Final Project Step 3

Mithil Patel

2022-06-04

Introduction

One of the leading causes of death in the United States every year is heart disease resulting in one out of every four deaths. Cardiovascular disease results in approximately 659,000 deaths yearly and while the CDC and other public health organizations have spent hundreds of millions of dollars on the budget for reducing heart disease-related deaths, it has still been a primary concern for the past several years. The lack of progression in the cardiovascular research field led me to pursue an alternate solution for the problem, one such solution being the usage of data science. With data science, we have the possibility of solving many of the world's issues such as cardiovascular disease with in-depth analysis. The utilization of many predictive tools in data science will 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 factor will be crucial in reducing the disease. A predictive model will be created to determine the risk factors that we've selected from our datasets and explored to see what drives the prediction. Feedback from the model can assist individuals to detect and monitor the long-term risk factors to reduce the chances of having heart disease. This study will be advantageous to doctors and patients who will be able to use the analysis to promote a healthier lifestyle.

The problem statement you addressed

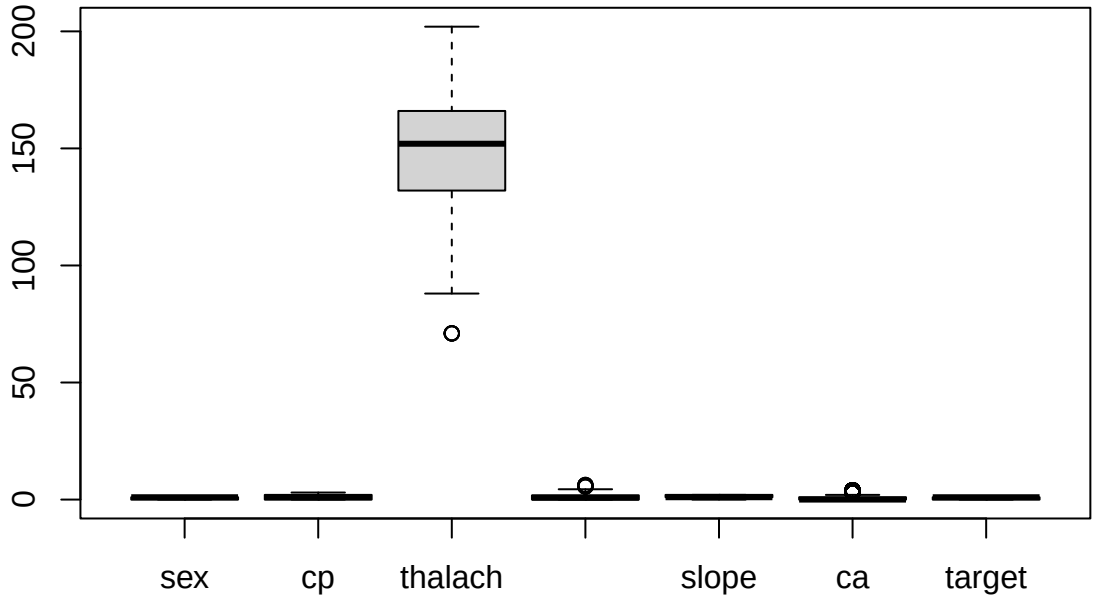
The purpose of the project is to investigate the leading cause of death in the U.S. and shed light on which health factors can significantly affect a patient's risk for heart disease. We shall create a logistic model as well as a k-nearest neighbors (KNN) model to explore the relationship between cardiovascular disease and potential risk factor to propose the likely cause of cardiovascular-related death and provide recommendations to reduce the risk of heart disease.

Procedure

Final Dataset sample

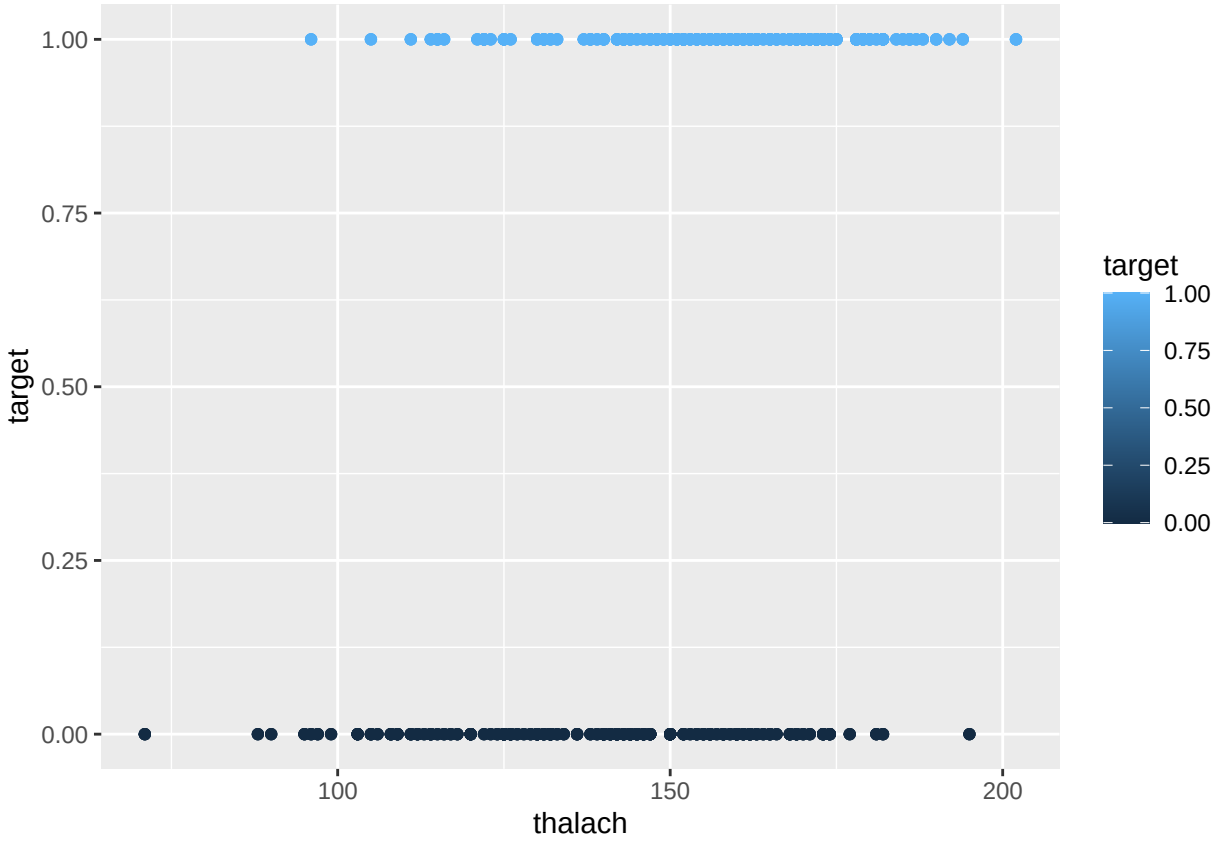
	sex	cp	thalach	oldpeak	slope	ca	target
2041	1	0	131	2.2	1	3	0
2042	1	3	174	1.4	1	1	0
2043	1	0	95	2.0	1	2	0
2044	1	0	158	0.0	2	0	0
2045	1	0	143	0.1	2	0	1
2046	1	1	164	0.0	2	0	1
2047	1	0	141	2.8	1	1	0
2048	1	0	118	1.0	1	1	0
2049	0	0	159	0.0	2	0	1
2050	1	0	113	1.4	1	1	0

Boxplots to detect outliers



The boxplot above shows several outliers exist in our dataset. We can handle the outlier values by simple removing the rows from our dataset.

Scatterplot: thalach vs target



Logistic Regression Summary

```
Call:
glm(formula = target ~ sex + cp + thalach + oldpeak + slope +
    ca, family = binomial(), data = combined_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2723  -0.4951   0.1722   0.6294   2.3694

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.954097   0.717932  -4.115 3.88e-05 ***
sex         -1.626687   0.209996  -7.746 9.46e-15 ***
cp           0.922050   0.090416  10.198  < 2e-16 ***
thalach      0.024610   0.004668   5.272 1.35e-07 ***
oldpeak     -0.680329   0.106020  -6.417 1.39e-10 ***
slope        0.540619   0.175040   3.089  0.00201 **
ca          -0.724593   0.093274  -7.768 7.95e-15 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1420.24  on 1024  degrees of freedom
Residual deviance:  815.57  on 1018  degrees of freedom
```

```
AIC: 829.57

Number of Fisher Scoring iterations: 5
```

Multicollinearity Condition

```
[1] 1.204538
```

```
[1] 0.8433963
```

The accuracy of our logistic model is:

```
[1] 82.65027
```

K-nearest neighbor (KNN) model accuracy table

```
  k_number accuracy
1        2 97.07317
2        3 85.36585
3        5 84.87805
4       10 85.36585
5       15 86.82927
6       20 82.92683
7       25 78.53659
```

How you addressed this problem statement

For the problem we want to address, I first needed to find a proper dataset to extract insight from the data and use its valuable information to make predictions about a particular topic. For our case, I needed 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, I needed a dataset that looked at the overall effect of exercise on a person's internal body such as blood pressure and cholesterol level.

