Prédiction de 5 classes de cancers par trois algorithmes de Machine Learning

21200834[1]

[1]Aix-Marseille Université, Marseille, France

Introduction

L’apprentissage automatique (*Machine Learning* ou *ML* en anglais) est un champ d’étude de l’intelligence artificielle qui vise à donner aux machines la capacité d’«apprendre» à partir des données, via des modèles mathématiques.

Dans ce projet, nous allons utiliser trois modèles de ML (arbre de décision, régression logistique et le réseau neuronal) pour prédire le type de cancer des patients en fonction du niveau d’expression des gènes.

Matériels et méthodes

Le jeu de données utilisé fait partie de l'ensemble de données RNA-Seq (HiSeq) PANCAN. Il contient 20 531 gènes et 801 patients atteints d’une des classes tumorales: BRCA (Breast Invasive Carcinoma), KIRC (Kidney Renal Clear Cell Carcinoma), COAD (Colon Adenocarcinoma), LUAD (Lung Adenocarcinoma) et PRAD (Prostate Adenocarcinoma).

Dans notre analyse, nous avons utilisé Python 3.8 avec Keras 2.11.0, Matplotlib 3.1.3, Numpy 1.20.1, Pandas 1.1.4 et Scikit-learn 1.2.0. Nous avons choisi trois algorithmes de classification : la régression logistique, l’arbre de décision (avec la librairie Scikit-learn) et le réseau neuronal artificiel (avec Keras). Le code est disponible dans le répertoire : ‘*github.com/Aimen-prog/Machine_Learning_Project’*

La régression logistique utilisée est multinomiale. Ce type de régression est parfaitement adapté à notre étude car la variable dépendante n'est pas limitée qu’à deux catégories.

Le réseau neuronal a été construit avec 3 couches (une couche d’entrée, de sortie et une intermédiaire). Pour la couche de sortie la fonction softmax a été choisi car elle convient aux classifications de type multi-classes. 100 epochs ont été choisi pour l’apprentissage car plus ce nombre est grand plus on devrait obtenir une bonne précision.

Résultats

Tout d’abord, une ACP (Analyse en composantes principales) a été réalisé à l’aide de la librairie Sckit-learn (Figure 1). L'analyse en composantes principales est l'une des méthodes d'analyse de données multivariées les plus fréquemment utilisées.

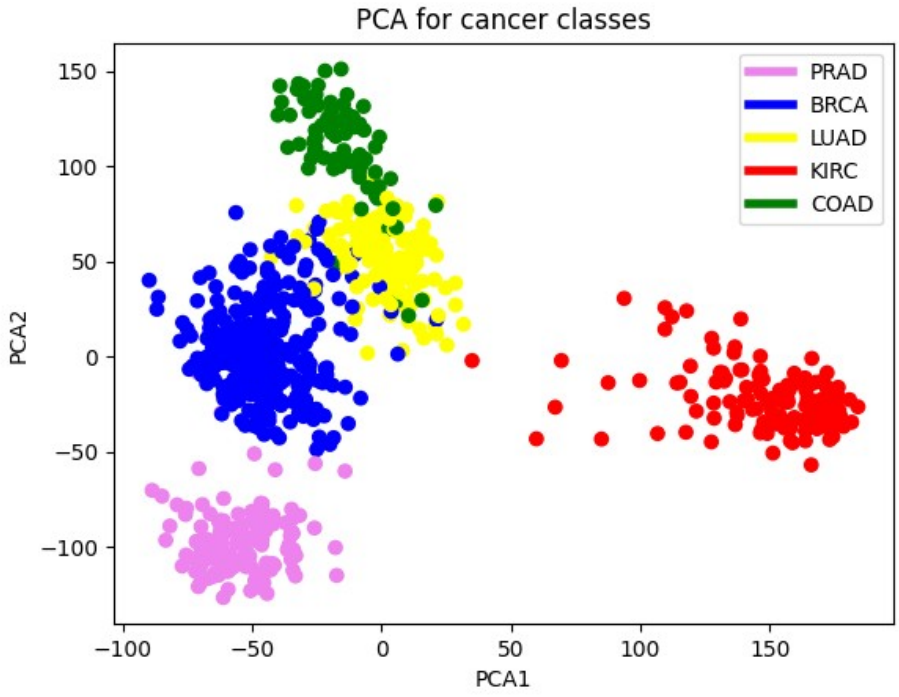

Figure 1 : Analyse en composantes principales (ACP) des classes de cancers PRAD, BRCA, LUAD, KIRC et COAD. Les deux composantes sont PC1 et PC2.

D’après les résultats, nous avons bien une stratification des patients en 5 régions. Chaque point correspond à un patient et chaque couleur correspond à une région, à un type de cancer. Nous pouvons voir que les types de cancers LUAD, COAD, PRAD et BRCA sont proches entre eux et que KIRC semble être plus différent.

