Introduction to Machine Learning for Epi

Assignment 2 due Tuesday January 26, 2021 by 5:29 PM

**Sydney Ng sn2863**

**Topic: Breast cancer screening**

**Predictive:**

Research Question: What baseline health characteristics can be used to predict breast cancer screening results in middle-aged women?

Yes, I would want to use data-driven methods to help answer my predictive question. When doing predictive modeling, I would be limited to working with the data that I have to identify which variables in my dataset would most efficiently and accurately predict breast cancer screening results. Hence, I would have to use data-driven methods at least a little bit! In order to do this, I would try to do some dimension-reducing technique like principle-components analysis which would knock out any extra variables not meaningful for predicting breast cancer screening results. Then I would use an unsupervised learning algorithm such as some sort of clustering to separate positive/negative breast cancer screening results because we are trying to identify the best list of baseline health characteristic variables that would separate these clusters as efficiently (minimizing MSE) and accurately as possible. This clustering algorithm is data-driven because it is splitting up the positive/negative clusters based on the dataset we feed into it.

**Explanatory:**

Research Question: Do adverse childhood experiences cause increased risk of screening positive for breast cancer in middle-aged women?

I would not want to use data-driven methods for creating my explanatory model. As Hsu and Healy (2019) explain, “unlike for prediction … the validity of causal inferences also depends on the adequacy of expert causal knowledge” such as drawing a DAG to identify causes of interest. Using a data-driven approach would also not be useful because quite a lot of research goes into breast cancer, so there is a lot known about it and we do not need data-driven methods to identify new possible contributing causes. On top of that, I am defining my exposure of interest to be adverse childhood experiences from the very start. Therefore, when addressing whether having ACEs increases the risk of having/screening for breast cancer, our research is guided by the question at hand and theory from which I identified the exposure of interest. Then, as we are all used to with our epidemiology studies, I would try to do a logistic regression (because the outcome would be positive/negative breast cancer screening result), identify confounders and interactions, and interpret the results using odds ratios and confidence intervals.