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,*"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).

Material: You can use the code from the exercises to see how the various methods learned in the course are used in Python. In particular, you should review exercise 1 to 4 in order to see how the various tasks can be carried out.

Deliverables: The group-based project work should result in a written report (pdf format) and associated code required to reproduce the results in the report.

The report should be maximum 10 pages long (A4, min. 11 pt font, min. 2 cm margin) including figures and tables and give a concise, correct and coherent introduction to and overview of the dataset you have chosen. An appendix may be included with additional results to backup statements/discussion in the main report but the assesment is based soley on the main report.

Group work and individual contributions: The project must, as default, be be carried out in a group of 3 students. Each student's contribution to the report must be clearly specified, thus for each section specify (in **a table on the frontpage**) who was responsible for it. Every team member is resonsible for the report and must (ideally) contribute to all parts of the report. For reports made by 3 students each section must have a student who is 40% or more responsible. For reports made by 2 students each section must have a student who is 60% or more responsible. Permission to work alone requires extraordinary circumstances and explicit approval from the main teacher.

Assesment: Submissions are evaluated based on the degree to which the report concisely, correctly and completely addresses the tasks and questions below. The submitted code is not assesed per say, but is used to validate correctness of the reported result.

Deadline: The **deadline for handin is no later than Thursday 2 October at 17:00 CET via DTU Learn**. Late handins will not be accepted under normal circumstances and without consent from the main teacher.

Submission checklist:

- Your submission should consist of exactly **two files**: A `.pdf` file containing the report, and a `.zip` file containing the code you have used (extensions: `.py`, `.R` or `.m`; do **not** upload your data). The reports are not evaluated based on the quality of the code (comments, etc.), however we ask the code is included to avoid any potential issues of illegal collaboration ¨between groups. Please do not compress or convert these files.
- Make sure the report clearly display the **names *and* study numbers** of all group members on the frontpage. Make absolutely sure study numbers are correct.
- Make sure you have (at least) addressed all the tasks and questions below in your report!

Description

Understanding the data you are trying to model well 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 and describe how you expect the data can be used in the later reports.

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"* . You should consider yourself as a new employee in a company who has just been given a data set. Your job is to make a useful description of the data set for your co-workers and make some basic plots. In particular, the report ***must*** include the following items and the report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality. For readability and brevity consider not using one subsection for each item.

1. **A description of your data set.**

   - Explain the overall problem of interest and the associated data.
   - Provide a reference to where you obtained the data.
   - Summarize previous analysis of the data. (i.e. go through one or two of the original source papers and read what they did to the data and summarize their results).
   - You will be asked to apply (1) classification and (2) regression on your data in the next report. For now, we want you to consider how this should be done. Therefore:

     – Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques?
     – Explain which attribute you wish to predict in the regression based on which other attributes?
     – Which class label will you predict based on which other attributes in the classification task?
     – Explain if you need to transform individual attribues in order to carry out these tasks (e.g. centering, standardization, discretization, log transform, etc.) and how you plan to do this.

     One of these tasks is likely more relevant than the rest and will be denoted the **main machine learning aim** in the following.

   The purpose of the following questions, which asks you to describe/visualize the data, is to allow you to reflect on the feasibility of the **main machine learning aim**.

2. **A detailed explanation of the attributes of the data.**

   - Describe if the attributes are discrete/continuous and whether they are nominal/ordinal/interval/ratio.
   - Give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so and how you will handle them.
   - Include relevant summary statistics of the attributes. Reflect on the values.

   If your data set contains many similar attributes, you may restrict yourself to describing a few representative features (apply common sense). You can place additional results in the appendix if needed.

3. **Data visualization(s) based on suitable visualization techniques.**
   Touch upon the following aspects, use visualizations when it appears sensible.
   Keep in mind the ACCENT principles and Tufte's guidelines when you visualize the data.

   - Are there issues with extreme values or outliers in the data?
   - How are the individual attributes distributed (e.g. normally distributed)?
   - Are the attributes correlated?

There are three aspects that needs to be addressed when you carry out the PCA analysis for the report:

- The principal directions of the considered PCA components. Plot and interpret the components in terms of the attributes.
- The amount of variance explained as a function of the number of PCA components included.
- The data projected onto the considered principal components, e.g. in 2D scatter plots (hint: it may be helpful to color code the points according to the value of the attribute you wish to predict).

Hint: If your attributes have very different scales, it may be helpful to standardized the data prior to the PCA.

4. **A discussion explaining what you have learned about the data.**
   Summarize the most important things you have learned about the data and give your thoughts on whether your primary machine learning aim appears to be feasible based on your visualization.

Collaboration and Plagiarism

The usual DTU rules for collaboration and plagishm applies for the reports. The main rule is that if you hand in a report, you must have authored or co-authored the content of the report for this assignment, and if your report contains text you did not write, then it must be with attribution. Notice in particular:

- You are of course allowed to use the scripts, etc. supplied in this course for the reports.
- If you have used AI tools (e.g. Copilot, ChatGPT) to assist in writing the report or code, it must be clearly stated in the report where and how the tools have been used.
- If you are authoring a report together with a person who has previously taken the course, you cannot re-use that report since you did not originally author it. We recommend that you simply choose another dataset and re-write the text such that the new report can be considered original joint work by both authors.
- If you are taking the course again, you are allowed to re-use content from a report that you previously authored or co-authored.

Discussions and collaboration related to all aspects of the problem is obviously strongly encouraged within the group. Automatic plagiarism checks will be carried out across all the submissions to ensure that there is no plagiarism amongst the groups and with all previously submitted reports in the course. Plagiarism among groups or from third parties will be reported.