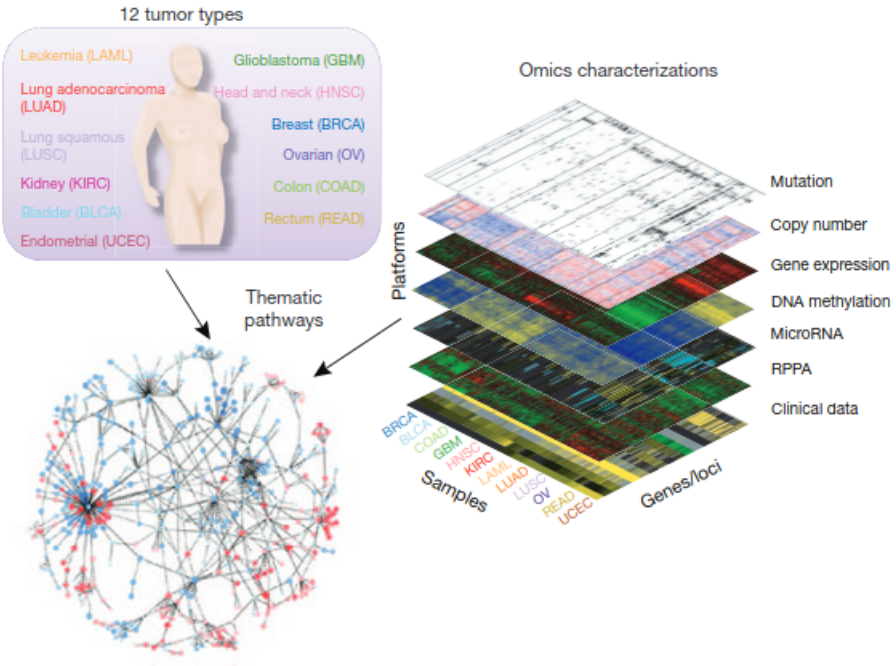
Machine Learning

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. Le but est de trouver une corrélation entre la valeur des gènes et le type de cancers et ainsi prédire à partir d'un jeu de données, le type de cancers auquel il correspond.

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 ces échantillons, pour travailler sur ce nombre important de variable nous commencerons par faire une ACP qui va nous permettre de faire une étude sur un espace de variables beaucoup plus petit mais qui conserve les informations fournis par les données. 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).

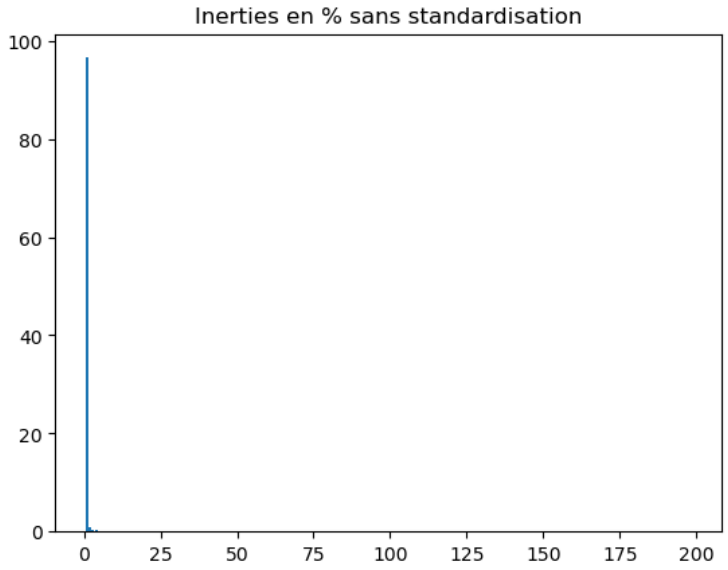


FIGURE 1 – ACP sans standardisation sur 200 variables

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 :

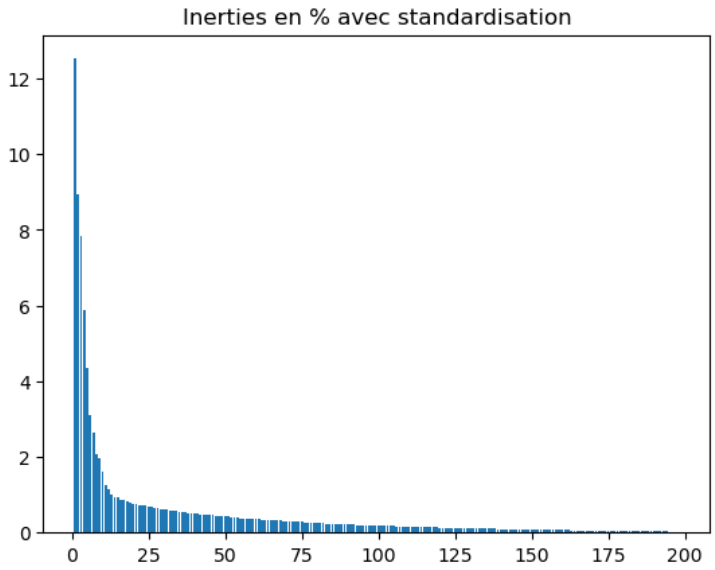


FIGURE 2 – ACP avec standardisation sur 200 variables

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

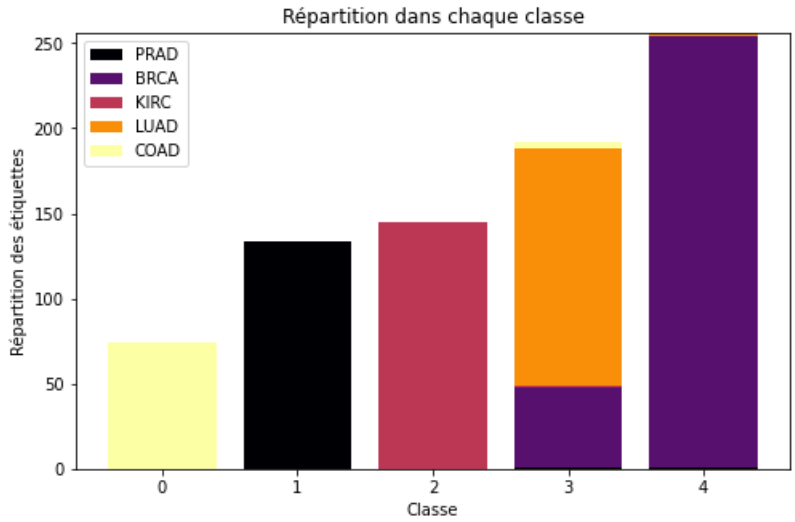
Nous avons fait une étude via l'algorithme de K-moyenne avec et sans validation croisée, et voici les résultats :

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

FIGURE 3 – 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 4 – Matrice de confusion K-Moyennes avec validation croisée



On obtient ainsi 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.

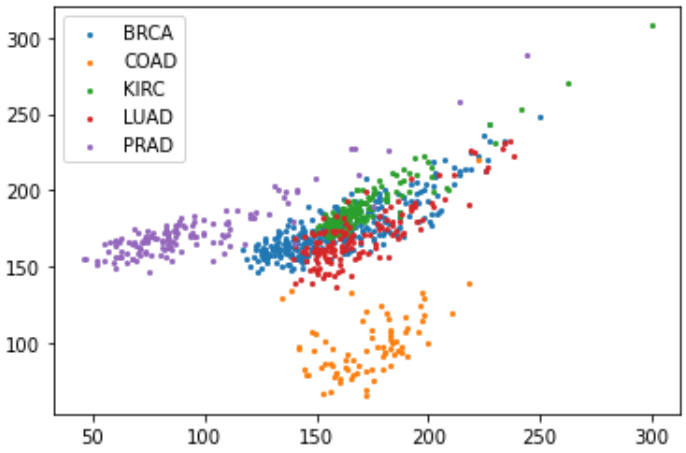
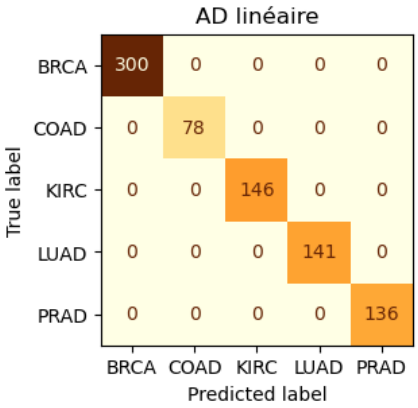


FIGURE 5 – Nuage de points K-moyennes

Analyse Discriminante

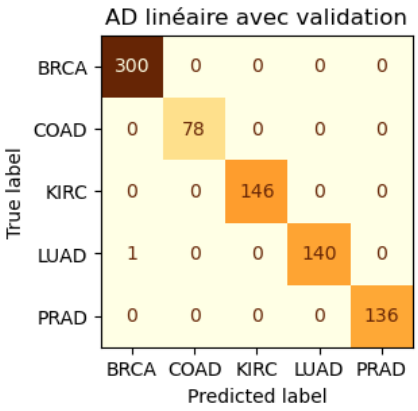
Linéaire

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



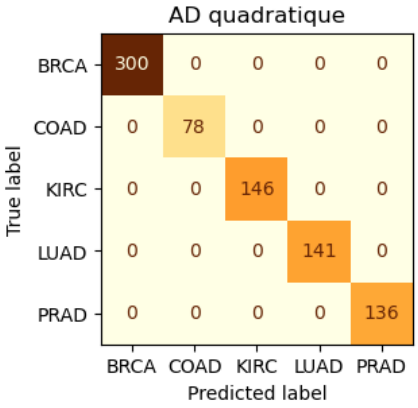
On réalise alors l'analyse discriminante avec validation croisée et on obtient un taux d'erreur = 0.004.

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

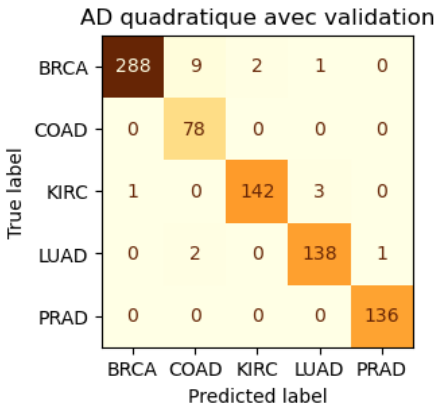


Quadratique

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). Le résultat est très peu satisfaisant car le nombre de variables est plus grand que le nombre d'individu. Ce qui cause de nombreux avertissements de variables colinéaires.



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 :

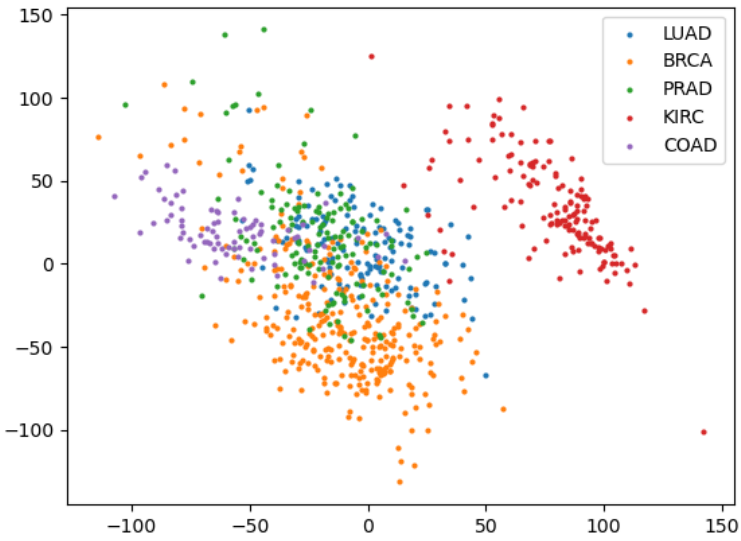
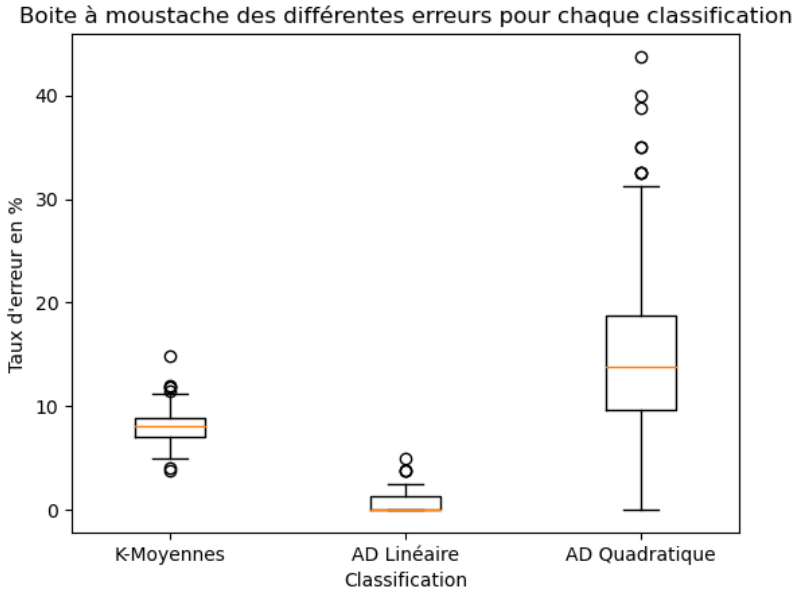


FIGURE 6 – Nuage de points Analyse en composantes principales

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.



L'algorithme d'ADQ marcherait nettement mieux si l'on réduisait le nombre de variables à une centaine. En faisant ainsi on obtient un taux d'erreur de 0.001. Le choix reste discutable et dépend de la situation, même si on réduit le nombre de variable l'algorithme d'ADQ reste coûteux par rapport à celui de l'ADL. Ici au vue du nombres de données, on choisira plutôt l'Analyse Discriminante Linéaire.

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

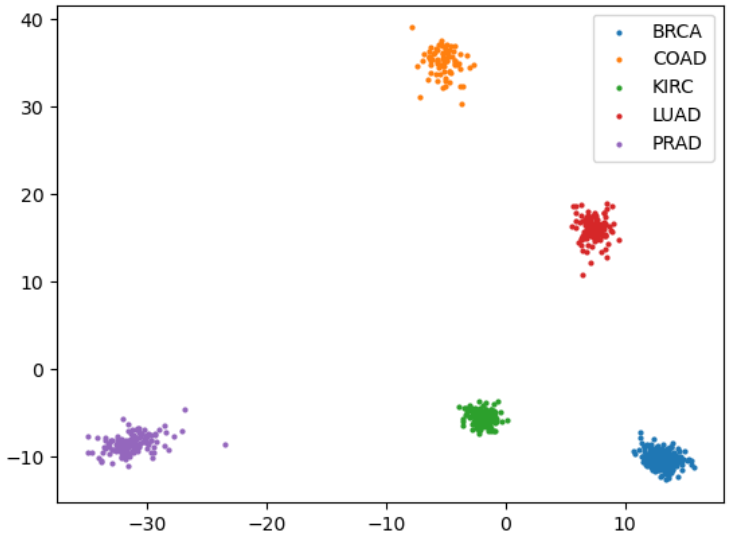


FIGURE 7 – Nuage de points Analyse discriminante linéaire