I/ Data Preparation

1. Descriptions des données

Le jeu de données all_data_original_copie comporte 21083 patients et 16 attributs. Les attributs sont entre autres les données nominatives des patients, les signes et les symptômes constatés par le médecin, le traitement proposé, le diagnostic final du médecin, les résultats du TDR (Test de Diagnostic Rapide) et le statut (hospitalisation, décès ou mis en observation) du patient.

Insérer tableau ici

On remarque que les attributs tels que signe_Symptome et Diagnostic sont multivariés. Ainsi nous devrons les splitter pour récupérer les informations qui sont pertinentes pour notre étude à savoir prédire si un patient présentant un certain nombre de signes et symptômes est atteint de paludisme ou non

2. Sélection des signes et symptômes du paludisme

Toutes les informations contenues dans la colonne signe_Symptome ne sont pas utiles pour diagnostiquer un paludisme. Ainsi pour notre étude nous avons extrait 12 nouveaux attributs qui sont les signes et les symptômes du paludisme. Ces attributs sont les suivants : manque d'appétit, fatigue, arthralgie, trouble digestif, vertiges, frisson, myalgie, diarrhée et douleur abdominale.

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. Ainsi la colonne Diagnostic reste multivariée avec les catégories suivantes: accès palustre, paludisme, simple paludisme, paludisme grave, accès paludisme grave, accès paludisme simple, syndrome palustre, paludisme suspect et pas de paludisme (se sont tous les diagnostics différents du paludisme).

Insérer tableau ici

3. Imputation des données manquantes

La plus part des attributs de notre jeu de données présente des valeurs manquantes. Ces valeurs manquantes peuvent avoir un impact négatif sur notre analyse future. Pour remplacer ces dernières, nous avons utilisé le package missForest du logiciel R.

Pour cela les attributs textuels comme les signes et symptômes ou les classes de l'attribut Diagnostic ont été remplacés par des valeurs numériques.

Avec un NRMSE de 0,01 nous pouvons dire que l'imputation a réussi mais aussi n'a pas altéré la structure des données.

4. Modèle de Prédiction
    a. Data Preprocessing

Dans l'optique de faire une régression logistique, les opérations suivantes ont été effectuées.

Si le patient présente l'un des symptômes ou signes suivants : manque appétit, fatigue, arthralgie, trouble digestif, vertige, frisson, myalgie, douleur abdominale, vomissement, nausée, Céphale ou

fièvre, on lui attribue la valeur 1 et si le patient ne présente pas le signe ou le symptôme, on lui attribue la valeur 0.

Pour le diagnostic, nous avons mis 1 lorsqu'un patient est atteint de l'un des types de paludisme énumérés ci-dessus et 0 pour la classe non paludisme.

b. Data Exploration

Les diagrammes ci-dessous donnent une idée de la distribution des données que nous disposons.

