Statistical Learning (1st semester 2022-2023): Assignment 2

The data for this part of the assignment are from a large epidemiological study on depressive and anxiety disorders among adults living in the Netherlands. The data are from a subsample of 1152 subjects, who were suffering from an anxiety and/or depressive disorder at the start of the study, 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 that were assessed at the start of the study.

The dataset is available on Brighspace as 'MHpredict.csv'. It contains 20 potential predictor variables, which were assessed at the start of the study, and are described in Appendix I. You can read it into R as follows:

```
MH_data <- read.table("MHpredict.csv", sep = ",", header = TRUE)
```

1. Select three supervised learning methods from those discussed in lectures 9 through 12 to analyze this dataset. Thoroughly justify why you chose each of these methods for this specific prediction problem.
2. Apply the three methods you selected to the dataset. But first, randomly split the dataset into a training ($n = 1000$) and test ($n = 152$) dataset. Use your student number to set the seed of the random number generator. Motivate your choice of the main model-fitting / hyper parameters. Thus, make a well-informed choice and/or use cross-validation to set their values (even if you choose to use default settings).
3. Assess and compare the predictive accuracy of each of the models using the test set. Which model predicts best? Bonus: Using a suitable approach, compute confidence intervals for the pairwise differences in predictive performance (not taught during the lectures).
4. Provide an interpretation of each of the resulting models (note that you only need to do this for the models for which this is possible):
    - Describe which variables are relatively (un)important for the predicted value of the outcome variable.
    - Describe the effect of the most important variables (e.g., describe the shape and direction of the effect on the outcome and/or provide and discuss plots of the variables' effects).
5. Based on 3 and 4: What is your overall conclusion regarding which predictors are related to the outcome?

6. A psychologist has seen David Sthymia, a 63-year old man, for an intake today. The psychologist wonders whether he should refer David to the intensive treatment program. The results of David's intake assessment are provided on Brightspace, in the file 'DSthymia.csv'. You can read it into R as follows:

   ```
   DS_data <- read.table("DSthymia.csv", sep = ",", header = TRUE)
   ```

   The psychologist asks you to provide them with an estimate of the severity of David's depressive symptoms in 12 months time. Patients with predicted depressive symptom severity equal to or greater than 17 are referred to the intensive treatment program. What would your estimate be? Should David be referred to the intensive treatment program?

Guidelines for the report

- Produce two separate documents: A textual report that answers all questions as well as an R (.R 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 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 can be found on Brightspace.

Good luck!

Appendix I

Variable name	Explanation / values
disType	Type of disorder (depressive disorder, anxiety disorder, or comorbid disorder (that is, 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)