BU510.650 – Data Analytics, Spring 2019

Project Description

The objective of this project is to use techniques taught in class to study the data patterns or to predict heart disease for the Cleveland dataset. Refer to the Appendix for the reference and description of the data set.

There are two deliverables for the group project:

(1) Delivering a 10-minute presentation (preferably using slides) in class on Class 7
(2) Submitting project material (your report, presentation, data, and R script) on the link titled "project submission" (inside the "project" located on the left-hand side menu of your Blackboard) at the end of the day when class 7 occurs

About the report:
The report should be at most 3 pages long (excluding the appendix, which is not subject to a page limit). In your report, please follow the outline below:

(1) Data Exploration (6 pts): Conduct a deep exploratory analysis for the data. What relationships seem to be interesting?
(2) Questions/Hypotheses (7 pts): From the data exploration phase, what relationships are you planning to further clarify with modeling? What other questions or hypotheses you want to understand? Note that 4 out of the 7 points will be awarded for originality in your hypothesis (e.g., using outcomes different from diagnosis of heart disease, or analyzing polynomial associations). Make sure that your hypotheses can be tested with the methodologies covered in class.
(3) Methodologies (10 pts): Write a complete, clear description of the analysis you performed. This should be sufficient for someone else to write an R program to reproduce your results. It will also be helpful to people who read your code later. *This section should tie your computations to your questions / hypotheses, indicating the type of analysis you are conducting to answer each question (Please make this explicit in your report – e.g., hypothesis 1 will be explored using model linear regression).* Make sure you use at least two of the approaches covered in class for supervised learning and one approach for unsupervised learning.
(4) Results and Conclusion (7 pts): Discuss your results. Focus in particular on the results that are most interesting, surprising, or important. Discuss the consequences or implications. Interpret the results: if the answers are unexpected, then see whether you can find an explanation for them, such as an external factor that your analysis did not account for.
(5) Appendix: Put plots, tables, technical details or other results in appendix if necessary. This part is optional, but including visual support (for example, plots and tables) is highly recommended.

About the presentation:
Each group should select one or multiple team members to present their projects in class. Each presentation should be no longer than 10 minutes. It is encouraged to use slides (e.g., Powerpoint). The slide deck should summarize the main points of your project, including motivation, research questions,

and results. During or after each presentation, there will be a question and answer session. Each member of the presenting group, not only the presenters, can answer the questions or give comments.

Grading criteria

The project will be graded on a curve. The average grade in points will be matched to the center of a normal distribution with mean 78/100. If all the groups get the same grade, then the grade for each group will be 78. Copying the work of other groups (or internet) might not help you to get a higher than average grade.

References

http://archive.ics.uci.edu/ml/datasets/heart+disease

https://rpubs.com/mbbrigitte/heartdisease

Appendix

The dataset can be found in the following link:

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

The description of the columns as they appear in the dataset is as follows:

1. age: age in years
   2. sex: sex (1 = male; 0 = female)
   3. cp: chest pain type
   -- Value 1: typical angina
   -- Value 2: atypical angina
   -- Value 3: non-anginal pain
   -- Value 4: asymptomatic
   4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
   5. chol: serum cholesterol in mg/dl
   6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
   7. Restecg: resting electrocardiographic results
   -- Value 0: normal
   -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
   8. thalach: maximum heart rate achieved
   9. exang: exercise induced angina (1 = yes; 0 = no)
   10. oldpeak: ST depression induced by exercise relative to rest
   11. slope: the slope of the peak exercise ST segment
   -- Value 1: upsloping
   -- Value 2: flat
   -- Value 3: downsloping
   12. ca: number of major vessels (0-3) colored by flouroscopy
   13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
   14. num: diagnosis of heart disease (angiographic disease status)
   -- Value 0: < 50% diameter narrowing
   -- Value 1: > 50% diameter narrowing
   (in any major vessel)