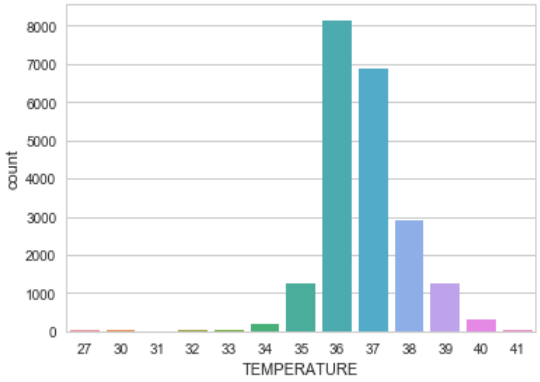


Figure 1

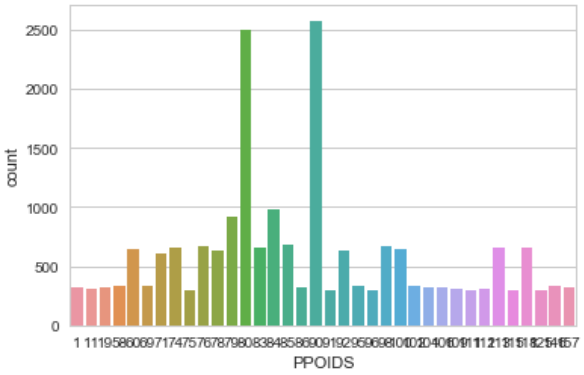


Figure 2

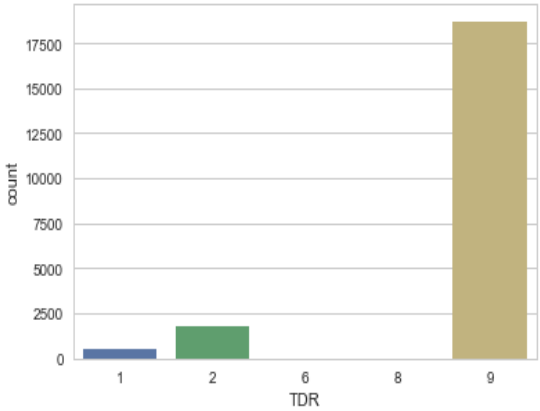


Figure 3

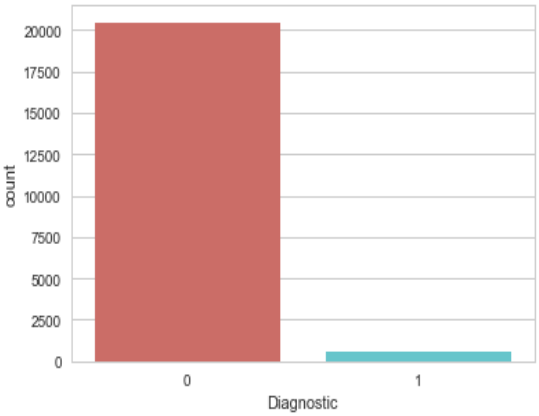


Figure 4

Nous remarquons que les classes de la colonne Diagnostic sont très déséquilibrées comme le montre d'ailleurs la figure 4.

Étant donné que la variable cible est l'attribut Diagnostic alors il risque de se produire. ce qu'on appelle un vote majoritaire, c'est-à-dire que l'algorithme aura tendance à prédire que tous les patients ne sont pas atteints de paludisme car la classe non paludisme (0) est trop majoritaire devant la classe paludisme (1).

Pour éviter cette situation nous proposons de rééquilibrer les données en faisant du 'over sampling'' ou sur échantillonnage.

Il 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. Pour cela nous avons implémenté l'algorithme SMOTE avec le **package imblearn** de python comme l'atteste cette figure.

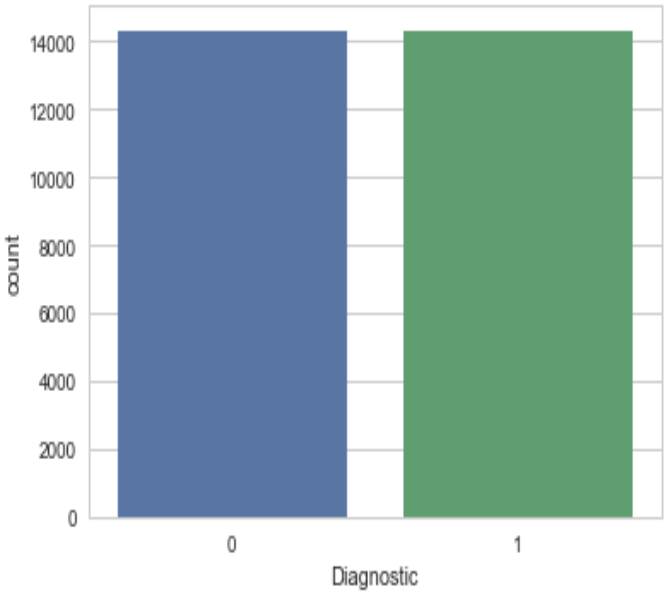


Figure 5

Cette figure 5 montre clairement que nous disposons maintenant des données qui sont parfaitement équilibrées et prêtes pour l'analyse.

Il faut aussi noter que le sur-échantillonnage effectué ne concerne que les données qui seront utilisées pour l'entrainement de notre modèle de prédiction. Ceci permettra de tester l'algorithme avec des données originales. Par conséquent il n'y aura pas perte d'information en testant le modèle.

c. Modèle de Prédiction

Après avoir rééquilibrer les données on les divise en deux parties : une partie pour entrainer le modèle et une autre partie pour tester le modèle.

