November 2020

Prediction of Malaria with Machine Learning Algorithms : An experimental Study

Oussseynou MBAYE [a,1], Mamadou Lamine BA [b] and Alassane SY [b]
[a] *ALioune Diop University of Bambey*
[b] *ALioune Diop University of Bambey*

Abstract. Still today, Malaria remains one of the most feared diseases in Sub- Saharan Africa and especially in Senegal. This is mainly due to inappropriate medical care support coupled with an often late and error-prone diagnosis from the medical staff. In addition, largely used diagnostic standards such as the Rapid Diagnosis Test is not fully reliable. With the development and increasing adoption of automated tools in the health field, machine learning applications might help medical actors in their decision-making process. In this paper, we propose an experimental study of six machine learning algorithms for the prediction of Malaria in Senegal. These algorithms aim at predicting whether or not a given patient suffers from Malaria based on his signs and symptoms. The performance of the algorithms have been extensively tested and evaluated over real data sets about patients in Senegal that suffer or not from Malaria. The algorithms are evaluated using four criteria: accuracy, Recall, F-measure, Precision and Specificity. The research has shown that there is not necessarily a single best classification tool, but instead the best performing algorithm will depend on the dataset to be analysed

Keywords.

1. Introduction

Malaria, also known as *des marais* in French, is an infectious disease caused by a mosquito of the type *Plasmodium*. In its severe form, the disease can lead to *yellow skin*, *seizures*, *coma* or *death*. As a result, Malaria is now recognized and treated as a serious health problem worldwide by the World Health Organization (WHO), and particularly pandemic in Sub-Saharan Africa. In 2018, an estimated 228 million cases of Malaria occurred worldwide, thanks to the 2019 World Malaria Report [**?**]. Most Malaria cases in 2018 were in the WHO African Region (213 million or 93%). In the special case of Senegal the problem is acute because of the use of inappropriate care support means coupled with an often late and error-prone diagnostic from the local

[1]Corresponding Author: Book Production Manager, IOS Press, Nieuwe Hemweg 6B, 1013 BG Amsterdam, The Netherlands; E-mail: bookproduction@iospress.nl.

medical staff. Setting up a reliable way to predict the disease when a patient visits a doctor become then crucial in order to avoid its evolution towards a critical state.

Over the past years, many efforts have been done by governmental and non governmental organizations to eradicate Malaria: actions continuously conducted by the WHO are real examples of those. In the research field, many studies, aiming at understanding the disease from the Plasmodium mosquito point of view or proposing automated detection tools, have been conducted [**?**,**?**,**?**,**?**]. The Rapid Diagnostic Test (RDT) [**?**] is one of the most successful and prominent introduced tool to automatically predict whether or not a given patient suffers from Malaria. It relies on the detection of specific Plasmodium proteins, PfHRP2, pLDH and aldolase. The RDT is largely used and adopted as a standard in many health structures in Sub-African countries because of its simplicity to utilize and does not require any specific domain knowledge. However as highlighted in [**?**] the RDT is not fully reliable: in Section 2.1 we show that the precision of the RDT is about 90% for the real datasets used in this study. With the development and increasing adoption of automated tools in the health field, machine learning (ML) [**?**,**?**] applications might help medical actors in their decision-making process. There are already some attempts to apply ML techniques for the prediction or a better understanding of various diseases, e.g. [**?**,**?**]. For example, machine learning is used to analyze blood data obtained from high definition microscopic screenshots in [**?**]. On the other hand, logistic regression has been tested in [**?**] for the prediction of Malaria and provides promising results. In this paper, we propose an extensive comparative study of six machine learning algorithms, among the most popular for the prediction of Maria in Senegal. The evaluated and compared ML algorithms are Naive Bayes (NB) [**?**], Logistic Regression(LR) [**?**], Decision Tree(DT) [**?**], Support Vector Machine(SVM) [**?**], Random Forest(RF) [**?**], and Artificial Neural Network(ANN) [**?**]. Whereas the four first algorithms are simple models, the two last ones are built on more complex learning strategies. RF is an ensembling model and ANN performs Deep Learning. We conducted experiments on five datasets based on the two real world datasets about Senegalese citizens that suffer or not from Malaria. These two datasets have been collected in two different contexts and contain clinical data such as sign, symptom and final diagnostic of patients living in distinct locations in Senegal (for the first dataset) or within the same area (for the second dataset). Those patients have been examined by doctors in given health services and their clinical data recorded: for each patient the final diagnostic is provided with the corresponding signs and symptoms. The outcome of the RDT is also provided. To evaluate the performance of every considered algorithm we have considered common measures of the accuracy of a prediction system that are *Precision*, *Recall*, *F1-score*, *True Positive Rate*, and *False Positive Rate* on both datasets augmented with semi-synthetic datasets which are obtained after imputation in order to deal with missing values.

