TOWARDS AN EFFICIENT PREDICTION MODEL OF MALARIA CASES IN SENEGAL

VERS UN MODELE DE PREVISON EFFICACE DES CAS DE PALUDISME AU SENEGAL

Ousseynou Mbaye, Mouhamadou lamine Ba, Gaoussou Camara, Alassane Sy
UADB
{ousseynou.mbaye, mouhamadoulamine.ba, gaoussou.camara, alassane.sy}@uadb.edu.sn

Résumé

Le paludisme est l'une des maladies les plus mortelles dans le monde plus particulièrement en Afrique subsaharien. La situation est critique dans des pays comme le Senegal à cause du manque de services de santé de qualité et de personnel médical qualifie et capable de faire des diagnostics précis des maladies dont souffrent les patients.
Ainsi la nécessité de trouver des outils automatisées pour aider les acteurs médicaux dans le processus de prise de décision après avoir réalisé le diagnostic.
Dans ce papier nous proposons les premières étapes vers la réalisation d'un algorithme de diagnostic du paludisme basé sur les signes et symptômes du patient en plus du TDR (Test de Diagnostic Rapide). Notre modèle de prédiction est basée sur la régression logistique.
Les résultats des premiers tests sur un jeu de données réel et semi-synthétique de patients se sont très prometteurs concernant l'efficacité de l'approche proposée.

Mots clés : Diagnostic, Paludisme, Modèle de Prédiction, Imputation des données

1 INTRODUCTION

Le paludisme est l'une des maladies les plus meurtrières au monde, en particulier dans les pays d'Afrique subsaharienne tels que le Sénégal. Le paludisme est causé par des microorganismes unicellulaires parasites appartenant au groupe Plasmodium; c'est une maladie infectieuse transmise à l'homme par les piqûres de moustiques Anophèles femelles infectées. Une personne atteinte de paludisme peut présenter des symptômes tels que : fièvre, fatigue, vomissements et maux de tête. Dans sa forme sévère, la maladie peut causer une peau jaune, des convulsions, le coma et la mort.

Problématique et Motivations

Selon le dernier rapport [25] sur le paludisme dans le monde, publié en novembre 2017 par l'OMS, il y a eu 216 millions de cas de paludisme en 2016. On constate une hausse significative comparé aux 211 millions de cas enregistré en 2015.
On estime à 445 000 le nombre de décès dus au paludisme en 2016, un chiffre similaire à celui de l'année précédente (446 000) malgré les efforts colossaux consentis par les états et organismes non gouvernementaux dans le cadre de l'amélioration des services de santé et des stratégies de sensibilisation en particulier dans les régions les plus affectées.
Une analyse profonde des statistiques ci-dessus montre le paludisme demeure toujours un facteur majeur de mortalité en Afrique. En effet la région OMS de l'Afrique supporte une part disproportionnée de la charge mondiale du paludisme. En 2016, 90% des cas de paludisme et 91% des décès dus à cette maladie sont survenus dans cette région. Plus précisément 80% de la charge de morbidité due au paludisme pesaient sur une quinzaine de pays tous situés en Afrique

subsaharienne, sauf l'Inde. Cela démontre que le paludisme reste un véritable fléau dans les pays d'Afrique subsaharienne et que le Sénégal n'est pas du tout épargné. Dans cette étude nous proposons une approche efficace pour prédire le paludisme en utilisant le machine learning (apprentissage machine) lors qu'un patient vient se faire consulte. A partir des signes et des symptômes du patient mais aussi du Test de Diagnostic Rapide (TDR) notre solution permettra de dire avec précision si un patient souffre du paludisme ou non.

Le paludisme est un grave problème au Sénégal, principalement en raison du manque de services soins de santé de qualité et un personnel médical bien formé, capable d'effectuer un diagnostic précis des maladies dont souffrent les patients.

Au cours des dernières années, le gouvernement, avec l'aide de la communauté internationale et des organisations non gouvernementale a essayé d'éradiquer le paludisme en mettant en œuvre divers mécanismes proactifs et des solutions réactives pour combler le fossé en termes de services de soins de santé et de ressources humaines. Cependant, le taux de mortalité reste encore très élevé, par exemple dans les zones mal desservies, zones sans soins de santé requis, zone ou la population n'est pas alphabétisée mais aussi dans les zones ou la population a un faible revenu, etc. La plupart de cas décès de paludisme signalés sont dus à un diagnostic inexact, parfois incomplet du type exact de paludisme. En revanche, la survenue du paludisme ou sa complication est souvent notée lors d'événements populaires (par exemple des événements religieux tels que comme le Grand Magal de Touba [19]) qui rassemble des milliers de personnes de partout dans le pays pendant une courte période. Au cours de ces événements populaires, les points de santé temporaires sont mis en place pour assister et soigner les personnes malades. Le personnel soignant est parfois uniquement composé de volontaires sans compétences médicales avancées.

Chacun de ces points médicaux peut recevoir et traiter des centaines de patients chaque jour dont certains d'entre eux sont potentiellement atteints de paludisme.

Ainsi la nécessité de trouver des outils automatisées pour aider les acteurs médicaux dans le processus de prise de décision après avoir réalisé le diagnostic.

Approche Proposée.

