Infectious Disease Modelling 7 (2022) 262–276

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Feature importance: Opening a soil-transmitted helminth machine learning model via SHAP

Carlos Matias Scavuzzo [a, e, *], Juan Manuel Scavuzzo [a], Micaela Natalia Campero [a, e], Melaku Anegagrie [b, c], Aranzazu Amor Aramendia [b, c], Agustín Benito [c], Victoria Periago [d, e]

[a] Instituto de Altos Estudios Espaciales Mario Gulich, Univesidad Nacional de Córdoba-Comisión Nacional de Actividades Espaciales, Argentina
[b] Fundación Mundo Sano, Madrid, Spain
[c] National Centre for Tropical Medicine, Institute of Health Carlos III, Madrid, Spain
[d] Fundación Mundo Sano, Buenos Aires, Argentina
[e] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 15 June 2021
Received in revised form 19 January 2022
Accepted 29 January 2022
Available online 3 February 2022
Handling Editor: DAIHAI HE

Keywords:
Shap
Shapley
Machine learning
Remote sensing
Hookworm
Ethiopia

ABSTRACT

In the field of landscape epidemiology, the contribution of machine learning (ML) to modeling of epidemiological risk scenarios presents itself as a good alternative. This study aims to break with the "black box" paradigm that underlies the application of automatic learning techniques by using SHAP to determine the contribution of each variable in ML models applied to geospatial health, using the prevalence of hookworms, intestinal parasites, in Ethiopia, where they are widely distributed; the country bears the third-highest burden of hookworm in Sub-Saharan Africa. XGBoost software was used, a very popular ML model, to fit and analyze the data. The Python SHAP library was used to understand the importance in the trained model, of the variables for predictions. The description of the contribution of these variables on a particular prediction was obtained, using different types of plot methods. The results show that the ML models are superior to the classical statistical models; not only demonstrating similar results but also explaining, by using the SHAP package, the influence and interactions between the variables in the generated models. This analysis provides information to help understand the epidemiological problem presented and provides a tool for similar studies.

1. Introduction

Intestinal parasites are a group of cosmopolitan parasites, including protozoa and helminth species, which are both urban and rural populations. In particular, soil transmitted helminths (STH) contaminate soil through eggs/larvae contained in human feces; this is why they are more common among people without access to water, hygiene, and/or basic sanitary

* Corresponding author. Instituto de Altos Estudios Espaciales Mario Gulich, Univesidad Nacional de Córdoba-Comisión Nacional de Actividades Espaciales, Spain.
E-mail address: matiasscavuzzo@fcm.unc.edu.ar (C.M. Scavuzzo).
Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2022.01.004

conditions (Campbell et al., 2014; Clasen et al., 2019; O'Reilly et al., 2008; Strunz et al., 2014). STH are among the most common parasites worldwide; according to the World Health Organization (WHO), 820 million people are infected with roundworms (Ascaris lumbricoides), 460 million with whipworms (Trichuris trichiura), 440 million with hookworms (Ancylostoma duodenale and Necator americanus) and over 600 million with the threadworm Strongyloides stercoralis (Organization, 2020). Thus, STH are on the list of twenty Neglected Tropical Diseases (NTDs) elaborated by the WHO (Organization, 2010).

These parasites are endemic in tropical and subtropical regions of the world; in particular, they are widely distributed in Sub-Saharan Africa (SSA), being Ethiopia the country with the third-highest burden of hookworms in the region (Abera et al., 2013; Amor et al., 2016; Anegagrie et al., 2020; Aramendia et al., 2020; Karagiannis-Voules et al., 2015; Muluneh et al., 2020; Nute et al., 2018). It is estimated that more than 80 percent of Ethiopia's over 107 million inhabitants live in rural areas that are endemic for STH (Amor et al., 2016; Anegagrie et al., 2020). Also, the prevalence of hookworm in the country was estimated at 78.7%. The knowledge of the distribution of STH in specific areas and the identification of the most relevant factors which influence their development and transmission would allow the adaptation of control policies, not only for the detection of the areas most likely to be affected (and consequently the allocation of resources) but also for the promotion of health, to avoid post-treatment reinfection (Anegagrie et al., 2020; Mengitsu et al., 2016).

The landscape epidemiology concept is widely used in the field of epidemiological problems related to environmental conditions (soil, climate, land cover) (Estallo et al., 2016a; Polop et al., 2007; Porcasi et al., 2012; Rotela et al., 2017). In recent years, these environmental conditions have been available and easily extracted from spatial information (satellite images and products) (Estallo et al., 2016b), therefore these satellite products may be useful in many areas, one of them being the construction of models that generate relevant information from environmental data. It is worth remembering that typically these models are developed on the basis of linear statistical or generalized linear approaches, however, the use of ML for problem modeling is innovative and highly promising (Weatherhead et al., 1998). ML is an effective empirical family of methods/algorithms for regression and/or classification of linear and non-linear systems and can involve thousands of variables. It is also ideal for solving problems where although theoretical knowledge is still incomplete, there are some observations available to train the model. ML is handy for a large number of applications in earth sciences and bio-geophysical information extraction algorithms (Azamathulla et al., 2012; Brown et al., 2008; Lary et al., 2009; Madadi et al., 2015; Yi & Prybutok, 1996; Zahabiyoun et al., 2013), but their effective use in applications is relatively new, and its prospects are extensive (Lary et al., 2016; Lundberg & Lee, 2017a; Peña-Barragán et al., 2014). Currently, some of the most widely used ML algorithms are artificial neural networks, support vector machines, decision trees, and random forests (Lary et al., 2016).

In the field of epidemiology, large-scale acquisition of massive field data is not always possible and it is very costly, so the contribution of artificial intelligence like ML, to the modeling of epidemiological risk scenarios is a good alternative (Scavuzzo et al., 2020). These kinds of tools are very useful to generate models that can learn from a small amount of data, so that the effect of predictor variables and their interaction in the model may be understood, enabling the development of a risk scenario that can be better adapted to the above limitations (Scavuzzo et al., 2020).

A search of the bibliographic base Scopus returns more than 4000 publications that include “epidemiology” and ”machine learning”, 311 of them in 2016. Of this total, 45% correspond to the area of “Sciences of the Earth”, 44% to ”Computer Science” and 35% to “Engineering”, being China, the United States, Italy, and India the countries with the highest scientific production in the area (Bose et al., 2016; Jafari Goldarag et al., 2016; Wang et al., 2016). It is remarkable to note that a classification criteria focused in the human health area does not exist; being that the application of spatial epidemiology is known to effectively contribute to a comprehensive approach to health-related problems, helping to identify the most vulnerable communities and to design public policies that respond to their particular needs (Souris, 2019). In addition, it is interesting to note that none of the most important authors that appear in the previous search have worked with Epidemiology, Remote Sensing or ML.

One of the most important objections to the use of ML is that this methodology is visualized as a black box, where we can find good models but not understand how they work. To overcome this difficulty, the use of SHapley Additive exPlanation (SHAP) (Lundberg & Lee, 2017b), represents an important advance in interpreting ML models. This is a state-of-the-art Python library commonly used in the feature engineering step in ML projects. It uses the classic Shapley value of game theory and its extension (Lundberg & Lee, 2017b) to link optimal credit allocation to local interpretation, allowing us to explain the results of ML models. SHAP was developed by Scott Lundberg and Su-In Lee in 2017 and combines several existing methodologies to create an intuitive and theoretically reliable way for explaining model predictions, by showing how estimations change after specific variables are removed. The SHAP value quantifies the magnitude and direction (positive or negative) of the feature's influence on the prediction (Gilbert, 2019; Lundberg & Lee, 2017a; Lundberg et al., 2018, 2020).

The Python SHAP package (https://github.com/slundberg/shap), allows us to calculate SHAP values for a selected model and it has already been widely used (Lundberg & Lee, 2017b; Lundberg et al., 2018, 2020). Recently, a new class charting tool, known as decision diagrams, has been added to the SHAP package. This instrument provides a detailed view of the inner workings of a model, which means that it allows us to understand how models make decisions.

In this study, an automatic learning model was applied to an epidemiological study of hookworm infection in Ethiopia (Anegagrie et al., 2021). The output of the model is the prediction of the number of individuals per house infected with hookworm in three rural villages from the Amhara Region. Additionally, in order to aid in the understanding of hookworm transmission, all the variables that were collected were analyzed using the SHAP package as well, to determine the influence of each variable on the occurrence of hookworm infection. The software tools and procedures used will be made available so

as to allow replicability of the model, with the intention of enhancing the application of these techniques for the community working with similar problems. Therefore, given the necessary passage of hookworm through the soil for its development, the analysis included environmental, soil, and socioeconomic variables; most of these available from satellite products.

Finally, this study aims to break with the "black box" paradigm that underlies the application of automatic ML techniques, using the SHAP package to analyze the universe of variables. By consequence, knowing the contribution of each variable in the final prediction of the models applied to geospatial health problems in particular and landscape epidemiology problems in general.

2. Materials and methods

2.1. Study area and field data

Ethiopia is a Federal Democratic Republic composed of ten regional states and two administrative cities, divided into 95 zones and 839 districts or woredas. The current study was conducted in the Amhara National Regional State (Fig. 1), in the northwest of the country, a subtropical region with a rainy season (June to September) and a dry season (February to May) (Aramendia et al., 2020). The region's capital, Bahir Dar (population approximately 2 50 000), is situated at the southern tip of Lake Tana, the biggest lake in the country. Within a radius of about 30 km from the city center, there is a central urban area and surrounding rural areas (Nute et al., 2018). The area of the study is a rural district located at an altitude of 1900 m above sea level and consists of 9 villages with a population of 11 300 inhabitants (data from the Health Center of the district) (Aramendia et al., 2020). A first study was conducted in the rural kebele of Zenzelema (ZE), which is located about 20 km east of the city of Bahir Dar. This kebele consists of nine small villages or gotts. Three of them were randomly selected for the epidemiological study, Zenzelema (ZE), Mazoria (MA), and Sesaberet (SE).

