Finisterra, *LVII*(121), 2022, pp. x-x
ISSN: 0430-5027
doi: 10.18055/Finis28635
Artigo

FATORES SOCIOECONÓMICOS E ALGORITMOS DE *MACHINE LEARNING* APLICADOS À PREDIÇÃO DE RISCO DE DOENÇAS NEGLIGENCIADAS.
ESTUDO DE CASO NOS MUNICÍPIOS DO ESTADO DE GOIÁS E DISTRITO FEDERAL, BRASIL

THAMY BARBARA GIOIA[1]
JULIANA RAMALHO BARROS[1]
RENATO RODRIGUES DA SILVA[2]

RESUMO – Analisar a relação entre variáveis socioeconómicas e doenças tropicais negligenciadas pode auxiliar os gestores no desenvolvimento de políticas públicas para a redução de casos. O objetivo deste trabalho foi avaliar, com base em algoritmos de *machine learning*, quais as variáveis socioeconómicas mais importantes para a classificação de risco de três doenças negligenciadas: hanseníase, leishmaniose tegumentar e dengue. Foram avaliados três algoritmos baseados em árvores de decisão: *Random Forest (RF)*, *XGBoost* e *C5.0*. Como área de estudo, delimitaram-se os municípios do Estado de Goiás e o Distrito Federal – Brasil. Para as classes de risco de dengue, tanto o algoritmo *RF* quanto o *XGBoost* apresentaram valores de exatidão acima de 0,6. Ambos destacaram como variáveis preditivas mais importantes as condições de baixa renda, alfabetização e raça. No caso das classes de risco de hanseníase, os três algoritmos apresentaram resultados de exatidão acima de 0,6 indicando como importantes as variáveis abastecimento de água, alfabetização, raça e moradia. No caso das classes de risco de leishmaniose tegumentar, os algoritmos apresentaram exatidão inferior a 0,4, inviabilizando a avaliação das possíveis variáveis preditivas ao modelo. Os três algoritmos avaliados apresentaram desempenho preditivo aproximado. No entanto, o *RF* foi ligeiramente superior. As variáveis socioeconómicas mais importantes para predição das classes de risco de dengue e hanseníase foram similares.

Palavras-chave: Doenças tropicas negligenciadas; determinantes sociais; *XGBoost*; *Random Forest*; *C5.0*.

ABSTRACT – SOCIOECONOMIC FACTORS AND MACHINE LEARNING ALGORITHMS APPLIED TO NEGLECTED DISEASES RISK PREDICTION. CASE STUDY IN THE MUNICIPALITIES OF THE GOIÁS STATE AND FEDERAL DISTRICT, BRAZIL. Analyzing the relation between socioeconomic variables and neglected tropical diseases can help managers in the conception of public policies to reduce cases. The objective of this study was to evaluate, based on machine learning algorithms, which socioeconomic variables are more important for the risk classification of three neglected diseases: leprosy, cutaneous leishmaniasis, and dengue. Three algorithms based on decision trees were evaluated: *Random Forest (RF)*, *XGBoost*, and *C5.0*. As a study area, the municipalities of the state of Goiás and of the Federal District – Brazil, were delimited. For the dengue risk classes, both the *RF* algorithm and the *XGBoost* showed accuracy values above 0.6. Both emphasizing the low-income conditions, literacy, and race as the most important predictive variables. In the leprosy risk classes case, the three algorithms presented accuracy results above 0.6, indicating the variables water supply, literacy, race, and housing as important. For the tegumentary leishmaniasis risk classes, the algorithms showed an accuracy lower than 0.4, making the evaluation of possible predictive variables to the model unfeasible. The three evaluated algorithms revealed approximate predictive performance; however, the *RF* was slightly higher. The most important socioeconomic variables for dengue and leprosy risk classes prediction were similar.

Keywords: Neglected tropical diseases; social determinants; *XGBoost*; *Random Forest*; *C5.0*.

RÉSUMÉ – FACTEURS SOCIOÉCONOMIQUES ET ALGORITHMES MACHINE LEARNING APPLIQUÉS À LA PRÉDICTION DES RISQUES DE MALADIES NÉGLIGÉES. ÉTUDE DE CAS DANS LES MUNICIPALITÉS DE L'ÉTAT DE GOIÁS ET DU DISTRICT FÉDÉRAL, BRÉSIL. Analyser la relation entre les variables socio-économiques et les maladies tropicales négligées peut accompagner les gestionnaires dans l'élaboration de politiques publiques pour réduire les cas. L'objectif de ce travail était d'évaluer, sur la base d'algorithmes machine learning, quelles variables socio-économiques les plus importantes pour la classification des risques de trois maladies négligées: la lèpre, la leishmaniose tégumentaire et la dengue. Trois algorithmes basés sur des arbres de décision ont été considérés: *Forêt Aléatoire (FA)*, *XGBoost* et *C5.0*. La zone d'étude délimitée sont les municipalités de la province de Goiás et le District Fédéral, situées dans la région Centre-Ouest du Brésil. Pour les classes de risque de dengue, l'algorithme FA et *XGBoost* ont présenté des valeurs de exactitude supérieures à 0,6. Les deux ressortent comme des variables plus prédictives facteurs tels que les conditions de faible revenu, l'alphabétisation et la race. Dans le cas des classes de risque de lèpre, les trois algorithmes ont présenté des résultats de exactitude supérieurs à 0,6, indiquant comment paramètres importants tels que l'approvisionnement en eau, l'alphabétisation, la race et les conditions de logement. Dans le cas des classes de risque de leishmaniose tégumentaire, les algorithmes

Recebido: 18/11/2022. Aceite: 20/12/2022. Publicado: 30/12/2022.
[1] Instituto de Estudos Socioambientais (IESA), Universidade Federal de Goiás, Av. Esperança, s/n, Samambaia, 74001-970, Goiânia, Goiás, Brasil. E-mail: thamygioia@gmail.com, juliana@ufg.br
[2] Instituto de Matemática e Estatística (IME), Universidade Federal de Goiás, Goiânia, Goiás, Brasil. E-mail: renato.rrsilva@ufg.br

ont adopté une exactitude inférieure à 0,4, rendant impossible l´évaluation d´eventuelles variables prédictives au modèle. Les trois algorithmes évalués ont montré performances prédictives approximatives, cependant, le FA était légèrement supérieur. Les variables socio-économiques les plus importantes pour prédire les classes de risque de dengue et de lèpre étaient similaires.