Dans cet article, nous présentons les premiers paliers vers une manière efficace de diagnostic automatique du paludisme ou non en fonction des signes et symptômes du patient et les résultats du Test de Diagnostic Rapide(TDR). Nous considérons ce problème de diagnostic comme un problème classique de classification binaire en considérant deux classes: ''Paludisme'' et ''Non-Paludisme''. Ainsi connaissant les données d'un patient, notre objectif principal est de trouver correctement à quelle classe appartient le patient. Pour résoudre ce problème de classification, nous utilisons l'apprentissage automatique avec la régression logistique comme algorithme de base de notre modèle de prévision. L'apprentissage automatique a été largement utilisé dans plusieurs domaines (par exemple, l'informatique médicale [8]) à des fins diverses, tandis que la régression logistique a démontré son efficacité pour traiter des problèmes de classification binaire tel que le nôtre. Pour le cas de notre étude nous nous intéressons à la prévision du paludisme au Senegal. Pour cela, nous utilisons un grand volume de jeux de données d'enregistrements de patients recueillies pendant le grand Magal de Touba ; l'un des plus grands événements religieux populaire au Sénégal. Les des différents points de santé installés à cet occasion, à savoir plus de vingt points reçoivent des centaines de patients. Pour les premiers tests de notre algorithme, nous introduisons une méthode de préparation de données afin (i) d'explorer l'ensemble de données pour une bonne labélisation; (ii) pour conserver uniquement les attributs liés au paludisme; (iii) Nettoyer et transformer les attributs, pour obtenir un jeu de données numérique composé uniquement que des attributs du paludisme; et (iv) imputer les valeurs manquantes (il y avait beaucoup de valeurs manquantes dans l' ensemble de données collectées comme le montre la section 3).

La préparation de données a été réalisé en utilisant OpenRefine (Google Refine) pour effectuer divers tâches de nettoyage, de traitement, de profilages de l'ensemble des données brutes de patients mais l'algorithme missForest qui est outil robuste pour imputer les données manquantes de différents types; voir la section 3 pour plus de détails. Les résultats des premiers tests sur un jeu de données réel et semi-synthétique de patients se sont très prometteurs concernant l'efficacité de l'approche proposée.

Plan de l'article. Le reste de l'article est organisé comme suit. Nous résumons les travaux correspondants à l'imputation des données et les méthodes de classification binaire dans la section 2. Dans la section 3, nous introduisons un pipeline de préparation de données brutes enregistrées sur les patients recueillies pour la phase de prédiction. Nous présentons ensuite notre modèle de prédiction pour les cas de paludisme dans la section 4 . Des expériences et analyse de performance de données réelles collectées, ainsi qu'un ensemble de données semi-synthétiques, sont détaillées dans la section 5 avant de conclure dans la section 6.

2. Revue de la littérature

Dans cette section, nous résumons l'état de la recherche sur le paludisme en général, et en particulier l'utilisation des techniques d'apprentissage automatique pour aborder les différents aspects liés à l'un des principaux problèmes de santé dans le monde, notamment le paludisme

Comme on le sait bien, le paludisme est causé par la piqure de l'anophèle femelle, dont la plus dangereuse espèce est le Plasmodium falciparum. De nombreux travaux préliminaires ont été ensuite consacrés à l'étude de l'évolution et de la répartition du moustique responsable, principalement dans le but de détecter ou de diagnostiquer la gravité de la maladie par rapport à un patient infecté donné [10 , 5]. Les recherches récentes sur le paludisme ont largement adopté le machine learning et ont démontré sa capacité à résoudre divers aspects de la maladie. La plupart de ces techniques basées sur le machine learning reposent sur l'analyse des données sanguines obtenues à partir de captures d'écran microscopiques de haute définition comme dans [12]. Les auteurs dans [12] proposent un algorithme d'apprentissage non supervisé qui détecte et détermine les types de cellules sanguines infectées. L'approche de prédiction utilisée consiste à quantifier la quantité de parasite plasmodium dans un frottis sanguin. Dans la même intuition de recherche d'exploitation du sang. 'The Jordan-Elman neural networks classifier' introduit dans [7] ,permet rapidement déterminer l'occurrence du paludisme et son niveau de gravité également. Elle est basée sur une analyse des caractéristiques des données sanguine des patients par le réseau de neurones. Toujours en utilisant le machine learning, DIAZ et al. ont proposé dans [9] un algorithme semi-supervisé permettant de quantifier et de classer les érythrocytes infectés par les parasites du paludisme à travers des images microscopiques. L'originalité de cette méthode vient de son efficacité même en présence de fines pellicules de sang infecté par le Plasmodium falciparum pour la quantification et de classification des parasites plasmodium infectés. Outre les données sanguines, des enregistrements de signes et de symptômes des patients ont également été utilisés pour étudier le paludisme avec les méthodes de machine learning. En effet, une approche basée sur les arbres de décision a été proposée au Nigeria [23] pour prédire la survenue du paludisme à partir des données de diagnostic. Cependant, un arbre de décision souffre de diverse limite en tant que classificateur. En effet**,** il peut facilement sur-adapter ou peut être extrêmement sensible aux petites variations dans les données. Quand bien même nous nous appuyons et sur les signes et sur les symptômes, le modèle de prédiction dans [23] diffère du nôtre sur de nombreuses facettes: notre modèle est construit sur la régression logistique et est entrainé en utilisant également les informations du test de diagnostic rapide. De plus, nous appliquons notre méthode dans le contexte de patients vivant au Sénégal. Un exemple du travail cité précédemment et qui a utilisé la régression logistique est celui de Farida et al. dans [3]. La régression logistique y est utilisée pour la sélection des attributs afin de construire des arbres de décision stables. Les arbres de décision sont ensuite utilisés pour prédire les critères de gravité du paludisme dans le contexte afghan.

Dans la lignée des travaux appliquant l'apprentissage automatique, dans [16] , Pranav et al. proposent un agent d'apprentissage par Reinforcement Learning (RL) capable de

prédire la probabilité qu'un individu présente un résultat positif au test du paludisme en posant des questions sur leur ménage. Cet agent est un Deep Q-network RL qui apprend une politique directement à partir des réponses aux questions, avec une action définie pour chaque question de sondage possible et pour chaque classe de prédiction possible. En outre, une classification fondée sur des règles statistiques améliorées et permettant de diagnostiquer le paludisme a été proposé dans [6].Un prototype correspondant intégrant les règles et les modèles statistiques a été mis en place; L'objectif principal de l'étude était de développer un prototype statistique permettant de réaliser un diagnostic clinique du paludisme, compte tenu de ses effets indésirables sur l'ensemble des soins de santé. Toutefois, son traitement reste très coûteux pour la majorité des patients.
À notre connaissance, il s'agit du premier travail au Sénégal qui tente de fournir un modèle de prédiction du paludisme à partir des données des patients.

