Rapport de projet

Groupe Malaria

Nom challenge : ***MediChal***
Nom de groupe : ***Malaria***
Membres : Charlotte TRUPIN <charlotte.trupin@u-psud.fr>, Mathilde LASSEIGNE <mathilde.lasseigne@u-psud.fr>, Alicia BEC <alicia.bec@u-psud.fr>, Sakina ATIA <sakina.atia@u-psud.fr>, Maxime VINCENT <maxime.vincent1@u-psud.fr>, Baptiste MAQUET <baptiste.maquet@u-psud.fr>.
URL challenge : *https://codalab.lri.fr/competitions/625*
Lien dépôt GitHub : *https://github.com/charlottetrupin/malaria*

Binôme 1 preprocessing : Alicia BEC & Charlotte TRUPIN
Binôme 2 model : Maxime VINCENT & Baptiste MAQUET
Binôme 3 visualization : Sakina ATIA & Mathilde LASSEIGNE

Contexte et motivation

La malaria autrement connue sous le nom de paludisme est une maladie infectieuse due à un parasite, propagé par certaines espèces de moustiques. Elle atteint majoritairement les enfants de moins de 5 ans et les femmes enceintes. Au total, près de 219 millions de personnes contaminées et 435 000 décès ont été enregistrés en 2017.
(Source : https://fr.wikipedia.org/wiki/Paludisme)
Le but de ce projet est donc de concevoir un algorithme simple permettant d'identifier et différencier les cellules saines des cellules infectées afin de faciliter la prise en charge de malades, tout particulièrement dans des zones à risque et ayant des problèmes sanitaires. Le recours à l'apprentissage automatique permettrait de dépister les personnes infectées à partir d'images de cellules, l'extraction de telles cellules demandant peu d'équipement.

Données et description du problème

Notre projet revient à traiter un problème de classification binaire. Nos données contiennent deux classes : les cellules infectées et les cellules saines, qu'il va nous falloir départager.
La répartition de nos données est telle que trouvée dans le **Tableau 1**.
Nous travaillons ce projet sur le document README.ipynb sur le dépot github donné plus haut.

Approche choisie

Pour commencer, nous avons choisi de faire une sélection de *features* pour exclure celles qui sont les moins utiles, une détection des outliers avec les méthodes fit et predict et une réduction de dimension avec une comparaison des méthodes *PCA (Principal Component Analysis)* et *TSNE (T-distributed Stochastic Neighbor Embedding)* qui permet de transformer les *features* en une dimension inférieure. Ce qui permet un gain de temps de calcul non négligeable lors de la classification.
Par la suite, nous avons sélectionné plusieurs classifieurs binaires que nous avons entraînés sur l'ensemble d'apprentissage à l'aide de *GridSearch* afin de déterminer les meilleurs hyperparamètres de chacun d'entre eux. Partant de ces algorithmes bien entraînés, nous avons cherché à savoir lesquels étaient en sur-apprentissage, en calculant leur score sur l'ensemble de test et à l'aide de l'outil de *cross-validation*. Notons que la méthode de calcul du score est basée sur l'AUC (aire sous la courbe ROC).

Le binôme visualization a choisi une approche un peu particulière, étant donné que nous jouons un rôle dans le travail de chaque groupe:
Nous avons choisi dans un premier temps d'analyser les méthodes de visualizations déjà présentes dans le notebook, comme par exemple les matrices de confusions ou encore l'histogramme représentant l'importance des features. Nous trouvions important et nécessaire d'étudier les visualizations déjà fournies et ne pas se précipiter dans la création de nouvelles visualizations pour comprendre quels étaient les résultats attendus et sous quelle forme.
Ensuite nous avons décidé d'implémenter de nouvelles visualizations en utilisant les bibliothèques pyplot et panda d'abord en modifiant celles proposées pour une visualization plus pertinente. Modifier des features déjà présentes nous permettait de mieux comprendre les données avec lesquelles nous travaillons et de tester des visualizations tout en ayant

connaissance du résultat attendu. Nous avons tenté d'améliorer le graphique représentant l'analyse des features en utilisant une représentation en violon. Toutefois, il s'est avéré que l'histogramme original était plus intéressant et plus facile à comprendre à l'œil nu.
Nous souhaitons par la suite, créer nos propres visualizations plus en adéquation avec les deux autres binômes, comme une comparaison des scores avec différents paramètres de processing.

