Finding a dataset for the reports and group registration

Objective: The assesment in 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 3 persons (unless agreed otherwise with your TA/staff) 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.
- `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`.

Notice: You are not allowed to choose the Wine Quality, Iris or Whisky dataset used in the in the lecture notes!

As a guideline, your dataset should have at least 60 observations (with no missing or erroneous values), and 5 attributes with at least 3 of the attributes being continuous variables (e.g. interval or ratio).

We recommend you read the description for project 1 which is available 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 must carry out on your dataset appear feasible. Consider what variables you will use for: 1) PCA / visualization (project 1), 2) regression (project 2), and 3) classification (project 2). Please read Chapter 2 in the lecture notes along with the following guidelines:

- PCA: The variables on which you want to apply PCA should typically be continuous (Interval, Ratio or, in special cases, Ordinal). You can attempt to include categorical (e.g., binary) variables, but we do not recommend this unless you have discussed it with your TA.
  For example, for the Iris dataset `https://archive.ics.uci.edu/ml/datasets/iris`, it would be relevant to consider a PCA analysis of the sepal length, sepal width, petal length, petal width which are all continuous (Ratio) variables.

- Regression: The variable you want to predict should be continuous attributes (Interval, Ratio, or Ordinal). If your intended variables are different from the ones suggested here, you probably need to consider if a transformation can be applied to make the attributes/data appropriate for a regression task.
  For example, for the Iris dataset `https://archive.ics.uci.edu/ml/datasets/iris`, it would be relevant to perform regression from say sepal length, sepal width, petal length to petal width, i.e. we are interested in predicting if petal width can be predicted from the other attributes (without knowing the class labels/Iris type).

- Classification: The class label you want to predict should typically be associated with a discrete variable (i.e., a Nominal attribute). You can often create a discrete from a continuous variable by quantizing it into intervals. If your identified attributes are different from those indicated here, you may need to consider if a transformation can be applied to make the attributes/data appropriate for classification.
  For example, for the Iris dataset `https://archive.ics.uci.edu/ml/datasets/iris`, it would be relevat to consider if we can predict the class label (i.e. if the Iris is Setosa, Versicolour or Virginica) from the sepal length, sepal width, petal length, and petal attributes. Alternatively, imagine we do not have the class label. In this case, we could construct a classification task by predicting if one of the continuous variables is *high* or *low* by discretizing, for example, petal width.

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

Avoid datasets consisting of images, sounds or time-series data as they will likely be unsuitable given the methods we consider in this course.

Group registration and dataset check:

You need to register your group on DTU Learn (under My Course→Groups) and inform your TA about the group number (from DTU Learn) and the dataset you have choosen.

Deadline: Please have your group registred on both DTU Learn and with your TA before lecture 4, 20 September, 2022.