3 Préparation des données

Dans cette section, nous détaillons le processus de préparation des données suivi pour l'obtention d'un jeu de données sur le paludisme pour la phase de prédiction. Nous commençons par présenter les techniques de nettoyages et de normalisation des données utilisées.

3.1 Nettoyage et labélisation des données

Dans le but de mettre en place un modèle de prédiction efficace des cas de paludisme au Sénégal, nous nous sommes appuyés sur un jeu de données de patients réel pour la validation. Le jeu de données a été extrait du Grand Magal de Touba de 2016 [19]. Environ 4-5 millions d'individus se réunissent chaque année dans la ville sainte de Touba, au Sénégal, lors de la cérémonie religieuse du Grand Magal. Plusieurs points de santé sont établis lors de cet événement religieux; chaque point reçoit chaque jour des centaines de patients, dont certains atteints de paludisme. Les données de patients que nous utilisons ici ont été recueillies manuellement à partir des registres de ces points puisqu' aucun système de gestion électronique de la santé n'existe. En détail, l'ensemble des données comprend des milliers d'enregistrements de patients comportant chacun 16 caractéristiques. Certaines de ces caractéristiques (également appelées fonctionnalités) comprennent des données personnelles sur le patient, mais aussi les signes et les symptômes du patient signalés par le médecin qui a pris le patient en charge. Les autres attributs décrivent des données cliniques telles que des informations sur le diagnostic final du médecin (la maladie dont souffre le patient), le résultat du test de diagnostic rapide (TDR) et le statut (c.-à-d. admission, décédé ou mis sous observation) du patient. Pour des raisons de confidentialité et certaines restrictions d'utilisation des données, nous avons ignoré les données personnelles sur le patient pendant ce travail. En raison du fait que les données des patients ont été collectées manuellement dans des registres, nous avons constaté de nombreuses incohérences telles que fautes d'orthographe, mêmes valeurs d'attributs avec des écritures différentes (par exemple, «DIARRHEE INFECTIEUSE "et" INFECTIEUSE DIARRHE "), et des attributs à valeurs multiples (par exemple la colonne signes et symptôme). Nous utilisons le logiciel OpenRefine [13 , 1] pour d'abord nettoyer puis normaliser les valeurs dans l'ensemble des données des patients. OpenRefine (Google raffiné) est un puissant outil open source qui permet aux chercheurs ou aux scientifiques d'accomplir l'activité de criblage des données, c'est-à-dire de travailler avec des données en désordre: les nettoyer; les transformer d'un format à un autre; et les compléter avec des services Web et des données externes. Nous avons utilisé les méthodes suivantes fournies par OpenRefine pour prétraiter notre jeu de données brutes.

- **Text filter function:** le filtre de texte permet d'explorer les valeurs des attributs, de les nettoyer et d'identifier ceux qui peuvent avoir de nombreuses variantes.
- **Transform functions**: OpenRefine fournit deux fonctions de transformation différentes: les fonctions de transformations prédéfinies (preset transformations functions) pour la résolution de problèmes de mise en forme triviaux tels que le rognage des espace blancs et des fonctions de transformations avancées (Advanced transformations) basées sur le langage (GREL) de OpenRefine pour normaliser les données par lots ou les fractionner. Cette deuxième classe de formations est très utile, en particulier lorsque le nombre de valeurs de données à normaliser est très important

(faire la même tâche manuellement prendrait beaucoup de temps et serait sujette aux erreurs). Par exemple, GREL permet d'utiliser une expression régulière simple pour toutes les variantes d'un symptôme dans la colonne Symptôme.
- **Cluster and edit function:** l'option de groupage dans OpenRefine fournit également aux utilisateurs des méthodes pour fusionner et normaliser les variations dans l'ensemble de données. La puissance de cette fonction est qu'il est capable de détecter automatiquement les petites variations de données qui suivent un certain modèle.

En ce qui concerne le cas particulier des attributs à valeurs multiples tels que les colonnes symptôme et diagnostic de notre jeu de données brutes, nous les avons divisés en plusieurs valeurs dans des colonnes distinctes. En effet, dans le jeu de données brutes, des informations telles que les symptômes dont un patient donné souffre ont été stockés dans une seule colonne, séparée par le caractère spécial '+', par exemple "DOULEUR ARTICULAIRE "+" DOULEUR PELVIENNE "+" VOMISSEMENTS ".
Après cette étape de nettoyage et de normalisation des données, nous avons procédé à l'extraction des caractéristiques du paludisme.

3.2 Sélection des caractéristiques du paludisme

Pour bien étudier le paludisme, il faut disposer d'un ensemble de données d'un patient comprenant les principales caractéristiques de la maladie. Malheureusement, certaines de ces caractéristiques du paludisme n'étaient pas explicitement spécifiées dans notre jeu de données brutes. En conséquence, nous avons déduit douze nouveaux attributs qui décrivent mieux les signes et les symptômes du paludisme selon les experts du monde de la santé. Ces nouveaux attributs sont**: manque d'appétit, fatigue, fièvre, céphalalgie, nausée, arthralgie, troubles digestifs, vertiges, frissons, myalgie, diarrhée et douleurs abdominales**. Nous avons ensuite ajouté les nouveaux attributs à notre jeu de données et transformé ce dernière en remplissant la valeur de chaque nouvel attribut en fonction de la liste des signes et des symptômes signalés pour chaque patient.
Le Diagnostic est le résultat des signes et symptômes confirmé par un examen médical en général. Dans la colonne Diagnostic, on voit plusieurs conclusions différentes. Ainsi tout diagnostic autre que le paludisme est remplacé par la classe non paludisme
À cette étape de notre processus de préparation de données, nous avons créé un jeu de données patient contenant les caractéristiques requises pour le paludisme. Cependant, notre jeu de données n'était pas encore complet ni prêt à cause de l'absence de valeurs. Enfin, nous avons complété notre ensemble de données en utilisant une approche d'imputation des données robuste.

