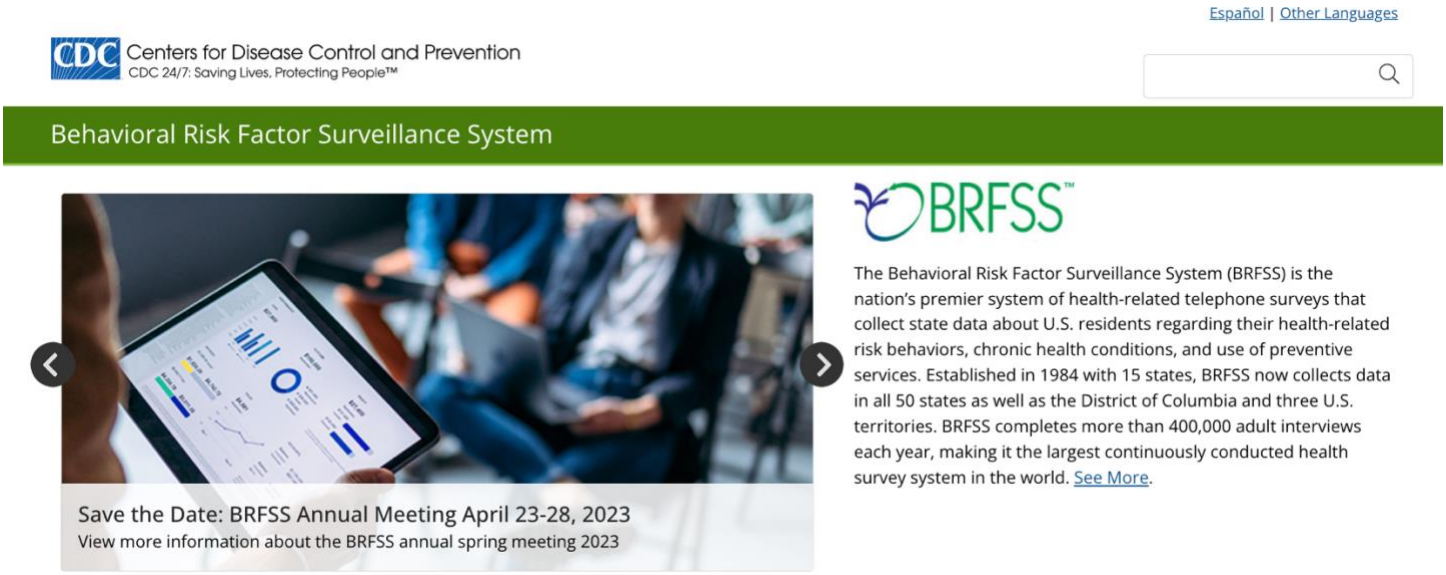
HW Assignment 6: Comprehensive Supervised Learning

CS6140: Machine Learning Spring 2023

Due Date: …, 2023

(10 points)



Scenario

BRFSS or Behavioral Risk Factor Surveillance System, is a system of ongoing health surveys conducted by the Centers for Disease Control and Prevention (CDC). BRFSS is the world's largest telephone survey, with over 400,000 adults surveyed each year across all 50 states, the District of Columbia, and U.S. territories. The primary purpose of BRFSS is to collect data on the prevalence of behavioral risk factors and chronic health conditions among *adults*. These risk factors include smoking, alcohol consumption, physical inactivity, poor diet, lack of preventive health care, and others. The data collected through BRFSS is used to identify trends in risk factors and health conditions, inform public health policies and programs, and evaluate the effectiveness of public health interventions.

BRFSS surveys are conducted via landline and cellular telephone interviews, using a standardized questionnaire that includes questions on demographics, health status, health behaviors, and access to health care. BRFSS also includes special modules or questions on specific health topics, such as oral health, diabetes, and cancer screening. The data are used by researchers, public health practitioners, and policymakers to inform and guide public health programs and policies.

