UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Prédiction de la schizophrénie

LEBRETON Louis, TARVERDIAN Mariam, VO Nguyen Thao Nhi

Sous la direction de M. DUCHESNAY Edouard

Master 2 MoSEF 2024 - 2025

19 février 2025

1 Introduction

La schizophrénie est une maladie qui entraîne une atrophie de la matière grise cérébrale. Les procédures actuelles ne permettent pas de diagnostiquer efficacement ces troubles, car il faut en moyenne 10 ans entre l'apparition des premiers symptômes et le diagnostic définitif de la maladie. Les techniques d'apprentissage automatique sont utilisées pour répondre à ce besoin. À partir de données de patients et de témoins sains, ce projet vise à concevoir un modèle prédictif de cette pathologie en exploitant les données cérébrales.

2 Baseline

Pour ce projet, nous avions à notre disposition deux ensembles de données : le premier portant sur les régions d'intérêt (ROIs) du cerveau, comprenant 284 variables, et le second constitué d'images 3D (VBM), représentant l'ensemble des voxels du cerveau. Afin d'évaluer la performance de notre modèle de classification, nous utilisons d'abord la métrique ROC-AUC, qui mesure la capacité du modèle à discriminer les classes. En complément, la balanced accuracy prend en compte les déséquilibres entre classes en calculant la moyenne de la sensibilité et de la spécificité.

3 Données utilisées

Dans notre phase de modélisation, nous avons utilisé uniquement les données issues des régions d'intérêt (ROIs) du cerveau, plutôt que l'ensemble des données voxels (VBM), pour prédire la schizophrénie. Les raisons sont les suivantes :

- **Hypothèses biologiques précises** : Certaines zones du cerveau sont connues pour leur implication dans la schizophrénie. L'utilisation des ROI permet de tester directement ces hypothèses avec des mesures spécifiques, sans considérer l'ensemble du cerveau.
- **Réduction des biais statistiques** : L'analyse VBM implique des milliers de comparaisons voxel par voxel, ce qui augmente le risque d'erreurs statistiques (fausses découvertes). En limitant l'analyse aux ROI, nous réduisons significativement ce risque.
- **Précision dans la délimitation des régions** : Les techniques VBM peuvent inclure partiellement des zones d'intérêt ou être affectées par des erreurs de segmentation. Les ROI, en revanche, permettent une délimitation plus précise.

4 Modélisation

Un processus de standardisation a été appliqué à l'aide de **StandardScaler** (centré et réduit). La standardisation est cruciale pour assurer que toutes les caractéristiques ont une échelle similaire, évitant que des valeurs à grande échelle ne dominent les modèles. Cela est particulièrement crucial pour les algorithmes basés sur des distances ou des gradients (comme les SVM ou la régression logistique).

Nous avons choisi un **VotingClassifier** avec une approche de vote « soft » permettant de combiner plusieurs modèles en prenant une décision collective. En utilisant les probabilités prédites par chaque modèle, le classificateur final prend une décision finale basée sur une moyenne pondérée.

Ainsi, les erreurs individuelles des modèles peuvent être compensées par les prédictions des autres, réduisant le risque d'erreur globale. De plus, chaque modèle capte des patterns différents dans les données :

— La **régression logistique** permet de modéliser les relations linéaires et d'obtenir des probabilités pour chaque classe. Afin d'améliorer sa performance dans des situations où il y a beaucoup de variables et où certaines peuvent être fortement corrélées entre elles, nous avons utilisé la régularisation Elastic Net. Cette méthode combine deux types de régularisation : L1 (Lasso) et L2 (Ridge). Le L1 ratio de 0.8 signifie que nous privilégions la régularisation L1, ce qui permet de rendre certains coefficients nuls et d'éliminer les variables les moins importantes. Cela aide à mieux gérer la multicolinéarité.

— Le **NuSVC** avec un noyau RBF permet de capturer des relations non linéaires complexes entre les variables, ce qui est particulièrement adapté aux données médicales souvent caractérisées par des corrélations complexes. Le paramètre nu régule le compromis entre la complexité du modèle et le taux d'erreur, offrant une meilleure flexibilité pour ajuster la frontière de décision.

— Le **SVC** avec un noyau poly permet de capturer des interactions complexes d'ordre supérieur entre les variables, et modéliser des interactions spécifiques que le noyau RBF ou la régression logistique ne pourraient pas identifier.

Certains hyperparamètres ont été fixés manuellement en fonction des bonnes pratiques, tandis que pour d'autres un **GridSearchCV** a été utilisé pour optimiser les paramètres. Cette approche a permis de trouver les meilleures combinaisons tout en limitant le temps de calcul.

