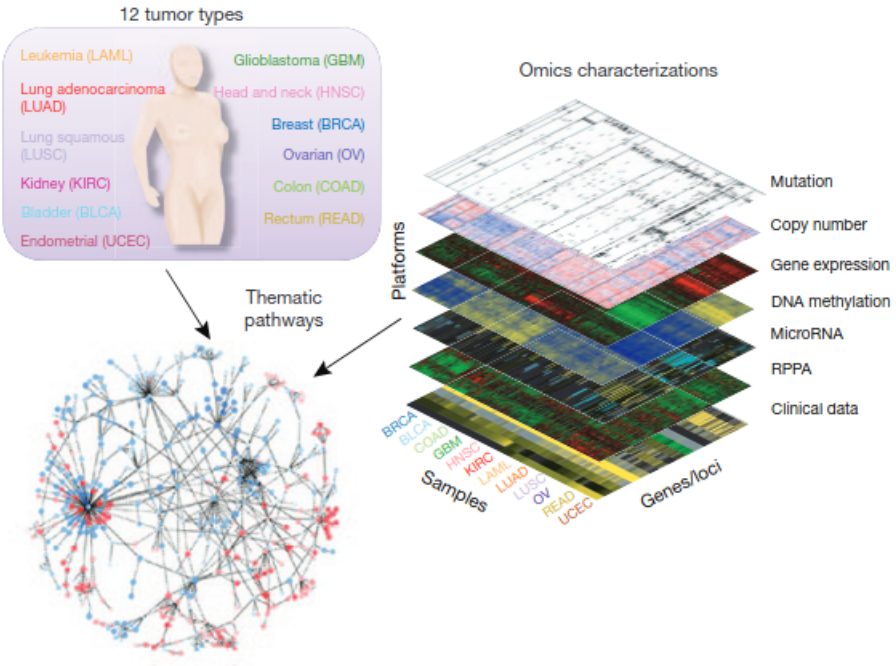
Machine Learning

Classification de cancer à partir de données ARN



Agathe Blanvillain [1], Kevin Tran [2]

1. agathe.blanvillain.1@etudiant.univ-rennes1.fr
2. kevin.tran@etudiant.univ-rennes1.fr

Introduction

Notre étude porte sur les différents types de cancer. Il s'agit de trouver une corrélation entre la valeur des gènes et le type de cancers.

Pour cela, on va utiliser les méthodes de classification apprises lors des 3 CMs et 3TPs de Machine Learning.

On étudiera les données qui correspondent à un échantillon de 801 individus et 20531 variables en enlevant les variables constantes, accompagné de cela les labels pour chaque individu.

L'étude se fera directement sur cet échantillon, pour travailler sur ce nombre de variable nous commencerons par faire une ACP et récolter les variables les plus importantes. Puis nous comparerons les résultats de 3 méthodes de classification :

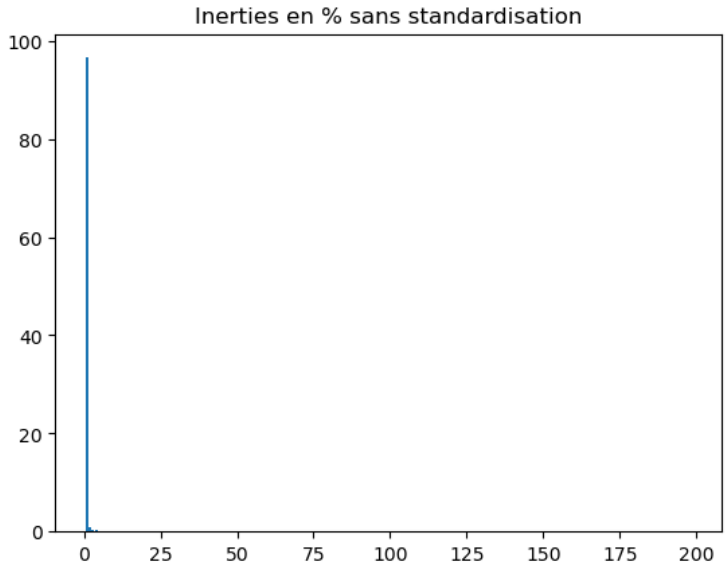
1. les k-moyennes ;
2. l'analyse discriminante linéaire ;
3. l'analyse discriminante quadratique.

Chaque classification sera faite avec une validation croisée et accompagnée par une matrice de confusion.

