# MA335 Final Project

## Due in: 12pm(noon) Wednesday 21st June 2023, week 38

Submit a copy via Faser

**Task:** Suppose you work as a data science consultant and you are asked to analyse a dataset which includes various characteristics of Alzheimer. The aim is to investigate the relationship between those characteristics and the diagnosis, i.e., Alzheimer (Demented) or not (Nondemented).

You should use R in order to conduct your statistical analysis. The project_data.csv file stores the data set used for this assignment. You should include the code as part of the Appendix of your report which should run without errors. When answering the questions you should explain the methods used and justify your answers. You should submit your report in pdf format. Zipped folders e.g. .zip or 7z will not be accepted. In order to analyse the dataset complete the following tasks (Before the tasks, convert M/F into numeric values, remove rows with Group = "Converted" and missing values):

1. Analyse using descriptive statistics (both graphical and numerical representations) on the dataset project_data.csv. Generate an appropriate table as summary and appropriate graphs, e.g., boxplots, histograms and scatterplots. **[20 marks]**
2. Implement clustering algorithms, demonstrate the results and comment on that. **[30 marks]**
3. Fit a logistic regression model using the remaining variables to predict variable `Group`. Describe the produced model and comment on what it demonstrates. **[20 marks]**
4. Implement a feature selection method to find the most important features, demonstrate your results and discuss on your findings. **[30 marks]**

The dataset includes the following characteristic variables:

| Variables | Description |
|---|---|
| Group | Group of the diagnosis (Nondemented, Demented, Other) |
| M/F | Gender |
| Age | Age |
| EDUC | Year of education |
| SES | Socioeconomic Status (1-5, 1-low, 5-high) |
| MMSE | Mini mental state examination |
| CDR | Clinical dementia rating |

| | |
|---|---|
| eTIV | Estimated total intracranial volume |
| nWBV | Normalize whole brain volume |
| ASF | Atlas scaling factor |

**General report guidelines:**

- Plan and structure your work. Structure your report, for example: Page 1: cover page (title, your name, date, . . . ). Page 2: abstract, contents and word count. Pages 3-7: introduction; preliminary analysis; analysis; discussion; conclusion; references. Page 8-10: appendix: R-code with explanations, etc..
- Use R. Put all R code, which was necessary for your report in an appendix and explain your code (add comments within the code). Do not include 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).
- Use an appropriate word processor (MS Word, Open office, ...) or type setter (Lyx, Latex,...).
- The report can have a maximum of 1800 words (without appendix). More than 1800 words or more than 6 pages (without counting the cover page and the appendix) will reduce the marking.
- Use point size 12, Times New Roman; line spacing 1.5.
- Do not use more than 5 figures and 3 tables within the main text. You may include further figures and tables into the appendix, if necessary.
- In addition, your report should include a clear account of any assumptions made in the analysis of the data.

Additionally, marks will be awarded as follows:

Report guide lines:

0 of 10: student did not follow the guide lines.

5 of 10: student followed the guide lines; but understanding of specific parts of the guidelines/report structure is weak; e.g. no table legends, citation style inappropriate, etc.

10 of 10: student followed the guide lines.