Final Project Instructions
STOR 565 Machine Learning, Spring 2021

Project Overview: The overall goal of the project is to apply supervised and unsupervised statistical learning methods to a dataset of your choice to answer a research question. The formulation of this question, which methods you will use, and how you interpret the results, are entirely up to you and your team. You will additionally be expected to research and implement a supervised learning technique not covered in the computing assignments, write a report on this technique, and use this technique in the analysis of your data.

Groups: You will work in groups of 3 and may choose your team members. Please email the instructor and instructional assistant with your group members as soon as you have decided. Those that do not form a group by 5:00 pm on Tuesday 4/13 will be randomly assigned to groups. However, the instructor reserves the right to make changes to the groups if necessary.

Project Components: Your group should submit the following to the instructors via email:

1. *Project Proposal:* No more than 1 page in length. This is expected to be a brief of your project plan, describing your choice of data in detail, and outlining your research question(s) and learning methods (both old and new, both supervised and unsupervised). In addition to explicitly stating each of these points of interest, this proposal should include a summary of your data that discusses its origin, who collected it, why it was collected, and what predictors are available. If you cannot explain what a given predictor measures then you probably should not use it in your analysis. You should also include a top-level explanation of whatever new supervised learning method you have chosen. Please also include the names of each person in the group.

2. *Presentation:* At the end of the semester, groups will be asked to give a presentation on their work. Presentations will be given live and are expected to be 10-15 minutes in length. We ask that each group use a slideshow format and avoid overloading slides with text. Presentations should be rehearsed beforehand to ensure they last for the proper amount of time. Each member of your group should speak for an approximately equal amount. You should expect to answer some questions from the instructor, instructional assistant and other students after your presentation. Grading will be based on clarity, thoroughness, the degree to which contributions seem equal, and ability to answer questions accurately and concisely. A copy of the slideshow presented by the group should be sent to the instructor and instructional assistant by the deadline listed in the table below.

3. *Final Report:* 10-12 pages in length (including graphics) written in R markdown. This should be structured as follows:

    - Introduction: Give a high-level overview of what your report is about. Describe what dataset you chose and what research question(s) you investigated.

- Data: Should include a detailed discussion of your data and your predictors. Please also elaborate upon why your team chose this data. You should also include a careful account of any data cleaning that was necessary, and any pre-processing you did. You are also encouraged to provide some exploratory analysis of the data in this section.
- Learning Methods: Should outline which supervised and unsupervised methods you used and discuss why they are relevant to your project aims. This section should include a very detailed report of the supervised learning technique not covered in the computing assignments. You should be able to discuss specifically what this new method does, how it does it and how it can be used for your classification or regression objectives.
- Results and Discussion: Should include detailed, informative visuals with a discussion of how you got them. You should also interpret the visuals you created in the context of your research question.
- Conclusion: Based on your analysis, what is your conclusion?

Project Component Grading and Due Dates: Your group will receive a single grade for the project depending on the quality and timeliness of each component. The percentage of the grade for each component as well as its due date are given in the table below.

Component	**Grade**	**Tentative Due Date**
Project Proposal	10%	4/20
Presentation	40%	4/29
Final Report	50%	5/5

Data: In this project, your group is free to choose what dataset to work with. If you do not have datasets in mind, you may want to browse the UCI Machine Learning Repository[1], which contains a wide variety of benchmark datasets commonly used to evaluate machine learning methods. If you have questions or concerns about the datasets your considering, please feel free to discuss with the instructional assistant at office hours or over email.

New Learning Method: In this project, your group is also free to choose what new machine learning method to employ. There are a number of interesting techniques for both unsupervised and supervised learning available to try. We suggest considering *SVM's*, *decision trees*, or *naive Bayes*. Groups considering larger datasets might also consider using a *neural network*. Details on most of these methods may be found in the recommended texts. If you are in need of additional references, please consult the instructor and/or the instructional assistant.

[1] https://archive.ics.uci.edu/ml/datasets.php