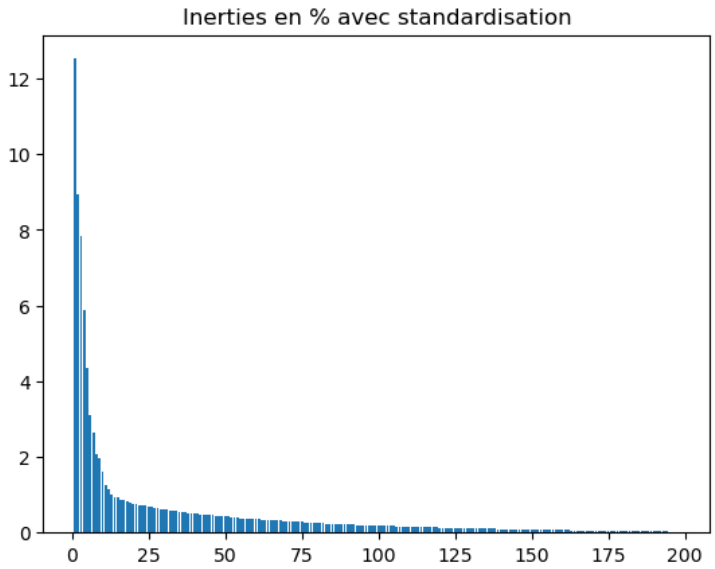
Analyse en composantes principales (ACP)

Au vu du nombre important de données et de la contrainte de temps, deux possibilités s'offrent à nous. Nous avions le choix entre sélectionner un nombre n de variables parmi les 20531 tirées de façon aléatoire et faire notre étude restreinte sur celles-ci ; ou de faire une ACP et faire une étude sur les composantes principales qui réunissaient un pourcentage conséquent de l'inertie.

Nous avons fait le choix de l'ACP, comme illustration de notre choix voici ce que donne notre ACP avec et sans standardisation (sur 200 variables, l'illustration étant plus parlante ainsi).



Cette image est trompeuse, on pourrait croire que toute l'information est contenue dans moins de 10 variables mais lorsque l'on regarde le graphique avec standardisation on obtient :



Ce qui est beaucoup plus cohérent et nous impose de faire un choix sur le pourcentage que l'on souhaite garder. Nous avons choisi 90% de l'inertie, ce qui correspond à 373 variables, cela nous permet d'avoir une grande partie de l'information, tout en faisant abstraction du "bruit". Avec ce pourcentage, nous trouvons des résultats satisfaisants pour nos différents algorithmes de classification.

Classification

On travaille maintenant sur les composantes principales sélectionner sur l'ACP et donc sur 373 variables. Le but dans cette partie est de comparer les différents modèles que nous allons étudier et de trouver le modèle le plus performant parmi ceux étudier.

K-moyennes

Nous avons fait une étude via l'algorithme de K-moyenne avec et sans validation croisée, et voici les résultats :

$$\begin{bmatrix} 74 & 0 & 0 & 4 & 0 \\ 0 & 134 & 0 & 1 & 1 \\ 0 & 0 & 145 & 1 & 0 \\ 0 & 0 & 0 & 139 & 2 \\ 0 & 0 & 0 & 47 & 253 \end{bmatrix}$$

FIGURE 1 – Matrice de confusion K-Moyennes

$$\begin{bmatrix} 38 & 1 & 0 & 4 & 0 \\ 0 & 64 & 0 & 1 & 0 \\ 0 & 0 & 71 & 0 & 0 \\ 0 & 29 & 0 & 123 & 0 \\ 0 & 0 & 0 & 1 & 71 \end{bmatrix}$$

FIGURE 2 – Matrice de confusion K-Moyennes avec validation croisée

On obtient aussi un taux d'erreur de 7% sans validation croisée et de 8% avec validation croisée. Ce résultat est appuyé avec le nuage de point proposé par l'algorithme des K-moyennes.

