# Lab03: Classification and Regression Prediction Models

**Handed out:** Wednesday, Oct. 30, 2019

**Return date:** Monday, Nov. 11, 2019, at the beginning of the class.

**Grades:** Lab03 counts 12 % towards your final grade.

**Objectives:** Comparison of different classification as well as regression tree models and the evaluation of their predictive properties.

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab03, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the bold task and sub-task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code the use mathematical type-setting when equations are required. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

## Part 1: Classification trees [7 points]

You will be using for this part the dataset **credit.csv** and split it into a ***training*** data-frame with 800 observations and ***test*** data-frame with 200 observations. The dependent variable is **default**.

**Task 1:** Build well-fitting logistic regression model predicting the probability of default in the ***training*** data-frame. [0.5 point]

**Task 2:** Build a classification tree, properly prune it and interpret the pruned tree. Show the pruned and unpruned trees. Use the ***training*** data-frame. [1.5 points]

**Task 3:** Build a predictive model using the **randomForest** function with bagging based on the parameter **ntree**. Evaluate the relevance of the features. Use the ***training*** data-frame. [1 point]

**Task 4:** Build a predictive random forest model using the **randomForest** function with an optimal hyper-parameter **mtry**. Use the ***training*** data-frame. [1 point]

**Task 5:** Build a predictive boosted model using the function **C5.0** and find the optimal depth hyper-parameter **trial**. Use the ***training*** data-frame. [1 point]

**Task 6:** Compare the models from tasks 1 to 5 for the ***test*** data-frame by using the functions **CrossTable**, **sensitivity** and **specificity** and the **auc** statistic. In addition, in one properly labelled plot show the ROC curves for all five models. Interpret the results. [2 points]

## Part 2: Regression trees [5 point]

You will be using for this part the dataset **redwines.csv** and split it into 80% ***training*** data and 20% ***test*** data. The dependent variable is **quality**.

**Task 7:** Use lasso regression to identify a set of relevant variables using the ***training*** data-frame. Calculate the model fit for the ***test*** data-frame. [1 point]

**Task 8:** For the identified relevant variables use generalized additive model estimation procedure with smoothing splines for the metric variables and plot the components of the resulting model. Decide for each metric variable whether a spline is needed to model the data. For model calibration use the ***training*** data-frame. Calculate the model fit for the ***test*** data-frame. [1 point]

**Task 9:** Build a pruned regression tree with all feature variables and interpret the pruned tree. Show the pruned and unpruned trees. For model calibration use the ***training*** data-frame. Calculate the model fit for the ***test*** data-frame. [1 point]

**Task 10:** Calibrate for the ***training*** data-frame with all feature variables a random forest model and identify its optimal hyper-parameter **mtry**. Calculate the model fit for the ***test*** data-frame. [1 point]

**Task 11:** Calibrate for the ***training*** data-frame with all feature variables a boosted model and identify its optimal depth hyper-parameter **trial**. Calculate the model fit for the ***test*** data-frame. [1 point]