Predição de Surtos de Dengue e Diagnóstico de Sífilis Congênita Utilizando Aprendizado de Máquina

Robson Aleixo[1], **Fabio Kon**[1], **Raphael Y. de Camargo**[2]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)
São Paulo – SP – Brasil

[2]Centro de Matemática, Computação e Cognição – Universidade Federal do ABC (UFABC)
Santo André – SP – Brasil

robson.aleixo@alumni.usp.br, kon@ime.usp.br, raphael.camargo@ufabc.edu.br

Resumo. *A sífilis congênita e a dengue são duas doenças que causam impactos significativos no Brasil e em outros países do Hemisfério Sul, afetando a saúde de milhões de pessoas. A sífilis é uma infecção sexualmente transmissível (IST) que ao ser transmitida em crianças durante o período da gestação, é chamada de sífilis congênita. Já a dengue é uma doença viral transmitida pelos mosquitos Aedes Aegypti e Aedes Albopictus. Nesta dissertação, desenvolvemos aplicações inovadoras de modelos de aprendizado de máquina para essas doenças. O primeiro deles estima a probabilidade de uma criança nascer com sífilis. O segundo prevê surtos de dengue com base em dados sociodemográficos, climáticos, série histórica de casos, número de unidades de saúde, índice de mensuração de mosquitos e séries históricas de zika e chikungunya. No caso da sífilis congênita, avaliamos os modelos pela métrica AUC (Area Under Curve) e o resultado foi bom mas não excelente, i.e., 0.68 para a predição de casos positivos, obtidos pelos modelos LightGBM e XGBoost. No que se refere à dengue, o modelo Catboost obteve resultados muito bons, identificando 75% dos surtos com três meses de antecedência. Parte significativa deste trabalho foi investida na explicabilidade das predições de dengue, o que torna o modelo um importante aliado para a desenho de políticas públicas de saúde.*

Abstract. *Congenital syphilis and dengue are two diseases that cause significant impacts in Brazil and other countries in the Southern Hemisphere, affecting the health of millions of people. Syphilis is a sexually transmitted infection (STI) that, when transmitted to children during pregnancy, is called congenital syphilis. Dengue is a viral disease transmitted by the Aedes Aegypti and Aedes Albopictus mosquitoes. In this thesis, we developed innovative applications of machine learning models for these diseases. The first one estimates the probability of a child being born with syphilis. The second predicts dengue outbreaks based on sociodemographic and climatic data, historical series of cases, number of health units, mosquito measurement index, and historical series of zika and chikungunya. In the case of congenital syphilis, we evaluated the models using the AUC (Area Under Curve) metric and the result was good but not excellent, i.e., 0.68 for the prediction of positive cases, obtained by the LightGBM and XGBoost models. With regard to dengue, the Catboost model obtained very good results, identifying 75% of outbreaks three months in advance. A signi-*

ficant part of this work was invested in the explanation of dengue predictions, which makes the model an important ally for the design of public health policies.

1. Introdução e Motivação

A sífilis é uma infecção sexualmente transmissível (IST) causada pela bactéria *Treponema pallidum*. O diagnóstico em um recém nascido é complexo e depende de diferentes critérios, tais como clínico, sorológico e radiográfico. Ao nascer, alguns dos sintomas que uma criança pode vir a apresentar são dores nos ossos e inflamação articular [Guinsburg and Santos 2010].

Já a dengue é uma doença viral e seus surtos ocorrem com frequência no Brasil. Ao comparar o boletim epidemiológico do Ministério da Saúde de 2022 com o boletim de 2023 no período entre Janeiro e Março, percebe-se que houve um aumento de 53% no número de casos. Ou seja, mesmo atualmente, o Brasil enfrenta dificuldades em lidar com a ocorrência de casos.

Com relação aos trabalhos relacionados, encontramos poucos que utilizaram aprendizado de máquina voltados para sífilis congênita, como [Liu et al. 2010] e [Lago et al. 2004]. A maioria focou na avaliação de fatores críticos relacionados à doença somente. Já no caso da dengue, existem muitos trabalhos que buscam prever o número de casos futuros. Segundo o levantamento [Siriyasatien et al. 2018], foram identificados 966 modelos. Entretanto, as predições de médio e longo prazo ainda são um desafio, principalmente na identificação dos picos de casos. Além disso, no caso de trabalhos que usam modelos mais complexos, não foram utilizados métodos de interpretabilidade, que são necessários para que os responsáveis por políticas públicas entendam as predições. O código gerado neste trabalho pode ser acessado em `https://gitlab.com/interscity/health/dengue-prediction`.

2. Objetivos

O objetivo deste trabalho foi desenvolver modelos de aprendizado de máquina para auxiliar no diagnóstico de sífilis congênita e na predição de casos de dengue. No primeiro caso, o objetivo foi determinar a probabilidade de uma criança nascer com sífilis, utilizando como base de entrada os dados sobre a gestante e o recém nascido. Em relação à dengue, o objetivo foi predizer o número de casos até três meses no futuro, utilizando informações sócio-demográficas, ambientais, de estabelecimentos de saúde por região e séries históricas de casos. Além disso, utilizamos também o número de casos em regiões vizinhas para que a propagação entre os bairros pudesse ser representada.

3. Contribuições do Trabalho

No caso da sífilis congênita, criamos um modelo de aprendizado de máquina para auxiliar os profissionais de saúde no diagnóstico da doença. O modelo desenvolvido obteve um bom desempenho, com área sobre a curva (AUC) igual à 0.68. Este valor não permite sua aplicação em casos práticos, mas esta abordagem pode ser futuramente melhorada com a inclusão de mais informações sobre o período de gestação.

Para a predição de dengue, obtivemos as seguintes contribuições: (1) criamos um modelo para predição de surtos que gera predições com até três meses de antecedência;

(2) avaliamos o modelo na cidade do Rio de Janeiro; (3) fornecemos informações de interpretabilidade, indicando como os preditores foram utilizados pelo modelo para gerar as predições.

4. Resultados

Nesta seção mostramos os resultados obtidos com o modelo de dengue. Não incluímos os resultados do modelo de Sífilis devido às restrições de espaço.

4.1. Dados de Entrada

Analisamos dados do município do Rio de Janeiro, contendo informações epidemiológicas, ambientais e demográficas. O principal preditor utilizado foi o histórico de casos de dengue. Para cada bairro, consideremos o número de casos nos últimos três meses (`t-1`, `t-2` e `t-3`), além da soma de casos entre todos os bairros vizinhos no mês anterior (`sum_vizinhos`). Assim, fornecemos informações temporais e espaciais sobre a dengue na região.

Incluídos também dados de precipitação, temperatura média e umidade média do ar em cada mês para capturar as condições ambientais. Também incluímos dados adicionais de saúde pública: LIRAa (indicador de infestação do Aedes aegypti) e o número de casos de zika e chikungunya, por serem transmitidos pelo mesmo vetor da dengue. Por fim, consideramos o número de estabelecimentos de saúde, por estar associado as ações de saúde que a gestão pública aplica na região. Todas esses dados foram agrupados por bairro, ano e mês, considerando o período de 2016 a 2020.