Our main result is that RF, SVM with Gaussian Kernel and ANN are promising and offer the best overall accuracy to predict the appearance or not of the disease with precision, recall and F1-score at least equal to 92%, 85% and 89% respectively on both datasets. More specifically, those three learning approaches outperform the RDT which represents the baseline automatic diagnostic tool largely adopted as a standard within the Health system in Senegal.

The rest of the paper is structured as follows.First we gives detail explanation of machine learning metthods and the different data sets methods used for this study in section 2. In section 3 our differents results are presented. Section 4 and 5 includes discussion of our results and conclusions of this paper.

2. Methods

In this part we discusse about the methods and the technic of machine learning used in this study

2.1. Real patient health datasets

In order to carry out our experiments in a real setting we have collected two real world datasets about patients living in Senegal. We describe each of them in the sequel.
Data collection. Our first dataset, that we to refer to it as DT1, contains medical records about patients living in distinct places in Senegal. It has been collected in 2016 during the **Grand Magal of Touba** which is one of the most popular religious event in Senegal. Such an event gathers every year several millions of persons that come from various areas around the country [**?**]. During the event several fixed and mobile health points are set up to enable the examination and treatment of ill persons. The second dataset, denoted by DT2, has been collected by drawing our attention on medical records about patients living in the same area. We focused on the district of Diourbel,Thies and Fatick [2] where the prevalence of Malaria is very high and collected patient records from its different health structures.
Data features. Table 1 contains the main characteristic of each dataset in terms of number of recorded variables (mainly clinical features), number of observations, variable types, number of observations per class (Malaria or Not Malaria), and the precision of the Rapid Diagnosis Test. In details, values for seventeen variables have been extracted for each observation in both datasets; two variables are basically of numerical types while the remaining are Boolean. Some of these variables (also called features or attributes) include personal data about the patient, but also signs and symptoms of the patient reported by the doctor who treated this later. The other attributes describe clinical data such as information about the doctor's final diagnosis (the patient's disease), the outcome of the Rapid Diagnosis Test and the patient's status (i.e. admission, death or observation). For privacy reasons and certain restrictions in the use of the data, we have ignored patient personal data during this study. In addition, we can observe that the first dataset is larger than the second one (21083 observations versus 5809 observations). Moreover, both datasets are unbalanced because the proportion of observations per class is largely unequal. As an example for dataset DT1 we have 614 observations in the first class and 5108 observation in the second class. Finally, we remarked that the precision of the Rapid Diagnosis Test is around 90% for both datasets, meaning that the systematically performed RDT in Senegal is not fully reliable.

On the other hand, Figure 1 shows that the raw datasets come with missing values for some variables on given observations.

[2] https://en.wikipedia.org/wiki/Diourbel_Region

Dataset	Variables	Observations	Variables types		Classes		Precision of RDT
			Numeric	Boolean	Malaria	not Malaria	
DT1	16	21083	2	14	614	20469	90.23%
DT2	16	5809	2	14	5108	701	90.49%

Table 1. Raw Data characteristics

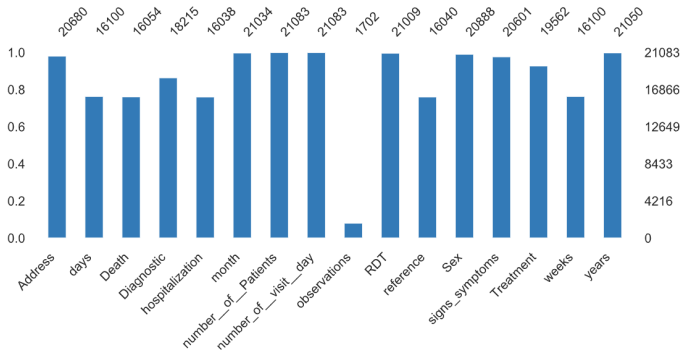

Figure 1. Proportion of missing values per variable

To resolve the problem of unbalanced datasets and data messness, we followed a data preparation pipeline describe in [**?**] in order to fit our datasets into the good format for our experimentation.