Description du groupe
Le binôme 1 s'occupe de la sélection des outliers (**Code 1**), de sélectionner les features les plus importantes (**Code 2**) en fonction de la **Figure 1** donnée et d'implémenter les méthodes *PCA* et *TSNE* avant de tester laquelle performe le mieux. Après plusieurs tests, nous avons conclu que la méthode PCA est la plus performante sans TSNE (avec n_components = 4).

Le binôme 2 se charge d'analyser les modèles et de sélectionner le modèle le plus efficace pour résoudre notre problème.

Le binôme 3 se charge de représenter les données et les résultats de manière à faire ressortir ce qui est étudié. En choisissant une bonne visualisation, non seulement les résultats seront plus compréhensibles, mais il sera également plus facile d'avancer sur le projet. C'est le cas du Pie chart représentant l'importance des features donné en **Figure 1**. Nous trouvions plus intéressant de l'avoir sous cette forme, car il est plus visuel : plus de couleurs, et on voit de suite ce qui est mis en avant . Nous avons aussi créé deux graphiques simples montrant la répartition de cellules infectées parmi nos données. De plus, nous avons également appliqué les graphiques déjà existants à une plus grande base de données, tel que la matrice de confusion que nous avons appliquée à tous les modèles testés par le binôme Model, ce qui les a aidés à sélectionner le plus efficace (**Figure 3**).

Premiers résultats
Notre notebook README.ipynb donne les premiers résultats sur lesquels nous avons travaillé. Dans notre problème, on doit prêter une attention toute particulière à ce que le nombre de personnes réellement infectées mais classées comme saines (VN = Vrai Négatif != FP = Faux Positif), soit le plus faible possible. En ce sens, l'algorithme qui présente le meilleur score est le *RandomForestClassifier*, score qui se confirme lorsqu'on observe la matrice de confusion associée à ce classifieur (**Figure 3**). En effet, le *RandomForestClassifier* a globalement les nombres de VN et de FP les plus équilibrés et les plus bas. De plus, ce classifieur n'est pas en *over-fitting*. Cela dit le Perceptron a le nombre de faux négatif le plus faible mais un nombre de faux positif assez élevé ce qui l'enlève de notre sélection du meilleur algorithme.

Figures :