In all these villages, except for Zenzelema, the houses are far apart from each other and surrounded by crops and forest areas. On the other hand, Zenzelema is a crowded slightly more urbanized area with the houses located adjacent to each other. In terms of the geographical area and water source, Zenzelema is located on a main road and the water source is scarce. Mazoria and Sesaberet can be reached on foot by dirt tracks, while the natural streams and currents are relatively small.

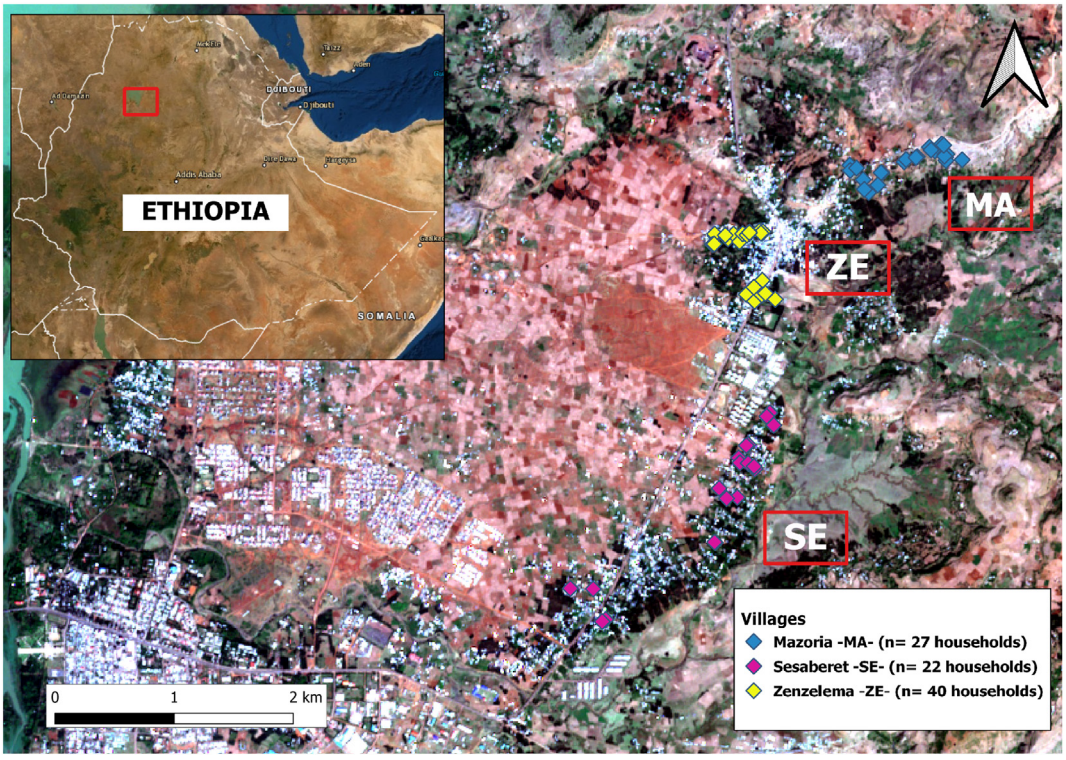

Fig. 1. Villages from Amhara Region (Ethiopia) included in the study. The initials of each village are highlighted in black and white, where MA: Mazoria, ZE: Zenzelema, and SE: Sesaberet. Map data ©2020 Google, base map obtained through QuickMapServices QGIS plugin - QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

Data for this study were collected as part of a larger survey to determine the prevalence of STH, with a specific focus on S. stercoralis at school (Amor et al., 2016), and community levels (Aramendia et al., 2020). Consequently, an area with high hookworm prevalence was found (Table 1). The field study was approved by the Amhara National Regional State Health Bureau Ethics Review Committee Reference number: 1/87/200. All residents over 5 years of age who have lived in the area for at least three months were invited t participate by providing fecal samples for analysis. Information on the characteristics of individuals (children and adults) and families were obtained through standardized WHO surveys that were adapted to the Ethiopian culture and translated into Amharic. All participants were asked to complete an individual personal questionnaire and the head of the household was also asked to complete a household questionnaire (Amor et al., 2016; Aramendia et al., 2020).

Fig. 2 shows the geolocation of infected individuals for each village included in the study, data corresponds to field data collected in previous studies (Aramendia et al., 2020; Amor et al., 2016).

2.2. Variables included in the study

Complete socioeconomic data were available only from the villages of SE, MA, and ZE, thus including 368 individuals and 89 households (Anegagrie et al.). Households were used as data-points to model the risk of infection (number of hookworm infected individuals per household). The variables related to this study involve socioeconomic and environmental parameters, as well as variables related to hookworm infection. It is important to clarify that the dataset used of this study is the same as the one used in a previous study (Anegagrie et al., 2021) and thus the specific data is not replicated herein.

2.3. Modeling

The dataset mentioned above was used to model the risk of infection (number of hookworm infected individuals per household), taking as inputs the value of environmental variables obtained from satellites and the socioeconomic variables that were collected in the field.

Likewise, the cross-validation technique was used to compare the performance of both models; which ran with 5 splits and a ratio of 80-20 for the separation of the data in training and validation. Mean Square Error (MSE) was used as the error metric at cross-validation step, therefore the mean score represents the average of errors from each instance of the dataset.

Soft hyperparameters tuning was used for the selected model by using the Cross-Validation Score to select the best hyperparameter combination. This way, we adjust the "maximum depth", used to control over-fitting, as more depth will allow the model to learn very specific relationships for a particular sample. Increasing this value will make the model more complex and more likely to be over-fit. In addition, for the regularization terms in the weights we adjust the alpha and lambda parameters, considering that the increase of this value will make the model more conservative, all from the booster set. For other settings, default settings were used.

For the analysis of the relative importance of the variables within the model, the Python SHAP library is available (Lundberg et al., 2020). To facilitate the use and integration with SHAP, the model was trained using the Scikit-Learn Wrapper interface provided by XGboost. A baseline XGBoost model was applied to analyze all the variables with the exception of the variable number of people per house, which was erased in order to avoid masking the behavior and interactions of the rest of the variables (environment, soil, and socioeconomic) (Chen & Guestrin, 2016a). Specifically, the contribution of the different variables in the prediction of the model is calculated through the SHAP values, an example is shown in Fig. 3.

To obtain the description of the contribution of the variables on a particular prediction, decision plots, which offer a detailed view of a model's inner workings, were used; these show a large number of feature effects clearly visualized through multi output predictions, displaying the cumulative effect of interactions and exploring feature effects for a range of feature values (Lundberg et al., 2018). In addition to the above-mentioned chart, the data were analyzed through a decision plot that shows how complex models arrive at their predictions (i.e. how models make decisions), therefore, it shows the important features involved in a model's output. A decision plot can be more helpful than a force plot when there are a large number of significant features involved. Furthermore, the data were analyzed using dependence plots that allow having information about the relationships or dependencies between the variables in the context of the developed model. Finally, an absolute summary plot and a summary plot were used to perform a global analysis of the impact of each of the variables on the model's prediction. This can be observed both point by point and in terms of its absolute value at the following site: (https://github.com/juansca/geohelmintos-modeling).

Table 1
Infection status of sampled people from the villages included in this study: Mazoria, Sesaberet and Zenzelema (Region of Amhara, Ethiopia).

	Infected people (n)	Prevalence (perc)	Inf. people per house (mean)
Zenzelema (n = 148)	95	64.2	2.3
Sesaberet (n = 193)	151	78.2	2.7
Mazoria (n = 152)	121	79.6	2.5

