iiiiiii
HEAD
=======
¿¿¿¿¿¿¿
7c1cec024806f0122dbd3ebada2b3244dc9de7ab
iiiiiii
HEAD
=======

¿¿¿¿¿¿¿
7c1cec024806f0122dbd3ebada2b3244dc9de7ab

iiiiiii HEAD

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 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.

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.

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. Methodology

2.2. Machine Learning algorithms

In the following we discuss about some of these methods. Those algorithms are chosen among the most used ones in the health field according to studies[**?**,**?**].
Decision tree (DT)[**?**] is a supervised classifier which is obtained by recursively partitioning the labelled set of observations. It is one of the most adopted classifiers, thanks to its simplicity and its straightforward interpretation. For CART algorithms, hyperparameters are the impurity criteria (entropy and gini), the maximum depth, the minimum samples to split and the minimum samples at a leaf
Random Forest (RF)[**?**] is an ensemble approach built upon many decision tree classifiers. It is a supervised classifier which requires the same hyper parameters as DT, plus the number of trees to create and the random number of features to look at when splitting the labelled data during the training step [**?**].
Naive Bayes classifier (NB) [**?**] is a*supervised* machine learning algorithm, i.e. requires to be trained, used for classifying observations to given distinct classes based on *input explanatory variables* (a.k.a feature or attribute). It is a classification technique based on the well-known *Bayes' theorem*[2] with strong and naive assumptions. It simplifies learning by assuming that features are independent of given class.
Logistic regression (LR) [**?**] is a statistical model used in the machine learning domain as a supervised classifier for binary classification [**?**]. It is based, in its basic form,

---

[2]https://en.wikipedia.org/wiki/Bayes%27_theorem

on a logistic function to describe a binary dependent variable[**?**,**?**] by considering as input qualitative or/and ordinal explanatory variables in order to measure the probability of a given class label. The greatest advantage of the logistic regression classifier is the fact that you can use continuous explanatory variables and it is easier to handle more than two explanatory variables simultaneously and its ability to quantify the strength of the relationship between each explicative variable and the variable to explain, given the other variables integrated to the model.

Support Vector Machine (SVM) [**?**] is a supervised classification approach whose intuition is to represent input data in a space and to determine the optimal hyperplane that divides that space in two regions depending on the targeted value.

An Artificial Neural Network (ANN) [**?**] is a computational approach also referred to as a Connectionist System used in Machine Learning. ANNs are loosely modeled after the biological neural network in an attempt to replicate the way in which we learn as humans. Think of it as a computing system, structured as a series of layers, each layer consisting of one or several neurons. The types of the layers comprise *input*, *output* and *hidden* layers [**?**,**?**].

3. Results of the experiments

3.1. 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 [3] where the prevalence of Malaria is very high and collected patient records from its different health structures.

Data features. Table 3 contains the main characteristic of each dataset. 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 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

---

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

Diagnosis Test is around 90% for both datasets, meaning that the systematically performed RDT in Senegal is not fully reliable.

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

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.

3.2. 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.

3.3. Results

Table 2 presents the results of the experiments with the different algorithms on our data on Malaria. More specifically, Table 2 contains the precision, the recall, the
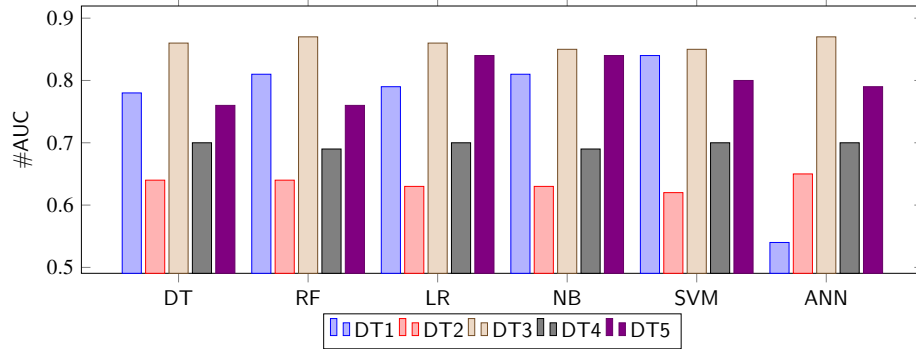
ML ALgorithms	Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
Decision Tree	DT1	0.97	1	0.98	0.78	97.04	0.05
	DT2	0.59	0.48	0.48	0.64	63.01	0.80
	DT3	0.89	0.85	0.87	0.86	80.86	0.69
	DT4	0.68	0.57	0.62	0.70	65.60	0.74
	DT5	0.99	0.84	0.91	0.76	83.41	0.58
Random Forest	DT1	0.97	1	0.99	0.81	97.13	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
	0.99	0.82	0.90	0.84	85.61	0.71	0.71
Support V 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 N 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 2. Performances measures of our classifiers over all datasets

