# ML01 – Spring 2021
# Lab 7: tree-based methods

## 1 Spam dataset

1. Load the `spam` dataset. Split the data into a training set (approximately 2/3 of the data) and a test set.
2. Build a classification tree to predict the class of an email (spam/email). Represent this tree graphically. Compute the corresponding confusion matrix and error rate.
3. Optimally prune the tree and represent the obtained pruned tree. Compute the corresponding confusion matrix and error rate. Did pruning improve the performance ? The interpretability ?
4. Apply random forests to this data set.
5. Compare the results to those obtained using LDA and logistic regression.
6. Which predictors are significant according to logistic regression ? Compare this list with the variable importance estimates from random forests.

## 2 Prostate data

Apply regression trees and random forests to predict `lpsa` from the `Prostate` dataset. Compare their performances to those of different variants of linear regression (least squares, ridge, lasso, variable selection). Compare the variables in the regression tree and the variable importance estimates of random forests to the variables selected using the lasso or stepwise selection methods. Interpret the results.