Mot clés: Maladies tropicales négligées; déterminants sociaux; *XGBoost*; *Forêts Aléatoires*; *C5.0*.

RESUMEN – FACTORES SOCIOECONÓMICOS Y ALGORITMOS DE MACHINE LEARNING APLICADOS A LA PREDICCIÓN DE RIESGO DE ENFERMEDADES DESATENDIDAS. ESTUDIO DE CASO EN LOS MUNICIPIOS DEL ESTADO DE GOIÁS Y DEL DISTRITO FEDERAL, BRASIL. Analizar la relación entre las variables socioeconómicas y las enfermedades tropicales desatendidas puede ayudar los gestores en la producción de políticas públicas para la reducción de casos. El objetivo de este trabajo fue evaluar, con base en algoritmos de machine learning, qué variables socioeconómicas son más importantes para la clasificación de riesgo de tres enfermedades desatendidas: lepra, leishmaniasis cutánea y dengue. Se evaluaron tres algoritmos basados en árboles de decisión: *Random Forest (RF)*, *XGBoost* y *C5.0*. Como área de estudio, fueron delimitados los municipios del Estado de Goiás y del Distrito Federal – Brasil. Para las clases de riesgo de dengue, tanto el algoritmo *RF* como el *XGBoost* presentaron valores de exactitud superiores a 0,6. Ambos resaltan como las variables predictivas más importantes las condiciones de baja renta, alfabetización y raza. En el caso de las clases de riesgo de lepra, los tres algoritmos presentaron resultados de exactitud superiores a 0,6, lo que indica que las variables suministro de agua, alfabetización, raza y vivienda son importantes. En el caso de las clases de riesgo de leishmaniasis cutánea, los algoritmos mostraron una exactitud inferior a 0,4, haciendo inviable la evaluación de posibles variables predictivas del modelo. Los tres algoritmos evaluados presentaron un rendimiento predictivo aproximado, sin embargo, el *RF* fue ligeramente superior. Las variables socioeconómicas más importantes para la predicción de las clases de riesgo de dengue y de lepra fueron similares.

Palavras clave: Enfermedades tropicales desatendidas; determinantes sociales; *XGBoost*; *Random Forest*; *C5.0*.

I. INTRODUÇÃO

As doenças negligenciadas correspondem a um grupo de doenças assim classificadas por receberem baixos investimentos em pesquisa e em produção de medicamentos, além de prevalecerem em condições de vulnerabilidade social e de serem mais frequentes em países em desenvolvimento (World Health Organization [WHO], 2020). A vulnerabilidade social em saúde pode ser avaliada a partir de variáveis socioeconómicas, também definidas como determinantes sociais em saúde, e que consideram aspectos tais como condições de renda, educação, saneamento básico e habitação (Barata, 2009; Souza *et al.*, 2015).

Hanseníase, leishmaniose tegumentar e dengue são doenças negligenciadas e que apresentam prevalência no Brasil. Em 2018, foram registados 36 766 casos de hanseníase no país, aproximadamente 1,76 casos a cada 10 000 habitantes, quando recomendações da WHO sugerem índices abaixo de um caso a cada 10 000 habitantes (WHO, 2020). No estado de Goiás, foram 1791 registos da doença para o mesmo ano, enquanto para o Distrito Federal foram 205 registos, colocando-os, respectivamente, em 8º e 22º lugar no *ranking* nacional (Sistema de Informação de Agravos de Notificação [SINAN], 2018). Para a dengue, no ano de 2018, registaram-se 265 460 casos no Brasil, sendo 91 530 no estado de Goiás e 2444 no Distrito Federal. Em 2018, Goiás ocupou o 1º lugar no *ranking* de casos no país, enquanto o Distrito Federal ocupou o 16º lugar (SINAN, 2018). Para os casos de leishmaniose tegumentar, em 2018, foram registados 17 950 casos da doença no Brasil, 323 no Estado de Goiás, colocando-o em 12º lugar no *ranking* nacional e 34 casos no Distrito Federal, colocando-o em 23º lugar (SINAN, 2018).

A leishmaniose tegumentar é uma doença infecciosa e não transmissível causada por diferentes espécies de protozoários do género *Leishmania*. É considerada uma infecção zoonótica, ou seja, a partir de um hospedeiro (animais silvestres ou domésticos) a doença é transmitida para seres humanos. Nos seres humanos a doença manifesta-se por meio de lesões cutâneas ou nas mucosas. A transmissão da doença está, por vezes, relacionada com alterações significativas da paisagem, como a expansão de atividades agrícolas, de mineração, ocupação de encostas e de áreas periféricas próximas a matas secundárias (BRASIL, 2017).

No caso da hanseníase, a transmissão ocorre pelas vias respiratórias, de humano para humano. A doença é causada pelo parasita *Mycobacterium leprae,* também conhecido como bacilo de Hansen. Os sintomas são caracteristicamente dermatológicos, com lesões na pele e nervos periféricos, podendo evoluir para incapacidades físicas. Além de condições individuais, situações de vulnerabilidade social

relacionadas com condições precárias de moradia e alta densidade habitacional, podem favorecer a transmissibilidade da doença (BRASIL, 2002).

Quanto à dengue, a doença é causada pelo vírus do gênero *Flavivirus*. A fonte de infecção é o ser humano que a transmite por meio de um vetor do gênero *Aedes* da espécie *Aedes aegypti*. Dentre os sintomas da doença destacam-se: febre, cefaleia, náuseas, vómitos e dores abdominais (BRASIL, 2015). De acordo com Valle (2021), os grandes centros urbanos costumam apresentar maiores índices de infestação da doença, considerando a associação de condições precárias de infraestrutura urbana e condições naturais de temperatura e precipitação.

Dentre os recursos técnicos aplicáveis para a modelação de dados estão os algoritmos de aprendizado de máquina (*machine learning*), que, tecnicamente, processam dados de entrada visando predizer resultados de classificação e/ou de regressão (Géron, 2019). Na saúde, os algoritmos têm sido usados na tentativa de predizer variáveis potenciais no diagnóstico de doenças, evolução de óbitos e contextos de vulnerabilidade (Santos *et al.*, 2020). Outro aspecto significativo por parte dos algoritmos de *machine learning* é a possibilidade de avaliar a relevância relativa de cada variável quando comparada a outras em uma predição (Géron, 2019).