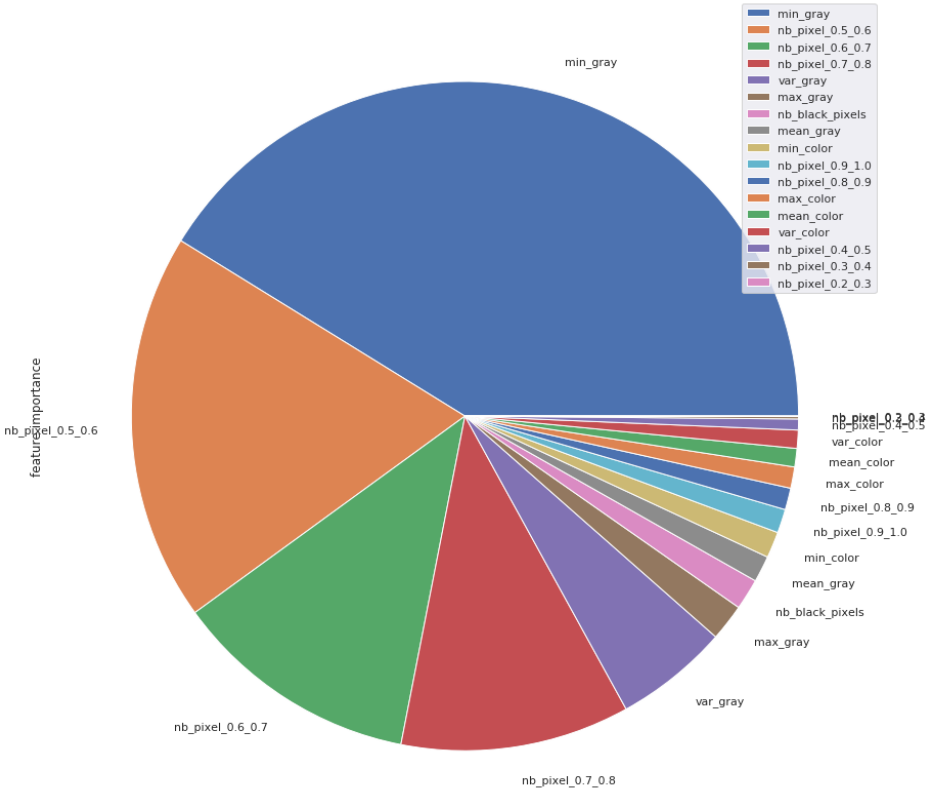


Figure 1: Importance des features proposées par le challenge

Figure 2 : Scores et sur-apprentissage des algorithmes

```
# Affichage des scores de chaque algorithme avec
# les meilleurs hyperparamètres
print(algorithm_scores)
# Meilleur modèle parmi ceux testés
best_model = max([(algorithm_scores[algo], algo) for algo in algorithm_scores.keys()])[1]
print("Best model =", best_model)
```

```
{'Perceptron': 0.8428046592896256, 'MLPClassifier': 0.9340020256250104, 'KNeighborsClassifier': 1.0, 'DecisionTree
Classifier': 0.9481340463512504, 'RandomForestClassifier': 0.9845173615546335}
Best model = KNeighborsClassifier
```

```
check_over_fitting()
```

```
{'Perceptron': False,
 'MLPClassifier': False,
 'KNeighborsClassifier': True,
 'DecisionTreeClassifier': False,
 'RandomForestClassifier': False}
```

Code 1: Sélection des valeurs non aberrantes:

```
import warnings
warnings.filterwarnings('ignore')
liste = []
from sklearn.ensemble import RandomForestClassifier
estimator = IsolationForest(n_estimators = 10)
estimator.fit(X_train)
for i in range(data.shape[0]):
        if estimator.predict(X_train)[i] != -1 :
        liste.append(i)
# on ne garde dans X_train que les valeurs de la liste
X_train = X_train[liste,:]
```

Code 2: Sélection des features:

```
X_train = X_train[:,[0,5,7,14,15,16]]
# la liste [0,5,7,14,15,16] correspond aux colonnes des features les plus importantes
```

Code 3:

```
# Notre propre algorithme d'apprentissage
class AlgorithmLearning:

        # Constructeur très complexe
        def __init__(self):
                self.learning = False
                self.learnt = [ ]

        # Apprentissage légèrement en sur-apprentissage
        def fit(self, data, label):
                self.learnt = label

        # Prédiction parfaitement dépendante des données
        def predict(self, data):
                return self.learnt
```

Code 4: Pour créer et afficher les matrices de confusion **Figure 3**

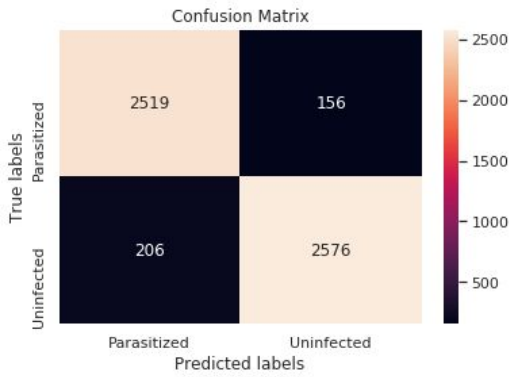
```
#Affiche la matrice de confusion sur l'ensemble de test

def plot_confusion_matrix(Y_test_predict):
    global Y_test
    ax = plt.subplot()
    #annot=True to annotate cells
    sns.heatmap(confusion_matrix(Y_test, Y_test_predict), annot=True, fmt='g', ax=ax)

    # labels, title and ticks
    ax.set_xlabel('Predicted labels');
    ax.set_ylabel('True labels');
    ax.set_title('Confusion Matrix');
    ax.xaxis.set_ticklabels(['Parasitized', 'Uninfected']);
    ax.yaxis.set_ticklabels(['Parasitized', 'Uninfected']);
    plt.show()

# Affichage des matrices de confusion pour chaque algorithme
for file_predict in files_predict:
    plot_confusion_matrix(nparray_predict(file_predict))
```