BRFSS uses a complex sampling and weighting scheme to measure prevalence of many health conditions, behavioral and lifestyle related risk factors and emerging health issues in states. The weighting scheme adjust analysis results to balance sampling bias, non-response bias, under / over representation of certain stratifications and much more. Weighting also allows the user to extrapolate results to the whole population. Weights are primarily used for correct estimation of prevalence of health conditions and their risk factors. Since the primary objective of this analysis is run ML algorithms and not estimation, we will not use the weight variable for this exercise and assume the data to be a large unbiased sample collected from all over the country.

BRFSS data can be used to understand comorbidity, which refers to the presence of multiple chronic conditions in an individual. ML algorithms can be used to analyze BRFSS data and identify patterns of comorbidity among the survey respondents. One approach for understanding behavioral factors behind comorbidity is to apply classification algorithms to the BRFSS data. Such algorithms can be trained to predict whether an individual has multiple chronic conditions based on their responses to the survey questions. By analyzing the features that are most important for predicting comorbidity, one can identify risk factors and inform the development of targeted prevention and intervention strategies.

Task

This assignment uses BRFSS data from two years (2019, 2021). The choice of years was inspired by the curiosity of discovering potential differences in the pre and post covid eras. The primary objective of this analysis is to understand the data using exploratory data analysis and classification techniques. You are expected to clean, standardize, and merge data and create two new categorical variables as described later. You will then *apply various classification techniques* to model the newly created columns with the given feature space, and then report your findings.

BRFSS collects information on prevalence of several chronic conditions that can be combined to study changes over time. The primary objective of this case is to map and predict comorbidity classes. Two comorbidity response variables will be created based on the responses to questions related to chronic conditions, which will then be used for classification tasks. Two separate year specific data for BRFSS files are provided for this analysis.

Data Dictionary for BRFSS can be found here:

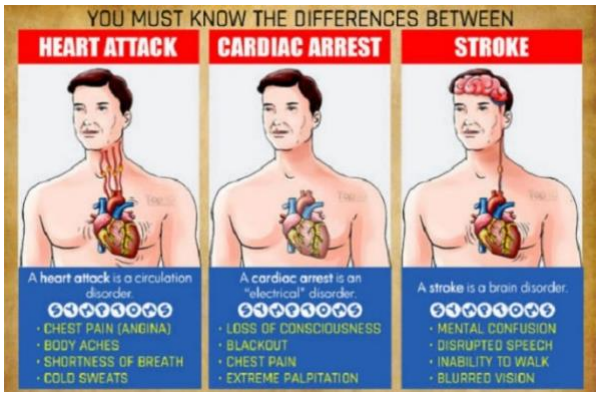
2021 codebook: https://www.cdc.gov/brfss/annual_data/2021/pdf/codebook21_llcp-v2-508.pdf

2019 codebook: https://www.cdc.gov/brfss/annual_data/2019/pdf/codebook19_llcp-v2-508.HTML

Technical Details

The chronic conditions include diabetes, asthma, heart related conditions, COPD, cancer, and depression. The following BRFSS variables need to be used for creating comorbidity variables:

Depression = (ADDEPEV)
Ever told Asthma = _CASTHM
COPD = (CHCCOPD)
Cancer = (combine CHCSCNCR and CHCOCNCR)
Ever told Heart Condition = combination of CVDCRHD4, CVDINFR4 and CVDSTRK3
Diabetes = DIABETE



Create two additional categorical columns as follows:

a) Comorbidity Variable 1 (Binary): This column indicates whether an individual has any of the six chronic conditions.
b) Comorbidity Variable 2 (Multiclass): Create a similar column as above with 0, 1, 2, 3+ categories.

Analysis Part 1:

Use both years of data for this part. Run Exploratory data analysis using crosstabs, visuals, and basic frequency distributions to understand how chronic conditions are distributed across geography and demography. Write a summary of the salient features of the healthiest and least healthy states in the country. You may also use the newly created comorbidity variables for this analysis. Discuss if you noticed any associations of the risk factors such as Age, Sex, Income, Education, Marital Status etc. with the level of comorbidities, while comparing the two years of data.

Analysis Part 2:

Use only 2021 data for this part. Use several classification algorithms (Logistic, KNN, RF, Gradient boosting, XGBoost, Catboost) to classify Comorbidity Variables 1 and 2. Write a short report describing the performance metrics. In the end, choose one model each for the classification of both categorical variables.