PROJET : PRÉDICTION DE LA SCHIZOPHRÉNIE

Alexis CHRISTIEN, Issame ABDELDJALIL
MASTER 2 MOSEF

Ce projet a pour but la classification d'individu en fonction de s'il sont ou non atteint de la maladie de la schizophrénie. Pour ce faire, nous disposons de 410 individus sur notre ensemble d'entraînement et de plus de 330.000 variables, divisées en 2 parties :

- Les ROIs (Regions Of Interest) contenant des informations sur la matière grise du cerveau (284 variables)
- Les cartes VBM 3D de la matière grise (331 695 variables)

La cible est une variable binaire : [control , schizophrenia] et est équilibrée : 54% et 46% respectivement.

Sélection des Features

Etant donnée la dimension de l'espace des Features, il est nécessaire de le réduire pour éviter du sur-ajustement et accélérer l'entraînement de nos modèles. Pour ce faire, nous ne gardons en premier lieu que les ROIs puisqu'il nous est indiqué que leurs informations sont redondantes avec celles des VBM. En second lieu, nous étudions les corrélations entre les variables ROIs afin de supprimer là aussi des potentielles redondances.

Pour étudier les corrélations entre les variables, nous avons mis en place différents tests, dépendant de la normalité ou non des variables ainsi que de leurs types (binaires, catégorielles ou continues). Pour nos 284 variables, aucune n'est binaire ni catégorielle mais 280 sont continues. les 4 dernières valent 0 pour tout les individus

Pour les variables continues (42), nous avons analysé la normalité des variables avec 3 tests : **Shapiro-Wilk**, **Kolmogorov Smirnov** et **Jarque-Bera**. Nous avons considéré comme normales les variables pour lesquelles les 3 tests étaient d'accord. Pour ces dernières, nous avons mis en place une corrélation de **Pearson** et un seuil à 0.9 pour détecter les redondances. Toutes paires de variables dépassant ce seuil sont considérées comme trop corrélées. Nous avons décidé de supprimer la variable la moins corrélée à la cible, corrélation réalisée à l'aide du test du point bisérial (puisque notre cible est binaire).

Pour les variables continues non normales (238), nous avons utilisé le test de **Spearman**, qui ne nécessite pas la normalité. A l'instar de l'analyse précédente, nous avons utilisé le seuil et le point bisérial pour faire notre sélection.

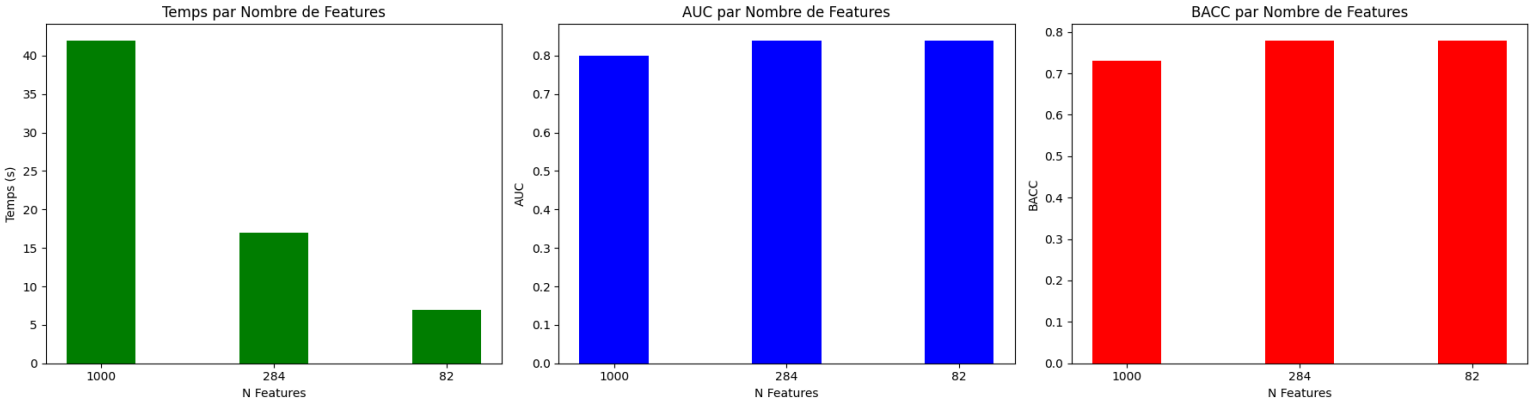
Après l'analyse des corrélations, nous passons de 280 variables à 214, ce qui est une réduction considérable mais qui doit encore être poussée. Pour ce faire, nous réalisons une régression logistique et sélectionnons les Features impactant le plus les prédictions, celles ayant les plus gros coefficients de régression. A la suite de cette analyse, nous avons 82 Features, que nous utiliserons dans nos modèles.