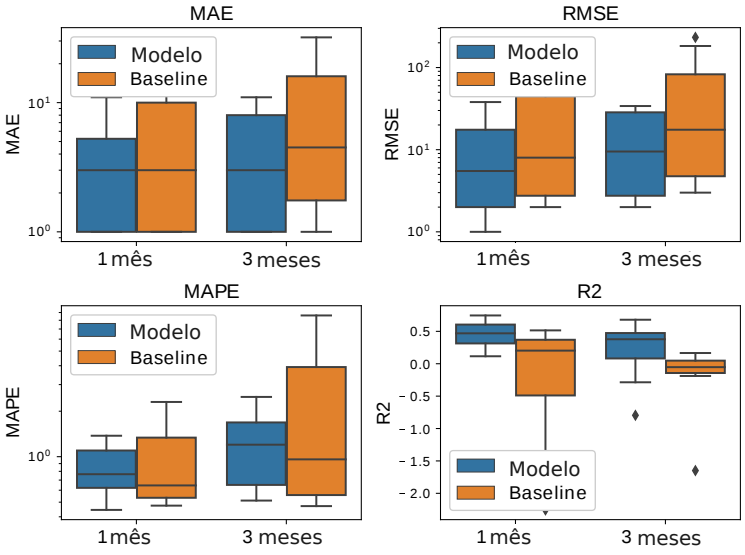
4.2. Desempenho do Modelo de Regressão

Utilizamos quatro métricas de avaliação de desempenho de modelos de regressão: R^2, MAE, MAPE e RMSE. As métricas MAE e RMSE avaliam o erro em valores absolutos e o MAPE em valores percentuais. Já o R^2 mostra o quanto da variabilidade dos dados é explicada pelo modelo.

Usamos o modelo *Catboost* para realizar as predições. Comparamos seus resultados com um modelo de *baseline*, o SARIMA [Hyndman and Athanasopoulos 2018]. A predição pelo SARIMA é baseada somente no histórico de casos de dengue e com isso podemos avaliar se a utilização de outras informações afetavam a qualidade da predição.

Os erros RMSE e MAPE do modelo seguem o mesmo padrão do MAE, com a diferença de que sua mediana para três meses foi maior do que para um mês (Figura 1). Para o R^2, valores maiores são melhores e, para um mês, o valor mediano foi próximo a 0,47, o que significa que o modelo previu cerca de 47% da variância no número de casos. No entanto, os resultados foram menos favoráveis para três meses, com uma mediana de 0,38 e o quartil 25% em 0,08.

Ao comparar os modelos catboost e SARIMA usando os erros MAE e RMSE, catboost apresentou valores medianos menores em todos os cenários, exceto para predições MAE de um mês, onde as medianas tiveram o mesmo valor. Ainda em comparação com o SARIMA, para predições de três meses, nosso modelo reduziu a mediana do MAE de 4,5 para 3,0 e a mediana do RMSE de 17,5 para 9,5. Em ambos os casos, reduziu também as amplitudes dos quartis 25% e 75%. Em outras palavras, nosso modelo reduziu tanto



Modelo (1 mês)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	1.0	0.44	0.12
25.0	1.0	2.0	0.62	0.31
50.0	3.0	6.0	0.77	0.47
75.0	5.0	18.0	1.1	0.6
100.0	11.0	38.0	1.38	0.75

Baseline (1 mês)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	2.0	0.47	-5.13
25.0	1.0	3.0	0.53	-1.15
50.0	3.0	8.0	0.65	0.11
75.0	10.0	45.0	1.34	0.3
100.0	16.0	63.0	2.3	0.51

Modelo (3 meses)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	2.0	0.51	-0.8
25.0	1.0	3.0	0.65	0.08
50.0	3.0	10.0	1.2	0.38
75.0	8.0	28.0	1.69	0.47
100.0	11.0	34.0	2.49	0.68

Baseline (3 meses)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	3.0	0.47	-86.49
25.0	2.0	5.0	0.56	-8.97
50.0	4.0	18.0	0.96	-0.16
75.0	16.0	83.0	3.92	-0.01
100.0	32.0	233.0	8.61	0.17

Figura 1. Avaliação de desempenho dos modelos *Catboost* e SARIMA pelas métricas MAE, RMSE, MAPE e R^2. A análise considera predições de um e três meses. Foram considerados os dados entre os anos de 2016 a 2020.

a mediana quanto a variabilidade dos erros. Os resultados foram menos claros para o MAPE, já que nosso modelo tendo erros MAPE medianos maiores, mas erros menores no quartil 75%. Os erros do MAPE são altamente influenciados pelo número de casos na baixa temporada, uma vez que o número de casos entra no denominador, e não é uma boa medida de erro para predições de surtos. Por fim, o SARIMA teve um desempenho muito ruim quando avaliado usando a métrica R^2 por três meses, mostrando que não pode explicar a variabilidade no número de casos de dengue.

