Final Project

The goal of this final project is to apply data mining and statistical learning techniques (covered in class) to some real data. Feel free to use any appropriate data et. You can get your dataset from your own work or other data sources such as one of the following online data libraries:

- UCI Machine Learning Repository
  `http://www.ics.uci.edu/~mlearn/MLRepository.html`
- UCI KDD Archive
  `http://kdd.ics.uci.edu/`
- Datasets for Data Mining at KDnuggets:
  `http://www.kdnuggets.com/datasets/`
- KDD Cup Datasets from 1997 to 2008:
  `http://www.sigkdd.org/kddcup/index.php`
- StatLib – `http://lib.stat.cmu.edu/`
- Data Analysis and Data Mining: An Introduction by Azzalini and Scarpa
  `http://azzalini.stat.unipd.it/Book-DM/data.html`
- Kaggle – `https://www.kaggle.com/datasets`

You are encouraged to collaborate on ideas but please work on the project independently – feel free to discuss your ideas with the instructor and/or your classmates. You may need to find appropriate references to better understand the scientific problem involved in the project. The results should be due to yourself though, e.g., if several of you work on the same data set, each of you needs to produce his or her own results even if the original ideas were developed jointly. Remember to always give credit where credit is due.

1 Deliverables

1.1 Recorded Presentation

You are expected to prepare and record a 10-minute video presentation and upload it to Blackboard. In the presentation, you should briefly introduce the data source, backgrounds of the data, clearly specify the scientific question(s), explain your analytic goals and statistical methods employed and then summarize the results and present interesting findings.

1.2 Project Report

Your project report should contain the following components:

- **The executive summary**

  The executive summary should briefly describe the source and background of your data and summarize the questions you addressed and your key results (e.g., explained in the context of your specific application). Also mention any caveats. Summarize your recommendations on how these results can or cannot be used. This should be roughly 10% of the length of the full report.

- **Main Report**

  The main part of your final report should be concise and contain only 3-10 pages. It should cover the following points:

  1. What problems you specifically addressed, including details in scientifical, technical, business, etc., terms
  2. The techniques you used and the procedure you followed.
  3. Interesting results. Include the results along with a practical interpretation. Describe possible scientific implications.
  4. Conclusions & References.

- **Appendix: R Codes, Procedures, and Detailed results**

  Give appendices that contain R codes/output and information that would allow someone else to reproduce your analyses (assuming a reasonable knowledge of the tools used, e.g., someone else in the class) and actual printouts of the results, annotated so that if someone did re-run your analysis, they would know how to get from the raw results to conclusions.

Please submit via Blackboard. You may use either R Markdown or plain R for this project.

2 Grading

Grading will be based on (in order of importance):

1. The Procedure You Followed: Is it correct (given the techniques you used), did you describe it well? This includes aspects such as data selection, preprocessing, exploratory data analysis (EDA) for understanding your data, etc.

2. Techniques Used: Given the scientific questions you chose to address, is it a supervised or unsupervised learning problem? Did you approach it in appropriate ways? Did you justify it?

3. Interpretation of Results: Did you correctly understand and interpret the raw results you obtained?

4. Quality of Presentation and Write-up: Did you present what you did in an understandable and usable manner?

The above questions are key factors for assessing your proficiency in data mining. The following questions will be of interest, but will have a lower impact on your final score:

- Subject or scientific questions addressed.
- Quality of results: Since this is largely an artifact of the data and your initial selection of problems, lack of interesting results won't have a major effect on your score.