LADDEM Laboratório de Análise de Dados em Demografia

UNICAMP

INSTITUTO FEDERAL São Paulo Campus Campinas

Annual Meeting 2022

ACUNS ACADEMIC COUNCIL ON THE UNITED NATIONS SYSTEM

THE GRADUATE INSTITUTE GENEVA — INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT — GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

United Nations

Evidence-Based Solutions for Intensifying Global Challenges

Geneva, Switzerland | Thursday-Saturday, 23-25 June 2022

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

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

[a] beluzo@ifsp.edu.br - *Federal Institute of São Paulo - Brazil*
[b] everton.silva@ifsp.edu.br - *Federal Institute of São Paulo - Brazil*
[c] lcalves@unicamp.br - *Data Analysis Lab in Demography/NEPO, University of Campinas - Brazil*

Author profile

Ph.D Candidate @ Demography Post Graduation Program - University of Campinas - UNICAMP

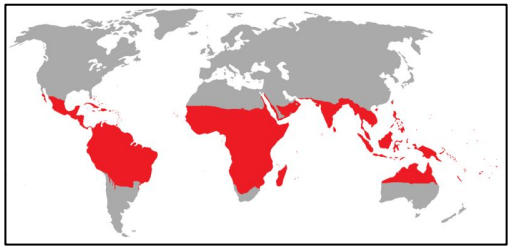
Supervisor: PhD Luciana Correia Alves

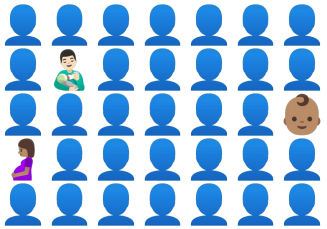
Data Science Researcher @ Project

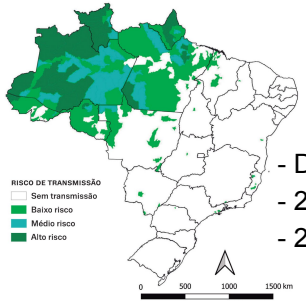
"*Data Science applied to epidemiological and demographic information as a strategy to simulation and malaria vigilance monitoring in the Brazilian Amazon*"

- Bill & Melinda Gates Foundation
- Brazilian Ministry of Health

Malaria Facts







- Desired drop until 2016
- 2017 - Raised 52%
- 2020 - 145k

Malaria treatment schemes

- Antimalarials prescription / dispensing needs lab results
- Disease management also by health agents
  - Enables routines and guidelines maintenance
- Treatment scheme just one of the actions
- SIVEP-Malaria
  - Brazilian Malaria Epidemiological Surveillance Information System
  - 29 predefined treatment schemes (2007-2019)
    - Dosages and regimens based on age and weight
- Malaria treatment guide - Details and cautions

Research proposal

- Machine learning model
  - Routine surveillance dataset
  - Recommendations on the use of the most frequent treatment scheme.
- 82% of the cases uses:

"*Infections with P. vivax, or P. ovale with chloroquine in 3 days and primaquine in 7 days*", referenced in this paper as "**Treatment Scheme 11**".

- Binary classification problem
- Trained on dataset with 1,2 mi records
  - Epidemiological, demographic, and socioeconomic information
  - Logistic Regression, Decision Tree, and Extreme Gradient Boost
- Feature importance procedure

Dataset: SIVEP-malaria

- *Brazilian National Program for Prevention and Control of Malaria*
- Epidemics spatial and temporal monitoring
- Assess diagnosis and treatments coverage
- Malaria surveillance platform for collecting and disseminating relevant data
- Used to build epidemiological characteristics knowledge
  - num of cases, API, lethality, and the demographic / socioeconomic profile.
- Initial dataset with 29+ mi notifications
  - Brazilian Legal Amazon (2007-2019)
  - 12% infected (3,6+ mi positive cases)

Proposed method

- Build knowledge regarding treatr
- Preprocessing/cleanups
  - 1,2 mi records
- 15 variables selected from 40 (o
  - Used as input features
  - Target variable indicating if **Trea**
- 3 models implemented, best per
- Models interpretability
- SHAP method to measure the in