3.3 Imputation de données manquantes

Comme le montre le tableau 1, nous avons observé de nombreuses valeurs manquantes dans notre jeu de données, affectant ainsi la majorité des attributs de données. Ces valeurs manquantes ne doivent pas être ignorées car les données d'autant plus que l'exhaustivité et la qualité sont très importantes pour traiter un problème de prédiction; ceci pourrait avoir un impact négatif sur la précision de nos prédictions et devrait être traité de manière appropriée. Il faut noter que l'apprentissage automatique repose sur un ensemble de données complet. Les sources et les types de valeurs manquantes peuvent être variés [22] . Dans notre contexte, les manquements ne sont pas complètement aléatoires et peuvent être dus à une connaissance incomplète des données du patient, au fait que le personnel médical ne spécifie pas une valeur d'attribut lorsqu'elle n'est pas observée, ou à une difficulté pour les patients à décrire correctement certaines informations (liées par exemple aux signes ou aux symptômes de leurs maladies) au moment du diagnostic. Puisqu'ils pourraient avoir une certaine relation entre les valeurs d'attribut pour le même patient, voire une corrélation entre les dossiers des patients, nous décidons de résoudre notre problème de valeurs manquantes en utilisant des algorithmes d'imputation au lieu de choisir des valeurs arbitraires ou de supprimer des enregistrements avec valeurs manquantes.

L'imputation des données est souvent utilisée dans le domaine de l'apprentissage automatique pour traiter les erreurs d'informations. De nombreux algorithmes ont été proposés dans la littérature [18 , 22], dépendant de la nature de l'absence ou du type de données. MissForest [21] a été prouvé très efficace en présence de divers types de données simultanément comme dans notre cas (par exemple données numériques, chaîne, données catégorielles, etc.). L'algorithme missForest s'appuie sur RandomForest qui est une méthode de prédiction non paramétrique capable de traiter des données mixtes et qui permet des effets de régression interactifs et non linéaires. Un tel algorithme d'imputation vise à traiter tous jeu de données d'information en minimisant (dans la mesure du possible) les présomptions sur les aspects structurels des données. Étant donné un ensemble de données d'information, missForest résout le problème de données manquantes en utilisant un schéma d'imputation itérative aléatoire sur les valeurs observées dans une première étape, suivie de la prédiction des valeurs manquantes puis procédant de manière itérative jusqu'à la convergence.
Nous avons appliqué missForest à notre jeu de données de patient du paludisme obtenu du paludisme précédemment en utilisant le logiciel Python [2].
Nous étudierons et prouverons dans la section 5 précisions de la prédiction faite par l'algorithme à l'aide la NRMSE (normalized root mean squared Error).

4 Modèle de Prédiction

Pour prédire le paludisme à partir du jeu de données labélisé obtenue étant donné un nouveau patient, nous utilisons de régression logistique prédicteur. Dans cette section, nous rappelons brièvement les bases de la fonction de régression logistique. Puisque notre problème est un problème de classification binaire, nous commençons par introduire le problème de classification binaire que nous devons résoudre dans l'étude.

Attribute name	#missing_values
Lack of appetite	21068
digestive disorders	21062
Loss of weight	21017
Arthragia	20940
Chill	20925
Nausea	20874
Myalgia	20870
Tiredness	20713
Diarrhea	20481
Vomit	20051
Abdominal pain	19770
Dizziness	19628
Fever	18245
Temperature	17636
Arterial pressure	16924
Cephalalgia	15370
Diagnostic	2875
Quick diagnostic test	76

Table 1. The number of missing values per attribute

4.1 Classification Binaire

Supposons deux classes de diagnostic du paludisme: le paludisme et le non-paludisme. Nous considérons également P et C comme l'ensemble des patients et un modèle de prédiction. Un patient p dans P est défini par un ensemble de paires $(a_1, v_1), (a_2, v_2), ..., (a_n, v_n)$ où a_i et v_i, pour chaque $1 \leq i \leq n$, correspond respectivement à une caractéristique donnée du paludisme et à sa valeur numérique associée définie comme suit.

$$v_i = \begin{cases} 1 \text{ si un i est observé} \\ 0 \text{ sinon} \end{cases} \quad (1)$$

Définition 1. (Notre problème de prédiction) Nous définissons notre problème de classification binaire pour la prédiction de la présence ou non du paludisme sur un jeu de données de patients donné, comme une fonction C qui chaque patient p dans P associe une et une seule classe dans {Paludisme, Non-paludisme}.

Mathématiquement on le note C: P ↦ {Paludisme, non paludisme}.

Dans les cas particulier de notre étude prenons C comme étant la régression logistique

4.2 Régression logistique

La régression logistique est une méthode statistique pour effectuer des classifications binaires [17]. Elle prend en entrée des variables prédictives qualitatives et/ou ordinales (par exemple, la présence ou non de fièvre chez un patient donné) et mesure la probabilité de la valeur de sortie (par exemple, la présence ou pas du paludisme) en utilisant la fonction sigmoïde . La figure 1 montre la forme de la courbe de la fonction Sigmoïde

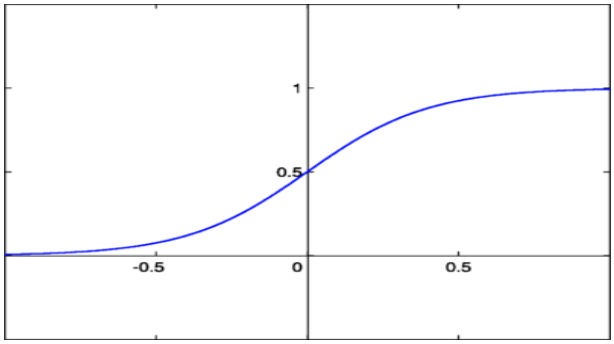


Figure 1 :

La régression logistique est un des modèles multivariables couramment utilisé en épidémiologie [4, 15] (c'est-à-dire l'étude de l'incidence, de la distribution et du contrôle possible des maladies et d'autres facteurs liés à des problèmes de santé pouvant affecter des groupes de population). Dans un tel contexte la variable dépendante est habituellement la survenue ou non d'un événement (maladie ou autre) et les variables indépendantes sont celles susceptibles d'influencer la survenue de cet événement c'est-à-dire les variables mesurant l'exposition à un facteur de risque ou à un facteur protecteur, ou variable représentant un facteur de confusion. L'intérêt majeur de cette technique est de quantifier la force de l'association entre chaque variable indépendante et la variable dépendante, en tenant compte de l'effet des autres variables intégrées dans le modèle[4].

