Data Analytics - Fall 2023
Exercise 2

DUE IN: Tuesday, 10.10.2023 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.

Breast Cancer Wisconsin data set

Download and load the Breast Cancer Wisconsin (Diagnostic) CSV dataset from kaggle. The dataset has 569 samples and 32 features.

1. Install and load the library `MASS` as well as the library `class`.
   (a) Look at the structure of the dataset and remove the "id" column.
   (b) Use

   ```
   table(dataset$diagnosis)
   ```

   to see the two classes.
   (c) Many classifiers in `R` require the target feature to be factors. Change the classes "B" and "M" into factors with the `factor()` function.
   (d) Turn the values of the target feature into percentages by doing the following:

   ```
   round(prop.table(table(wbcd$diagnosis)) * 100, digits = 1)
   ```

2. Now, you explore and preprocess the data a bit.
   (a) Look into the statistics of the data with summary()
   (b) Normalize the data by writing your own function in R. The math for the normalization is the following:

   $$\frac{(x - min(x))}{(max(x) - min(x))}.$$

   (c) Afterwards apply the created function onto the independent variables:

   ```
   as.data.frame(lapply(DATASET[2:31], normalize))
   ```

   (d) Now split the dataset into training (`DATASET[1:469,]`) and test set(`DATASET[470:569,]`) and then add the diagnosis column back to the datasets by using `cbind()`

   Side note: You should read a bit into `lapply, sapply,` etc. it's powerful!

3. It's showtime!

   (a) Train the following classification algorithms on the training data and perform them on the testing data:

      i. LDA (MASS library)
      ii. QDA (MASS library)
      iii. Logistic Regression Classifier
      iv. k-NN (class library) (try out different k values)
      v. random forest (package randomForest)

   (b) Show the test error of each classifier. Which classifier performs best in this case and which value of k is the best for the k-NNs tried out?.