Once datasets were gathered, I went through the process of cleaning and combining datasets into one using their similar variables/columns. Afterward, I received rows with empty values that did not match and removed them to not skew the dataset. Then, I used the cor() function to generate a correlation matrix to determine the variables that were correlated to the 'target' variables (closer to 1 meaning positive for heart disease, 0 indicating otherwise). To determine relevant columns, the variables with a correlation coefficient above 0.20 or below -0.20 were selected. Box plots were used after the final dataset was ready to detect outliers in each column. Given the abundant number of data points, the rows with outliers were removed from the dataset.

To perform logistic regression, the dataset was split into training and testing sets with 80% of the data going into training while 20% going into testing our model. Before the accuracy of our model is determined, I used the VIF function and tolerance to test for no multicollinearity in our dataset. As shown above, the mean VIF value is close to 1 and the mean tolerance is above 0.2, we can conclude that the no multicollinearity condition is met, thus our dataset is not biased. It also suggests that the predictors are not highly correlated with one another and they can accurately provide useful information about the dependent variable. To test the accuracy of our model, the confusion matrix was calculated to be approximately 82%. To further improve the accuracy, a supervised machine learning algorithm called k-nearest neighbors (KNN) to create a model. Based on the accuracy of our KNN model with different k numbers, it is evident that KNN outperformed the logistic model with a model accuracy of 97.07% when k = 2.

Analysis

The logistic model summary seems to suggest that the cp (chest pain) predictor has the greatest effect on heart disease. Analyzing the performance of our both model, it is evident that the KNN model was more accurate in predicting the heart disease based on our independent variable compare to the logistic model. Considering that the output variable is categorical, logistic model will typically raise or lower the predicted probability to a discrete value and, as a result, lowering the accuracy of the model.

While evaluating variables to include to our model using the correlation matrix, I was surprised to find variables not strongly correlated with heart disease. Diastolic and systolic blood pressure, which measures the blood pressure in the arteries when the heart rests and when the heart beats, respectively, were not strongly correlated with body performance. Subsequently, a model exploring a relationship between body performance (i.e. sit-ups) and heart disease could not be created. Hence, I was unable to provide patients and doctors, with overwhelming statistical evidence, several lifestyle changes such as doing sit-ups daily to potentially lower the risk of cardiovascular disease.

Implications

Our analysis shows that the predictors used in our model are a notable contributor to cardiovascular disease. Therefore, individuals can keep track of these variables (i.e. maximum heart rate or chest pain) and hopefully be able to take measures to avoid life-threatening consequences. Individuals with the inflated blood vessels of the coronary artery or achieving maximum heart rate on daily basis shall understand the underlying risk associated with their condition and should seek medical attention. Another implication from our analysis, though not so prominent, is that body performance such as performing set-ups is not correlated with Diastolic and systolic blood pressure - blood pressure measurements to determine a health heart. Consequently, several body performance seems to have little to no effect on heart disease.

Limitations

There were many limitations I came across while working on my research project, from finding the right dataset to cleaning and merging different datasets. One such limitation would be when I had to combine different data sets, which resulted in disregarding multiple variables from each such as smoking, cholesterol, and alcohol consumption. Two different datasets were stacked on top of each with one common column 'target', it was difficult to include other variables to the final dataset as multiple entries of the combined dataset were filled with empty (N/A) values that could not be processed. For the future steps, it would be helpful to find an efficient means to merge the two datasets by similar variables types/columns to avoid a plethora of empty values within the combination. Another limitation I encountered was not being able to find desired datasets to investigate several factors (i.e. running or food) that could directly reduce heart disease. Factors that individuals can incorporate into their lives that could significantly benefit their body. Ergo, there exist a lack of datasets available to process for an accurate relationship.

Concluding Remarks

As I learned that the cardiovascular disease is the leading cause of death in the United States, I wanted to find ways to reduce the number of deaths and spread awareness for the potential risk factors. By creating a predictive model using a logistic regression, I can investigate which health factors can significantly affect a patient's risk for heart disease and what steps an individual can take to lead a potentially better lifestyle. The relatively high accuracy of our model (82 percent) suggest that the variables (sex, chest pain, maximum heart rate, ST depression, peak ST segment, and the number of inflated major vessels) used in our model are good predictors of heart disease. A change in one of the predictors can significantly affect the likelihood of an individual having a heart disease. For example, an increase in chest pain can be a good indicator that one is suffering from a cardiovascular disease. Our conclusion can be further supported by the high accuracy

(97 percent) of our KNN model. While the results should not be revered as a guideline for an individual's view on heart disease, it can be the extra step taken for them to identify the concerns. Future data science enthusiast have plenty of opportunity to build upon this study with proper datasets. For example, they can find an effective method to merge multiple datasets with similar column while preserving the number of rows and columns. Frankly, there is always a room for improvement.