## Discussion

### Interprétation des résultats

Comme indiqué dans la section précédente, nos meilleurs scores viennent des **méthodes d’ensemble.**

Pour rappel, notre implémentation du Voting réunit et fait voter les quatre modèles optimisés suivants :

* LogReg
* DT
* RF
* KNN

Notre implémentation du Stacking, qui affiche également de belles performances, est construite sur ces mêmes modèles de base, suivie de la désignation de LogReg comme estimateur final.

Le pouvoir de ces approches résident en leur capacité à combiner des modèles très différents : elles sont en mesure de tirer profit des forces individuelles de ces derniers, tout en maîtrisant leurs faiblesses respectives. Cette diversité, que nous avons essayé de respecter au mieux dans nos implémentations (en choisissant des modèles aussi variés que possible), est un atout face notre jeu de données. En effet, la complexité, le déséquilibre, les dimensions et l’hétérogénéité de ce dernier mettent à mal les modèles classiques, qui ne parviennent pas à en capturer les caractéristiques plurielles et la structure complexe sous-jacente. [15]

Cependant, malgré leurs bons scores, ces modèles comportent aussi des inconvénients. Les deux principaux sont les suivants :

* Ils sont coûteux en ressources de calcul. [16] À titre d’exemple, notre implémentation du Voting revient à faire quatre modélisations, ce qui représente une opération conséquente compte tenu des dimensions de notre jeu de données.
* Leur complexité se traduit par une interprétabilité plus faible. [17] Dans le domaine de l’apprentissage automatique, la question de l’interprétabilité des modèles est loin d’être anodine. En fonction du secteur d’activité, des parties prenantes et d’autres éléments de contexte, elle peut même être centrale et contraindre le choix de modèle (souvent au détriment de la performance, laquelle augmente généralement avec la sophistication d’un modèle). Le plus souvent, le besoin ou le désir de développer la confiance des usagers dans un modèle et ses prédictions incite à en préférer un avec une bonne interprétabilité, quitte à faire des concessions sur la performance. [18] L’environnement réglementaire influe également sur les parties prenantes : le RGPD donne, aux personnes dont les données sont traitées, le droit à une explication des décisions algorithmiques et le droit d’être informées. [19] Enfin, le fait de proposer des interprétations et des explications des algorithmes souvent perçus ou caractérisés comme des « boîtes noires » ont, en soi, une valeur sociale et éthique. [20]

Il n’est donc pas évident de préférer les méthodes d’ensemble aux autres approches que nous avons également testées, notamment celles basées sur les **arbres de décision,** car ces dernières proposent des scores qui restent corrects, tout en étant plus facilement interprétables.

Afin de nous assurer d’avoir les meilleurs scores possibles en main, nous avons effectué diverses **optimisations** (mise à l’échelle, rééquilibrage, optimisation du seuil). Ces dernières fonctionnent très bien sur la plupart des algorithmes que nous avons implémentés. En revanche, elles semblent produire l’**effet contraire** sur ceux issus de la **classification des séries temporelles.** Cela est peut-être dû à notre emploi d’**approches naïves** au sein des étapes suivantes de prétraitement des données :

* **Mise à l’échelle.** Il existe de nombreuses méthodes de normalisation des séries temporelles, chacune adaptée à des problématiques différentes. [21] Par manque de temps et d’expérience, nous nous sommes limités aux méthodes proposées par scikit-learn (StandardScaler, MinMaxScaler). Par ailleurs, bien que la normalisation soit généralement nécessaire dans l’étude des séries temporelles, elle peut aussi détruire des informations et contribuer à la dégradation des analyses. [22]
* **Rééquilibrage.** Notre approche ne tient pas compte de la difficulté inhérente à la classification des séries temporelles déséquilibrées, une thématique de recherche qui se révèle assez riche et complexe. [23] En effet, la dimensionnalité élevée et la forte corrélation inter-variable qui caractérisent ces ensembles de données ne sont pas prises en compte par les méthodes classiques proposées par scikit-learn (SMOTE, RandomUnderSampler). [24]

Compte tenu de tous les éléments exposés ci-dessus, nous avons décidé de retenir **l’algorithme nº 2 (RF avec StandardScaler + SMOTE + ROC)** comme modèle d’apprentissage automatique pour notre projet de classification binaire. Même s’il se classe légèrement en dessous de l’algorithme nº 1 (Hard Voting optimisé avec StandardScaler + SMOTE) au niveau des scores, il est plus facile à interpréter et à expliquer, ce qui en fait un choix global plus équilibré, et meilleur, selon nous.