Malaria treatment scheme model recommendation using Machine Learning and routine surveillance data

Carlos Beluzo[a,b], Everton Silva[a], Luciana Alves[b]

[a] *Federal Institute of São Paulo, Campinas - SP, Brazil*
[b] *Population Studies Center Elza Berquó, University of Campinas, Campinas/SP - Brazil*

Abstract

Malaria is a worldwide public health problem affecting some of the largest populations in the world. It is a parasitic disease predominating in less developed regions. Among response actions that can be taken out for infected individuals, the definition of which treatment scheme to be used is crucial for the effective and efficient recovery of patients, as well as for public health planning activities. Currently, according to data from Brazilian Malaria Epidemiological Surveillance Information System, there are 28 predefined treatment schemes in use, in addition to the category ”other treatment schemes”, when a customized scheme is defined for a specific case. Considering that, despite the diversity of predefined schemes, more than 82% of the cases observed in the dataset use the same treatment scheme. In this paper we present a machine learning model that, based on routine surveillance dataset, performs recommendation to use the most frequent treatment scheme or not. The model was trained using a dataset having more than 1,2 mi records with epidemiological, demographic and socioeconomic information, and was implemented using Logistic Regression, Decision Tree and Extreme Gradient Boost algorithms. In the best performance the model achieved metrics of 91% on Accuracy, 93% on Sensitivity, 82% on Specificity and an AUC value of 0.87, which are considered very acceptable metrics of a good model. Besides that, to better understand the model decisions, it was also applied a feature importance procedure, which identified as the most relevant the features exam result, gestation time, other hemoparasites researches result, active/passive detection, and patient age. Obviously it’s not the intent to propose the model as substitute for the health agents and protocols, but provide a decision support tool and pointing out the possibilities that can be achieved. Furthermore, from a demographic point of view, studies based on data analysis are valuable to corroborate important statements, previously settled by using small populations without an expressive statistical sample.

Keywords: malaria, treatment, machine learning, demographic features, models interpretability, features importance

1. Introduction

Malaria is a serious disease caused by parasites that are transmitted to people through the bite of infected female *Anopheles* mosquitoes and it is present in practically all tropical and subtropical regions, and continues being a worldwide public health problem. Parasitic diseases predominate in less developed regions, affecting some of the largest populations in the world,

influencing the development of countries where the levels of education and basic sanitation are low, causing many deaths [1]. Some population groups are at considerably higher risk of contracting malaria, and developing severe disease, than others, these include infants, children under 5 years of age, pregnant women, and patients with HIV/AIDS, as well as non-immune migrants, mobile populations, and travelers [2].

According to World Malaria Report 229 million cases of malaria were estimated in the world in 2019, in 87 endemic countries, a decrease compared to the situation in 2000, which indicated 238 million cases and 108 countries with endemic malaria situation. In addition, the estimated number of deaths from malaria continues to decline in line with the historical series beginning in 2000, where 736,000 deaths were estimated, rising to 594,000 in 2010 and reaching 409,000 in 2019 [3].

Over the same period, in the Americas, malaria cases were reduced by 40%, case incidence by 53%, deaths rate reduced by 39% and mortality by 50%. Despite that, the disease remains a major challenge for Brazil, which despite accumulated knowledge about the disease and years of national campaigns to fight it, together with Colombia and Venezuela, account for 86% of the total cases in this region [3]. It should be emphasized that the current organization of the health system and socioeconomic conditions contribute to sustained vulnerability to transmission, since vector control measures are not carried out and adapted according to local needs, which is very difficult work in some geographical locations like at Amazon forest for example.

In Brazil, the definition of treatment scheme to be used is one of the several actions taken by public health program in response to the infected individuals. According to the information available for the period 2007 to 2019 in the Brazilian Malaria Epidemiological Surveillance Information System (SIVEP - Malaria), 28 different predefined treatment schemes can be identified, in addition to the category ”Other Treatment Schemes”, that means when a customized treatment is defined for a specific cases.

Despite the diversity of predefined schemes, 82.74% of the cases observed in the dataset use the same treatment scheme: *”Infections with P. vivax, or P. ovale with chloroquine in 3 days and primaquine in 7 days”* that will be referenced on this paper as ”Treatment Scheme 11”. For this reason, we decided to lead this problem as a binary classification problem. Then, the purpose is to create a Machine Learning (ML) model capable of indicating if the treatment scheme to be used is the most frequent one or not. As depicted on Figure 1 the dataset is imbalanced and there are a huge samples for Treatment Schema 11 compared with all other schemes.

