INFO 1998: Final Project

TIMELINE:
Project Guideline Released: 4/10
Form a group (2-3) and submit a proposal: 4/17
Sign-up for a proposal meeting time-slot on CMS: 4/17
Proposal meetings will happen between: 4/18 - 4/23
Final Project due: 5/1

OVERVIEW
The final project will consist of three components: a **project proposal,** a **final coded project**, and a **written report**. The project is expected to be an original concept that uses the models and visualization tools you have worked with all semester. Teams are permitted to use any dataset publicly available and are encouraged to use a dataset they find interesting. The project is intentionally left open-ended so that you can perform data science on a topic in which you are interested in. Teams may consist of groups of 2- 3 people and should follow the guidelines below. We expect that this project will take some time to finish so we encourage you to start the project early and reach out to course staff if you need help.

DATASET
We would like you to have an interest in the research you are doing, so you may use any dataset you want for this project. There are resources available to help you find datasets such as Google's dataset search engine (https://toolbox.google.com/datasetsearch) and Kaggle website (https://www.kaggle.com/datasets). In addition, if you would like any suggestions on datasets to use, don't hesitate to ask. All our TAs have worked on similar projects before and have several datasets in mind that could be good for the project.

There are two submissions: a project proposal (due **April 17th at midnight**) and a final report with code (due **May 1st at midnight**).

PROJECT PROPOSAL (10/110 points)
To make sure that you start on the right track, we're requiring that you write a proposal, provide your dataset, and schedule a short meeting with a TA. All of these will be will be due on CMS. You should form your groups before submitting.
The proposal document should include the following two sections:

- *The Project Objective*: What will be your target variable(s)? How might your results be important/impactful?
- *The Dataset*: If your dataset is available online, you can provide a link in your proposal document and don't need to submit anything for the "Dataset" file on CMS. Otherwise, please submit your raw dataset under "Dataset" in the CMS assignment. If your dataset is above the file size limit on CMS, please cut out some rows of the dataset to make it smaller (i.e. randomly sample the data).

- *Priors & Plans*: What kind of relationships and patterns are you expecting, based on logic or preliminary analysis? How have you started (or how will you start) analyzing the dataset? Do you have any intuition as to what models might be most effective?
- *Division of Labor*: How are you planning on dividing the work? Do you have a timeline to complete the project?

The proposal should be no more than 300 words. The proposal and meeting are relatively casual -- you should have a thoughtful plan, but we won't criticize your grammar or ask you to give a formal presentation. The grade for the proposal and meeting combined will account for 10 points.
Later on, you do not have to follow your proposal exactly; however, if you decide to change your dataset or project objective, you should let a TA know.

FINAL REPORT (100/110 points)
The final report will be a Jupyter Notebook (and, optionally, a separate paper), including visualizations, models, and written descriptions of the process.

Preprocessing and Visualizations (30/110 points)
Most likely, you will need to preprocess your data before analyzing it, visualizing it, and using it to train a model. You are required to preprocess and manipulate your data as needed (if you didn't have to process your data, please state the reasons in the write-up), and to create two meaningful visualizations of your data. These can be any of the visualizations we have covered (see lecture 3 if you need a refresher). The two visualizations cannot be the same type; i.e., if your first graph is a histogram, your second graph must be something different. Please remember to label all relevant components: title, axes, include a key if applicable. It must be clear what purpose your visualizations serve, why they are meaningful, and what insights can be gained from them. The plot of accuracy scores against k values in the lecture 6 demo for kNN is a good example.

- 12 points for each visualization (half off for duplicate graph type)
- 6 points for preprocessing

Models (40/110 points)
You can choose any two (or more) models covered in the course to gain different meaningful insights in the dataset you chose. You may use the same target variable for both models or you may use different ones. You must use two different types of models. While metrics may vary between datasets, be sure that you are able to justify why and how the model you chose solves some particular problem with regards to the dataset in your write-up, with a small, high-level explanation of how the model works. Remember our discussion of baselines when examining accuracy.

- 15 points for each model (half off for duplicate model type)

Write-up (30/110 points)

The written portion of this report (i.e., excluding the code) should be at least 500 - 750 words and should supplement your visualizations and models. We think that the best way to do this is to write text in markdown cells throughout your Jupyter Notebook, but writing a separate paper is fine. Some ideas on what to write about: any data processing done, significance and insights from visualizations, how you chose model parameters, why you chose specific models or cross validation procedures, etc. Your write-up will be graded based on how much it helps readers understand your code/process and interpret the results.

AWARD

The top 3 teams will be chosen and given the following awards:
1st: Google Home Mini
2nd: Amazon Gift Card (tentative)
3rd: Starbucks Gift Card (tentative)