Dataset	Variables	Observations	Variables types		Classes	
			Numeric	Boolean	Malaria	not Malaria
DT1	16	61396	2	14	30698	30698
DT2	16	14336	2	14	7168	7168

Table 2. Data characteristics after preparation step

From DT1 and DT2 we built three news datasets DT3, DT4 and DT5 data sets as below.

DT3: It is obtained by concatenating the DT1 and DT2 datasets. Thus it concerns 37,175 patients of which 9,837 are diagnosed positive for malaria.

DT4: It is obtained by considering the 16,092 patients in the DT2 data set (including 9,223 patients with malaria). Since this DT2 is unbalanced, we randomly selected 2354 patients who tested negative for malaria from the DT1 data set at the end of the rebalance. Thus it concerns 18,446 patients, 9,223 of whom are suffering from malaria.

DT5: is obtained by the over sampling of DT1 by the SMOTE method of python. This method consists first of dividing DT1 into two parts, one for training (train set) and the other for testing (test set). The train set being unbalanced, then we apply the SMOTE method to remedy it. Thus we obtain a new train set comprising 30,369 patients, half of whom tested positive for malaria.

2.2. Experimentation Setting

In this section data is available for applying classification algorithm. After model creation from training data, classification operation is performed on test data. All the performed tests have been done in the same machine and the same operating system. To test the performance of our six chosen ML algorithms, we relied on their Python implementations available through the scikit-learn library. Scikit-learn is an open source simple and efficient tool for predictive data analysis that implements most of the existing ML algorithms

Then some of the most important performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. For the details about the description of each parameter of ML we refer to the official documentation of the implementation of these algorithms in scikit-learn7. Concerning the segmentation of both datasets for the training of our ML algorithms and their testing we have considered the stratified-5-fold cross-validation in classification model construction and efficiency evaluation. This method is very useful to handle data with an unbalanced class distribution, increases the validation of classification and prevents from random and invalid results.

2.3. Results of the experiments

This section presents the results of the experimentation on each real dataset for each of the six classifiers.

2.4. Discussion

In this study, the algorithms DT, RF, LR, NB, SVM and ANN were applied on five datasets concerning patients with or without malaria and living in regions of Senegal namely: Diourbel, Thies and Fatick. Indeed, in order to offer a new technique for diagnosing and predicting malaria, it is important to know the performance of those existing through our datasets. Analysing in details the performance of our six classifiers across the five datasets, the results show that there is not necessarily a single best classification algorithm, but that the best performing algorithm will depend on the characteristics of the dataset to analyze. Indeed we notice that all the algorithms produce their best precision on the DT1, DT3, and DT5 data sets. These values, which reach 97% at times, outperform the Rapid Diagnosis Test which is the standard diagnostic tool largely adopted in the healthcare system in Senegal. However, on these same datasets, the algorithms often present very low specificities, for example 0.05 on DT1. This shows that our best performing classifiers are only able to predict a single class: either the patient has malaria or he does not, but not in both spots. This is because the DT1 and DT3 datasets are very unbalanced. In fact in these datasets either the number of patients with malaria is greater than those who are not or the opposite is true. Furthermore, we note that on the DT2 and DT4 datasets all the algorithms present specificities and Sensivity that are significant and quite similar. Contrary to what is quoted a little above, on these datasets the algorithms are efficient on the prediction tasks of the two classes. Looking closely at the results in terms of precision, recall and F-measure we observe that the classifiers RF, LR,

Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
Decision Tree						
DT1	0.97	1	0.98	0.78	97.04	0.05
DT2	0.59	0.48	0.48	0.64	63.01	0.80
DT3	0.89	0.85	0.87	0.86	80.86	0.69
DT4	0.68	0.57	0.62	0.70	65.60	0.74
DT5	0.99	0.84	0.91	0.76	83.41	0.58
Random forest						
DT1	0.97	1	0.99	0.81	97.13	0.07
DT2	0.63	0.34	0.44	0.64	63.33	0.85
DT3	0.89	0.85	0.87	087	80.86	0.70
DT4	0.68	0.56	0.62	0.70	65.82	0.74
DT5	0.99	0.84	0.91	0.76	78.35	0.60
Logistic Regression						
DT1	0.97	1	0.99	0.79	97.19	0.05
DT2	0.58	0.36	0.44	0.63	61.96	0.81
DT3	0.85	0.88	0.86	0.86	79.59	0.55
DT4	0.98	0.56	0.92	0.70	65.82	0.72
DT5	0.90	0.78	0.88	0.84	81.86	0.75
NAive Bays						
DT1	0.97	1	0.99	0.81	97.13	0.00
DT2	0.60	0.34	0.43	0.63	62.86	0.83
DT3	0.86	0.87	0.86	0.85	79.94	0.60
DT4	0.68	0.59	0.63	0.70	65.63	0.73
DT5	0.99	0.82	0.90	0.84	85.61	0.71
Support Vector Machine						
DT1	0.97	1	0.99	0.84	97.13	0.00
DT2	0.58	0.05	0.09	0.62	62.86	0.97
DT3	0.57	0.86	0.86	0.85	79.94	0.64
DT4	0.68	0.58	0.62	0.70	65.63	0.73
DT5	0.99	0.86	0.92	0.80	85.61	0.62
Artificial Neural Network						
DT1	0.97	1	0.99	0.84	97.15	0.04
DT2	0.59	0.40	0.48	0.65	62.86	0.80
DT3	0.89	0.85	0.87	0.87	86.68	0.69
DT4	0.68	0.58	0.62	0.70	0.70	0.75
DT5	0.99	0.84	0.91	0.79	83.26	0.65

Table 3. Performances measures of our classifiers over all datasets

SVM and ANN generally outperform the others for each dataset. Indeed, for the dataset DT1, which contains observations on patients living in different regions of Senegal, these four classifiers have an accuracy of 99%, a recall greater than 92% and an F-measure greater than 95%. We note the same trend with the DT2 dataset which contains observations on patients living in the same area in Senegal. It can also be noted that RF, LR, SVM and ANN have better precision than the rapid diagnostic test carried out and systematically used in the majority of health structures