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

Results

- Models performance
- Model can provide a recommendation with confidentially
- SHAP results
- 'exam result', 'patient age'
    - Also used in the formal protocol

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

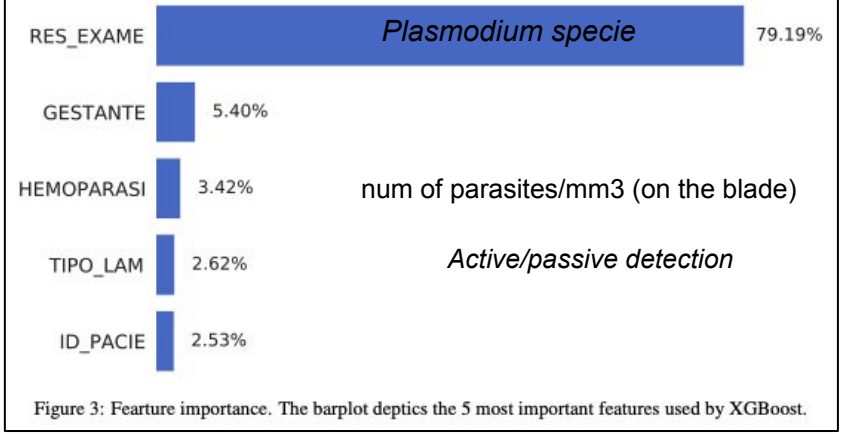


Figure 3: Fearture importance. The barplot deptics the 5 most important features used by XGBoost.

Conclusions

- Proposition of a new method to recommend a specific treatment scheme for malaria
  - Based on a combination of ML classifiers and routine surveillance information
- Created a dataset comprising more than 1,2 mi samples of malaria-positive cases
  - With public data collected from the Brazilian government
- ***The model built does the classification without human analyses***
  - ***Recommendations adhered protocols, without knowing them in advance***
  - ***Model can learn the protocol based on the dataset***
- Model is not substitute for health agents and protocols, is a Decision Support Tool
- Data Analysis studies are valuable to corroborate statements about the problem

Acknowledgment

Fulfillment

Support

Funding

This work is funded by Bill & Melinda Gates Foundation [ID INV 003970], Brazilian Ministry of Health and Brazilian National Council for Scientific and Technological Development [443048/2019-3].

Thank you for your presence!

PVD Malária

LADDEM Laboratório de Análise de Dados em Demografia

initial | About | dashboard | Views | Models | productions | Contact

Data Visualization Platform

The Malaria Data Visualization Platform - PVD Malaria - is a web page that seeks to facilitate managers, researchers and the general public to better understand the contamination and spread of Malaria in the Brazilian Legal Amazon.

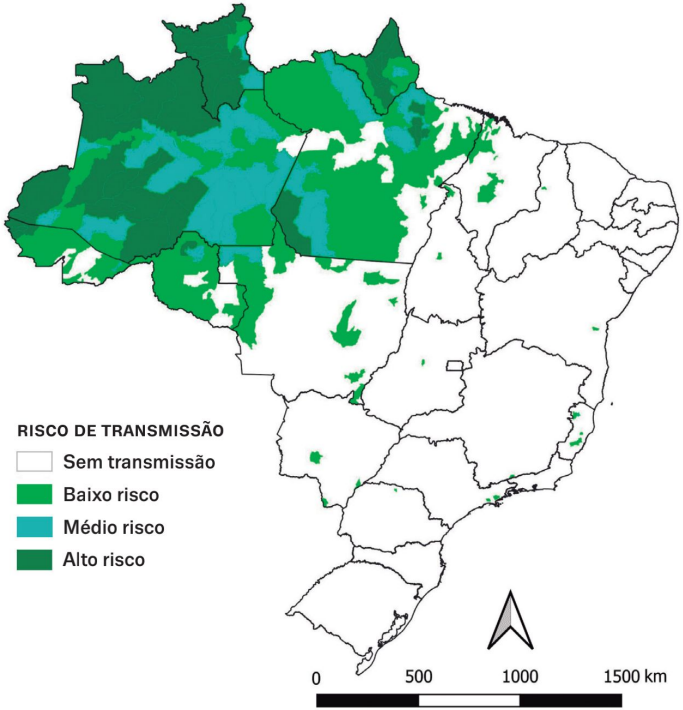
Through this platform it is possible to explore different views, both static and interactive, and create others according to your needs and interests.