2. Data

National malaria control programs need to take special measures to protect population groups from infection, taking into consideration their specific circumstances. SIVEP-Malaria is the main tool used by the Brazilian National Program for Prevention and Control of Malaria, for the prevention and control of the disease and to improve the quality of information produced about it [4, 5]. Thus, the dataset generated and maintained by this system, allows spatial and temporal monitoring of epidemics but also serves to assess the coverage of diagnosis and treatment [6].

SIVEP-Malaria is a platform for collecting and disseminating relevant data to malaria surveillance in Brazil. Much of the information available today about the epidemiological

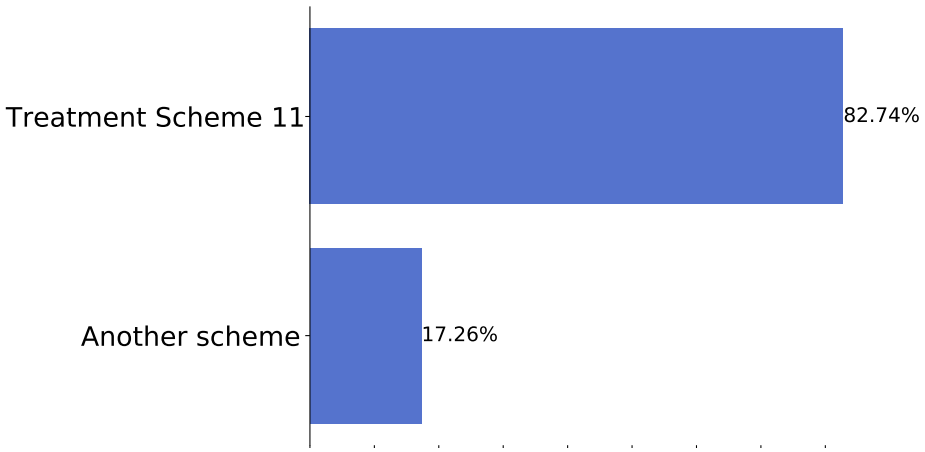

Figure 1: Dataset class distribution. The bartplot shows that the dadaset is imbalanced and there are a huge samples for Treatment Schema 11 compared with all other schemes.

characteristics of the disease originates from the dataset itself, such as the number of cases, annual parasitic index (IPA), lethality of the disease and the demographic and socioeconomic profile of the patients.

The initial dataset extraction had a total of 29,414,406 records of notifications in the period of 2007 to 2019, from the Brazilian Legal Amazon. From these, 12.5% refers to notified individuals that were then confirmed as infected (positive cases). For this paper, only the positive cases were used (3,677,019 notifications), as the model intends to evaluate treatment schemes used for these individuals. After dataset preprocessing, the final dataset had 1,223,696 records, each of them uniquely identified by the key fields' notification number, notification date and notifying municipality.

The original dataset comprises 40 variables that can be classified in three major groups: administrative data, patient information and epidemiological/laboratory data [5]. For this paper, variables having more than 70% of missing values were dropped. Also, variables related to location and dates were discarded. The variables select for this paper are described on Table 1.

2.1. Treatment Schemes

In Brazil, diseases with an endemic profile that cause a socioeconomic impact on the population, including malaria, are targets of specific public policies for their control, which includes free availability of diagnostic and therapeutic resources. In addition, the prescription and dispensing of antimalarials should only be done with a confirmatory laboratory result. The management of malaria is not private to the physician and, therefore, it can be performed by professionals from other areas of Health, which enables the maintenance of actions recommended in routines, guidelines and clinical guidelines [7].

Medication dosages from different treatment regimens already take into account information such as the patient's age group and patient weight whenever possible. The combination of chloroquine and primaquine, used in the most frequent treatment regimen, aims to cure both blood

Table 1: Variables from SIVEP-Malaria selected for the ML model.

