EPID 8451: Final Assignment

DUE Friday, May 1st at 5PM

Instructions: Complete all parts of this project. Within each part, you should be sure to perform any necessary cleaning of data, make decisions regarding how you will treat missing data (if any) and determine if any provided features should be excluded prior to the analysis. You should turn in a single .docx file that was generated by knitr in R. That document should contain all programming code and output, as well as specific text answers when requested. Label sections using the subheadings used in this assignment (i.e Part 1, Question 1..etc.) and clearly state which analytic option you have selected in Part 2. You may discuss this assignment with students in this class, the TAs or the instructor. However, the final work and write-up must be performed individually. Unless otherwise specified, all answers must be written in complete sentences. Questions should not be answered within code chunks. This assignment needs to be uploaded to the course Courseworks site by Friday, May 1st at 5PM. No late assignments will be accepted.

Part 1: Neighborhood environments in New York City

This part will utilize the dataset nyc_env.csv.

Researchers are interested in determining how the social and physical environments within communities influence health. The team of researchers approached community leaders to inquire which environmental or social factor they perceived was most relevant to their health and the health of their community. Rather than choose a single factor, the community leaders discussed their concerns about the presence of multiple factors within their community. They expressed a strong belief that is was the totality of factors, not a single "bad actor", that led their community to have adverse health outcomes. The researchers listen to the community's input and decide to compile data on multiple factors across communities in New York City. For every census tract in New York City, they compile information on ambient air concentrations of diesel exhaust, proximity to hazardous waste sites (NPL sites), proportion of housing stock built before 1960 as a proxy for potential lead exposures, and the proportion of the neighborhood with household income lower than the city median. They ask you, their data analyst, to run an analysis that will be able to determine if there is evidence for distinct patterns in multiple factors across communities and, what those patterns may be.

Question 1: Run an appropriate unsupervised learning analysis that will address the above research question. Your analysis should include a data-driven approach to determine the optimal number of outputs that are retained/produced by the chosen learning technique.

Question 2: Describe the outputs of the analysis in terms of their composition of the input features.

Question 3: List a subsequent research question that could be addressed using the output of this analysis. The output can be used as an exposure, outcome or confounding variable. State what type of research question it is: descriptive, explanatory or predictive.

Part 2: Choose your own supervised adventure

For this part, you will choose one of the two research questions detailed below. Each question will require a different analytic pipeline that is consistent with the goal of the research question. No matter which option you choose: you must do the following:

- Implement an appropriate data-driven analytic pipeline to address your specific question
- Tune hyperparameters across a broad range of values (not just package defaults)
- Validate, evaluate and/or interpret your final model in a way that is appropriate for your research question

A) Construct a risk score for developing a myocardial infarction using the data from the Framingham Heart Study (dataset: framingham.csv) For this question, you should try three different algorithms before selecting the best model.
B) Researchers are interested in uncovering nutritional factors that influence the development of diabetes. To generate hypotheses to guide future research, they ask you to utilize feature selection methods to identify the factors that are most relevant to fasting blood glucose. (The dataset you are given is glucose.csv.) You should try two different algorithms for feature selection and comment if they identify the same set of variables as being important for serum glucose.

Part 3: Ethical considerations of data-driven analyses in social epidemiology

The following is an excerpt from an article on the ethical tensions in using social media data to characterize individuals' and communities' mental health. After reading the brief excerpt, address both questions listed below. The response (to both questions) should be limited to one page.

"Powered by machine learning techniques, social media provides an unobtrusive lens into individual behaviors, emotions, and psychological states. Recent research has successfully employed social media data to predict mental health states of individuals, ranging from the presence and severity of mental disorders like depression to the risk of suicide. These algorithmic inferences hold great potential in supporting early detection and treatment of mental disorders and in the design of interventions. At the same time, the outcomes of this research can pose great risk…"

Question 1: Describe one potential risk to either individuals, communities or specific populations that could arise from research or public health practice that utilizes data-driven analyses of social-media data.

Question 2: Describe one potential safeguard that could be implemented to prevent the risk you describe.