Data Analytics - Fall 2023
Exercise 3

DUE IN: Tuesday, 17.10.2023 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.

Insurance data set

We continue with the analysis of the insurance dataset from the following link `https://www.kaggle.com/mirichoi0218/insurance`.

1. Install and load the library `tidyverse` as well as the library `caret`. Use the commands `help(‘‘% > %’’), help(dummyVars)` and `?preProcess` to get some basic idea about the pipe operator, dummy variable coding, and the preProcess command.
2. First, split the data into training and testing data using a 70-30 split.
3. Use the `dummyVars` command to turn the categorical predictors into dummy variables. Which are the categorical predictors? What happens when you turn them into dummy variables?
4. Use the pre-processing commands `nearZeroVar, findCorrelation` and `findLinearCombos` in the caret package to perform the following data cleaning steps on the training data and test data separately:
   (a) check for zero and near-zero variance variables (and remove if needed)
   (b) check for highly correlated variables
   (c) check for linear combinations
5. Read about Feature Scaling methods: ”Standardization”, ”Normalization”, ”BoxCox- Transformation”, and ”Yeo-Johnson Tranformation”.
   (a) Which feature scaling method is more sensitive against outliers? Standardization or Normalization?
   (b) Whats the main difference between BoxCox-Transformation and Yeo-Johnson-Tranformation?
6. Use the `preProcess` command to perform the following data transformation/feature scaling steps (you can do it in on ego) for trainign and testing data separately:
   (a) Centering

(b) Scaling

(c) Box-Cox transformation

7. Add the response variable charges to the training data frame obtained in the previous step and perform a multiple linear regression with charges as the response and all other variables as predictors.

   (a) Use the `summary()` function to print the results. Comment on the output. For instance:

       - Is there a relationship between the predictors and the response?
       - Which predictors appear to have a statistically significant relationship to the response?
       - What does the coefficient for the age variable indicate?

   (b) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

8. Now, compute a regression tree for the training data.

9. Compare the obtained regression tree model with the linear regression model on the test data. Which model performs better?