MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE CÁRIE NA PRIMEIRA INFÂNCIA

ANA BEATRIZ LIMA DE QUEIROZ[1]; LUIZ ALEXANDRE CHISINI[2]; FLAVIO FERNANDO DEMARCO[3]; MARCOS BRITTO CORRÊA[4]

[1]Universidade Federal de Pelotas – queiroz.abl@gmail.com
[2]Universidade Federal de Pelotas – alexandrechisini@gmail.com
[3]Universidade Federal de Pelotas – ffdemarco@gmail.com
[4]Universidade Federal de Pelotas – marcosbrittocorrea@hotmail.com

1. INTRODUÇÃO

A cárie dentária é a doença não-transmissível mais prevalente no mundo, e é considerada um problema global de saúde pública. De acordo com o reporte de saúde bucal da Organização Mundial da Saúde (OMS), também é a doença crônica mais prevalente na infância, afetando 514 milhões de crianças (OMS, 2022). Ademais, a Associação Americana de Odontopediatria (AAPD) classifica a presença de um ou mais dentes cariados em crianças abaixo de seis anos de idade como cárie na primeira infância (AAPD, 2020). A cárie severa na primeira infância é o termo utilizado pela AAPD para designar a presença de maior número de lesões de cárie de estágio mais avançado em crianças com idade de até 5 anos (AAPD, 2020).

Na infância, e especialmente nos primeiros anos de vida, há uma relação de dependência integral do bebê aos cuidados. A criança, na maioria das vezes, tem na figura materna a sua provedora primária de cuidados; portanto, os comportamentos e ações da mãe têm impacto direto na saúde da criança (MARTÍNEZ-RICO, 2024). Esse dado traz consigo inúmeros questionamentos, incluindo a possibilidade de identificar risco de cárie na criança a partir de dados de maternos, visando ao monitoramento, à prevenção e à intervenção precoce para proporcionar saúde bucal.

O objetivo deste estudo foi investigar se um conjunto de dados de saúde bucal materna, fatores comportamentais e socioeconômicos registrados durante os primeiros mil dias de vida é um bom preditor de cárie na primeira infância da criança.

2. METODOLOGIA

Este estudo foi realizado com dados da Coorte de nascimentos de 2015 de Pelotas (HALLAL, 2018). O estudo foi reportado de acordo com o *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis* (TRIPOD) AI (COLLINS, 2024). Os desfechos adotados foram a cárie na primeira infância (ECC) e cárie severa na primeira infância (S-ECC), coletadas por meio de exame de saúde bucal na criança aos 48 meses de idade e classificadas em um desfecho dicotômico (presença ou ausência da condição) de acordo com as definições da AAPD (2020). Os preditores utilizados foram dados de saúde bucal, fatores socioeconômicos e autorrelatados relacionados à saúde bucal da mãe. Os dados de saúde bucal das mães foram coletados por meio de exame clínico e classificado de acordo com quantidade de superfícies dentárias acometidas por

cárie, ausentes ou obturadas (CPO-S de 0 a 192). As covariáveis adotadas foram fatores socioeconômicos (idade materna, escolaridade materna e renda familiar) e relacionados à saúde bucal (ansiedade odontológica materna, consulta odontológica durante a gestação, uso de fio dental e orientações de saúde bucal até os doze meses de vida da criança), que foram coletadas por meio de questionário de múltipla escolha.

Variáveis categóricas com mais de duas categorias foram pré-processadas usando *one-hot encoding*, enquanto variáveis contínuas foram padronizadas com *Z-Scores*. Valores ausentes foram corrigidos por Imputação Multivariada por Equações Encadeadas (MICE), utilizando um algoritmo de *random forest* para imputação no conjunto de treinamento. Este procedimento foi conduzido no RStudio versão 4.3.0 (RStudio Team, MA, EUA). O conjunto de dados foi dividido aleatoriamente em dois subconjuntos independentes, com 70% alocados para treinamento e 30% para teste.