Desta forma, o objetivo deste trabalho foi avaliar, com base em algoritmos de *machine learning*, quais as variáveis socioeconómicas mais importantes para a classificação de risco de três doenças negligenciadas: hanseníase, leishmaniose tegumentar e dengue. Estas doenças foram selecionadas considerando a lista de doenças negligenciadas da Organização Mundial da Saúde (WHO, 2020) e levando em consideração as taxas de prevalência observadas nos municípios do Estado de Goiás e Distrito Federal, área de estudo deste trabalho, e para os períodos com dados oficiais disponíveis (2001-2018) (SINAN, 2018).

Por fim, considerando que leishmaniose tegumentar, dengue e hanseníase são classificadas como doenças negligenciadas e que estas prevalecem em condições de vulnerabilidade social (WHO, 2020), optou-se por avaliar apenas as condições socioeconómicas das áreas de estudo. A hipótese é que as áreas em análise mostrarão similaridade no que diz respeito às variáveis socioeconómicas mais pertinentes para a predição das classes de risco para as três doenças analisadas.

II. METODOLOGIA

Como área de estudo, delimitaram-se os 246 municípios do estado de Goiás e o Distrito Federal, ambos localizados na região Centro-Oeste do Brasil (fig. 1).

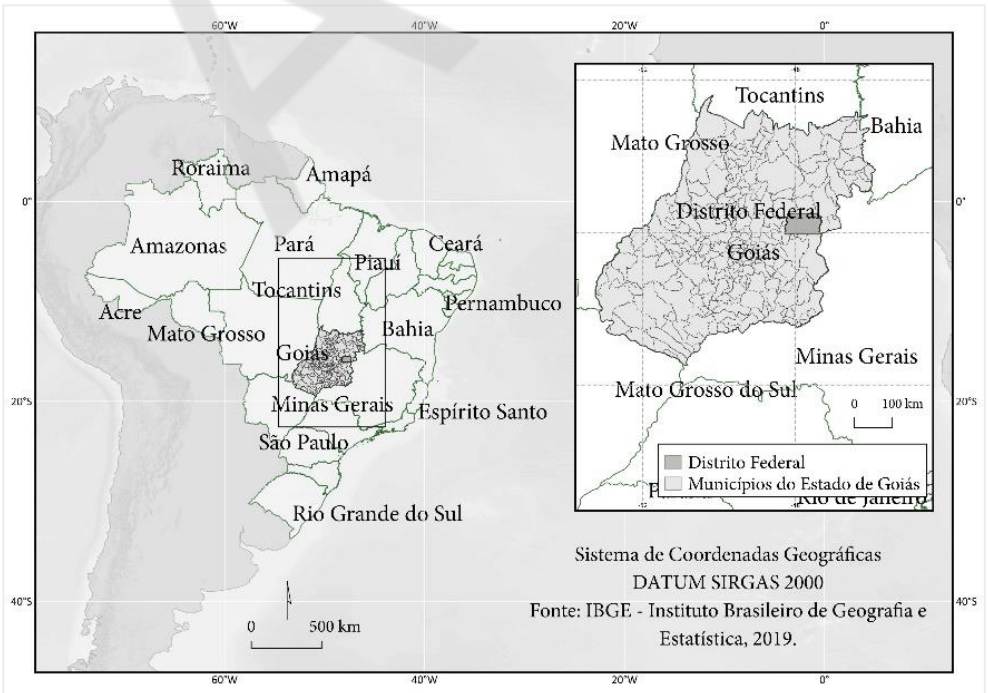

Fig. 1 – Localização do estado de Goiás e do Distrito Federal.

Fig. 1 – Goiás and Federal District location.

Para o cálculo das taxas de prevalência, foram utilizados os registos de casos de hanseníase, de leishmaniose tegumentar e de dengue e as projeções populacionais referentes ao período de 2001 a 2018, disponíveis no SINAN (2018).

As variáveis socioeconómicas empregadas para a avaliação de risco foram obtidas no Sistema de Recuperação Automática (SIDRA) do Instituto Brasileiro de Geografia e Estatística (IBGE), referente aos censos demográficos dos anos 2000 e 2010.

Por fim, as variáveis socioeconómicas preditivas avaliadas seguiram o trabalho produzido pela Gerência de Epidemiologia e Informação (GEEPI) do município de Belo Horizonte – MG (2003) e são ilustradas no quadro I.

Quadro I – Variáveis independentes utilizadas para modelação dos algoritmos.

Table I – Independent variables used to modeling the algorithms.

Cód.	Descrição
V01	Média de moradores por domicílio
V02	População acima de 10 anos com classe de renda até 1 salário-mínimo (%)
V03	População acima de 10 anos com classe de renda de mais de um a dois salários-mínimos (%)
V04	População acima de 10 anos com classe de renda de mais de dois a três salários-mínimos (%)
V05	População acima de 10 anos com classe de renda de mais de três a cinco salários-mínimos (%)
V06	População acima de 10 anos com classe de renda de mais de cinco a dez salários-mínimos (%)
V07	População acima de 10 anos com classe de renda de mais de dez a vinte salários-mínimos (%)
V08	População acima de 10 anos com classe de renda superior a vinte salários-mínimos (%)
V09	População acima de 10 anos sem rendimento (%)
V10	População autodeclarada branca (%)
V11	População autodeclarada preta (%)
V12	População autodeclarada amarela (%)
V13	População autodeclarada parda (%)
V14	População autodeclarada indígena (%)
V15	Domicílios com rede geral de esgoto ou pluvial (%)
V16	Domicílios com fossas sépticas (%)
V17	Domicílios com tipo de esgotamento inadequado (%)
V18	Domicílios com coleta de resíduos de serviço público (%)
V19	Domicílios com disposição inadequada de resíduos (%)
V20	Domicílios com abastecimento de água por rede pública (%)
V21	Domicílios com abastecimento de água via poço na propriedade (%)
V22	Domicílios com abastecimento inadequado de água (%)
V23	População alfabetizada (%)
V24	População não alfabetizada (%)

1. Pré-processamento das taxas

As taxas de prevalência das três doenças analisadas foram calculadas a partir da média de dados disponíveis para dois períodos: 2001-2009 e 2010-2018, com o intuito de descartar possíveis flutuações aleatórias nos registos e de compatibilizar com os dados socioeconómicos oficiais disponibilizados relativos ao censo demográfico brasileiro nos anos de 2000 e de 2010.

Na sequência, as taxas foram padronizadas a cada 100 000 habitantes e classificadas em categorias de risco para cada doença negligenciada conforme documentos de orientação do Ministério da Saúde do Brasil e da Organização Mundial da Saúde (Brasil, 2015; Departamento de Informática do SUS [DATASUS], (n.d); WHO, 2019) (quadro II).