Group	Feature	Description
ADMISTRATIVE DATA	TIPO_LAM	Active/passive
PATIENT DATA	ID_PACIE	Patient age
	SEXO	Patient gender
	GESTANTE	Gestation time
	NIV_ESCO	School level
	RACA	Skin color
	COD_OCUP	Employment
EPIDEMIOLOGICAL / LABORIATORIAL DATA	VIVAX	Patient under *P. vivax* treatment
	FALCIPARUM	Patient under *P. falciparum* treatment
	EXAME	Exam method
	RES_EXAME	Exam result
	QTD_CRUZ	Parasitaemia
	HEMOPARASI	Other hemoparasites researched
	SINTOMAS	Symptoms
	ESQUEMA	Treatment scheme code

and liver forms and thus prevent recrudescence and relapse, respectively [7]. This is one of the reasons that it is the most used treatment.

Anyway, the Malaria treatment guide in Brazil presents a complete details on malaria treatment schemes, presenting important cautions for each treatment scheme, for example the fact that pregnant women and children under 6 months of age cannot use primaquine [7]. In this manner, the ML model built on this work must not be used as a substitute for the malaria treatment protocols, instead of that we pretend to provide a decision support tool, pointing out the possibilities that can be achieved using the proposed methodology.

3. Proposed Method

Doing an exploratory analyses on SIVEP-Malaria dataset it was identified 28 different predefined treatment schemes, in addition to the category used when a customized treatment scheme is defined for specific cases, "Other Treatment Schemes". In Table 2 it is listed the description of the 5 most frequent treatment schemes, and in Figure 2 is depicted the proportions of them. As we are addressing this problem as a binary classification task, all the treatment schemes different from Treatment Scheme 11 have been grouped under one category referred on this paper as "Other Treatment Scheme".

In this paper we present a ML model that, based on routine surveillance dataset, performs recommendation to use the most common treatment or not. This model was built using a dataset having more than 1,2 mi records with information about the epidemiological characteristics of the disease, and demographic and socioeconomic profile of the patients currently available in SIVEP-Malaria dataset, that classifies if a patient diagnosed with malaria should be treated with "Treatment Scheme 11", or not.

As an initial benchmark, Extreme Gradient Boost algorithm (XGBoost) was used to create the classifier model, due to its efficiency on problems with similar designs, achieving very acceptable results. It was also implemented models using Logistic Regression and Decision Tree classifiers

Table 2: Description of the 5 most used treatment schemes.

Treatment Scheme ID	Description
10	Infections with *P. vivax*, or *P. ovale* with chloroquine in 3 days and primaquine in 14 days
11	Infections with *P. vivax*, or *P. ovale* with chloroquine in 3 days and primaquine in 7 days
16	*P. falciparum* infections with the fixed combination of Artesunate + Mefloquine in 3 days and primaquine in a single dose
17	*P. falciparum* infections with fixed combination of artemether+lumefantrine in 3 days
24	Other Scheme used

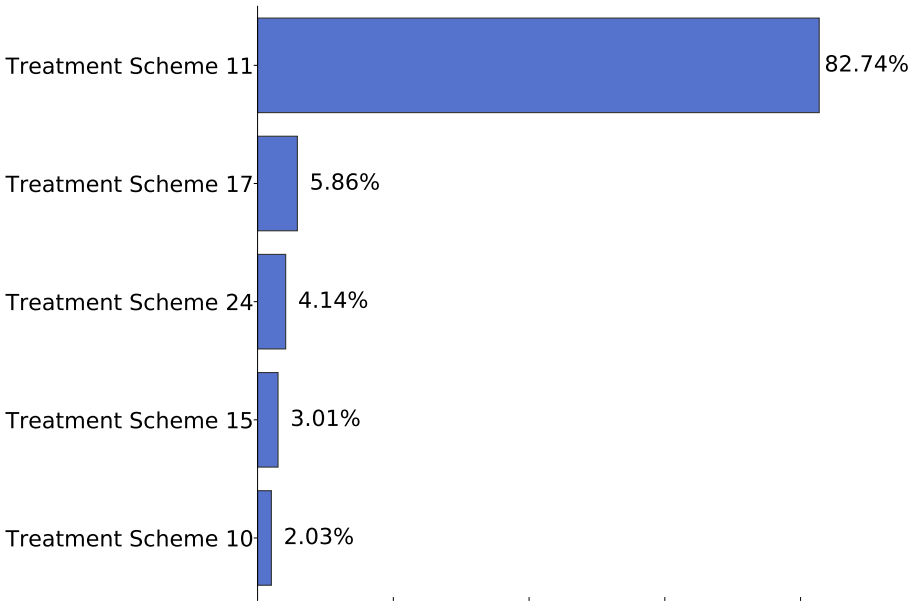

Figure 2: Treatment scheme frequency. The barplot deptics the frequency of the 5 most used treatment schemes.

in order to compare performance among then. To build the ML models, the selected SIVEP-Malaria variables were used as input features, and a new column (target feature) has been included, indicating if the current individual was treated with Treatment Scheme 11 (class 1) or with Another scheme (class 0).