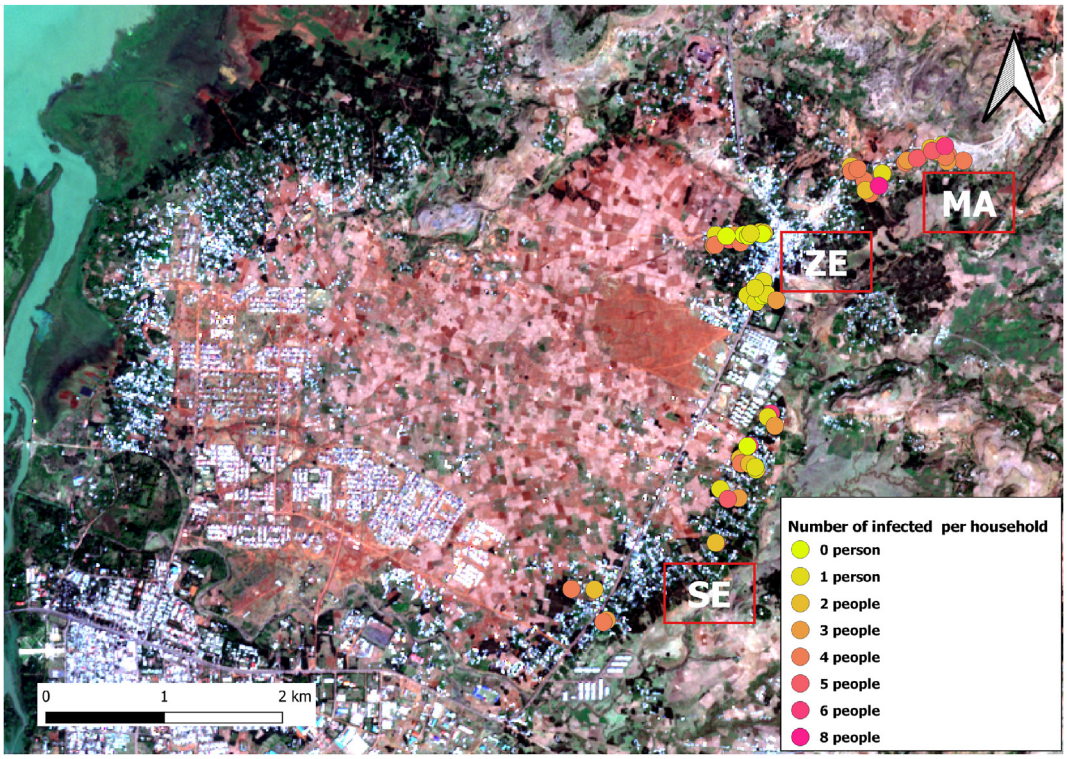

Fig. 2. Spatial distribution of hookworm infection in the villages from Amhara Region (Ethiopia) included in the study. The different colored dots represent the number of infected individuals per household. The initials of each village are highlighted in black and white, where MA: Mazoria, ZE: Zenzelema, and SE: Sesaberet. Map data ©2020 Google, base map obtained through QuickMapServices QGIS plugin - QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

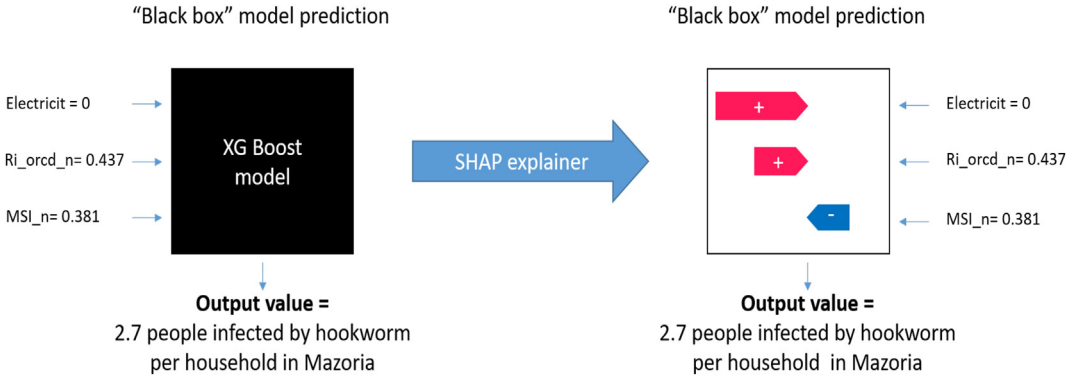

Fig. 3. Diagrammatic representation of SHAP. Figure adapted from Lundberg et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56–67. The values and variables shown in this figure were taken from the force plot presented in Fig. 6.

3. Results

3.1. XGBoost model

Based on the concepts of the classical statistics, the residuals of our initial XGBoost model results were examined. It is important to remember that in classical models a normal and cero centered distribution is expected, which can be expressed with a histogram. In this regard, as a complementary validation of the model's performance to fit the analyzed dataset, a histogram of the residuals of the XGBoost baseline model's output was performed; obtaining as a result a histogram with normal distribution centered at zero.

Also in order to evaluate the aforementioned model, plots were made to show the predictions of the model in comparison to the real infection values, both in the training set and in the test set. It is relevant to note that the model was given only the training target data. In other words, the model did not receive the test's data points before the prediction. In the training routine the model predicts prudently to take care of its performance metrics and minimize the error, so it presents point estimates dispersed in a relatively small variance range around the mean. In this way the model achieves better efficiency in estimating the predicted values in the test routine, in which it is validated with a portion of the data set with which it had no contact in the training routine. Also we observed that the distance between the predicted and actual values decreases considerably, obtaining an acceptable metric.

To determine contribution of the variables Fig. 4 represents decision plot, where a straight vertical line marks the model's base value while the colored line is the prediction. Feature values are printed next to the prediction line for reference. Starting at the bottom of the plot, the prediction line shows how the SHAP values accumulate from the base value to arrive at the model's final score at the top of the plot. On the x-axis at the bottom of the plot, the average prediction of the model is less than 2.6 hookworm infected individuals per house, while on the y-axis, the variables are ordered from highest to lowest depending on their influence on the prediction of the model. According to this decision plot, lack of electricity is the most important variable followed by certain soil characteristics (i.e. Soil Organic Carbon Content - ORCDRC, Bare Soil Index - BSI, etc.). If the house increases the average value of the final prediction of the model, the line is red. On the contrary, if it decreases the average value of the final prediction, the line is blue. In summary, each line represents an observation or house included in the study, and the behavior of the entire set as it descends, ends up throwing the average of the prediction which in total is that of approximately 2.5 hookworm infected individuals per household.

Fig. 5 shows in more detail the multiple interactions of each variable within the proposed model when an average household from each village is taken and analyzed individually. The average household from each village was chosen

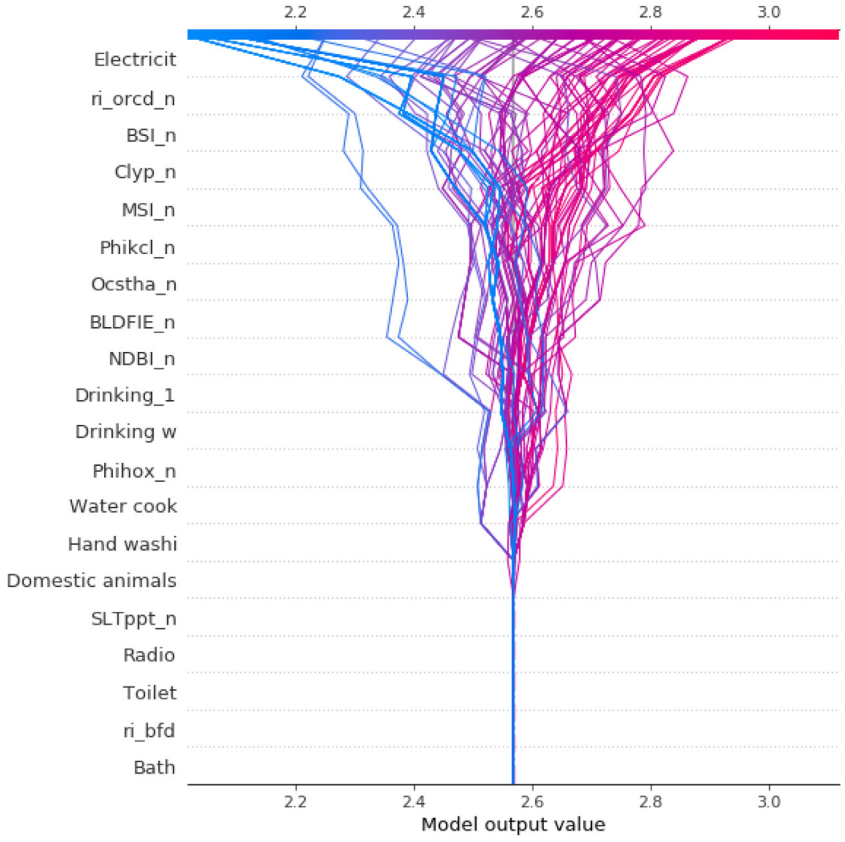

Fig. 4. Decision plot of the complete data set. The y-axis represents the variables used in the study, which refer to: Electricit = Presence or absence of electricity; ri_orcd_n = Constructed risk from soil organic carbon content; BSI_n = Bare Soil Index; Clyp_n = Weight percentage of the clay particles; MSI_n = Moisture Stress Index; Phikcl_n = pH index measured in KCl solution; Ocstha_n = Soil organic carbon stock; BLDFIE_n = Bulk density; NDBI_n = Normalized Difference Builtup; Drinking_1 = Source of drinking water during the rainy season; Drinking w = Source of drinking water during drought periods; Phihox_n = pH index measured in water solution; Water cook = Source of water used for cooking; Hand washi = Hand washing, source of water used for hand washing; Domestic animals = Presence or absence of domestic animals; SLTppt_n = Weight percentage of the silt particles; Radio = Presence or absence of radio; Toilet = Type of bath; ri_bfd = Constructed risk from bulk density; Bath = Origin of water used in the bath. The x-axis represents the number of hookworm infected individuals per household which is the output value of the model.

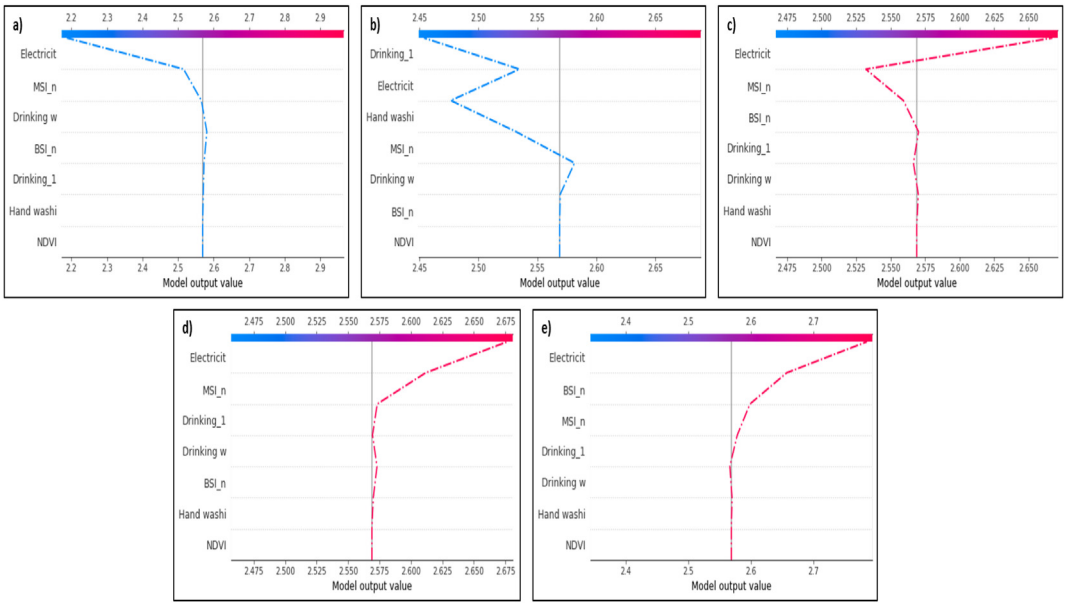

Fig. 5. Decision plot for individual and typical households. These plots refer to the prediction made by the model for each particular observation, thus Fig. 5a refers to an average house in the village of Zenzelema, Fig. 5b refers to an average house in the village of Sesaberet, and Fig. 5c refers to an average house in the village of Mazoria; for the three cases mentioned above the median number of cases per village was taken into account. On the other hand, Fig. 5d and e refer to those houses that had the minimum and the maximum number of hookworm infected persons per house, both of these located in the village of Mazoria.

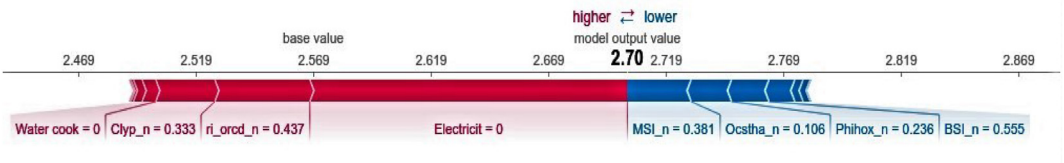

Fig. 6. Force plot of a typical household of Mazoria where the variables that increase the prediction of the model the most are shown in red and those that reduce the prediction of the model the most in blue. The numbers in the black line represent the number of hookworm infected individuals per household according to the base value of the training set (2.57) and the output value of the model's prediction (2.70). Variables: presence or absence of electricity; ri_orcd_n = Constructed risk from soil organic carbon content; Clyp_n = Weight percentage of the clay particles; MSI_n = Moisture Stress Index; Ocstha_n = Soil organic carbon stock; Phihox_n = pH index measured in water solution.

considering the median number of infected individuals per household. For example, in the case of the electricity variable, a contrasting difference can be observed between a house in the village of Zenzelema (Fig. 5a) that is located in the main road, where electricity decreases the model's prediction, and a house in the village of Mazoria (Fig. 5c); a much more remote area, where the same variable increases the model's prediction. Moreover, by isolating the analysis on a single observation or typical house from a particular village and comparing the different villages, the order of importance in the priority of the most influential variables in the average of the model's prediction can vary. For example, for the house observed in Sesaberet (Fig. 5b), the variable of drinking water in the rainy season was taken as the most influential variable in the model instead of the electricity variable. Nonetheless, regardless of the importance of each of the variables, in all these plots, the number of hookworm infected individuals per household remains between 2.5 and 2.6.

In the following force plots, a typical house in each village containing the average number of hookworm-infected individuals per household is represented. Thus, it can be seen in each graph, that for the three cases a base value of 2.5 means that per house there is an average of 2.5 infected individuals; and on the basis of that value the model calculates a different output value for each village. In this type of graphs it can be seen a risk explanation bar that shows red features that push the risk higher (pointing to the right) and blue features that push the risk lower (pointing to the other side), and so depending on how long the bar of each variable is the power of influence it has.

The force plot for each of these houses is shown in Figs. 6 and 7. In Fig. 6, the “base value” for Mazoria (corresponding to the model's average prediction of the training set) is 2.57 hookworm infected individuals per house; while the "output value" (model's prediction) is 2.70. The absence of electricity is also observed as the variable that most increases the prediction of the model due to the bar extension referring to this variable, followed by variables that characterize the organic carbon content (ri_orcd_n) and clay content in the soil (Clyp_n). On the other hand, the moisture stress index (MSI) is the variable that

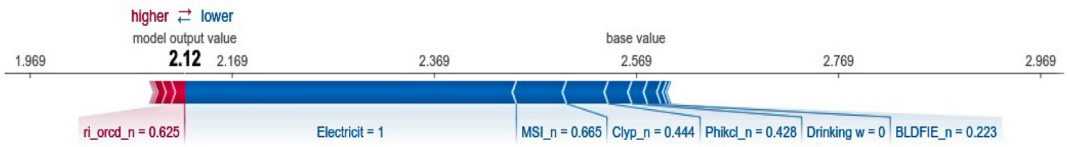

Fig. 7. Force plot of a typical household of Zenzelema where the variables that increase the prediction of the model the most are shown in red and those that reduce the prediction of the model the most in blue. The numbers in the black line represent the number of hookworm infected individuals per household according to the base value of the training set (2.57) and the output value of the model's prediction (2.12). Variables: Electricit = Presence or absence of electricity; Clyp_n = Weight percentage of the clay particles; MSI_n = Moisture Stress Index; Phikcl_n = pH index measured in KCl solution; Drinking w = Source of drinking water during drought periods.

reduces most the model's prediction, followed by the organic carbon stock (Ocstha_n) and the soil pH index measured in water solution (Phihox_n).

Although the force plot is not shown for the village of Sesaberet (due to the similarity of intervening variables), it is relevant to note that the value predicted by the model is 2.40 hookworm-infected individuals per house. Like the village of Mazoria, the absence of electricity is the variable that most increases the prediction of the model, followed by the variables that characterize the organic carbon content (Ocstha_n) and the soil bulk density (BLDFIE). However, in this village, the consumption of water from the pipes during the rainy season (Drinking_1) is the variable that most reduces the prediction of the model, followed by the organic carbon content in the soil (ri_orcd_n), the use of water from pipes for hand washing (Hand washi), water stress index (MSI), clay content (Clyp_n) and soil pH index measured in KCl (potassium chloride) solution (Phikcl_n).

Finally, the predicted values for Zenzelema (Fig. 7) show that the average prediction of the model on the training set is 2.56 hookworm infected individuals per house; while the predicted value of the model is 2.12. Again, the presence of electricity is the variable that reduces most the prediction of the model, followed by variables that characterize water stress (MSI), clay (Clyp_n) and soil ph index measured in KCl solution (Phikcl_n), and finally the consumption of well water for drinking in the dry season (Drinking w). With respect to the characterization of the villages previously carried out, it should be noted that the variables present in Mazoria and Sesaberet, which increase the number of hookworm infected individuals per house, are weighted so low in Zenzelema, that the force plot does not manage to categorize them and therefore they are not considered in the analysis (Fig. 7).

In the dependence plot shown below (Fig. 8), the vertical axis represents the SHAP value (or importance of the variable) while the horizontal axis shows the actual value of the variable. Also, in these graphs, each point is presented with the color palette on the right side of the graph which represents the scale of values of the second variable at each point (not its SHAPs). As observed in Fig. 8, the blue points are those in which the NDVI took low values while the red points are those in which the NDVI took high values. From the same figure, it can be interpreted that when the NDVI is low, the influence of the presence of

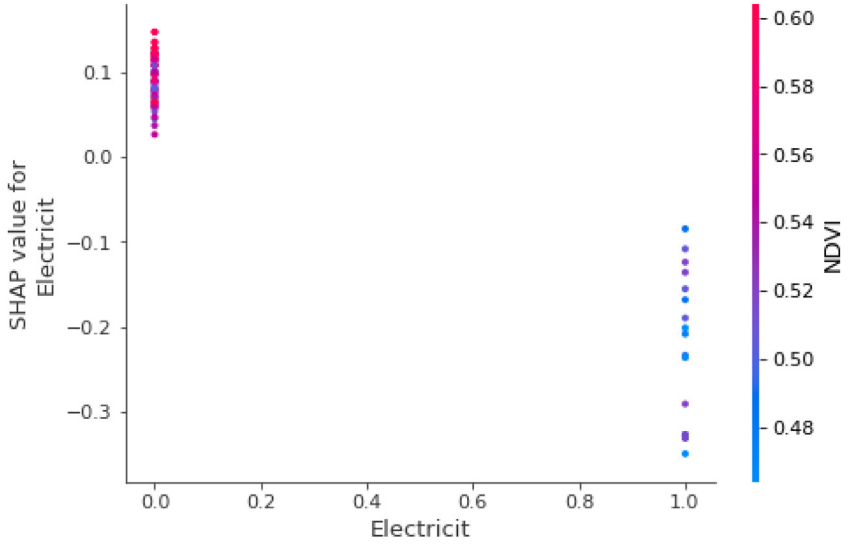

Fig. 8. Dependence plot for SHAP values of electricity and the relation with NDVI. The vertical axis represents the SHAP value (or importance of the variable) while the horizontal axis shows the actual value of the variable. Each point is presented with the color palette detailed for NDVI on the right side of the graph.

