**EPID 8451: Final Project**

**DUE Monday, May 2nd at 5PM**

**Instructions:** For the final course project, students may choose one of the three options below. Requirements for each option are clearly listed below. For all options, students should be sure to perform and document any necessary cleaning of data, describe decisions regarding how missing data will be treated (if any) and describe any strategies for feature selection. The final project should be a full report, with both quantitative analysis and interpretation. This assignment needs to be uploaded to the course Courseworks site by Monday, May 2nd at 5PM. **Late assignments will not be accepted.**

**Formatting:** All projects must be submitted as a single .docx or pdf file that was generated by knitr in R. That document should contain a header that includes the option number the student is addressing, the name of the dataset being used, and the specific research question being addressed. All programming code and necessary output, as well as specific text answers to meet all requirements must be included. Sections must have clear sub-headings to illuminate the specific area of the pipeline or project the section addresses. You do not need to include intermediate output unless it is relevant for answering a specific question (e.g. Rather than printing out all of the results from 10-fold cross-validation, you can just show the best tune value.) Unless otherwise specified, all answers and interpretations must be written in complete sentences. Questions should not be answered within code chunks.

**Collaboration:** You may discuss this assignment with any students in this class, the TAs or the instructor. You may not consult external faculty or students. When choosing the option for a group project, each member of the group must submit their own assignment and clearly state the other members of the group in the header.

Option 1: Choose your own adventure

Students choosing this option will select their own dataset and address their specified research question within that dataset. Requirements for the dataset, research question and analysis are detailed below.

Dataset: At **minimum**, the dataset must:

* Contain at least 1000 observations and 30 features **OR** contain at least 100 features and 100 features
* Include both continuous and categorical features
* Contain features and/or outcomes relevant to public health

Research Question: At **minimum,** the question must

* Be relevant to public health
* Be specific and clearly interpreted as either predictive, explanatory or descriptive
* Be able to be addressed using a machine learning algorithm and a data-driven analytic pipeline

Analysis: At **minimum,** the analysis must

* Use at least two machine learning algorithms covered in class
* Encompass a full analytic pipeline appropriate for the specific research question (e.g. a predictive question requires a predictive pipeline with partitioning, cross-validation, tuning of hyperparameters, evaluation in a holdout dataset)
* Tune hyperparameters across a broad range of values (not just package defaults)

*Option 2: Explore the exposome and health*

Students choosing this option will use the set of exposome and health datasets provided in class. For this option, students will develop their own specific research question within one of the four categories below. Similar to above, the research question must:

* Be relevant to public health
* Be specific and clearly interpreted as either predictive, explanatory or descriptive
* Be able to be addressed using a machine learning algorithm and a data-driven analytic pipeline

The four categories of research questions that can be addressed within this option are:

1. ~~Construct a risk score for a specific outcome to use to identify individuals in need of medical/clinical/public health intervention. You must specify the intervention.~~
2. ~~Hypotheses generation for targeted epidemiologic studies. You must include a brief description of the targeted epidemiology study that you are trying to inform with your analysis.~~
3. ~~Construct a propensity score to aid in an epidemiologic analysis of an etiologic question and then perform an analysis using the propensity score to address that etiologic question (choice of weighting or matching is yours)~~
4. **Construct a prediction model to identify individuals with an outcome of interest in an unseen dataset. The rationale for the need of the prediction model must be presented.**

Similar to Option 1, the analysis must meet the following **minimum** criteria:

* Compare performance of least two machine learning algorithms covered in class
* Encompass a full analytic pipeline appropriate for the specific research question (e.g. the construction of a risk score requires typical components of a prediction pipeline but then additionally assesses calibration)
* Tune hyperparameters across a broad range of values (not just package defaults)

Option 3: Group Assignment

Students can choose to work in a group of up to 3 people. Students choosing this option can further choose Option 1 or 2. In addition to meeting all of the requirements for either Option 1 or Option 2, students choosing Option 3 must meet the following criteria:

* Use/compare at least **three machine learning algorithms** covered in class
* Construct a new variable using an unsupervised analysis; this variable can be used in the main analysis or not but a description of the utility of the new variable must be included. In addition, students must interpret the new variable in the context of its input features, as previously done in class

**All students must address the following questions/items when reporting their project**

1. Provide a brief rationale for your research question. Why is it important?
2. Interpret results in the context of your research question. Showing the quantitative analysis is not sufficient. You must interpret your results in text.
3. Describe at least one analytic limitation of your project **and** at least one ethical consideration that arises from your project.

Grading Rubric

The final project is worth 35% of your total grade. It will be graded out of 100 points. 90 points are for content and meeting all of the required criteria outlined in this document. 10 points are for clarity of presentation, including appropriate use of sub-headings and formatting of output.

1. Traditional logistic regression
2. CART (Classification Tree 🡪 single tree)
3. Random forest/bagging (ensemble methods)
4. Support Vector Machine

**NOTES**

Feature selection: variable importance from *random forest & Lasso*

*NOTE: classifier performance does NOT increase monotonically as the # of features increases*

or a variable titled 'cluster' which identifies which cluster an observation belongs to.

Unsupervised 🡪 features has to be numerical

be sure that your question is clear as to whether

the primary goal **is**

1. **to predict who has asthma** or
2. **to identify factors that predict asthma**