Final Project Draft

Mithil Patel

2022-05-15

Introduction

One major problem that contributes to the leading cause of death in the United States every year is heart disease. One of out of every four deaths in the United States is caused by cardiovascular disease resulting in approximately 659,000 death yearly. While the CDC and public health organization have spent hundreds of millions of dollars to budget for reducing heart disease related deaths, it has still lead to be a highlighted concern for the past several years. However, data science can provide in-depth analysis to solve many of the issues we have and be a potential solution to the heart disease issue around the world. The assistance of predictive tools in data science can help us understand what certain health factors can affect a patient's risk for heart disease. For example, one of the leading risk factors for heart disease is cholesterol and exploring data for this risk factor can help to reduce heart disease. A predictive model can be created for the selected risk factors and explored to see what drives the prediction. As a result, doctors and patients will be to use the data from these predictive models to promote a healthier lifestyle.

Research Questions

1. What are the leading risk factors that cause heart disease?
2. What are some variables that can help reduce heart disease?
3. What more can I add to the model to make it representative of the entire world, not just the U.S. population? Perhaps more data from different parts of the world or their diet?
4. How effective is the model when considering family history and genetics?
5. Is the risk of cardiovascular disease higher in one sex compared to the other?
6. How can factors such as menopause in women affect the criteria of the model?
7. Are there other types of predictive model that will give us better results?
8. Are the risk factors influenced by people's profession?

Approach

Overview: The purpose of the project is to investigate the leading cause of death in the U.S. and shed light on which health factors can significantly affect a patient's risk for heart disease. We shall use a multiple regression model to analyse a correlation between cardiovascular disease and potential risk factor.

Detailed explanation: Any data scientist needs a good dataset in order to extract insight from data and use the valuable information to make predictions about a particular topic. In our case, we need a dataset with physiological characteristics (height, weight, etc), glucose level, cholesterol level, blood pressure, daily habits (smoking, alcohol consumption, exercising) as well as whether a person has cardiovascular disease. In addition, we will need a dataset that looks at the overall effect of exercise on a person's internal body such as blood pressure and cholesterol level. Once we have gathered our dataset, we will generate bar graphs and histograms of how cardiovascular disease is related to each variable. Based on our visual analysis, we can filter the variables if there exists a relationship between heart disease and potential risk factors. After applying data wrangling and transformation techniques to deal with empty rows, we will perform multiple regression analysis to generate a model to statistically evaluate a linear relationship between heart disease and the predictors. To determine whether our model is unbiased, we will be looking to perform certain tests and employ data visualization tools to ensure our model can accurately represent the population model. The multiple regression model can identify correlations that can help predict the outcome based on the indicators (independent variables). Given more time, we wish to create a more advanced model to identify all relationships between the variables using machine learning algorithms such as logistic regression or a Random Forest Classifier. Lastly, once we have all the relationships established we can evaluate our model to propose the likely cause of cardiovascular-related death and provide recommendations to reduce the risk of heart issues.

Datasets

Heart issue dataset link:

1. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
2. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Exercise dataset:

3. https://www.kaggle.com/datasets/kukuroo3/body-performance-data

Required Packages

- dplyr
- purr
- readxl
- readr
- car
- tidyr
- stringr
- knitr
- ggplot2
- ppcor
- caTools
- ROCR
- lm.beta
- pander
- randomForest

Needed Plots and Tables

- Bar graphs
- Histogram
- Box plot
- Scatter plot
- Pie chart
- Clusters
- Heat map

Questions for future steps

Some earlier variables that I tested for such as R-squared estimation or checking for conditions to see if the model is unbiased were difficult to interpret and I would find it hard to apply these results to the multiple linear regression to be used for the predictive model. Additionally, I encountered contradictory results based on reading from the book which makes me doubt the findings and methods used. For example, during one of my earlier assignments, I was testing for residual independence using the Durbin Watson test and I noticed that simple shuffling of the data set yielded a completely different conclusion, which made me question whether or not to transform the original data. In essence, I need to learn the necessary techniques to extract insightful information from statistics to make appropriate conclusions when creating a multiple regression model.

I need to get comfortable creating useful data visualization graph by properly tuning the parameters.

I have never used/created a machine learning model in r, so I need to learn how to develop a model and the necessary libraries and packages required.