UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Projet de prédiction de l'âge à partir de la matière grise cérébrale

Introduction:

Dans ce rapport, nous allons décrire notre méthodologie pour construire un modèle de prédiction de l'âge à partir d'images vectoriel de la matière grise cérébrale. Un tel modèle revêt une importance significative pour la recherche en neurosciences, il peut servir d'outil précieux pour détecter des écarts par rapport à l'âge cérébral attendu, ce qui pourrait indiquer la présence de processus neurobiologiques pathologiques.

Nous disposons de deux types de données d'entrée :

1. Régions d'Intérêt (ROIs) de la Matière Grise (GM) ajustées pour le Volume Intracrânien Total (TIV) contiennent 284 caractéristiques. Trois mesures supplémentaires ont été dérivées des données portant sur : le volume total du liquide céphalorachidien (CSF_Vol), de la matière grise (GM_Vol) et de la matière blanche (WM_Vol).
2. Morphométrie basée sur les voxels (VBM) : Cartes ou Images 3D de la GM VBM : des images 3D de forme (121, 145, 121). Ces données incluent le masque 3D et la transformation affine vers le référentiel MNI, fournissant 331 695 caractéristiques d'entrée (voxels) "plates" pour chaque participant.

La variable cible est l'âge des participants. Le jeu de données d'entraînement contient 357 échantillons et le jeu de données de test en contient 90.

Le principal défi auquel nous sommes confrontés dans ce problème est le nombre élevé de variables dans nos données d'entrée par rapport à la taille de l'échantillon. Ce problème de haute dimensionnalité présente plusieurs défis, notamment en termes de complexité computationnelle, de surajustement potentiel du modèle et de difficulté à interpréter les résultats.

Réduction de dimension :

Pour surmonter ce défi, nous avons choisi d'effectuer une réduction de dimension sur les données VBM en utilisant une Analyse en Composantes Principales (PCA). Cette technique permet de réduire le nombre de variables tout en conservant autant d'informations que possible. En réduisant la dimensionnalité des données VBM, nous parvenons à extraire les caractéristiques les plus significatives tout en réduisant le risque de surajustement du modèle.

Modélisation :

En ce qui concerne la méthode de modélisation, nous avons opté pour l'approche du stacking. Le stacking est une technique d'ensemble qui combine les prédictions de plusieurs modèles de base pour produire une prédiction finale. Dans notre cas, nous avons utilisé un Stacking Regressor avec trois modèles de base différents : Random Forest, XGBoost et Bayesian Ridge. Cette approche présente plusieurs avantages, notamment la possibilité de capturer la complexité du problème à l'aide de modèles variés et de combiner leurs forces pour améliorer les performances prédictives globales. De plus, le stacking permet de réduire le risque de surajustement en combinant les prédictions de modèles diversifiés. En régressant les prédictions des modèles de base, le modèle de stacking est moins sensible aux fluctuations aléatoires dans les données d'entraînement, ce qui peut contribuer à une meilleure généralisation aux données de test.

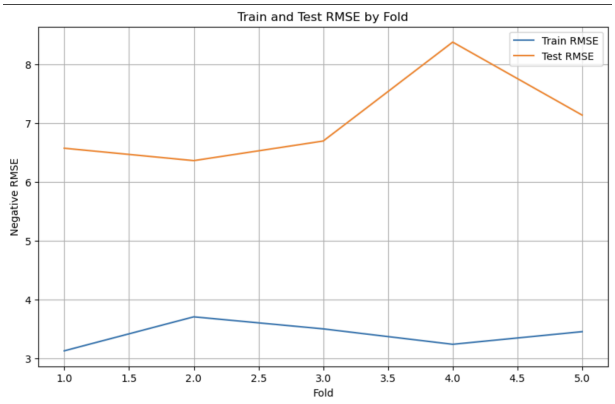
Bien que le modèle final de stacking combine les prédictions de plusieurs modèles de base, il reste relativement interprétable. En utilisant une régression linéaire comme modèle final, les coefficients associés à chaque modèle de base peuvent être examinés pour évaluer leur contribution relative à la prédiction finale.

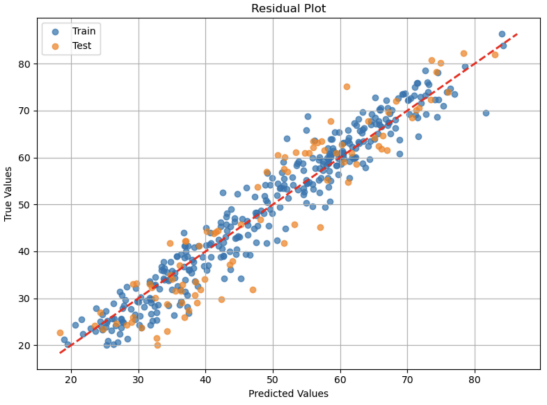
Le paramétrage des modèles individuels nous permet de contrôler la complexité de chaque modèle et d'éviter le surajustement, ce qui est crucial dans un tel contexte de haute dimensionnalité. En déterminant les meilleurs paramètres pour chaque modèle grâce à un grid search, nous avons pu optimiser les performances de chaque composant du modèle de stacking. Une fois que les meilleurs paramètres ont été déterminés pour chaque modèle, nous les avons inclus dans notre modèle de stacking en tant que composants individuels.

Tableau des performances des modèles testés :

Modèle	RMSE CV	RMSE TEST
Staking	6,06	6,88
Bayesian Ridge	6,32	7,01
Random Forest	7,49	7,96
XGBOOST	7,35	8,01
SVR	16,14	17,79
MLP	15,24	21,38

Représentation des performances du modèle stacking retenue :





Après avoir analysé les résidus de chacun de nos modèles testés, nous avons identifié un risque de surapprentissage. Cependant, ce risque a été considérablement atténué par l'approche du stacking, qui s'est avérée plus robuste en combinant les avantages des méthodes ensemblistes telles que XGBoost et Random Forest, qui capturent les relations complexes dans nos données, avec la capacité de régularisation de la méthode de régression bayésienne.

Lors de l'analyse de l'importance des prédicteurs pour le modèle de stacking sélectionné, nous avons examiné les coefficients de régression associés à chaque modèle inclus. Ici, le calcul de l'importance des prédicteurs varie selon les modèles, par exemple, entre le modèle bayésien et le Random Forest. Nous avons spécifiquement mis en évidence l'importance du modèle bayésien en analysant ses coefficients de régression spécifiques, ainsi que les variables les plus contributives à chaque modèle individuel (Random Forest, XGBoost, Bayesian Ridge). En croisant ces informations, nous avons identifié des variables communes jouant un rôle significatif dans la prédiction de l'âge pour l'ensemble du modèle de stacking : VBM_0, 'lThaPro_GM_Vol,'rMedFroCbr_GM_Vol','rPla_CSF_Vol, Cau_GM_Vo, 'lExtCbe_GM_Vol, 'rSupFroGy_CSF_Vol', 'lExtCbe_GM_Vol, lFroOpe_GM_Vol, 'lFroPo_GM_Vol'.

Mounira ARBOUCH

Mélanie YAKOUB