Exercise

Tree-based methods

Use the data from
`http://archive.ics.uci.edu/ml/datasets/Bank+Marketing`,
which are also available on our TUWEL course. Load the smaller data set using
`d <- read.csv2("bank.csv")`. The data contain information about direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit or not. This information is contained in the binary variable y (last one).

1. *Classification trees*: function `rpart()` from the R package `rpart`
   (a) Select randomly a training set of a reasonable size and apply a tree T_0 (see `help(rpart)` or lecture notes).
   (b) Visualize the tree with the function `plot()` and `text()`, and interpret the results.
   (c) Predict the group membership for the test set (see `help(predict.rpart)` or lecture notes). How high is the resulting misclassification rate?
   (d) Show and interpret results of cross-validation obtained by using `printcp()` und `plotcp()`. What is the optimal complexity?
   (e) Prune the tree T_0 of the optimal complexity using `prune()`. Visualize und interpret the results.
   (f) Predict the group membership for the test set and calculate the resulting misclassification rate. Do we observe any improvement?
2. *Random forests*: function `randomForest()` from the R package `randomForest`
   (a) Use the option `importance=TRUE` in the function `randomForest()`, and plot the result object with `plot()` and `varImpPlot()`. How can you interpret these plots?
   (b) Look at the misclassification error. Try to make the error of the "yes" clients smaller (by keeping the overall misclassificatione error still small) by using different strategies.
      i. Undersampling: randomly select from the bigger group the same number of observations that is available in the smaller group.
      ii. Modify the parameter `sampsize` in the `randomForest()` function. What is it doing?
      iii. Modify the parameter `cutoff` in the `randomForest()` function. What is it doing?

      Which approach leads to the overall best solution?