**DATS 7860 Statistical and Machine Learning for Big Data**

**Instructions for the final project**

**Summary**

The goal of the course final project is to let you study and evaluate some statistical learning methods through simulation studies and/or real-world datasets. It will serve as the final exam of this course and will contain the following two components: an oral presentation by you as an video file (10% of the final grade) and a written report (15% of the final grade)

My expectation is that you should: (1) read a few (1-3) applied (or even methodological, for PhD students) papers on the statistical learning methods that you are interested in to gain an in-depth understanding of them, (2) check the software implementations of them and read the reference pages of the software tools to understand how to apply them, (3) apply and evaluate those methods in simulated datasets and/or real-world datasets, (4) summarize your results and what you have learned in a formal written report, and (5) present what you have learned and your results.

**Requirements – please read carefully!**

**1.** You are required to apply at least 3 different statistical learning methods, and at least one method should be developed on or after the year 2000.

**Note 1:** Among the methods covered in the textbook (ISL), random forests and BART were developed after 2000 and thus meet the above requirements (Please see “**History\_of\_Methods.docx**” file for the years when each method was developed). Other methods were developed before 2000, so the original versions of them as presented in the textbook do NOT meet the above requirements. But some extended methods based on them were developed after 2000 (e.g. elastic net, group lasso, extensions on neural networks, etc.), which meet this requirement. If you are not sure about whether the methods that you want to compare meet this requirement, you are welcome to discuss with me.

**Note 2:** You are encouraged to learn and apply some new methods not covered in the textbook. If you apply them correctly, you will receive 2-3 bonus points, but your grade of the course paper will not exceed 25% of the total grade.

**2.** You can compare and evaluate the methods you choose through either simulation studies or real-world datasets, or both of them. The real-world dataset can be from any field, not limited to the biomedical fields. If you purely use simulations without real-world datasets, please try to do a relative comprehensive and thorough comparisons and evaluations. For instance, you may need to simulate data using several different models (e.g. both linear and non-linear models) and also try several different model parameter settings.

**3.** You need to give an oral presentation summarizing the methods you compared and your results: Please make some slides (in PDF or PowerPoint format) and record your presentation as a video file and submit them via D2L. There is no limit on the length of the presentation, but the suggested length is between 15-30 minutes. Please see the file “**Tips\_for\_presentation.pdf**” for more suggestions.

**3.** You need to write a report for this final project. Your written report is expected to contain the following parts:

- A review of the methods you are comparing: please clearly cite the methodological papers where the methods were developed and which software packages you are using. You can use mathematical notations, formulas and tables to describe the methods.

- If simulations are used, please describe the simulation settings thoroughly and clearly. If a real-world dataset is used, please provide a description of background of the data and the variables (predictors) in the dataset.

- Summary of your results: please use tables and figures to summarize your results.

Your written report should follow the format and standards of a scientific paper and the academic conduct regulations of the university. Plagiarism is strictly prohibited and subjected to university policies.

I uploaded the presentations from previous classes (see “**Examples\_previous\_years**” folder), the reports from previous classes (““**Previous\_reports\_students**” folder”) and also several papers comparing methods published in academic journals (see “**Example\_papers\_published**” folder) on D2L. They are just for your information. You don’t need to follow the exact format of previous years and your written report is not expected to achieve the quality of those papers published in academic journals (Just try your best will be OK!).

**What to submit (please submit to corresponding folders via D2L)**

1. **The video and slides of your oral presentation:** it will be shared to the whole class and possibly the class in the future.

2. **The written report for the final project.**

**Late submission and extension**

I understand this course is intense, so if you are not able to complete the final project before the due date (end of this semester), please contact me ([yshi@augusta.edu](mailto:yshi@augusta.edu)) in advance and I can extend the due date. **Please note:** if you request an extension, you will be assigned an incomplete grade (“I”) temporarily, and it will be changed to a letter grade until you submit your presentation and written report and have them graded by me.