La régression logistique est utilisée ici, car notre variable cible est binaire et la plus part des variables explicatives sont multinomiales. Elle est très souvent utilisée dans le domaine médical, car elle permet d'isoler les effets de chaque variable, c'est-à-dire d'identifier les effets résiduels d'une variable explicative sur une variable d'intérêt.

5) Résultat Expérimental

Nous appliquons deux fois la régression logistique.

D’abord, nous allons faire la prédiction avec les signes et les symptômes, le poids, l’âge et la température corporelle mais sans tenir en compte le TDR (Test de Diagnostic Rapide).

Ensuite nous faisons la prédiction avec les attribues précédents en plus du TDR ce qui nous permet de comparer les résultats obtenus

a. Première Expérience :

Après la première expérience, nous obtenons la matrice de confusion suivante :

[[2955 623]

[919 2670]]

Cette matrice montre qu’il y’a 2955 + 2670 prédictions qui sont correctes c'est à dire que 5625 patients ont eu un diagnostic adéquat et 623 + 919 patients mal classés c’est-à-dire que 1542 patients ont été mal diagnostiqués.

La précision sur le test est de 78,48% alors qu’il est de 81% sur l’entrainement

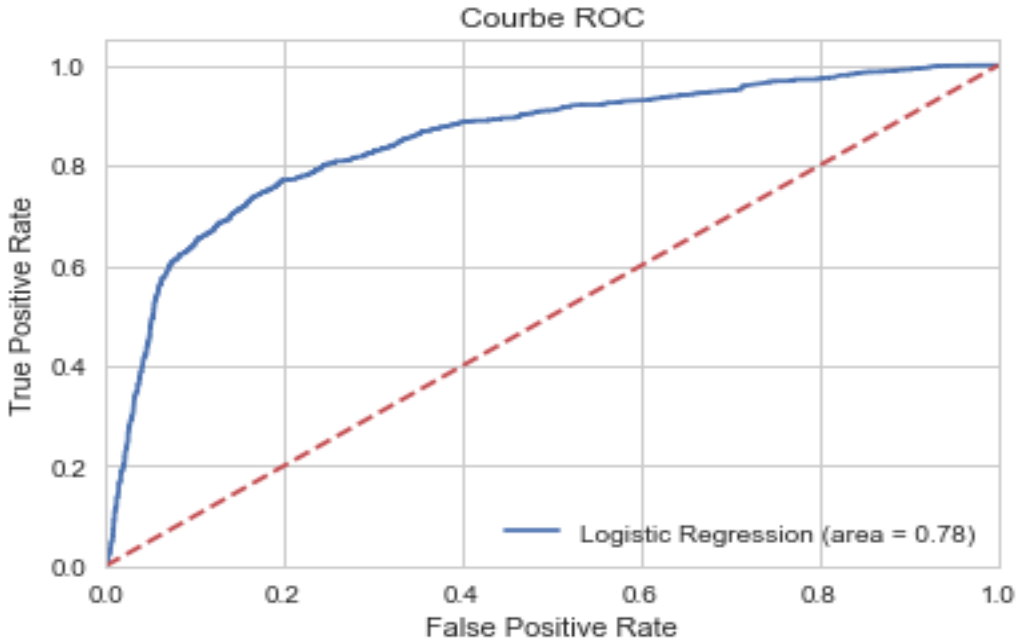
	Precision	Recall	f1-score	Support
0	0.76	0.83	0.79	3578
1	0.81	0.74	0.78	3589
avg / total	0.79	0.78	0.78	7167

La précision est le rapport du nombre total des vrais positifs par total des faux positifs.

Le recall (rappel) est le rapport du nombre total des vrais positifs par le nombre total de faux négatifs.

Ainsi en se basant uniquement sur les signes et les symptômes d’un patient sans faire aucun examen médical, nous pouvons prédire à hauteur de 78,48% qu’un patient a le paludisme.

La courbe ROC (Receiver Operating Chacacteristic) ci-dessous montre et confirme les résultats fournis par le calcul de la précision.