The platform also offers diverse demographic, social and economic information on the Amazon region and the states that comprise it and an interactive service for analyzing indicators using artificial intelligence. In addition, a set of projections were calculated, which allows an anticipation of the future scenario and serves as a subsidy for decision making.

Brazilian Legal Amazon

https://pvd-malaria-develop-w4.web.app/producoes/boletins/amazonia-legal

Risk areas for malaria x Annual Parasitic Incidence



https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/publicacoes-svs/malaria/guia_tratamento_malaria_2nov21_isbn_site.pdf/view

SIVEP-Malaria

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

Variable	Description	Variable	Description	Variable	Description	Variable	Description
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			

https://www.longdom.org/open-access/data-visualization-for-epidemiological-and-demographic-data-for-malaria-surveillance-in-the-brazilian-amazon-20072019.pdf

Malaria most frequent treatment schemes

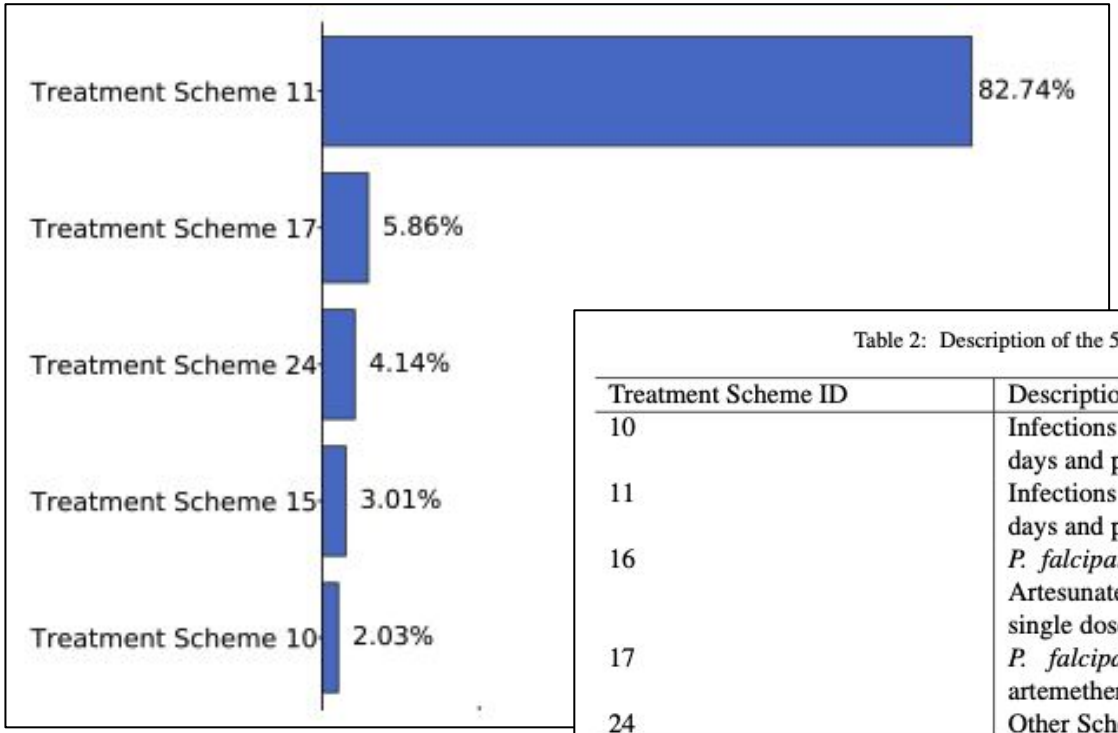
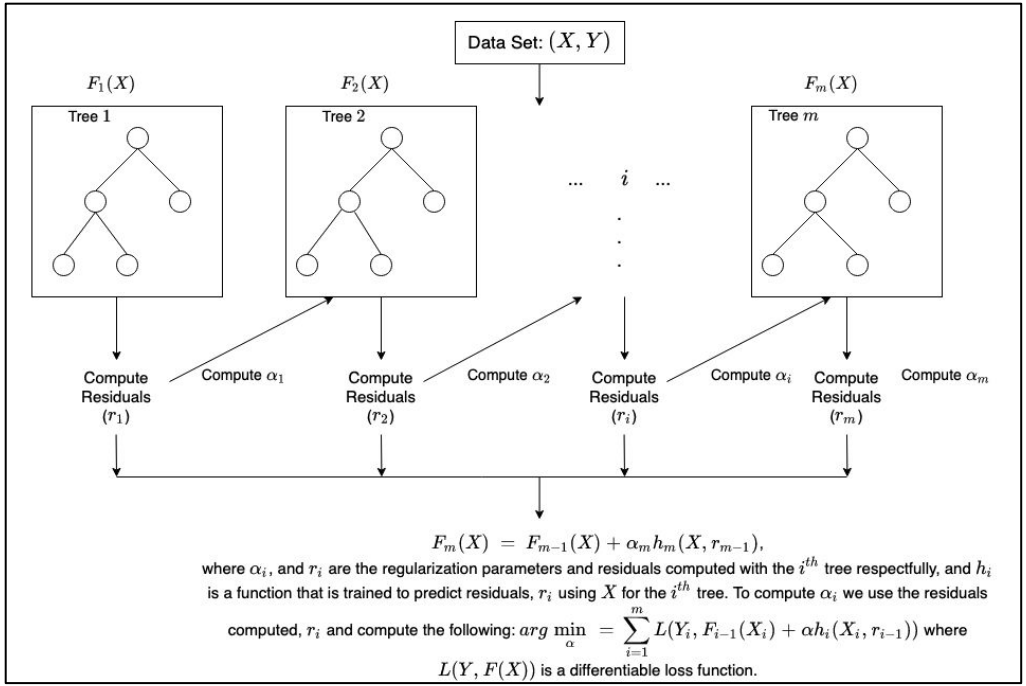


Table 2: Description of the 5 most used treatment schemes.

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

XGBoost

A popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. When using gradient boosting for regression, the weak learners are regression trees, and each regression tree maps an input data point to one of its leafs that contains a continuous score. XGBoost minimizes a regularized (L1 and L2) objective function that combines a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity (in other words, the regression tree functions). The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It's called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

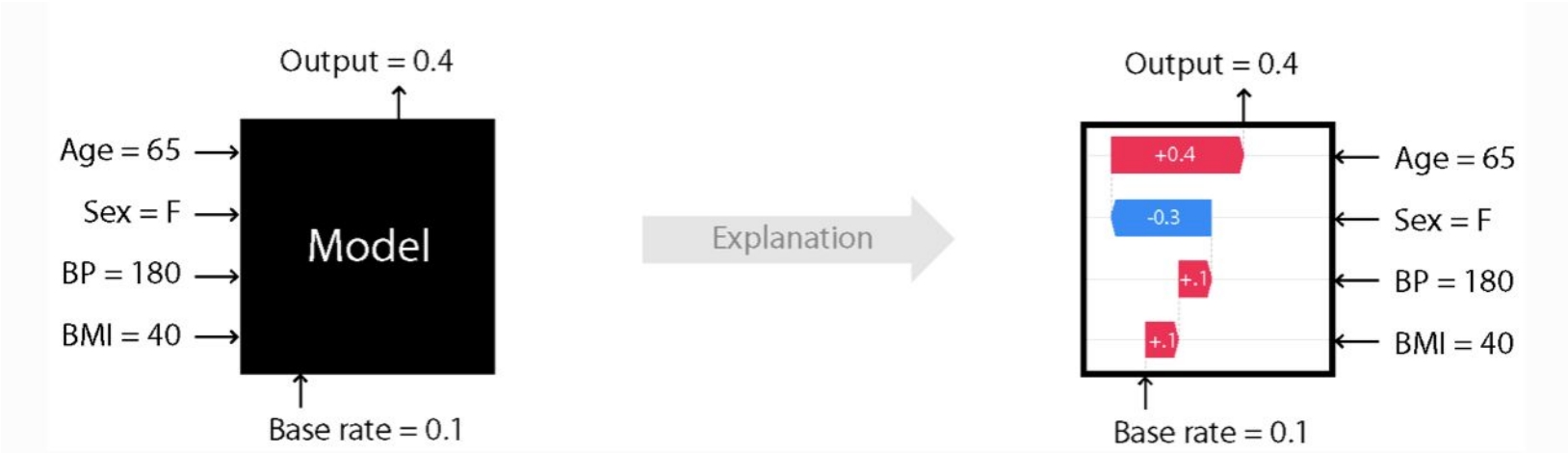


https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

SHAP

https://shap.readthedocs.io/en/latest/index.htm
https://github.com/slundberg/shap#citations

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.



