# Submission deadline is 4 January 2024,14:00

*(Early hand-ins are OK, however, please note that the hand-in button on LearnIT will only open 14 days before the official deadline)*

## Objective

The objective of the exam paper is to apply the methods you have learned in the course on a dataset of your choice in order to solve a relevant statistical problem or modelling task of your choice.

## The dataset

Pick a dataset of your own choosing. It can be something you want to work with for your MSc project, something you have access to via your employer, or something you have previously worked with. You may also use the dataset you're creating in the *Data Wrangling in the Wild* class, provided the analysis you make is different from that submitted in the other class, or you might start from a published academic paper that comes with raw data and reanalyse the results.

The size of the dataset should be appropriate for the modelling method that you're planning to use. You also need to ensure that the dataset its relevant for the problem you want to solve. As such, getting an initial understanding of your dataset is very important. You can apply very sophisticated machine learning methods but if you are not aware of potential issues with the data further modelling will be difficult.

## The task

Identify and describe one or more statistical problems/questions which you want to solve using techniques from the course with the found dataset. Problems can vary in type from: using multivariate statistics to predict customer churn, segmenting modes of failure on heavy machinery from operational data, identifying fraud in financial data, finding patterns in text or image data, to building models to estimate poverty from digital data. However, it is vital you ensure the dataset can be used to solve the problem. A good project will have a stringent connection from the research question asked to the methods chosen, the experiments, their evaluation and interpretation.

## Supervision and abstract submission

The two course meetings on 1 December and 8 December are reserved for project supervision. Each group can book a meeting with either Christian or Michael on each of these days. By 14:00 on Monday 27 November, you should submit the title of your project and a short abstract (no more than a page) that briefly describes the dataset and analysis methods you're planning to use. This can still be revised later.

## The hand-in

You are a group of 3-4 students (students are responsible for forming groups), and the exam hand-in is a written report. The report should be formatted using the IEEE conference template (A4 format, Word or LaTeX at your option). The template can be found here: https://www.ieee.org/conferences/publishing/templates.html

The report will be 6-10 pages long including figures and tables, but excluding bibliography and appendices, and contain:

- A precise and coherent introduction to the problem and overview of the dataset(s) you have chosen (*e.g. where did you get the data, what variables does it contain, what is the problem you are trying to solve, and why is it important to solve this problem*).

- A description of the method(s) you have used to analyse and model the data (*e.g. regression, classification, clustering, anomaly detection, etc. specify which type of model you are using and why, how hyper parameters were selected, etc).* Reason about why you have selected the specific method(s) and why they are appropriate to solve the problem.

- Present and interpret the results from your analysis and outline your conclusions (*did you manage to achieve what you set out to do and did you solve the primary problem(s)*). If available, relate your results to results others have found from analysing the same data.

- A discussion explaining what you have learned about the data. Summarise here the most important things you have learned and critically reflect on whether your dataset of choice can be used to solve your statistical modelling task(s).

- If you choose to use the dataset you created in the *Data Wrangling in the Wild* class, you should also include a short paragraph that highlights in what way the analysis in this report is different from that submitted in the *Data Wrangling* exam.

On LearnIT, you'll find an example with a suggested structure for the report. This is a suggestion, and you're not required to follow it strictly.

Your only deliverable is a report, and you don't need to include any code or data in your submission. However, the report must provide enough information for us to understand exactly what has been done, what data has been used, etc., so the work should be **reproducible** for a competent reader (given the same data) even without access to your files, code, etc.

The report may include appendices if necessary. This would be a good place for information on detailed parameter settings that are important for replicability without being central to the argument made in the paper, or you might include cite a short,

central code snippet to illustrate exactly how something was done. But the evaluation of your work will primarily be based on the main text of your report.

## Who to pitch your report to?

The target reader of your report is the "competent reader" mentioned above. You may assume that they have the same level of expertise as you, but haven't necessarily taken exactly the same classes. For instance, you could imagine someone who has a degree in data science from a different university. Write in a way that would allow such a person to follow your reasoning and replicate your work without trouble. Bibliography and references should be formatted consistently in accordance with standard practices of academic writing. Follow the IEEE style that matches the style files.

## Group grading

Group members are expected to work together on the hand-in documents, but each student will have individual main responsibility for a part of the analysis and a section of the report. Students must clearly identify which respective parts of the submitted work they are responsible for us to be able to establish a solid foundation for individual grading. The oral exam has both group and individual components. At the exam, you should be prepared to answer detailed questions about your own and any shared parts of the report.