STA 551 Term Project Guidelines

General Requirements

The idea is to perform an end-to-end data science project on a realistic task. To this end, we set up a web page with some real-world data sets. You can also use a different data set if you wish which should have enough information for both statistical models and machine learning algorithms.

- ***Requirements of data set***
  - *Feature Variables*: The data set should have at least 10 feature variables that include both numerical and categorical variables.
  - *Sample Size*: The size of the data should be at least 10 times the number of variables in the data set.
  - *Response/Outcome*: The response variable must be categorical.

- **Requirements of Models, Algorithms, and Methods**
  - *Statistical Model(s)*: Binary logistic regression model as a predictive model.
  - *ML Algorithms*: at least one classification algorithm should be used in the analysis.
  - *Methods*: Training-testing, cross-validation, feature selection and extraction, performance evaluation, etc.

Project Proposal

Explore the data set you selected for your project and formulate the analytics problems you want to address. The proposal should outline:

1. A high-level statement of the problems you intend to address.
2. The data source(s) you intended to use and a description of the data (i.e., variables definition and size, data challenges such missing values, etc.).
3. The goals of your analysis.
4. A description of the data analysis methods and tools (i.e., software and programming language) you plan to use.
5. The outcomes of the project ideally include visualizations.

Project Report

Your project report is the formal description of your project. The format should be similar to statistical journal articles. The re[ort should have 6-10 pages in length and contains sections on:

- *Problem Statement and Background*
  - Give a clear and complete statement of the problem. Where does the data come from, what are its characteristics?

- Include an informal description of quality and performance measures (e.g. accuracy on cross-validated data) that you planned to use.
- Include background material as appropriate: who cares about this problem, what impact it has, what implications better solutions might have.

- *Methods*
  - Describe the methods you explored (usually algorithms, or data cleaning or feature selection, or extraction).
  - Justify your methods in terms of the problem statement.
  - Describe the methods suitable but used in the analysis.
  - Detail every method you tried.
  - Detail evaluation quality and performance metrics you will use.
  - Outline the process of searching for the final model.

- *Analytical Tools*
  - Describe the tools such as programming languages, software tools that you used, and the reasons for their choice.
  - Platforms and tools are needed when the final model is deployed.

  Tools will probably include those used for machine learning, and possibly data wrangling and visualization.

- *Results and Discussion*

Give a detailed summary of the results of your work. Here is where you specify the exact performance measures you used.

- Pick only important outputs (figures and tables) that support your argument. Please do not simply copy the part of the table from the output and paste it into the report. The table should be like those in journal articles.
- Interpret the tables and figures that you included in the report.
- Provide accuracy or quality measures that justify the models and algorithms you identified to address the research questions.
- There may also be a performance such as a runtime measure.

Please use visualizations whenever possible. Include links to interactive visualizations if you built them.

- *Appendix*

  This is where you put your extra outputs and computer code.