Machine learning for malaria treatment schema recommendation using routine surveillance data

Carlos Eduardo Beluzo, Everton Josué da Silva, Natália Martins Arruda, Vinícius de Souza Maia, Bianca Cechetto Carlos, Tiago Carvalho and Luciana Correia Alves

Abstract

Malaria, a parasitic disease predominating in less developed regions, is a worldwide public health problem, affecting some of the largest populations in the world. Among the various factors that favor response actions to infected individuals, the treatment schema selected to be used in infected individuals are crucial for the effective and efficient recovery of patients, as well as for public health planning activities. According to information from Brazilian Malaria Epidemiological Surveillance System (SIVEP-Malaria), currently, there are approximately 22 treatment schemes that can be recommended to patients, and their selection is determined by several factors. In this paper we present a machine learning model to recommend which treatment to use in patients diagnosed with malaria. The model is trained with epidemiological, demographic and socioeconomic information. In preliminary experiments, using the Extreme Gradient Boost algorithm, two models were built, for the two most frequent treatment schemas. The models achieved an accuracy value of 0.94 with an AUC value of 0.914 when predicting "Schema treatment 6" against all other schemas, and accuracy value of 0.93 with an AUC value of 0.911 when predicting "Schema treatment 17".

Introduction Malaria is still a worldwide public health problem. It 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. According to the latest World Malaria Report, there were 228 million cases in 2018 compared to 231 million cases in 2017 (WHO, 2019). The estimated number of malaria deaths stood at 405,000 in 2018, compared with 416,000 deaths in 2017 (WHO, 2020).

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 (França et al., 2008).

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 (WHO, 2019).

The disease continues to be a great challenge to Brazil, that despite accumulated knowledge about it and years of national campaigns to fight it, concentrates about 34.4% of the disease cases registered in the American continent (WHO, 2017). 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.

Among the various factors that favor response actions to infected individuals is the treatment scheme to be used. Currently, there are approximately 22 different treatments that can be recommended, determined by several factors. The objective of this work is to implement a Machine Learning (ML) model to recommend the type of treatment to be used in patients diagnosed with malaria, using information about the epidemiological characteristics of the disease, and demographic and socioeconomic profile of the patients currently available in the Brazilian Malaria Epidemiological Surveillance System (SIVEP-Malaria). Two models were created, providing recommendations for the most two frequent treatments identified in the SIVEP-Malaria dataset: **a)** treatment schema "*Infections with P. vivax, or P. ovale with chloroquine in 3 days and primaquine in 7 days (short schedule)*", identified here as SCHEME 6; and **b)** treatment schema "*Another scheme used (by doctor)*" identified here as SCHEME 17.

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 (Pina-Costa et al., 2014; Wiefels et al., 2016). 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 (Oliveira-Ferreira et al., 2010).

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 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.

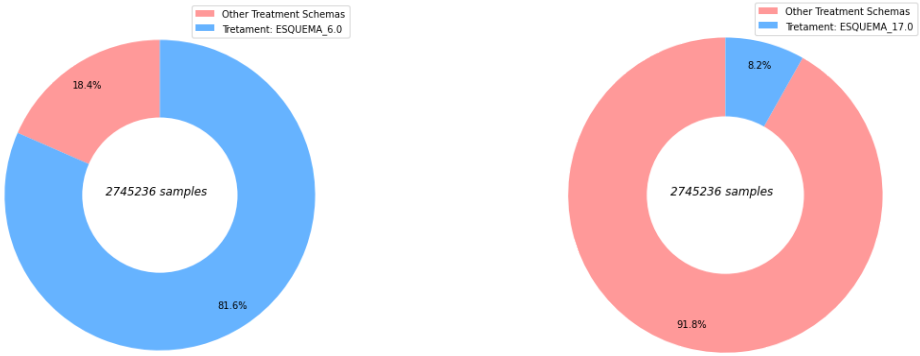
The initial dataset extraction comprises a total of 29,414,406 of notifications in the period of 2007 to 2019, from the Brazilian Legal Amazon. From these, 12.5% refers to notified individuals that were then confirmed as infected (positive cases). For this paper, only the positive cases were used (3,677,019 notifications), as the model intends to evaluate treatment schemas used for these individuals. In total, there are 40 variables that can be classified in three major groups: administrative data, patient information and epidemiology and cure control. The complete list of variables is presented in Table 1.

After dataset preprocessing, the final dataset comprises 2,745,236 records, each of them uniquely identified by the key fields' notification number, notification date and notifying municipality. Figure 1, depicts the dataset distribution for each model proposed in this paper: 1) classify between SCHEME 6 and the other schemes; and 2) classify between SCHEME 17 and the other scheme. In the first scenario, 81% of the dataset refers to the positive class (SCHEMA 6). In the second scenario the dataset will have only 8.2% of the positive class (SCHEMA 17).

Table 1. SIVEP data description table - Adapted from WIEFELS et al. (2016)

Variable	Definition	Variable	Definition	Variable	Definition	Variable	Definition
COD_NOTI	Notification number	DT_NASCI	Birth date	MUN_RESI	Municipality of residence	LOC_INFE	Locality of infection
DT_NOTIF	Notification date	ID_PACIE	Patient age	LOC_RESI	Locality of residence	DT_EXAME	Examination date
TIPO_LAM	Active/passive	ID_DIMEA	Age writing format	DT_SINTO	First symptoms date	EXAME	Examination method
UF_NOTIF	State of notification	SEXO	Sex	DT_TRATA	Date of treatmen	RES_EXAM	Examination results
MUN_NOTI	Municipality of notification	GESTANTE	Pregnancy length	VIVAX	Patient is under Vivax treatment	QTD_CRUZ	Parasitaemia
COD_UNIN	Health unit of notification	NIV_ESCO	Schooling level	FALCIPARUM	Falciparum treatment	QTD_PARA	Parasites by mm^3
COD_AGEN	Health agent code	RACA	race/skin color of the patient	ID_LVC	Follow-up consultation	HEMOPARASI	Hemoparasites
SEM_NOTI	Notification week	COD_OCUP	Employment	PAIS_INF	Country of infection	EXAMINADOR	Examiner code
DT_DIGIT	Date of digitalization	PAIS_RES	Country of residence	UF_INFEC	State of infection	Treatment schedule	
DT_ENVLO	Data entering into National database date	UF_RESID	State of residence	MUN_INFE	Municipality of infection	SINTOMAS	Symptoms
Administrative data		Patient data		Epidemiological and laboratorial data			

Figure 1. Dataset distribution among predicting class.



Methods As an initial benchmark, Extreme *Gradient Boost* algorithm was used, due to its efficiency on problems with similar designs, achieving acceptable results. To build the ML model, SIVEP-Malaria variables were used as input features, and two new columns (SCHEMA_6, SCHEMA_17) were added, indicating if the current individual was treated with one of these two treatment schemes. To build the model for recommended treatment schema 6, column SCHEMA_6 was set as the target feature to be predicted, and for treatment schema 17, column SCHEMA_17 was used instead.

Preliminary Results Initially two rounds of experiments were performed. In the first one, the model was created to predict individuals that should use treatment schema 6, in contrast with all other schemas. Similarly, in the second round, a model was created to predict individuals that should use treatment schema 17, in contrast with all other schemas. The results of model creation are depicted on Figures 2 and 3, for model predicting treatment schema 6 and 17 respectively. The model predicting SCHEMA_6 achieved an accuracy value of 0.94 with an AUC value of 0.914, and the model predicting SCHEMA_17 achieved an accuracy value of 0.93 with an AUC value of 0.911. The results of models' execution are detailed in the classification report, Table 2. Although these metrics indicate promising results, further analysis needs to be done against the metrics presented in the confusion matrix (Table 3).

Figure 2. ROC curve for model predicting treatment schemes 6 and 17 respectively.