specificity, the AUC mesure, the score and F-measure of each algorithm tested while Figure 1 shows their respective ROC curve.

3.4. 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,
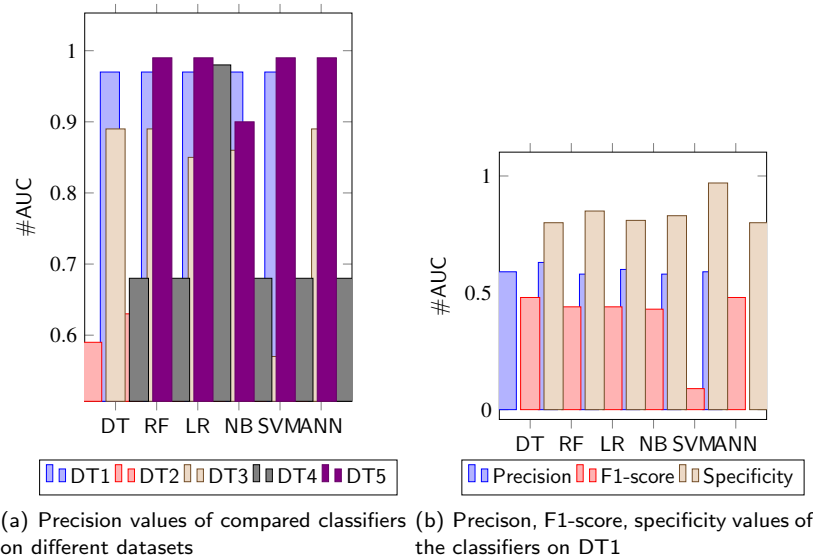
Figure 1. Comparison of the ROC Curves of the classifiers on differents datasets

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, 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 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

4. 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 pre-

(a) Precision values of compared classifiers on different datasets

(b) Precison, F1-score, specificity values of the classifiers on DT1

dicting 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

======= tikz

On the efficiency of Machine Learning Models in predicting Malaria using Symptôm Data

Ousseynou MBAYE, Mouhamadou Lamine BA [4] and Alassane SY

LIMA, Université Alioune Diop, Bambey, Senegal

---

[4]Corresponding Author: Mouhamadou Lamine BA, LIMA, Université Alioune Diop, BP.3400 Bambey, Senegal; E-mail: mouhamadoulamine.ba@uadb.edu.sn.

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. electronic camera-ready manuscript, IOS Press, LATEX, book, layout

5. Introduction

Malaria, also known as *"fiévre 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 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 8 we show that the precision of the RDT is about 90% for the real datasets used in this study. The Liverpool Model on Malaria (LMM) extended in [**?**] is an example of mathematical model that tries to model the parasite life cycle. It simulates the spread of Malaria at a daily resolution using the average daily temperature and the accumulated precipitation over 10 days. the final goal is to build a climate- or weather-driven Malaria model, allowing for a better understanding of Malaria transmission dynamics.

Clearly LMM is not a diagnostic system. We defer the reader to Section 6 for an exhaustive review of the literature. Despite existing works, the accuracy of Malaria prediction is still a concern: used mechanisms, e.g. domain knowledge and RDT, in Senegal are error-prone.

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. We first review the literature of existing research works dealing with Malaria in Section 6. In Section 8, we then provide a detailed description of our two real world datasets which contain medical records about patients living in Senegal. More precisely, we present the characteristics of each dataset, their imputation to deal with missing values and the precision values of RDT. In the sequel, we briefly describe in Section 7 the six ML algorithms that are evaluated and compared in this study. We present our experimentation setting, considered performance measures and discuss about the results of the experiments in Section 9. Finally, we conclude this paper in Section 10.

