BST 263 Final Project

Overview

The final project will revolve around data analysis and the thoughtful application of the methods learned in the course to real datasets. Working together in groups, students will select a real dataset of their choice and introduce scientific objectives they wish to pursue using the data. They will then propose appropriate statistical learning techniques to accomplish those objectives, and will submit a 1-page outline for feedback from the teaching staff. They will implement the analyses and will report, interpret, and visualize their results. The write-up of the analysis and results will be presented in the format of a scientific paper, with length no more than 8 pages, single-spaced (includes tables and figures, excludes references). Students will give a 10 min presentation about their project to the class.

Timeline

April 13	Group formation due (as part of HW5)
April 20	1-page outline due
May 9	Final write-up of the project due
May 9, 11	In-class presentations

Selecting a dataset

Please be creative when selecting a dataset and scientific question to address! The dataset does not need to directly measure health outcomes but the project should have biomedical or public-health relevance. You are welcome to use a dataset associated with a group member's research. It is recommended to choose a dataset that all group members can access. However, we will allow teams to base their project on a dataset that only one (or a subset) of the group members can access (due to privacy restrictions). If using a privacy-protected dataset, the group must seek the data owner's permission to share results of the analysis with the class.

Don't know of a dataset to use? Check out these free, open access datasets relevant to public health research (you are also free to use others not listed here):

- MIMIC ICU data from Beth Israel Deaconess Medical Center: https://mimic.physionet.org/
    - Note: Need to have completed CITI training to access
- eICU Collaborative Research Database: https://www.physionet.org/content/eicu-crd/2.0/
    - Note: Need to have completed CITI training to access
- Other open access health data on physionet: https://www.physionet.org/about/database/
- CDC's national environmental public health tracking data: https://www.cdc.gov/nceh/tracking/Topics.htm
- COVID-19 data: https://www.nature.com/collections/ebaiehhfhg/
- 1,000 Genomes Project: https://www.internationalgenome.org/data

- US census data available through the tidycensus R package: https://walker-data.com/tidycensus/articles/basic-usage.html
- Twitter data: https://developer.twitter.com/en/solutions/academic-research/resources

Dividing responsibilities

Your group can divide the responsibilities among the members in any way you like. One option is to have each group member contribute to many/all components of the project. However, in some cases that system may not be feasible or efficient (e.g., if data are privacy-protected and can only be accessed by one group member). Alternatively, you can assign each person a single component or a small number of components to work on—for example, one person could be responsible for the analyses, one person for the visualization of results, one person for the writing, one person for the presentation, etc… At the end of the project write-up, you will need to specify the component(s) of the project to which each group member contributed.

Descriptions of each project component/deliverable

Outline

The outline should be approx. 1 page long and should contain the following information:

- 1 short paragraph of background about the topic of your project
  - Why is it important?
  - What are the knowledge gaps that need to be filled?
- 1 short paragraph describing the dataset you plan to use
  - Who will have access to the raw data and when?
  - What is the sample size?
  - What variables will you be using?
- Clearly state the scientific objective(s) of your project
  - This is the most important part!
  - What product or scientific insights will you provide? For example…
    - We will build a classification model to predict Alzheimer's risk based on patients' demographic features and a set of over 100,000 genetic variants. This predictive model could be used to identify high-risk individuals and recommend them for preventative treatments.
    - We will characterize the relationship between a collection of environmental contaminant exposures and bladder cancer, investigating possible non-linear relationships and interactions between exposures and adjusting for many potential confounders. The insights gained from this analysis could be used to inform policy to regulate levels of the most dangerous pollutants.
- List the methods you will consider for your analysis
  - It is recommended to try several different methods we have learned in the class and compare them.
  - Can also consider statistical learning methods not covered in class (e.g., neural nets) or variants/combinations of methods learned in class.
- Specify how you will choose any tuning parameters and evaluate/compare model performance

Write-up

The final project write-up should be no longer than 8 single spaced pages (with 1-inch margins). The page limit includes any tables and figures, but excludes references/bibliography. You don't *need* to write 8 pages, this is just an upper limit—the more concisely you can convey the details of your analysis, the better. The write-up should be structured like an substantive paper in public health / biomedicine, with the following sections:

1. Introduction
    - Provide motivating background about the topic addressed by your project
    - Clearly state your scientific objective(s)
2. Methods
    - Describe the dataset you are using, specifically identifying the predictor variables and, if using supervised learning, the outcome variables to be employed in your analyses
    - Explain any data pre-processing, including removal of missing values, transformations, etc
    - Describe the statistical learning methods you will apply on your data, how you will select any tuning parameters, and how you will evaluate/compare model performance
    - Mention sensitivity analyses you will conduct (if any)
3. Results
    - Provide a Table 1 with summary statistics from your data
    - Present the results of your analyses, including tables and/or visualizations and model comparison statistics
    - Identify a "best model" that you think should be used for inference and/or prediction
        - Note that, in some contexts, it may be worth sacrificing a bit of predictive accuracy to gain interpretability, so you should justify your choice of the "best" model by weighing these considerations
4. Discussion
    - Summarize your study and the key takeaways
    - Discuss the impact of your findings/product
    - Discuss the fairness of your data/models/analyses
    - Discuss strength and limitations of your analyses

At the end of the paper, please include a paragraph stating each group member's contributions. For example:

- Group member W led the data curation and analyses. Group member X created visualizations, made tables, and contributed to the creation of slides for the presentation. Group member Y led the paper-writing. Group member Z contributed to the creation of slides for the presentation and delivered the in-class presentation.

Also include a bibliography/references (you can use whatever citation style you like).

<u>Presentation</u>

Each group will present a 10-minute overview in class describing their project and their results. Groups should present using slides, and the slides should follow the same structure as the write-up (Introduction, Methods, Results, Discussion). We ask all class members make every attempt to attend class these two days, and each group should be prepared to answer questions from the audience at the end of their presentation. Groups can choose to have multiple presenters or a single presenter.