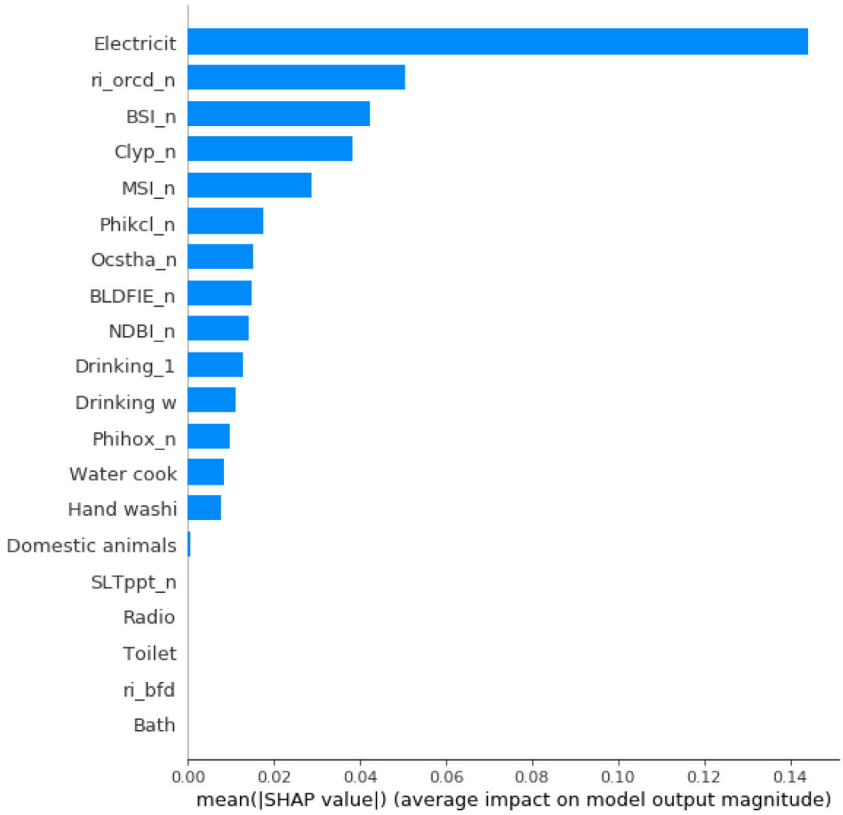

Fig. 9. Absolut summary plot of the complete dataset, where the average absolute value of the SHAP values for each variable is taken in order to obtain a bar chart as a function of the contribution of each variable to the prediction of the model. The variables are ordered from most (top) to least (bottom) important. The y-axis represents the variables used in the study, which refer to: Electricit = Presence or absence of electricity; ri_orcd_n = Constructed risk from soil organic carbon content; BSI_n = Bare Soil Index; Clyp_n = Weight percentage of the clay particles; MSI_n = Moisture Stress Index; Phikcl_n = pH index measured in KCl solution; Ocstha_n = Soil organic carbon stock; BLDFIE_n = Bulk density; NDBI_n = Normalized Difference Builtup; Drinking_1 = Source of drinking water during the rainy season; Drinking w = Source of drinking water during drought periods; Phihox_n = pH index measured in water solution; Water cook = Source of water used for cooking; Hand washi = Hand washing, source of water used for hand washing; Domestic animals = Presence or absence of domestic animals; SLTppt_n = Weight percentage of the silt particles; Radio = Presence or absence of radio; Toilet = Type of bath; ri_bfd = Constructed risk from bulk density; Bath = Origin of water used in the bath. The x-axis represents the number of hookworm infected individuals per household which is the output value of the model.

electricity on the model decreases even more. Otherwise, at high values of the NDVI, the influence of the absence of electricity in the model increases the prediction.

The absolute summary plot is presented in Fig. 9. In this figure, the average absolute value of the SHAP values for each variable is taken in order to obtain a bar chart as a function of the contribution of each variable to the prediction of the model. This way, the relative importance of each variable in contributing to the prediction of the number of hookworm infected individuals per house is observed. From this plot, it can be inferred that the most influential variables in the model were electricity, the amount of soil carbon (n_ocrd_n), and the index of bare soil (BSI), from greater to lesser presence, respectively.

Finally, Fig. 10 presents a summary plot in which the contribution of each variable to the model is displayed taking into account all of the values of each of the variables. This figure includes all the variables entered into the model, where the magnitude is represented by the colored line on the right. The horizontal axis represents the SHAP values of each of the variables by which the model predicts the number of hookworm infected individuals per house. Following the previous plot, the most influential variables in the model are shown. The interpretation of the summary plot shows that the presence of the variable electricity (higher values visible in red on the horizontal bar) implies a decrease in the predicted number of infected individuals per house (expressed on the scale at the bottom of the figure); conversely, lack of electricity (visible in blue) is associated with an increase in the predicted value of the number of hookworm infected individuals per house. The same analysis can be applied to the rest of the variables. Moreover, in the case of the soil variables, the values appear more heterogeneous due to their continuous nature (more violet instead of blue and red) unlike electricity which is a dichotomous variable and polarizes its colors (only blue and red).

It is important to remark here that the aim of this study is not to implement the best predictive model but to understand how the variables interact within the ML models through the SHAP package. In order to achieve this, a baseline XGBoost model was applied and analyzed up to here, including all the variables with the exception of the number of people per house.

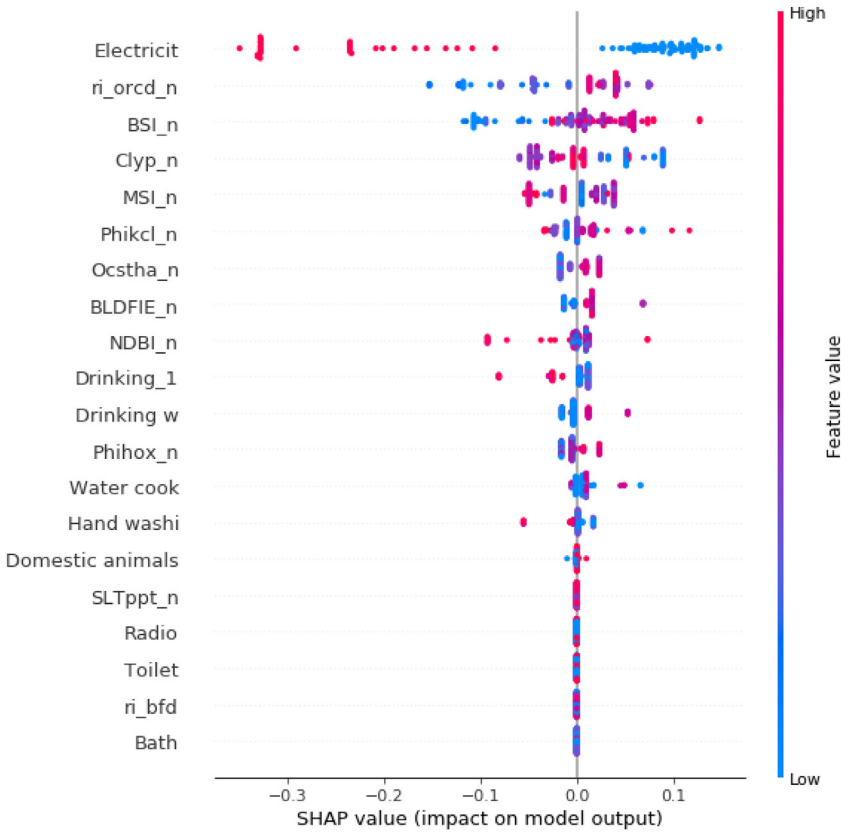

Fig. 10. Summary plot of the complete dataset. The horizontal axis represents the SHAP values of each of the variables by which the model predicts the number of hookworm infected individuals per household while the vertical axis lists all of the variables used in the study. The y-axis represents the variables used in the study, which refer to: Electricit = Presence or absence of electricity; ri_orcd_n = Constructed risk from soil organic carbon content; BSI_n = Bare Soil Index; Clyp_n = Weight percentage of the clay particles; MSI_n = Moisture Stress Index; Phikcl_n = pH index measured in KCl solution; Ocstha_n = Soil organic carbon stock; BLDFIE_n = Bulk density; NDBI_n = Normalized Difference Builtup; Drinking_1 = Source of drinking water during the rainy season; Drinking w = Source of drinking water during drought periods; Phihox_n = pH index measured in water solution; Water cook = Source of water used for cooking; Hand washi = Hand washing, source of water used for hand washing; Domestic animals = Presence or absence of domestic animals; SLTppt_n = Weight percentage of the silt particles; Radio = Presence or absence of radio; Toilet = Type of bath; ri_bfd = Constructed risk from bulk density; Bath = Origin of water used in the bath. The x-axis represents the number of hookworm infected individuals per household which is the output value of the model.

The main reason is that this variable appears as the most important one; which responds to the epidemiology of hookworm infection, as previously documented (Milano et al., 2007; Parija et al., 2017; Chen & Guestrin, 2016b,a; Romero-Sandoval et al., 2017). Therefore, in order to avoid masking the behavior and interactions of the rest of the variables (environment, soil, and socioeconomic), this variable was removed from the analysis for this first version of the model.

The performance metrics, for both training and test data subsets are presented in Table 2: the linear regression model applied by Anegagrie et al. (using the same dataset) and two variants of the XGBoost model; the baseline model, and a final version including only the most important variables that were identified in the baseline analysis (number of people per household, electricity, risk of soil organic carbon content, the index of bare soil and the moisture stress index). The final version of the ML model presents better performance metrics (R^2 and Mean Square Error -MSE-) both for training and testing.

