# Compulsory Exercise 2

TMA4268 Statistical Learning V2024

Kenneth Aase, Daesoo Lee, Stefanie Muff, Sara Martino
Department of Mathematical Sciences, NTNU

---

**The submission deadline is: Monday April 15, 23:59h**

## Introduction

In this project, you will be working on analyzing a data set by developing *prediction/classification* models using statistical learning techniques.

The goal of this relatively open project is to give you hands-on experience, with the methods learned in the course, and to help you develop your skills in data pre-processing, model selection, and evaluation.

To get started, here is a broad guideline for your project:

1. **Choose a data set for your project.** We list two possible data sets below, or you can find one on your own.
    - **Heart Failure:** The data set is provided by user *Fedesoriano* on Kaggle. More information can be found in the provided link.
    - **AirBnB Prices in European Cities:** The data set is provided by user *The Devastator* on Kaggle. More information can be found in the provided link. Note that the description says that the dataset is designed for inference, but we will use it for prediction or classification. You can choose to focus on data from one city, or combine data from different cities.
    - **Choosing your own data set:** If you choose to use another dataset, ensure that that it is diverse enough and contains enough data points to train a good model. There are many example data sets provided by `R` packages (e.g. `carData`), and you can find a variety of data sets on Kaggle or TidyTuesday. Try searching for a topic that you find interesting and would like to work with. To find a "well-organized" data on Kaggle set we recommend choosing a data set with sufficiently-large amount of upvotes. Otherwise, a data set could be often unorganized or poorly structured with missing values. We do not want you to spend much time on data cleaning!
2. **Decide on a prediction/classification task.** What are you trying to predict/classify with the data set? Decide on this *before* you try out any of the models - it's fine if it turns out in the end that you are not able to get good predictions, as long as you have done everything correctly. Some ideas are provided here, but in principle you can decide on anything that makes sense.
    - For the heart failure data set you could try to classify whether a patient gets heart disease or not, predict their age, or predict their cholesterol.
    - For the AirBnB price data set, you could try to predict the price or the rating of of accommodation, or classify the room type or "superhost" status of the host.
    - If you chose another data set you should come up with a prediction or classification task that makes sense from the data.
3. **Data pre-processing.** Before building your model, you might need to pre-process the data (data wrangling), depending on what format your methods require.

4. **Choose appropriate** models and methods for your project. You should use methods that you have learned about in the course. Make sure to justify your choice of algorithm, check model assumptions and, if relevant, tune the hyperparameters to improve model performance. **Use methods from at least two different modules of the course**. You can also consider transformations of the variables in your data set, and interaction effects. Make sure that you do the model selection in a *valid* way, as you have learned in the course, and remember the bias-variance tradeoff.
    - If you are solving a prediction task, you could try methods such as multiple linear regression, GAMs, ridge/lasso regression, trees, random forests, boosting *etc.*
    - If you are solving a classification task, you could try methods such as logistic regression, LDA, QDA, KNN, regression trees, random forests and boosting *etc.*
5. **Model assessment.** Evaluate your model's performance using appropriate metrics or evaluation tools, such as accuracy, MSE, sensitivity, specificity, *etc.* Again, make sure that you do this in a *valid* way, as you have learned in the course.
6. **Reporting.** Present your results and state your findings and interpretation of the results.

We hope you find this project both challenging and rewarding. Best of luck!

## Grading

Maximal score is 100 points and the number of points given for each section of your report are indicated below. The grading is given by PASS/FAIL. To pass the compulsory exercise, your score must be at least 60.

## Supervision

We will use the times where we would have lectures and exercises for supervision.

Supervision hours (in the usual lecture rooms):

- Thursday, April 11, 8:15-10:00 and 16:15-18:00
- Friday April 12, 12:15-14:00

Remember that there is also the Mattelab forum, and we strongly encourage you to use it for your questions outside the supervision hours – this ensures that all other students benefit from the answers (try to avoid emailing the course staff).