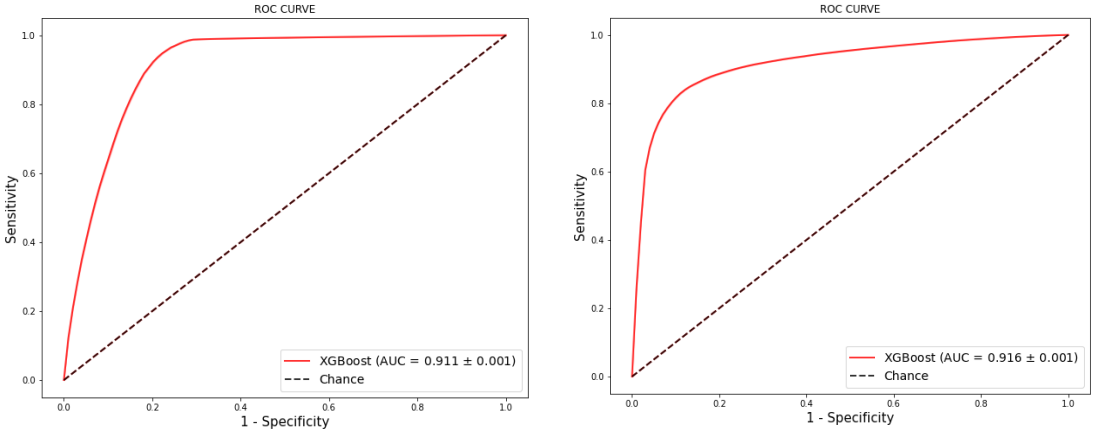


Table 2 – Model classification report

	Treatment Schema 6				Treatment Schema 17			
	precision	***recall***	***f1-score***	***support***	***precision***	***recall***	***f1-score***	***support***
Negative class	*0.84*	*0.75*	*0.79*	*505,965*	*0.98*	*0.89*	*0.93*	*2,520,826*
Positive class	*0.94*	*0.97*	*0.96*	*2,239,271*	*0.40*	*0.83*	*0.54*	*224,41*
accuracy			*0.93*	*2,745,236*			*0.89*	*2,745,236*
macro avg	*0.89*	*0.86*	*0.87*	*2,745,236*	*0.69*	*0.86*	*0.74*	*2,745,236*
weighted avg	*0.93*	*0.93*	*0.93*	*2,745,236*	*0.94*	*0.89*	*0.90*	*2,745,236*

Table 3. Confusion Matrix

Treatment *Scheme* 6		Predicted	
		0	**1**
Actual	**0**	*379.367*	*126.598*
	1	*70.906*	*2.168.365*

Treatment *Scheme* 17		Predicted	
		0	**1**
Actual	**0**	*2.244.691*	*276.135*
	1	*38.735*	*185.675*

References

A. d. Pina-Costa, P. Brasil, S. M. D. Santi, M. P. d. Araujo, M. C. Sua´rez-Mutis, A. C. F.e. S. Santelli, J. Oliveira-Ferreira, R. Lourenço de Oliveira, and C. T. Daniel-Ribeiro.Malaria in brazil: what happens outside the amazonian endemic region. 109(5):618–633, 2014.

J. Oliveira-Ferreira, M. V. G. Lacerda, P. Brasil, J. L. B. Ladislau, P. L. Tauil, and C. T. Daniel-Ribeiro. Malaria in brazil: an overview. 9:115, 2010. ISSN 1475-2875.

T. C. C. França, M. G. d. Santos, and J. D. Figueroa-Villar. Malária: aspectos históricos e quimioterapia. 31(5):1271–1278, 2008.

WHO. *World Malaria Report 2017*. WHO, 2017. ISBN 9789241565523.

WHO. *World Malaria Report 2019*. WHO, 2019. ISBN 9789241565721.

WHO. Malaria in western pacific, 2020. https://www.who.int/westernpacific/health-topics/malaria. Visited on 01 Mai 2020.

A. Wiefels, B. Wolfarth-COUTO, N. Filizola, L. Durieux, M. Mangeas, A. Wiefels, B. Wolfarth-COUTO, N. Filizola, L. Durieux, and M. Mangeas. Accuracy of the malaria epidemiological surveillance system data in the state of amazonas. 46(4): 383–390, 2016.