Modèles

Concernant les modèles, nous avons dans un premier temps consulté la documentation sur certains articles scientifiques prédisant la schizophrénie. Nous nous sommes ensuite inspirés des méthodes utilisées afin d'effectuer des modèles issus de la librairie scikit-learn.
Nous avons décidé de tester un par un plusieurs modèles non linéaires issus des différentes familles de modèles : Un Support Vector Machine Classifier, un Random Forest, un Gradient Boosting, et un MLP.
Ayant remarqué que le Gradient Boosting fonctionnait très bien (ROC AUC de 0.83), nous avons décidé de le combiner à d'autres modèles. Notre modèle final reposait donc sur une combinaison des modèles suivants: Gradient Boosting, Support Vector Machine Classifier ainsi que Multi-Layer Perceptron qui sont tous des modèles très différents et complémentaires.

Notre modèle de Gradient Boosting a la capacité de capturer des relations complexes entre les variables grâce à son approche d'ensemble d'arbres de décision. Nous avons décidé d'ajouter par la suite le MLP car il est capable de modéliser des relations complexes et non linéaires mais permet également d'extraire des représentations avancées des données, ce qui est pertinent dans notre sujet compte tenu du nombre élevé de variables. Finalement, nous avons combiné les 2 modèles précédents avec le SVC car ce modèle a permis de stabiliser les prédictions finales en nous permettant de séparer les classes schizophrènes et contrôles.
Finalement, nous avons longuement hésité entre choisir un Stacking ou bien un Voting Classifier pour combiner nos modèles mais nous avons fini par choisir le Voting Classifier car il est plus facile à interpréter vu que chaque modèle contribue directement au résultat. Nous avons également voulu réduire les risques d'overfitting étant donné la complexité des données cérébrales et le Voting Classifier était le meilleur choix car il ne ré-entraîne pas un modèle supplémentaire (à la différence du metamodel du Stacking) sur les prédictions des autres modèles.
Sur ce modèle final, on obtient un score de 0.85 pour l'AUC et un score de 0.76 pour la Balanced Accuracy.

Nous avons lancé nos modèles pour différents nombre de Features (avec 1000 features, avec les 284 initiales, avec les 82 finalement choisies) et nous voyons bien l'impact sur les métriques. La prise en compte de toute les variables augmente le risque d'Overfitting et nous voyons bien que notre modèle est plus performant sur l'ensemble d'entraînement et moins sur l'ensemble de test que lorsque l'on sélectionne des variables, ce qui traduit bien le fait que le modèle sur-rapprend et parvient moins bien à généraliser.
Pour illustrer notre propos, nous avons enregistré nos résultats en local du temps d'apprentissage, de l'AUC et de la BACC en fonction du nombre de Features. On voit clairement que 82 features prennent moins de temps tout en étant le plus performant.



Shapley Values et LIME values:

Nous remarquons bien que la variable rPal_GM_Vol se distingue et est la plus importante à travers ces 2 méthodes. Les autres variables restent assez similaires en termes d'importance.