BME 580 Course Project

**Guidance for Project Proposal** (due 11:59 PM ET, 2/23/2023):

One person should submit the pdfs for each of the deliverables electronically on Sakai, following the file naming convention of the homeworks: *date\_lastname\_firstname\_[group/proposal /finalReport/presentation].pdf.* For example, my assignment would be named *20230223\_Dunn\_Jessilyn\_proposal.pdf* (**5 points will be deducted for files named incorrectly** and an **additional 3 points will be deducted for files formatted incorrectly** (e.g., a word document as opposed to a pdf)). You only need to use the name of the group member who is submitting on behalf of the team.

The goal of the initial project proposal is to check that the proposed idea fits these two criteria. Be creative about the problem you want to address; the project should be something that engages you!

**Page limit:** There is a 3-page limit to the proposal *including* figures and tables but *excluding* references. The document should be single spaced, 11+ point font, 1 inch margins, Arial or Times New Roman. Any figures should have no smaller than 9pt font when printed out. Make sure to include page numbers at the bottom right of the pages. Be succinct, but convey all of the necessary information.

**Figures and Tables:** All figures should have a minimum font size of 9 (including axis titles and tick labels) and should be high resolution. Every figure and table should have a caption, as you would see in a published paper. Colors and symbols should be used appropriately as needed to convey your message, but not used without purpose. Be sure to avoid color schemes that are not colorblind-friendly. Remove figure backgrounds and gridlines except where absolutely needed to make your point.

**Grading Breakdown:**

* 10%: Proposal Clarity.

o 5%: Is the proposal succinct but complete? Is it clearly written *without jargon* such that a student not in this class, but familiar broadly with biomedical data types, probability/statistics, and overarching machine learning and data science principles could understand it?

* For graduate students, if necessary (for example, if writing in English poses some challenges for your team), have you run the proposal by the Duke GCIP team to ensure that it is written clearly and understandably?

o 5%: Are tables and figures readable and easily understandable by a general audience? Are they placed and referenced appropriately in the text? Note that all figures should be called out the main text, and should be numbered in the order in which they appear in the text. You should have a title page with your project title, group member names, the date, assignment type, and link to your Github repository. All pages except for the title page should have numbers in the bottom right.

* 20%: Problem Description o 10%: What is the problem you are addressing? Be sure to do a sufficient literature review to substantiate your claims and clearly articulate the gap that your work aims to address. Make sure that you cite your references appropriately and that you are thorough in building your case from this literature. If the literature on your exact question is limited, are there other related problems with an existing literature? I expect to see somewhere between 5-10 citations for your problem setup. o 5%: What is the exact question you’re asking of the data? i.e.

what is your hypothesis? How does the literature support this?

* + 2.5%: Why is this an interesting problem? o 2.5%: What is the potential impact of solving this problem? If you could solve this problem, how would it change current health or medicine practice?

* 15%: Data Description o 2.5%: What data are you using for this project? o 5%: Where is the data coming from? What is the source of the data (database, organization, measurement modality(ies) that generated the data)
  + 2.5%: How does the data help answer the question(s) that you posed? o 5%: Describe the data (dimensions, number and type of fields, units of each data type). Are variables categorical, numerical, ordinal, etc? Are there any concerns you may have at the outset about the data?

* 30% Exploratory Data Analysis on your training dataset (you may choose 60-80% of the total data for this analysis) (refer to the *Figures and Tables* section above) o 15%: Perform visualization of your data and present the figure(s) that best tell your story about the data that you are working with. What are the distributions of the variables you are most interested in (e.g., histograms, boxplots, violin plots, etc)? What do you notice about them? Are there clear groupings (e.g. PCA, Clustering)? Do you see any early signs of relationships in the data (e.g., scatterplot matrix; correlation matrix)?
  + 15%: Overall, what does your data look like? Are there any obvious trends you see from a first look at your data using the EDA tools we learned in class? Describe what you see and what it means to you for how you will approach the next steps in your project. For example, will you need to perform any preprocessing, handle outliers or missingness, etc.? o Include in this section the link to your Github repository where you will be uploading your code and data for this project. Please make sure that both

Dr. Dunn and your teaching assistant can access your repository. To test access, you may want to check with a classmate who is not a part of your group to be sure they can access your repo.

* 20%: Tentative Predictive Model Building Approach o 10%: What is your planned approach? Note that this just has to be a rational approach. You will not be held to this, but it should represent your planned first pass. We still have a lot of ground to cover on the details on approaches in the upcoming classes, but you should be able to distinguish whether you will use supervised or unsupervised learning, or both, and whether you will approach your question as a classification or regression problem, or try both. o 10%: Why is this approach feasible? This should be your **best guess** as to why this is feasible. Clearly you won’t really know until you actually do the problem.

o How will work be broken down amongst your team members?

* 5%: References.

o 2.5%: Cite reliable references (e.g. statistics from a government-sourced document or a clinical journal) – **be sure you demonstrate sufficient evidence to support the claims in your proposal**.

* Do not cite my course notes or Wikipedia.
* If you decide to use ChatGPT for any aspects of the writing, please cite and mention the sections where it was used. It is ok to use the text generated by ChatGPT as a launching point for your writing, but **do not copy/paste** and do make sure you **verify and correct** any mistakes.

o 2.5%: Include a link to your Github repository that you cite somewhere in your document. Make sure you have populated your README page and have begun some data preprocessing and EDA methods that you have committed to your repo.

**Planned Feedback:**

Our goal is to determine (i) whether your problem is interesting and demanding enough to be a good problem for the course project, and (ii) whether your problem and the proposed approach is feasible. We don’t want anyone to get stuck on an infeasible project or attempt to accomplish too much (or too little) in a relatively short project period. Note that there is a big difference between an infeasible project and negative results! **You will not be penalized for negative results.**