CNAM-MEDAS
Année 2020-2021

Projet Fouilles de données

Données :

Trois bases sont à votre disposition pour réaliser ce projet et pour répondre aux questions suivantes. La 1ère base « SwissLabor ; N=872 » contient des données sur la participation au marché du travail, la seconde « barro ; N=161 » des données sur le produit intérieur brut (PIB) d'un ensemble de pays et la troisième « gss_wages ; N=61697 » des données sur le revenu.

Les données ainsi que le descriptif de chaque base sont disponibles via le lien suivant (https://vincentarelbundock.github.io/Rdatasets/datasets.html)

Objectif du projet :

En fonction des données choisies, trouver la meilleure méthode en termes de pouvoir prédictive.

Stratégie à adopter :

1/ Définir un échantillon d'entrainement et un échantillon test

2 / Analyser les liaisons entre les prédicteurs et la variable d'intérêt

3/ Choisir au moins deux méthodes qui permettrait de prédire efficacement la variable réponse

4/ Pour chaque méthode considérée, tester plusieurs modèles et s'assurer de la qualité du modèle choisi pour la prédiction

5/ Comparer les performances des différentes méthodes sélectionnées en fonction de la qualité de la prédiction (sont à considérer 1 ou 2 critères d'évaluation).

Instructions pour le travail à rendre

Le travail à rendre prendra la forme d'un rapport où la démarche d'analyse est explicitée, les étapes d'analyses argumentées et les résultats commentés. Le travail est à exécuter sur le logiciel R dont le script sera rendu avec le rapport.