CEG
Centro de Estudos Geográficos

Finisterra, *LVII*(121), 2022, pp. 109-123
ISSN: 0430-5027
doi: 10.18055/Finis28635
Artigo

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

Thamy Barbara Gioia[1]
Juliana Ramalho Barros[1]
Renato Rodrigues da Silva[2]

ABSTRACT – 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 cutaneous 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.

Received: 18/11/2022. Accepted: 18/12/2022. Published: 15/12/2022.

[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

RESUMO – FATORES SOCIOECONÔMICOS 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. Analisar a relação entre variáveis socioeconômicas e doenças tropicais negligenciadas pode auxiliar 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 socioeconômicas mais importantes para a classificação de risco de três doenças negligenciadas: hanseníase, leishmaniose tegumentar e dengue. Foram avaliados três algoritmos embasados 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 acurácia 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 acurácia 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 acurácia inferior a 0,4 inviabilizando a avaliação das possíveis variáveis preditivas ao modelo. Os três algoritmos avaliados apresentaram desempenho preditivo aproximados, no entanto, o *RF* foi ligeiramente superior. As variáveis socioeconômicas 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*.

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 sont le plus important 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: *Aléatoire Forêt (AF)*, *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 *AF* et *XGBoost* ont présenté des valeurs de précision supérieures à 0,6. Les deux ressortent comme des variables plus prédictives de 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 précision 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 cours risque de leishmaniose tégumentaire, les algorithmes ont adopté une précision inférieure à 0,4, rendant évaluation des variables prédictives possibles au modèle. Les trois algorithmes évalués performances prédictives approximatives, cependant, le *AF* était supérieur résistant. Les variables les variables socio-économiques les plus importantes pour prédire les classes de risque de dengue et de lèpre étaient similaire.

Mot clés: Maladies tropicales négligées; déterminants sociaux; *XGBoost*; *Random Forest*; *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 precisión 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 precisión 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 precisión 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. INTRODUCTION

Neglected diseases are a diseases group classified as such because they receive low investments in research and drug production, in addition to prevailing in social vulnerability conditions and being more frequent in developing countries (World Health Organization [WHO], 2020). Health social vulnerability can be assessed based on socioeconomic variables, also defined as social determinants of health, which take into account aspects such as income, education, basic sanitation, and housing (Barata, 2009; Souza *et al.*, 2015).

Leprosy, cutaneous leishmaniasis, and dengue are neglected diseases and are prevalent in Brazil. In 2018, 36 766 cases of leprosy were registered in the country, approximately 1.76 cases per 10 000 inhabitants, when WHO recommendations suggest rates below one case per 10 000 inhabitants (WHO, 2020). In the state of Goiás, there were 1791 disease records for the same year, while in the Federal District there were 205 records, placing them, respectively, in 8th and 22nd place in the national ranking (Sistema de Informação de Agravos de Notificação [SINAN], 2018). For dengue, in 2018, 265 460 cases were registered in Brazil, with 91 530 in the state of Goiás and 2444 in the Federal District. In 2018, Goiás ranked 1st in the cases ranking in the country, while the Federal

District ranked 16th (SINAN, 2018). For cutaneous leishmaniasis, in 2018, 17 950 cases of the disease were registered in Brazil, 323 in the State of Goiás, placing it in 12th in the national ranking, and 34 cases in the Federal District, placing it in 23rd (SINAN, 2018).

Cutaneous leishmaniasis is an infectious and non-transmissible disease caused by different protozoa species of the *Leishmania* genus. It is considered a zoonotic infection. In other words, from a host (wild or domestic animals), the disease is transmitted to humans, in which it manifests itself through cutaneous or mucosa injuries. Its transmission is sometimes related to significant changes in the landscape, such as agricultural activities, mining, slopes and peripheral areas occupation close to second-growth forests expansion (Brasil, 2017).

In the leprosy case, transmission occurs through respiratory tract, from human to human. The disease is caused by the parasite *Mycobacterium leprae*, also known as Hansen's bacillus. Symptoms are characteristically dermatological, with cutaneous and peripheral nerves injuries, which may progress to physical disabilities. In addition to individual conditions, social vulnerability situations related to precarious housing conditions and high housing density may favor the disease transmissibility (Brasil, 2002).

