

# The data mining project:

- 1) Your third project task is to choose data from my kaggle.com project list:  
[https://github.com/pwasiewi/dokerz/blob/master/rstudio/00kaggle\\_projects\\_to\\_learn.R](https://github.com/pwasiewi/dokerz/blob/master/rstudio/00kaggle_projects_to_learn.R)  
(I have selected just smaller sets – maybe not all links work or not all sets are suitable for case studies)  
or from <http://archive.ics.uci.edu/ml/datasets.html> and **make a case study**.
  - a) Do not choose datasets used during my lectures or too simple to make your case study (I mean datasets with a small number of attributes taken from games or simulations).
  - b) The used datasets should be obtained from the real world research. I think they are more interesting and appropriate.
  - c) I proposed some datasets, but you should choose the appropriate one for your case study (fitting your classifiers e.g. trees, svm, regression and so on).
- 2) First you may read and run (maybe using my docker 42n4/rstudio?) some exemplary case studies from the Cichosz book commented during my lectures:  
[https://github.com/pwasiewi/dokerz/blob/master/rstudio/dm\\_casestudy01.R](https://github.com/pwasiewi/dokerz/blob/master/rstudio/dm_casestudy01.R)  
[https://github.com/pwasiewi/dokerz/blob/master/rstudio/dm\\_casestudy02.R](https://github.com/pwasiewi/dokerz/blob/master/rstudio/dm_casestudy02.R)  
[https://github.com/pwasiewi/dokerz/blob/master/rstudio/dm\\_casestudy03.R](https://github.com/pwasiewi/dokerz/blob/master/rstudio/dm_casestudy03.R)
- 3) You should provide a (short - max 2-3 pages) documentation file (tex, word, txt) to my email containing:
  - a) the description of attributes (columns) e.g. the number of discrete and continuous attributes (make factors from discrete ones, get rid of useless ones or remove one attribute from each pair of strongly correlated attributes). Find or make your target class attribute.
  - b) the description of used classification methods and their validation process (changing their parameters e.g. minsplit, cp, k-fold crossvalidation, boosting, bagging, adding cost matrices to them, modifying input data e.g. standardization i.e. standard score, normalization, removing NA and so on).
  - c) the validation process summary including ROC plots and their comparison.
  - d) You may use some clustering methods for not labeled datasets (without an obvious target attribute). After this operation you can learn classifiers utilising input data and obtained cluster labels.
  - e) Enclose please your R code with comments and links to data used in R code (if it was changed before read.csv then dropbox links to your new data).
- 4) All projects should be done individually and may be done in pairs on that condition that you divide your work into two parts belonging to individual persons, but each person should make some classifier validations.
- 5) **Deadline for at least preliminary version: the 10<sup>th</sup> of June at 2 o'clock room 25A**, the final version deadline – the end of August by email. Of course, it can be sent earlier.
- 6) The suggested language is R (ask for permission to use another programming language such as python).  
The proposed classifier library wrapper: caret
- 7) My e-learning materials are provided at these sites (among them R scripts from lectures based on the Cichosz book samples e.g. dm\_models.R, dm\_roc.R, dm\_tree\_basics.R):  
<https://github.com/pwasiewi/earin>  
<https://github.com/pwasiewi/dokerz/tree/master/rstudio>  
The latter site has instructions to run a docker with all needed libraries for three lecture case studies.
- 8) My office hours on Mondays at 2 o'clock pm in room 25A. Just email me and I will help you to choose your dataset, if you have some doubts or just to make further research.

My email: [pwasiemi@elka.pw.edu.pl](mailto:pwasiemi@elka.pw.edu.pl)