Définition fonctionnelle de notre modèle de régression.

On supposera que la variable Y à laquelle on s'intéresse est la survenue ou non du paludisme, dont les deux catégories seront notées M + et M-.

Dans le cas particulier d'une seule variable X explicative (équivalent d'une régression simple), le modèle s'écrit :

$$\Pr(M+ \mid a) = \frac{e^{\alpha+\beta\times a}}{1+e^{\alpha+\beta\times a}} \qquad (2)$$

où les coefficients α et β sont les paramètres du modèle.

P R (M + | a) mesure la probabilité d'apparition du paludisme si la variable a est observé. La figure 1 représente la fonction logistique correspondante f (a). Encore une fois, l'intérêt principal de cette fonction réside dans la simplicité d'atteindre une estimation de la cote ratio (OR) qui mesure la force de l'association entre la maladie M et une variable d'exposition dans une analyse de régression. En effet, si l'exposition est codée en 0 (la variable n'est pas observée) et 1(la variable est observée) comme dans notre cas, le modèle permet d'arriver après simplification à OR = exp (β). Le coefficient β de la variable d'exposition dans le modèle logistique est donc le

logarithme de l'odds-ratio mesurant l'association entre cette variable (signe ou symptôme) et la maladie (paludisme), ce qui permet d'interpréter facilement les résultats d'une régression logistique.
L'extension vers un modèle à plusieurs variables (régression multiple) se fait très simplement comme le montre la formule ci-dessous

$$\Pr(M+ \mid a_1, a_2, \ldots, a_n) = \frac{e^{\alpha + \sum_{i=1}^{n} \beta_i \times a_i}}{1 + e^{\alpha + \sum_{i=1}^{n} \beta_i \times a_i}} \qquad (3)$$

À chaque variable X_i est associé un coefficient β_i et OR_i (mesurant l'association entre X_i et M_+) se calcule par exp (β_i).

Optimisation du modèle. La question qui se pose généralement lorsqu'on utilise une régression multiple consiste à savoir comment sélectionner l'ensemble minimal de variables parmi les a_i qui explique mieux la variable Y. Plusieurs stratégies d'optimisation sont possibles pour obtenir le meilleur modèle de prédiction finale qui prend en compte le maximum d'informations tout en restreignant autant que possible le nombre de variables explicatives afin de faciliter l'analyse des résultats. Les plus employées sont les procédures dites « pas à pas descendantes ou pas à pas ascendantes ». Les deux approches appliquent une régression itérative en incluant d'abord dans le modèle la variable qui présente le meilleur coefficient de détermination, puis en ajoutant la variable qui améliore ce coefficient et ainsi de suite pour les méthodes ou pas à pas ascendantes. Pour les méthodes ou pas à pas descendantes, l'ensemble des variables est considéré au début et les variables sont progressivement exclus du modèle, en fonction de ceux qui n'ont pas significativement amélioré le coefficient de détermination.

Dans la section suivante nous présentons les résultats de nos expérimentations obtenues en utilisant notre modèle de régression logistique sur des données réelles de patients

5. Expérimentation et Résultats

Dans cette section, nous présentons les performances de notre modèle de prédiction du paludisme à travers une analyse des résultats des expérimentations que nous avons menées sur des jeux de données du réelles de patients et un jeu de données semi-synthétique obtenue à partir du jeu de données réelles. Nous commençons par présenter les conditions d'expérimentation

5.1 Conditions d'expérimentations

Nous avons testé le modèle sur trois jeux de données différents en implémentant l'algorithme de la régression logistique avec le logiciel Python. Pour imputer les données manquantes, nous avons utilisé le package missForest du logiciel R est utilisé.

Nos jeux de données. Nous avons collecté et utilisé un jeu de données patient réels provenant de différents points de santé qui ont été définis lors du Grand Magal de Touba en 2016. Nous avons également généré et utilisé deux variantes de ce jeu de données de patient réels. La description des caractéristiques de notre jeu de données brutes de patients réels, ainsi que le processus de préparation des données que nous avions proposées pour nettoyer, normaliser et imputer les informations, sont données dans la section 3. Nous notons DT1 ce jeu de données.
La première variante, notée DT2 est obtenue en supprimant les enregistrements avec les attributs manquants dans DT1 au lieu d'utiliser un algorithme d'imputation qui prédit les valeurs des informations manquantes.

Une telle variante aidera à étudier l'impact de la suppression des enregistrements avec des valeurs manquantes dans l'exactitude de la prévision.
La deuxième variante, appelée DT3, est un jeu de données semi-synthétique qui a été généré en utilisant une stratégie de sur échantillonnage sur le jeu de données brutes TD1. En effet une analyse explicative effectuée sur le jeu de données DT1 a révélé que les données sont assez déséquilibrés, c'est-à-dire qu'il montre un déséquilibre important entre les classes; le nombre de patients atteints de paludisme était largement inférieur au nombre de patients qui ne souffrent pas de paludisme, comme illustré à la figure 5. 1. L'exploitation des approches d'échantillonnage peut permettre d'obtenir un jeu de données équilibré concernant les deux classes à prédire.
Pour cela nous avons implémenté l'algorithme SMOTE [24], avec le **package imblearn** [14]. SMOTE consiste à créer un échantillon de données semi synthétiques à partir de la valeur dépendante diagnostic au lieu de faire des copies des valeurs existantes. Ensuite, choisir de manière aléatoire, l'un des k plus proches voisins et l'utiliser pour créer de nouvelles observations similaires, mais au hasard.
Dossier donné afin de créer de nouvelles observations au hasard. Nous avons appliqué un sur échantillonnage de la classe minoritaire dans notre jeu de données patient pour générer un ensemble semi-synthétique de données DT3 contenant le même nombre d'enregistrements pour les deux classes.

