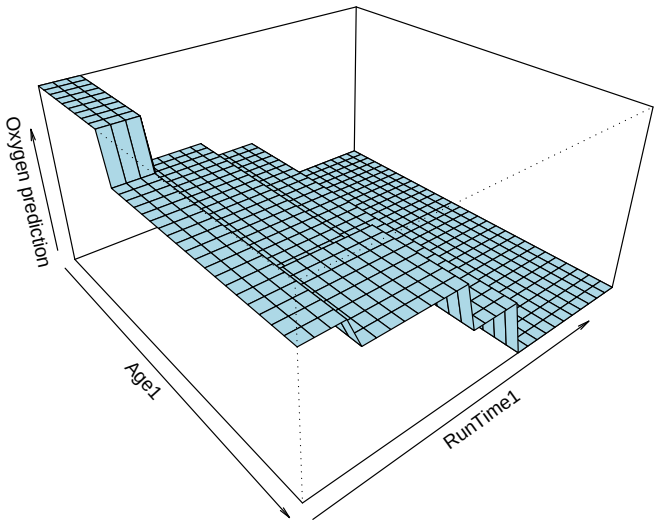
Classification and Regression Trees

Tasks (Lab 8):

(**all tasks are scored**)

1. Classification trees. Consider dataset *SAheart.data* (South African Heart Disease) containing information about patients in Age between 15 and 64. Target variable **chd** indicates the presence of myocardial infarction. The detailed description you can find in file *SAheart.info*.
   (a) Fit classification tree. Check how different parameters affect the structure and the size of the tree.
   (b) Draw a structure of the tree.
   (c) Choose the optimal tree using cost-complexity criterion.
2. Dataset *fitness.txt* corresponds to men's performance parameters measured in the 1.5 mile run. We consider the following variables:
   - **Oxygen** oxygen uptake intensity (TARGET VARIABLE),
   - **Age** age,
   - **Weight** weight,
   - **RunTime** run time,
   - **RestPulse** resting pulse,
   - **RunPulse** averaged pulse while running,
   - **MaxPulse** maximal pulse while running

   (a) Fit regression tree using default parameters. Visualize the structure of the tree.
   (b) Using the fitted model, answer the question: for which runner the oxygen consumption is assessed as the highest?
   (c) Make a prediction for observation described by feature vector x_0, for which coordinates are equal to the means of the variables (so x_0 is a typical runner).
   (d) Choose the optimal sub-tree, you can use e.g. cost-complexity criterion.
   (e) Fit a tree model using only two variables: **RunTime** and **Age**. Make a visualization of the predicted values **Oxygen**. Example visualization is depicted below.
3. Implement your own version of bagging algorithm. You can use available implementations of decision trees (or other base learners). Compare the accuracy (compute for different train/test splits) of bagging and single tree using the two above datasetes.


Oxygen prediction
Age1
RunTime1