# Exercise 4

## Advanced Methods for Regression and Classification

November 24, 2022

Download the file `hcvdat1.csv` from our TUWEL course, and load these data into R with

`d <- read.csv("hcvdat1.csv")`. This data set has been taken from the UCI Machine Learning Repository, and it contains laboratory values of blood donors and Hepatitis C patients, as well as information on age and sex. The variable `Category` contains information about the diagnosis: BloodDonor (healthy), Cirrhosis, and Hepatitis. Remove observations which contain missings (with `na.omit()`).

We want to apply discriminant analysis for the grouping variable `Category`, by using all remaining variables.

1. Start with visualizing the data using PCA (principal Component Analysis): compute the first 2 PCA scores by using all variables except `Category`, and visualize these PCA scores, with color information by `Category`. What can you see? Which difficulties could we expect for classification?

2. Select a training set of about 2/3 of the observations. The remaining test set observations should be used for the evaluation (misclassification rate).

   (a) *Linear Discrimant Analysis (LDA)*: function `lda` from `library(MASS)` Compare the test set error with a CV-error.

   (b) *Quadratic Discrimant Analysis (QDA)*: function `qda` from `library(MASS)` Compare the test set error with a CV-error.

3. (Now the more tricky part of the exercise ...)

   (a) You will have realized that the group sizes are very different. Is there a way to improve the misclassification error by using a different sampling scheme (i.e. a different strategy to select training and test data)? And is there an evaluation measure with also somehow accounts for the heavily imbalanced group sizes.

   (b) Especially QDA will be instable for very small sample sizes (compared to the number of variables). Could we reduce the number of variables? How?

   (c) Could we generally figure out which variables are important for the classification task, and which ones essentially represent "noise"?