Prediction Model on BRFSS Data

Authors: Escanord Le (dhl64), Trung Nguyen (tnn18), Max Tjen (mkt38), Tung Tran (ttt45), Audrey Zhu (aaz25)

# Abstract

In this study, we aim to use the big data processing capabilities of Hadoop to process a large data set and train a supervised deep-learning model to predict the GENHLTH variable. The GENHLTH variable, which ranks a respondent’s general health on a scale of 1-5 with 1 being excellent and 5 being poor, uses all the other non-identification variables in the data set as predictors. The data set we will use is the 2020 data set from the Behavioral Risk Factor Surveillance System, which is a collaboration between the United States Centers for Disease Control and various US state health departments with the goal of collecting annual data on various health-related behaviors and diagnoses. The 2020 BRFSS survey data, found [here](https://www.cdc.gov/brfss/annual_data/annual_2020.html), contains records for over 200,000 randomly sampled United States residents on over 250 variables. Given the size and diversity of the data collected in this survey, researchers rarely use the entire data set all at once, which represents a missed opportunity to explore potential connections between the variables that are not immediately obvious. The goal of this model will be to explore how well the data collected in the BRFSS can predict general health and which variables have the most predictive power with respect to general health.