With respect to dengue, the disease is caused by a *Flavivirus* genus virus. The infection source is the human being, who transmits it through a vector of the *Aedes* genus, of the *Aedes aegypti species*. Among the symptoms of the disease, there are: fever, headache, nausea, vomiting, and abdominal pain (Brasil, 2015). According to Valle (2021), large urban centers usually present higher rates of disease infestation, considering the precarious conditions of urban infrastructure and the natural conditions of temperature and precipitation association.

Among the technical resources applicable to data modeling are *machine learning* algorithms, which technically process input data in order to predict classification and/or regression results (Géron, 2019). In health, algorithms have been used in an effort to predict potential variables in diseases diagnosis, death evolution, and vulnerability contexts (Santos *et al*., 2020). Another significant *machine learning* algorithms aspect is the ability to assess the relative relevance of each variable when compared to others in a prediction (Géron, 2019).

Therefore, the aim of this article was to evaluate, based on *machine learning* algorithms, which are the most important socioeconomic variables for the risk classification of three neglected diseases: leprosy, cutaneous leishmaniasis, and dengue. These diseases were selected by deeming the neglected diseases list of the World Health Organization (WHO, 2020) and taking into account the prevalence rates observed in the municipalities of the State of Goiás and of the Federal District, this work's study area and the periods with official data available (SINAN, 2018).

Finally, considering that cutaneous leishmaniasis, dengue and leprosy are classified as neglected diseases and prevail in social vulnerability conditions (WHO, 2020), it was decided to evaluate only the socioeconomic conditions of the study areas. The hypothesis is that the analyzed areas will reveal similarity about the most relevant socioeconomic variables for risk classes prediction for the three analyzed diseases.

II. METHODOLOGY

As a study area, the 246 municipalities of Goiás and of the Federal District were delimited, both located in Brazil Midwest (fig. 1).

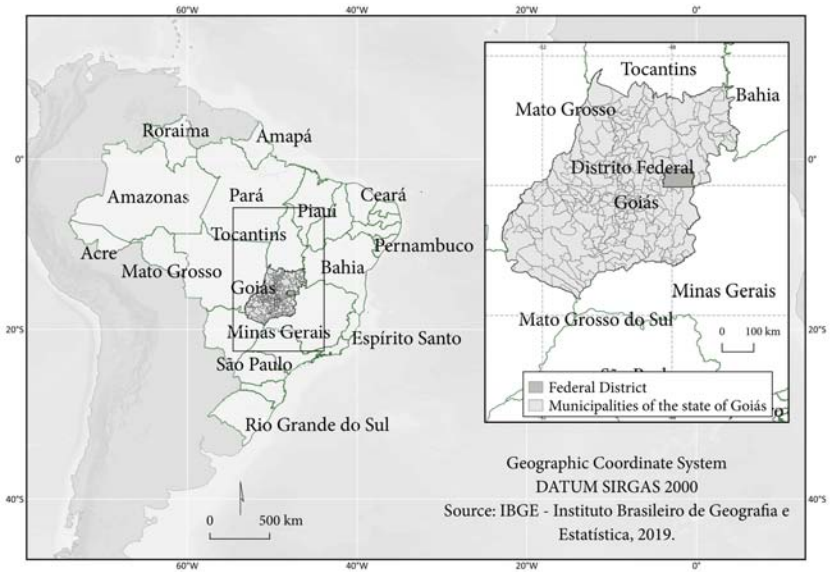

Fig. 1 – Goiás and Federal District location.

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

To estimate the prevalence rates, records of leprosy, cutaneous leishmaniasis, and dengue cases, besides population projections for the period from 2001 to 2018, available on SINAN (2018), were used.

The socioeconomic variables used for risk assessment were obtained from the Sistema de Recuperação Automática (SIDRA) [Automatic Recovery System] of the Instituto Brasileiro de Geografia e Estatística (IBGE) [Brazilian Institute of Geography and Statistics], referring to the demographic census of the years 2000 and 2010.

