MACHINE LEARNING SUPERVISÉ

Estimer le coût de la couverture médicale d'un.e américain.e

1

ANTOINE MEYER -- ALEKSANDRINA STOYANOVA-CHRISTEN

Sommaire

Analyse des données

- Variables qualitatives et quantitatives
- Statistique descriptive

Modélisation

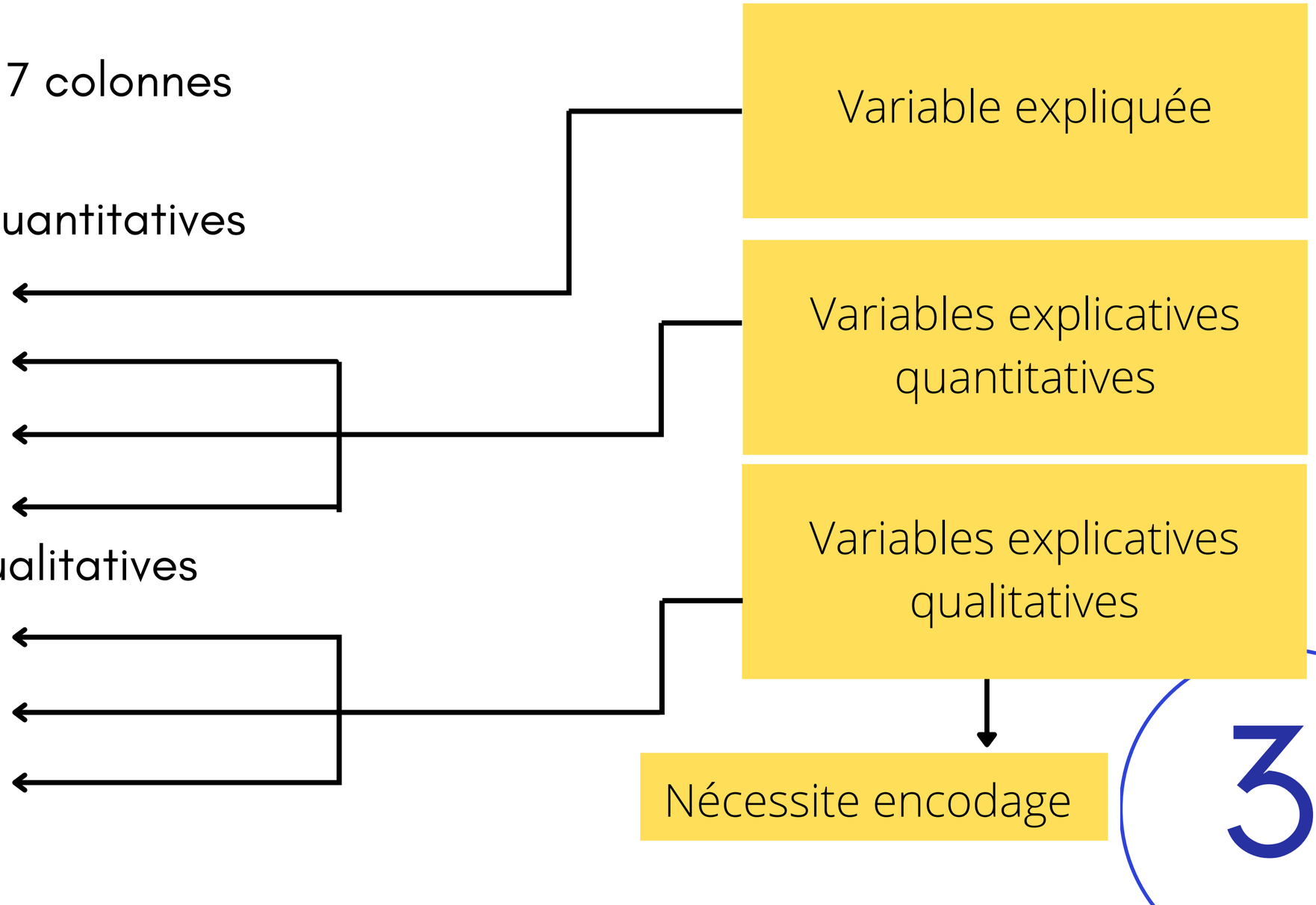
- Régression Linaire Multiple
- Forêts Aléatoires