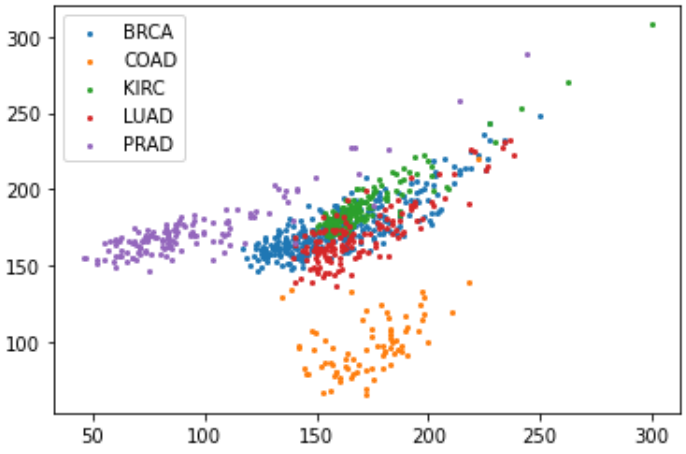
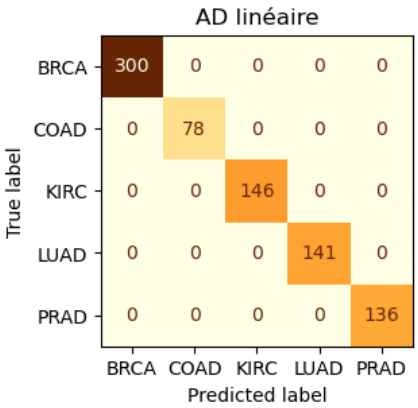


FIGURE 3 – Nuage de points K-moyennes

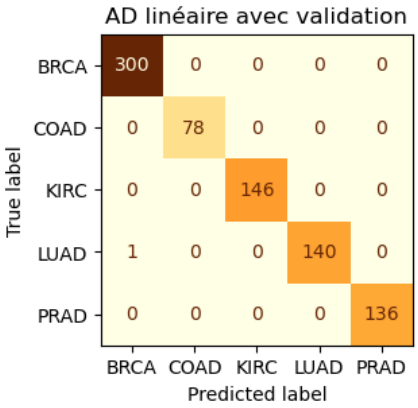
Analyse Discriminante

Linéaire

En analyse discriminante linéaire, sans validation croisée, on obtient un taux d'erreur = 0.0.



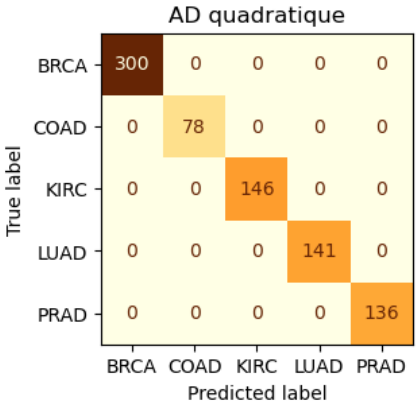
On réalise alors l'analyse discriminante avec validation croisée et on obtient un taux d'erreur = 0.004.



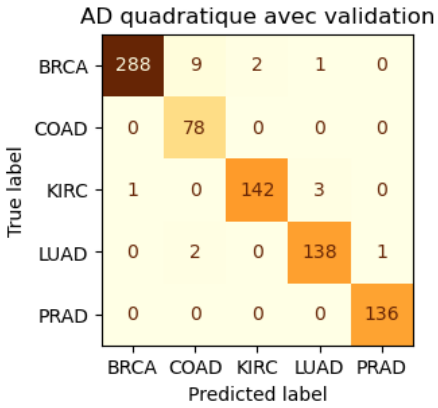
L'analyse discriminante linéaire semble donc être une méthode assez fiable.

Quadratique

En comparaison, on réalise l'analyse discriminante quadratique. Sans validation croisée, on obtient aussi un taux d'erreur = 0.0.



Cependant, avec la validation croisée, le taux d'erreur est beaucoup plus grand (≈ 0.143). De plus, l'ADQ a un temps de calcul plus long que l'ADL, avec de nombreux avertissements de variables colinéaires.



L'analyse discriminante quadratique est donc bien moins efficace que l'analyse discriminante linéaire.

Conclusion

Avec l'analyse des composantes principales (ACP), nous avons pu obtenir un nuage de points confus qui ne nous permettait pas de faire de classification :

