# Statistique en Grande Dimension et Apprentissage
# TP-Projet

Ce travail est individuel et compte pour 50% de la note finale. Le but est d'illustrer et de tester les méthodes d'apprentissage statistique vues tout au long du semestre.

Un rapport de vos travaux (dont une partie peut être réalisée en Notebook) est à rendre pour le vendredi 10 janvier 2025 (la date est lointaine mais je vous conseille de commencer assez rapidement). Une très courte soutenance sera organisée la semaine du 20 janvier 2025 (semaine après les examens).

**Exercice 1** (Régression vectorielle)**.** .

1. Programmez à la main un modèle d'arbre de régression à sorties vectorielles. On suppose pour simplifier que les entrées sont uniquement quantitatives. La fonction devra prendre en sortie un vecteur de taille $r$ et utilisera la norme
$$\|y\|_q = \left( \sum_{j=1}^{q} |y_j|^q \right)^{\frac{1}{q}}.$$
Pour le choix du seuil de coupure, on pourra classer chaque variable dans l'ordre croissant et tester de manière exhaustive toutes les séparations possibles : si on note $x_1 < x_2 \ldots < x_n$ une variable classée, on testera les coupures $\alpha_k = \frac{x_k + x_{k+1}}{2}$ pour $k = 1, \ldots, n-1$ (si un choix moins coûteux est trouvé, il n'est pas interdit de l'utiliser non plus !).
On fixera une profondeur maximale et des seuils d'arrêt (homogénéité inférieure à un seuil $\alpha$, nombre d'individus inférieur par noeud inférieur à $n_0$ donné,...).
2. Testez ce modèle sur un exemple simulé : on peut par exemple supposer que les données satisfont :
$$Y^j = \langle X, \theta^j \rangle + \varepsilon_j, \quad j = 1, \ldots, r$$
où $X$ est à valeurs dans $\mathbb{R}^{r \times p}$ est une vecteur composé d'entrée Rademacher de paramètre 1/2, les $\varepsilon_j$ sont indépendants de loi $\mathcal{N}(0, \sigma^2)$ et $\theta^j = \mathbf{1}_{S_j}$ où $S_j$ est un ensemble de $s$ indices tirés au hasard parmi les $p$. On pourra renvoyer le score de validation croisée ainsi que le score de test. On pourra aussi choisir des données de régression avec sortie vectorielle.
3. Implémentez une méthode de gradient boosting sur ce modèle à nouveau programmée "à la main". On pourra afficher l'évolution de l'erreur d'entraînement et de l'erreur de validation.
4. Optionnel : Programmez l'algorithme de gradient boosting pour la classification multi-classes (à l'aide l'arbre de régression à sorties vectorielles).

**Exercice 2.** Dans ce second exercice, on s'intéresse à une base de données liée au cancer sein (Attention, les données sous un format $p \times n$, *i.e.* variable $\times$ individu). Les bases de données sont téléchargeables au lien suivant : `https://plmbox.math.cnrs.fr/f/fedcac32b2a949198dce/`
Dans cette base de données, l'objectif de prédire la réaction au traitement. La variable à prédire est "treatment_response" à partir de données génétiques et d'autres caractéristiques (âge/ethnie/Stade de la tumeur T/N). On se contentera des données génétiques afin de simplifier le problème et d'éviter de mettre sur un même plan des variables "cliniques" probablement très corrélées avec la réponse et des variables génétiques dont la complexité nécessite des méthodes plus complexes.

L'objectif de cet exercice d'optimiser l'erreur de classification puis le $F_1$-score associé au modèle. Testez quelques algorithmes à votre disposition pour tenter d'optimiser vos scores.

**Exercice 3** (Data Challenge)**.** Ce challenge sera organisé du 9 décembre 2024 au 10 janvier 2025. Il s'agit du Data Challenge "Prédire l'interaction molécule - protéine pour la découverte de médicaments", proposé par https://challengedata.ens.fr/. Ce challenge est un problème supervisé de régression, l'objectif étant de prédire l'inhibition d'une protéine par différentes molécules. Sa description est accessible ici. Vous devrez vous inscrire sur la plateforme et charger la base de données à disposition. Si le score obtenu est évidemment important dans la note associée à ce projet, la description des démarches effectuées ainsi que de l'algorithme choisi pour obtenir le résultat est néanmoins attendue.