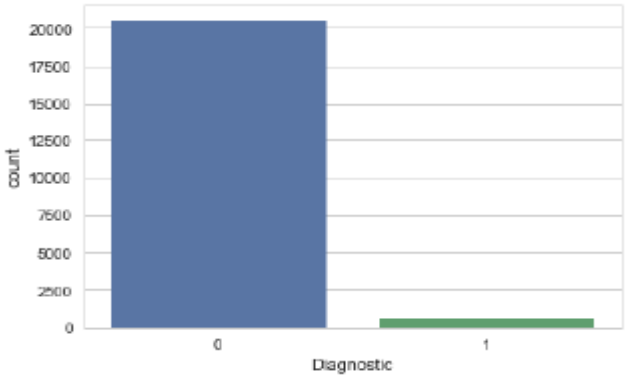


Fig. 2. The number of records by class

Paramètres du modèle de prédiction.
. Afin de mettre en place notre modèle de classification basée sur la régression logistique, la librairie sklearn[1] de Python. Ce package de python définit les paramètres par défaut de la régression logistique, ainsi que des stratégies d'optimisation, pour effectuer correctement la classification binaire en utilisant le meilleur modèle final. Pour les besoins de nos tests, nous avons utilisé les paramètres d'entrée de la régression logistique suivante:
- **random state:** modélise le l'état initial du générateur de nombres pseudo aléatoires à utiliser lors du mélange des données. Sa valeur est définie à 0 car nous n'avons pas besoin de mélanger les données dans notre expérimentation.
- **class_weight:** c'est le poids associés aux classes. Nous le réglons sur '**None**', c'est-à-dire que toutes les classes sont censées avoir le poids qui est égale à 1.
– dual. Il n'est mise en œuvre que pour les problèmes avec une pénalisation l2. Ce paramètre est défini sur **Faux** car le nombre d'échantillons est supérieur au nombre de fonctionnalités.
- **fit_intercept:** utile si une constante (ou biais) est ajoutée à la fonction de prédiction. Par conséquent, nous avons fixé l'interception d'ajustement à **Vrai**.
Intercept_scaling: ce paramètre, défini sur 1, n'est utile que lorsque le solveur «liblinear» est utilisé et **fit_intercept** est défini sur **Vrai**.
- **max_iter:** nombre maximum d'itérations prises pour que les solveurs convergent.
- **multi class:** si l'option choisie est **ovr**, alors un problème binaire est correct.

1 https://scikit-learn.org/stable/modules/generated/sklearn.linearmodel.LogisticRegression.html

- **n_jobs:** nombre de processeurs cpu utilisés lors de la parallélisation de classes si multi class = "ovr". Ce paramètre est ignoré lorsque le solveur est défini sur “liblinear” que «multiclass» soit spécifié ou non.

-Penalty: ce paramètre est utilisé pour spécifier la norme dans la pénalisation. Nous avons fixé la pénalité à sa valeur par défaut 1/2.
- **solver:** il permet de spécifier la stratégie utilisée pour résoudre l'optimisation sous-jacente de notre modèle. Il est fixé à liblinear.
- **tol:** tolérance pour le critère d'arrêt définie sur 0.0001.
- **verbose:** pour le liblinear solver, définissez verbose sur un nombre positif
- **warm_start:** lorsqu'il est défini sur **True**, réutilise la solution de l'appel précédent pour l'adapter à l'initialisation, sinon, effacez la solution précédente. Inutile pour le liblinear solver.
Comme la régression logistique effectue un apprentissage supervisé, nous avons utilisé 60% du jeu de données pour l’entrainement du modèle et 30% du jeu de données pour de test.

5.2 Les Mesures de performance.
Pour évaluer la performance de notre modèle de prédiction sur les différents jeux de données utilisés, nous avons calculé la précision, le rappel (ou la sensibilité),
la F-mesure et la spécificité des classes prédites. Nous avons également tracer la courbe ROC (Receiver operating Chacacteristic) de la régression logistique pour étudier sa forme. La sensibilité, la spécificité et la courbe ROC sont souvent utilisés dans le domaine de la médecine en tant que mesure de la performance pour l’évaluation des modèles de prédiction.
Precision. La précision p, ou taux de valeur positive, pour une classe est le nombre de vrais positifs (c’est-à-dire le nombre de cas correctement étiquetés comme appartenant à la classe M +) divisés par le nombre total de cas étiquetées comme appartenant à la classe M + (c’est-à-dire la somme des positifs et faux positifs). Les faux positifs sont les cas qui prédit dans vraies alors qu’ils sont faux.

$$p = \frac{\sum_{i=1}^{|R|} Entity(i)}{R}. \qquad (4)$$

Entity (i) est une fonction binaire qui renvoie **true** si la classe prédite pour le ième cas est correct et **false** sinon. R est la somme du nombre de vrai positifs et des faux positifs.
Recall (Rappel). Le rappel r (également appelé sensibilité) est défini comme le nombre de vrais positifs divisé par le nombre total de cas qui appartiennent réellement à la classe M + (c’est-à-dire la somme des vrais positifs et faux négatifs, qui sont des cas qui n'ont pas été étiquetés comme appartenant à la classe M + mais aurait dû être).