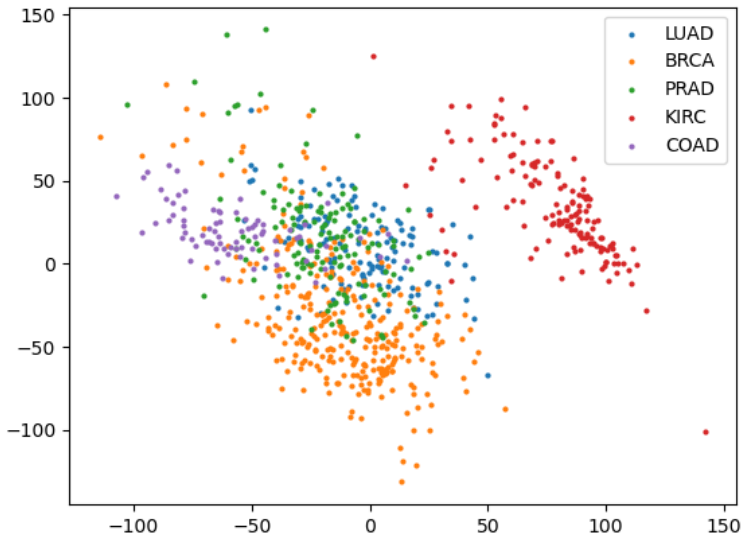
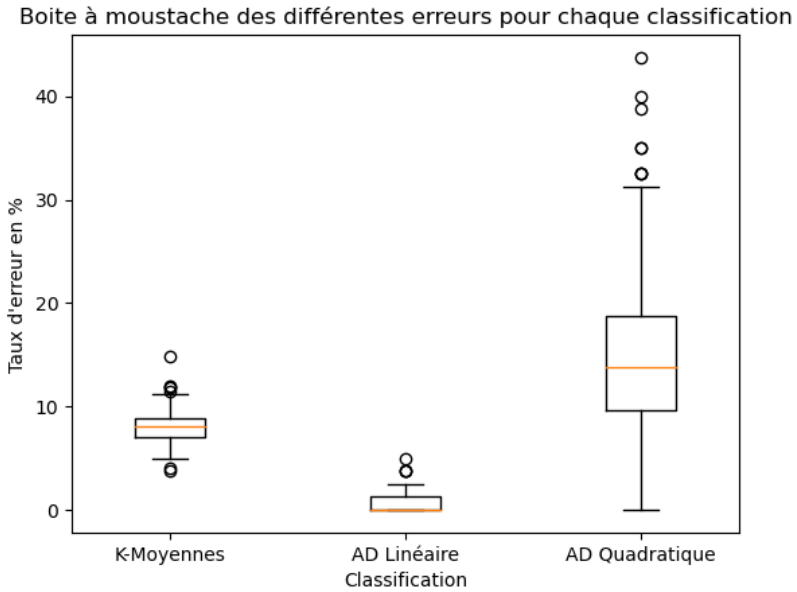


FIGURE 4 – Nuage de points Analyse en composantes principales

Les résultats obtenus avec les précédentes méthodes nous montrent que l'analyse discriminante linéaire est la plus efficace, devant les K-moyennes et l'analyse discriminante quadratique. On peut l'illustrer par la distribution du taux d'erreur de chacune des classifications.



On peut alors tracer les nuages de points avec l'analyse linéaire suivant les différentes composantes principales. Sur les graphes obtenus, on peut observer des groupes de points bien séparés selon la classe à laquelle ils appartiennent.

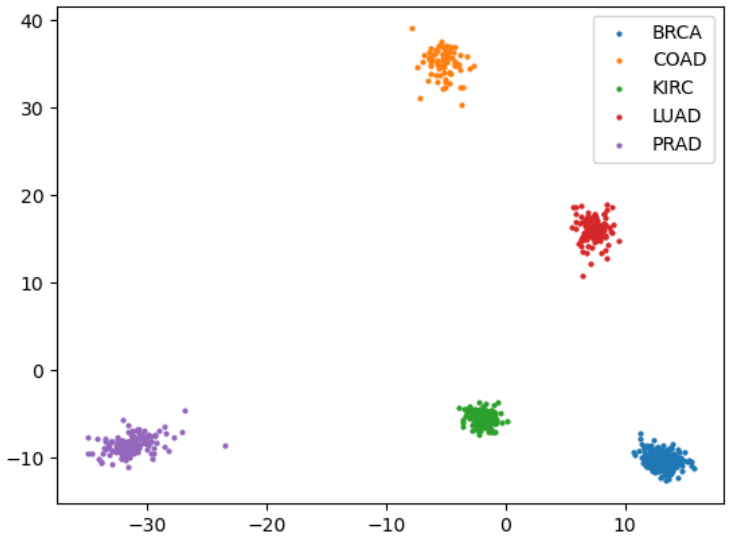


FIGURE 5 – Nuage de points Analyse discriminante linéaire

On remarque alors l'importance du traitement des données en comparant ce nuage de point à celui obtenu après ACP.