Assignment #01

IT University of Copenhagen (ITU)
Data Mining, KSD (DAMIN)
(Autumn 2024)

Deadline: October 24, 2024 at 23:59

Introduction For this mandatory assignment, you will complete a small-scale data mining project. You will work with a dataset similar to the one created during Lecture #02 based on the questionnaire you all filled out. You must use this data to answer one or more questions you define. Then, you will write and analyze the necessary code to answer those questions.

Purpose This assignment aims to give you a taste of what a data mining project looks like. You will work with a dataset, write code to analyze it and write a report to present your findings. This assignment will also help you understand the importance of defining a straightforward question before starting a data mining project.

General Concepts You must apply at least two pre-processing methods and one supervised learning method (classification or regression). For example, you could use normalization, missing value replacement, and linear regression. However, many other combinations are possible. You must write all the code for these algorithms. Additionally, your source code will only be accepted if it includes reasonable and clear comments.

Deadline The deadline for submitting Assignment #01 is Thursday, October 24, 2024, at 23:59. Please submit your solution via learnIT.

Teamwork You must complete this project in a group of 2-4 students. Therefore, you will work together to formulate the questions, write the code, conduct the experiments, and write the report. By default, all group members will receive the same grade (*passed* or *failed*). To get started, please inform your group members as soon as possible and record your group's composition in the Group Self-Selection resource on learnIT before the project submission deadline. Since group work can sometimes be uneven, you must include a one-page document titled "*Work Allocation*," which details who was responsible for each part of the coding and report.

Submission Information Your submission must include (i) *source code* and (ii) *report*. Please compress these two parts into **a single zip file** named `groupXY_solution.zip`, where

`XY` represents your group's ID (e.g., `group01_solution.zip`). Only one group member must submit the solution on learnIT.

Source Code The source code must be written in Python and include all the necessary code to answer your defined questions. The code must be well-commented and easy to read. Minor technical errors will not be a cause for failure, at least as long as they are acknowledged within the report. You must also include a `requirements.txt` file with all the necessary dependencies to run your code.

Report You must submit a two-page report as a .pdf file. Each page should contain up to 2,400 units (including spaces and notes). In the report, describe the questions you aimed to answer, the methods you used, the results you obtained, any problems you encountered (such as issues with the dataset), and any other relevant observations. Include numerical results and, if possible, graphs. If your report exceeds two pages due to layout or graphs, please include the total unit count (including spaces and notes) in the report.

Grading We will grade your assignment based on a pass/fail basis. Completing the assignment will not impact your final grade. If you fail the assignment or miss the submission deadline, you can resubmit it after the group project deadline. However, the resubmission will require you to implement additional algorithms, and the new deadline may interfere with your exam preparation. Please note that passing the resubmission is mandatory to be eligible for the final oral examination.

Plagiarism and Code Reuse Plagiarism will not be tolerated and result in the plagiarist failing the assignment. On the other hand, the assignment is a natural continuation of the work done during the labs. Therefore, you can reuse the code you wrote during the labs and even some data set analysis you performed. If you completed all the labs, you will have most of the work on the assignment. Therefore, use the lab sessions effectively, and don't be afraid of asking the TAs and teacher questions!