Finding a dataset for the reports and group registration

Objective: The exam of this course includes two written group reports to be completed during the semester:

1. Data: Feature extraction, and visualization
2. Supervised learning: Classification and regression

The reports must be completed in groups of no more than 3 persons and will make use of a dataset you choose. This can either be your own dataset, or one selected from the resources given below. After you have selected a dataset, contact a teaching assistants to register the group and discuss any potential issues with your choice.

- `http://archive.ics.uci.edu/ml/datasets.html` Examples of data sets that could be interesting to analyze: Ecoli Data Set, Glass Identification Data Set, Concrete Compressive Strength Data Set. **Notice: Do not take the Wine Quality data set as this will be used in the course!**
- `https://web.stanford.edu/~hastie/ElemStatLearn/` Examples of data sets that could be interesting to analyze: Los Angeles Ozone, Marketing, NCI (microarray), Phoneme, Prostate, Protein flowcytometry data, SRBCT microarray data, South African Heart Disease, Spam, Vowel.
- `http://www.kdnuggets.com/datasets/index.html`
- `http://www.statsci.org/datasets.html`
- For SAS-bachelors the following source is also relevant: `http://www.cengage.com/aise/economics/wooldridge_3e_datasets/`, see the `excelfiles.zip` link which contains datasets and their descriptions in separate files. Examples of data sets that could be interesting to analyze: `AIRFAIR`, `HPRICE2`, and `LOANAPP`.

As a guideline, your dataset should have at least 60 observations and 5 attributes with at least two of the attributes being interval or ratio.

We recommend you read the text for project 1 and consider if the tasks you are required to carry out on your dataset appears, i.e. what variables will you be interested in predicting, etc. Note no single dataset will be ideally suited for all methods, and an aspect of the report work will be to make meaningful choices and interpretations of the results along the way. Talk to the instructors if you have doubt. Please avoid datasets consisting of images, sounds or time-series data as they will likely be unsuitable.

Approval and group registration: Once you have found a dataset you find interesting, you need to have the dataset approved by one of the *teaching assistants* of the course. Please discuss any questions you have about your dataset with your teaching assistant and then send him/her an email with details about your group. *Make sure the email contains the study numbers of all members of the group.*

This information will later be used to inform how reports are distributed amongst instructors and make subsequent feedback on project work easier[1]. The emails of the teaching assistants can be found on campusnet or the course homepage `http://www2.imm.dtu.dk/courses/02450/`

Deadline: Please have your dataset approved at the end of exercise 3, 18 February, 2020.

[1]Note reports may be re-distributed to even the workload