4.3. Predição de Surtos

A respeito da definição de surto, não há um consenso na literatura, portanto definimos percentis que representassem os picos de casos que ocorreram no passado. A partir disso, definiu-se três grupos. No grupo 1, estão os valores abaixo do percentil 95, 35 casos por mês em cada região. No grupo 2, estão os valores entre os percentis 95 e 99, sendo o valor do percentil 99 igual a 120 casos. Para esse grupo denominou-se surtos leves. Por fim, no grupo 3, estão os valores acima do percentil 99, considerados surtos graves. Fizemos o mapeamento das predições dos modelos em cada um dos 3 grupos utilizando o número de casos preditos.

O modelo classificou corretamente a maioria dos surtos (Figura 2), mantendo os falsos positivos baixos mesmo com três meses de antecedência. Ele classificou corretamente que não haveria surtos em 97% das vezes (1ª linha da matriz C). Além disso, nas

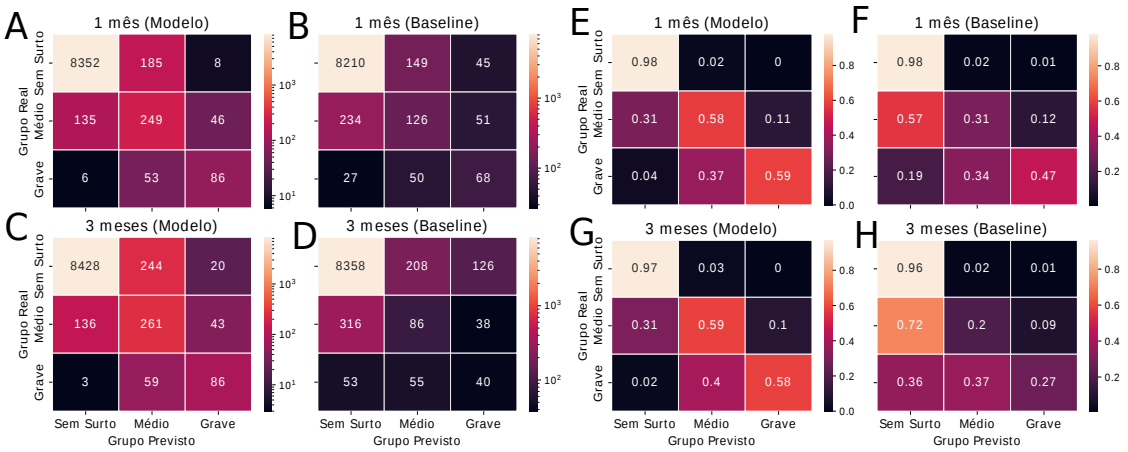


Figura 2. Matrizes de confusão para análise de surtos. Comparação entre os modelos *Catboost* e SARIMA. Nessa análise foram considerados os anos de 2016 a 2020, utilizando validação cruzada.

predições de surtos leves, houve uma chance de 57% de um surto leve ou grave (320 em 564), e quando predito como grave, uma chance de 86% de um surto (129 de 149), conforme mostrado na 2ª e 3ª colunas da matriz G. De maneira consolidada, o modelo foi capaz de prever 76% de todos os surtos futuros (448 de 587), estimados usando a combinação do 2º e 3ª linhas da matriz G.

4.4. Interpretabilidade do Modelo

Para a análise de interpretabilidade, utilizou-se o método *SHAP Values* [Lundberg et al. 2020]. Na Figura 3 são apresentados dois tipos de análises. No lado esquerdo, temos a avaliação global onde podemos observar que a quantidade de casos ocorridos no mês anterior, representado pelo t-1, foi a variável de maior importância na predição, atribuindo valores que vão de -10 até 800 aproximadamente. A segunda variável mais importante foi a *casos_criticos* que representa os bairros críticos da base de treino através da soma total de casos. Os valores de impacto proporcionados por essa variável estão entre -25 e 180. A importância dessa variável mostra que a alta ocorrência de casos costuma acontecer nas mesmas regiões em diferentes anos.