Cinco algoritmos de aprendizado de máquina para dados estruturados foram empregados (Gradient Boosting Classifier, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost Classifier (CatBoost) e Ada Boost Classifier). O ajuste de hiperparâmetros foi otimizado por meio de validação cruzada de 10 *k-folds* com 50 iterações. Os algoritmos foram rodados no Jupyter notebook em linguagem de programação Python para cada um dos desfechos. A contribuição das variáveis foi avaliada com Shapley Additive Explanations (SHAP) (LUNDBERG & LEE, 2017). O desempenho do modelo foi avaliado com base na área sob a curva ROC (Receiver Operating Characteristic), com intervalos de confiança de 95%. Métricas adicionais, incluindo acurácia, *recall* (sensibilidade), precisão e escore F1, foram computadas. As diferenças entre as curvas ROC foram testadas usando o método de DeLong. Para examinar a imparcialidade dos algoritmos, foram conduzidas análises estratificadas por sexo e raça. A significância estatística de 0,05 foi estabelecida para todas as análises estatísticas.

3. RESULTADOS E DISCUSSÃO

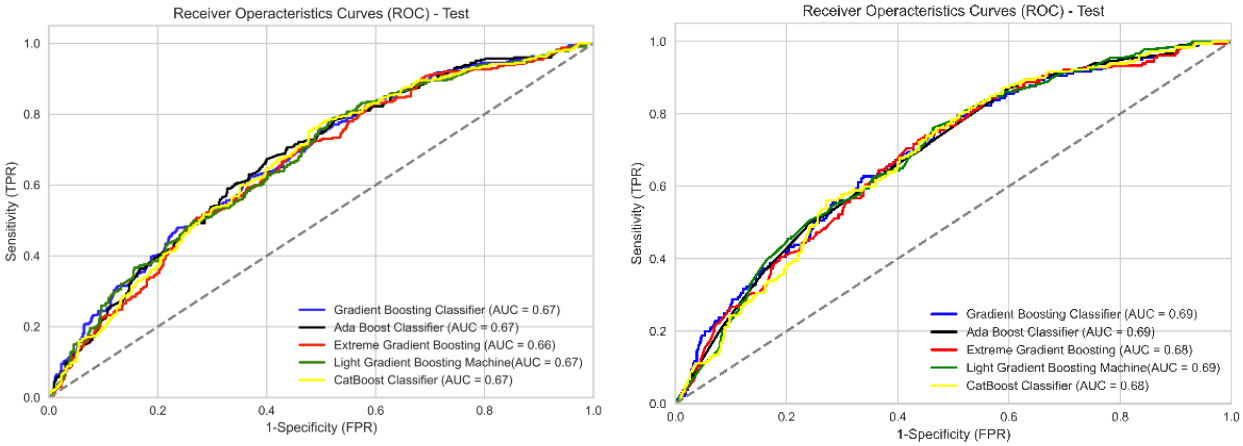
Dados de um total de 2.196 pares mãe-criança foram utilizados neste estudo. A Figura 1 apresenta as curvas ROC para os dados de teste para ECC e S-ECC. No modelo de predição de ECC, a área sob a curva ROC no treino e teste variou entre 0.65 e 0.72, precisão entre 0.37 e 0.50, recall entre 0.60 e 1.00, acurácia entre 0.38 e 0.63, escore F1 entre 0.54 e 0.56. No modelo de predição de S-ECC, a área sob a curva ROC no treino e teste variou entre 0.66 e 0.73, precisão entre 0.21 e 0.36, *recall* entre 0.39 e 1.00, acurácia entre 0.23 e 073, escore F1 entre 0.34 e 0.44. Assim, embora ambos os modelos apresentem valores semelhantes de área sob a curva ROC, com desempenho ligeiramente superior para S-ECC, o modelo para S-ECC mostrou-se menos efetivo na classificação, pois apresentou menor precisão e escores F1 mais baixos em comparação ao modelo de ECC, o que indica maior propensão a falsos positivos e menor equilíbrio global entre sensibilidade e especificidade.

As cinco variáveis que mais contribuíram para ambos os modelos de predição foram escolaridade materna, renda familiar, uso de fio dental por parte da mãe, idade materna no nascimento da criança e o CPO-S materno. O ordenamento foi levemente diferente entre os modelos, mas ambos apresentaram o nível de escolaridade materna como preditor mais forte, observando que quanto mais alto o nível de escolaridade materna, menor a probabilidade do diagnóstico de ECC e S-ECC. Esses achados foram consistentes com demais estudos que aplicaram

técnicas analíticas de aprendizado de máquina similares (PARK, 2021; TOLEDO REYES, 2023; EUSUFZAI, 2025). A contribuição do CPO-S materno foi acentuada para a performance do modelo de ECC e residual para o de S-ECC.

