Probability and Statisitcal Inference
Continuous Assessment Assignment 1

Due date: Sunday, 12th December 2021, 23:59
Marks: 60% of module total

Overview

The assignment is to perform a detailed analysis, with the use of R, of the **Student Performance Data Set** and to report on it. Information about the data set can be found at the UCI Machine Learning Repository (link). Further information about the data set can be found in a paper by Cortes and Silva (2008), who supplied it (link). The data set is provided (file *dataset.csv*) in the assignment specification file bundle.

Submission

The analysis should be reported in the form of an R Markdown Website. A template for this is in *caa1_site_template*. To start working on the template, open project *rmd_ca1a_site_template.Rproj* in RStudio. To submit, create a zip archive from the directory *caa1_site_template* and upload at the Brightspace assignment page created for this purpose.

Requirements

Format requirements

Your submission must include:

- the entire R Markdown Website project (not only the html files), with runnable R code in the R markdown (.Rmd) files, including required library calls on packages
- your information in the table in file *index.Rmd*
- your signature at the foot of the declaration of own work (the easiest way to do this is to replace the file *signature.png* in the template)
- a table of contents with hyperlinks to all the pages of your website
- citations and references for any material or ideas not your own but used in the report, including the paper by Cortes and Silva (2008) that describes the data set

The submission must comply with:

- the APA style for reporting statistics (a guide can be found here)
- either the APA or Harvard style of citing and referencing (but must be consistent)

Otherwise, you may organise your website in any way you wish, bearing in mind that this is an aspect of the submission that will be graded in as much as it affects user experience and readability.

Standard requirements

In your report, consisting of R code and text organised as a website, you must demonstrate:

- a systematic and in-depth understanding of concepts, methods and methodology

and the ability to:

- follow a standard methodology
- select correct methods in different contexts, justify your choices, discuss difficulties and limitations
- correctly apply selected methods
- interpret and critically evaluate outcomes of method application
- draw conclusions from outcomes and synthesise their domain-level implications

through

- clear and accurate written communication, following prescribed rules

with all of the above

- at NFQ level 9 standard (see NFQ level indicators here)

Content requirements

Your work for this assignment should span the entire data analysis methodology to the extent and depth covered in the module, as long as it is applicable to the data set.

Your analysis should incorporate:

- statement of research question(s)

  This may have to be an iterative exercise if your first research question does not provide enough scope for sufficient analysis 'bulk' (see exploration and predictive modelling requirements). Usually sub-questions arise that naturally increase the amount of analysis required. However, if this does not turn out to be the case, you may have to state another question to repeat the analysis process, thus expanding the amount of material included in the submission.

- collection and preparation

  While you will not be collecting the data, study and discuss how this was carried out, sampling concerns (validity, reliability etc.) and preparation steps, within the constraints of available information.

- description

  This will include data type classification (from different aspects e.g. discrete/continuous or interval/ratio scale), characterisation with appropriate measures (e.g. central tendency, variability), visualisation in graph and tabular form.

- exploration

  In this section hypotheses must be stated regarding relationships in the data. At least 5 different relationship measures must be used. More than five may be used if required

to answer the research question. The use of the tests must be fully justified and any qualification testing (e.g. for normality of a distribution) performed.

- predictive modelling

  One linear regression model and one logistic regression model must be fitted and evaluated.

How the different types of requirement interact

Note that the three types of requirement (format, standard and content) are mutually orthogonal and hence simultaneously applicable. Each content section needs to be treated with adherence to the required standards and formatting.

For example, what the quality requirements stipulate is that it is not sufficient to 'dump' a graph with no introduction or explanation into the *description* section. The visualisation type needs to be introduced in terms of purpose and utility, its use in the particular context justified, legends provided, a commentary using the right language included to accompany it and any new information or conclusions drawn from it stated.

Assessment schedule

- **Week 10:** you will be invited to submit any work you have completed and will receive feedback within the week
- **Week 12:** submission of assignments
- **Week 13:** ten-minute one-on-one meetings where the lecturer will ask questions about your submission

Marking scheme

research qestion(s) and hypotheses	10%
collection and preparation	5%
description	15%
exploration	20%
prediction	30%
overall presentation quality	10%
overall quality of R code	10%

Tips for successfully completing this assignment

- Work continually - many iterations may be needed for the final product to take shape.
- Take advantage of the opportunity to get feedback in week 10.
- When in doubt about anything, ask!