6. relate dwork

We mainly provide a brief overview of Machine Learning and Deep Learning models for healthcare applications.

In healthcare, Machine Learning apps can help better understand each patient's care journey, medical decisions, or the impact of new drugs. The survey of [**?**] explores the usefulness of various data mining techniques such as classification, grouping, association, regression in the health field. This survey also highlights the applications, challenges and future issues of data mining in healthcare. The recommendation regarding the appropriate choice of available data mining technique is also discussed in this article. The authors of [2] explain the fundamental principles of logistic regression and the stages of its application. Using two examples (the quality of follow-up care for diabetics and hospital mortality after acute myocardial infarction), they demonstrate the value that this statistical tool can have in studies carried out by the medical service of the national health fund, especially in studies aimed at evaluating professional practice. Another example of previous work that has used logistic regression is that of Farida et al. [**?**]. The logistic regression is exploited there for the selection of features in order to construct stable decision trees. The decision trees are then used to predict the severity criteria of Malaria in the context of Afghanistan. In [18] Uddin et al, provide a broad overview of the relative performance of different variants of supervised machine learning algorithms for disease prediction. Thus, their results showed that the Support Vector Machine (SVM) algorithm is applied most frequently (in 29 studies) followed by the Naïve Bayes algorithm (in 23 studies). However, the Random Forest (RF) algorithm has shown comparatively higher accuracy. Of the 17 studies where it has been applied, RF has shown the highest accuracy in 9 of them, or 53%. This was followed by SVM which exceeded 41% of the studies it considered.[**?**] Presents a reference of 7 machine learning algorithms used on binary classification tasks and applied to hospital data. In the study [**?**] Data mining acts as a solution to many health problems and it is useful for predicting early stage heart disease. The Naive Bayes algorithm is one such data mining technique that helps predict heart disease in patients. Gharehchopogh et al. [**?**] explain the use of medical data mining in determining methods of medical operation. They show that the decision tree algorithm designed for this case study generates a correct prediction for more than 86% of test cases. Indeed, decision trees based approach has been proposed in Nigeria [**?**] to predict the occurrence of Malaria given diagnostic data. In the same line of works applying machine learning, in [**?**], Pranav et al. propose Malaria likelihood prediction model built on a deep reinforcement learning (RL) agent. Such a RL predicts the probability of a patient testing positive for Malaria using answers from questions about their household. In the presented approach the authors have also dealt with the problem of determining the right question to ask next as well as the length of the survey, dynamically.

6.1. Description des données

7. Review of evaluated ML algorithms

We mainly provide a brief overview of Machine Learning and Deep Learning models for healthcare applications.

In healthcare, Machine Learning apps can help better understand each patient's care journey, medical decisions, or the impact of new drugs. Today researchers are using machine learning algorithms in the diagnosis of several diseases such as diabetes, stroke, cancer, malaria and heart disease. In the following we discuss about some of these methods. Those algorithms are chosen among the most used ones in the health field according to studies[?,?].

Decision tree (DT)[?] is a supervised classifier which is obtained by recursively partitioning the labelled set of observations. It is one of the most adopted classifiers, thanks to its simplicity and its straightforward interpretation. For CART algorithms, hyperparameters are the impurity criteria (entropy and gini), the maximum depth, the minimum samples to split and the minimum samples at a leaf Decision tree algorithm has been applied in many medical tasks, for examples, in increasing quality of dermatologic diagnosis[11], predicting essential hypertension [12], and predicting cardiovascular disease [13], predict and diagnose of heart disease [14].Decision tree is one of the most popular tools for classification and prediction.

Random Forest (RF)[?] is an ensemble approach built upon many decision tree classifiers. It is a supervised classifier which requires the same hyper parameters as DT, plus the number of trees to create and the random number of features to look at when splitting the labelled data during the training step [?].

Naive Bayes classifier (NB) [?] is a*supervised* machine learning algorithm, i.e. requires to be trained, used for classifying observations to given distinct classes based on *input explanatory variables* (a.k.a feature or attribute). It is a classification technique based on the well-known *Bayes' theorem*[5] with strong and naive assumptions. It simplifies learning by assuming that features are independent of given class. The Bayesian classifier has been applied in many medical issues, for examples, in measuring quality of care in psychiatric emergencies [21], predicting and diagnosis heart disease [14]. and assisting diagnosis of breast cancer [22].

