Prédiction des prix des logements à Boston. Le fichier BostonHousing.csv contient des informations recueillies par le Bureau du recensement des États-Unis concernant le logement dans la région de Boston, Massachusetts. L'ensemble de données inclut des informations sur 506 secteurs de recensement du logement dans la région de Boston. L'objectif est de prédire le prix médian des maisons dans de nouveaux secteurs en fonction d'informations telles que le taux de criminalité, la pollution et le nombre de pièces. L'ensemble de données contient 13 prédicteurs, et la réponse est le prix médian de la maison (MEDV). Le tableau 6.9 décrit chacun des prédicteurs et la réponse.

a. Pourquoi les données doivent-elles être partitionnées en ensembles d'entraînement et de validation ? À quoi servira l'ensemble d'entraînement ? À quoi servira l'ensemble de validation

b. Ajustez un modèle de régression linéaire multiple au prix médian des maisons (MEDV) en fonction de CRIM, CHAS et RM. Écrivez l'équation pour prédire le prix médian des maisons à partir des prédicteurs dans le modèle.

c. En utilisant le modèle de régression estimé, quel est le prix médian des maisons prédit pour un secteur de la région de Boston qui ne borde pas la rivière Charles, a un taux de criminalité de 0,1, et où le nombre moyen de pièces par maison est de 6 ? Quelle est l'erreur de prédiction ?

d. Réduisez le nombre de prédicteurs :

> i. Quels prédicteurs sont susceptibles de mesurer la même chose parmi les 13 prédicteurs ? Discutez des relations entre INDUS, NOX et TAX.

ii. Calculez le tableau de corrélation pour les 12 prédicteurs numériques et recherchez les paires fortement corrélées. Celles-ci présentent une redondance potentielle et peuvent provoquer de la multicollinéarité. Choisissez ceux à retirer en fonction de ce tableau.

iii. Utilisez la régression pas à pas avec les trois options (rétrograde, progressive, les deux) pour réduire les prédicteurs restants comme suit : Exécutez une régression pas à pas sur l'ensemble d'entraînement. Choisissez le meilleur modèle de chaque exécution pas à pas. Utilisez ensuite chacun de ces modèles séparément pour prédire l'ensemble de validation. Comparez le RMSE.