DIRECTION: Big Data and Analytics

COURSE: Big Data Programming and Infrastructure: Python and New budget

Individual Work Acad. Years 2022-2023

1. In general

This assignment will help you apply the knowledge you have acquired during the semester to a practical machine learning problem using the tools of the Python language.

In this specific assignment, you will deal with the area of   analysis of a set of clinical data.

In this task you are asked to confirm one of the methods already applied as well as apply a method to derive results which is not contained in the literature given.

2. Data

2.1. Data Sources

The program you will develop will accept as input the data file found on the website:

https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

The file contains a set of variables affecting the health of heart failure patients and a target variable (survival or non-survival).

The analysis of the results has been carried out in the publication which can be downloaded from here:
https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

2.2. Data analysis and processing

1. In the first stage of the work you will need to assess the quality of the ensemble data and may be required to clean/fill your data and confirm the correctness of the tables

2.3 of the publication. You may provide graphs of mean values and deviations that you think will be more complete than the descriptive elements in table 3 of the publication for the quantitative variables and some corresponding description for the categorical variables of the problem (table 2).

2. Then exclude the recheck time and choose two of the binary classification methods used by the authors and use the outcome evaluation metrics used by the authors to compare your results with those of the authors (table 4). For categorical features of your data set you can try not to include them or use one of the quantization methods (eg one-hot-encoding). Compare your results with those of the authors and explain what the discrepancies you find might be due to. Experiment with the methods of normalization/standardization of your data, hyperparameter optimization mentioned in the tutorial (authors use GridSearch) and/or different number of iterations and/or different method of cross validation and splitting the data set

3. The authors try to reduce the number of features they use with different methods finally calculating some values to rank them according to their importance (table 7). In this assignment you will use an alternative approach. As the goal is the final prediction of a patient's survival, you will use the PCA method and find the ranking of the variables (coefficients) that will result from the PCA (in descending order of importance).

4. The authors conclude that the first 2 variables in table 7 can be used to improve the results over the rest due to the noise they introduce. In your assignment you will create a scree plot of the cumulative variability explained by the PCA components as a function of the number of each component. Normalizing[1]components you can build a model of the percentage of variability that each component explains in your data set. By creating the above diagram find the value that creates an elbow which expresses the point after which the components contribute less variability to your data. Determine if the elbow exists and at what point and note the number of components (is it greater or less than 2?

5. Implement the three methods used by the authors and summarized in table 9 (you can use your own parameters and results validation method) using the first two components of PCA. Repeat for the number of components corresponding to the elbow point you calculated above. Are your results comparable to those of the publication? What are the possible reasons for deviation?

[1] https://towardsdatascience.com/a-step-by-step-implementation-of-principal-component-analysis-5520cc6cd598

3. Deliverables

Your submission should contain the following files

1. **File or Code Files.**The code you submit should be executed taking as input only the keywords provided by the user. Your code will create the appropriate bases. The code should automatically perform all the steps you implement and produce any diagrams included in your report and presentation. The code should be adequately commented.

2. **Presentation.**The slides you will present for the assignment. You can use as many slides as you wish or any other form of presentation, however the available time will have an upper limit of 10 minutes (you can use less than that) and 5 minutes will be allocated for 2-3 questions. In your presentation you should include your main results, techniques you used to optimize time - results, limitations and anything else you wish in the above time limit.

3. **Report.**Your report will have**maximum**size**10 A4 pages**with the arrangement that works best for you and Calibri body text type/size 11. Your report should contain at least the following parts

   a. Cover page
   b. Table of contents
   c. Description of the problem and available data
   d. A brief description of the system you developed and the techniques you chose for each stage as well as a brief description of the main parts of the code and in particular parts where you had the main contribution (eg parts of code that you developed). Description of the use case you selected
   e. Results : Results of each stage, graphs
   f. Conclusions: Overall commentary – evaluation of results of your work, limitations, optimizations you made, possible extensions.

Abs are not counted in the total of pages. Tables and graphs that will not be contained in your presentation (deliverable 2) can be included in appendices which also do not count towards the total pages.
The assignment files should be submitted as a compressed (.zip) file exclusively in electronic format through eclass (Assignments section) by the date specified there.

Good luck