$$r = \frac{\sum_{i=1}^{|R|} Entity(i)}{G} \qquad (5)$$

G est la somme du nombre de vrais positifs prédits et du nombre de faux négatifs prédits.
F-measure.la F-measure notée F1, est une métrique qui mesure la précision d'un test en analyse statistique d'une classification binaire. Il est calculé en utilisant à la fois la précision p et le rappel r du test en tant que rapport du nombre de réponses positives et le nombre de tous les résultats positifs renvoyés par le classificateur.

$$F_1 = 2 \times \frac{p \times r}{p + r} \qquad (6)$$

Specificity (Spécificité). La spécificité, également connue sous le nom de taux négatif réel, mesure la proportion de réels négatifs correctement identifiés en tant que tels (par exemple, le pourcentage de personnes ne souffrant pas de paludisme et qui sont correctement identifiées comme n’ayant pas la maladie).
Receiver operating Chacacteristic. La courbe ROC montre la capacité de diagnostic d'un système de classificateur binaire dont le seuil de discrimination est

varié. La courbe ROC est obtenu en traçant le graphe vrai positif (par exemple, sensibilité ou rappel en apprentissage automatique) par rapport au taux de faux positifs (1- spécificité) à différents seuils.

5.3 Expériences et analyse des résultats

Pour chaque ensemble de données considéré, nous avons effectué deux types de tests avec notre modèle de prédiction: un test sans inclure le résultat du test de diagnostic rapide(TDR) et un autre test avec le résultat du test de diagnostic rapide (TDR) parmi les attributs d'entrée. Nous avons d'abord décrit ci-dessous les résultats obtenus pour chaque ensemble de données, puis présenté une analyse comparative

Expérience avec DT1. Le tableau 2 et la figure 3 montrent respectivement les mesures de performance et la courbe ROC des résultats de notre approche de classification testée sur le jeu de données DT1. Les résultats sur le tableau 2a) et la figure 3a) sont obtenus sans tenir compte du résultat de test diagnostic rapide(TDR) en contraste avec les mesures du tableau 2b) et de la figure 3b). On peut facilement voir que l'exactitude de notre modèle de prédiction est sensiblement la même lorsque l'on considère ou non le TDR; cette prédiction est assez bonne comme le prouve la précision supérieure à 90%.

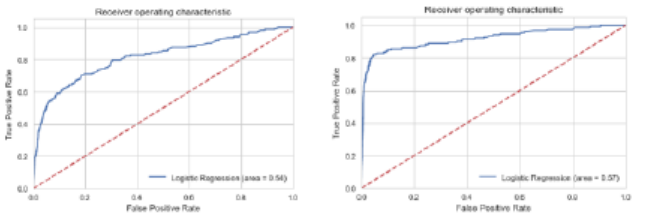
(a) Prediction without the QDT outcome

Precision	Recall	F-measure
0.97	1.0	0.99

(b) Prediction with the QDT outcome

Precision	Recall	F-measure
0.98	1.0	0.99

Table 2. Performance measures of the prediction on DT1



(a) Prediction without the QDT outcome (b) Prediction with the QDT outcome

Fig. 3. The curve of the Receiver Operating Characteristic for prediction on DT1

Expériences avec DT2. Le tableau 3 et la figure 4 montrent respectivement les mesures de performance et la courbe ROC des résultats (avec ou sans prise en compte du TDR) de notre modèle de prédiction testée sur le jeu de données DT2. Les mesures de précision sur la table 3a) et la figure 4) respectivement à celles du tableau 3b) et de la figure 4b) montrent que notre classificateur réussit bien lorsque l'on considère le TDR comme un attribut alors que sans le TDR la précision de la prédiction diminue.

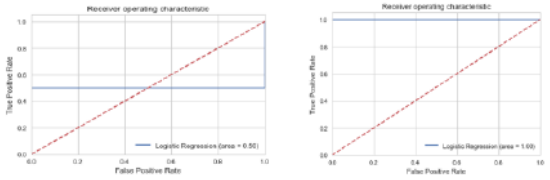
(a) Prediction without the QDT outcome

Precision	Recall	F-measure
0.75	1.0	0.86

(b) Prediction with the QDT outcome

Precision	Recall	F-measure
1.0	1.0	1.0

Table 3. Performance measures of the prediction on DT2



(a) Prediction without the QDT outcome (b) Prediction with the QDT outcome

Fig. 4. The curve of the Receiver Operating Characteristic for prediction on DT2

Expériences avec DT3. De manière similaire aux résultats sur DT2, les mesures de performance sur DT3 montre une meilleure précision de prédiction (voir le tableau 4 et la figure 5) lorsque le TDR est pris en compte comme un attribut

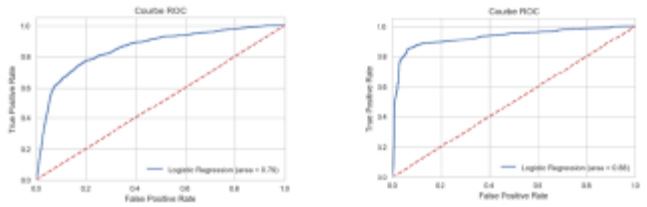
(a) Prediction without the QDT outcome

Precision	Recall	F-measure
0.75	1.0	0.86

(b) Prediction with the QDT outcome

Precision	Recall	F-measure
1.0	1.0	1.0

Table 3. Performance measures of the prediction on DT2



(a) Prediction without the QDT outcome (b) Prediction with the QDT outcome

Fig. 5. The curve of the Receiver Operating Characteristic for prediction on DT3

Analyse comparative des résultats Les expériences prouvent que notre modèle de prévision basé sur la régression logistique fonctionne bien en général et en particulier lorsque le résultat du test de diagnostic rapide est considéré comme une caractéristique du processus d'apprentissage, plus précisément, la précision de nos prédictions est supérieure à 90% pour les jeux de données DT1 et DT2, atteignant 100% pour DT1. Pour le cas spécifique du jeu de données TD1 contenant des données manquantes à l'aide d'un algorithme d'imputation, cette précision ne diminue pas, même si le TDR n'est pas pris en compte pendant le processus de prévision, et seules les caractéristiques du paludisme, c'est-à-dire les signes et les symptômes, sont prises en compte. En conséquence, nous pensons qu'il est possible de construire un modèle de prévision efficace et robuste du paludisme sans qu'il soit nécessaire d'effectuer le test de diagnostic rapide ; c'est-à-dire prédire si un patient donné est atteint ou non de paludisme

6. Conclusion

Dans cet article, nous avons étudié le problème de la prédiction de la présence ou non de paludisme chez un patient considéré comme atteint dans le contexte du Sénégal et en utilisant des techniques d'apprentissage automatique.

