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 of dataset.

- `https://archive.ics.uci.edu/ml/index.php` 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 (with no missing or erroneous values), and 5 attributes with ideally 3 of the attributes being interval or ratio.

We recommend you read the description for project 1 which is availble on DTU Learn, and additionally consider that you will be asked to do regression and classifiation on your dataset in project 2.

You should consider if the tasks you are required to carry out on your dataset appears feasible. Consider what variables you will use for 1) PCA / visualization (project 1), 2) regression (project 2), and 3) classification (project 2). A few general guidelines are provided below:

- PCA: The variables on which you want to apply PCA should typically be Interval, Ratio (or in special cases Ordinal).
- Regression: The variable you want to predict should typically be Interval, Ratio or in special cases Ordinal attributes. The attributes from which you intend to predict the class label are typically Interval, Ordinal or Ratio. If your intended variables are different from indicated here, you probably need to consider if a transformation can be applied to make the attributes/data appropriate for a regression task.
- Classification: The class label you want to predict should typically be associated with a Nominal attribute. The attributes from which you intend to predict the class label should typically be Interval, Ordinal or Ratio. If your identified attributes are different from indicated here, you may need to consider if a transformation can be applied to make the attributes/data appropriate for a classification.

Important: No single dataset will be ideally suited for all methods, and an aspect of the project work will be to make meaningful choices, transformations and interpretations of the results along the way. Talk to the instructors if you have doubt about the dataset.

Please avoid datasets consisting of images, sounds or time-series data as they will likely be unsuitable.

Dataset registration and approval: Once your Group has found a dataset you find interesting, please register it on DTU Learn. We suggest you discuss it with your instructor before you enter the data on DTU Learn. Once you have registred the data let you instructor know (with your group id) and they will do a quick inspection and let you know if there are any obvious issues.

This information will be used to catch issues early on in the process and later used to inform how reports are distributed amongst instructors and make subsequent feedback on project work easier[1].

Deadline: Please have your dataset registred on DTU Learn before lecture 4, 22 February, 2022.

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