Na análise de imparcialidade, observaram-se diferenças de desempenho entre grupos raciais nos dois modelos. Ambos modelos apresentaram menor *recall* e maior acurácia para as crianças brancas comparado às pardas e pretas, refletindo um desequilíbrio entre sensibilidade e especificidade entre os grupos. Apesar dessas variações, a precisão foi consistentemente baixa em todos os grupos, evidenciando um problema sistemático de superestimação de casos positivos. Sobretudo, ambos modelos apresentaram desempenho preditivo aceitável, com potencial aplicabilidade em estratégias de prevenção precoce a partir de variáveis facilmente coletáveis na rotina odontológica.

Figura 1 - Curvas ROC para os dados de teste nos modelos para predição de ECC (à esquerda) e S-ECC (à direita).



4. CONCLUSÕES

Os resultados deste estudo suportam a hipótese de que dados de saúde bucal, fatores socioeconômicos e autorrelatados relacionados à saúde bucal da mãe contribuem para a detecção de cárie na primeira infância. Os modelos de aprendizado de máquina sugerem que esse conjunto de dados maternos têm valor preditivo aceitável, mas podem não ser suficientes para previsões altamente precisas. Sobretudo, ressalta-se o potencial da implementação do aprendizado de máquina para predição de cárie na primeira infância e identificação de grupos de risco a partir de preditores facilmente coletáveis nos primeiros mil dias de vida.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AMERICAN ACADEMY OF PEDIATRIC DENTISTRY. Policy on early childhood caries (ECC): Classifications, consequences, and preventive strategies. The Reference Manual of Pediatric Dentistry. Chicago, Ill.: **American Academy of Pediatric Dentistry**; 2020:79-81. Acessado em 20 ago. 2024. Online. Disponível em:https://www.aapd.org/globalassets/media/policies_guidelines/p_eccconsequences.pdf

COLLINS, G. S.; MOONS, K. G. M.; DHIMAN, P.; RILEY, R. D.; BEAM, A. L.; VAN CALSTER, B.; GHASSEMI, M.; LIU, X.; REITSMA, J. B.; VAN SMEDEN, M.; BOULESTEIX, A. L.; CAMARADOU, J. C.; CELI, L. A.; DENAXAS, S.; DENNISTON, A. K.; GLOCKER, B.; GOLUB, R. M.; HARVEY, H.; HEINZE, G.; HOFFMAN, M. M.; KENGNE, A. P.; LAM, E.; LEE, N.; LODER, E. W.; MAIER-HEIN, L.; MATEEN, B. A.; MCCRADDEN, M. D.; OAKDEN-RAYNER, L.; ORDISH, J.; PARNELL, R.; ROSE, S.; SINGH, K.; WYNANTS, L.; LOGULLO, P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. **BMJ, London**, v. 385, p. 1-11, 2024.

EUSUFZAI, S. Z.; JAMAYET, N. B.; AHMED, S.; ISLAM, M. B.; AHMAD, W. M. A. W.; ALAM, M. K. Development and evaluation of an early childhood caries prediction model: a deep learning-based hybrid statistical modelling approach. **European Archives of Paediatric Dentistry**, Berlin, p. 1-11, 2025.

LUNDBERG, S.; LEE, S. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, San Diego, v. 30, p. 4766–4775, 2017.

MARTÍNEZ-RICO, G.; ARGENTE-TORMO, J.; CALERO-PLAZA, J.; GONZÁLEZ-GARCÍA, R.J. The role of women in the field of early intervention. **Heliyon**, 2024: v.10, n.10, e31571.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Global oral health status report: Towards universal health coverage for oral health by 2030**. 2022. Acessado em 20 ago. 2024. Online. Disponível em: https://www.who.int/publications/i/item/97892400614840

PARK, Y. H.; KIM, S. H.; CHOI, Y. Y. Prediction Models of Early Childhood Caries Based on Machine Learning Algorithms. **International Journal of Environmental Research and Public Health**, Basel, v. 18, n. 16, p. 1-13, 2021.

TOLEDO REYES, L.; KNORST, J. K.; ORTIZ, F. R.; BRONDANI, B.; EMMANUELLI, B.; SARAIVA GUEDES, R.; MENDES, F. M.; ARDENGHI, T. M. Early Childhood Predictors for Dental Caries: A Machine Learning Approach. **Journal of Dental Research**, Thousand Oaks, v. 102, n. 9, p. 999–1006, 2023.