Évaluation et comparaison des modèles

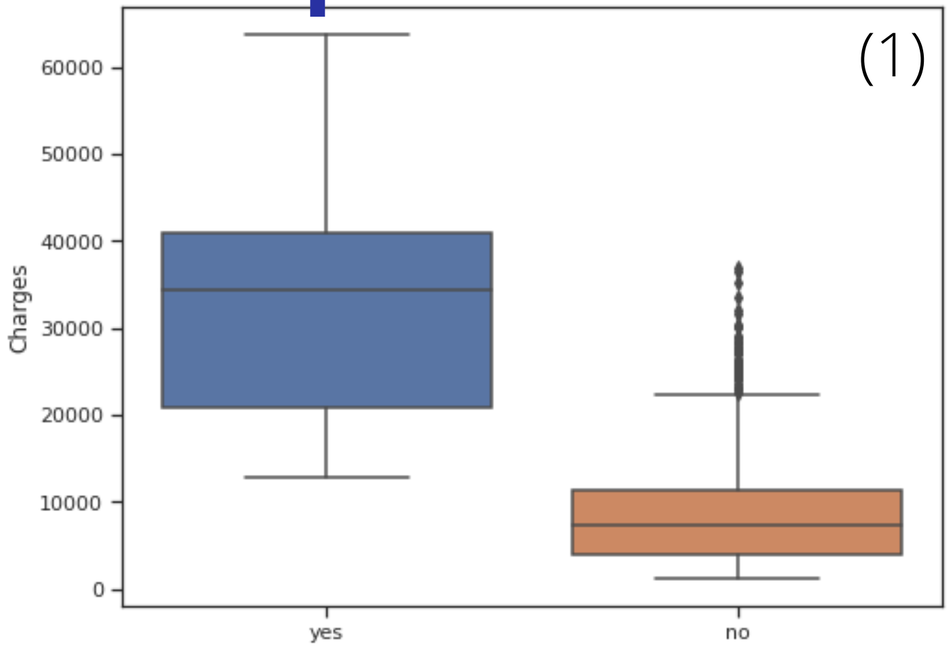
2

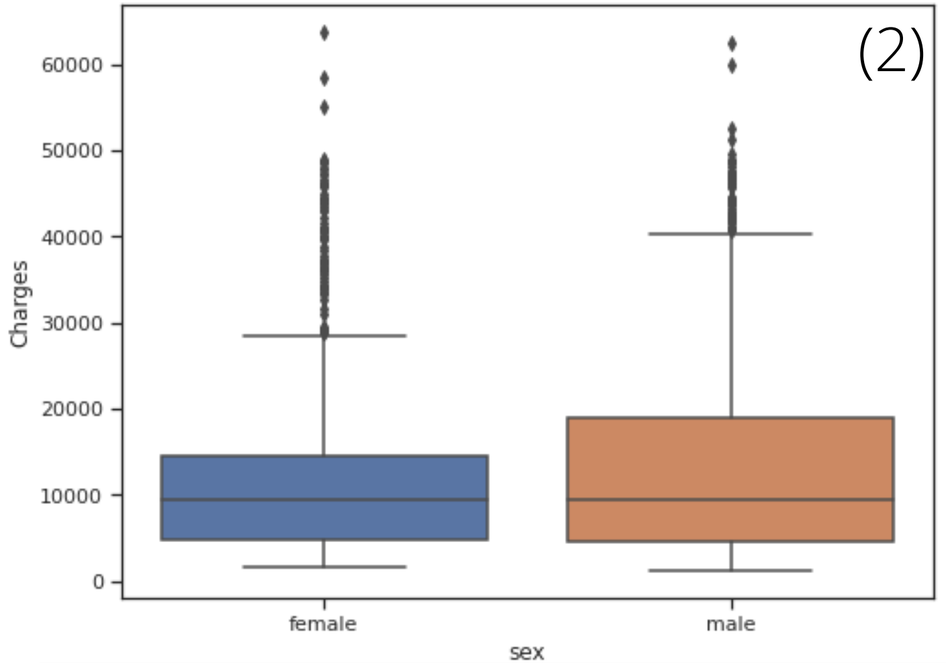
Analyse des données

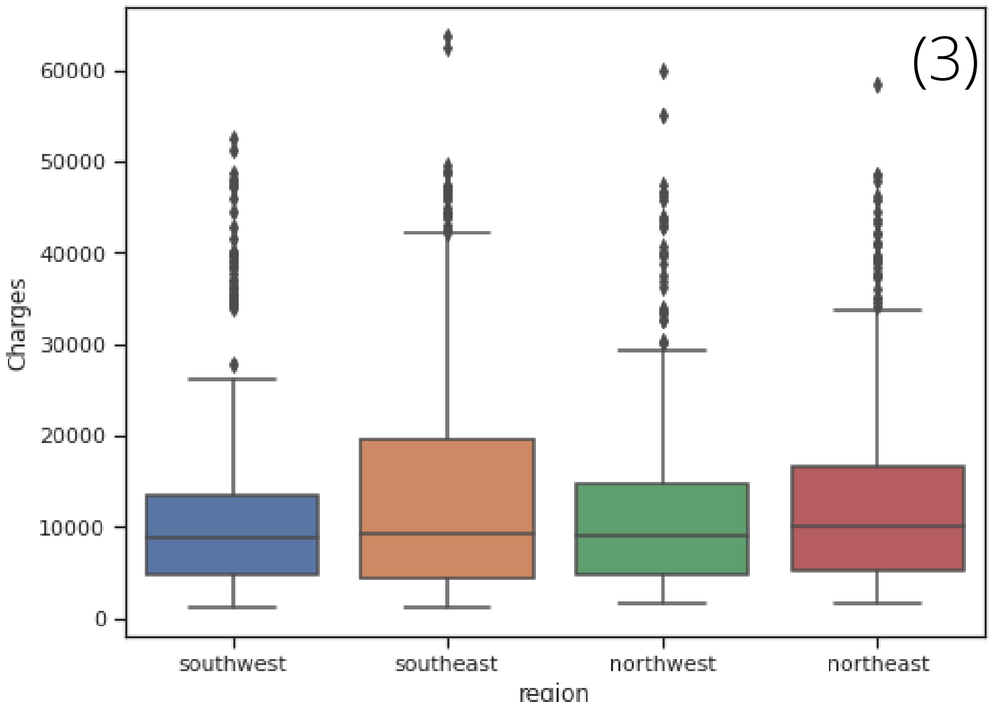
- 1338 lignes / 7 colonnes
- Complet
- 4 variables quantitatives
  - **charge**
  - **bmi**
  - **children**
  - **age**
- 3 variables qualitatives
  - **sex**
  - **smoker**
  - **region**