Nos quatro gráficos a direita, tem-se a avaliação dos impactos das variáveis em pares. O eixo x possui os valores da variável em análise, o eixo y possui os valores de impacto na predição. No gráfico superior a esquerda, sobre a variável de temperatura, é possível observar que a mudança entre os impactos negativos e positivos na predição ocorrem entre os valores de 25°C e 26°C. No gráfico ao lado, que analisa a precipitação, de maneira semelhante à temperatura, possui um intervalo entre 160mm e 180mm, onde o impacto deixa de ser negativo e passa a ser positivo.

5. Conclusão

A principal contribuição dessa pesquisa foi o desenvolvimento do modelo de aprendizado de máquina para a predição de casos de dengue. Mostramos que ele consegue identificar surtos com boa precisão com até três meses de antecedência. Todo esse processo foi sustentado por técnicas de explicabilidade dos resultados, permitindo ao gestor entender melhor os contextos que favorecem a propagação da doença. Este trabalho gerou

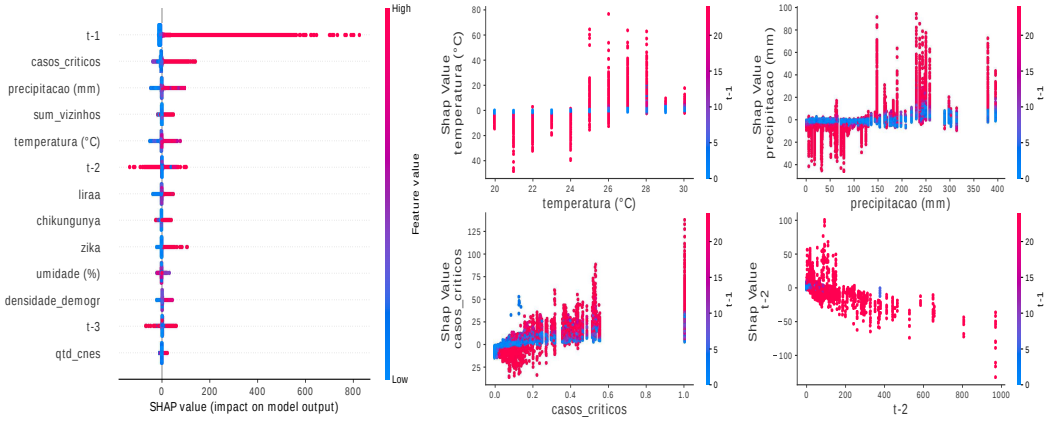


Figura 3. Análise de importância de variáveis do modelo *Catboost*. A análise considera a visão geral e em pares. Variáveis de temperatura, precipitação, casos_criticos e t-2 foram relacionadas com a variável t-1.

a publicação do artigo [Aleixo et al. 2022] que foi premiado como melhor artigo no primeiro Workshop Internacional *Artificial Intelligence for Health* (AI4Health), que ocorreu em conjunto com o CCGrid em 2022.

Referências

Aleixo, R., Kon, F., Rocha, R., Camargo, M., and de Camargo, R. (2022). Predicting dengue outbreaks with explainable machine learning. In *1st Int. Workshop on Artificial Intelligence for Health (AI4Health 2022)*, Taormina, Italy. IEEE Computer Society. Disponível em `https://ieeexplore.ieee.org/abstract/document/9826002`.

Guinsburg, R. and Santos, A. M. N. d. (2010). Critérios diagnósticos e tratamento da sífilis congênita.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Lago, E. G., Rodrigues, L. C., Fiori, R. M., and Stein, A. T. (2004). Congenital syphilis: identification of two distinct profiles of maternal characteristics associated with risk. *Sexually transmitted diseases*, 31(1):33–37.

Liu, J.-B., Hong, F.-C., Pan, P., Zhou, H., Yang, F., Cai, Y.-M., Wen, L.-Z., Lai, Y.-H., Lin, L.-J., and Zeegers, M. P. (2010). A risk model for congenital syphilis in infants born to mothers with syphilis treated in gestation: a prospective cohort study. *Sexually Transmitted Infections*, 86(4):292–296.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

Siriyasatien, P., Chadsuthi, S., Jampachaisri, K., and Kesorn, K. (2018). Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes. *IEEE Access*, 6:53757–53795.