Capstone Two – Project Proposal

Hio Wa Mak

04/25/2024

Problem statement

How can the Center for Disease Control (CDC) develop an algorithm to predict whether a patient had a heart attack with greater than 80% accuracy by the end of this year based on individuals' demographic characteristics, medical history, and other health and lifestyle factors.

Context

Heart disease is the leading cause of death in the United States (CDC, 2021). There is someone who has a heart attack every 40 seconds. About 805,000 people have a heart attack each year in the United States. Heart disease costs a huge burden of $239.9 billion each year. It is therefore important to identify specific demographic, medical, and health and lifestyle risk profiles of heart attack to inform more targeted prevention and interventions.

Criteria for success

Successfully identified demographic, medical, and lifestyle risk profiles by the end of this year that can predict whether an individual had a heart attack with 80% of accuracy.

Scope of solution space

Focus on specific demographic (e.g., sex, age, race), medical (e.g., comorbidities), and health and lifestyle (e.g., BMI, sleep, mental health, physical fitness) risk factors.

Constraints

The data I had included approximately 40 features that could potentially be used to predict whether individuals had a heart attack. However, I do not have individuals' physiological data (e.g., blood pressure, heart rate, health rate variability) that are important features in predicting heart attack. However, results from this study could be used in conjunction with electronic health record (EHR) in hospitals that are likely to have this data to further improve model prediction accuracy.

Stakeholders

Stakeholders include CDC and hospitals. Results will be presented to CDC and hospitals so that targeted prevention and interventions can be developed for individuals with high-risk profiles.

Data sources

The dataset was obtained from Kaggle (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data).

The dataset originally came from the CDC Behavioral Risk Factor Surveillance System (BRFSS) 2022 and was reduced from 300+ to 40 variables that are potentially of high relevance to heart attack (the procedures in converting the original dataset to the current dataset was documented by Kamil Pytlak on a GitHub repository: https://github.com/kamilpytlak/data-science-projects/blob/main/heart-disease-prediction/2022/notebooks/data_processing.ipynb). The BRFSS is an annual telephone surveys on U.S. residents regarding their health status. The survey consisted of responses from more than 400,000 adults each year.