Os dados foram organizados numa planilha, contendo a taxa média de prevalência classificada por risco para as três doenças analisadas e as variáveis socioeconómicas do quadro I, concernentes aos anos 2000 e 2010. Para as taxas do período de 2001 a 2009, avaliaram-se os dados socioeconómicos de 2000 e, para o período de 2010 a 2018, as variáveis de 2010.

Quadro II – Classificação de risco a partir das taxas de prevalência para dengue, hanseníase e leishmaniose tegumentar.

Table II – Risk classification based on prevalence rates for dengue, leprosy, and cutaneous leishmaniasis.

Classes	Dengue*	Hanseníase*	Leishmaniose tegumentar*
Baixa	Até 100	Menor que 2,00	Até 0,95
Média	101 - 299	2,00 - 9,99	0,96 - 4,94
Alta	300 - 599	10,00 - 19,99	4,95 - 12,69
Muito Alta	600 - 799	20,00 - 39,99	12,70 - 26,71
Hiperendêmica	Acima de 800	Maior igual a 40	26,72 - 46,50

*Casos por 100 000 habitantes.

Fonte: Dengue (Brasil, 2015); Hanseníase (DATASUS, n.d); Leishmaniose tegumentar (WHO, 2019). Adaptado pelo autor.

2. Algoritmos de *Machine Learning*

Para a predição de classes de risco das três doenças analisadas, foram utilizados três algoritmos de *machine learning* para classificação supervisionada baseados em árvores de decisão: o *Random Forest* (Breiman, 2001), o *XGBoost* (Chen & Guestrin, 2016) e o *C5.0* (Quinlan, 1993). Na classificação, uma árvore de decisão funciona como um conjunto de regras hierárquicas de divisão de variáveis (chamados de nós) em subconjuntos por meio de regras (denominados de ramos) até que se obtenha um subconjunto homogéneo o suficiente para ser classificado como uma mesma classe, obtendo, assim, um nó terminal (chamado de folha) (fig. 2).

Os três algoritmos utilizados abordam estratégias que geram uma série de árvores de decisão, os quais permitem uma modelação mais robusta do que a produção de uma única árvore. O que diferencia os algoritmos são as características pertinentes ao modo de treinamento dos modelos e suas características operacionais (Breiman, 2001; Chen & Guestrin, 2016).

No caso do *RF*, a decisão final para classificação consiste na combinação de árvores criadas de forma independente, em que cada árvore é ajustada a partir de um vetor de atributos amostrados a partir do método *bootstrap*. Os hiperparâmetros desse algoritmo são: o número de árvores (*ntree*) a serem criadas e o número de variáveis (*mtry*) testadas (Breiman, 2001). Neste estudo foram utilizados como parâmetros o valor de *ntree* igual a 500 e *mtry* a raiz quadrada do número total de variáveis ($\sqrt{N}$) de entrada, condições-padrão do pacote estatístico R.

O algoritmo *C5.0* é um tipo de árvore de decisão construída a partir do particionamento recursivo dos dados. Para melhorar a performance de classificação, usa-se o método *boosting*. Esse método consiste em ajustar sequencialmente vários ensaios do algoritmo, atribuindo pesos para as observações que foram classificadas incorretamente. O hiperparâmetro otimizado foi o número de ensaios, que nesse estudo foi definido igual a 20 (Kuhn & Johnson, 2016).

O *XGBoost* trabalha a partir de árvores sequenciais para chegar aos resultados de classificação, numa abordagem conhecida como *Gradient Boosting*. Cada árvore é construída sequencialmente considerando e corrigindo o erro da árvore anterior, ou seja, a árvore inicial estará associada a um erro residual, a próxima árvore será construída com ajuste ao erro residual da etapa anterior. Os resultados anteriores serão combinados para a construção de uma nova árvore onde a raiz do erro quadrático médio será menor que a antecessora. O processo é contínuo até se obter o menor erro residual (Chen & Guestrin, 2016; Espinosa-Zuniga, 2020).

A modelação foi realizada no *software* livre *R* versão 4.0.3, por meio do pacote *Classification and Regression Training* – CARET (Kuhn *et al.*, 2021).

Para o treinamento da modelação, 70% dos dados foram utilizados como amostra de treinamento e calibrados com a validação cruzada *k-fold* para k=5. Os demais 30% de amostras foram separados para teste, com o propósito de validar os algoritmos com um conjunto de dados independente do usado no treino. A métrica de validação empregada foi a exatidão, com intervalo de confiança de 95% a partir do teste *Exact Binomial* (Clopper & Pearson, 1934).

Finalmente, examinou-se quais as variáveis mais importantes no processo de predição das classes de risco das três doenças negligenciadas analisadas. Cada algoritmo possui métricas distintas para ordenar a importância das variáveis (Kuhn *et al.*, 2021). Todas as métricas têm como princípio atribuir maior peso às variáveis que estão nos nós superiores da árvore de decisão. Como essa ordenação de importância de variáveis permite a normalização das variáveis entre 0 e 100, isso possibilita a comparação relativa entre os algoritmos.

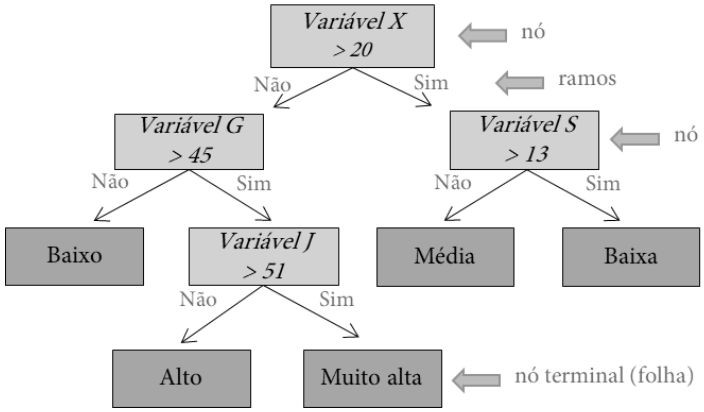

Fig. 2 – Exemplo de árvore de decisão com variáveis e valores hipotéticos.
Fig. 2 – Example of decision tree with hypothetical variables and values.

III. RESULTADOS

No quadro III exibem-se os resultados para as amostras de treino (validação cruzada) e de teste, com o intervalo de confiança de exatidão para os três algoritmos e para a categorização das três taxas de prevalência classificadas.