Logistic regression (LR) [?] is a statistical model used in the machine learning domain as a supervised classifier for binary classification [?]. It is based, in its basic form, on a logistic function to describe a binary dependent variable[?,?] by considering as input qualitative or/and ordinal explanatory variables in order to measure the probability of a given class label. The greatest advantage of the logistic regression classifier is the fact that you can use continuous explanatory variables and it is easier to handle more than two explanatory variables simultaneously and its ability to quantify the strength of the relationship between each explicative variable and the variable to explain, given the other variables integrated to the model. One the other hand, the logistic regression is one of the most used multi-valued models in epidemiology[23]. In such a context, the variable to explain is often the occurrence or not of an event like a disease and the explanatory variables, i.e. the features, are those that highly impact the occurrence of this event, i.e. variables assessing the exposure to a risk factor or a protective factor, or a variable representing the confusion factor. The logistic regression is applied to predict malaria in [26] and in identification of at-risk populations in public health research and outreach [27] and the results are are very relevant.

---

[5]https://en.wikipedia.org/wiki/Bayes%27_theorem

Support Vector Machine (SVM) [**?**] is a supervised classification approach whose intuition is to represent input data in a space and to determine the optimal hyperplane that divides that space in two regions depending on the targeted value. SVM is used to studied the diagnosis of coronary artery disease[28].

An Artificial Neural Network (ANN) [**?**] is a computational approach also referred to as a Connectionist System used in Machine Learning. ANNs are loosely modeled after the biological neural network in an attempt to replicate the way in which we learn as humans. Think of it as a computing system, structured as a series of layers, each layer consisting of one or several neurons. The types of the layers comprise *input*, *output* and *hidden* layers [**?**,**?**].

8. 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 [6] where the prevalence of Malaria is very high and collected patient records from its different health structures.

Data features.

Table 3 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

---

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

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 2 shows that the raw datasets come with missing values for some variables on given observations. To resolve the problem of unbalanced datasets and data messness, we followed a data preparation pipeline in order to fit our datasets into the good format for our experimentation; we discuss about such a data preparation step next.

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 3. Raw Data characteristics
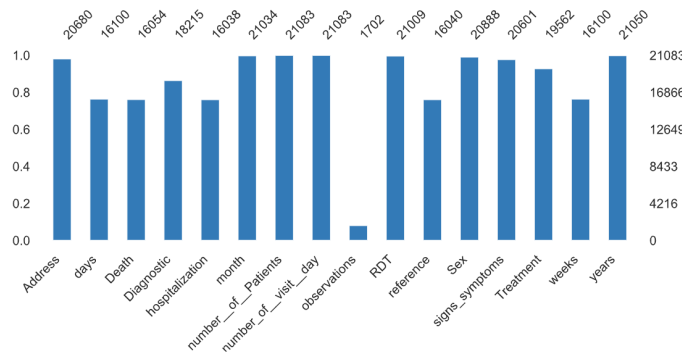
Figure 2. Proportion of missing values per variable

Data preparation.

We have followed the same process as in [**?**] in order to cleanse, normalize, impute, and balance information in our real datasets. Firstly, the raw datasets come with lot of inconsistencies due to the way the information were originally collected within the health structures. Indeed information about patients are manually recorded in the majority of health structures in Senegal. Second, when we did an explanatory analysis of the set of real-world data, they have revealed that the datasets were not balanced and come with missing values as mentioned above. We then used *OpenRefine*[7] to first clean and normalize information in our datasets. After that, we resolved our problem of missing values and unbalanced datasets by respectively using a K-Nearest Neighbours based imputation algorithm and an oversampling of the minority class: for more details we defer the reader to [**?**]. Figure 4 summarizes the new characteristics of DT1 and DT2 after the data preparation step.

---

[7]https://openrefine.org/

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 4. 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.

9. Experimentation and results

We detail and analyze in the section the results of the experimentation we performed using the six ML algorithms presented in Section 7 over the two real datasets described in Section 8. We start by presenting our experimentation setting.

9.1. 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.

