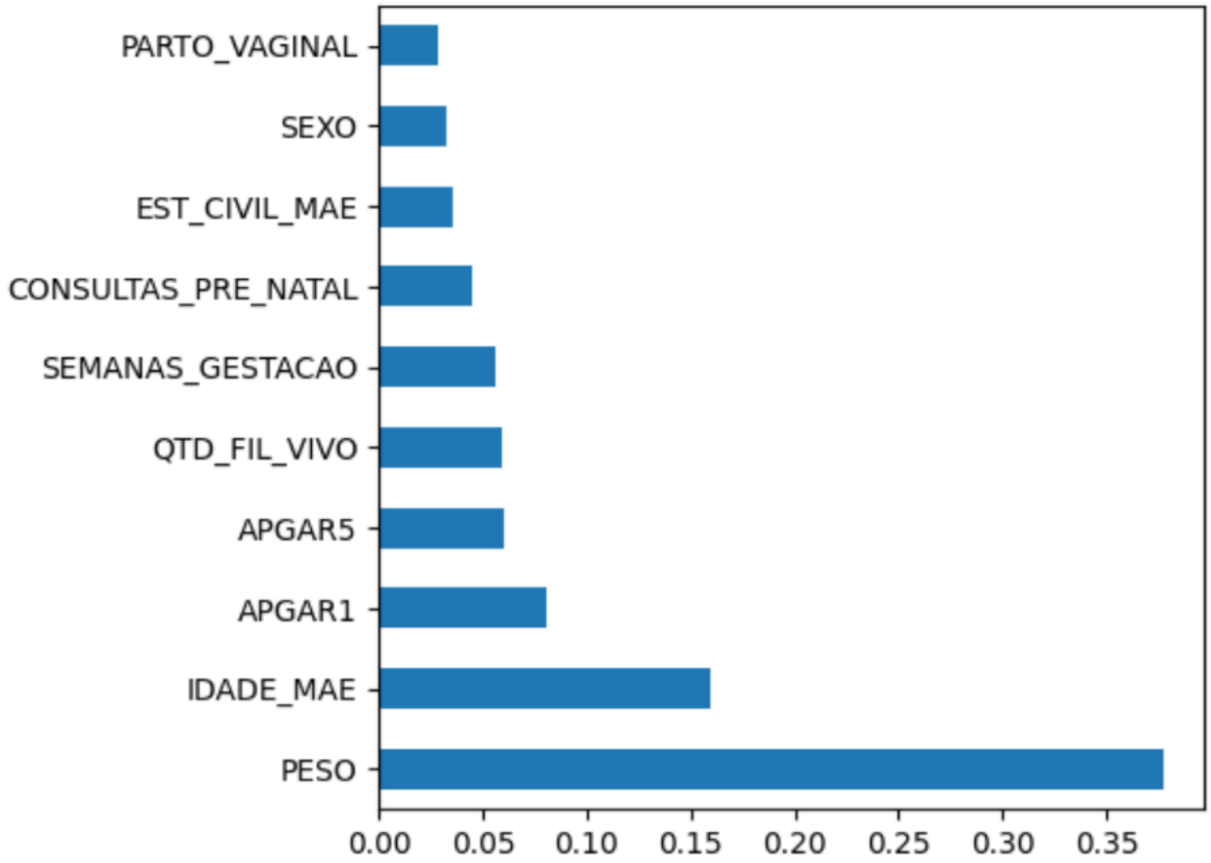
A) SHAP: Impacto Médio de cada variável no XGBClassifier.



B) SHAP: Dispersão de Densidade dos valores SHAP para cada recurso no XGBClassifier.



C) Resultado de Importância de Recursos pelo Modelo RandomForestClassifier.



Fonte: Elaborada pelo autor.

Legenda: PESO (Peso ao Nascer), APGAR1 (APGAR no Primeiro Minuto de Vida), APGAR5 (APGAR no Quinto Minuto de Vida), SEMANAS GESTAÇÃO (Idade Gestacional), ANOMALIA (Presença de Anomalia Congênita), EST CIVIL MÃE (Estado Civil da Mãe), QTD FIL VIVO (Número de Filhos Vivos), CONSULTAS PRÉ-NATAL (Número de Consultas Pré-Natal), PARTO VAGINAL (se o parto foi vaginal ou cesáreo), IDADE MÃE (Idade da Mãe), SEXO (Sexo), TIPO GRAVIDEZ (única ou mais de um filho), QTD FIL MORTO (Número de Filhos Mortos).