A partir dos resultados, é possível observar que:

- Para as classes de risco de dengue, os algoritmos *Random Forest* (*RF*) e *XGBoost* revelaram desempenho pouco superior se comparado ao algoritmo *C5.0*. Os resultados de exatidão (dados de teste) para o algoritmo *RF* foram de 0,58, com intervalo de confiança de 0,50-0,66; de 0,56 para o algoritmo *XGBoost*, com intervalo de confiança de 0,48-0,65; e de 0,46 para o algoritmo *C5.0*, com intervalo de confiança de 0,38-0,56. Logo, tanto o algoritmo *RF* quanto o *XGBoost* permitiram explicar cerca de 60% da variância dos dados das classes de risco de dengue, a partir das variáveis preditivas selecionadas para modelação;
- Para as classes de risco das taxas de hanseníase, os três algoritmos tiveram resultados de exatidão dos dados de teste em intervalo de confiança acima de 0,6. Frisa-se que os algoritmos *RF* e *XGBoost* permitiram explicar cerca de 70% da variância dos dados acerca das classes de risco de hanseníase;
- Para as classes de risco de leishmaniose tegumentar, os algoritmos apresentaram exatidão (dados de teste) abaixo de 0,4. Dessa forma, optou-se por prosseguir com as análises referentes à importância das variáveis preditivas apenas para as classes de risco de dengue e de hanseníase.

Quadro III – Resultados de exatidão para os algoritmos avaliados.
Table III – Accuracy results for the evaluated algorithms.

Algoritmos	***RF***			***XGBOOST***			***C5.0***		
	TDC*	**THC**	**TLC**	**TDC**	**THC**	**TLC**	**TDC**	**THC**	**TLC**
Exatidão dos dados de Treino	0,53	0,64	0,32	0,52	0,63	0,33	0,49	0,62	0,29
Exatidão dos dados de Teste	**0,58**	**0,66**	0,28	**0,56**	**0,64**	0,26	0,46	**0,59**	0,27
Intervalo de confiança dos dados de teste (95%)	0,50-0,66	0,58-0,74	0,21-0,36	0,48-0,65	0,55-0,71	0,19-0,34	0,38-0,55	0,51-0,67	0,20-0,35

* TDC – Classes da taxa de prevalência de dengue; THC – Classes da taxa de prevalência de hanseníase; TLC – Classes da taxa de prevalência de leishmaniose tegumentar classificadas.

Para as classes de risco de dengue, verifica-se que, considerando as cinco primeiras variáveis mais importantes, tanto o algoritmo *RF* quanto o *XGBoost* identificaram como de maior importância a variável V09 – População acima de 10 anos sem rendimento (%), seguida das variáveis V03 – População acima de 10 anos com classe de renda de mais de um a dois salários-mínimos (%) e V12 – População autodeclarada amarela (%), e V23 – População alfabetizada (%) (fig. 3). Para a modelação

do algoritmo *RF*, enfatiza-se, ainda, a variável V01 – Média de moradores por domicílio, e, para o algoritmo *XGBoost*, a variável V13 – População autodeclarada parda (%).

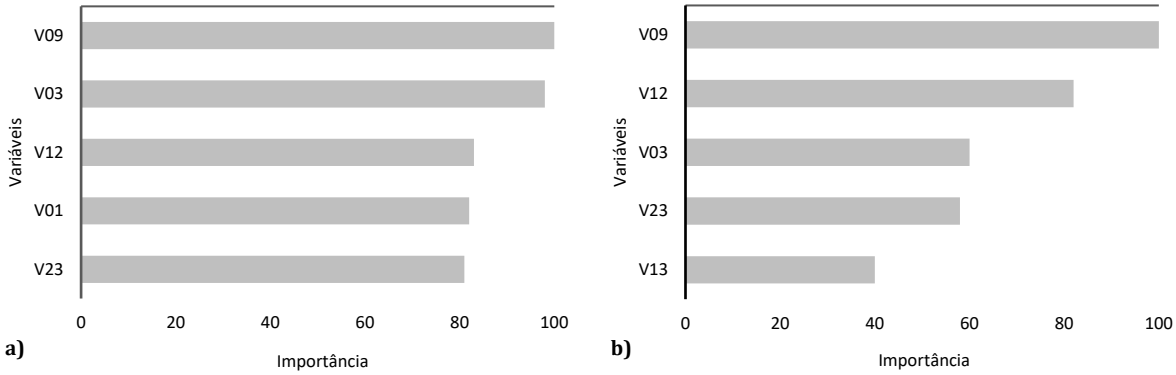

Fig. 3 – Importância das variáveis preditivas para as classes de risco de Dengue; a) Modelação a partir do algoritmo *Random Forest*; e b) Modelação a partir do algoritmo *XGBoost*.

Fig. 3 – Importance of predictive variables for Dengue risk classes; a) Modeling from the Random Forest algorithm; and b) Modeling from the XGBoost algorithm.

Os resultados atinentes às classes de risco para hanseníase indicaram como variáveis preditivas mais importantes a V23 – População alfabetizada (%) e a V13 – População autodeclarada parda (%), que aparecem entre as cinco primeiras variáveis de importância para os três algoritmos (fig. 4). A variável V01 – Média de moradores por domicílio é apontada como muito importante tanto para o algoritmo *RF* quanto para o *XGBoost* (figs. 4a e 4b). As variáveis V22 – Domicílios com abastecimento inadequado de água (%) e V05 – População acima de 10 anos com classe de renda de mais de três a cinco salários-mínimos (%) são sinalizadas como de maior importância para o algoritmo *C5.0*, e, por fim, a variável V11 – População autodeclarada preta (%) aparece entre as cinco primeiras variáveis de importância para os algoritmos *RF* e *C5.0* (figs. 4a e 4c).

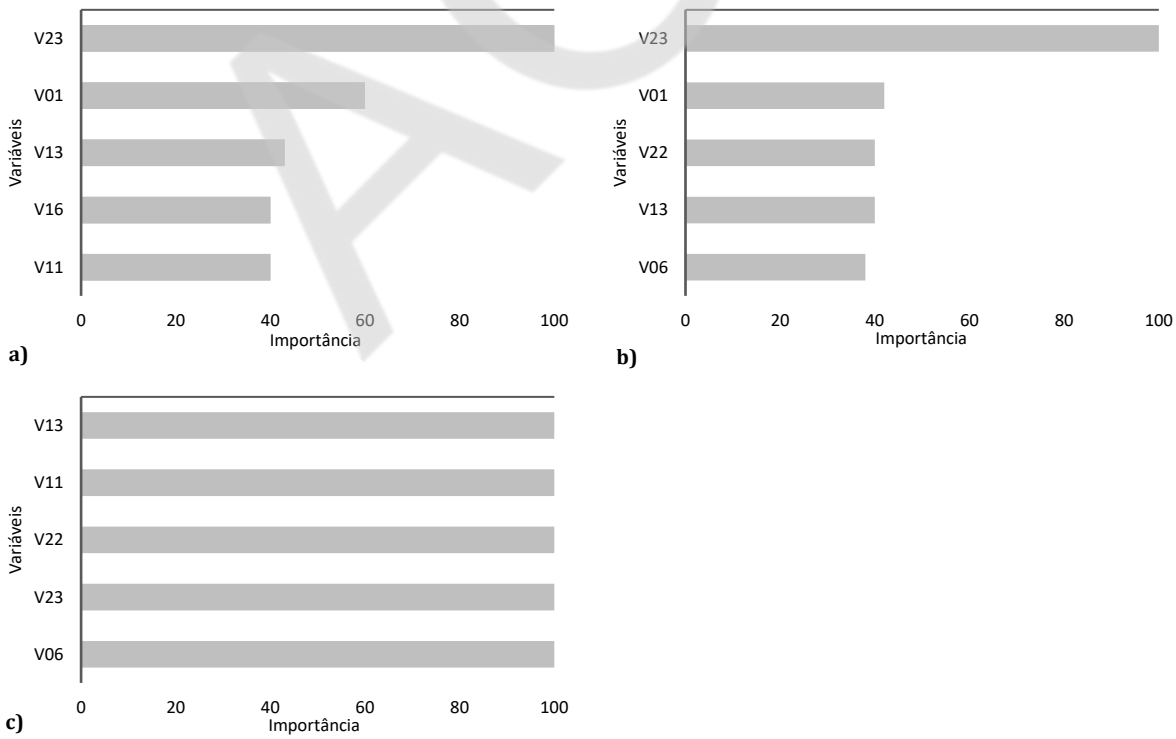

Fig. 4 – Importância das variáveis preditivas para as classes de risco de Hanseníase. a) Modelação a partir do algoritmo *Random Forest*; b) Modelação a partir do algoritmo *XGBoost*; e c) Modelação a partir do algoritmo *C5.0*.

Fig. 4 – Importance of predictive variables for leprosy risk classes. a) Modeling from the Random Forest algorithm; b) Modeling from the XGBoost algorithm; and c) Modeling from the C5.0 algorithm.

IV. DISCUSSÃO

Os municípios do Estado de Goiás e Distrito Federal apresentaram condições de renda distintas para predição das classes de risco de dengue e hanseníase. Baixas condições salariais (% de população sem rendimento e de um até dois salários-mínimos) foram variáveis importantes para a predição das classes de risco de dengue, diferentemente do observado para as classes de risco de hanseníase, em que a condição de renda superior – três a cinco salários-mínimos foi identificada como preditiva para a modelação no algoritmo *C5.0*.

Baixas condições de renda associadas à dengue foram constatadas em estudos realizados por Honorato *et al.* (2014) no estado do Espírito Santo, onde variáveis como população com renda abaixo de três salários e coleta deficiente de resíduos sólidos apresentaram o melhor desempenho, ao que o autor chamou de modelo de efeito espacial. Verifica-se, todavia, que essa condição não é padrão. Num estudo conduzido no estado da Paraíba, altas incidências de dengue foram encontradas em áreas com melhores condições de renda (Silva *et al.*, 2020).

No caso da hanseníase, variáveis concernentes à condição de renda inferiores não foram as mais pertinentes para a predição, mas a percentagem de população alfabetizada aparece com maior importância dentre as variáveis para os três algoritmos analisados, seguida de condições de abastecimento de água inadequado e de percentagem de população de raça parda e negra, modeladas a partir do algoritmo *C5.0*. Ressalta-se que condições negativas de saneamento básico estão frequentemente atreladas a condições negativas de renda e de escolaridade, se levados em conta os contextos de vulnerabilidade social (Gerência de Epidemiologia e Informação, 2003; Instituto de Pesquisa Econômica Aplicada, 2015; WHO, 2020), de forma que tais resultados devem ser mais bem explorados.

Condições de saneamento básico inadequadas associadas às taxas de incidência de hanseníase foram averiguadas num estudo realizado por Monteiro *et al.* (2017) no Estado do Tocantis no período de 2001 a 2012. Domicílios com menor proporção de atendimento por água encanada e com presença de banheiro (<61,7%), apresentaram IRR (Razão da taxa de incidência) de 0,627. Para a variável coleta de resíduos, o IRR foi de um para domicílios com proporção de coleta inferior a 88,8%.

Quanto à condição etária e étnico-racial, uma pesquisa feita nas regiões Norte e Nordeste do Brasil detectou risco acrescido de mortalidade para pacientes com hanseníase do sexo masculino, com idade acima de 60 anos e de raça-cor parda e preta. De acordo com os autores, considerando o caráter crónico da hanseníase, o maior risco de mortalidade na faixa etária acima de 60 anos pode indicar baixa qualidade de vida. Com relação à raça-cor parda, a associação encontrada com as taxas de hanseníase possivelmente reflete condições de desigualdade social, mas também condições clínicas relacionadas com padrões diferenciais de resposta imune (Ferreira *et al.*, 2019).

Com relação à variável V01 – Média de moradores por domicílio, dentre as cinco de maior importância, tanto para as classes de risco de dengue (algoritmo *RF*) quanto para as classes de risco de hanseníase (algoritmos *RF* e *XGBoost*), estudos realizados no município de Niterói – RJ e para o estado de Sergipe, identificaram associações entre a densidade de moradores por domicílio e a taxa anual de incidência de dengue (Araújo *et al.*, 2020; Resendes *et al*., 2010). Áreas com menores condições de infraestrutura, por vezes, estão associadas a incrementos populacionais significativos, de forma que, ao mesmo tempo que condições ambientais favorecem a disseminação do vetor transmissor da doença, mais pessoas ficam suceptíveis a adquiri-la (Resendes *et al*., 2010).

Relativamente à hanseníase e à média de moradores por domicílio, a condição de transmissibilidade pode ser facilitada pela densidade habitacional (Brasil, 2002). Num estudo ecológico descritivo realizado para os 27 estados brasileiros, chegou-se à conclusão que a incidência de hanseníase apresenta tendência de aumento proporcional em domicílios com densidades habitacionais mais elevadas (Castro *et al.*, 2016). Esta condição está diretamente relacionada com a sua forma de transmissibilidade pelas vias respiratórias (Brasil, 2002).

A condição de renda observada no *ranking* de importância das variáveis para o algoritmo *C5.0* e as classes de risco de hanseníase suscitam a indispensabilidade de uma avaliação mais aprofundada dos resultados. Dadas as condições a princípio contraditórias, se considerado que são investigadas relações diretas e positivas entre condições de baixa renda, analfabetismo e condições de saneamento básico inadequado. Em parte, é preciso ter em mente, nesse caso, que cada algoritmo utiliza critérios distintos para avaliar as variáveis mais importantes (Kuhn *et al.*, 2021), o que pode ser uma das explicações para esse caso e deverá ser futuramente examinado.

Por fim, é importante ressaltar que neste trabalho avaliou-se especificamente variáveis socioeconómicas e a sua relação com as classes de risco para dengue, hanseníase e leishmaniose tegumentar. No entanto, sabe-se que a transmissibilidade da dengue e da leishmaniose tegumentar podem estar relacionadas com condições físicas-geográficas como temperatura e precipitação no caso da dengue (Mussumeci & Coelho, 2020; Souza *et al.* 2010), e mudanças de uso e cobertura do solo no caso da leishmaniose tegumentar (Negrão & Ferreira, 2013), de forma que tais condições poderão também ser avaliadas em trabalhos futuros.

V. CONSIDERAÇÕES FINAIS

O uso dos algoritmos de *machine learning* no estudo da predição de classes de risco entre três doenças negligenciadas no Brasil e variáveis socioeconómicas mostrou-se parcialmente eficiente. No modelo proposto neste artigo, valores de exatidão abaixo de 0,4 apontaram a necessidade de reavaliação do método para trabalhos respeitantes à leishmaniose tegumentar. No que diz respeito às classes de risco de dengue e de hanseníase, os resultados de exatidão exibiram valores acima 0,6.

Quanto às variáveis de maior importância, em partes, observaram-se variáveis similares para dengue e hanseníase, a destacar a condição étnico-racial, média de moradores por habitação e alfabetização. Entretanto, condições de renda identificaram estratos diferentes para as classes de risco dessas doenças, indicando condições de baixa renda como mais importante para dengue, ao contrário do observado nas classes de risco para hanseníase, em que o estrato de renda foi superior, de três a cinco salários-mínimos. Assim, os resultados confirmam, parcialmente, a hipótese de referência inicial desta pesquisa.

Em razão dos resultados promissores encontrados para as classes de risco de dengue e de hanseníase com base nos indicadores de exatidão, julga-se imprescindível avançar nas pesquisas relacionadas ao uso dos algoritmos de *machine learning*, aprofundando-se em referenciais relativos às variáveis preditivas para as doenças negligenciadas e realizando novos testes, que contemplem outras variáveis ou perspectivas específicas de vulnerabilidade em saúde, o que poderá gerar resultados mais positivos também para as classes de risco de leishmaniose tegumentar.

AGRADECIMENTOS

O artigo resulta de pesquisas referentes a elaboração de tese de doutorado da primeira autora no Programa de Doutorado do Instituto de Estudos Socioambientais - IESA da Universidade Federal de Goiás - UFG, financiado pelo CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil.

ORCID iD

Thamy Barbara Gioia https://orcid.org/0000-0001-6431-6096
Juliana Ramalho Barros https://orcid.org/0000-0002-9264-2785
Renato Rodrigues da Silva https://orcid.org/0000-0002-1934-8141

CONTRIBUTOS DOS/AS AUTORES/AS

Thamy Barbara Gioia: Conceptualização; Metodologia; Software; Validação; Análise formal; Investigação; Escrita – preparação do esboço original; Redação – revisão e edição; Validação. **Juliana Ramalho Barros**: Conceptualização; Metodologia; Validação; Análise formal; Redação – revisão e edição; Visualização; Supervisão. **Renato Rodrigues da Silva**: Conceptualização; Metodologia; Software; Redação – revisão e edição; Visualização; Supervisão.

REFERÊNCIAS BIBLIOGRÁFICAS

Araújo, D. C., Santos, A. D., Lima, S. V. M. A., Vaez, A. C., Cunha, J. O., & Araújo, K. C. G. M. (2020). Determining the association between dengue and social inequality factors in north-eastern Brazil: A spatial modelling. *Geospatial Health*, *15*(1), 854. https://doi.org/10.4081/gh.2020.854

Barata, R. B. (2009). *Como e por que as desigualdades sociais fazem mal à saúde* [How and why social inequalities are bad for health]. Fiocruz.

Brasil. (2002). *Guia para controle da hanseníase* [Guide to leprosy control]. (Série A. Normas e Manuais Técnicos; n. 111). Ministério da Saúde. Secretaria de Políticas de Saúde. Departamento de Atenção Básica. https://bvsms.saude.gov.br/bvs/publicacoes/guia_de_hanseniase.pdf

Brasil. (2015). *Guia de Vigilância Epidemiológica* [Guide to epidemiological surveillance]. (Série A. Normas e Manuais Técnicos). Ministério da Saúde. Secretaria de Políticas de Saúde. Departamento de Atenção Básica.

Brasil. (2017). *Manual de Vigilância da Leishmaniose Tegumentar* [Manual to Tegumentary Leishmaniasis surveillance]. Ministério da Saúde.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-32. https://doi.org/10.1023/A:1010933404324

Castro, S. S., Santos, J. P. P., Abreu, G. B., Oliveira, V. R., & Fernandes, L. F. R. M. (2016). Leprosy incidence, characterization of cases and correlation with household and cases variables of the Brazilian states in 2010. *Anais Brasileiros de Dermatologia*, *91*(1), 28-33. https://doi.org/10.1590%2Fabd1806-4841.20164360

Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Association for Computing Machinery Proceedings* (pp. 785-794)[Proceedings]. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, august 2016, New York, USA. https://doi.org/10.1145/2939672.2939785

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404-413. https://doi.org/10.2307/2331986

Departamento de Informática do SUS. (n.d.). *Taxa de incidência de hanseníase – D.2.6* [Leprosy incidence rate – D.2.6]. http://tabnet.datasus.gov.br/tabdata/LivroIDB/2edrev/d0206.pdf

