

Benjamin Faulknor

Project Abstract

Heart Attack Indicators | Supervisor: Ceni Babaoglu| 22-Jan-23

# Introduction

The dataset I have chosen to analyze using a classification and regression theme is the 2021 survey data from the Behavioural Risk Factor Surveillance System (BRFSS)[[1]](#footnote-1). This survey is completed annually via telephone across the United States of America by the Centers for Disease Control and Prevention (CDC). This survey contains a total of 438,693 records and 303 columns, and can be downloaded as a .XPT file which can then be loaded into R/Python. The CDC resource also contains detailed documentation on the methods that were used in completing the survey as well as the definitions of the variables within the data set and the manner in which those questions were asked.

In the United States, someone has a heart attack every 40 seconds and is the leading cause of death for people of most racial and ethnic groups in the United States[[2]](#footnote-2). My goal with this data set will be to attempt to predict whether a survey respondent has or has not ever been told they have had a heart attack/myocardial infarction (variable: CVDINFR4) and to provide an analysis on the combination of variable which provide the largest indicator to the outcome of this variable.

A challenging element to this data will be to reduce the number of dimensions, however I do believe this challenge is reasonable due to the excellent documentation resources provided. I intend to complete this through a variety of techniques including:

* Self-filtering of data that is more administrative in nature (e.g. Interview Day)
* Considering the variation in individual variables (e.g. CCLGHOUS which is either missing or Yes)
* Correlation analysis (e.g. \_AGEG5YR and \_AGE65YR) to remove highly correlated variables
* Principal Component Analysis as a potential final measure of reduction if required

I will then work towards cleaning the existing rows of data either via:

* Imputing values where missing based on a logical measure such as median or mode or where alternative columns may provide insight
* Removing rows of data – While I will need to keep in mind the potential for bias, this likely makes sense in some situations such as when CTELENM1 indicates incorrect phone number or DISPCODE indicates the survey was not completed

Another item of note I will need to complete in the early phases will be to consider the balancing of the data, as the target variable is currently 5.20% Yes (22,831) and 94.19% No (413,207).

Once the data has been reduced, cleaned and balanced I will investigate potential models for forecasting the variable CVDINFR4. I intend to consider multiple categorical algorithms e.g. Decision Trees, K Nearest Neighbors and logistic regression, however as I progress I will evaluate these models using relevant test statistics to confirm whether any algorithm is better/worse at a statistically significant level. In order to attempt to reduce the likelihood of overfitting the predictions, these algorithms will be trained on a separate training set, and when evaluated, will be evaluated against a test set that the algorithm has not utilized previously. The majority of my algorithms will likely be implemented through various Python libraries.

1. <https://www.cdc.gov/brfss/annual_data/annual_2021.html> [↑](#footnote-ref-1)
2. https://www.cdc.gov/heartdisease/facts.htm#:~:text=Every%20year%2C%20about%20805%2C000%20people,States%20have%20a%20heart%20attack.&text=Of%20these%2C,are%20a%20first%20heart%20attack&text=200%2C000%20happen%20to%20people%20who%20have%20already%20had%20a%20heart%20attack [↑](#footnote-ref-2)