We worked on Linear Regression on two different datasets, the Boston Housing Dataset and Prostate cancer Dataset. The first one contains 14 features, the feature to explain is the last column : medv, other are explicative variables. It contains N = 506 observations.

We developed general methods that can be applied to any datasets.

Handle missing values :

After loading the data, the first step is to replace missing values. We chose to replace each misssing value by the mean of the corresponding feature.

Split train/test :

We then split the data between training and test set. The Training set is used to train the model but also to choose hyper-parameters. In our case the hyper-parameter involved is the number of features selected to train the model. The test set is only used at the very end to test the performances. To split the data we first shuffle it and then selected the first (1-p)\*N observations for training set and others for test set.

Normalization :

We the normalize the data with a min max scaler so that each feature has the same range between 0 and 1. It is not compulsory since normalization is a linear operator, so it can be found by the Linear regression model, but it can speed up the training process.

Feature selection :

We used two different method to select features