**Description des données:**

Pour mener à bien les expériences de cette étude, nous avons utilisé deux ensembles de données. Le premier ensemble de données que nous appelons DT1, collecté lors du grand Magal de Touba en 2016 et qui concerne des patients de tout le Sénégal. Alors que le deuxième ensemble de données notait DT2 plus homogène concernant uniquement les patients des districts sanitaires de la région médicale de Diourbel, de Thiès et de Fatick. L'ensemble de données, ainsi que le pipeline de préparation de données que nous avons proposé pour nettoyer, normaliser et imputer des informations, sont donnés dans [16]. Le tableau 1 ci-dessous montre les caractéristiques de nos ensembles de données.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | **Variables** | **observations** | **Classe** | |
| **Malaria** | **Not malaria** |
| **DT1** | **16** | **21083** | **614** | **20469** |
| **DT2** | **16** | **16092** | **9223** | **6869** |

A partir des jeux données DT1 et DT2 nous avons construit les jeux de données DT3, DT4 et DT5 suivants.

**DT3:** Il est obtenu en combinant les jeux de données DT1 et DT2. Ainsi il concerne 37175 patients dont 9837 sont diagnostiqués positifs au paludisme.

**DT4:** Il est obtenu en considérant les 16092 patients du jeu de données DT2 (dont 9223 malades de paludisme). Etant données que ce DT2 est déséquilibré, nous avons tires de manière aléatoire 2354 patients testés négatif au paludisme dans le jeu de données DT1 en fin de le rééquilibré. Ainsi il concerne 18446 patients dont 9223 sont atteintes de paludisme.

**DT5:** estobtenu par sur échantillonnage de DT1 par la méthode SMOTE de python. Cette méthode consiste d’abord à diviser DT1 en deux parties dont une pour l’entrainement (train set) et l’autre pour le test (test set). Le train set étant déséquilibré, alors on y applique la méthode SMOTE pou y remédier. Ainsi on obtient un nouveau train set comportant 30369 patients dont la moitié est testée positif au paludisme

**Expérimentation avec DT1**

Dans cette partie nous avons utilisé le jeu données DT1 avec les classes déséquilibrées. C’est-à-dire le nombre de patients atteints de paludisme est très minoritaire devant ceux diagnostiqués négatif au paludisme

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F-score | AUC |
| DT |  |  |  |  |
| 0 | 0.97 | 1.00 | 0.98 | 0.78 |
| 1 | 0.38 | 0.0.5 | 0.09 |  |
| RF |  |  |  |  |
| 0 | 0.97 | 1.00 | 0.99 | 0.81 |
| 1 | 0.45 | 0.07 | 0.12 |  |
| LR |  |  |  |  |
| 0 | 0.97 | 1.00 | 0.99 | 0.79 |
| 1 | 0.62 | 0.05 | 0.10 |  |
| NB |  |  |  |  |
| 0 | 0.97 | 1.00 | 0.99 | 0.82 |
| 1 | 0.00 | 0.00 | 0.00 |  |
| SVM |  |  |  |  |
| 0 | 0.97 | 1.00 | 0.99 | 0.52 |
| 1 | 0.00 | 0.00 | 0.00 |  |
| ANN |  |  |  |  |
| 0 | 0.97 | 1.00 | 0.99 | 0.84 |
| 1 | 0.55 | 0.04 | 0.07 |  |

**Expérimentation avec DT2**

Dans cette partie nous avons utilisé le jeu données DT2 avec les classes déséquilibrées. C’est-à-dire le nombre de patients atteints de paludisme est très majoritaire devant ceux diagnostiqués négatif au paludisme

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F-score | AUC |
| DT |  |  |  |  |
| 0 | 0.59 | 0.48 | 0.48 | 0.64 |
| 1 | 0.64 | 0.80 | 0.71 |  |
| RF |  |  |  |  |
| 0 | 0.63 | 0.34 | 0.44 | 0.64 |
| 1 | 0.64 | 0.85 | 0.73 |  |
| LR |  |  |  |  |
| 0 | 0.58 | 0.36 | 0.44 | 0.63 |
| 1 | 0.63 | 0.81 | 0.71 |  |
| NB |  |  |  |  |
| 0 | 0.60 | 0.34 | 0.43 | 0.63 |
| 1 | 0.63 | 0.83 | 0.72 |  |
| SVM |  |  |  |  |
| 0 | 0.58 | 0.05 | 0.09 | 0.62 |
| 1 | 0.58 | 0.97 | 0.73 |  |
| ANN |  |  |  |  |
| 0 | 0.59 | 0.40 | 0.48 | 0.65 |
| 1 | 0.64 | 0.80 | 0.71 |  |