41 DISCUSSÃO

Na era da tecnologia do século XXI, com a ampla adoção do Big Data e da IA, o setor de saúde reconheceu a importância de coletar e organizar os dados de forma digitalizada. Isso traz inúmeros benefícios, como o aumento da eficiência das tarefas, a facilitação da pesquisa e tomada de decisões, o acompanhamento e monitoramento dos pacientes, além do uso e desenvolvimento da IA (PANCH *et al*., 2019; HAMET *et al.*, 2017). No entanto, a falta de infraestrutura para coletar as variáveis necessárias para treinar os algoritmos tem sido um obstáculo para a implementação efetiva do aprendizado de máquina na prática clínica.

Estudos de meta-análise são o topo da pirâmide de confiabilidade nos resultados, mas para permitir uma boa metanálise há necessidade de resultados amplos e geralmente a seleção de um grande número de estudos com qualidade técnica e padronização no método, o que pode custar muito tempo e recursos para realizar os estudos originais e selecionar os artigos durante o processo de revisão (HERNANDEZ, *et al.,* 2020).

A aplicação de técnicas de Mineração de Dados como seleção de características, previsão de eventos determinísticos e para explicar fenómenos que envolvem a saúde infantil pode ser uma maneira mais eficiente e rápida. (2010, VIANNA, R. *et al.*) A aplicação de algoritmos para selecionar as melhores características relacionadas à mortalidade infantil já é utilizada com o modelo *Random Forest* juntamente com seu bom poder preditor (2022, SILVA ROCHA, E.D. et al), caso optem pelo método, isso SHAP pode auxiliar os profissionais que lidam diretamente com o desfecho a terem resultados muito comparáveis a grandes metanálise e auxiliar nas orientações epidemiológicas que podem ser disponibilizadas a partir da análise dos dados.

Com base nas características identificadas pelo método SHAP, como apresentado na Tabela 1 e nas Figuras 1-A e 1-B, o estudo de Veloso e outros (2019), por meio de uma revisão sistemática e meta-análise, também identificou fatores relacionados à mortalidade infantil. Esses fatores incluem a ausência de consultas pré-natais, baixo peso ao nascer, presença de anomalias congênitas, pontuação APGAR no quinto minuto e número de semanas de gestação.

Além disso, a escolaridade da mãe, estado civil da mãe, idade da mãe, sexo do recém-nascido e tipo de gravidez (múltipla ou única) foram considerados fatores de risco médio para a mortalidade infantil (VELOSO *et al.,* 2019).

Batista e outros (2021) utilizaram a população de São Paula para utilizar aprendizado de máquina e o método SHAP para prever e identificar os principais fatores relacionados à mortalidade neonatal, que identificou que as cinco variáveis mais importantes foram Apgar no 5º minuto, peso ao nascer, Apgar no 1º minuto, presença de anomalia congênita e idade gestacional, respectivamente (BATISTA *et al.,* 2021).

Apesar do estudo de Batista et. al ter focado na melhoria da métrica de predição do modelo, uma importante evidência foi levantada, que aponta para o fato de que o uso de um número menor de variáveis (cinco indicadores perinatais mínimos da OMS) não diminuiu significativamente o desempenho do algoritmo XGBoost. Este fato direciona para a interação forte entre os fatores menos importante e mais importantes para a predição, ou seja, mesmo com menos importância relativa, os outros fatores ainda precisam ser analisados para uma boa predição do modelo de aprendizado de máquina. A comparação dos resultados de interpretabilidade encontrados pode ser vista na Tabela 1.

A metodologia apresentada foi aplicada a dados de natalidade e mortalidade infantil, mas pode ser utilizada em diferentes áreas e problemas de saúde. Por exemplo, para encontrar fatores de risco relacionados ao COVID-19 (Sars-CoV-2). Um algoritmo semelhante foi desenvolvido para melhorar o entendimento das características clínicas importantes para a triagem de COVID-19 em diversos ambientes, incluindo hospitais, clínicas e locais de trabalho. Os autores também acreditam que o modelo poderia ser usado para desenvolver novos testes diagnósticos para COVID-19 no caso de falta de testes. (2021, FERNANDES, F. T., et al; PINASCO et al., 2022). No caso da pandemia, por se tratar de uma doença recente e com prognóstico imprevisível, foi necessário encontrar padrões entre os pacientes, nos quais estudos tradicionais seriam inviabilizados pela demora em sua realização.

Outra aplicação dessas técnicas é a individualização do risco de pacientes cardíacos, que já superou em precisão e interpretabilidade do escore criado pelo grande estudo de Framingham (2019, ALAA *et al.*).

WICHMANN e outros (2022) em seu estudo visou entrevistar médicos para questionar como eles gostariam de receber a informação de uma Inteligência Artificial, a pesquisa revelou que os médicos têm uma abertura geral para receber resultados preditivos de ML. No entanto, eles possuem preferências específicas sobre a forma como esses resultados devem ser apresentados. No artigo, os resultados apontam para o fato de que os médicos preferem que os resultados sejam claros, concisos e acompanhados de explicações sobre o raciocínio do modelo.

Tendo isso em mente, a utilização de técnicas de interpretabilidade, como o SHAP, para modelos de Machine Learning pode ser uma forma eficiente de selecionar e identificar características como Peso, APGAR, Idade Gestacional e Presença de Anomalias, que estão diretamente relacionadas à mortalidade infantil.

Essa abordagem proporciona resultados semelhantes aos obtidos em metanálise, porém de maneira mais rápida e sem a necessidade de realizar múltiplos estudos e incorporá-los em uma revisão. Além disso, a aplicação desses algoritmos neste estudo ressalta a sua simplicidade de execução, especialmente com a popularização da programação e da utilização de inteligências artificiais sem código (*no-code*), tornando-os uma alternativa mais viável para a maioria dos pesquisadores.

É importante destacar que as variáveis identificadas como fatores de risco pelas técnicas de predição utilizadas neste estudo não estabelecem uma relação causal (2022, COLODETTE, A. L., *et al*). Como o estudo é observacional, a correlação entre variáveis não implica em causalidade devido à presença de variáveis ocultas e fatores aleatórios que também podem influenciar o resultado (2020, PROSPERI, M., et al.). Portanto, o desfecho não pode ser atribuído exclusivamente às variáveis identificadas no estudo.

Além disso, uma limitação adicional está relacionada à integração de dados. Embora o linkage determinístico com variáveis pessoais seja mais robusto, em casos em que há falta de informações precisas, é possível realizar o relacionamento probabilístico entre as bases de dados. (2008, COUTINHO, R. G. *et al*.).

Esse método tem sido amplamente utilizado em estudos de coorte para monitorar desfechos. Ele permite a integração de bancos de dados de naturezas diferentes, mesmo na ausência de um identificador único. Isso é alcançado através do uso de campos comuns nas bases relacionadas para estimar a probabilidade de que um par de registros se refira ao mesmo indivíduo.

Como próximo passo, o algoritmo pode evoluir para uma calculadora de risco na prática clínica neonatal (2022, TEJI, J., et al), na qual possa trazer interpretabilidade prática em tempo real para apoiar a prática médica baseada no topo nas evidências clínicas.

5 CONCLUSÃO

Pode-se concluir, portanto, que o uso de técnicas de Inteligência Artificial, mais especificamente o Machine Learning, aplicado a modelos de explicabilidade e interpretabilidade, como o SHAP, apresentam um grande potencial na seleção e identificação de fatores de risco populacionais associados à mortalidade infantil, aproveitando-se de bancos de dados existentes e eliminando a necessidade de conduzir novos estudos populacionais. Nesse sentido, essa abordagem oferece uma maneira eficiente de obter evidências e levanta informações relevantes para auxiliar gestores e profissionais de saúde na tomada de decisões em saúde pública.

REFERÊNCIAS

LAWN, J. E., *et al.* (2014). Every newborn: progress, priorities, and potential beyond survival. **Lancet,** 384, 189-205. http://dx.doi.org/10.1016/S0140-6736(14)60496-7.

MATIJASEVICH, A., *et al.* (2016). Método para estimação de indicadores de mortalidade infantil e baixo peso ao nascer para municípios do Brasil, 2012. **Epidemiol Serv Saúde**, 25, 637-646. https://doi.org/10.5123/S1679-49742016000300020.