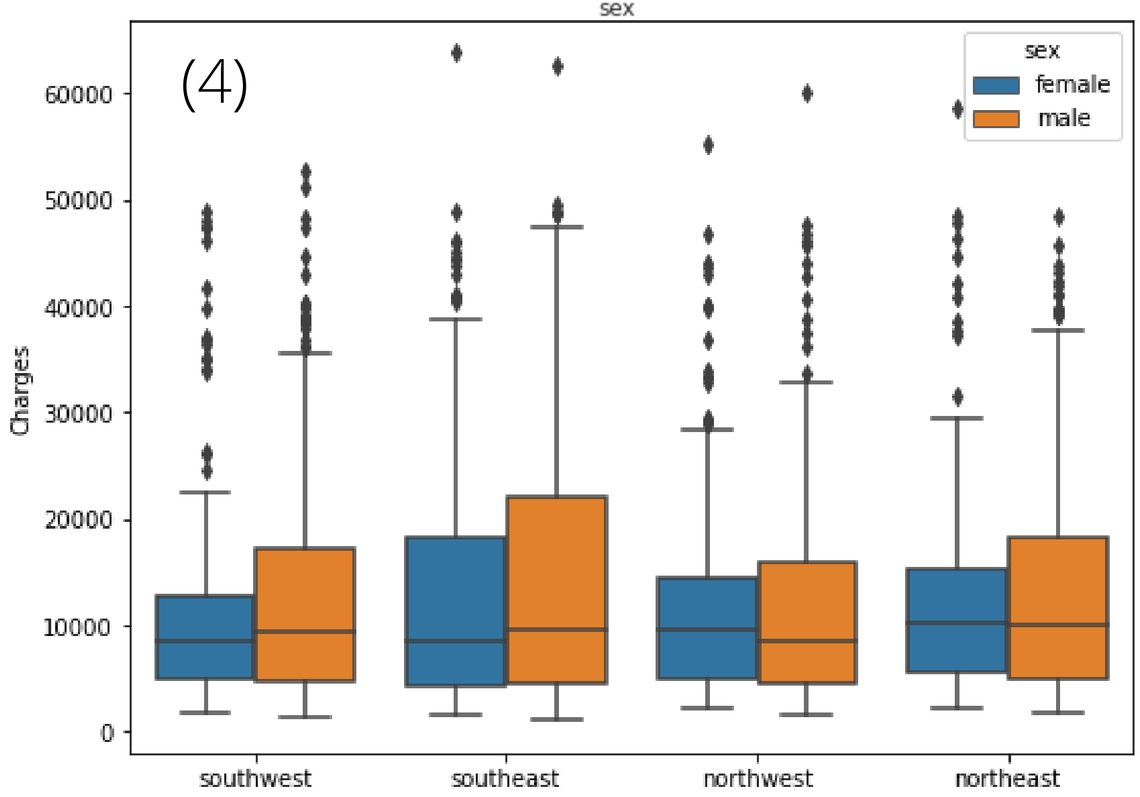


Variables qualitatives

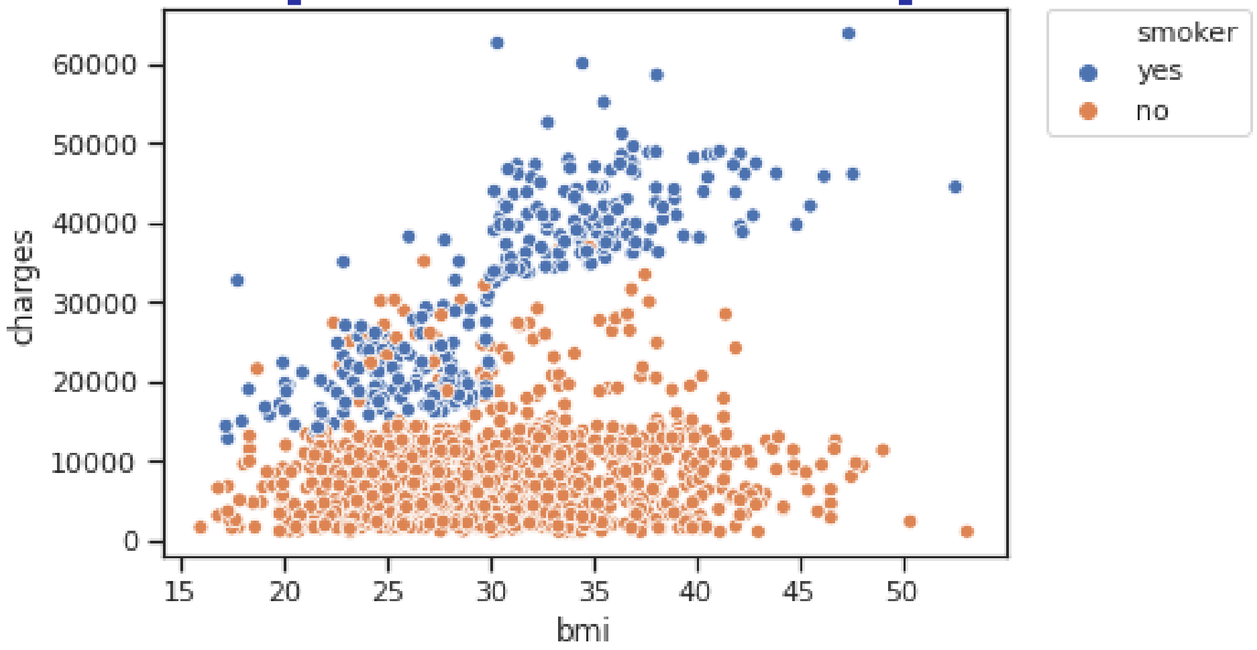


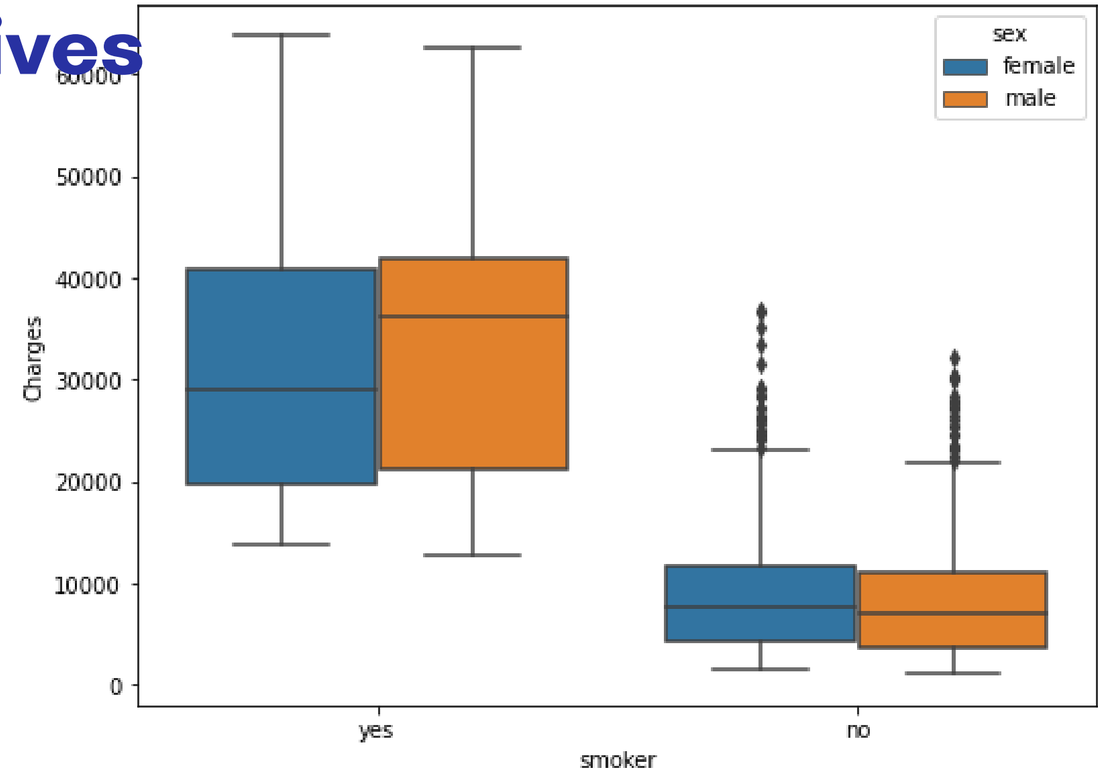


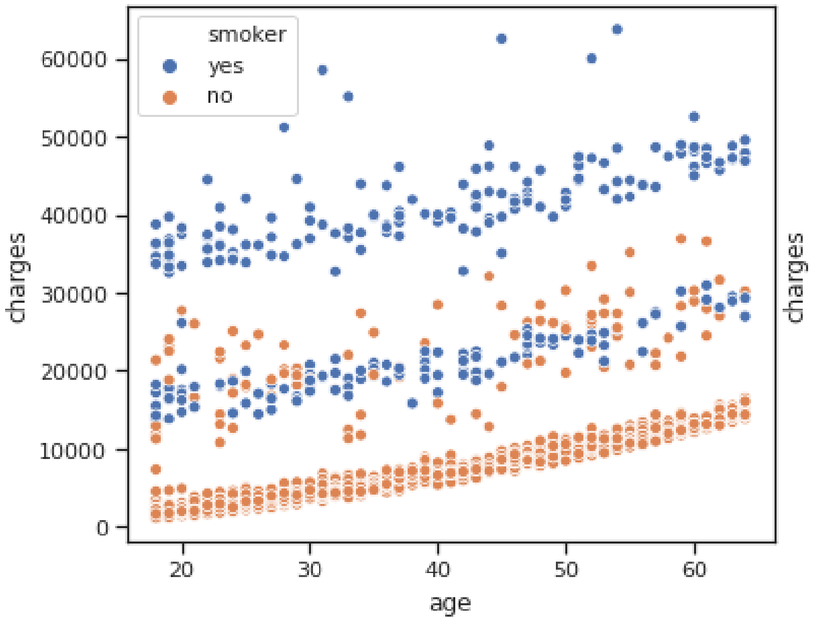


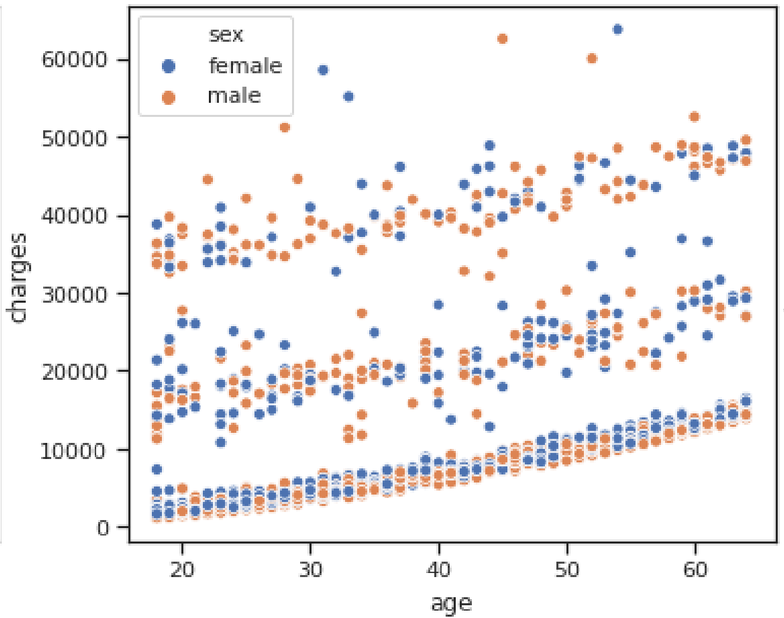


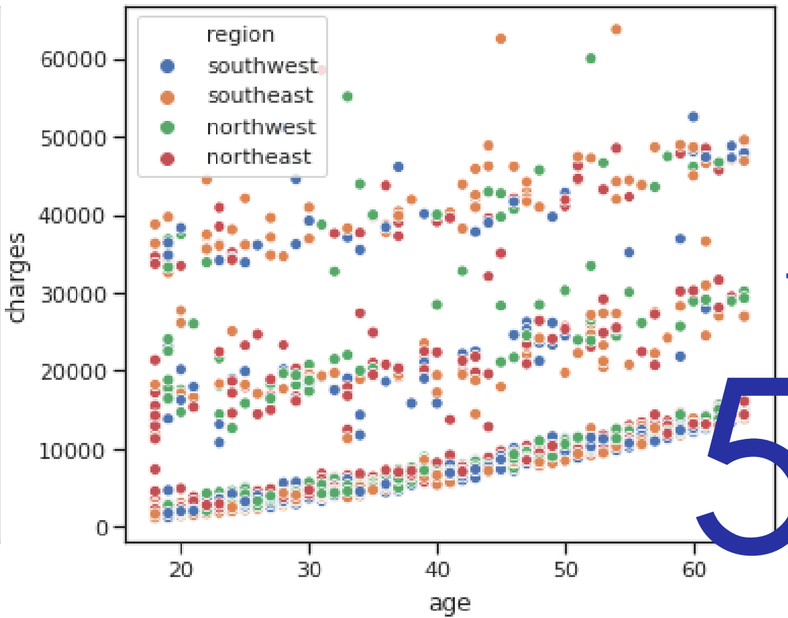
Variables qualitatives et quantitatives