Table 2
Performance metrics of the different models used. For the original data a linear multiple regression model was used (Anegagrie et al., 2021), and then the baseline (which included all the variables except the number of people per house) and final XGBoost model (including the top five variables and the number of people per household) for the current study. In adition: R^2 train refers to R-squared of the training data set, MSE train refers to the Mean Square Error of the test data set, R^2 test refers to R-squared of the training data set, MSE test refers to the Mean Square Error of the test data set.

	R^2 train	MSE train	R^2 test	MSE test
Linear Multiple regression model	0.76	0.68	0.71	0.64
Baseline XGBoost model	0.37	1.72	−0.10	1.94
Final XGBoost model	0.99	0.01	0.79	0.36

R2 train: R-squared of the training data set; MSE train: Mean Square Error of the training data set, R2 test: R-squared of the training data set; MSE test: Mean Square Error of the training data set.

Finally, Fig. 11 shows the true and predicted values of the number of hookworm infected individuals per house when using the final version of the XGboost model.

These last results validate the purpose of the "SHAP" package for an adequate analysis of variables (importance, dependences, etc.), once this process has been carried out by means of the SHAP method, a model containing the most influential variables can be obtained using a small dataset, which present very high-performance metrics in comparison to the linear model originally performed with the field data (Anegagrie et al., 2021).

4. Discussion

It is widely known that disease modeling is usually developed on the basis of linear or generalized linear statistical approaches; however, the use of ML for modeling problems, in general and in this work in particular, is innovative and very promising. As shown herein, ML tools were and are useful for solving problems where limited data are available to train the model, a very common situation in the field of epidemiology. This is why, in a global context of growing needs and limited resources, open data science and ML techniques play a key role in contributing to the generation of research products and the promotion of decision making focused on local health priorities. In this work, a public health problem was addressed using these novel tools in a Region of Ethiopia, where the prevalence of STH infection is on the list of Neglected Tropical Diseases (NTDs) elaborated by the WHO (Organization, 2010); thus evidencing the high potential of approaches with this methodology.

ML tools are increasingly used in many fields to model regression and/or classification of data, with special emphasis on nonlinear systems that involve a high number of variables. In the area of epidemiology, it is essential to optimize resources for data collection in the field. This is why ML models applied to health are of great importance to generate efficient models that can learn from small datasets. This promotes the adoption of these technologies in entire communities facing similar problems (Bates et al., 2014; Gebreyes et al., 2014; Han et al., 2015; Roski et al., 2014; Wiens & Shenoy, 2018).

In this study, two ML models with applications in epidemiology are presented, using SHAP analysis techniques to examine the influence of the predictor variables and their interactions in the proposed model. The SHAP analyses helped to better understand the behavior of the predictor variables within the model, and thus enable developments that best fit the dataset in question. The predicted SHAP values for each variable allowed us to visualize and weigh those that were most influential, and, through this, to determine the nature of the variable and decide on future courses of action. It should also be noted that the variables in these ML models modify both their weight and their sign throughout the simulated data set.

The aim of the current study was to break the "black box" paradigm using ML technology, which uses SHAP to determine the contribution of each variable in the ML model applied to geospatial health. On the other hand, as described in the result section, the construction of a histogram of the residuals or errors is in accordance with that observed in other studies (Emsley et al., 2010; Mayer & Butler, 1993; Baddeley et al., 2005) where the use of this type of plots of residuals is proposed as an adequate tool for validation of a model. Although the results of the baseline XGBoost model are not fully disclosed, the number of individuals per household was the most weighted variable in the prediction, as in previous studies, where it was shown that hookworm infection seems to be determined by the number of individuals per household, since infection tends to be more prevalent when there's overcrowding (Milano et al., 2007; Parija et al., 2017; Chen & Guestrin, 2016b,a; Romero-Sandoval et al., 2017). However, this variable was excluded from the model in order to observe the interactions of the rest of the variables in greater detail.

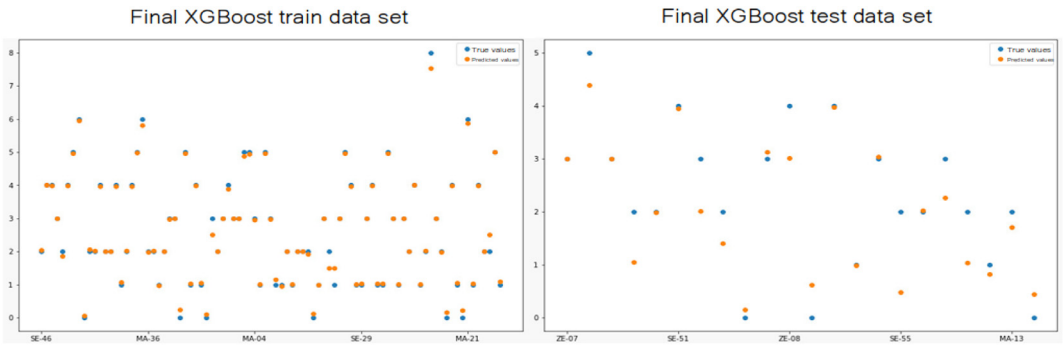

Fig. 11. Observed values vs. predicted values plot on the TRAIN vs TEST of XGBoost final model. It can be observed in the left panel the plot corresponding to the train set, and in the right panel the plot for the test. For both graphs, the blue dots represents the real number of hookworm infected individuals per household, and orange plots shows model's prediction of the number of hookworm infected individuals per household.

As for the parameters and their configuration, we can say that Booster, General and Task are the three main parameter sets of XGBoost. Although XGBoost has many parameters and different combinations of parameters obtain different evaluation scores, in this work it has shown excellent results in all aspects. It should be clarified that in this manuscript only some parameters of the Booster set were modified, being that in most cases, Booster is used to define the details of the boosting tree. Thus, the definition of each tree can be precise; and we do not need to adjust all the parameters (Chen et al., 2015; Chen and Guestrin, 2016a, 2016b; Jiang et al., 2019; Scavuzzo et al., 2018, 2020).

In the SHAP methods decision plot, summary and absolute summary plots, the most influential variables in the prediction of the model were electricity, the risk constructed from the ORCD (Soil organic carbon content), the BSI (Bare Soil Index), the CLYP (Weight percentage of the clay particles), and the MSI (Moisture Stress Index); this is in agreement with the results of the model applied in the work of Anegagrie et al. where it is specified that the same variables were statistically significant. For the village of Sesaberet, the variables "drinking_1: 2" (origin of water for consumption during the rainy season: pipeline) and "hand wash: 2" (origin of water used for handwashing: pipeline) were the variables that most decreased the prediction together with a soil variable, being the absence of electricity the one that most increased it (Fig. 5b). The same results were obtained in the force plot explained for the village of Sesaberet, which also shows a similar distribution and interaction. This is in agreement with previous studies (Anegagrie et al., 2021; Morales-Espinoza et al., 2003; Periago et al., 2018; Chen & Guestrin, 2016a,b; Molla & Mamo, 2018; Tekalign et al., 2019; Anunobi et al., 2019; Grimes et al., 2016; Loukouri et al., 2019; Muluneh et al., 2020; Oswald et al., 2017). Moreover, for the village of Mazoria, the force plot method applied (Fig. 6) showed that low Moisture Stress Index (MSI) decreased the prediction, the absence of electricity increased the number of predicted infected individuals, and this village also presented a lower NDBI and a higher NDVI (Anegagrie et al., 2021). This agrees with what is shown in the dependence plot (Fig. 8), where electricity and vegetation cover are intimately related, as also observed in previous studies (Alvarez Di Fino et al., 2020; Chaiyos et al., 2018; Knopp et al., 2008; Mudenda et al., 2012; Oluwole et al., 2015; Ovutor et al., 2017; Sedionoto & Anamnart, 2018).

On the other hand, Zenzelema differs in its characteristics in comparison to the other two villages included in the study as already described above and in the study conducted by Anegagrie et al., with the lowest number of hookworm infected individuals. Based on results shown in Fig. 5a and the force plot (Fig. 7), the presence of electricity interacts with a lower MSI, decreasing the prediction of the model. This is consistent with our results, where we can see that the "output value" in the force plot or final model prediction is the lowest compared to the other two villages.

The SHAP methods applied in this study showed a good approximation to the real situation of each village. This package allowed us to explain why the model predicts specific risks, allowing us to plan appropriate evidence-based interventions. It should be clarified that in each graph obtained, the most important characteristics influencing the risk are shown for quick reference. Each group of characteristics is ranked according to the magnitude of their impact and the characteristics with the greatest influence on the variable response. We not only provide the model with the characteristics we consider important, but we allow the model to use the characteristics it chooses. This implies that the package may encounter characteristics that it was not expected to predict in the first place (for this purpose, for some of these characteristics, it is useful to label them with indicators of their relationship to risk) (Lundberg & Lee, 2017a; Lundberg et al., 2018, 2020).

This was useful to understand the epidemiological problem presented, given that sometimes it is not feasible to obtain massive training data on a large scale, it is important to generate models that can learn from small volumes of data. It is also important to mention that the routines and techniques used are described and made available to others who may be interested in using this methodology in the github repository, including the source codes of the main routines. Therefore, based on the results obtained here, we can state that ML models are superior to classical statistical models, not only demonstrating similar results but also explaining the influence and interactions between variables within the generated models.