Pour résoudre ce problème, nous avons d'abord présenté une méthode de préparation de données qui permet de nettoyer, normaliser et imputer les valeurs manquantes à partir d'un jeu de données réel en utilisant des outils et des algorithmes efficaces. Nous avons également introduit un moyen d'extraire les fonctionnalités qui caractérisent le paludisme. Nous avons ensuite proposé un modèle de prédiction basé sur la régression logistique pour déterminer l'apparition du paludisme. La performance d'un tel modèle a été démontrée à travers de nombreuses expérimentations sur le monde réel et un ensemble de données semi-synthétiques. Comme perspective de recherche, nous prévoyons d'abord d'inclure une prévalence facteur dans notre fonction de prédiction afin d'améliorer sa précision. Deuxièmement, nous allons utiliser d'autres modèles de classification binaire tels que Support Vector Machine (ou SVM en abrégé) et comparer leurs résultats à ceux obtenus avec le modèle basé sur la régression logistique.

References

1. Openrefine. http://openrefine.org/, online; accessed 30 Octobery 2018
2. Python implementation of missforest. https://pypi.org/project/predictive imputer/, online; accessed 31 October 2018
3. Adimi, F., Soebiyanto, R.P., Safi, N., Kiang, R.: Towards malaria risk prediction in afghanistan using remote sensing. Malaria Journal 9(1), 125 (May 2010)
4. Aminot I, D.M.: The use of logistic regression in the analysis of data concerning good medical practice. Rev Med Ass Maladie 33(2), 157–143 (2002)
5. AS, A., AM, V., SH., K.: Malaria parasite development in the mosquito and infection of the mammalian host pp. 195–221 (2009)
6. Bbosa, F.,Wesonga, R., Jehopio, P.: Clinical malaria diagnosis: rule-based classification statistical prototype. SpringerPlus 5(1), 939 (Jun 2016)
7. Chiroma, H., Abdul-kareem, S., Ibrahim, U., Ahmad, I.G., Garba, A., Abubakar, A., Hamza, M.F., Herawan, T.: Malaria severity classification through jordan-elman neural network based on features extracted from thick blood smear. In: Neural Network World (2015)
8. Dua, S., Acharya, U.R., Dua, P.: Machine Learning in Healthcare Informatics. Springer Publishing Company, Incorporated (2013)
9. Daz, G., Gonzlez, F.A., Romero, E.: A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. Journal of Biomedical Informatics 42(2), 296 – 307 (2009)
10. Ferguson, H.M., Mackinnon, M.J., Chan, B.H., Read, A.F.: Mosquito mortality and the evolution of malaria virulence. Evolution 57(12), 2792–2804 (2003)
11. Hosmer, D.W., Lemeshow, S.: Applied logistic regression. John Wiley and Sons (2000)
12. Kunwar, S.: Malaria Detection Using Image Processing and Machine Learning. ArXiv eprints (Jan 2018)
13. Kusumasari, T.F., Fitria: Data profiling for data quality improvement with openrefine. In: 2016 International Conference on Information Technology Systems and Innovation (ICITSI). pp. 1–6 (Oct 2016)
14. Lemaˆıtre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 18(17), 1–5 (2017)
15. Preux, P., Odermatt, P., Perna, A., Marin, B., Vergnengre, A.: Qu'est-ce qu'une regression
logistique ? Revue des Maladies Respiratoires 22(1, Part 1), 159 – 162 (2005)
16. Rajpurkar, P., Polamreddi, V., Balakrishnan, A.: Malaria likelihood prediction by effectively surveying households using deep reinforcement learning. CoRR abs/1711.09223 (2017)
17. Robert, C.: Machine learning, a probabilistic perspective. CHANCE 27(2), 62–63 (2014)
18. Silva, L.O., Z´arate, L.E.: A brief review of the main approaches for treatment of missing data. Intell. Data Anal. 18(6), 1177–1198 (Nov 2014)
19. Sokhna, C., Mboup, B.M., Sow, P.G., Camara, G., Dieng, M., Sylla, M., Gueye, L., Sow, D., Diallo, A., Parola, P., Raoult, D., Gautret, P.: Communicable and non-communicable disease risks at the Grand Magal of Touba: The largest mass gathering in Senegal. Travel Medicine and Infectious Disease 19, 56–60 (Sep 2017)

20. Sperandei, S.: Understanding logistic regression analysis. Biochem Med 24, 12–18 (feb 2014)
21. Stekhoven, D.J., Bhlmann, P.: Missforest: non-parametric missing value imputation for mixed-type data. Bioinformatics 28(1), 112–118 (2012)
22. Swalin, A.: How to handle missing data. https://towardsdatascience.com/how-to-handlemissing- data-8646b18db0d4, online; accessed 31 October 2018
23. Ugwu, C., Onyejegbu, N.L., Obagbuwa, I.C.: The application of machine learning technique for malaria diagnosis. Int. J. Green Comput. 1(1), 68–77 (Jan 2010)
24. Wang, J., Xu, M., Wang, H., Zhang, J.: Classification of imbalanced data by using the smote algorithm and locally linear embedding. In: 2006 8th international Conference on Signal
Processing. vol. 3 (Nov 2006)
25. WHO: World malaria report in 2017 (2017)
.