Information about project

Data Mining and Machine Learning I

Deadline: 3 July 2020

Drug use data

In order to help with costs of hospital admissions and other healthcare costs, it is important to understand possible factors that can lead to various drug addictions. You will be provided with a dataset relating to drug use and various features recorded on the patients. The data is presented in the file **student_XXXXXXXdata.csv** which can be found in the Project datasets folder in the Online submission of coursework section of Moodle. Here the XXXXXX should be replaced by your student ID number. So each of you will have your own individual dataset.

The datasets contain information on 600 patients from the original dataset (these will be different for all students). The class or group label variable for drug use is: "Never Used" (a value of 0), or "Used at some point" (a value of 1). This can be found in the 'Class' column of the dataset. For each respondent, 11 features are known. The features are summarised below:

1. F1 age;
2. F2 education;
3. F3 country of origin;
4. F4 Ethnicity;
5. F5 nscore;
6. F6 escore;
7. F7 oscore;
8. F8 ascore;
9. F9 cscore;
10. F10 impulsivity;
11. F11 sensation seeing.

If you would like more information about these, read the 'Database' section of the reference paper also in the moodle project folder.

Project goal

The project aim is to fit a variety of classification algorithms that you have been taught during weeks 3 to 5 to your dataset and decide which is the best at predicting the use of drugs by future patients.

Project assessment

The project is assessed on a report and corresponding R code submitted. This will be worth 40% of the overall grade for the course. On Moodle you should upload two files. One should refer to your report (preferably a pdf document) and the other one should be an R script that allows us to reproduce all the statistical analysis you refer to within your report. The deadline for submission is **Friday July 3rd 2020 at 23:00 BST**. Your assessment report must be uploaded on Moodle at the following link:

https://moodle.gla.ac.uk/mod/assign/view.php?id=1531324. Please note that ONLY pdf format documents will be accepted.

Report and R code - tasks

After you create the training, validation and test data sets, you have to:

- Perform exploratory analysis on the training data set.
- Apply all classification techniques that you have learned during weeks 3 to 5 (e.g. k-nearest neighbours, tree-based methods and support vector machines).
- Create appropriate graphs or summaries that communicate the results from these methods.
- Comment on these and interpret the results.
- Compare the results to choose a final, "best" classification model (with justification given for the choice) and comment on this model's overall classification rate, sensitivity and specificity for future sample predictions.
- Create a 10 page report that includes all the previous information. (This refers to the maximum number of pages, and does not include a title page if you wish to have one or the R code)
- Suggested report structure: Introduction, Exploratory Analysis, Results, Discussion sections

The R code should allow us to:

- install (and load) any packages you might have used,
- clearly show which set of variables you were working on (since we will not be able to reproduce this),
- reproduce any models you have worked on and
- recreate any of the graphs and summaries you have on your report.

The R code should be well enough commented that the code is understandable to one who hasn't written it.