MA321-7-SU Team Project assignment

(27th of May 2025)

In this project you will be working in teams (see team allocation via moodle). As the project is done in teams of four or five students, it cannot be anonymous, so just give your teams's number and the names (first names, surnames plus registration numbers, emails) of each member of your team on the project report.

- **Submission of the Team Project Report via FASER by 12 noon, Friday, 20th June 2025 (Week 38).**
- **Oral presentations of the Team Project** are scheduled in **week 38 (CLAa01** Thursday 10:00 to 12:00, 19th June 2025, **CLAa02** 12:00 to 14:00 Friday, 20th June 2025).

All members of each team should participate in the analysis of the data using R, editing and writing of the submitted version of the Team Project Report and in the oral presentation (15 minutes each team). The allocation of the marks between the team members will be based on the **written statement listing the contribution of each member of the team** (if not equal, please provide percentages adding up to 100%), which must be included in the Team Project Report **and the contribution of the members of the team to the oral presentation**. If a team does not agree, each student can submit via FASER independently marks (percentages adding up to 100%) of the work done on their team project by each member of the team (including themselves) by 20th June 2025 (Week 38).

The teams analyse data identified by the R code and the data sets provided via moodle *MA321-7-SU Applied Statistics* (*'Assessment folder'*; sub-folder '*Team project Deadline 20th June 2025 at noon and Presentations on 19th June (CLAa01) and 20th June (CLAa02)'*:

```
team-project-task-MA321-7-2025.R
gene-expression-invasive-vs-noninvasive-cancer.csv
teamsubsets.csv
```

Four research questions of the Team Project Assignment to be analysed and answered:

I) Consider *supervised dimension reduction/supervised feature selection* of the 500 observed gene expression variables (features) in your data set. Use as label the variable `class` with `class==2` '*invasive cancer*' and `class==1` '*non-invasive cancer'*.

II) Use *supervised learning models/classification* to predict the variable `class` with `class==2` '*invasive cancer*' and `class==1` '*non-invasive cancer'* of future patients. Apply LDA, QDA, Random Forest and SVM. Discuss how and why you choose specific hyper parameters of a supervised learning model. You may add one or two further supervised learning methods to the investigation. Use resampling techniques as repeated 10-fold cross validation, jack-knife or bootstrap to compare the machine learning models applied without and with *supervised dimension reduction/supervised feature selection*. Discuss why resampling is better than sample splitting. Suggest and justify your 'best' machine learning model.

III) Use *unsupervised learning models/clustering* to investigate clusters/groups of genes and clusters/groups of patients. Apply k-means clustering and hierarchical clustering. You may add one further method. Discuss the stability/variability of the unsupervised learning results [Note: Make sure that you do not include the information provided by the label variable `class` for your unsupervised analysis and data narrative.]

IV) Investigate if clusters established under III) improve your 'best' machine learning model. Use resampling techniques as repeated 10-fold cross validation, jack-knife or bootstrap to compare the machine learning models applied. Suggest and justify your updated 'best' machine learning model.

General rules and hints:

- Plan and structure your work. Structure your report, for example: Page 1: cover page (title, your team number and names …, date, …). Page 2: abstract, contents, word count and the contribution of each member of the team. Pages 3-14: introduction; preliminary analysis; analysis of the research question I, research question II, research question III and research question IV; discussion; conclusion; references. Page 15 onwards: Appendix: R-code with explanations, etc..
- Use R. Put all R code, which was necessary for your report in an Appendix and explain your R code (add comments within the R code). Do not include R code of an analysis which is not used for your report. Make sure, that YOU wrote the R code (the use of some R code, without citing the source, can be viewed as ***plagiarism***).
- The report can have a length of 2000 to 4000 words (without cover page and Appendix). Not more than 12 pages without counting the cover page, title/abstract page, and Appendix. *More than 4000 words* **or** *more than 12 pages (without counting the cover page, title/abstract page and the Appendix) will reduce the marking. In cases where layout or formatting causes the report to go slightly beyond 12 pages, this will not result in a marking penalty as long as the 4000-word limit is respected, and the content is concise.*
- Use point size 12, Times New Roman; line spacing 1.5.
- Do not use more than **10 figures** and **10 tables** within the main text. You may include further figures and tables into the Appendix, if necessary. Add legends to the text of your report under each figure or table which provide a title and explain the figure or table in sufficient detail. Do not plot the title as part of the figure.

In addition, your report should include a clear account of any assumptions made in the analysis of the data.

Submit two files via FASER by 20th June 2025. Noon:

1. Report of the data narrative consisting of text explaining the results (data narrative), tables and figures with legends. **File format of the report .pdf.**

2. R code with comments explaining each line of R code used for the submitted data narrative. **File format of the R code .txt or .R file.**

Hint: If you use the **package** `rmarkdown`, you can save the Rmd file, which has txt format. Use Latex or MS Word to write the report document and to create the pdf format of the document.

Marking scheme:
Preliminary analysis (e.g. focus on description of base line data):
0 of 12: is missing.
8 of 12: narrative describing a preliminary analysis, which was suggested in the lectures and classes; at least one table showing means/sd/… and one figure showing boxplots/… are given and explained.
12 of 12: narrative describing a preliminary analysis, which includes justification of assumptions, provides further tables or figures which support the argument of the report, ...
Tasks 1 to 4 (each):
0 of 12: is missing or makes no sense.
8 of 12: narrative describing a main analysis, which was suggested in the lectures and classes; tables and/or figures should support the results.
12 of 12: narrative describing a main analysis, which includes justification of assumptions, provides further tables or figures which support the argument of the report, ...
Discussion / conclusion:
0 of 10: is missing or makes no sense.
5 of 10: discussion and conclusion summarises the data analysis and results of the study.
10 of 10: additional overall excellent quality of the report.
Oral presentation:
0 of 30: Oral presentation not well prepared and not well delivered.
20 of 30: Oral presentation well prepared and coherent presentation.
30 of 30: Excellent presentation.