Finally, the predictive socioeconomic variables assessed followed the studies produced by the Gerência de Epidemiologia e Informação (GEEPI) [Epidemiology and Information Management] in the city of Belo Horizonte – MG (2003), and are illustrated in table I.

1. Fee pre-processing

The prevalence rates of the three analyzed diseases were calculated from the available data average for two periods: 2001-2009 and 2010-2018, in order to exclude possible random fluctuations in the records and to make them compatible with the official socioeconomic data available regarding the Brazilian demographic census in the years 2000 and 2010.

Table I – Independent variables used to modeling the algorithms.

Quadro I – Variáveis independentes utilizadas para modelação dos algoritmos.

Code	Description
V01	Inhabitant average per home
V02	Population over 10 years old with income class up to 1 minimum wage (%)
V03	Population over 10 years old with income class of more than 1 to 2 minimum wages (%)
V04	Population over 10 years old with income class of more than 2 to 3 minimum wages (%)
V05	Population over 10 years old with income class of more than 3 to 5 minimum wages (%)
V06	Population over 10 years old with income class of more than 5 to 10 minimum wages (%)
V07	Population over 10 years old with income class of more than 10 to 20 minimum wages (%)
V08	Population over 10 years old with income class above 20 minimum wages (%)
V09	Population over 10 years without income (%)
V10	White self-declared population (%)
V11	Black self-declared population (%)
V12	Asian self-declared population (%)
V13	Brown-skinned self-declared population (%)
V14	Indigenous self-declared population (%)
V15	Housing with general sewerage or drainage network (%)
V16	Housing with septic tanks (%)
V17	Housing with inadequate sewage system (%)
V18	Housing with waste collection public service (%)
V19	Housing with inadequate waste disposal (%)
V20	Housing with public water supply (%)
V21	Housing with water supply from a well on the property (%)
V22	Housing with inadequate water supply (%)
V23	Literate population (%)
V24	Illiterate population (%)

Subsequently, the rates were standardized per 100 000 inhabitants and classified into risk categories for each neglected disease according to guidance documents from the Brazilian Ministry of Health and the World Health Organization (Brasil, 2015; Departamento de Informática do SUS [DATASUS], (n.d); WHO, 2019) (table II).

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

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

Classes	Dengue*	Leprosy*	Cutaneous leishmaniasis*
Low	Up to 100	Less than 2	Up to 0,95
Medium	101-299	2-9,99	0,96-4,94
High	300-599	10-19,99	4,95-12,69
Very High	600-799	20-39,99	12,70-26,71
Hyperendemic	Above 800	Higher/Equal to 40	26,72-46,50

*Cases per 100 000 inhabitants.

Source: Dengue (Brasil, 2015); Leprosy (DATASUS, n.d); Cutaneous leishmaniasis (WHO, 2019). Adapted by the author

Data were organized in a spreadsheet, containing the average prevalence rate classified by risk for the three analyzed diseases and the socioeconomic variables in table I, concerning the years 2000 and 2010. For the rates for the 2001 to 2009 period, socioeconomic data from 2000 were evaluated, and for the period from 2010 to 2018 the variables from 2010 were evaluated.

2. *Machine learning* algorithms

For the prediction of risk classes of the three analyzed diseases, three *machine learning* algorithms for supervised classification based on decision trees were used: *Random Forest* (Breiman, 2001), *XGBoost* (Chen & Guestrin, 2016) and *C5.0* (Quinlan, 1993). In classification, a decision tree works as a set of hierarchical rules for variables division (called node) into subsets through rules (called branches) until a subset is obtained, which is homogeneous enough to be classified as the same class, thus obtaining a terminal node (called leaf) (fig. 2).

The three used algorithms address strategies that generate a series of decision trees, which allow a more robust modeling than a single tree production. What differentiates the algorithms are the characteristics relevant to the models training mode and their operational characteristics (Breiman, 2001; Chen & Guestrin, 2016).

