

## Project Guidelines for Opportunities and Challenges of “Big Data”

A project is required for Biostat 202. The project will illustrate the steps of a data science investigation including: accessing the data, merging and cleaning the data, creating descriptions of the data with tabular and/or visualization techniques, and using machine learning algorithms to analyze the data. Individual components of the project will be due during the session (see below for due dates) as well as a final description combining all components. Individual components are graded only as 0 (not turned in on time), 1 (partially completed), 2 (completed). The final description will be graded by letter grade (A+, A, A-, B+, B, B-, C, D, F).

### Component 1: Specify the question, access and clean the dataset [2 pages or less]

*Due date*: Midnight, Sunday 8/8/2021

*Description*: Using one of the course datasets (see the CLE for the datasets and their descriptions as well as sample questions), pick a question of interest for your course project. Access the dataset you propose to use for your project.

*Details*: List the course dataset you will use, describe the source of the data, and specify the question of interest. If your focus will be on prediction (supervised learning) give the variable to be predicted. If your focus will be on finding structure in the data (unsupervised learning), briefly describe what you hope you will find. Feel free to propose questions that might require both supervised and unsupervised learning methods, if applicable. Describe the steps you went through to read in the data, including any problems you encountered. Provide a small table describing some of the key variables (mean, minimum, maximum and standard deviation for numeric variables and frequencies of categories for categorical variables).

### Component 2: Describe the data [2 pages or less]

*Due date*: Midnight, Sunday 8/15/2021

*Description*: Using both a table as well as a figure or graphic, describe the data in a way that informs the question of interest.

*Details*: Provide a table and an interpretation of the results in the table. Provide a figure or graphic and an interpretation of the results.

### Component 3: Analyze the data [1 page]

*Due date*: Midnight, Sunday 8/22/2021

*Description*: Apply two (or more) of the big data analytic algorithms to the chosen dataset to answer the question of interest.

*Details*: Describe the steps you went through to analyze the data, including any problems you encountered, why you choose that method and what steps you took to avoid overfitting (where appropriate). When quantifying prediction error,

make sure to explain how you performed training and validation and your choice of evaluation metric.

Final project: Combine components 1 through 3

*Due date:* Midnight, Thursday 9/2/2021

*Description:* Provide a brief write-up of all components on your project.

*Details:* Define the key issue, question or hypothesis you investigated. Explain the algorithms you used and why you chose it, including any problems or difficulties and how you handled them. Provide a summary of relevant results in tabular or graphical form and a discussion of what they indicate about your question of interest. Include a brief conclusion. Discuss limitations to the analysis including possible sources of bias and overfitting. **The text should be 1500 words or less and the project should have no more than 3 tables or figures.** Include a title and references but no abstract.

### Final Project Writeup Guidance and Suggestions

**Background (10 points)** – Provide some information that helps to identify why you have chosen your particular research question. This section does not have to be long, but needs to show that you have thought about why you are looking at the research question. The research question does not have to be novel or profound. You may not need references.

**Research question (5 points)** – State specifically your research question of interest, preferably in the form of: Does xxx (predictors) predict/associated with (yyy) outcome among your particular population of interest (for supervised learning)? Or Determine which xxx characteristics are associated with yyy outcome among your population of interest (for unsupervised learning).

**Variables/Predictors (5 points)** – Please identify your predictors of your model, especially any predictors that you have had to derive from the dataset. Remember to include any definitions that you have used to define derived predictors especially ones that aren't commonly used (commonly used being BMI, for example).

**Outcome** (for supervised learning algorithms) – Please list your primary outcome(s) and secondary outcome(s), if any.

**Methods (20 points)** – **includes information from component 1** – Identify whether you intend to perform a supervised or unsupervised learning algorithm and list the algorithms that you will use to perform your analysis in Orange (you should be using multiple algorithms given that it is easy to do so in Orange). Also identify any particular challenges or concerns you faced when importing and cleaning your data in Orange. If you performed the data cleaning procedures in an alternative program, please state which program you used. Identify how you will evaluate and validate your model.

**Results (30 points) – includes your table and figure from component 2**

- Comment on your “Table 1” or baseline characteristics of your population of interest, especially any particular points of interest.
- List your overall model results including results from various models (if you utilized several different models).
- State any sensitivity analysis results that you may have (e.g. testing or validation). May not be applicable for all projects.

**Discussion (30 points) – includes items from your components 2 and 3**

- Comment upon your results and which model you felt performed the best to look at your specific question of interest.
- Comment upon the performance of the model and what evaluation metrics you used to assess model performance.
- Comment upon any limitations that you can think of with respect to analysis of your data and to the interpretation of the results.
- Conclude with some comments specifically addressing whether or not your research question was adequately answered.