Glossary

https://developers.google.com/machine-learning/glossary#classification_model

supervised machine learning - Training a model from input data and its corresponding labels. Supervised machine learning is analogous to a student learning a subject by studying a set of questions and their corresponding answers. After mastering the mapping between questions and answers, the student can then provide answers to new (never-before-seen) questions on the same topic. Compare with unsupervised machine learning.

binary classification - A type of classification task that outputs one of two mutually exclusive classes. For example, a machine learning model that evaluates email messages and outputs either "spam" or "not spam" is a binary classifier.

classification model - A type of model that distinguishes among two or more discrete classes. For example, a natural language processing classification model could determine whether an input sentence was in French, Spanish, or Italian.

classification threshold - A scalar-value criterion that is compared to a model's predicted score in order to separate the positive class from the negative class. Used when mapping logistic regression results to binary classification. For example, consider a logistic regression model that determines the probability of a given email message being spam. If the classification threshold is 0.9, then logistic regression values above 0.9 are classified as spam and those below 0.9 are classified as not spam.

class-imbalanced dataset - A binary classification problem in which the labels for the two classes have significantly different frequencies. For example, a disease dataset in which 0.0001 of examples have positive labels and 0.9999 have negative labels is a class-imbalanced problem, but a football game predictor in which 0.51 of examples label one team winning and 0.49 label the other team winning is not a class-imbalanced problem.

Glossary

https://developers.google.com/machine-learning/glossary#classification_model

true negative (TN) - An example in which the model correctly predicted the negative class. For example, the model inferred that a particular email message was not spam, and that email message really was not spam.

true positive (TP) - An example in which the model correctly predicted the positive class. For example, the model inferred that a particular email message was spam, and that email message really was spam.

false negative (FN) - An example in which the model mistakenly predicted the negative class. For example, the model inferred that a particular email message was not spam (the negative class), but that email message actually was spam.

false positive (FP) - An example in which the model mistakenly predicted the positive class. For example, the model inferred that a particular email message was spam (the positive class), but that email message was actually not spam.

AUC (Area under the ROC Curve) - An evaluation metric that considers all possible classification thresholds. The Area Under the ROC curve is the probability that a classifier will be more confident that a randomly chosen positive example is actually positive than that a randomly chosen negative example is positive.

Sensitivity and Specificity

https://developers.google.com/machine-learning/glossary#classification_model
https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Sensitivity (true positive rate) - refers to the probability of a positive test, conditioned on truly being positive.

Specificity (true negative rate) - refers to the probability of a negative test, conditioned on truly being negative.

**Sensitivity and specificity mathematically describe the accuracy of a test which reports the presence or absence of a condition. Individuals for which the condition is satisfied are considered "positive" and those for which it is not are considered "negative".*

true positive rate (TPR)

Synonym for **recall**. That is:

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

True positive rate is the y-axis in an **ROC curve**.

false positive rate (FPR)

The x-axis in an **ROC curve**. The false positive rate is defined as follows:

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

false negative rate

The proportion of actual positive examples for which the negative class is predicted. False negative rate is calculated as follows:

$$\text{False Negative Rate} = \frac{\text{False Negatives}}{\text{False Negatives} + \text{True Positives}}$$

Accuracy, precision and recall

https://developers.google.com/machine-learning/glossary#classification_model

accuracy

The fraction of **predictions** that a **classification model** got right. In **multi-class classification**, accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Number Of Examples}}$$

In **binary classification**, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number Of Examples}}$$

See **true positive** and **true negative**. Contrast **accuracy** with **precision** and **recall**.

Click the icon for additional notes.

In a **class-imbalanced dataset**, great accuracy does not always imply a great model. For example, snow falls approximately 24 times per century in a certain subtropical city. So, a binary classification snow forecasting model that automatically predicted "no snow" every day would be about 99.93% accurate. Although 99.93% accuracy seems very high, the model actually has no predictive power.

Accuracy is just one of many metrics for determining how valuable a classification model's predictions are. For example, **precision** and **recall** are usually more useful metrics than **accuracy** for assessing class-imbalanced datasets.

precision

A metric for **classification models**. Precision identifies the frequency with which a model was correct when predicting the **positive class**. That is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

true positive rate (TPR)

Synonym for **recall**. That is:

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

True positive rate is the y-axis in an **ROC curve**.

ROC curve and AUC

https://developers.google.com/machine-learning/glossary#classification_model

ROC curve

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (**FPR**) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

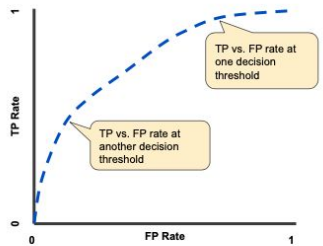


Figure 4. TP vs. FP rate at different classification thresholds.

AUC: Area Under the ROC Curve

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

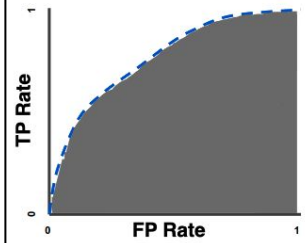


Figure 5. AUC (Area under the ROC Curve).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of logistic regression predictions:

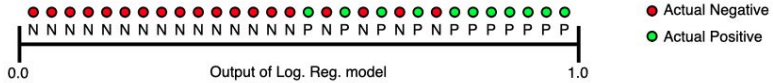


Figure 6. Predictions ranked in ascending order of logistic regression score.

AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example.

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

Resources

- *Machine Learning Glossary | Google Developers*
- *Malária — Português (Brasil)*
- *Guia de tratamento da malária no Brasil*
- *Lista de municípios pertencentes às áreas de risco ou endêmicas para malária*
- *https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/m/malaria/arquivos/fichanotificacaomalaria-sinan.pdf*
- *ESQUEMAS RECOMENDADOS PARA O TRATAMENTO DA MALÁRIA NÃO COMPLICADA NO BRASIL*
- *Orientações para o preenchimento do SIVEP-Malária*
- *Manual de Diagnóstico Laboratorial da Malária*
-

Field "*Exam Result*" = Plasmodium specie

▪ **Campo 40 – RESULTADO DO EXAME**

Preencher com código correspondente ao resultado do exame de sangue para malária.

No caso do exame da lâmina por microscopia	Para exames feitos com teste rápido
1- Negativo 2- F (P. falciparum) 3- F+FG (P. falciparum + gametócitos de P. falciparum) 4- V (P. vivax) 5- F+V (P. falciparum + P. vivax) 6- V+FG (P. vivax + gametócitos de P. falciparum) 7- FG (gametócitos de P. falciparum) 8- M (P. malariae) 9- F+M (P. falciparum + P. malariae) 10- Ov (P. ovale)	1- Negativo 2- F (P. falciparum) 5- F+V (P. falciparum + P. vivax) 11- Não F (não falciparum)

1. Resultados positivos

Os resultados do exame parasitológico da gota espessa para as diferentes espécies de plasmódios são registrados nos formulários do Sistema de Informação de Vigilância Epidemiológica – Sivep/Malária, pelas respectivas iniciais, conforme o quadro a seguir:

Registro	Descrição do resultado
V	*P. vivax*
F	*P. falciparum*
M	*P. malariae*
Ov	*P. ovale*
F + Fg	Formas assexuadas + sexuadas (gametócitos) de *P. falciparum*
Fg	Somente gametócitos
V+Fg	Formas de *P. vivax* + gametócitos de *P. falciparum*
F+V F+M V+M	Para diferentes combinações de infecções mistas

Obs: Em caso de infecção mista, registrar em primeiro lugar a inicial da espécie dominante. Exemplo:
15.000F-3Fg-500V (+++F 3Fg +V);
5.000V-60M (++V 60M)