For the *RF*, the final decision for classification consists of combining trees created independently, in which each tree is adjusted from an attribute vector sampled from the bootstrap method. This algorithm hyperparameters are: trees number (ntree) to be created, and variables number (mtry) tested (Breiman, 2001). In this study, the *ntree* value equal to 500 and the *mtry* to the total number square root of input variables ($\sqrt{N}$) were used as parameters, besides standard conditions of the statistical package *R*.

The *C5.0* algorithm is a type of decision tree developed from recursive data partitioning. To improve the classification performance, the boosting method is used, which consists of sequentially adjusting several algorithm trials, assigning weights to observations that were incorrectly classified. The optimized hyperparameter was the number of trials, which in this study was defined equal to 20 (Kuhn & Johnson, 2016).

XGBoost works from sequential trees to arrive at the classification results, in an approach known as *Gradient Boosting*. Each tree is sequentially built, considering and correcting the previous tree error, that is, the initial tree will be associated with a residual error, and the next tree will be built with an adjustment to the residual error of the previous step. The previous results will be combined to develop a new tree, in which the medium quadratic error root will be smaller than the predecessor. The process is continuous until the smallest residual error is obtained (Chen & Guestrin, 2016; Espinosa-Zuniga, 2020).

Modeling was performed by using the free software *R* version 4.0.3, through the *Classification and Regression Training – CARET package* (Kuhn *et al.*, 2021).

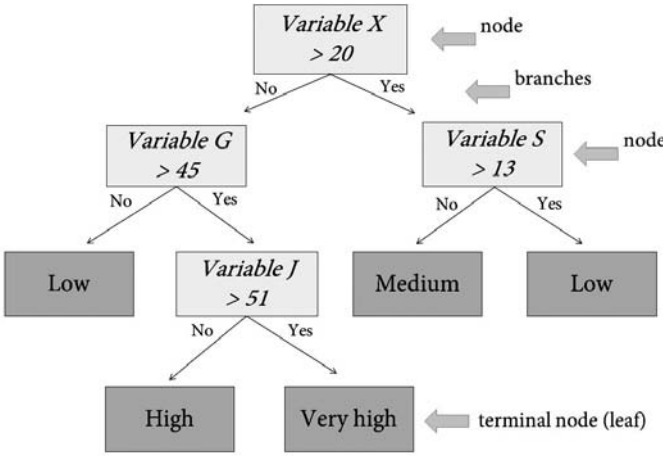

Fig. 2 – Example of decision tree with hypothetical variables and values.
Fig. 2 – Exemplo de árvore de decisão com variáveis e valores hipotéticos.

For modeling training, 70% of the data were used as a training sample and calibrated with *k-fold* cross validation for k=5. The remaining 30% of samples were separated for testing, aiming to validate the algorithms with a dataset independent of the one used in training. The validation metric used was accuracy, with a 95% confidence interval based on the *Exact Binomial* test (Clopper & Pearson, 1934).

Finally, it was examined the most important variables in the prediction process of the three neglected diseases analyzed risk classes. Each algorithm has different metrics to order the variables importance (Kuhn *et al*., 2021). All metrics have the principle of attributing greater weight to the variables that are in the upper nodes of the decision tree. As this variables importance ordering allows the variables normalization between zero and 100, a relative comparison between algorithms is possible.

III. RESULTS

Table III shows the results for the training (cross validation) and test samples, with the accuracy confidence interval for the three algorithms and for the three classified prevalence rates categorization.

From the results, it is possible to note that:

– For dengue risk classes, the *Random Forest (RF)* and *XGBoost* algorithms revealed slightly better performance compared to the *C5.0* algorithm. The accuracy results (test data) for the *RF* algorithm were 0.58, with a confidence interval of 0.50-0.66; 0.56 for the *XGBoost* algorithm, with a confidence interval of 0.48-0.65; and 0.46 for the *C5.0* algorithm, with a confidence interval of 0.38-0.56. Therefore, both the *RF* algorithm and the *XGBoost* allowed explaining about 60% of the dengue risk classes data variance, based on the predictive variables selected for modeling;

– For risk classes of leprosy rates, the three algorithms had test data accuracy results in confidence intervals above 0.6. It should be noted that the *RF* and *XGBoost* algorithms allowed explaining about 70% of the leprosy risk classes data variance;
– For cutaneous leishmaniasis risk classes, the algorithms presented accuracy (test data) below 0.4. Thus, it was decided to proceed with the analyzes referring to the predictive variables importance only for the dengue and leprosy risk classes.

