Heart Disease Prediction with Emphasis on Factors Causing It

Milestone: Project Proposal

Group 63

Krishna Hemant

Yanming Liu

617-834-9804

617-959-5728

durgasharan.k@northeastern.edu

liu.yanmi@northeastern.edu

Percentage of Effort Contributed by student 1: 50%

Percentage of Effort Contributed by student 2: 50%

Signature of Student 1: *Krishna Hemant*

Signature of Student 2: *Yanming Liu*

Submission Date: 2/14/2022

Problem Setting

Cardiovascular health has been one of the most important aspects of human beings over centuries and science and medicine have been trying to predict, diagnose and solve problems as efficiently as possible. As technology and the amount of data generated have increased multifold, computer science started playing a role in helping doctors to cure heart related problems with a deeper understanding of it. Data science, especially data mining algorithms, has piqued interest in academia or healthcare research for detecting patterns, classifying or predicting values in various areas which could improve a person's health and also contribute data for further understanding of our complex body.

Problem Definition

The purpose of the project lies in predicting if the person has heart disease or not using a collection of features from patients and healthy volunteers. All the features can be obtained by the healthcare app, so as to avoid tedious medical examinations in hospitals. Furthermore, we aim to find the most influential factors contributing to people's disease using exploratory data analysis, which can help in preventing serious implications to the patients in the future. We aim to explore and compare different data mining techniques to further understand the data and come up with an accurate model. Lastly, we would try to do a correlation analysis between heart disease and covid with additional relevant datasets and will show if we need to take special measures on protecting patients with heart disease.

Data Sources

Kaggle is a website for data scientists and machine learning to learn, discuss, compete, and solve real problems. The heart disease dataset can be found in the following link: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Data Description

The heart disease dataset consists of 70000 records of patient data, 11 attributes and one target column. It can be categorised into 3 parts, objective, examination, and subjective. Objective attributes are basic information, like Age, Height, Weight, and Gender. Examination contains Systolic and Diastolic blood pressure, Cholesterol and Glucose levels. Subjective information refers to whether the person smokes, drinks, or actively exercises. The target variable is if the patient has heart disease or not. The dataset contains numeric and categorical variables.