## Practical issues (Please read carefully)

- You should work in the same groups as for compulsory exercise 1.
- Remember to write your names and group number on top of your submission file.
- The exercise should be handed in as two files: **one R Markdown file and a pdf-compiled version** of the `R` Markdown file (if you are not able to produce a pdf-file directly make an html-file, open it in your browser and save as pdf in portrait format). We will read the pdf-file and use the Rmd file in case we need to check details in your submission.
- Do not include the text from the file that you are reading now. We want your (relevant) `R` code, plots and written solutions - use the attached template `Compulsory2_template.Rmd`.
- Please **no more than 14 pages** in your pdf-file. **We will stop reading your report after page 14.** Keep this in mind when choosing what `R`-code/output to include, and when sizing your figures.
- In the `R`-chunks, carefully consider when to set `echo` and `eval` to `TRUE` or `FALSE`, to make it simpler for us to read and grade. Use common sense, and do not include irrelevant output and code.
- Please save us time and **do not submit word or zip**, and do not submit only the Rmd or pdf (submit **both**). This only results in extra work for us.
- **Bonus hint: Neat reports are easier to understand and may result in a better grade - simply because we cannot give full points if things are unclear, ambiguous or messy.** Pretend that the task you decided on was given to you by your boss at a company, and the report is

what you will deliver to the boss. When writing the report, keep in mind that boss has limited time and attention, and has not spent as much time as you have on getting familiar with the problem.

---

# Guideline for the Template

Please use the template `Compulsory2_template.Rmd` that we provide on the course website (under the *Compulsory Exercise 2* tab).

## Title

Replace the placeholder title by an informative title.

## Abstract (max. 350 words) (5 points)

The purpose of the abstract it to give a short and concise summary of your project. It is a stand-alone text that is given before the actual report starts. It includes the following components:

1. Begin your abstract by clearly stating the purpose of your project. What problem are you trying to solve? What question do you want to answer? It is important to be concise and to the point.
2. In the next few sentences, describe the data and methods you used to conduct your study. What kind of data set did you use? How did you analyze it? What tools, techniques, or methods did you use? Be specific, but avoid going into too much detail.
3. Summarize your key findings: In the main part of your abstract, summarize the most important results of your project and interpret them briefly (i.e., what does this mean?). Highlight the most significant findings, and provide enough detail to give the reader a sense of what you discovered.
4. (optional) Emphasize the significance of your results: Explain why or/and how your finding(s) is/are important. Highlight any novel or unexpected findings, and explain how they add to our understanding of the topic.

## Introduction: Scope and purpose of your project (15 points)

- Briefly introduce the broad idea of the problem or task that you chose and the respective data set that you use. This could be a classification task (e.g., predicting whether a patient gets heart disease or not) or a prediction task (e.g., predicting the price of an AirBnB). Clearly define the scope of your project. What specific problem are you trying to solve?
- Describe the source and give a reference to where the data set is coming from.
- Describe the purpose of your project in more detail. What are the specific question that you want to answer in your project? Are you trying to find the best performing method or a good performing and light method that is easy to use? Who is your audience? Are you trying to discover the relations between different variables? Are you trying to find important predictors for your classification? Are you trying to draw some insightful understanding in a particular topic/domain?

## Descriptive data analysis/statistics (15 points)

Conduct descriptive data analysis to get an overview over your data (see this example for inspiration). Try to **focus on what will be relevant for your modelling** and use common sense. For example, too much detail, or figures without any explanation or axis labels, are not useful to the reader.

For example:

- Report measures such as mean, median, range, standard deviation, and variance to describe the central tendency, variability, and distribution of a data set.

- Scatter plots and correlation matrices across different variables and histograms of variables (see this example).
- Box plots of variables.

## Methods (30 points)

- Describe the methods that you are using in your project and explain in detail how you applied them. You should use several methods for your problem so that you can compare their performance.
- Explain briefly how each method works, what its strengths and weaknesses are, both in general but also in the light of your project (how suitable is the method *in your case*?).
- Describe which hyperparameters are optimized for the methods (e.g., the shrinkage factor is a hyperparameter in Lasso regression).
- Describe clearly how you evaluate the performance of the different models and methods (accuracy, MSE, misclassification error, CV error,...). Explain how each metric is calculated, and why it is a useful measure of model performance.
- (optional) Consider and describe potential limitations of the methods and the chosen evaluation metrics.

## Results and interpretation (30 points)

1. Present your results in a *clear and organized* manner. This could include tables, graphs, or other visualizations that help to convey your findings. Report also all the hyperparameters, the performance (e.g., test error) etc. that you introduced in the Methods section.
2. Interpret the results. **You can compare the different methods in terms of aspects such as performance, computational cost, flexibility, bias-variance trade-off, etc.** The interpretation should depend on the prediction/classification task you decided on.
3. Discuss any limitations or caveats that are important to keep in mind when interpreting your results.
4. (Optional) Give an outlook on potential alternative/better ways to analyze your data in the future.

## Summary (5 points)

Summarize the main findings of your project. What did you discover, and what were the key insights that you gained from your analysis?