4. Experiment Results

This section describes performed experiments to validate proposed method as well as results obtained. We used the cross validation process (5 folds) at 80% of the full dataset, picking the samples randomly. The other 20% will be used as a test set. After that, the final models were created using each of the 3 algorithms defined on the methodology of this paper. The results

obtained, both in the cross-validation and in the test set, are presented below.

Table 3 presents the performance results for Accuracy, Specificity, Sensitivity and Area Under the Receiver Operating Characteristic Curve (ROC AUC) metrics. The models using XGBoost and Logistic Regression algorithms presented the highest Accuracy and AUC values (93% and 0.92 respectively), followed by and Decision Tree algorithms (83% and 0.73 respectively).

Table 3: Performance metrics for each algorithm using 5-folds in cross validation process.

S/N	Algorithm	Accuracy	Sensitivity	Specificity	AUC
1	XGBoost	93%	94%	92%	0.92
2	Logistic Regression	93%	93%	92%	0.92
3	Decision Tree	91%	93%	79%	0.86

After models have been created, they were tested in the dataset slice reserved for test (20% of the initial dataset, which were not used during training step). The results shown in Table 4 present metrics for the test step and confirmed that the XGBoost algorithm has a better predictive performance compared to the others.

Table 4: Performance metrics for each algorithm in the test set.

S/N	Algorithm	Accuracy	Sensitivity	Specificity	AUC
1	XGBoost	93%	99%	67%	0.82
2	Logistic Regression	93%	99%	66%	0.82
3	Decision Tree	83%	83%	85%	0.84

Finally a feature importance procedure was performed in order to identify which of the input features are the most important for the model decision. Interpretability of ML models are a concern and for public health and demography this kind of characteristic is specially important. An expert which holds its recommendation using as help a ML model, needs to explain and justify the presented conclusions.

In this sense, besides the results of the proposed method execution, on this paper it was also applied the **SH**apley **A**dditive ex**P**lanation (SHAP) Values [8] method, to measure features importance and provide a better interpretation of results. Formally, a SHAP value measures the the influence of a feature to the output. The Figure 3 depicts the the top 5 most important features used by XGBoost model to classification task. The feature exam result (RES_EXAME) was identified as the most relevant to the model decision, and its value is been used by the model in 79% of the classifications performed.

5. Discussion

Along this paper, we proposed a new method to recommend the use of one specific treatment scheme for malaria, based in a combination of ML classifiers and routine surveillance information.

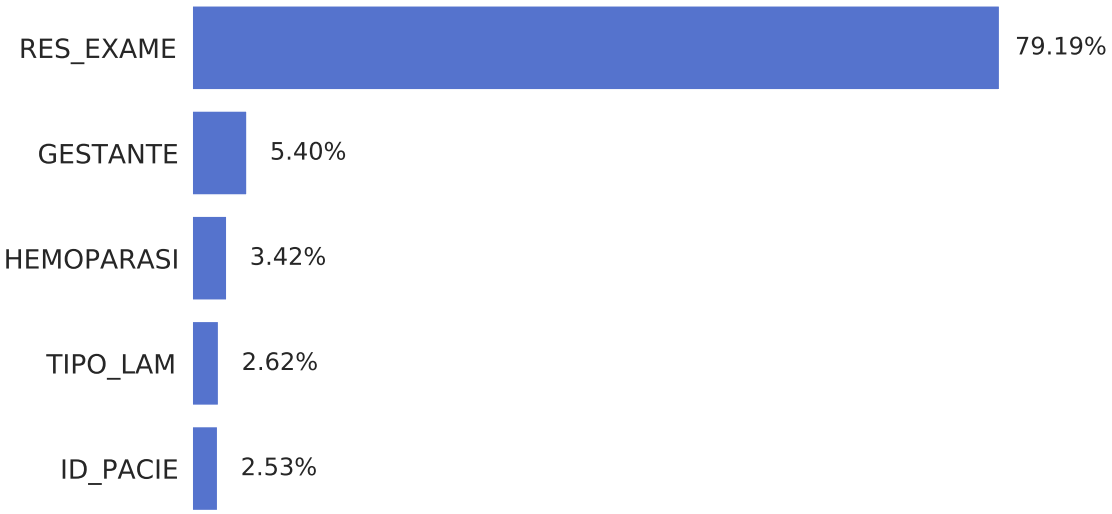

Figure 3: Fearture importance. The barplot deptics the 5 most important features used by XGBoost.

From public data collected from Brazilian government, we created a dataset, comprising more than 1.2 million samples of malaria positive cases.

With results exceeding 92% AUC when using XGBoost as final classifier, the method is able to provide both a recommendation of treatment scheme and an interpretation of the result obtained through the use of SHAP values.

Similarity, between results with two more different ML classifiers, points toward expressiveness of features, being exam result (RES_EXAME), gestation time (GESTANTE), other hemoparasites researched (HEMOPARASI), active/passive detection (TIP_LAM) and patient age (ID_PACIE) respectively the five more relevant features, as indicated by an specific analysis using SHAP values.

When evaluation the the features that the model is using when doing classification decisions, they are aligned with the protocols provided by Brazilian Ministry of Health for health agents responsible to recommends treatment scheme for an positive individual. Its important to notice that, the model built on this paper is doing the classification without humans analyses and are doing recommendations that is adherent to the protocols, without knowing them in advance, that means that the model is able to learn the protocol based on the dataset. For example the features exam result and patient age are pointed by SHAP as important features, as well as in the formal protocols.

Obviously the resultant ML model must not be used as a final substitute for the malaria treatment health agents and protocols, instead of that, we pretend to provide a decision support tool, pointing out the possibilities that can be achieved using the proposed methodology. As a decision support tool, this kind of method can be used to help health agents to take decisions on which treatment scheme recommend to a patient. Additionally, from a demographic point of view, studies based on data analysis are valuable to corroborate important statements, once most of studies are performed in small populations without an expressive statistical sample.

Data Statement and Source Code

This work uses data in the public domain, made available upon request for access to information (request id: 25820005204202066) to the Brazilian Ministry of Health, in accordance with Law No. 12,527, of November 18, 2011 (L12527). The data is de-identified, and the results present only aggregated information, therefore, this work is exempt from being evaluated by the Ethics and Research Committee, according to Resolution No. 510, of April 7, 2016, of the National Health Council (Resolution CNS/MS 510/16).

The experiments were implemented using Python Jupyter Notebook and source code is available on a public repository [9].

References

[1] França TCC, Santos MGd, Figueroa-Villar JD. Malária: aspectos históricos e quimioterapia. Química Nova;31(5):1271–1278.

[2] World Health Organization. World malaria report 2019. World Health Organization;.

[3] World Health Organization. World malaria report 2020: 20 years of global progress and challenges. World Health Organization;.

[4] Pina-Costa Ad, Brasil P, Santi SMD, Araujo MPd, Suárez-Mutis MC, Santelli ACFeS, et al. Malaria in Brazil: what happens outside the Amazonian endemic region. Memórias do Instituto Oswaldo Cruz;109(5):618–633.

[5] Wiefels A, Wolfarth-COUTO B, Filizola N, Durieux L, Mangeas M. Accuracy of the malaria epidemiological surveillance system data in the state of Amazonas. Acta Amazonica;46(4):383–390.

[6] Oliveira-Ferreira J, Lacerda MV, Brasil P, Ladislau JL, Tauil PL, Daniel-Ribeiro CT. Malaria in Brazil: an overview. Malaria Journal;9(1):115.

[7] Brasil. Ministério da Saúde. Guia de tratamento da malária no Brasil. 2020. Secretaria de Vigilância em Saúde. Departamento de Imunização e Doenças Transmissíveis;. Available from: https://portalarquivos2.saude.gov.br/images/pdf/2020/janeiro/29/af-guia-tratamento-malaria-28jan20-isbn.pdf.

[8] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765–4774.

[9] Beluzo CE, Silva E, Alves LC. Malaria treatment source code [Python Jupyter Notebook]; 2022. Available from: https://drive.google.com/file/d/1hD2lSN-ylPQcNEd6zdq96w-oSj2Crwff/view?usp=sharing.