BIOS 735 Project Proposal

Group 2

Di Hu, Eunchong Kang, Wanting Jin, Andrew Walther, Feiming Wei

March 25, 2022

1. Introduction

According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). Many health status indicators are found related to heart disease such as high blood pressure, high cholesterol, smoking, diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare.

The dataset to be investigated comes from the 2020 CDC annual survey data and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. The dataset contains 319,795 observations in total. We choose the presence of Heart Disease as the outcome for our classification problem and 17 key indicators for health status are included as potential predictors. A brief summary about the dataset is shown as below.

a. Dataset: Personal Key Indicators of Heart Disease

b. Table 1/Data dictionary

Variable	**Description**	**Type**
HeartDisease	Respondent has had coronary heart disease or myocardial infarction	Binary (Yes/No)
BMI	Body Mass Index	Continuous
Smoking	Smoked >= 100 cigarettes in lifetime	Binary (Yes/No)
AlcoholDrinking	Respondent is a heavy drinker (men: >14 drinks per week, women: > 7 drinks per week)	Binary (Yes/No)
Stroke	Respondent has had a stroke	Binary (Yes/No)
PhysicalHealth	Days out of past 30 where physical health was not good	Continuous
MentalHealth	Days out of past 30 where mental health was not good	Continuous
DiffWalking	Respondent has difficulty walking or climbing stairs	Binary (Yes/No)

Sex	Male or Female	Factor
AgeCategory	Age of respondent	14 discrete levels
Race	AmericanIndian/Asian/Black/Hispanic/ White/Other	Factor
Diabetic	Respondent has diabetes	Binary (Yes/No)
PhysicalActivity	Respondents played any sports (running, biking, etc.) in the past month	Binary (Yes/No)
GenHealth	Excellent/Very good/ Good/Fair/Poor	Factor
SleepTime	Average sleep hours	Continuous
Asthma	Respondent has Asthma	Binary (Yes/No)
KidneyDisease	Respondent has kidney disease	Binary (Yes/No)
SkinCancer	Respondent has skin cancer	Binary (Yes/No)

2. Study Aims
    a. Aim 1 - determine relationship between heart disease prevalence and all other factors
    b. Aim 2 - determine relationship between BMI & heart disease
    c. Aim 3 - determine relationship between risky behaviors (smoking & drinking) with heart disease or stroke (poor health outcomes)
    d. Aim 4 - Compare the model fitting performance between logistic regression model and other machine learning methods. And propose predictions for likelihood of heart disease given health condition indicators
3. Methods
    For AIM 1: Logistic regression model adjusting for all the factors in Table 1.
    a. Logistic regression model
        i. Utilize BFGS algorithm or SGD to find MLE for parameters of logistic regression model
        ii. Using rcpp-armadillo for calculation
    For AIM 2-3: Test about significance of effects for covariates in the fitted logistic model
    For AIM 4: Apply Machine Learning Method for Heart Disease classification
    a. Module 3 method such as SVM or Random Forest
    b. Compares logistic regression and machine learning methods for predicting disease by using ROC and AUC

4. Analysis Plan
    a. Data transformation
        i. Binary variable: Yes (1), No(0)
        ii. Ordinal data: AgeCategory, GenHealth
        iii. Categorical data: Sex, Race
        iv. Need to check whether standardization of covariates is necessary.
        v. MentalHealth and PhysicalHealth
            1. The distribution of these variables are extremely zero inflated. One possible solution is to change this variables to a binary variable. For example, 0 days to No, >= 1 days to Yes.
    b. One possible analysis plan
        i. Split the data into the training and test set.
        ii. Fit the model with logistic regression and machine learning method by using the training set
        iii. Compare two methods by using the test set

5. References
    a. Kaggle dataset: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
    b. Any relevant literature