Challenge - Méthodes de Régression Avancées

On dispose d'un échantillon statistique de **n = 100** données conjointes d'une variable y et de variables $x^{(1)}$, …, $x^{(p)}$ avec p = 200. Ces données sont dans le fichier **'data.txt'**.

Le but est de construire le meilleur modèle prédictif de la variable *y* à partir des variables $x^{(1)}$, …, $x^{(p)}$ sachant qu'il faudra au final calculer et retourner les prédictions associées au fichier **'Xtest.txt'** de taille 100×200 (100 prédictions à réaliser, une par ligne, cf. dernière étape ci-dessous).

Le critère utilisé pour évaluer ce challenge est le critère d'erreur de prédiction usuel (Root Mean Square Error) :

$$\mathbf{RMSE} = \sqrt{\mathbf{MSE}} \quad \mathbf{où\ MSE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2$$

où n_{test} = 100 désigne le nombre de prédictions à réaliser, y_i les valeurs réellement observées pour y et $\hat{y}_i$ vos prédictions. On pourra s'aider au début des indications suivantes :

0. Charger uniquement le fichier de données **'data.txt'** qui sera utilisé pour mettre au point un modèle de prédiction.

1. Calculer l'écart-type de la variable y comme premier **RMSE de référence** qui consisterait à prédire par la moyenne de y (prédiction constante qui ne tient compte d'aucun prédicteur).

2. Peut-on envisager une régression linéaire multiple ? (essayer) Que pourrait-on faire de simple ?

3. Compte tenu du nombre important de prédicteurs, envisager une méthode de type LASSO.

4. Réfléchir à des améliorations possibles.

5. Charger enfin le fichier **'Xtest.txt'** et calculer les prédictions correspondant à votre meilleur modèle.

On déposera sur Campus le fichier texte associé à vos prédictions (fichier comportant donc une seule colonne formée de vos 100 prédictions). Mettre ce fichier au format **NOM.txt**.