Statistical Learning 2025 – Graded Assignment 2

Part A: Supervised learning

You will be working with the same data as in Part B of Assignment 1.

The data for this part of the assignment are from a large epidemiological study on the course of depressive and anxiety disorders among adults living in the Netherlands. The data are from a subsample of 1500 subjects, who at the start of the study were suffering from an anxiety and/or depressive disorder and were diagnosed as such. Twelve months after the start of the study, the severity of each subject's depressive symptoms was assessed again (variable dep_sev_fu; short for depression severity at follow-up). The goal of the analyses is to predict the severity of depressive symptoms after twelve months, using the characteristics assessed at the start of the study.

The dataset is available on Brightspace as 'MHpredict.csv'. It contains 20 potential predictor variables, which were assessed at the start of the study, and are described in Appendix I.

1. Select **one** supervised learning method from those that were discussed in **weeks 7 and 8** for analyzing this dataset. The chosen method must be different from the ones in Assignment 1! Justify why you would select this method for this specific prediction problem. (Use max. 200-250 words.)
2. Now apply this method you selected to the dataset. Beforehand, randomly split the dataset into a training (n=1000) and test (n=500) dataset. Use your student number to set the seed of the random number generator.
3. Justify your choice of the main model-fitting parameters. Thus, make a well-informed choice for a fixed value of each parameter and/or use cross-validation to set their values. Your answer should reflect understanding of what each parameter does. (Use max. 200-250 words.)
4. Provide an interpretation of the resulting model:
    a. Describe which variables are most important in predicting the value of the outcome variable, and which measure(s) you used to determine their relative importance.
    b. Describe the effect of the most important variables (e.g., describe the shape and direction of the effect on the predictions and/or provide and discuss plots of the variables' effects). (Use max. 150-200 words)
    c. Assess the predictive accuracy of the model using the test set.
5. Now create a Table **comparing the two models in your Assignment 1 and the new model fitted in Assignment 2**. How do the predictive accuracies compare? Which model predicts best? (Use max. 200 words.)

6. Based on 3-5: Provide a short overall conclusion regarding which predictors are related to the outcome and which model you would recommend and why (Use max. 200 words.)
7. A psychologist has seen David Edgar Pression for an intake today. The psychologist wonders whether they should refer David to an intensive depression treatment program. The results of David's intake assessment are provided on Brightspace, in the file 'Patient.csv'.

   You can read it into R as follows: pat_dat <- read.table("Patient.csv", sep = ",", header = TRUE, stringsAsFactors = TRUE)

   Hint: Check whether factor variables are correctly coded. To assign a variable in pat_dat the same factor levels as a variable in MH_dat (use of function levels has a different effect):

   pat_dat$fact <- factor(pat_dat$fact, levels = levels(MH_data$fact))

   The psychologist asks you to provide them with an estimate of the severity of David's depressive symptoms in 12 months. Patients with predicted depressive symptom severity equal to or greater than 17 are referred to the intensive treatment program. What is your estimate? Should David be referred to the intensive treatment program? Bonus: Using a suitable approach, quantify the uncertainty of your estimate (not specifically taught during the lectures). (Use max. 100 words, max. 200 words including bonus.)

Appendix I

Variable name	Explanation / values
disType	Type of disorder (depressive disorder, anxiety disorder, or comorbid disorder (i.e., having a diagnosis for both types of disorder))
Sexe	Male or female
Age	Age in years
Aedu	Years of education completed
IDS	Testscore on the Inventory of Depressive Symptomatology
BAI	Testscore on Beck's Anxiety Inventory
FQ	Total score on the Fear Questionnaire
LCImax	Percentage of time in which symptoms of anxiety and/or depressive disorders were present during the past four years
Pedigree	Presence of a first-degree relative with an anxiety and/or depressive disorder
Alcohol	Alcohol disorder diagnosis
bTypeDep	Subtype of depression
bSocPhob	Diagnosis of social phobia
bGAD	Diagnosis of generalized anxiety disorder
bPanic	Diagnosis of panic disorder
bAgo	Diagnosis of agoraphobia
AO	Age at onset of the disorder
RemDis	Whether the anxiety and/or depressive disorder is currently in remission
Sample	Whether subject is a patient in specialized mental health care, a patient in primary care, or not currently receiving (mental) healthcare
ADuse	Whether subject uses anti-depressant medication
PsychTreat	Whether subject receives psychological treatment for the disorder(s)

Part B. Unsupervised learning

For this part of the assignment, please use the attached dataset (*data.US.csv* or *data.US.txt*). The dataset consists of 1000 individuals who have been measured on 30 personality variables (facets). The variables are from V2 to V31: anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability, warmth, gregariousness, assertiveness, activity, excitement-seeking, positive emotions, fantasy, aesthetics, feelings, ideas, actions, values, trust, straightforwardness, altruism, compliance, modesty, tendermindedness, competence, order, dutifulness, achievement striving, self-discipline, deliberation. The variables already have been standardised.

Our aim is to identify groups of individuals with similar personalities. However, some of the variables are highly correlated so some dimension reduction may be needed.
Analyse the dataset and answer the following questions:

B1. Do you think that reducing the variables to a smaller set of (new/derived) variables may be a good idea? Which statistical technique can be used to achieve such a dimension reduction? Explain why this technique is appropriate.

B2. How many new/derived variables should be computed to capture the most important part of the information in the original variables? Use at least four different methods to select the number of new/derived variables. Explain each method and thoroughly justify your final decision (e.g., by providing relevant figures).

B3. Can you give a meaning to these new/derived variables?

B4. Can you somehow quantify the degree to which the new/derived variables capture the information in the original variables?

Is it possible to group the study participants in terms of their personalities? To this end, use the new/derived variables (and not the original ones).

B5. Which statistical technique(s) can be used to group the participants? Explain why it is/they are appropriate.

B6. How many groups are there? How did you determine this? Thoroughly justify your answer (e.g., figures). If you identified more than one technique in the previous question, choose two and compare the techniques and the obtained results.

B7. How large is each group?

B8. If you identified more than one technique, what are the main differences between your chosen methods?

B9. What are the main differences between the groups in terms of personality (give a substantive interpretation of the obtained clustering)?

Guidelines for the report:

Produce two separate documents: A textual report that answers all questions as well as an R (.R or .Rmd file) or Python (.py file) script.

Do not refer to the script in your report. Your answers must be self-contained.

Your report should be aimed at a broad audience of researchers who might be interested in these analyses. Assume they have some knowledge of statistics. Write in full sentences. Do not use code language (e.g., variable names like “dep_sev_fu” are somewhat arbitrary and have no meaning outside of the code).

In the report, clearly divide the answers to each of the numbered assignments above (e.g., use section numbering or numbered headings).

In the script, clearly divide the code into one section for each assignment. Make sure that your code is readable (meaningful variable names, comments etc.).

Upload both the report and the script via Brightspace.

The deadline is 20:00 on 15 May 2025.

Good luck!