MATH 6333: Statistical Learning/ MATH 8333 Adv. Statistical Learning
Final Project Instructions
Fall 2022
Due on: 9th Dec 2022

The premise of this project is very simple. You are to pick a real data set for which you believe there are interesting questions to answer. You will then try out all the different statistical learning approaches that we have covered in this course to try to find the best way to answer these questions. The project is expected to be completed individually.

Deliverables: For the Section MATH 6333-90L this will be an individual project. For the sections MATH 6333-01 and MATH 8333-01R this will be a group project.
Each individual/group needs to determine their project topics and datasets for the analysis by the 5th Nov 2022.

Analysis can be done using any statistical software (preferably R, SAS, SPSS or Minitab), Each student needs to submit a written report by the last week of classes.

The report should consist of the following sub sections.

1. **Introduction:** Here, give a brief introduction to the problem that you are working on. Any background information that is connected to the current problem can be included.
2. **Literature Review:** Discuss about previous work and their findings related to the problem that you are working on.
3. **Methods:** Discuss about any statistical data mining methods that you used. Present the underlying methodologies, assumptions, etc
4. **Results:** Include descriptive statistical summaries which are important. Afterwards, discuss about any important findings that you discovered using advanced statistical methods that you implemented. Make sure to label figures and tables and to add a caption detailing what it explain. Figure titles and corresponding captions should be listed on the bottom of the figure. Table titles and captions should be on the top of each table.
5. **Discussion:** Include a brief discussion on what you found, its importance in general and any limitations in your study.
6. **Appendix:** Include any R codes, tables and figures that might be important.

The final report might contain at least 5 pages excluding the appendix. Try to be concise and only to present important findings (Although, you may have performed many different analyses, you won't need to add them unless you feel it adds something to the study findings). Also, please **include a title page including your name and the title of the project**. **Please add the page numbers. (If you are submitting your report as R markdown file, then it would enough to include the name and the title of the project, at the heading)**

Following are some data repositories that you might need to explore, to find a suitable dataset for the project.

Data Repositories

1. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
2. KDD Nugets: http://www.kdnuggets.com/datasets/
3. StatLib: http://lib.stat.cmu.edu
4. DASL: https://dasl.datadescription.com/datafiles/
5. TwitteR: http://cran.r-project.org/web/packages/twitteR/index.html
6. rfigshare: http://figshare.com, http://cran.r-project.org/web/packages/rfigshare/index.html
7. Kaggel: www.kaggle.com
8. Miscellaneous Datasets: http://users.stat.ufl.edu/%7Ewinner/datasets.html
9. US Census Data: https://www.census.gov/data/data-tools.html
10. Open Gov. Data: www.data.gov, www.data.gov.uk, www.data.gov.fr, http://opengovernmentdata.org/data/catalogues/

Suggestions

1. I highly recommend starting to work on your project R code as soon as you get my approval on your dataset. If your proposal is ready before the deadline, please feel free to send it to me for approval.

2. Since you will be applying the methods we learned in this class on your datasets, your assignments' R code should be very helpful!