**palu\_cross\_vali\_morbidite\_impute**

**Expérimentation avec DT3**

**Palu\_cross\_val\_append**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F-score | AUC |
| DT |  |  |  |  |
| 0 | 0.89 | 0.85 | 0.87 | 0.86 |
| 1 | 0.62 | 069 | 0.65 |  |
| RF |  |  |  |  |
| 0 | 089 | 0.85 | 0.87 | 0.87 |
| 1 | 0.62 | 0.70 | 0.65 |  |
| LR |  |  |  |  |
| 0 | 0.85 | 0.88 | 0.86 | 0.86 |
| 1 | 0.62 | 0.55 | 0.58 |  |
| NB |  |  |  |  |
| 0 | 0.86 | 0.87 | 0.86 | 0.85 |
| 1 | 0.62 | 0.60 | 0.61 |  |
| SVM |  |  |  |  |
| 0 | 0.87 | 0.86 | 0.86 | 0.85 |
| 1 | 0.61 | 0.64 | 0.62 |  |
| ANN |  |  |  |  |
| 0 | 0.89 | 0.85 | 0.87 | 0.87 |
| 1 | 0.61 | 0.69 | 0.65 |  |
|  |  |  |  |  |

**Expérimentation avec DT4**

**Palu\_cross\_val\_append\_balance (Morbidité & Magal)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F-score | AUC |
| DT |  |  |  |  |
| 0 | 0.68 | 0.57 | 0.62 | 0.70 |
| 1 | 0.64 | 0..74 | 0.69 |  |
| RF |  |  |  |  |
| 0 | 0.68 | 0.56 | 0.62 | 0.70 |
| 1 | 0.64 | 0.74 | 0.69 |  |
| LR |  |  |  |  |
| 0 | 0.67 | 0.59 | 0.63 | 0.70 |
| 1 | 0.64 | 0.72 | 0.68 |  |
| NB |  |  |  |  |
| 0 | 0.68 | 0.59 | 0.63 | 0.70 |
| 1 | 0.65 | 0.73 | 0.68 |  |
| SVM |  |  |  |  |
| 0 | 0.68 | 0.58 | 0.62 | 0.69 |
| 1 | 0.64 | 0.73 | 0.68 |  |
| ANN |  |  |  |  |
| 0 | 0.68 | 0.56 | 0.62 | 0.70 |
| 1 | 0.64 | 0.75 | 0.69 |  |

**Expérimentation avec DT5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F-score | AUC |
| DT |  |  |  |  |
| 0 | 0.99 | 0.84 | 0.91 | 0.76 |
| 1 | 0.10 | 0.58 | 0.17 |  |
| RF |  |  |  |  |
| 0 | 0.99 | 0.84 | 0.91 | 0. |
| 1 | 0.10 | 0.60 | 0.17 |  |
| LR |  |  |  |  |
| 0 | 0.9 | 0.78 | 0.88 | 0.84 |
| 1 | 0.09 | 0.75 | 0.17 |  |
| NB |  |  |  |  |
| 0 | 0.99 | 0.82 | 0.90 | 0.84 |
| 1 | 0.11 | 0.71 | 0.18 |  |
| SVM |  |  |  |  |
| 0 | 0.99 | 0.86 | 0.92 | 0.80 |
| 1 | 0.12 | 0.62 | 0.20 |  |
| ANN |  |  |  |  |
| 0 | 0.99 | 0.84 | 0.91 | 0.79 |
| 1 | 0.11 | 0.65 | 0.18 |  |

**CV scores**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | DT | RF | LR | NB | SV M | ANN |
| DT1 | 0.9704 | 0.9713 | 0.9719 | 0.9713 | 0.9713 | 0.9715 |
| DT2 | 0.6301 | 0.6333 | 0.6196 | 0.6224 | 0.6286 | 0.6286 |
| DT3 | 0.8086 | 0.8086 | 0.7959 | 0.7983 | 0.7994 | 0.8668 |
| DT4 | 0.6560 | 0.6555 | 0.6582 | 0.6598 | 0.6563 |  |
| DT5 | 0.8341 | 0.8343 | 0.7835 | 0.8186 | 0.8561 | 0.8326 |
|  |  |  |  |  |  |  |

**AUC**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | DT | RF | LR | NB | SV M | ANN |
| DT1 | 0.78 | 0.81 | 0.79 | 0.82 | 0.84 | 0.84 |
| DT2 | 0.64 | 0.64 | 0.63 | 0.63 | 0.62 | 0.65 |
| DT3 | 0.86 | 0.87 | 0.86 | 0.85 | 0.85 | 0.87 |
| DT4 | 0.70 | 0.69 | 0.70 | 0.69 | 0.70 | 0.70 |
| DT5 | 0.76 | 0.77 | 0.84 | 0.84 | 0.80 | 0.79 |
|  |  |  |  |  |  |  |

Table 3 below shows the performance measures (precision, recall, F-measure and precision) of the results of our classifiers after experimentation in all our datasets. The observation shows that the best scores of our classifiers are achieved on the datasets DT1, DT3 and DT5 in general

**Méthodes ensemblistes sur DT4**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Precision | Recall | F-score | score | AUC |
| Max Voting |  |  |  | 74 | 0.85 |
| 0 | 74 |  |  |  |  |
| 1 | 90 |  |  |  |  |
|  |  |  |  |  |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |