# Projet 3 Apprentissage Statistique : Méthodes non paramétriques

Léo Pichery, Adèle Dejoie

Avril 2025

## Table des matières

# Introduction

Nous nous intéressons ici à un jeu de données portant sur 5 types de tumeur extrait de la base synapse (Weinstein, John N., et al. 'The cancer genome atlas pan-cancer analysis project.' Nature genetics 45.10 (2013)). L'objectif de ce projet est de proposer des méthodes pour classer différents types de tumeurs selon l'expression de certains gènes.

Nous utiliserons pour cela différentes méthodes telle que la méthode des k-plus proches voisins, des méthodes d'abre de décisions et d'agrégations de forêts aléatoire ainsi que AdaBoost.

# 1 Analyse des données

## 1.1 Description

La base de données que nous utilisons étudie 12 différentes tumeurs mais nous ne considérerons qu'une restriction de ce jeu avec 420 échantillons (sur plus de 5000) et 5 tumeurs différentes présentées dans le tableau ci-dessous. La figure 1 représente la répartition des échantillons selon les différentes tumeurs.

| Symboles | PRAD | LUAD | BRCA | KIRC | COAD |
|---|---|---|---|---|---|
| Partie du corps | Prostate | Poumon | Sein | Rein | Colon |
| Nombre d'échantillons (sur 420) | 72 | 80 | 150 | 86 | 32 |

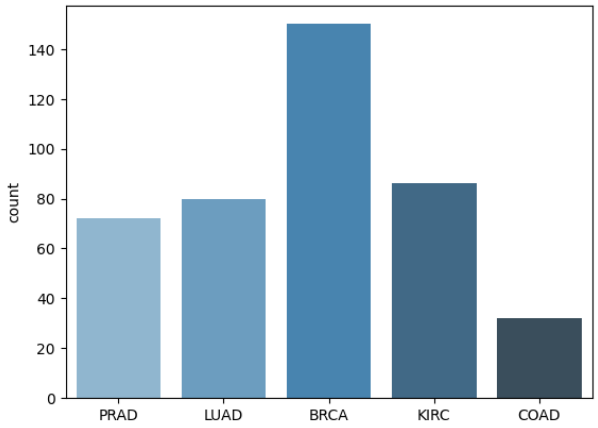


FIGURE 1 – Répartition des différents types de tumeur

En ce qui concerne les variables explicatives, nous avons pour chacun des échantillons considérés leur "expression" pour 20531 gènes différents. Ces expressions sont traduites dans les données par des nombres, de l'ordre de l'unité, et sont issus d'un séquençage de l'ARN.

En étudiant les moyennes et écart-type des valeurs associées aux différents gènes, nous pouvons extraire des couples de gènes grâce auxquels il semble être possible de classifier les échantillons (dans le sens où nous avons l'impression de pouvoir séparer linéairement les tumeurs, ce qui serait gage de réussite des méthodes de LDA et d'arbres de décisions). On peut voir 2 exemples de ces couples sur les figures 2 et 3.

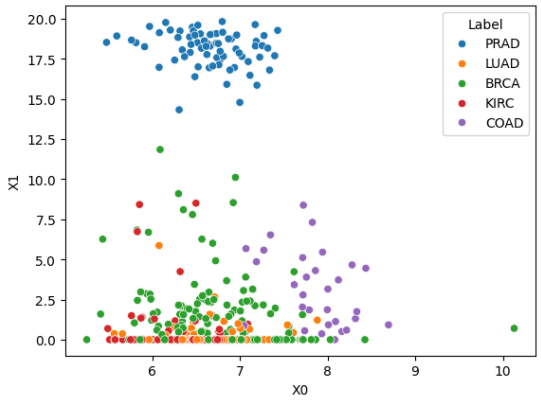


FIGURE 2 – Couple des gènes 3 et 9176

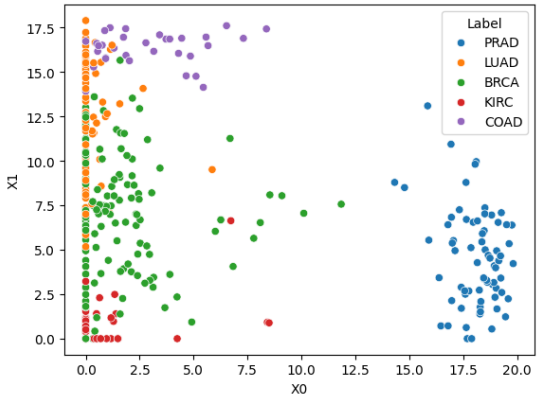


FIGURE 3 – Couple des gènes 9176 et 3540

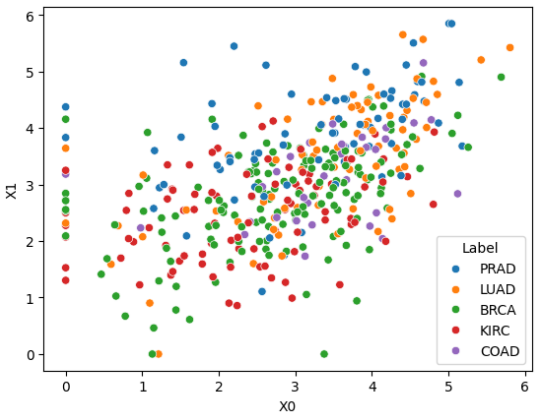


FIGURE 4 – Exemple de couple de gènes (1 et 2) qui ne semble pas permettre de classifier les tumeurs

Il paraît alors raisonable de penser que nous pourrions être capable de classifier les différentes tumeurs en ne considérant que certains gènes particulier ou composantes principales issues de l'ACP.

## 1.2 Standardisation et normalisation

Les expressions de chaque échantillon sur les différents gènes sont du même ordre de grandeur (au plus de la dizaine) et donc la standardisation ne semble pas nécessaire. Par ailleurs, standardiser les données donnerait la même importance à chaques gènes alors que nous avons vu que certains pouvaient permettre de d'isoler des tumeurs : nous décidons donc de ni standardiser, ni normaliser les données (sauf pour la méthode des k-plus proches voisons où cela sera nécessaire).

## 1.3 Analyse en composante principale

Nous avons donc fait une Analyse en Composante Principale pour déterminer si nous pouvions réduire le nombre de variables explicatives, sans sacrifier une quantité trop importante d'informations.

Les pourcentages d'inertie expliquées pour les 20 premières composantes principales sont représentés sur la figure 5.

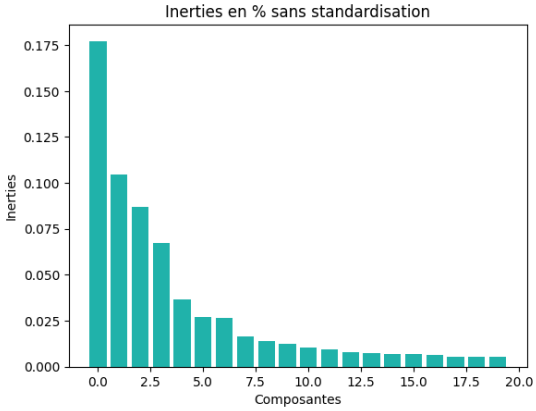


FIGURE 5 – Diagramme représentant l'inertie expliquée selon les 20 premières composantes principales.

En effectuant la somme cumulée de ces pourcentages d'inertie, les 20 premières composantes expliquent 64% de l'information totale ce qui témoigne d'une perte assez importante d'information. Néanmoins, dans ce cadre, le nombre de variables est divisé par 1000 donc nous décidons de tout de même étudier nos données avec uniquement ces 20 premières composantes pour voir si nous pouvons obtenir des résultats satisfaisant.

La représentation des données dans le plan factoriel délimité par les deux premières composantes principales nous donne la figure 6.

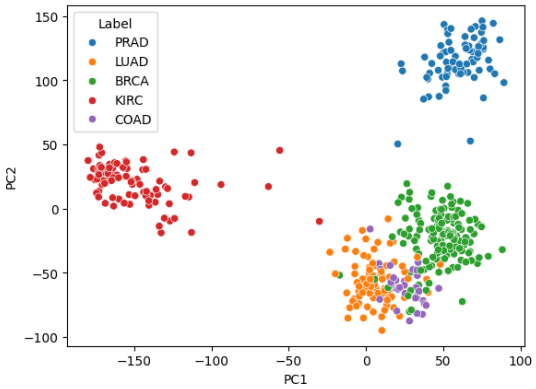


FIGURE 6 – Représentation des points dans le plan des deux premières composantes principales de l'ACP.

# 2 Une classe pour chaque tumeur

Après avoir observé et analysé la structure de notre jeu de données, nous allons essayer de classifier nos 5 tumeurs.

## 2.1 LDA

On effectue tout d'abord une analyse discriminante linéaire. Pour cela, parmi l'échantillon qui nous est fourni, on ne garde que 30% des données pour entraîner un classifieur,

et on le teste sur les 70% restantes afin d'éviter le risque de sur-entraînement. On obtient alors la matrice de confusion de la figure 7.

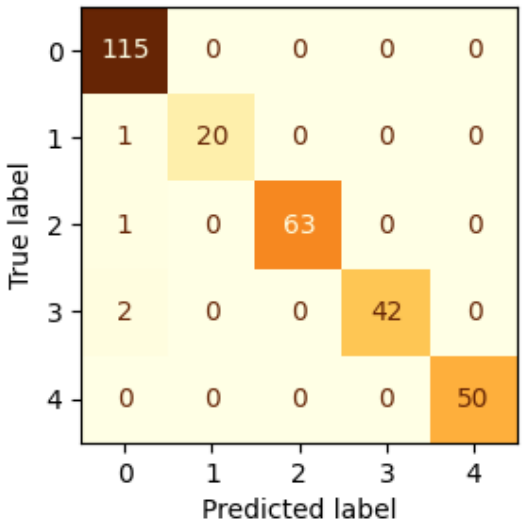


FIGURE 7 – Matrice de confusion pour la LDA

On peut constater quelques erreurs mais le résultat reste très bon dans l'ensemble. Cela est confirmé par le score en validation croisée qui montre que le taux de tumeurs bien prédite en moyenne est de 0,993.

Nous pourrions envisager de ne pas considérer d'autres méthodes puisque celle-ci marche déjà très bien mais la LDA étant notre méthode la moins coûteuse en terme de calcul, cela semble donc prometteur pour la suite de notre étude. En effet, nous pourrions trouver des méthodes encore plus précises et donc plus fiables pour classifier les différentes tumeurs. Par ailleurs, la LDA ne nous permet pas d'avoir d'informations sur les différents gènes alors qu'une méthode comme les arbres de décisions pourra nous permettre d'interpréter les gènes les plus importants à la classification.

## 2.2 k-plus proches voisins

Nous avons, dans un second temps, utilisé la méthode des k-plus proches voisins. Comme pour la LDA, nous entraînons le classifieur sur 30% des données et testons sur les 70% restantes afin d'éviter le risque de sur-entraînement.

Pour cette méthode, nous utilisons les données standardisées. Il nous faut alors trouver la meilleure métrique (f1-score, accuracy..) ainsi que le meilleur paramètre k, qui représente le nombre de voisins.

Au départ, nous avons pris un nombre k=10 arbitraire pour visualiser ce que la méthode donnait comme résultats et nous avons tout de même obtenu de très bons résultats. Dans notre cas, nous n'avons pas de type d'erreur à prioriser mes les courbes ROC sont très satisfaisantes comme le montre la figure 8.

On utilise ensuite la méthode de validation croisée pour déterminer la meilleure métrique ainsi que le meilleur paramètre k (qui permet de réduire l'erreur moyenne et la variance

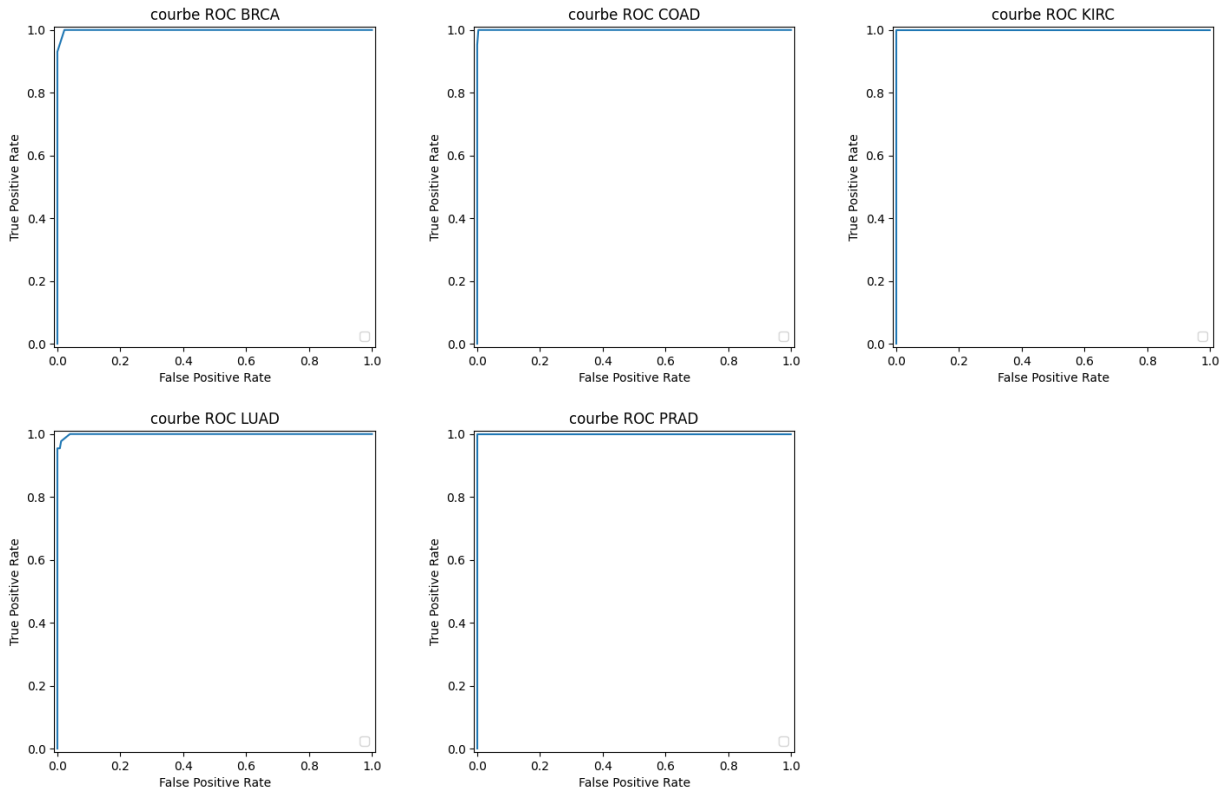


FIGURE 8 – Courbes ROC avec $k = 10$

des résultats) et on obtient pour ces paramètres la matrice de confusion suivante :

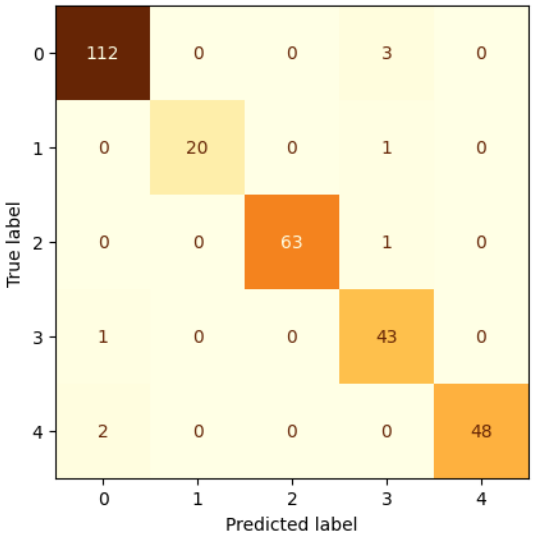


FIGURE 9 – Matrice de confusion pour k=1

Les scores associés aux différentes métriques sont :

- Accuracy : 0.97
- F1-score : 0.97
- Précision : 0.97
- Recall : 0.97

Ces scores, bien qu'il n'y ait ni vrais positifs, ni faux positifs etc parce qu'il y a cinq classes

différentes, sont calculés en faisant une moyenne. On a mis pour cela "average='weighted'" dans la commande pour le calcul de ces scores.

## 2.3 Abres de décisons, Agrégation de forêts aléatoires

Après la méthode des plus proches voisins, on essaie maintenant de classifier nos données avec la méthode des arbres.

Cette méthode a l'avantage de ne pas nécessiter de standardiser nos données contrairement à celle des k-plus proches voisins, mais elle nous permet surtout d'interpréter les résultats que l'on obtient.

On travaille encore une fois sur nos mêmes échantillons *train/test*. Nous essayons tout d'abord sur un premier abre de profondeur arbitraire prise égale à 4 et on obtient l'arbre de la figure 10. On observe que dans les dernières "boîtes", la liste *value* ne contient qu'un élément (qui représente donc une tumeur) ce qui indique que l'arbre sépare parfaitement les différentes tumeurs sur l'échantillon d'entraînement .

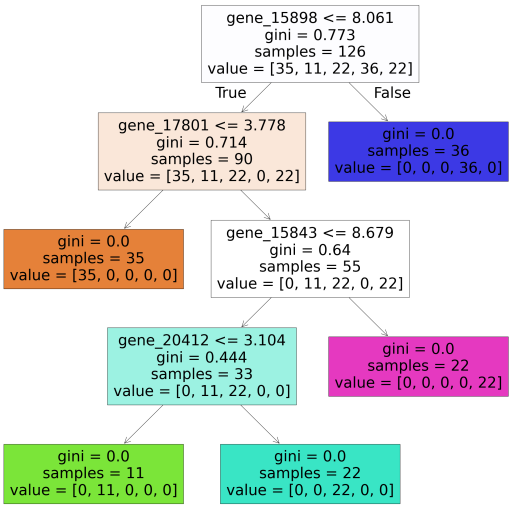


FIGURE 10 – Arbre de profondeur 4

On peut alors observer les différents gênes utiles pour séparer et donc classifier ces données en efectuant un arbre sur l'ensemble de nos données. Ce sont donc les gènes 6875, 12983, 15896 et 18746 qui sont déterminants. On observe cela sur la figure 11. En effet, les deux graphiques supérieurs montrent bien qu'il est facile de séparer BRCA et KIRC des autres tumeurs. Le troisième graphique comme BRCA et KIRC ont déjà été triés, il faut juste comparer LUAD (en orange) à COAD et PRAD (respectivement en violet et bleu). On observe alors qu'on peut encore bien les séparer. On effectue encore le même raisonnement pour le dernier graphique entre LUAD et COAD.

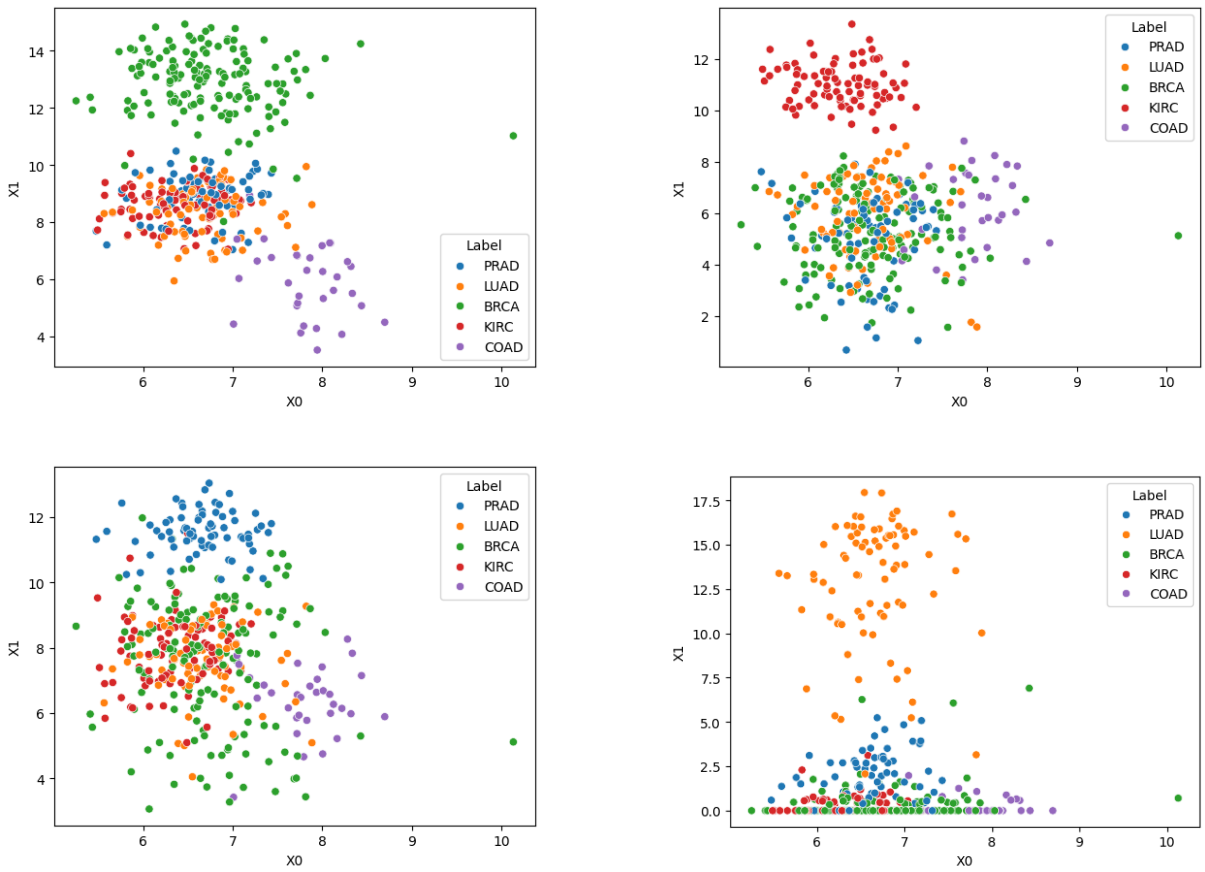


FIGURE 11 – Séparation en fonction des gènes

Avec l'arbre de la figure 10, on obtient la matrice de confusion représentée en figure 12 ainsi que les résultats en score suivant :

- Accuracy : 0.92
- F1-score : 0.92
- Précision : 0.92
- Recall : 0.92

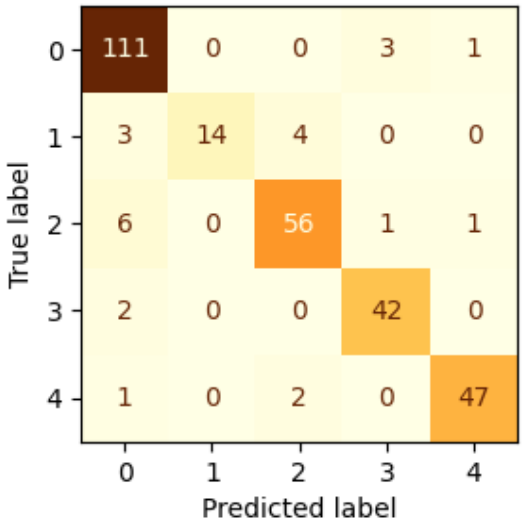


FIGURE 12 – Matrice de confusion pour l'arbre de profondeur 4

Ces résultats, bien que moins bons que ceux obtenus par la méthode des k-plus proches voisins, ont comme intérêt de nous renseigner sur les gènes jouants un rôle dans la classification des tumeurs.

On cherche maintenant à l'aide de la commande *GridSearchCV* si l'on peut obtenir un arbre plus complet. On obtient que le meilleur estimateur serait l'arbre avec comme profondeur 14 et max_feature = 13 que l'on observe sur la figure 14. Sa matrice de confusion est représentée en figure 13.

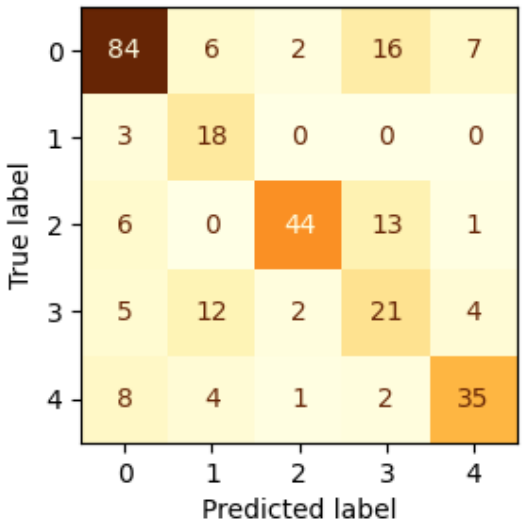


FIGURE 13 – Matrice de confusion pour l'arbre obtenu avec GridSearchCV

On peut également comparer les scores obtenus :

- Accuracy : 0.66
- F1-score : 0.66
- Précision : 0.68
- Recall : 0.67

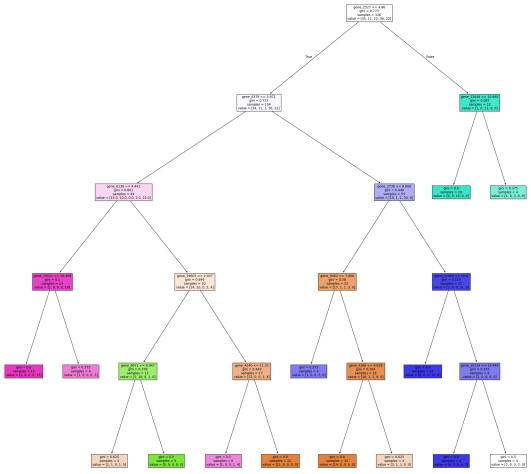

FIGURE 14 – Arbre obtenu avec GridSearchCV

Il reste maintenant à essayer la méthode d'agrégation des forêts qui devrait améliorer les résultats et c'est ce que l'on observe dans la figure 15. On obtient alors les scores suivants qui sont nettement plus élevés que ceux obtenus avec un seul arbre :

- Accuracy : 0.93
- F1-score : 0.93
- Précision : 0.94
- Recall : 0.93

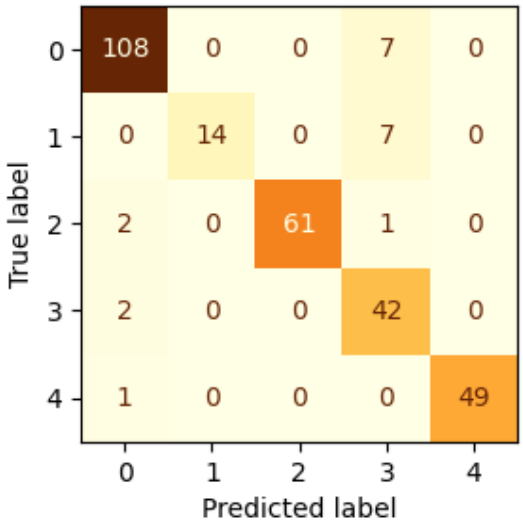


FIGURE 15 – Matrice de confusion obtenue avec la forêt

## 2.4 AdaBoost

Nous avons ensuite appliqué la méthode Adaboost pour classer les tumeurs. Toutefois, son utilisation directe sur l'ensemble des variables explicatives s'est révélée trop lente et coûteuse en ressources. Pour contourner cette difficulté, nous avons réduit la dimensionnalité des données à l'aide de l'Analyse en Composantes Principales (ACP), ce qui n'était pas envisageable dans le cas des arbres afin de préserver l'interprétabilité des résultats.

Nous avons ensuite utilisé la commande `GridSearchCV` pour rechercher les meilleurs hyperparamètres du modèle. L'optimisation a conduit à la configuration suivante : `n_estimators = 300` et `learning_rate = 0.01`.

Avec ces paramètres, les performances du modèle sont excellentes, comme le montre la matrice de confusion de la figure 16. Les scores obtenus sont :

— Accuracy : 0.97
— F1-score : 0.97
— Précision : 0.98
— Recall : 0.97

Ces résultats démontrent qu'Adaboost est une méthode très efficace pour la classification des tumeurs dans notre jeu de données.

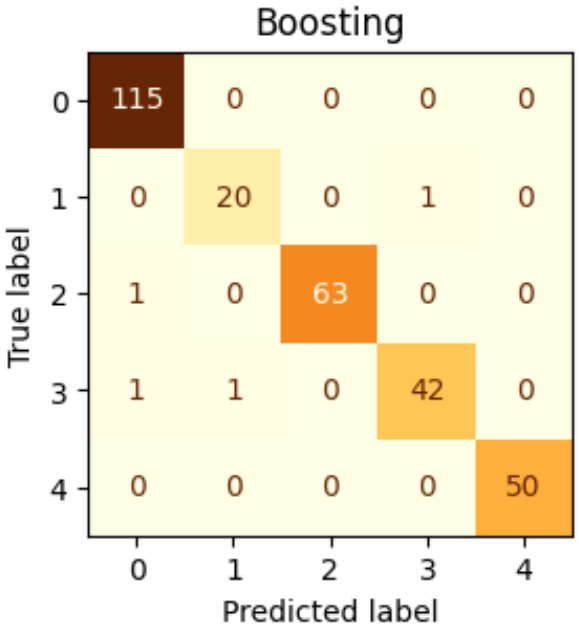


FIGURE 16 – Matrice de confusion par la méthode Adaboost

# 3 Seule contre tous

Après avoir tenté de différencier les cinq types de tumeurs entre eux, nous adoptons ici une approche binaire, visant à identifier une tumeur particulière parmi l'ensemble des autres. Cette stratégie repose sur la création d'un classifieur opposant une classe cible à toutes les autres, regroupées sans distinction.

Dans cette étude, nous choisissons de détecter spécifiquement les tumeurs de type BRCA (cancer du sein). Les autres types de tumeurs (LUAD, PRAD, KIRC, COAD) sont réunis dans une seule classe appelée OTHER. Dans les matrices de confusion, la classe BRCA sera codée par 0, et les autres par 1.

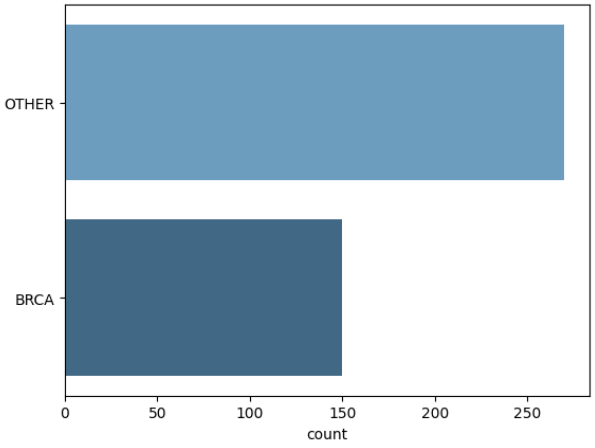


FIGURE 17 – BRCA vs OTHER

Nous allons réappliquer les différentes méthodes de classification utilisées précédemment afin d'évaluer leurs performances dans ce cadre binaire.

## 3.1 ACP

Comme dans la première partie, nous commençons par une Analyse en Composantes Principales afin de visualiser la séparabilité des classes dans un espace réduit. La projection des données sur les deux premières composantes principales révèle que les échantillons de la classe BRCA semblent se distinguer assez nettement des autres tumeurs.

Cela suggère qu'une classification binaire pourrait être efficacement menée, même avec une réduction de dimension importante.

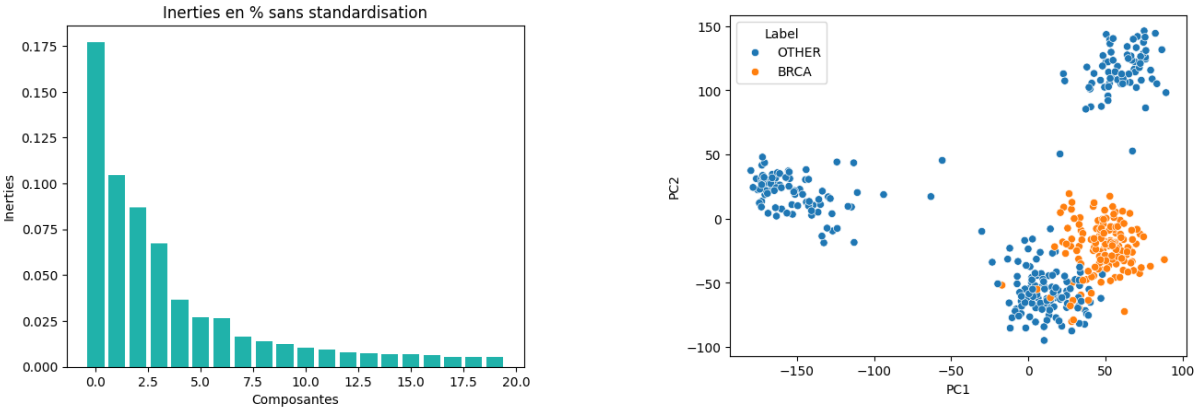


FIGURE 18 – ACP

## 3.2 LDA

Nous appliquons ensuite l'Analyse Discriminante Linéaire pour évaluer sa capacité à distinguer la classe BRCA des autres. Comme précédemment, nous utilisons une validation croisée pour estimer la performance du modèle.

Le taux moyen de bonne classification obtenu atteint environ 99,7%, confirmant l'efficacité de la LDA dans ce contexte binaire.

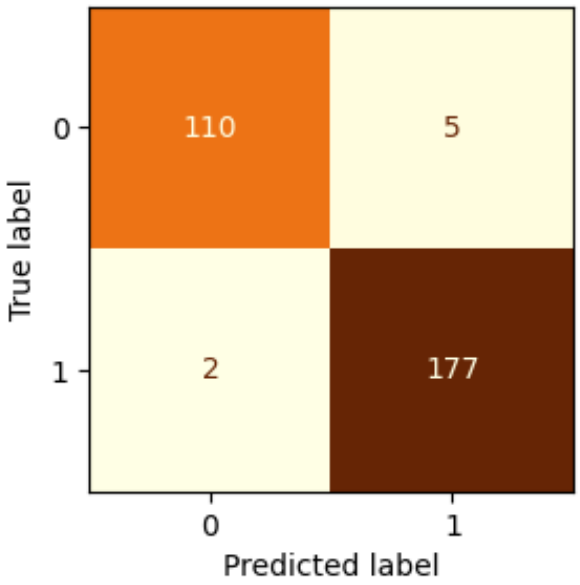


FIGURE 19 – LDA pour les deux classes choisies (BRCA et OTHER)

## 3.3 k-pus proches voisins

Nous testons ensuite la méthode des $k$-plus proches voisins. Dans un premier temps, un choix arbitraire de $k = 10$ permet déjà d'obtenir une courbe ROC très satisfaisante.

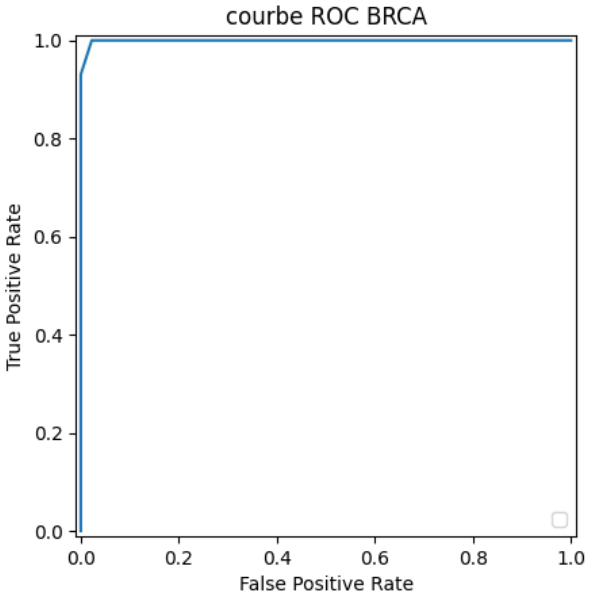


FIGURE 20 – Courbe ROC BRCA

En affinant le paramètre $k$ via validation croisée, nous trouvons que la meilleure valeur est $k = 3$. La matrice de confusion associée et les scores obtenus sont excellents :

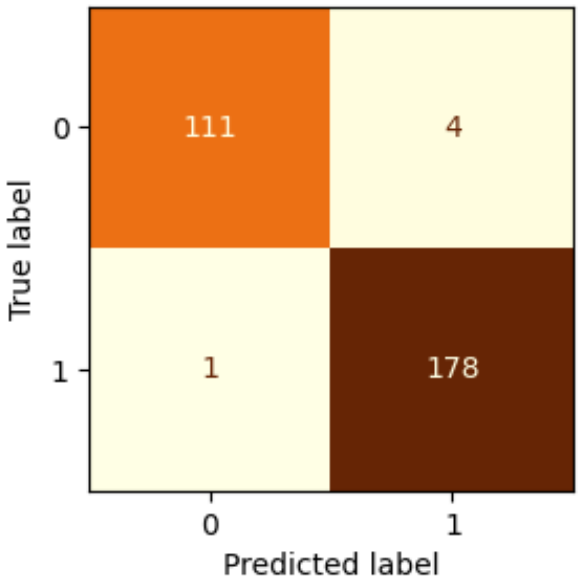


FIGURE 21 – k-NN pour deux classes et $k = 3$

— Accuracy : 0.983
— F1-score : 0.983
— Précision : 0.983
— Recall : 0.933

Ces résultats confirment que le classifieur $k$-NN est particulièrement efficace pour isoler les tumeurs BRCA.

## 3.4 Abres de décisons

Nous explorons ensuite les arbres de décision. En imposant une profondeur maximale de 4, on observe que l'arbre n'utilise en fait que deux niveaux. Le gène n°17801 est notamment un facteur de séparation très efficace.

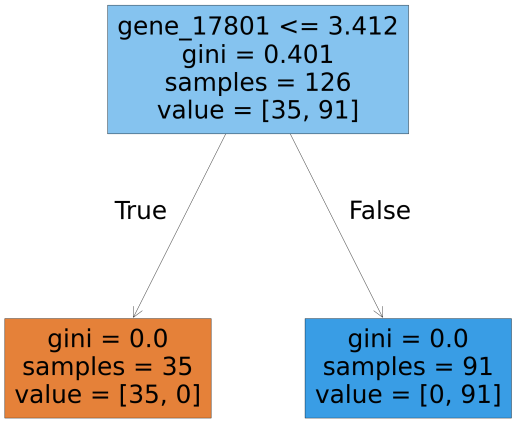


FIGURE 22 – Arbre de profondeur maximale 4 trouvé

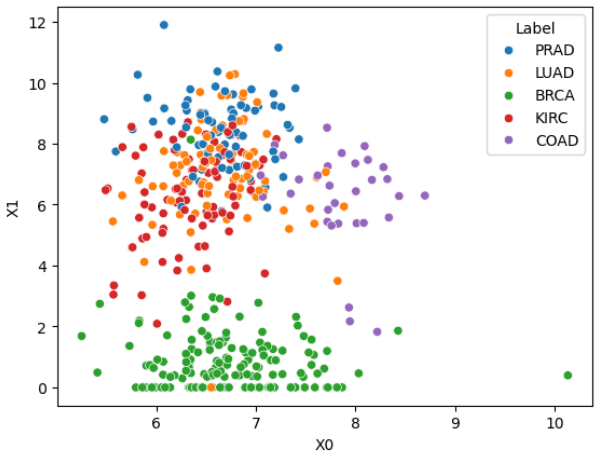


FIGURE 23 – Séparation en fonction du gêne 17801

Les performances de cet arbre simple sont déjà très bonnes :

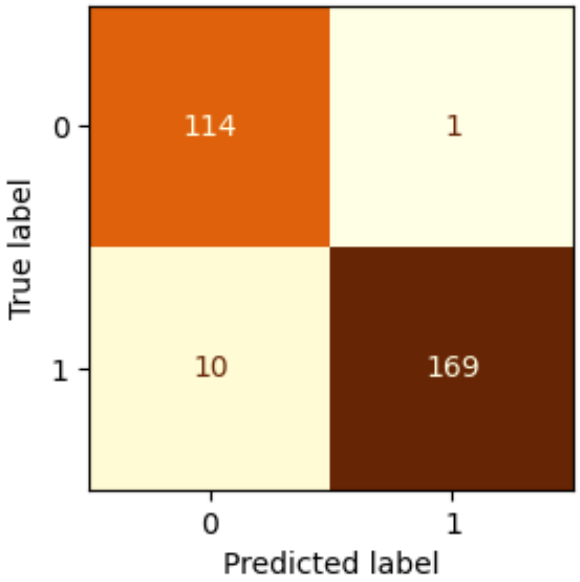


FIGURE 24 – Matrice de confusion pour l'arbre de profondeur 2

— Accuracy : 0.963
— F1-score : 0.963
— Précision : 0.965
— Recall : 0.963

Une tentative d'optimisation via `GridSearchCV` aboutit à des hyperparamètres (profondeur = 6, `max_features` = 6, `min_samples_leaf` = 6) qui donnent cependant des résultats inférieurs :

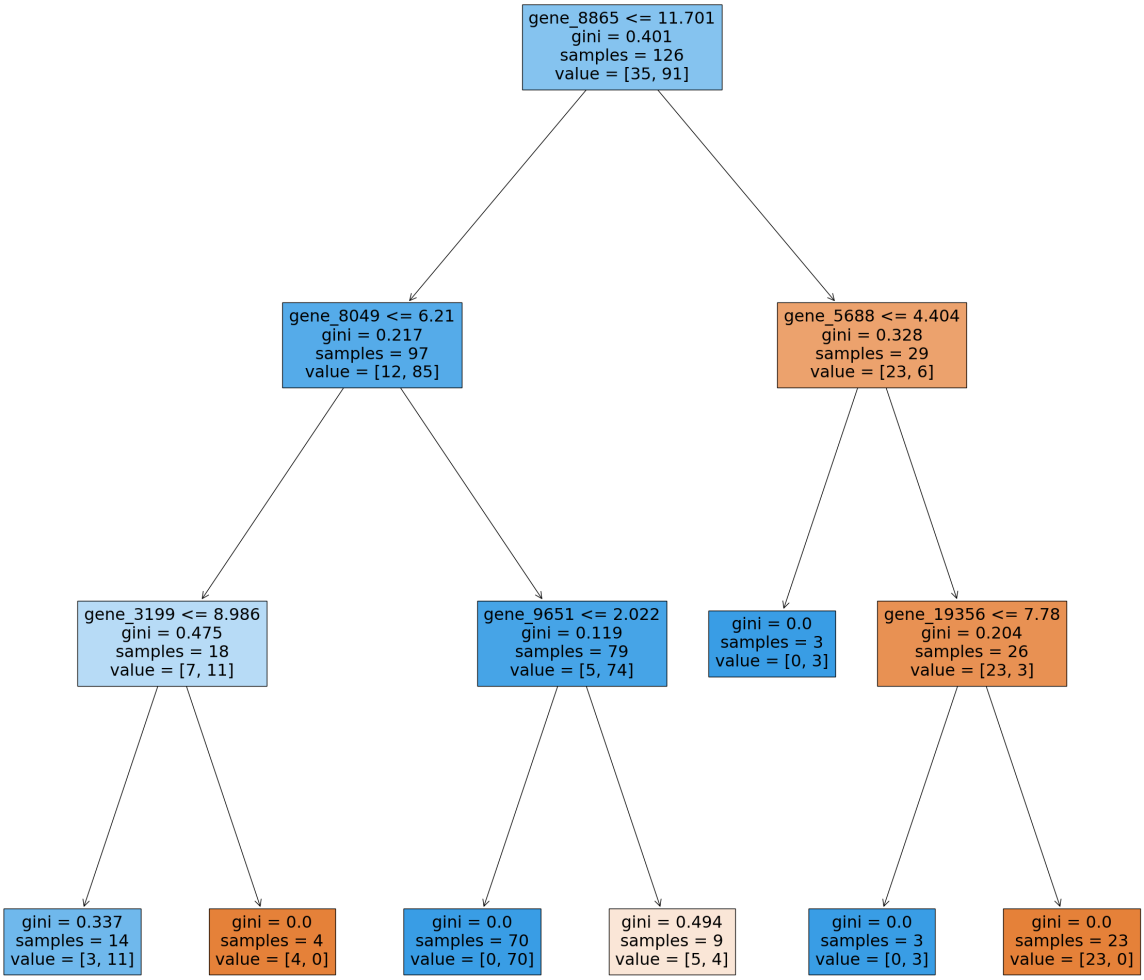


FIGURE 25 – Deuxième arbre obtenu

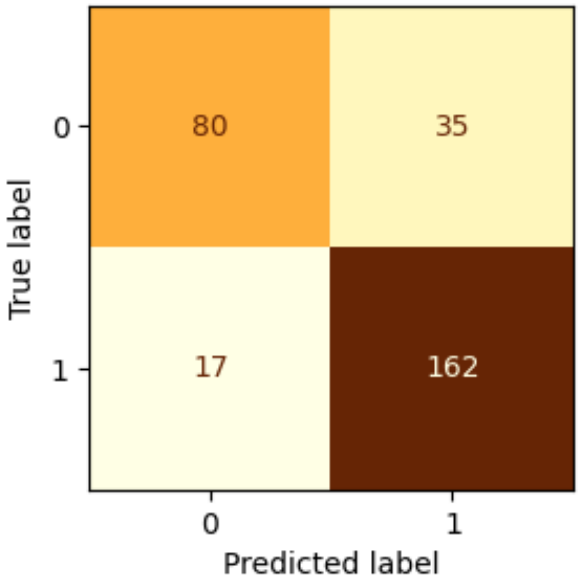


FIGURE 26 – Matrice de confusion pour le deuxième arbre

— Accuracy : 0.823
— F1-score : 0.853
— Précision : 0.82
— Recall : 0.823

Ces résultats paradoxaux montre qu'une simple structure d'arbre peut parfois surpasser des modèles plus complexes.

## 3.5 AdaBoost

Enfin, nous appliquons à nouveau la méthode Adaboost, qui avait montré d'excellentes performances dans la classification multi-classes. Après optimisation des hyperparamètres par `GridSearchCV` (`n_estimators = 100`, `learning_rate = 0.01`), nous obtenons :

— Accuracy : 0.925
— F1-score : 0.925
— Précision : 0.925
— Recall : 0.925
— Taux de bonne classification moyen (validation croisée) : 0.979

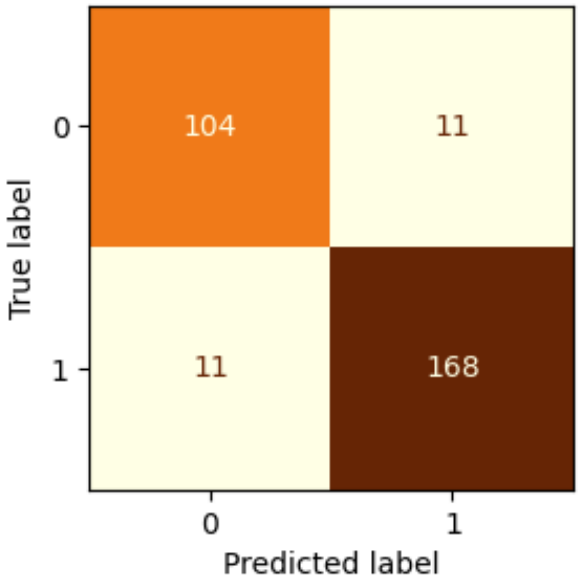


FIGURE 27 – Matrice de confusion obtenue à partir de Adaboost

Ces résultats confirment la robustesse et l'efficacité d'Adaboost pour détecter les tumeurs BRCA dans une tâche de classification binaire.

# 4 Conclusion

Que ce soit dans le cadre de la classification multi-classes (cinq types de tumeurs) ou dans une approche binaire (identifier une tumeur spécifique parmi les autres), les différentes méthodes de classification testées se sont révélées globalement très performantes.

L'Analyse Discriminante Linéaire et Adaboost se distinguent particulièrement en offrant d'excellents taux de prédiction, tout en conservant des coûts de calcul relativement faibles. Ces méthodes constituent donc des choix pertinents pour une application à grande échelle ou en contexte médical, où la rapidité d'exécution et la fiabilité sont essentielles.

Par ailleurs, l'utilisation des arbres de décision, bien que légèrement moins performante en termes de scores, a permis d'identifier certains gènes particulièrement discriminants pour différencier les types de tumeurs. Cette capacité d'interprétation représente un atout important, notamment pour la recherche biomédicale.

Ces résultats laissent entrevoir de nombreuses perspectives. En particulier, on peut envisager d'étendre ces approches à un nombre plus important de types de tumeurs — comme les 20 classes présentes dans la base complète de l'Atlas du Génome du Cancer — pour aider à affiner le diagnostic et améliorer la compréhension des mécanismes génétiques associés aux cancers. Les méthodes d'apprentissage supervisé apparaissent ainsi comme un outil puissant au service de la médecine personnalisée.