Régression Linéaire Multiple

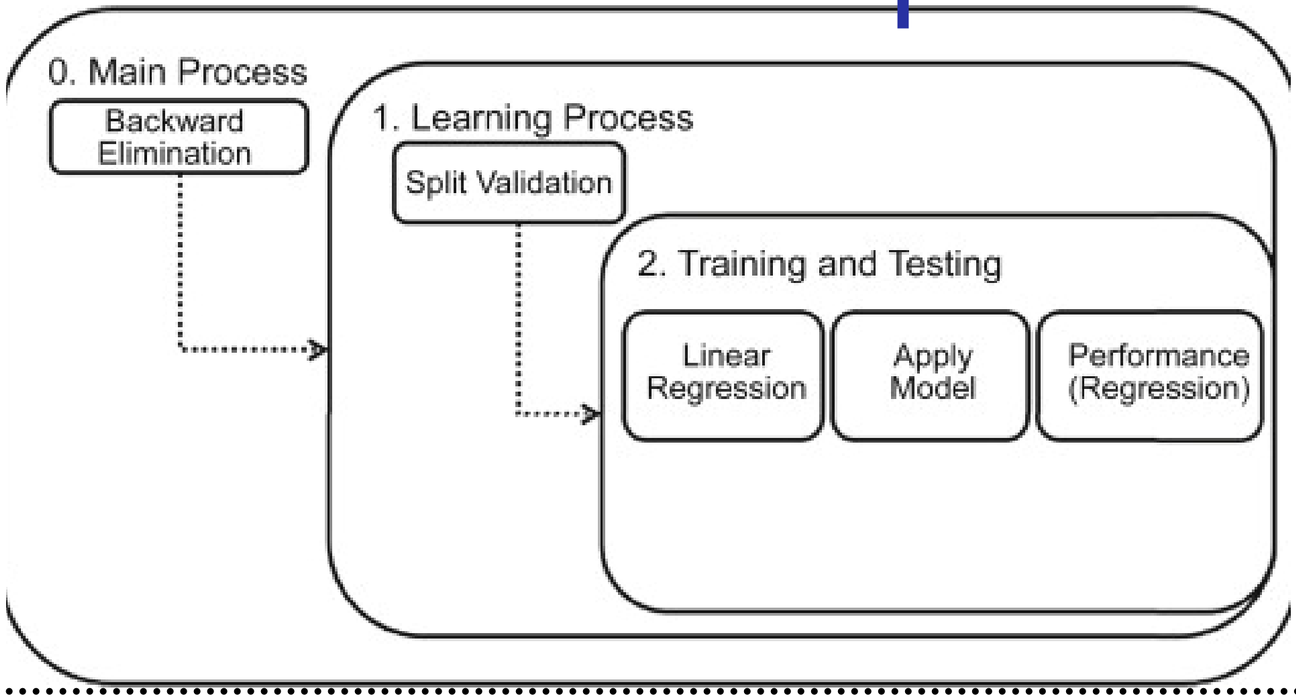
$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \in$$

coefficients

target

inputs

random error



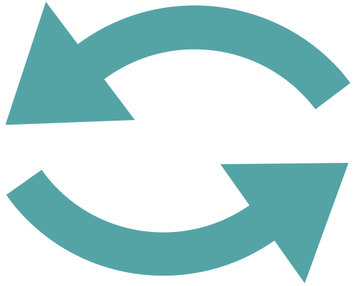
age/bmi/children/sex/smoker/region

(R^2-ajusté) égale à 0.762 et une précision de 53 % vs (R^2-ajusté) égale à 0.763 et une précision de 54 %. age/bmi/children/smoker/region

Charges = -12466.156 + 23347.627*__smoker_yes__ + 880.117*__region_northeast__ + +244.7259*__age__ + 340.0924*__bmi__ + 619.166*__children__.

Random Forests

Création du modèle



Trouver ses paramètres optimaux

Paramètre par défaut

SCORE R^2 :

0.831

Paramètres optimisés

SCORE R^2 :

0.854

'max_depth': 4

"n_estimators': 38':

Traitement des features

SCORE R^2 :

0.852

"smoker" 70%

"bmi" 16%

'"age" 11%

Comparaison

Régression linéaire

SCORE R^2 :

0.76

VS

Forêts aléatoires

SCORE R^2 :

0.85

Le modèle "Forêts aléatoires" l'emporte !

ANTOINE MEYER -- ALEKSANDRINA STOYANOVA-CHRISTEN