

Course project

Biostat 216

The course project

- For the course project, you will analyze a dataset using a machine learning algorithm to answer some question or hypothesis.
- You can choose between:
 - Analyzing the Behavioral Risk Factor Surveillance System (BRFSS) dataset that we have provided.
 - Picking your own dataset.
- Due date: March 17th, 2021

BRFSS dataset

- BRFSS collects state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services
- The data is collected from a random sample of adults through a telephone survey.
- Data from 2014-2015 is posted in CLE: ~5000 observations, ~300 variables
- A larger dataset is available on Kaggle: <https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system>
- The full dataset can be found on the CDC website: https://www.cdc.gov/brfss/annual_data/annual_data.htm
- You decide what problem you are interested in and how to use machine learning to answer your question.

What you'll need to turn in

- A project writeup (1500 words or less)
 - Describe the study and variables in the dataset.
 - Apply one or more machine learning algorithms and present your analysis.
- Code for reproducing your analysis
 - Python and R are both acceptable

General steps for doing the project

- Decide what question you are trying to answer and which dataset you will use.
- Download the data.
- Check the data. Visualize the data. Perform any data cleaning needed.
- Choose what machine learning algorithms you will use based on your question(s) of interest (classification, regression, clustering, or dimension-reduction). Decide how you will address potential problems like over-fitting.
- Analyze the data and interpret the results.
- Discuss limitations in your analysis, such as potential bias.

Choose the algorithm based on your question of interest:

