Presenting groups: 15 and 16, Date: 07.07.2021

The aim of this exercise: compare prediction properties of a normal regression classification tree with ensemble methods. You can use pre-installed packages such as **caret** for boosting.

Exercise 1:

Consider the following data generating process in which we have $n = 500$ observations and p covariates $X_j \sim \mathcal{N}(0, \sigma_j^2)$ for $j = 1, \ldots, p$. The values of p and σ_j^2 are chosen by you (be ready to have some basic explanation). Response variable y is generated by some nonlinear function of $\mathbf{X}$ that should include some (possibly higher order) interaction terms. Again, the choice of the specific function is up to you.

1. Generate the data according to the dgp described above and fit a classification tree with optimal pruning and a boosted tree.
2. Compare the test classification errors for both methods.

Exercise 2 (Simulation study)**:**

The goal here is to think about the way how a regression tree makes its predictions and how boosting improves these properties.

1. Propose a dgp that will improve the boosting classification error vs. the traditional regression tree. Illustrate the properties of your dgp in some simplified graphs.
2. Propose and implement another ensemble method of your choice (either bagging or random forests) and benchmark these methods against boosting within the framework above.