5. Conclusion

The SHAP methods used in this study unraveled the black box paradigm underlying the application of ML techniques. The figures revealed the interactions of all the variables with each other and how this relationship is reflected in the model for proper analysis of the variables to arrive at a better model with higher performance metrics, which allowed concluding that the variables of electricity, soil and environment are the most influential in modeling hookworm infected per household.

The study of variables obtained from satellite information was crucial in modeling in general and hookworm etiology in particular, therefore is important to recognize that satellital data is available in open access. We believe that we have presented an ML model analysis methodology that gives the possibility to explore the potential of remote sensing in the area of epidemiology from a health sciences perspective.

Regarding the limitations of the study, first of all we would like to mention the sample size of the dataset used, which could be considered to have a small N to be able to take full advantage of the statistical power of these ML tools. In any case, it was possible to complete the proposed routines without major inconveniences. However, it should be noted that with further training input, the model would improve its performance metrics.

Another limitation of the study is to recognize that these routines were adjusted and trained for a defined and particular area and contextual reality, which makes it necessary to consider that each model will be trained with social and cultural dimensions particular to the population under study, and that it cannot be extrapolated to other populations that do not share these dimensions.

Credit author statement

C.M.S.:Methodology, Software, Formal analysis, Writing - Original Draft, Visualization.
J.M.S.:Methodology, Software, Formal analysis, Writing - Original Draft, Visualization, Data Curation.
M.N·C.:Methodology, Formal analysis, Writing - Original Draft, Visualization.
M.A.:Conceptualization, Investigation, Resources, Writing - Review and Editing.
A.A.A.:Conceptualization, Investigation, Resources, Writing - Review and Editing.
A.B.:Investigation, Resources, Writing - Review and Editing.
V·P.:Conceptualization, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review and Editing, Visualization, Supervision, Project administration, Funding acquisition.

Availability of data and materials

All data and methods generated or analyzed during this study are included in this published article.

Funding

This study was funded by Fundación Mundo Sano and Instituto de Salud Carlos III. The funders had no roles in the design of the study or collection, analysis and interpretation of the data. C.M.S. and M.N.C. had a PhD scholarship from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the Amhara National Regional State Health Bureau in Bahar Dar for its collaboration and support in this study. We appreciate the support of the Zenzelema Health Center director Mr. Tadesse Meseret Kokeb and all the staff for their collaboration. We are most grateful to the community leaders for facilitating the participation and contact with the community. We would also like to thank Marcelo Abril and Marcelo Scavuzzo for their institutional support, directors of Fundacion Mundo Sano and Instituto Gulich, respectively.

References

Abera, B., Alem, G., Yimer, M., & Herrador, Z. (2013). Epidemiology of soil-transmitted helminths, schistosoma mansoni, and haematocrit values among schoolchildren in Ethiopia. *J Infect Dev Ctries, 3*(7), 253–260.
Alvarez Di Fino, E. M., Rubio, J., Abril, M. C., Porcasi, X., & Periago, M. V. (2020). Risk map development for soil-transmitted helminth infections in Argentina. *PLoS Neglected Tropical Diseases, 14*(2), Article e0008000.
Amor, A., Rodriguez, E., Saugar, J. M., Arroyo, A., López-Quintana, B., & Abera, B. (2016). High prevalence of strongyloides stercoralis in school-aged children in a rural highland of north-western Ethiopia: The role of intensive diagnostic work-up. *Parasites & Vectors, 1*(9), 6–17.
Anegagrie, M., Lanfri, S., Amor Aramendia, A., Scavuzzo, C. M., Herrador, Z., Benito, A., & Periago, M. V. (2020). Environmental characteristics around the household are strongly associated with hookworm infection in rural communities from bahir dar, amhara region, Ethiopia. *Actualizar, 1*(1), 1–2.
Anegagrie, M., Lanfri, S., Aramendia, A. A., Scavuzzo, C. M., Herrador, Z., Benito, A., & Periago, M. V. (2021). Environmental characteristics around the household and their association with hookworm infection in rural communities from bahir dar, amhara region, Ethiopia. *PLoS Neglected Tropical Diseases, 15*(6), Article e0009466.
Anunobi, J. T., Okoye, I. C., Aguzie, I. O., Ndukwe, Y. E., & Okpasuo, O. J. (2019). Risk of soil-transmitted helminthiasis among agrarian communities of kogi state, Nigeria. *Annals of global health, 85*(1).
Aramendia, A. A., Anegagrie, M., Zewdie, D., Dacal, E., Saugar, J. M., & Herrador, Z. (2020). Epidemiology of intestinal helminthiases in a rural community of ethiopia: Is it time to expand control programs to include strongyloides stercoralis and the entire community? *PLoS Neglected Tropical Diseases, 6*(14).
Azamathulla, H. M., Ab Ghani, A., & Fei, S. Y. (2012). ANFIS-based approach for predicting sediment transport in clean sewer. *Applied Soft Computing Journal, 12*(3), 1227–1230.
Baddeley, A., Turner, R., Møller, J., & Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B, 67*(5), 617–666.
Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123–1131.
Bose, P., Kasabov, N. K., Bruzzone, L., & Hartono, R. N. (2016). Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. *IEEE Transactions on Geoscience and Remote Sensing, 54*(11), 6563–6573.
Brown, M. E., Lary, D. J., Vrieling, A., Stathakis, D., & Mussa, H. (2008). Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS. *International Journal of Remote Sensing, 29*(24), 7141–7158.
Campbell, S. J., Savage, G. B., Gray, D. J., Atkinson, J. A., Soares Magalhaes, R. J., & Nery, S. V. (2014). Water, sanitation, and hygiene (wash): a critical component for sustainable soil-transmitted helminth and schistosomiasis control. *PLoS Neglected Tropical Diseases, 4*(8), Article e2651.
Chaiyos, J., Suwannatrai, K., Thinkhamrop, K., Pratumchart, K., Sereewong, C., Tesana, S., Kaewkes, S., Sripa, B., Wongsaroj, T., & Suwannatrai, A. (2018). Maxent modeling of soil-transmitted helminth infection distributions in Thailand. *Parasitology Research, 117*(11), 3507–3517.
Chen, T., & Guestrin, C. (2016a). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international Conference on knowledge Discovery and data mining, KDD '16* (pp. 785–794). New York, NY, USA: ACM.

