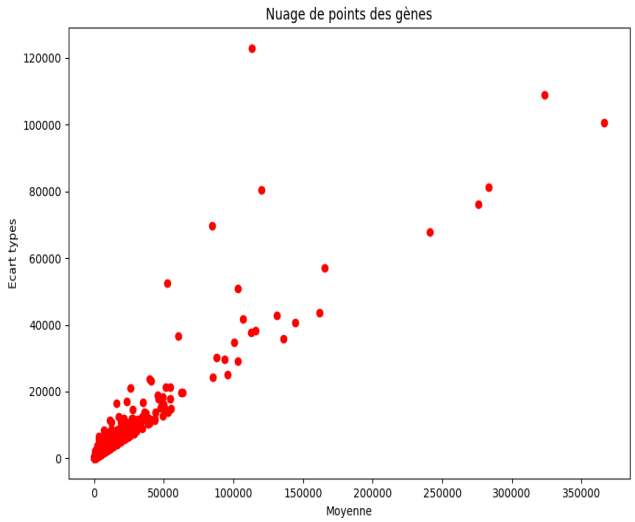
Sam VALLET
Thomas SINAPI

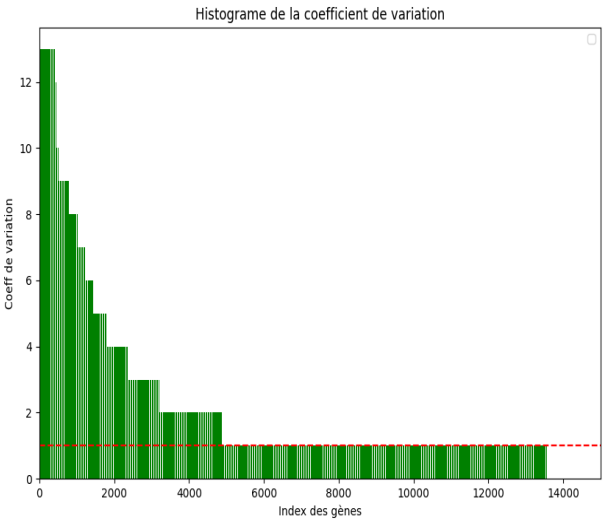
Rapport Projet Bio Info

La sclérose latérale amyotrophique (SLA) est une maladie neurodégénérative grave qui affecte les motoneurones, lesquels sont responsables de la commande des muscles volontaires. Elle peut entraîner une paralysie progressive et aboutir au décès en quelques années. A ce jour, il n'existe aucun traitement curatif pour cette maladie.

Dans le cadre de ce projet de bioinformatique, notre objectif était d'identifier des biomarqueurs associés à la maladie de la SLA en analysant des donnés de séquençage RNA-Seq à partir de biopsies du cortex cérébral post-mortem de personnes atteintes de la SLA et de personne non atteinte. On disposait donc d'informations sur le nombre d'ARN de chaque gène pour plusieurs individus, qu'ils soient atteints de la SLA ou non. Ces données proviennent du séquençage RNA-Seq de biopsies du cortex cérébral post-mortem.

Pour commencer, nous avons créé un dataframe à partir de ces données et nous avons réalisé les premières observations.

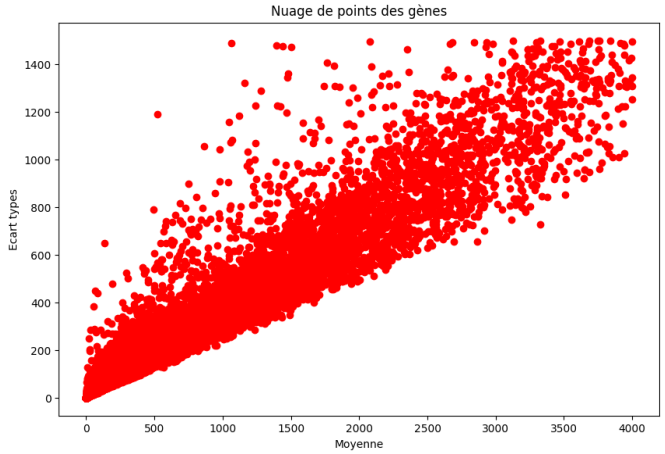


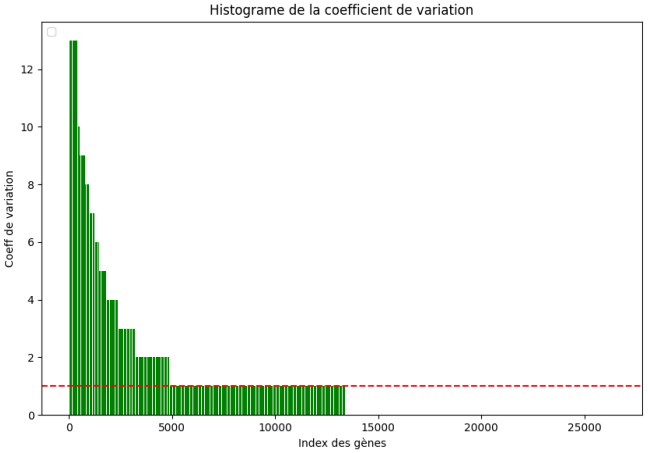


On a commencé par regarder le coefficient de variation des différents gènes (l'écart type en fonction de la moyenne). On peut déjà remarquer que certains gènes ont des valeurs particulièrement grandes par rapport aux autres. Nous allons donc chercher à créer un dataframe sans ces gènes-là.

Cela nous amène donc à la création de notre dataframe V5 (c’est la cinquième version nous avons réalisé plusieurs data frame ou on enlevait au fur et a mesur les gènes trop éloigné des autres)

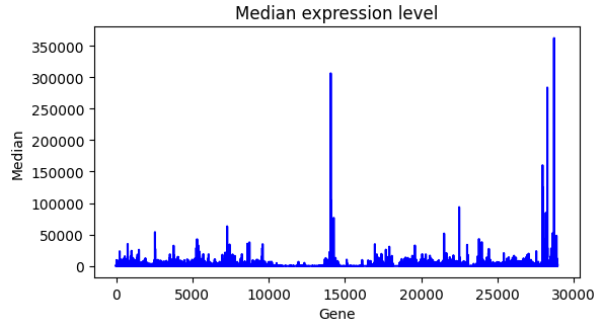
Sam VALLET
Thomas SINAPI

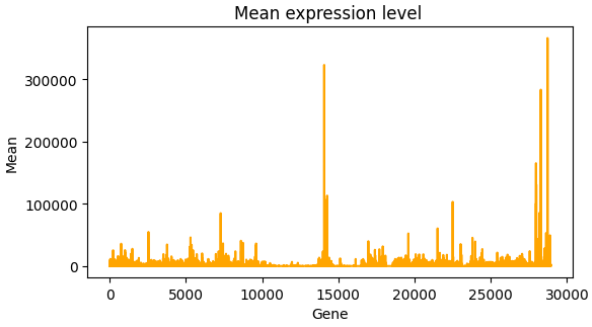


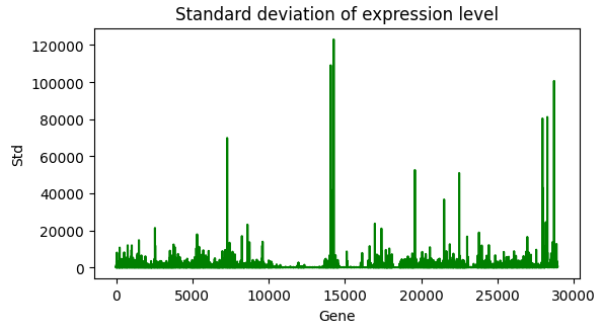


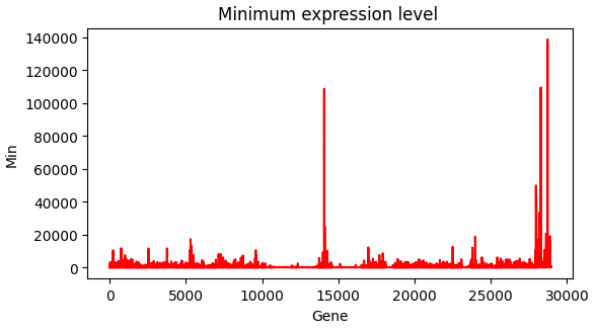
Voici les observations du data frame V5, on voit que les points sont proche les uns des autres contrairement au data frame v1. On a donc fait le choix de poursuivre nos observations sur ce data frame.

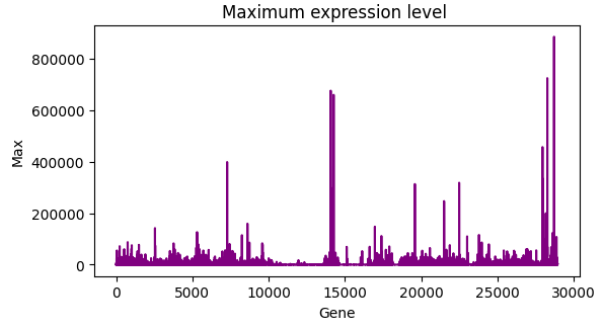
Le data frame V1 possède 28953 gênes et le data frame V5 27633. L'écart entre les deux n'est donc pas énorme.
Nous allons maintenant réaliser d'autre observations de base sur les data frame V1 et V5 V1 :



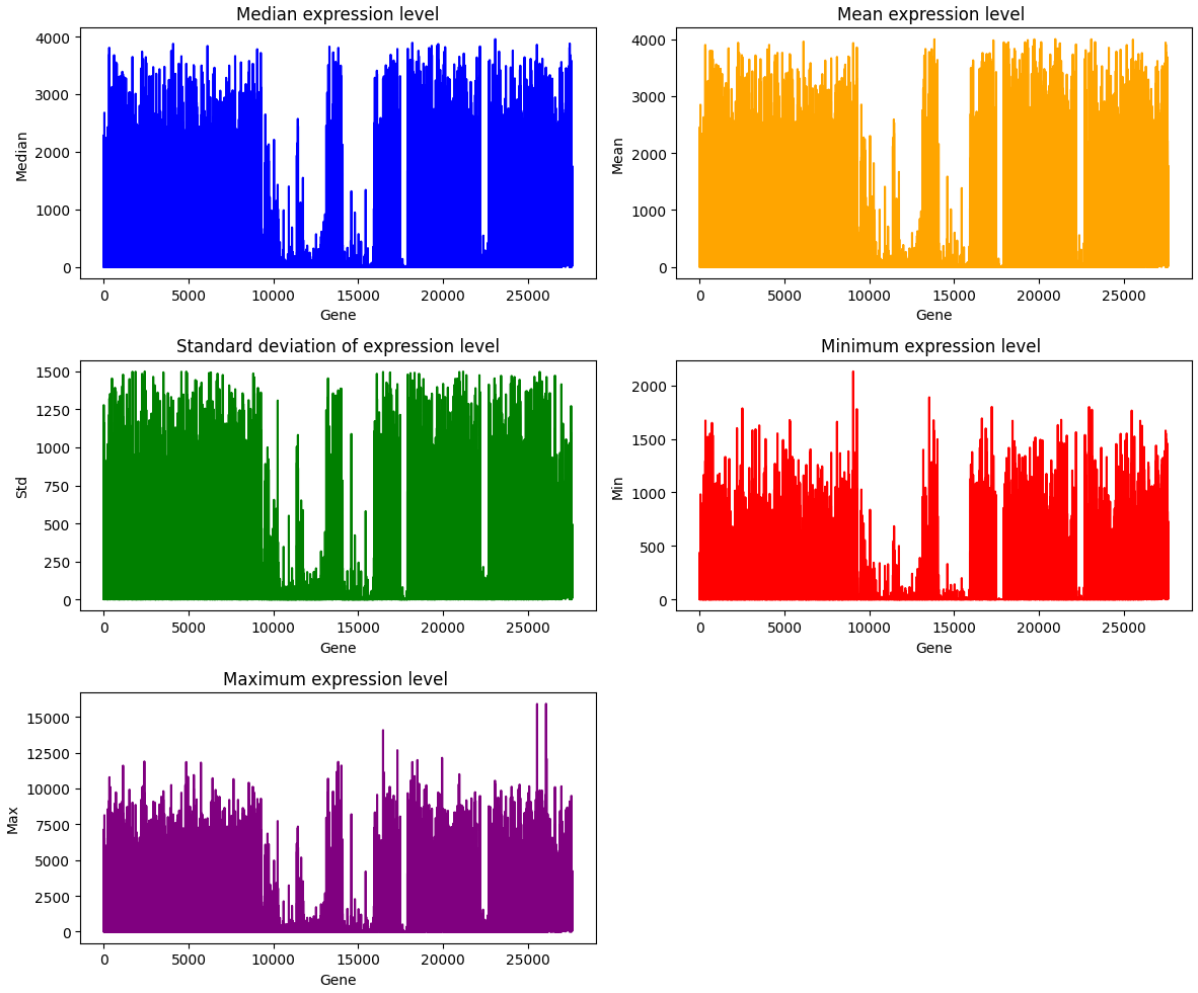






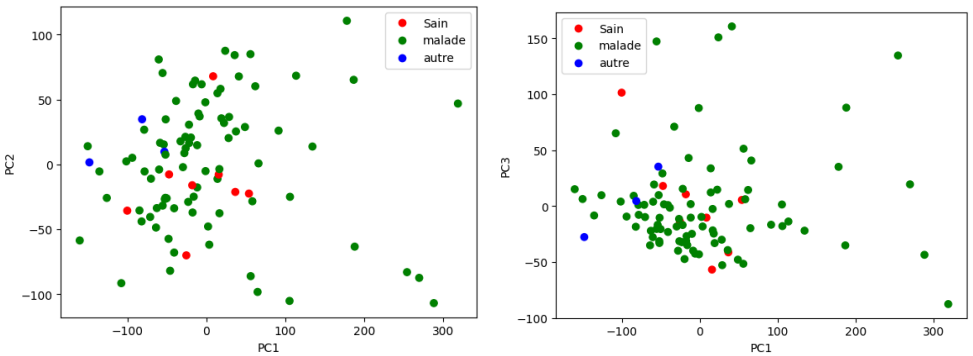


Sam VALLET
Thomas SINAPI



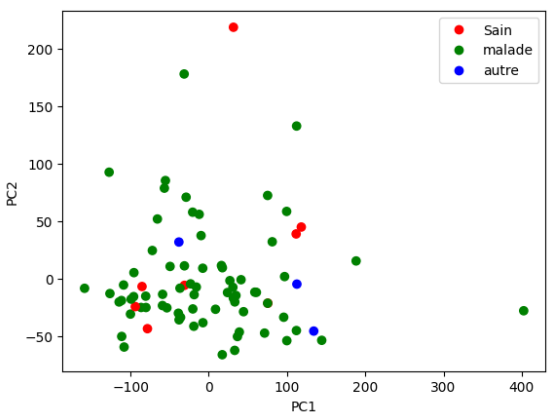
V5 ci-dessus.
On remarque que les gènes avec des valeurs extrêmement élevées dans le data frame V1 ne sont plus présents dans le V5, les gènes dans le V5 ont à peu près tous les mêmes valeurs sauf quelques gènes avec des valeurs relativement basses.

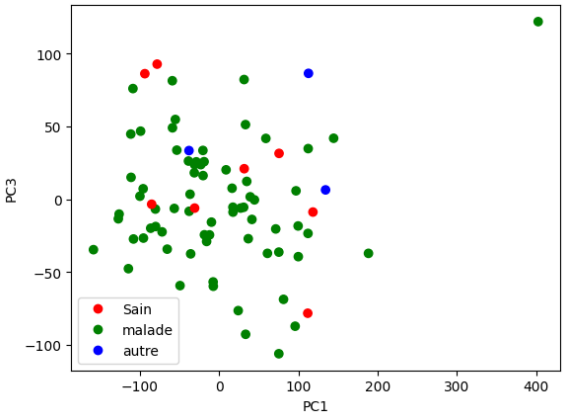
Nous avons ensuite réalisé une PCA sur nos données dans le but de pouvoir séparer les malades des non malades. Pour cela nous avons créé deux autres data frame, V8 et V9. Le V8 réunit les échantillons qui correspondent à la région “Frontal Cortex” et V9 la région ”Motor Cortex”. On a donc fait une PCA pour V8,V9 et V5
PCA V8 :



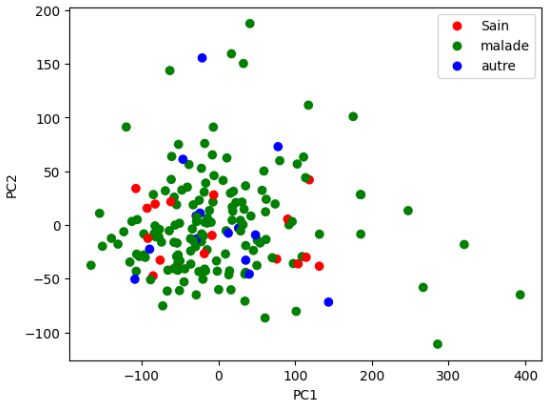
Sam VALLET
Thomas SINAPI

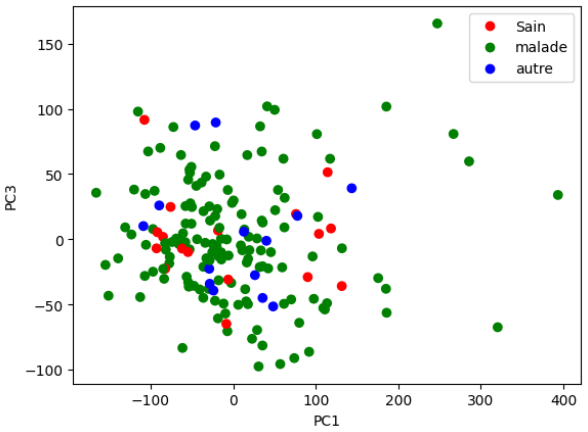
PCA V9 :





PCA V5 :



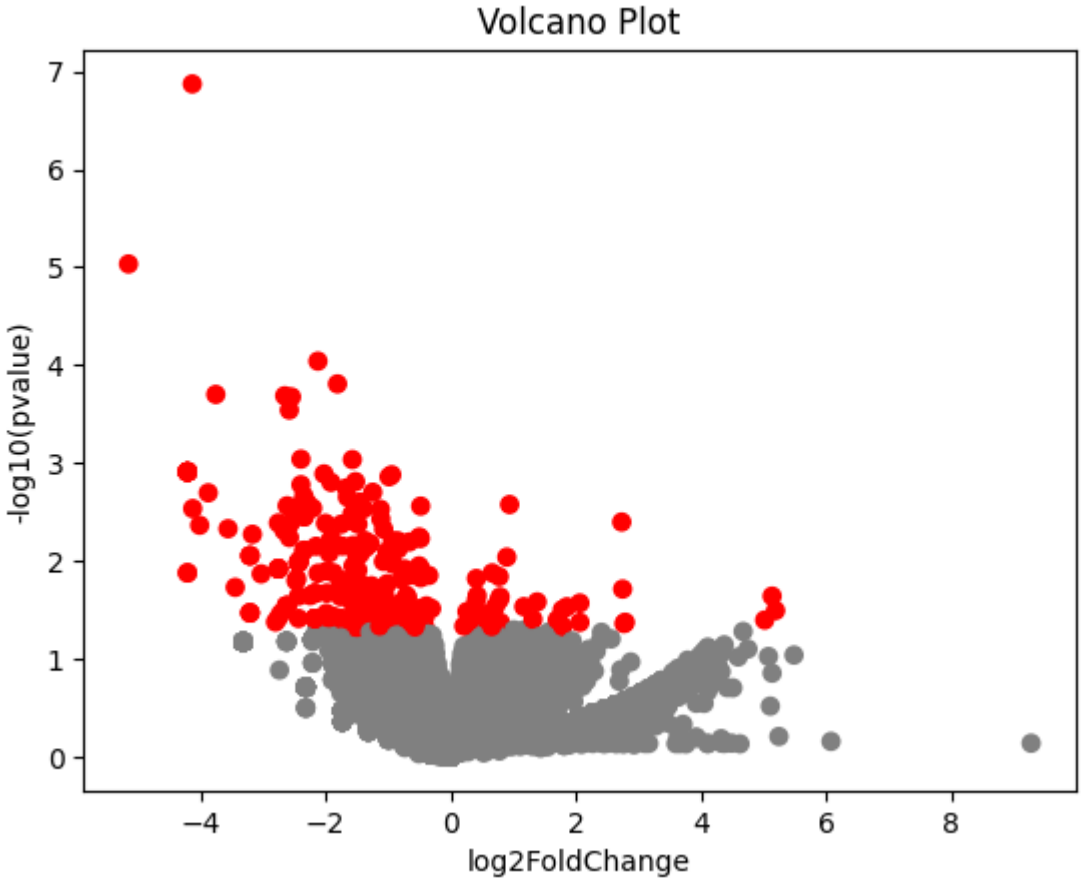


(Il y a plus d'image dans la partie code)

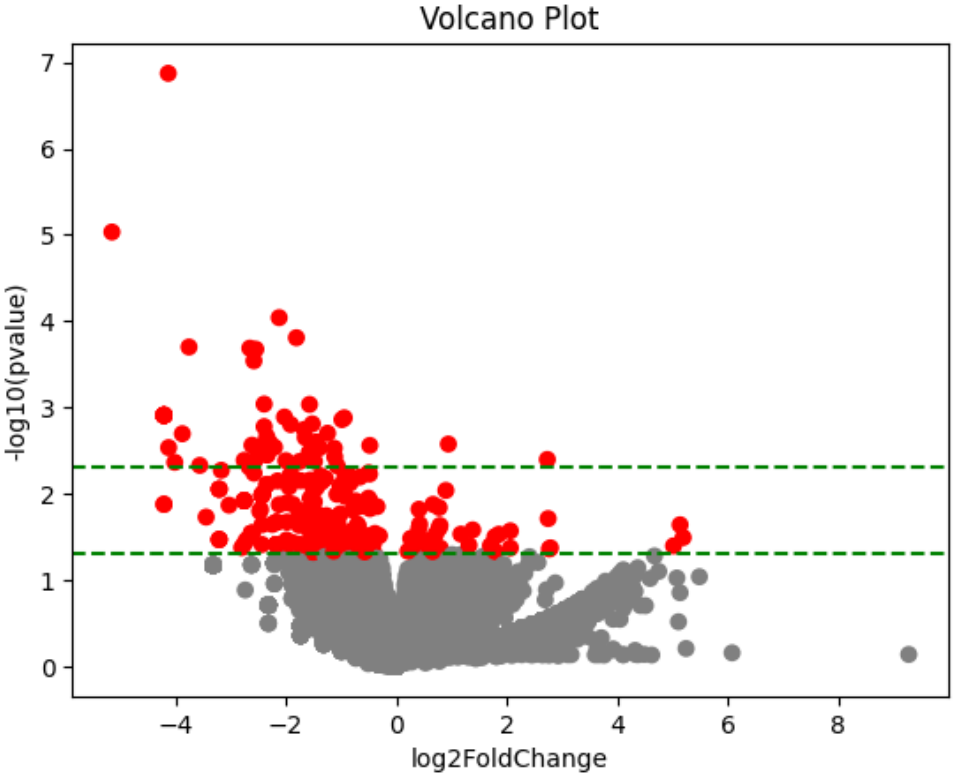
Dans les différentes PCA nous n'avons pas observé une séparation évidente entre les points vert et rouge (malade et non malade).

Sam VALLET
Thomas SINAPI

On a ensuite regardé les P-Values et réalisé un volcano plot. Les



Les points rouges correspondent aux gènes qui ont une p-value inférieur à 0,05.



Les gènes qui nous intéressent sont donc ceux entre les deux traits verts (On considère que ceux au-dessus sont des valeurs trop éloignées pour les regarder). On a donc créé un nouveau data frame V9 qui n'a que ces gènes. On a donc un nouveau data frame avec 214 gènes.

Dans la partie 6, une régression logistique avec Elastic Net a été appliquée sur les données d'entraînement. L'objectif principal de cette étape était de construire un modèle capable de prédire si un échantillon appartient à un groupe de contrôle ou à un groupe de patients atteints de la maladie d'ALS. Pour ce faire, le modèle a été évalué en utilisant une validation croisée à 3 plis et en testant différentes valeurs pour les hyperparamètres **l1_ratios** et **Cs**. L'ensemble des données a été normalisé à l'aide de la méthode fit_transform() du StandardScaler.

Après avoir entraîné le modèle, il a été évalué sur l'ensemble des données d'entraînement et de test. Les résultats obtenus sont les suivants :

Précision du modèle sur les données d'entraînement : 100%
Précision du modèle sur les données de test : 65.789%

Ces résultats montrent que le modèle est capable de prédire correctement les étiquettes de groupe pour la majorité des échantillons. Cependant, il est important de noter que la précision sur les données d'entraînement est de 100%, ce qui peut indiquer un overfitting du modèle.

Dans la deuxième partie de la partie 6, les mêmes étapes ont été effectuées en utilisant le sous-ensemble de gènes sélectionnés dans la partie 5 (dataV9).
En utilisant ce sous-ensemble de gènes, les résultats obtenus sont les suivants :

Précision du modèle sur les données d'entraînement (dataV9) : 100%
Précision du modèle sur les données de test (dataV9) : 71.053%

Ces résultats montrent une légère amélioration de la précision sur les données de test par rapport au modèle précédent. Toutefois, la précision sur les données d'entraînement reste à 100%, ce qui suggère que le modèle peut encore présenter un overfitting.

Dans la partie 7, un modèle de classification XGBoost a été entraîné sur l'ensemble des données d'entraînement (dataV9) et évalué.
Une fois le modèle entraîné, nous avons extrait les importances de chaque gène pour identifier les gènes les plus importants. Les importances des sont des scores qui indiquent l'importance relative de chaque feature (gène) pour la prédiction du label (malade/contrôles).

Les 100 premiers gènes en termes d'importance ont été sélectionnés. Cette liste de gènes est utile pour une analyse ultérieure, par exemple pour l'identification de biomarqueurs potentiels pour le diagnostic de la maladie d'ALS.

Sam VALLET
Thomas SINAPI

Il est important de noter que, bien que les modèles de régression logistique avec Elastic Net et XGBoost aient été en mesure de prédire les étiquettes de groupe avec une précision relativement élevée, il existe des différences dans les résultats obtenus. Le modèle XGBoost a permis d'identifier les gènes les plus importants pour la prédiction, tandis que la régression logistique avec Elastic Net a donné des résultats plus simples à interpréter et à comprendre.

En conclusion, les deux approches présentées dans les parties 6 et 7 ont montré leur capacité à prédire les étiquettes de groupe (contrôle ou ALS) pour la majorité des échantillons. Toutefois, il est important de garder à l'esprit que la précision sur les données d'entraînement était de 100% pour les deux modèles, ce qui peut indiquer un overfitting. Par ailleurs,nous avons choisi de sélectionner nos gènes avec les importances des caractéristiques extraites à partir du modèle XGBoost mais d’autres critères peuvent être utilisées pour des analyses ultérieures et pour l'identification de gènes candidats en tant que biomarqueurs potentiels pour le diagnostic de la maladie d'ALS.
Ainsi, notre liste de gènes candidats est disponible dans le fichier top_100_genes.csv.