VICTORA, C., *et al.* (2011). Maternal and child health in Brazil: progress and challenges. **Lancet,** 377, 1863-1876. http://dx.doi.org/ 10.1016/S0140-6736(11)60138-4.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. (2018). **Cidades IBGE Gov Br**. Acesso em 31/05/2023. Disponível em: https://cidades.ibge.gov.br/brasil/es/vitoria/panorama.

MOSLEY, W. H., & CHEN, L. C. (2003). An analytical framework for the study of child survival in developing countries. **Bulletin of the World Health Organization,** 81, 140-145.

HOSMER Jr., D.W and LEMESHOW, S. (2000) **Applied logistic regression**. 2nd Edition, John Wiley & Sons, Inc., New York. http://dx.doi.org/10.1002/0471722146.

BORGES, T. S., & VAYEGO, S. A. (2015). Risk factors for neonatal mortality in a county in the Southern region. **Ciência e Saúde (Paraná**), 8(1), 7-14. doi: https://doi.org/10.15448/1983-652X.2015.1.21010.

LANSKY, S., *et al.* (2014). Pesquisa nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. **Cadernos de Saúde Pública,** 30, S192-S207. https://doi.org/10.1590/0102-311X00133213.

GAIVA, M. A., *et al.* (2016). Maternal and child risk factors associated with neonatal mortality. **Texto & Contexto Enfermagem**, 25. https://doi.org/10.1590/0104-07072016002290015.

ALMEIDA, M. F. *et al.* (2011). Sobrevida e fatores de risco para mortalidade neonatal em uma coorte de nascidos vivos de muito baixo peso ao nascer, na Região Sul do Município de São Paulo, Brasil**. Cadernos de Saúde Pública**, 27, 1088-1098. https://doi.org/10.1590/S0102-311X2011000600006.

LAI, S., *et al.* (2017). Perinatal risk factors for low and moderate five-minute Apgar scores at term. **European Journal of Obstetrics & Gynecology and Reproductive Biology,** 210, 251-256. http://dx.doi.org/10.1016/j.ejogrb.2017.01.008.

GARCIA, L. P. *et al.* (2018). Risk factors for neonatal death in the capital city with the lowest infant mortality rate in Brazil. **J Pediatr (Rio J)**. 95(2):194-200. http://dx.doi.org/10.1016/j.jped.2017.12.007.

VELOSO, F. *et al.* (2019). Analysis of neonatal mortality risk factors in Brazil: a systematic review and meta-analysis of observational studies. **Jornal de Pediatria**, 95(5), 519-530. http://dx.doi.org/10.1016/j.jped.2018.12.014.

KASSAR, S. B., et al. (2013). Determinants of neonatal death with emphasis on health care during pregnancy, childbirth, and reproductive history. **Jornal de Pediatria (Rio de Janeiro),** 89(3), 269-277. http://dx.doi.org/10.1016/j.jped.2012.11.005

SIM, I., *et al.* (2001). Clinical decision support systems for the practice of evidence-based medicine**. Journal of the American Medical Informatics Association**, 8(6), 527-534. http://dx.doi.org/10.1136/jamia.2001.0080527.

GUIDES: ANSWERING CLINICAL QUESTIONS: HIERARCHY OF EVIDENCE. (2023). Acesso em 31/05/2023. Disponível em: https://guides.library.uwa.edu.au/c.php?g=800622.

BATISTA, A. F. M., et al. (2021). Neonatal mortality prediction with routinely collected data: a machine learning approach. BMC Pediatrics, 21(1), 322. doi: 10.1186/s12887-021-02788-9.

HAMET, P., & TREMBLAY, J. (2017). **Artificial intelligence in medicine.** Metabolism, 69S, S36-S40. http://dx.doi.org/10.1016/j.metabol.2017.01.011.

WORLD HEALTH ORGANIZATION. (2020). World health statistics 2020: Monitoring health for the SDGs, sustainable development goals. Geneva: **World Health Organization.** License: CC BY-NC-SA 3.0 IGO.

EL DIB, R. (2022). Como interpretar uma metanálise?. **Jornal Vascular Brasileiro**, 21. doi: 10.1590/1677-5449.202200431.

