Guidelines for Project

This project is a full analysis of a (moderately) large dataset using the data preparation tools and ML methods that you have learnt throughout the course to show your understanding of the materials covered in this course.

Generally, you are expected to perform analysis using appropriate methods on **ONE** data set, which can be selected based on your interests from the data sources given at the end of this guideline. Your data analysis is somewhat open-ended. At a minimum, you should have

- **Get familiar with the dataset** by describing and visualizing data of interest
- **ML techniques** used for at least two of the purposes: prediction, classification and clustering.

.

You are required to propose **your own project objective(s) and study/research questions** that are worthwhile based on the background of the data.

Write up a report

Summarize all your analyses into a short report, **at most 4 pages**, including

1. Introduction (1-3 paragraphs) – You are expected to present the main objective(s) and study question(s) of your project, some idea why it is interesting and provide relevant background information (based on the dataset used).
   A citation of the original data source should also be included. Additional relevant citations for background are encouraged.
2. Methods (2-3 pages) – This section allows readers to fully understand what you did and why you choose ML methods. Specifically, it should include
   a. Data preparation: extracting data of interest in the project, indicating sample size and the names and types of each variable in the study, identifying outcome variable. If variables were excluded from the analysis at this stage (e.g., due to irrelevance to the study questions), then it should be explained and justified.
   b. A description of any preliminary exploratory analyses, providing descriptive (or summary) statistics of variables and visualization of relationships between variables in your extracted dataset. Do not include large numbers of exploratory graphs in the main report. Necessary graphs (e.g., correlation matrix or scatterplot matrix) may be included. You should other graphs in the Appendix.
   c. Methodology: a description of the machine learning techniques that you used, including different models/methods considered, how you set training and test data, evaluation the performance of each model/method and the final models you choose. Based on the chosen models, summarize analysis results into tables and/or visualizing them by graphs, interpret results, and highlight results that have special relevance to your research question(s).

3. Conclusion & discussion (1-3 paragraphs) – Brief summary of your findings and how your objective achieves, limitations of methods used and your suggestion for possible improvement.

Besides, you may include references if you have cited to support any factual statements made in your report that are not directly based on the data or your own analysis, and an appendix including key R code and relevant computer outputs used in your report. These do not count towards the page limit. **Keep your report short and to the point**. **A good graph is worth a thousand words**.

Team member

Each team can have **at most 6 members** who may be from different tutorial/lab groups. You are encouraged to work in groups. If any student really would like to work alone, please let me know the reason by email.

Evaluation

The project report will be graded on **clarity** of your project purpose or study question(s), **appropriateness** of data preparation, **correctness** of the use of ML methods and data analysis, as well as the **clarity and reasonableness** of the conclusions that you reach.

This project is worth **20 marks** out of your overall course grade. I expect all members will be active in their group work. Ordinarily, team members will be given the same grade for the project - though I reserve the right to assign different grades to different team members in case of egregiously unequal contributions.

Each group must do the project independently. If any two reports look almost the same or same in most discussion parts, both groups will be penalized.

Project report submission

A softcopy (pdf file) of the Project Report should be submitted under Assignments on NTULearn by one of the group members by **deadline Sun 18 April, 11:59pm,** and should include names of all group members and the contribution of each at the top of the first page.

Data sources

The data can be found in the following

1. R package "mlbench" including 38 datasets. See details at https://cran.r-project.org/web/packages/mlbench/mlbench.pdf

Note: you should ensure the datasets having at least 9 variables. You are not allowed to use datasets “BreastCancer” and “BostonHousing” that have addressed in tutorial/lecture notes.

2. The UCI Repository Of Machine Learning Databases at http://www.ics.uci.edu/~mlearn/MLRepository.html, or Statlib at http://lib.stat.cmu.edu/datasets/