9.2. Results of the experiments

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

9.2.1. Decision Tree

Table 1 below shows the performance measures (precision, recall, F measure and precision) of the results of our Decision Tree classifier after experimentation on all our datasets. The observation shows that the best scores of our classifier are achieved on the datasets DT1, DT3 and DT5 which are 97.04%, 80.86% and 83.41% respectively. Also AUC (Area Under the Curve) values are higher for these same datasets which are 0.78, 0.86 and 0.76 respectively. However, we note that the sensitivity values are higher than the specificity values on the datasets DT1, DT3 and DT5 whereas they are substantially lower than thoses of datasets DT2 and DT4. This means that DT is more inclined to predict as well whether a given patient has malaria or he doesn't, on the datasets DT2, DT3 and DT4, while our classifier on the datasets DT1 and DT5 our classifier is only efficient in predicting whether a given patient has malaria. This same trend is observed on the F-scores which higher values varying between 0.91 and 0.98 on the datasets DT1, DT3 and DT5.

Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
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

Table 5. Performances measures of DT over all datasets

9.2.2. Random Forest

The performance of the random forest varied throughout the study depending on the dataset, although overall it performed well as shown in Table 2. Notice that best accuracy are achieved by random forest classifier on the datasets DT1, DT2 and DT5 which are respectively 97.13%, 80.86% and 78.35%. In contrast with the results obtained with the DT classifier, the Sensivity values are higher than specificity values on datasets DT1 and DT5 whereas the inverse is noticed on the dataset DT3. At the same time we note that these values are roughly identical on the datasets DT3 and D4

9.2.3. Logistic Regression

Logistic regression In table 3 we show the performance measures LR classifier experimented on our five datasets. We notice that our classifier have overall precision which are vary between 58% and 98%.

We observe that the higher precision is obtained with DT4 dataset while the corresponding score is equal to 65.82% is the lowest of all other datasets. Also we

Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
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

Table 6. Performances measures of RF over all datasets

Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
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

Table 7. Performances measures of LR over all datasets

notice that the LR presents homogeneous results on the DT3 dataset with an accuracy of 85%, a sensitivity equal to 88%, an F-score of 92%, an AUC which is 0.86 and a score equal to 79.59%. . We also note that the best AUC and the best F-score are obtained by LR on the DT3 dataset.

9.2.4. Naives Bayes

In contrast with the results above, NB classifier presents very heterogeneous performances regarding the performance measures used as shown in Table 4. In fact, we observe that the best precision is achieved on the dataset DT5 which is 99%, although the best F-score and the higher accuracy are obtained on the dataset DT1 which are 0.99 and 97.13% respectively and finally the best AUC is observed on the dataset DT3 which is 0.85. We also note that the best specificity is obtained on DT4 and varies between 0.65 and 0.70 (see appendices).

Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
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

Table 8. Performances measures of NB over all datasets

9.2.5. Support Vector Machine

Table 5 shows the performance measures of the SVM classifier. Table 1 shows the performance measures. The observation shows that the best score, precision and F1-score are obtained on the datasets DT1, DT3 and DT5. However the higher AUC and the best specificity are observer on the datasets DT1, DT3 and DT4.

Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
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

Table 9. Performances measures of SVM over all datasets
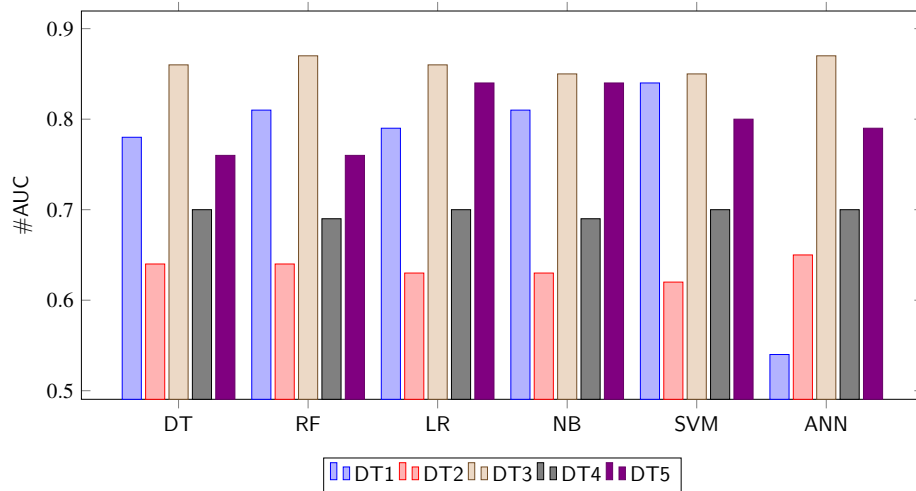
9.2.6. Artificial Neural Nework

