Projet de Machine Learning : Prédiction de l'âge à partir de données cérébrales

Yanis Ahdjoudj, Lucas Diaz, El Mehdi Agunaou

January 19, 2021

1 Rationnel scientifique

Le vieillissement de l'homme est en moyenne marqué par des pertes progressives des performances cognitives. Ces effets résultent principalement de l'atrophie, la diminution de la densité de la matière grise. Comme nous pouvons le voir à partir des données fournies (GM_ratio = Part de la matière grise dans le poids total du cerveau)

Grâce à l'imagerie médical, il est possible de visualiser l'évolution des caractéristiques anatomiques associées au vieillissement normal d'un cerveau.

Il serait donc bon d'estimer l'âge cérébral à l'aide des mesures de la matière neuronale sur des cerveaux d'une population d'individus en bonne santé.

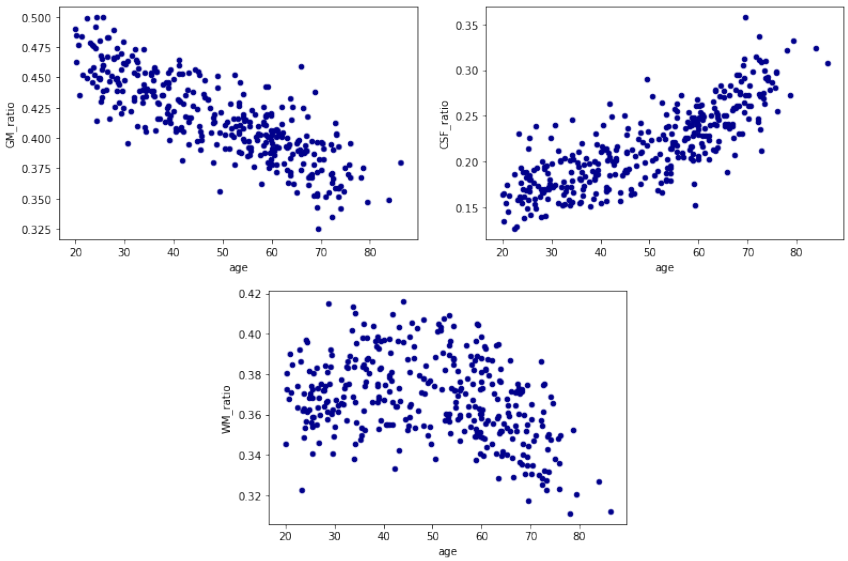

Figure 1: Volume de matière cérébrale en fonction de l'âge

On constate une relation linéaire décroissante entre le volume de matière grise dans le cerveau et l'âge des individus de notre base.
Cela renforce notre idée d'analyse de cette métrique dans notre estimation de l'âge.
Concernant le volume de matière blanche, on observe une tendance en deux temps, une montée suivie d'une baisse autour de la cinquantaine. On a une relation non linéaire avec l'âge.

On a une relation linéaire croissante entre l'âge et le liquide cérébrospinal(ancien céphalo-rachidien).

A l'aide du "mask" disponible sur le jupyter notebook on a pu redimensionner les données VBM en 5D pour pouvoir afficher quelques coupes axiales 2D du cerveau. Nous avons sélectionné quelques individus avec des attributs extrêmement opposés pour illustrer les différences sur le scan. Parmi ces attributs nous avons sélectionné : **age min (19)** *vs* **age max (86)**, **Volume GM min** *vs* **Volume GM max**. Les graphiques suivants affichent les couches axiales numéro 56, 57 et 58 des IRM de deux personnes (86 ans et 19 ans).

La concentration de matière grise est plus importante chez l'individu de 19 ans.

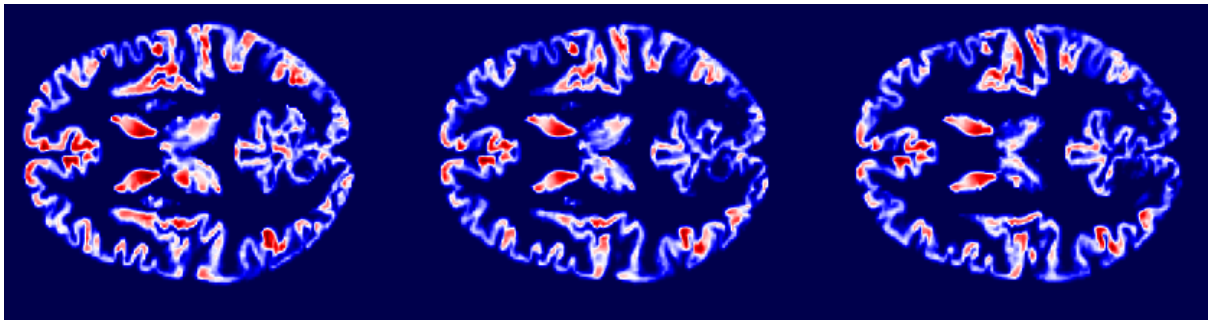

Figure 2: Trois couches du Scan VBM age : 86, id : 464

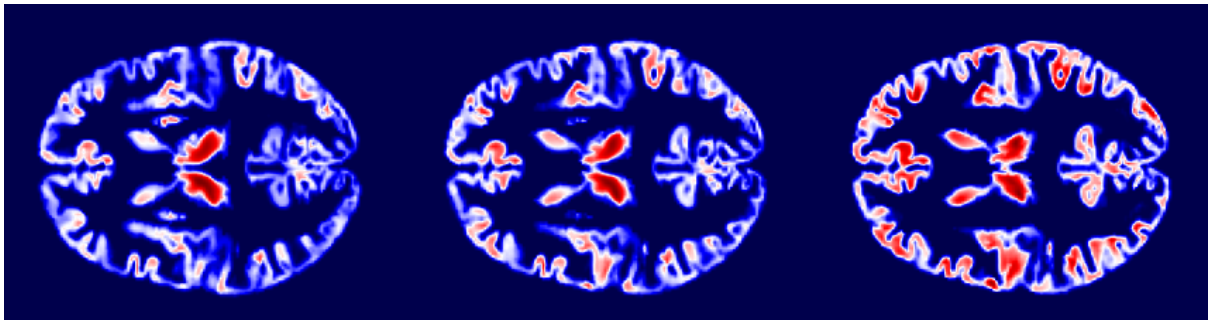

Figure 3: Trois couches du Scan VBM age : 19, id : 425

Les graphiques suivants affichent les couches axiales numéro 56, 57 et 58 du scan de deux individus ayant les volume de matière grise maximal et minimal.