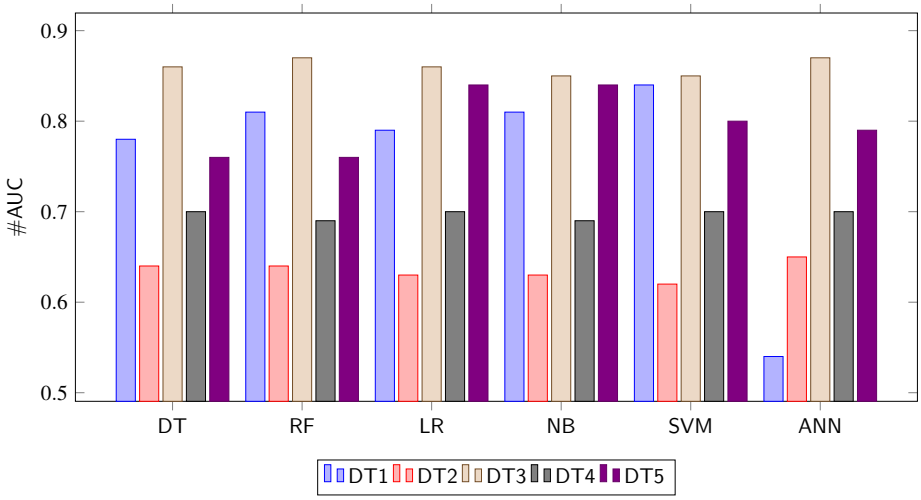

Figure 2. Comparison of the ROC Curves of the classifiers on differents datasets

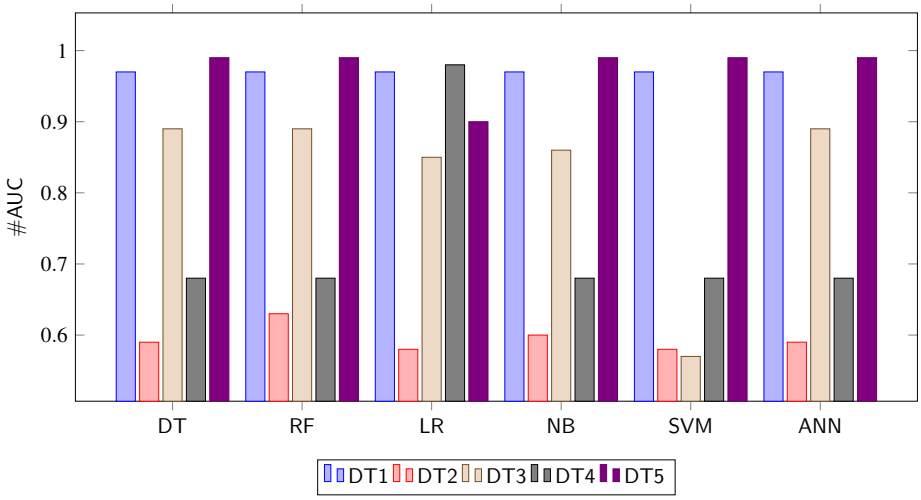

Figure 3. Precision values of compared classifiers on different datasets

in Senegal. This observation remains true with DT4 which is a perfectly balanced dataset. In conclusion, it is very difficult or even impossible for us to say definitively which algorithm is more efficient for the task of predicting malaria, but the choice of this one will strongly depend on the choice of the data set. However, this study shows that our classification problem has been taken care of. A method integrating several models and various datasets is necessary

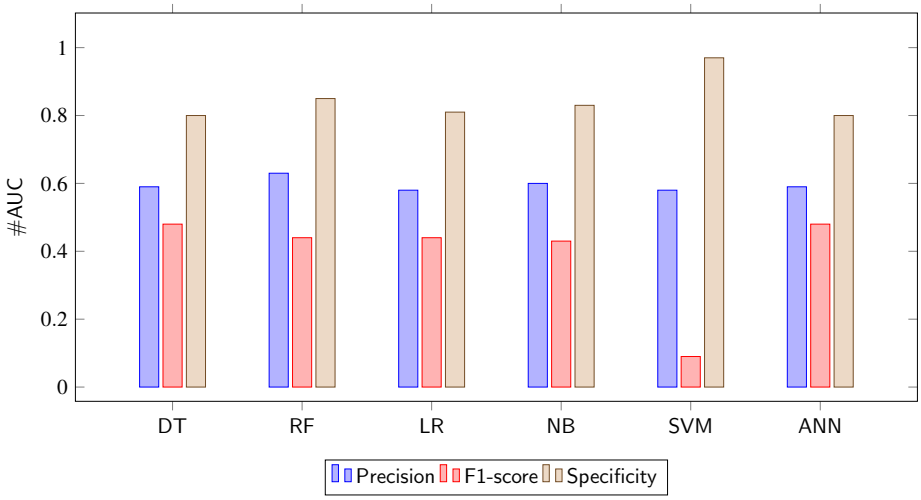

Figure 4. Precison, F1-score, specificity values of the classifiers on DT1

3. Conclusion

In this study, six classifiers using a wide variety of operating procedures have been extensively tested and compared over real world health datasets in order to evaluate their performance for the task of predicting the occurrence or not of Malaria in a patient knowing his signs and symptoms. The results obtained show that the algorithms RF, LR, SVM with Gaussian kernel and ANN present the best performances in predicting the occurrence or not of Malaria. In addition those four algorithms outperform the Rapid Diagnosis Test which is the standard diagnostic tool largely adopted in the health system in Senegal. This research has indicated that in practice there is no single best classification tool, but instead the best technique will on the characteristics of the dataset to be analysed. Future work consists in the study and the implementation of an ensemble method for predicting the occurrence or not of malaria based on the classifiers offering the best performances in our present study. But also to compare these performances with the ensemble methods for their validation

References