Table III – Accuracy results for the evaluated algorithms.

Quadro III – Resultados de acurácia para os algoritmos avaliados.

Algorithms	*RF*			*XGBOOST*			*C5.0*		
	TDC*	THC	TLC	TDC	THC	TLC	TDC	THC	TLC
Training data accuracy	0,53	0,64	0,32	0,52	0,63	0,33	0,49	0,62	0,29
Test data accuracy	**0,58**	**0,66**	**0,28**	**0,56**	**0,64**	**0,26**	**0,46**	**0,59**	0,27
Test data confidence interval (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 – Dengue prevalence rate classes; THC – Leprosy prevalence rate classes; TLC – Classified cutaneous leishmaniasis prevalence rate classes.

For dengue risk classes, it is verified that, considering the first five most important variables, both the *RF* algorithm and the *XGBoost* identified the variable V09 – Population over ten years without income (%) as the most important, followed by variables V03 – Population over ten years old with income class of more than one to two minimum wages (%), V12 – Asian self-declared population (%), and V23 – Literate population (%) (fig. 3). For the *RF* algorithm modeling, the variable V01 – Inhabitant average per home is also emphasized, and for the *XGBoost* algorithm the variable V13 – Brown-skinned self-declared population (%) is highlighted.

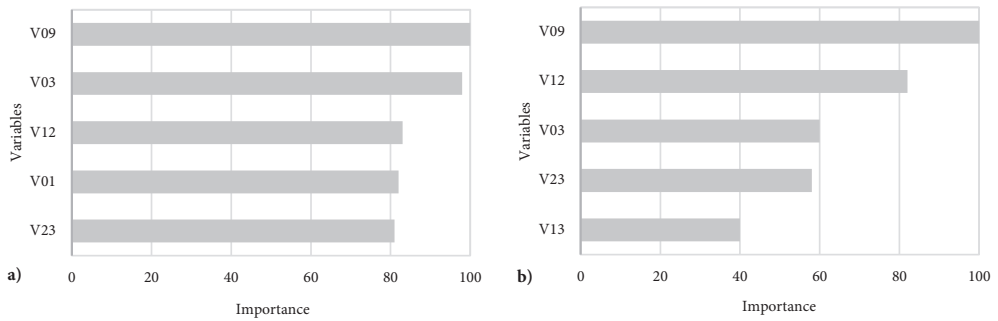

Fig. 3 – Importance of predictive variables for Dengue risk classes; a) Modeling from the *Random Forest* algorithm; and b) Modeling from the *XGBoost* algorithm.

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.

The results related to the leprosy risk classes indicated as the most important predictive variables V23 – Literate population (%) and V13 – Brown-skinned self-declared population (%), which appear among the first five importance variables for the three

algorithms (fig. 4). The variable V01 – Inhabitant average per home is identified as very important both for the *RF* algorithm and for *XGBoost* (figs. 4a and 4b). Variables V22 – Housing with inadequate water supply (%) and V05 – Population over ten years old with income class of more than three to five minimum wages (%) are flagged as of greater importance for algorithm *C5.0*, and, finally, the variable V11 – Black self-declared population (%) appears among the first five importance variables for the *RF* and *C5.0* algorithms (figs. 4a and 4c).