Chen, T., & Guestrin, C. (2016b). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. *R package version, 1*(4), 1–4, 0.4-2.
Clasen, T., Boisson, S., Routray, P., Cumming, O., Jenkins, M., & Ensink, J. H. (2019). The effect of improved rural sanitation on diarrhoea and helminth infection: Design of a cluster-randomized trial in Orissa, India. *Emerging Themes in Epidemiology, 1*(9), 7.
Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of coot. *Acta Crystallographica Section D Biological Crystallography, 66*(4), 486–501.
Estallo, E. L., Benitez, E. M., Lanfri, M. A., Scavuzzo, C. M., & Almirón, W. R. (2016a). MODIS environmental data to assess Chikungunya, Dengue, and Zika diseases through Aedes (Stegomia) aegypti oviposition activity estimation. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9*(12), 5461–5466.
Estallo, E. L., Benitez, E. M., Lanfri, M. A., Scavuzzo, C. M., & Almirón, W. R. (2016b). Modis environmental data to assess chikungunya, dengue, and zika diseases through aedes (stegomia) aegypti oviposition activity estimation. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9*(12), 5461–5466.
Gebreyes, W. A., Dupouy-Camet, J., Newport, M. J., Oliveira, C. J., Schlesinger, L. S., Saif, Y. M., Kariuki, S., Saif, L. J., Saville, W., Wittum, T., et al. (2014). The global one health paradigm: Challenges and opportunities for tackling infectious diseases at the human, animal, and environment interface in low-resource settings. *PLoS Neglected Tropical Diseases, 8*(11), Article e3257.
Gilbert, F. (2019). *Introducing shap decision plots visualize the inner workings of machine learning models with greater detail and flexibility*.
Grimes, J. E., Tadesse, G., Mekete, K., Wuletaw, Y., Gebretsadik, A., French, M. D., Harrison, W. E., Drake, L. J., Gardiner, I. A., Yard, E., et al. (2016). School water, sanitation, and hygiene, soil-transmitted helminths, and schistosomes: National mapping in ethiopia. *PLoS Neglected Tropical Diseases, 10*(3), Article e0004515.
Han, B. A., Schmidt, J. P., Bowden, S. E., & Drake, J. M. (2015). Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences, 112*(22), 7039–7044.
Jafari Goldarag, Y., Mohammadzadeh, A., & Ardakani, A. S. (2016). Fire risk assessment using neural network and logistic regression. *Journal of the Indian Society of Remote Sensing, 44*(6), 885–894.
Jiang, Y., Tong, G., Yin, H., & Xiong, N. (2019). A pedestrian detection method based on genetic algorithm for optimize xgboost training parameters. *IEEE Access, 7*, 118310–118321.
Karagiannis-Voules, D. A., Biedermann, P., Ekpo, U. F., Garba, A., Langer, E., & Mathieu, E. (2015). Spatial and temporal distribution of soil-transmitted helminth infection in sub-saharan africa: a systematic review and geostatistical meta-analysis. *The Lancet Infectious Diseases, 14*(15), 74–84.
Knopp, S., Khalfan, A. M., Khamis, I., Mgeni, A. F., Stothard, J. R., Rollinson, D., Marti, H., & Utzinger, J. (2008). Spatial distribution of soil-transmitted helminths, including strongyloides stercoralis, among children in Zanzibar. *Geospatial health, 3*(1), 47–56.
Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers, 7*(1), 3–10 (Special Issue: Progress of Machine Learning in Geosciences).
Lary, D. J., Remer, L. A., MacNeill, D., Roscoe, B., & Paradise, S. (2009). Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geoscience and Remote Sensing Letters, 6*(4), 694–698.
Loukouri, A., Méité, A., Kouadio, O. K., Djè, N. N., Trayé-Bi, G., Koudou, B. G., & N'Goran, E. K. (2019). Prevalence, intensity of soil-transmitted helminths, and factors associated with infection: Importance in control program with ivermectin and albendazole in Eastern Côte d'ivoire. *Journal of Tropical Medicine, 2019*, 1–10, 2019.
Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence, 2*(1), 2522–5839.
Lundberg, S. M., & Lee, S. (2017a). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.
Lundberg, S. M., & Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.
Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering, 2*(10), 749–760.
Madadi, M. R., Azamathulla, H. M., & Yakhkeshi, M. (2015). Application of Google Earth to investigate the change of flood inundation area due to flood detention dam. *Earth Science India, 8*(3), 627–638.
Mayer, D., & Butler, D. (1993). Statistical validation. *Ecological Modelling, 68*(1–2), 21–32.
Mengitsu, B., Shafi, O., Kebede, B., Kebede, F., Worku, D. T., & Herero, M. (2016). Ethiopia and its steps to mobilize resources to achieve 2020 elimination and control goals for neglected tropical diseases: Spider webs joined can tie a lion. *Int Health, 1*(8), 134–152.
Milano, A., Oscherov, E. B., Palladino, A. C., & Bar, A. R. (2007). Children enteroparasitosis in north east argentine urban area. *Medicina, 67*(3), 238–242.
Molla, E., & Mamo, H. (2018). Soil-transmitted helminth infections, anemia and undernutrition among schoolchildren in yirgacheffee, South Ethiopia. *BMC Research Notes, 11*(1), 1–7.
Morales-Espinoza, E. M., Sánchez-Pérez, H. J., del Mar García-Gil, M., Vargas-Morales, G., Méndez-Sánchez, J. D., & Pérez-Ramírez, M. (2003). Intestinal parasites in children, in highly deprived areas in the border region of chiapas, Mexico. *salud pública de méxico, 45*(5), 379–388.
Mudenda, N. B., Malone, J. B., Kearney, M. T., Mischler, P. D., del Mar Nieto, P., McCarroll, J. C., & Vounatsou, P. (2012). Modelling the ecological niche of hookworm in Brazil based on climate. *Geospatial health, 6*(3), S111–S123.
Muluneh, C., Hailu, T., & Alemu, G. (2020). Prevalence and associated factors or soil-transmitted helminth infection among children living with and without open defecation practices in northwest ethiopia: A comparative cross-sectional study. *The American Journal of Tropical Medicine and Hygiene, 1*(103), 266–272.
Nute, A. W., Endeshaw, T., Stewart, A. E. P., Sata, E., Bayissasse, B., & Zerihun, M. (2018). Prevalence of soil- transmitted helminths and schistosoma mansoni among a population-based sample of school-age children in amhara region, Ethiopia. *Parasites & Vectors, 1*(11).
Oluwole, A. S., Ekpo, U. F., Karagiannis-Voules, D.-A., Abe, E. M., Olamiju, F. O., Isiyaku, S., Okoronkwo, C., Saka, Y., Nebe, O. J., Braide, E. I., et al. (2015). Bayesian geostatistical model-based estimates of soil-transmitted helminth infection in Nigeria, including annual deworming requirements. *PLoS Neglected Tropical Diseases, 9*(4), Article e0003740.
O'Reilly, C. E., Freeman, M. C., Ravani, M., Migele, J., Mwaki, A., & Ayalo, M. (2008). The impact of a school-based safe water and hygiene programme on knowledge and practices of students and their parents: Nyanza province, western Kenya. *Epidemiology and Infection, 1*(136), 80–91.
Organization, W. H. (2010). *First who report on neglected tropical diseases: Working to overcome the global impact of neglected tropical diseases* (Vol. 1). Switzerland: WHO Geneva.
Organization, W. H. (2020). *Ending the neglect to attain the sustainable development goals: A road map for neglected tropical diseases 2021–2030. Technical report*. World Health Organization.
Oswald, W. E., Stewart, A. E., Kramer, M. R., Endeshaw, T., Zerihun, M., Melak, B., Sata, E., Gessese, D., Teferi, T., Tadesse, Z., et al. (2017). Association of community sanitation usage with soil-transmitted helminth infections among school-aged children in amhara region, Ethiopia. *Parasites & Vectors, 10*(1), 1–13.
Ovutor, O., Helen, I., & Awi-waadu, G. D. (2017). Assessment of physico-chemical parameters of soils in fallowing farmlands and pit toilet environments as it affects the abundance of geohelminthes in emohua local government area, rivers state, Nigeria. *Annual Research & Review in Biology*, 1–10.
Parija, S. C., Chidambaram, M., & Mandal, J. (2017). Epidemiology and clinical features of soil-transmitted helminths. *Tropical parasitology, 7*(2), 81.
Peña-Barragán, J., Gutiérrez, P. A., Hervás-Martínez, C., Six, J., Plant, R. E., & López-Granados, F. (2014). Object-based image classification of summer crops with machine learning methods. *Remote Sensing, 6*(6), 5019–5041.

Periago, M., García, R., Astudillo, O., Cabrera, M., & Abril, M. (2018). Prevalence of intestinal parasites and the absence of soil-transmitted helminths in añatuya, santiago del estero, argentina. *Parasites & Vectors, 1*(11).
Polop, F., Provensal, C., Scavuzzo, M., Lamfri, M., Calderón, G., & Polop, J. (2007). On the relationship between the environmental history and the epidemiological situation of Argentine hemorrhagic fever. *Ecological Research, 23*(1), 217.
Porcasi, X., Rotela, C. H., Introini, M. V., Frutos, N., Lanfri, S., Peralta, G., De Elia, E. A., Lanfri, M. A., & Scavuzzo, C. M. (2012). An operative dengue risk stratification system in Argentina based on geospatial technology. *Geospatial Health, 6*(3 SUPPL), S31–S42.
Romero-Sandoval, N., Ortiz-Rico, C., Sánchez-Pérez, H. J., Valdivieso, D., Sandoval, C., Pástor, J., & Martín, M. (2017). Soil transmitted helminthiasis in indigenous groups. a community cross sectional study in the amazonian southern border region of ecuador. *BMJ Open, 7*(3), Article e013626.
Roski, J., Bo-Linn, G. W., & Andrews, T. A. (2014). Creating value in health care through big data: Opportunities and policy implications. *Health Affairs, 33*(7), 1115–1122.
Rotela, C., Lopez, L., Frías Céspedes, M., Lighezzolo, A., Porcasi, X., Lanfri, M., Scavuzzo, C., & Gorla, D. (2017). *Analytical report of the 2016 dengue outbreak in Córdoba city, Argentina* (Vol. 12). Geospatial Health.
Scavuzzo, J. M., Scavuzzo, C. M., Espinosa, M., Andreo, V., Campero, M. N., Periago, V., & Abril, M. (2020). Estimación de la importancia de variables predictoras en modelos epidemiológicos de aprendizaje automático utilizando shap. In *2020 IEEE Congreso Bienal de Argentina (ARGENCON)* (pp. 1–6).
Scavuzzo, J. M., Trucco, F., Espinosa, M., Tauro, C., Abril, M., Scavuzzo, C. M., & Frery, A. C. (2018). Modeling dengue vector population using remotely sensed data and machine learning. *Acta Tropica, 185*, 167–175.
Sedionoto, B., & Anamnart, W. (2018). Prevalence of hookworm infection and strongyloidiasis in cats and potential risk factor of human diseases. In *E3S web of conferences* (Vol. 31)EDP Sciences, 06002.
Souris, M. (2019). *Epidemiology and geography: Principles, methods and tools of spatial analysis*. John Wiley and Sons.
Strunz, E. C., Addiss, D. G., Stocks, M. E., Ogden, S., Utzinger, J., & Freeman, M. C. (2014). Water, sanitation, hygiene, and soil-transmitted helminth infection: a systematic review and meta-analysis. *PLoS Medicine, 3*(11).
Tekalign, E., Bajiro, M., Ayana, M., Tiruneh, A., & Belay, T. (2019). Prevalence and intensity of soil-transmitted helminth infection among rural community of southwest ethiopia: a community-based study. *BioMed Research International, 2019*, 1–7, 2019.
Wang, D., Li, Y., & Gao, B. (2016). Neural network technology and semi-analytical approach combined model for remote sensing chlorophyll-a concentration. In *2016 IEEE international geoscience and remote sensing symposium (IGARSS)* (pp. 5852–5855).
Weatherhead, E. C., Reinsel, G. C., Tiao, G. C., Meng, X.-L., Choi, D., Cheang, W.-K., Keller, T., DeLuisi, J., Wuebbles, D. J., Kerr, J. B., et al. (1998). Factors affecting the detection of trends: Statistical considerations and applications to environmental data. *Journal of Geophysical Research: Atmospheres, 103*(D14), 17149–17161.
Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases, 66*(1), 149–153.
Yi, J., & Prybutok, V. R. (1996). A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution, 92*(3), 349–357.
Zahabiyoun, B., Goodarzi, M. R., Bavani, A. R. M., & Azamathulla, H. M. (2013). Assessment of climate change impact on the Gharesou river basin using SWAT hydrological model. *Clean - Soil, Air, Water, 41*(6), 601–609.