Data Analytics - Fall 2023
Exercise 4

DUE IN: Tuesday, 24.10.2023 in class,

FORMAT: You are strongly encouraged to work on the assignments and to provide a solution for them. For each week a different team is in charge, see file **Teams.pdf** on Microsoft Teams. The team in charge shall prepare a 15 minutes presentation showing and explaining their solution approach to the assignment.

The assignments will be created with R in mind as tool for using. For each assignment two teams will be in charge: one preparing a solution in R, the other one in Python.

Bagging, Random Forest and Boosting

For doing this assignment in R you need the R packages `MASS, tree, randomForest`, and `gbm`.

1. Briefly answer the following questions:
   (a) What is the general idea of ensemble learning?
   (b) What is the difference between bagging and boosting?
   (c) What are the advantages and disadvantages of both?
2. The dataset for this assignment is the `California Housing` dataset from `Kaggle`. The output variable is `median_house_value` (Regression task). The data can be downloaded from `https://www.kaggle.com/datasets/camnugent/california-housing-prices`.
   (a) Check for missing values and impute any missings by median value imputation.
   (b) Split the dataset into training set (80%) and test set (the remaining 20%)
   (c) Train a decision tree with the `tree()` function from the `tree` library and predict the output variable from the test set.
   (d) Calculate the MSE for the test data.
3. Bagging and Random Forest
   (a) Use the `randomForest()` function from the library `randomForest` to train the bagging regression model.The `randomForest` function has multiple arguments, some of them referring to hyperparameters of the bagging model. The hyperparameter `mtry` selects the number of features to be included in each trained model. As the dataset has in total 13 features (predictors) what number should you select for `mtry`, to get a bagging model? Why? Run a bagging model and check the MSE in the test data for this model.
   (b) What is the default number of bootstrap replications used in the `randomForest` function? Train another bagging model with 30 bootstrap replications by specifying the corresponding parameter in the `randomForest()` function. How does the MSE for the test data change? Did you expect this result?

(c) Now fit a random forest model using the default settings of the function `randomForest`. Compute the MSE for this model.

(d) Now vary the parameter `mtry` in the `randomForest` function in a loop from the default value to the maximum possible value in this case. Which `mtry` value yields the smallest MSE in the test data

(e) For this "best" model set the hyperparameter `importance = TRUE` in the `randomForest()` function. Run `importance(YOUR_randomForest_FUNCTION)`. What do the two measures of variable importance tell? Explain briefly.

4. Boosting

   Use the `gbm()` function from the `gbm` library to build a boosting model with 5000 trees and the number 4 for interaction.depth, which sets the depth of the tree models included. Check the MSE.

5. Which of the models computed in this assignment gives the best MSE for the test data?