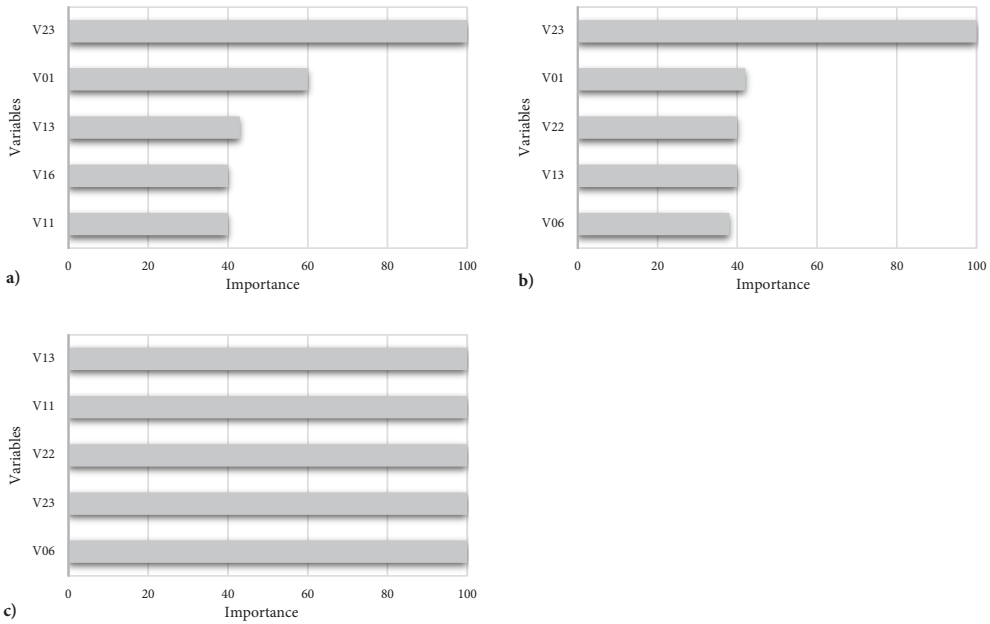

Fig. 4 – Importance of predictive variables for leprosy risk classes;
a) Modeling from the *Random Forest* algorithm; b) Modeling from the *XGBoost* algorithm;
and c) Modeling from the *C5.0* algorithm.

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.

IV. DISCUSSION

Municipalities in the state of Goiás and in the Federal District showed different income conditions for dengue and leprosy risk classes prediction. Low wage conditions (% of population with no income and one to two minimum wages) were important variables for dengue risk classes prediction, unlike what was observed for the leprosy risk classes, in which the higher income condition – three to five minimum wages – was identified as predictive for modeling in the *C5.0* algorithm.

Low-income conditions associated with dengue were found in studies carried out by Honorato *et al*. (2014) in the state of Espírito Santo, where variables such as population with income below three wages and poor solid waste collection presented the best performance, what the author called the spatial effect model. However, it is verified that this condition is not standard. In a study conducted in the state of Paraíba, high incidences of dengue were found in areas with better income conditions (Silva *et al*., 2020).

For leprosy, variables related to lower income conditions were not the most relevant for the prediction, but the literate population percentage appears with greater importance among the variables for the three analyzed algorithms, followed by inadequate water supply conditions and brown-skinned and black population percentage, modeled with the *C5.0* algorithm. It should be noted that negative basic sanitation conditions are often linked to negative income and schooling conditions, if social vulnerability contexts are considered (Gerência de Epidemiologia e Informação, 2003; Instituto de Pesquisa Econômica Aplicada, 2015; WHO, 2020), so such results should be further explored.

Inadequate basic sanitation conditions associated with leprosy incidence rates were verified in a study carried out by Monteiro *et al*. (2017) in the state of Tocantis from 2001 to 2012. Housings with a lower access proportion to piped water and the presence of a bathroom (<61.7%) had an IRR (Incidence Rate Ratio) of 0.627. For the waste collection variable, the IRR was 1 for housings with a collection proportion of less than 88.8%.

Regarding the age and ethnic-racial condition, a survey carried out in the North and Northeast regions of Brazil detected an increased risk of mortality for male leprosy patients aged over 60 years and of brown-skinned and black race-color. According to the authors, considering the leprosy chronic nature, the higher risk of mortality in the age group above 60 years may indicate low quality of life. About brown-skinned race-color, the association found with leprosy rates possibly reflects social inequality conditions, but also clinical conditions related to differential patterns of immune response (Ferreira *et al*., 2019).