The performance of the ANN varied throughout the study depending on the dataset, although overall it performed well. A large amount of initial effort was required to train and validate the model. Notice that best precision are achieved by ANN classifier on the datasets DT1, DT3 and DT5 which are respectively 97%, 89% and 99%. While the higher AUC and the best scores are obtained on the datasets DT1 and DT3. The Sensivity values are higher than specificity values on datasets

Datasets	Precision	Recall	F1-score	AUC	Score	Specificity
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 10. Performances measures of ANN over all datasets

9.3. 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

Figure 3. Comparison of the ROC Curves of the classifiers on differents datasets

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, 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 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

10. 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
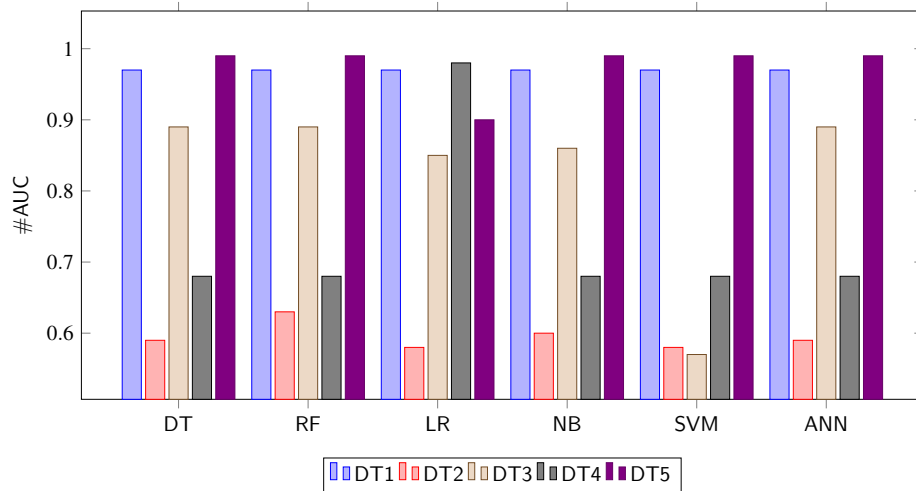
Figure 4. Precision values of compared classifiers on different datasets
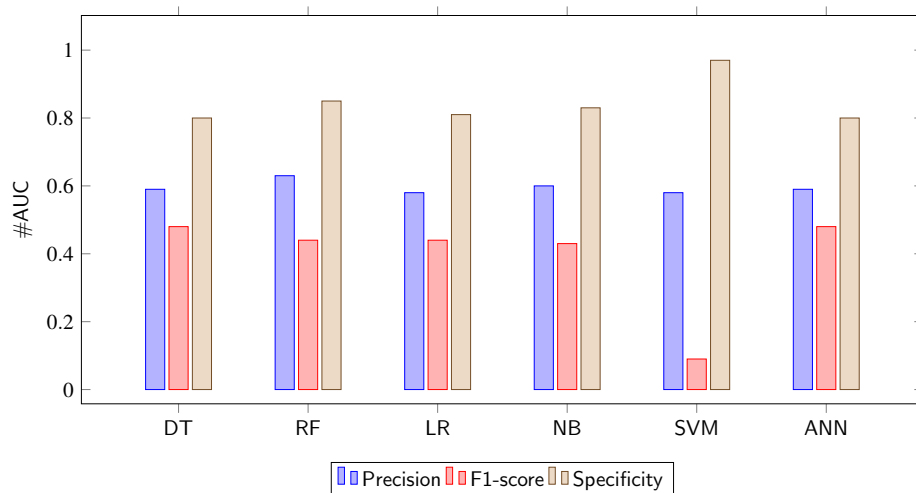
Figure 5. Precison, F1-score, specificity values of the classifiers on DT1

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

¿¿¿¿¿¿¿ 7c1cec024806f0122dbd3ebada2b3244dc9de7ab