02450 Project 1:
Project description for report 1

Objective: The objective of this report is to apply the methods you have learned in the first section of the course on "Data: Feature extraction, and visualization" on your own data set to get a basic understanding of your data prior to the further analysis (project report 2 and 3).

Resources: You can use the 02450Toolbox on Campusnet to see how the various methods learned in the course are used in Matlab, R or Python. In particular, you should review exercise 1 to 4 in order to see how the various tasks can be carried out.

Preparation: Exercise 1-4.

Understanding well the data you are to model is very important. You can apply very sophisticated machine learning methods but if you are not aware of potential issues with the data the further modeling will be difficult. Thus, the aim of this first project is to get a thorough understanding of your data.

Project report 1 should cover what you have learned in the lectures and exercises of week 1 to 4 covering the section "Data: Feature extraction, and visualization". In particular, the report ***must*** include the following:

1. **A description of your data set.**
   What is the problem of interest, where did you obtain the data, what has previously been done to the data (i.e. if available go through some of the original source papers and read what they did to the data and summarize what were their results). Explain what the primary machine learning modeling aim is as well as how you envision the data can be analyzed in terms of; a classification, a regression, a clustering, an association mining, and an anomaly detection problem. (You need to outline how all the methods may be applied to your data).

2. **A detailed explanation of the attributes of the data.**
   (I.e. discrete/continous, Nominal/Ordinal/Interval/Ratio), data issues (i.e. missing values, corrupted data), the basic summary statistics of the attributes.

3. **Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).**
   Are there issues with outliers in the data, do the attributes appear to be normal distributed, are variables correlated, does the primary machine learning modeling aim appear to be feasible based on your visualizations.
   Keep in mind the ACCENT principles and Tufte's guidelines when you visualize the data. Notice also that there are three aspects that needs to be described when you carry out the PCA analysis for the report: 1) The amount of variation explained as a function of the number of PCA components included, 2)

the principal directions of the considered PCA components, 3) the data projected onto the considered principal components. (If your attributes have very different scales it may be relevant to standardize the data prior to the PCA analysis.)

4. **A discussion explaining what you have learned about the data.** Summarize here the most important things you have learned about the data and give also your thoughts on whether your primary modeling task(s) appear to be feasible.

The report should be 5-10 pages long including figures and tables and give a precise and coherent introduction to and overview of the dataset you have chosen. A printed copy of the report is to be handed in at the beginning of the lecture Tuesday March 5th.