Pour le modèle de régression logistique multinominale, nous avons découpé nos données en 70% des données pour l’entraînement du modèle et 30% pour tester les performances de ce dernier. Nous avons obtenu une ‘accuracy’ parfaite (=1) pour l'entraînement et quasi-parfaite (=0,9958) aussi pour le test. Cette métrique (accuracy) permet de connaître la proportion de bonnes prédictions par rapport à toutes les prédictions. Une mauvaise classification d’un individu réellement LUAD en BRCA a causé cette non obtention de 1 pour l’accuracy. (Figure 2 à gauche).

```
Accuracy training: 1.0
Accuracy test: 0.995850622406639
####
1       BRCA  COAD  KIRC  LUAD  PRAD
row_0
BRCA      96     0     0     1     0
COAD       0    18     0     0     0
KIRC       0     0    42     0     0
LUAD       0     0     0    40     0
PRAD       0     0     0     0    44
```

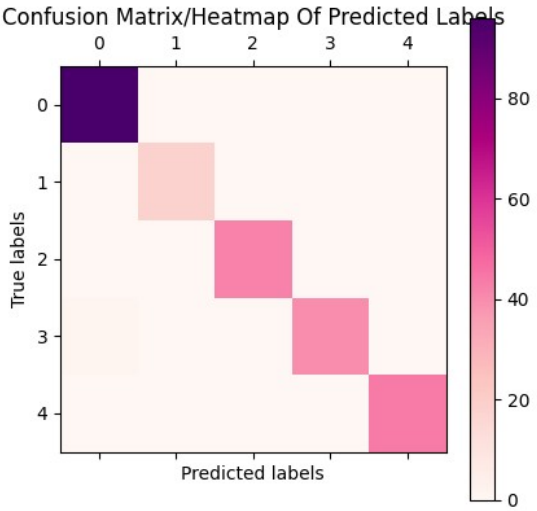

Figure 2 : Tableau de contingence du modèle de régression logistique (à gauche) et valeurs d'accuracy ; Tableau de contingence en forme de de heatmap mais qui n'est pas très descriptif à cause du faible nombre de données mal classées (à droite)

Concernant le modèle d'arbre de décision, le tableau de contingence donne des pourcentages de classification des classes de cancers vont de 0,90 jusqu'à 1 pour la partie test (30%). L'accuracy du test est égale à 0,95 dans ce cas. Ce modèle semble moins performant que la régression logistique mais reste tout de même un très bon prédicteur de classes (Figure 3).

```
Accuracy training: 1.0
Accuracy testing :  0.950207468879668
```

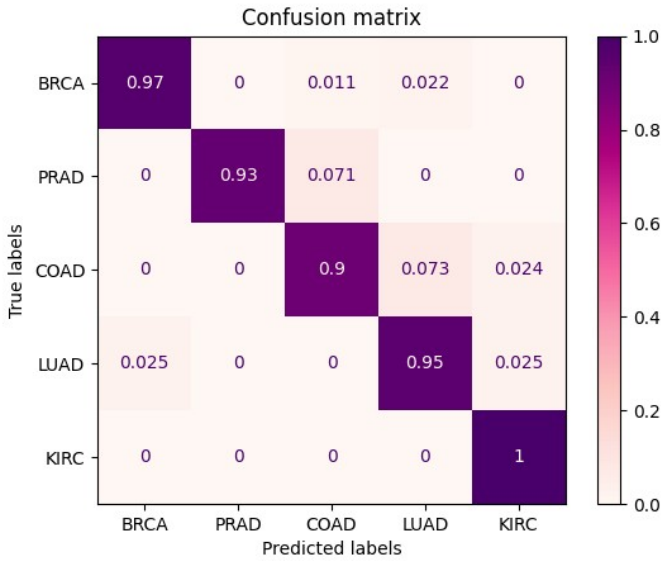

Figure 3 : Tableau de contingence du modèle d'arbre de décision et valeurs d'accuracy des classes de cancers PRAD, BRCA, LUAD, KIRC et COAD

En ce qui concerne l'ANN (A*rtificial neural network)*, on trouve une accuracy de test similaire à celle trouvé en régression logistique (=0,9958) avec toujours une seule mauvaise classification de LUAD en BRCA (Figure 4).

```
####
Accuracy test: 0.9958333333333333
####
BRCA = 0, PRAD=1, LUAD=2, KIRC=3, COAD=4
col_0   0   1   2   3   4
row_0
0      90   0   0   0   0
1       0  41   0   0   0
2       1   0  41   0   0
3       0   0   0  44   0
4       0   0   0   0  23
```

Figure 4 : Tableau de contingence de l'ANN et valeur d'accuracy pour les 5 classes de cancers qui sont codés comme suit : BRCA = 0, PRAD=1, LUAD=2, KIRC=3 et COAD=4

Conclusion

Pour conclure, les modèles utilisés ont montré de bonnes performances avec un faible taux d'erreur i.e. une seule mauvaise prédiction de BRCA à la place de LUAD. La régression et le réseau neuronal artificiel ont été les meilleurs prédicteurs de classes de cancers dans cette étude.

En somme, à travers cet exemple nous pouvons constater que l'apprentissage automatique peut être un outil très important pour le diagnostic, surtout avec la quantité de plus en plus importante de données à traiter.