Espinosa-Zuniga, J. J. (2020). Aplicacion de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito [Application of Random Forest and XGBoost algorithms based on a credit card applications database]. *Ingeniería, Investigacion y Tecnologia*, *21*(3), 1-16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Ferreira, A. F., Souza, E. A., Lima, M. S., García, G. S. M., Corona, F., Andrade, E. S. N., … Ramos, A. N. Jr. (2019). Mortalidade por hanseníase em contextos de alta endemicidade: análise espaço-temporal integrada no Brasil [Mortality from leprosy in highly endemic contexts: integrated temporal-spatial analysis in Brazil]. *Pan American Journal of Public Health, 43,* e87. https://doi.org/10.26633/RPSP.2019.87

Gerência de Epidemiologia e Informação. (2003). *Índice de vulnerabilidade à saúde* [Health Vulnerability Index]. Prefeitura Municipal de Belo Horizonte. http://www.pbh.gov.br/smsa/biblioteca/gabinete/risco2003

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (*2nd ed.). O'Reilly.

Honorato, T., Lapa, P. P. A., Sales, C. M. M., Reis-Santos, B., Tristão-Sá, R., Bertolde, A. I., & Maciel, E. L. N. (2014). Análise espacial do risco de dengue no Espírito Santo, Brasil, 2010: uso de modelação completamente bayesiana [Spatial analysis of distribution of dengue cases in Espírito Santo, Brazil, in 2010: use of Bayesian model]. *Revista Brasileira de Epidemiologia, 17*(2), 150-159. https://doi.org/10.1590/1809-4503201400060013

Instituto Brasileiro de Geografia e Estatística. (2019). *Downloads de bases cartográficas. Geociências* [Geoscience]. https://www.ibge.gov.br/geociencias/downloads-geociencias.html

Instituto Brasileiro de Geografia e Estatística. (n.d.). *Sidra. Geociências* [Geoscience]. https://sidra.ibge.gov.br/home/pms/brasil

Instituto de Pesquisa Econômica Aplicada. (2015). *Atlas da vulnerabilidade social nos municípios brasileiros* [Atlas of social vulnerability in Brazilian municipalities]. IPEA. http://ivs.ipea.gov.br/images/publicacoes/Ivs/publicacao_atlas_ivs.pdf

Kuhn, M., & Johnson, K. (2016). *Applied Predictive Modeling*. Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Hunt, T. (2021). *Caret: classification and regression training. R package (version 6.0-90).* CRAN.

Monteiro, L. D., Mota, R. M., Martins-Melo, F. R., Alencar, C. H., & Heukelbach, J. (2017). Determinantes sociais da hanseníase em um estado hiperendêmico da região Norte do Brasil [Social determinants of leprosy in a hyperendemic State in North Brazil]. *Revista de Saúde Pública, 51*, 70. https://doi.org/10.1590/S1518-8787.2017051006655

Mussumeci, E., & Coelho, F. C. (2020). Large-scale multivariate forecasting models for Dengue – LSTM versus random forest regression. *Spatial and Spatio-temporal Epidemiology, 35*, 100372. https://doi.org/10.1016/j.sste.2020.100372

Negrão, G. N., & Ferreira, M. E. M. C. (2013). Circuitos Espaciais da leishmaniose tegumentar americana no estado do Paraná [Spatial circuits of American Tegumentary Leishmaniasis in the state of Paraná]. *Hygeia – Revista Brasileira de Geografia Médica e da Saúde*, *9*(17), 74-94. https://doi.org/10.14393/Hygeia923164

Quinlan, R. (1993). C5.0: an informal tutorial. Rulequest. https://www.rulequest.com/see5-unix.html

Resendes, A. P. C, Silveira, N. A. P. R., Sabroza, P. C., & Souza-Santos, R. (2010). Determinação de áreas prioritárias para ações de controle da dengue [Determination of priority areas for dengue control actions]. *Revista de Saúde Pública*, *44*(2), 274-82. https://doi.org/10.1590/S0034-89102010000200007

Santos, H. G., Zampieri, F. G., Normilio-Silva, K., Silva, G. T., Lima, A. C. P., Cavalcanti, A. B., & Chiavegatto, A. D. P. F. (2020). Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer. *Journal of Critical Care, 55,* 73-78. https://doi.org/10.1016/j.jcrc.2019.10.015

Silva, E. T. C., Olinda, R. A., Pacha, A. S., Costa, A. O., Brito, A. L., & Pedraza, D. F. (2020). Análise espacial da distribuição dos casos de dengue e sua relação com fatores socioambientais no estado da Paraíba, Brasil, 2007-2016 [Spatial analysis of the distribution of dengue cases and its relationship with socio-environmental factors in the state of Paraíba, Brazil, 2007-2016]. *Saúde em Debate, 44*(125), 465-477. https://doi.org/10.1590/0103-1104202012514

Sistema de Informação de Agravos de Notificação. (2018). *Dados epidemiológicos Sinan* [Sinan epidemiological data]. http://portalsinan.saude.gov.br/dados-epidemiologicos-sinan

Sistema de Informação de Agravos de Notificação. (n.d). *Dados epidemiológicos Sinan* [Sinan epidemiological data]. http://portalsinan.saude.gov.br/dados-epidemiologicos-sinan

Souza, C. M. N., Costa, A. M., Moraes, L. R. S., & Freitas, C. M. (2015). *Saneamento: promoção da saúde, qualidade de vida e sustentabilidade ambiental* [Sanitation:

health promotion, quality of life and environmental sustainability]. Fiocruz.

Souza, S. S. de, Silva, I. G., & Silva, H. H. G. (2010). Associação entre incidência de dengue, pluviosidade e densidade larvária de Aedes aegypti, no Estado de Goiás [Association between dengue incidence, rainfall and larval density of Aedes aegypti, in the State of Goiás]. *Revista da Sociedade Brasileira de Medicina Tropical*, *43*(2),152-155. https://doi.org/10.1590/S0037-86822010000200009

Valle, D. (2021). *Aedes de A a Z* [Aedes from A to Z]. Fiocruz.

World Health Organization. (2019). *Leishmanioses*: *Informe Epidemiológico nas Américas* [Leishmaniasis: Epidemiological Report in the Americas]. Organização Pan-Americana da Saúde. https://iris.paho.org/handle/10665.2/51738

World Health Organization. (2020). *Ending the neglect to attain the Sustainable Development Goals: a road map for neglected tropical diseases 2021–2030.* WHO. https://www.who.int/publications/i/item/9789240010352