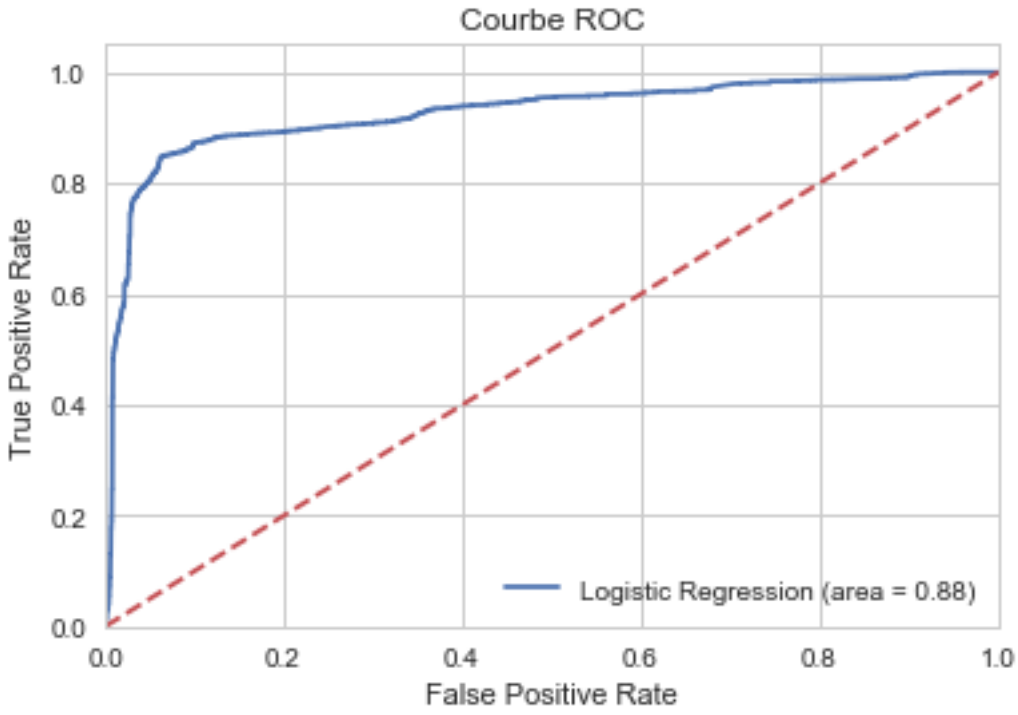
b. Deuxième expérience

Nous reprenons l'expérience précèdent mais nous incluons cette fois les résultats des TDR sur les données d'entrainement et de test. On remarque le nombre de faux positifs et le nombre faux négatifs diminuent au moment, alors que les vrais positifs et les vrais négatifs augmentent comme le montre les matrices ci-dessous

[[3231 347]
[488 3101]]

	Precision	Recall	f1-score	Support
0	0.87	0.90	0.89	3578
1	0.90	0.86	0.88	3589
avg / total	0.88	0.88	0.88	7167

Ainsi le score de précision passe 78,48% à 88,35% pour le test et de 81% à 90% pour l'entrainement. La courbe ROC correspondant est donnée sur la figure ci-dessous



On constate qu'un diagnostic suivi d'un examen médical est plus fiable qu'un diagnostic basé uniquement sur les signes et les symptômes d'un patient.

c. Conclusion

En somme, ce travail peut constituer un bon outil d'aide à la décision médical surtout dans des zones où il manque de personnels qualifiés comme les pays subsaharien. En effet l'algorithme que nous avons proposé permet de prédire avec une certitude de 88 ,35% si un patient est atteint de paludisme ou non. Cependant ce score pourra être amélioré dans les prochaines études en prenant en compte la localité, l'âge ou encore si le patient est décédé du paludisme ou non.

En effet un patient qui habite dans une zone à fort taux de prévalence présente beaucoup plus de risque d'attraper le paludisme qu'un patient qui habite dans une zone à faible taux de prévalence même si ces deux cas de patients présentent les mêmes symptômes. Ainsi le taux de mortalité, le taux de morbidité et le taux de létalité pourront être combiné pour fournir un coefficient de correction ou facteur de risque qui sera injecté dans le model pour augmenter sa précision.