MASSAD, Eduardo *et al.* (2004). **Métodos quantitativos em medicina**. São Paulo: Manole. Acesso em: 31 maio 2023.

CUMMINGS, S. R., et al. (2008). Delineando estudos de coorte. In: Stephen B. Hulley et al. **Delineamento da pesquisa clínica**: uma abordagem epidemiológica (3rd ed., pp. 115-120). Porto Alegre: Artmed.

RODRÍGUEZ-PÉREZ, R., & BAJORATH, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. **Journal Of Computer-Aided Molecular Design**, 34(10), 1013-1026. doi: 10.1007/s10822-020-00314-0.

PEDREGOSA *et al.* (2011). Scikit-learn: Machine Learning in Python. **The Journal of Machine Learning Research.** 12, pp. 2825-2830. https://doi.org/ 10.5555/1953048.2078195.

KULKARNI, V.Y. *et al.* (2013) Random Forest Classifiers: A Survey and Future Research Directions. **International Journal of Advanced Computing**, ISSN:2051-0845, Vol.36. Acesso em 31/05/2023. Disponível em: https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Random-Forest-Classifiers_A-Survey-and-Future.pdf.

LUNDBERG, S. M., & LEE, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st **International Conference on Neural Information Processing Systems** (NIPS'17) (pp. 4768-4777). NIPS 2017. Curran Associates Inc., Red Hook, NY, USA.

CHEN, T. *et al.* (2016). XGBoost: A Scalable Tree Boosting System. **KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. 785–794. https://doi.org/10.1145/2939672.2939785.

PANCH, T., et al. (2019). The "inconvenient truth" about AI in healthcare. **NPJ Digital Medicine,** 2, 77. doi: 10.1038/s41746-019-0155-4.

HAMET, P. et al. (2017) Tremblay, J. Artificial intelligence in medicine. **Metabolism**. 69S:S36-S40. doi:10.1016/j.metabol.2017.01.011.

HERNANDEZ, A. V. et al. (2020), Marti, K. M., Roman, Y. M. Meta-Analysis. **Chest.**158(1S):S97-S102. doi:10.1016/j.chest.2020.03.003.

VIANNA, R. et al. (2010). Mineração de dados e características da mortalidade infantil. **Cadernos De Saúde Pública**, 26(3), 535-542. doi: 10.1590/s0102-311x2010000300011.

SILVA ROCHA, E.d., et al (2022). On usage of artificial intelligence for predicting mortality during and post-pregnancy: a systematic review of literature. **BMC Med Inform Decis Mak**. 22, 334. https://doi.org/10.1186/s12911-022-02082-3

FERNANDES, F. T., et al. (2021). A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil. **Sci Rep** 11, 3343. https://doi.org/10.1038/s41598-021-82885-y.

CARREIRO PINASCO, G. et al. (2022). An interpretable machine learning model for covid-19 screening. **Journal Of Human Growth And Development**, 32(2), 268-274. doi: 10.36311/jhgd.v32.13324.

ALAA, A. M., et al. (2019) Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants**. PLoS One.1**4(5):e0213653. doi: 10.1371/journal.pone.0213653.

COUTINHO, R. G. et al. (2008). Sensibilidade do linkage probabilístico na identificação de nascimentos informados: Estudo Pró Saúde. **Rev Saúde Pública** 2008;42(6):1097-100.

COLODETTE, A. L., et al (2022). Feature Selection for Identification of Risk Factors Associated with Infant Mortality. **In Computational Advances in Bio and Medical Sciences** (pp. 8-15). Lecture Notes in Computer Science, vol. 13254. Springer, Cham. http://dx.doi.org/10.1007/978-3-031-17531-2_8.

PROSPERI, M., et al. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. **Nature Machine Intelligence**, 2(7), 369-375. doi: 10.1038/s42256-020-0197-y

WICHMANN, R., et al. (2022). Physician preference for receiving machine learning predictive results: A cross-sectional multicentric study. **PLOS ONE,** 17(12), e0278397. doi: 10.1371/journal.pone.0278397.

TEJI, J., et al (2022). NeoAI 1.0: Machine learning-based paradigm for prediction of neonatal and infant risk of death. **Computers In Biology And Medicine**, 147, 105639. doi: 10.1016/j.compbiomed.2022.105639.