Project groups - dataset registration

To ensure you (as a group) think carefully about the requirements for the dataset and for teaching staff to catch issues early, we ask that you register your group's dataset on DTU Learn as early as possible.

You do not necessarily have to answer all questions in detail at this point (it will be part of the report).

Question 1

Group number [1-500]

490 - Lenka

Question 2

Dataset URL

https://gist.github.com/slopp

Question 3

Number of observations in total [integer]:

344

Question 4

Number of observation without missing values (consider if missing values are going to be a problem) :

333

Question 5

Number of variables in total:

8

Question 6

Number of attributes of type "Nominal":

3

Question 7

Number of attributes of type "Interval":

1

Question 8

Number of attributes of type "Ratio":

4

Question 9

Explain what your data is about, i.e., what is the overall problem of interest?

The dataset contains three penguins' species and their anatomical information.

Question 10

Summarize previous analysis of the data. (i.e., go through one or two of the original source papers and determine what they did to the data and summarize their results).

The data were used for linear regression and classification to find the relationship between the values for each species. Since only few values are missing, the results were accurate.

Question 11

Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques?

Our main goal for this project is to accomplish further analysis of the dataset and find anatomical relations between the species. After finishing the task we would like to have deeper understanding of how to implement linear regression and classification in Python.

Question 12

Explain which attribute you wish to predict in the regression based on which other attributes? Notice that the attribute you want to predict should typically be Interval or Ratio (or in som cases Ordinal) attribute. The attributes from which you intend to predict the class label are typically Interval, Ordinal or Ratio. If your identified variables are different from indicated here, you probably need to consider if a tranformation can be applied to make the specific attributes approiate for a regression task.

Using linear regression we will predict bill depth for one of the species using the interval attributes (bill length, flipper length and body mass).

Question 13

Which class label will you predict based on which other attributes in the classification task? Notice that a class label will typically be associated with a Nominal attribute. The attributes from which you intend to predict the class label are typically Interval, Ordinal or Ratio. If you identified attributes are different from indicated here, you may need to consider if a tranformation can be applied to make the attributes approiate for a classificaiton task.

Our main goal for classification is to estimate the missing sex values based on all other attributes except the year and most likely island, as these two attributes does not affect the sex.

Question 14

If you need to transform the data in order to carry out the regression or classificaiton tasks, explain roughly how you plan to do this

We are not tranforming the data in this assignment.