Tableau 1 : Statistiques des données sur l'algorithme du RandomForest:

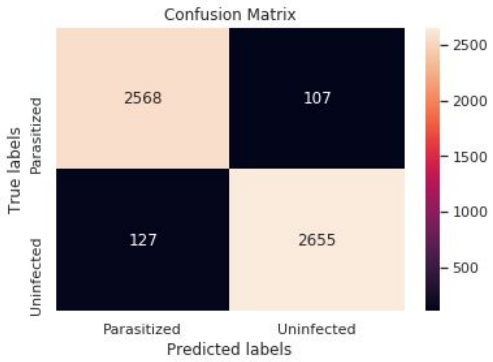
Dataset	Nombre d'exemples	Nombre de features	A des données manquantes ?	Nombre d'exemples dans chaque classe Bien classé/ mal classé	
Training	11077	19 (label compris)	Non	10 905	172
Test	5457	18	Non	5203	254

Figure 3 : Matrices de confusion

Perceptron Multicouche:

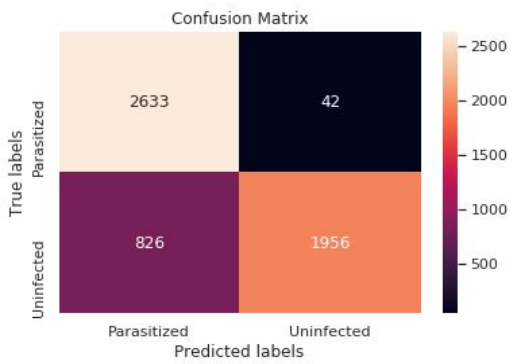


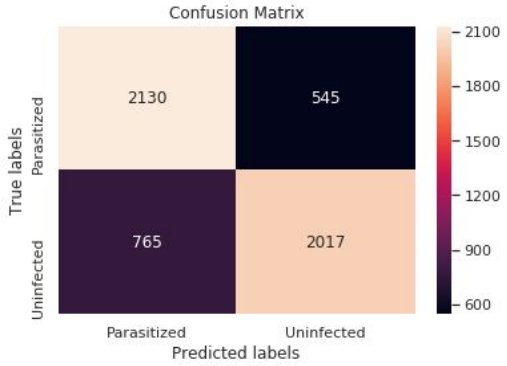
*Forêt d'arbres de décisio*n:



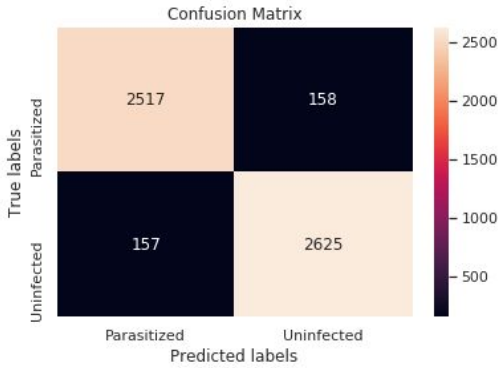
K-voisins:

Perceptron:





Arbre de décision:



Bibliographie :

- https://fr.wikipedia.org/wiki/Paludisme
- https://scikit-learn.org/stable/modules/outlier_detection.html
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- https://scikit-learn.org/stable/modules/feature_selection.html
- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
- https://pandas.pydata.org/pandas-docs/stable/index.html
- https://matplotlib.org/api/pyplot_api.html
-