5 Valeur SHAP

Les valeurs **SHAP** evaluent l'impact de chaque variable en mesurant la différence dans les prédictions lorsque cette variable est permutée ou modifiée, tout en gardant les autres constantes. Ainsi, voici **les 10 prédicteurs les plus importants** :

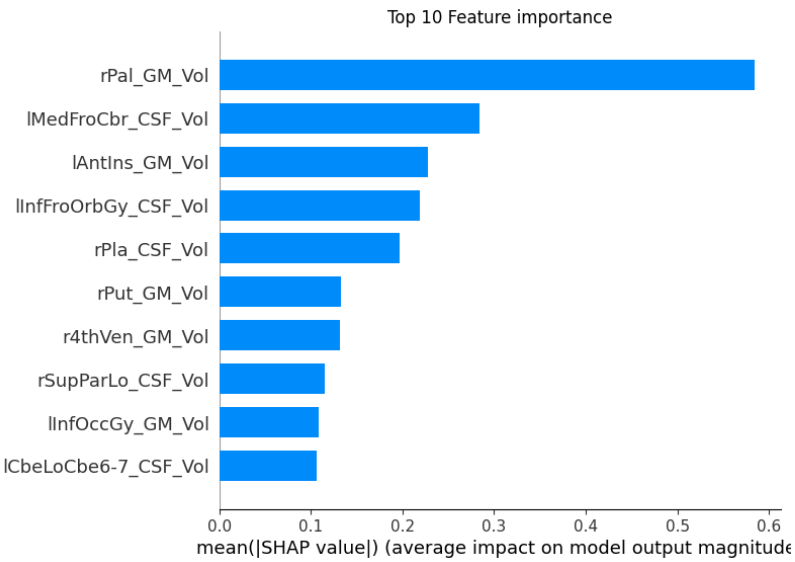

FIGURE 1 – Moyennes des Valeurs SHAP

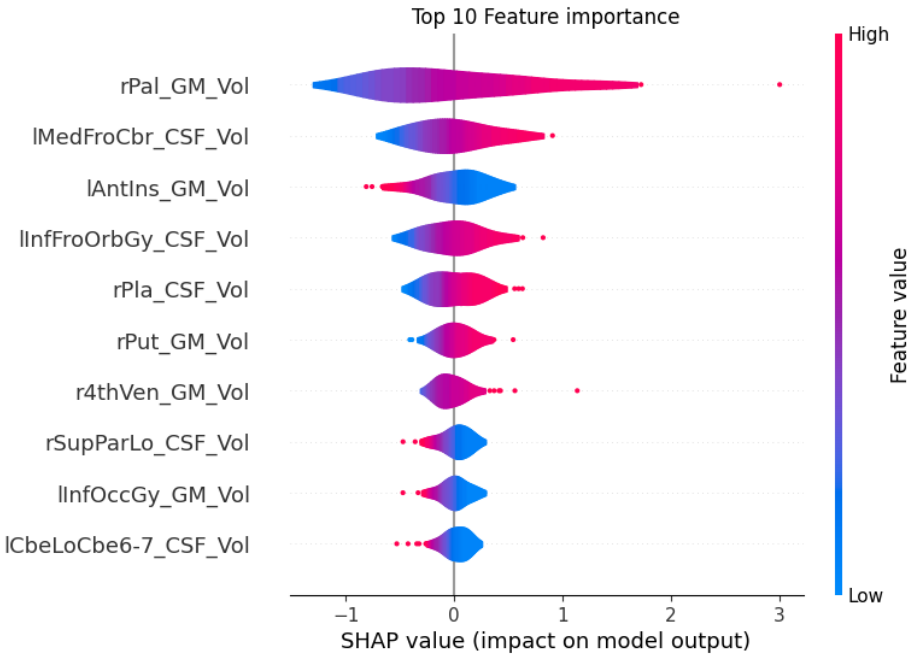

FIGURE 2 – Valeurs SHAP

— Sur l'axe des abscisses, les valeurs SHAP situées à droite de 0 augmentent la probabilité que la cible soit égale à 1 (dans notre cas, que le patient soit atteint de schizophrénie), tandis que les valeurs à gauche de 0 la diminuent.

— Sur l'axe des ordonnées, les caractéristiques sont classées par ordre d'importance moyenne, calculée à partir de la valeur absolue des SHAP. Les caractéristiques les plus influentes sont positionnées en haut du graphique.

— La couleur du graphique SHAP indique la valeur des caractéristiques : le rouge représente une caractéristique avec une valeur élevée, tandis que le bleu correspond à une valeur faible.

Cette visualisation permet ainsi d'identifier clairement les caractéristiques clés du modèle et de comprendre leur impact sur les prédictions. Avec notre modèle, le volume de la matière grise (GM) dans la région du pallidum droit joue un rôle déterminant dans la prédiction de la schizophrénie, tandis que le volume de liquide céphalo-rachidien (CSF) dans les régions du cervelet inférieur est un contributeur moins significatif pour distinguer ou prédire la schizophrénie dans le cadre de ce modèle spécifique.

6 Conclusion

Notre pipeline, construit à partir des données ROI et d'un ensemble de modèles prédictifs, a démontré une excellente capacité discriminante avec un AUC de **0,84** (sur la plateforme RAMP), reflétant une performance fiable dans la classification de la schizophrénie. La balanced accuracy pour le mondèle que nous avons choisi est de **0,76**.En optimisant les hyperparamètres via GridSearch et en intégrant une validation croisée, nous avons renforcé la robustesse du modèle tout en limitant le risque de surapprentissage. Cette approche souligne la pertinence des données ROI et de l'approche ensembliste pour répondre aux défis complexes de classification dans le domaine de la santé mentale.