Concerning the variable V01 – Inhabitant average per home, among the five most important both for dengue (*RF* algorithm) and leprosy risk classes (*RF* and *XGBoost* algorithm), studies carried out in the municipality of Niterói – RJ and in the state of Sergipe identified associations between the residents density per household and the annual dengue incidence rate (Araújo *et al*., 2020; Resendes *et al*., 2010). Areas with lower infrastructure conditions are sometimes associated with significant population increases, so that, simultaneously, environmental conditions favor the transmissible vector dissemination, and more people are susceptible to acquiring it (Resendes *et al*., 2010).

Regarding leprosy and the inhabitant average per home, the transmissibility condition can be facilitated by housing density (Brasil, 2002). In a descriptive ecological study carried out for the 27 Brazilian states, it was concluded that the leprosy incidence tends to increase proportionally in domiciles with higher housing densities (Castro *et al*., 2016). This condition is directly related to its transmissibility mode through the airways (Brasil, 2002).

The income condition observed in the variables importance ranking for the *C5.0* algorithm and the leprosy risk classes raise the indispensability of a more in-depth assessment of the results, given the initially contradictory conditions, if one considers that

direct and positive relations are being investigated between low-income conditions, illiteracy and inadequate basic sanitation conditions. Partly, it is necessary to keep in mind, in this case, that each algorithm uses different criteria to evaluate the most important variables (Kuhn *et al*., 2021), which may be one of the explanations for this case and should be examined in the future.

Finally, it is relevant to point out that this study specifically assessed socioeconomic variables and their connection with the risk classes for dengue, leprosy, and cutaneous leishmaniasis. However, it is known that the dengue and cutaneous leishmaniasis transmissibility may be related to physical-geographical conditions, such as temperature and precipitation, for dengue (Mussumeci & Coelho, 2020; Souza *et al*. 2010), in addition to changes in use and ground cover, for cutaneous leishmaniasis (Negrão & Ferreira, 2013), so that such conditions may also be evaluated in future studies.

V. CONCLUSION

The use of *machine learning* algorithms in the study of the risk classes prediction among three neglected diseases in Brazil and socioeconomic variables proved to be partially efficient. In the model proposed in this article, accuracy values below 0.4 point to the need to reassess the method for works related to cutaneous leishmaniasis. Regarding the dengue and leprosy risk classes, the accuracy results showed values above 0.6.

Concerning the most important variables, in parts, similar variables were observed for dengue and leprosy, highlighting ethnic-racial condition, inhabitants average per home and literacy. However, income conditions identified different strata for these diseases risk classes, indicating low-income conditions as more important for dengue, contrary to what was observed in the leprosy risk classes, in which the income stratum was higher, from three to five minimum wages. Thus, the results partially confirm the initial reference hypothesis of this research.

Due to the promising results found for the dengue and leprosy risk classes based on the accuracy indicators, it is considered essential to advance in research related to the use of *machine learning* algorithms, deepening in references related to the predictive variables for the neglected diseases and carrying out new tests, which include other variables or specific perspectives of health vulnerability, which may also generate more positive results for the cutaneous leishmaniasis risk classes.

ACKNOWLEDGMENTS

This article results from research referring to the doctoral dissertation elaboration of the first author in the Doctoral Program of the Instituto de Estudos Socioambientais (IESA) [Socioenvironmental Studies Institute] of the Federal University of Goiás – UFG, funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq) [Brazilian National Council for Scientific and Technological Development].

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

AUTHORS CONTRIBUTIONS

Thamy Barbara Gioia: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Writing –original draft preparation; Writing – review and editing; Validation. **Juliana Ramalho Barros:** Conceptualization; Methodology; Validation; Formal analysis; Writing – review and editing; Visualization; Supervision. **Renato Rodrigues da Silva:** Conceptualization; Methodology; Software; Writing – review and editing; Visualization; Supervision.

REFERENCES

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 Cutaneous 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 Cutaneous leishmaniasisin 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