La concentration de matière grise est plus importante dans la figure 4.

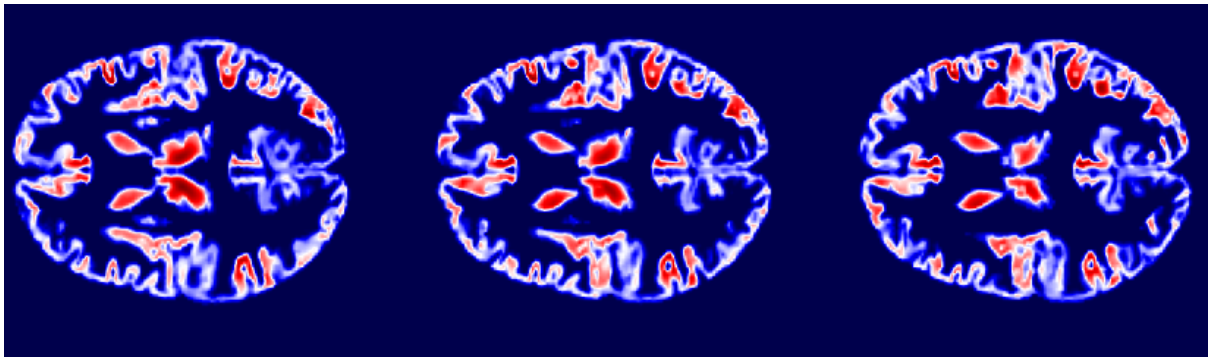

Figure 4: Trois couches du Scan VBM / VolGM : max, id : 257

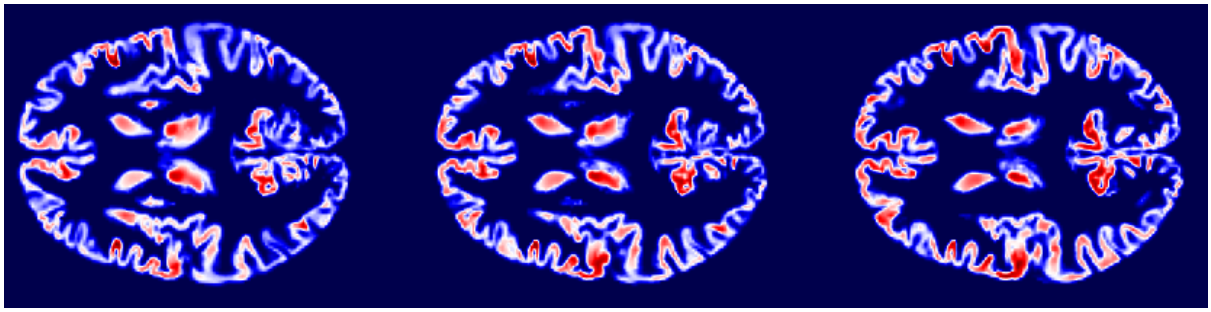

Figure 5: Trois couches du Scan VBM / VolGM : min, id : 214

2 Modèle choisi

Nous avons décidé de travailler avec les données d'imagerie médicale. On est dans le cas d'un modèle complexe, on à un nombre important de variables relativement au nombre d'observations. Il y a un risque de surapprentissage.

De plus on est dans une situation de modèle à haute dimension, le nombre de variables est supérieur au nombre d'observations. Il est plus difficile d'estimer les densités de l'échantillon, et les prédictions sont plus difficile en raison des observations en bord d'échantillon. Cela conduit aussi à du surapprenstissage.

Pour résoudre ce problème d'overfitting, nous allons procéder à des modèles de régression linéaire régularisée.

En effet, à travers le compromis biais-variance, on baisse la capacité d'apprentissage (et donc de surapprentissage) et on rend les solutions plus simples. On évite donc les solutions trop complexes, tenant compte du bruit ou des corrélations.

$$\text{Elastic Net (w)} = \left\|\mathbf{y} - \mathbf{X}^T\mathbf{w}\right\|_2^2 + \alpha\left(\rho\|\mathbf{w}\|_1 + (1-\rho)\|\mathbf{w}\|_2^2\right)$$

En raison de meilleures performances prédictives on retient le modèle de regression Elastic-Net.

Le terme de régularisation L1 à été choisi en fonction des performances du modèle et à donc été fixé à 0.05.

3 Résultats

Nous obtenons au final les résultats suivant à partir de la soumission sur ramp :

team	submission	rmse	train time [s]	validation time [s]
lucasdz	AhdAguDia_test6	6.55	65.964787	5.643768

Avec une RMSE de 6.55 nous sommes satisfaits des résultats car nous pouvons considérer une application pratique du modèle. En effet avec cette précision relative, nous pouvons en pratique prédire au minima une tranche d'âge pour un cerveau donné.
L'application des CNN pour cette problématique pourrait donner de meilleurs résultats, mais ce genre de modèles est très énergivore et chronophage.