Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Métodos de aprendizado de máquina aplicados para a previsão de prematuridade e do peso ao nascer no Brasil com base em dados públicos.

Saulo A. Araujo[1], João V. R. dos Santos[2]
IMECC, Campinas, SP
Vinícius C. Gandolfi[3], Cristiano Torezzan[4]
FCA/UNICAMP, Limeira, SP

1 Introdução

O crescente uso de ferramentas de Aprendizado de Máquina (AM), ou de forma mais geral, de técnicas de Inteligência Artificial, tem impactado diversas áreas do conhecimento. Um dos grandes desafios para o uso dessas ferramentas é a disponibilidade de dados para o treinamento dos modelos. Esse problema é ainda mais relevante na área da saúde, pois há muitos dados sensíveis ou privados, que demandam cuidados específicos para sua divulgação.

Como forma de atenuar esse problema, o DATASUS (Departamento de Informática do Sistema Único de Saúde) disponibiliza um vasto conjunto de dados públicos que podem ser utilizados o treinamento e testes de algoritmos de aprendizado de máquina. Uma das bases disponibilizadas pelo DATASUS é o Sistema de Informações sobre Nascidos Vivos (SINASC), que reúne informações epidemiológicas referentes aos nascimentos informados em todo território nacional.

Neste trabalho, a base de dados do SINASC será utilizada para avaliar a viabilidade e a acurácia de métodos de aprendizado de máquina para a predição de prematuridade e de baixo peso ao nascer no Brasil. Tais variáveis são determinantes para a saúde do recém-nascido e fortemente associadas a elevados custos de serviços saúde como UTI neonatal [1]. Assim, métodos de AM podem ser utilizados tanto para a predição/prevenção de casos específicos, quanto para a formulação de políticas de saúde pública [2].

2 Metodologia

Para este estudo foram selecionadas como entradas para os modelos as seguintes variáveis: escolaridade da mãe, estado civil da mãe, raça/cor da mãe, número de consultas, tipo de gravidez, IDH do município, idade da mãe, quantidade de gestações anteriores, quantidade de partos normais anteriores, quantidade de partos cesários anteriores, mês de início de pré-natal, quantidade de filhos vivos, quantidade de filhos mortos. A escolha de tais variáveis se deve ao fato de que elas podem ser obtidas durante a gestação, período o qual é de interesse para realizar as predições. Além disso, as variáveis categóricas foram subdivididas em várias classes binárias, obtendo assim um total de 35 variáveis.

[1] s211290@dac.unicamp.br
[2] j237809@dac.unicamp.br
[3] v245274@dac.unicamp.br
[4] torezzan@unicamp.br

Os problemas foram modelados sob o paradigma de classificação binária supervisionada. Para isso, segundo [3], nascimentos com menos de 37 semanas de gestação foram rotulados como prematuros e recém-nascidos abaixo de 2500 gramas foram considerados de baixo peso. Como é usual, os dados foram divididos de forma aleatória em conjuntos de treinamento de teste, na proporção 70% e 30%, respectivamente. Além disso, foi feito uma subamostragem nos conjuntos de treino e teste para tratar o desbalanceamento das classes que foi observado após testes iniciais.

O estudo comparou o desempenho dos métodos de Regressão Logística (Logit), Random Forest e Multilayer Perceptron (MLP), que foram implementados em linguagem Python, com auxílio das bibliotecas scikitlearn e tensorflow [4].

3 Resultados

A Tabela 1 apresenta uma comparação dos métodos investigados para os problemas de predição de prematuridade e de baixo peso de nascidos no Brasil. As métricas apresentadas referem-se as médias da acurácia e F1-score dos conjuntos de testes.

Tabela 1: Acurácia dos métodos no conjunto de teste.

Problema	**Métrica**	**Logit**	**RF**	**MLP**
Prematuridade	Acurácia	0.8693	0.9155	0.8740
	F1-score	0.0210	0.6070	0.1362
Baixo peso	Acurácia	0.9274	0.9563	0.9294
	F1-score	0.0	0.5992	0.1352

Os resultados apresentados na Tabela 1 mostram o método Random Forest obteve resultados superiores aos métodos MLP e Regressão Logística. Para averiguar a consistências dos resultados foi aplicado o teste t de student [5] sobre as médias das acurácias, em todas as comparações o teste t de student obteve p-value menor que 0.05, indicando que existe uma diferença estatisticamente significativa entre as acurácias dos 3 modelos, evidenciando o método Random Forest como um melhor classificador para os problemas de predição abordados.

O conjunto completo de resultados deste trabalho inclui uma investigação sobre correlação entre as variáveis socio-econômicas e os fenômenos estudados, bem como uma abordagem de interpretabilidade para os métodos Random Forest e MLP.

Referências

[1] H. Â. D. C. Ramos e R. K. N Cuman. "Fatores de risco para prematuridade: pesquisa documental". Em: **Escola Anna Nery Revista de Enfermagem** 13 (2009), pp. 297–304.

[2] A. F. M. Batista e A. D. P. Chiavegatto. "Machine Learning Aplicado à Saúde". Em: **Minicursos do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. Sociedade Brasileira de Computação, 2019. Cap. 1, pp. 1–42. ISBN: 978-85-7669-472-4.

[3] S. S. Carvalho e et al. "Fatores maternos para o nascimento de recém-nascidos com baixo peso e prematuros: estudo caso-controle". Em: **Ciência e Saúde** 9.2 (2016), pp. 76–82.

[4] A. Géron. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. 2a. ed. O'Reilly Media, 2019.

[5] A. F. Oliveira. "Testes estatísticos para comparação de médias". Em: **Revista Eletrônica Nutritime** 5.6 (2008), pp. 777–788.