Heart Disease Prediction

Hema Mallik Musunuru

15/06/2022

Summary

There is an increase in the day-to-day cases of heart failure disease at a very high rate, which has contributed to the need to have an efficient and effective way of predicting heart failure disease. Therefore this project has focused on developing a machine learning algorithm that will precisely and efficiently predict which patient is more likely to have heart disease by considering various medical attributes. The heart disease prediction system in this project will help to predict if a given patient, based on their medical history, is likely to be diagnosed with heart disease or not. The project has implemented the logistic regression machine learning algorithm for this task. This algorithm has been motivated by the gap identified in the literature review that proposes a model that can be used to improve the accuracy of predicting Heart Attacks for any patient. The results from this project are important as they will help enhance medical care and reduce costs for heart disease patients. the project also provides important knowledge essential in predicting patients with heart disease. The dataset used for this project has been retrieved from the Kaggle.com website.

Introduction

Heart disease is one of the primary reasons for course death in adults. These days cardiovascular diseases have become very common, and they describe a range of conditions that can affect one's heart. There was an estimate of 17.9 million global deaths from Cardiovascular disease by the World Health Organization. Therefore, this project intends to help predict patients or individuals who are likely to be diagnosed with heart disease through the help of their medical history. The project utilizes high blood pressure or chest pain to recognize patients who have heart disease symptoms. This will help diagnose heart disease with effective treatments and fewer medical tests to ensure that patients are cured accordingly. Therefore, the main objective of this project will be to check whether a patient is likely to be diagnosed with heart disease with the look of their medical attributes like high blood pressure, gender, fasting sugar level, chest pain, age, and many more. The project will utilize a dataset retrieved from the Kaggle.com website, which contains patients' medical attributes and history.

Literature Review

The project has been motivated by a significant amount of related work to diagnosing heart disease through the utilization of machine learning algorithms. In their study, Reddy et al. (2019) utilized Random Forest, Neural Network, and k-Nearest Neighbor to classify cardiac heart diseases as either normal or abnormal. From the results of their study, the authors were able to identify that the Random Forest machine learning algorithm could accomplish an optimal performance. In another study by Atallah & Al-Mousa, the author's utilized logistic regression, Random Forest, K-Nearest Neighbor, and stochastic gradient descent in predicting the diagnosis of heart disease. They were able to identify that the K-Nearest Neighbor algorithm achieved the highest accuracy of 90% compared to the other algorithms. In another study by Kannan & Vasanthi (2019), the authors implemented four machine learning algorithms to predict heart disease. These algorithms include logistic regression, stochastic gradient boosting, Random Forests, and support vector machines. From the analysis and prediction, it was identified that logistic regression had the highest accuracy of 86.5%. From the identified related works, it is clear that several machine learning algorithms were implemented, and most of the literature concentrated on the accuracy of these algorithms. However, this leaves a gap in identifying the main contributing factors to heart disease for individuals. Therefore, this project intends to perform heart disease predictions and identify and predict the main contributing medical attributes to heart disease. The project

Theory

H1: More males than females are diagnosed with heart disease.

H2: Older people are more likely to be diagnosed with heart disease.

Data

The dataset for this project was downloaded from Kaggle.com website through <https://www.kaggle.com/code/shravankumar94/heart-disease-analysis-with-r/data>. The data includes :

## Rows: 303

## Columns: 14

## $ age <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5…

## $ sex <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1…

## $ cp <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0…

## $ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1…

## $ chol <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2…

## $ fbs <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0…

## $ restecg <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1…

## $ thalach <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1…

## $ exang <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0…

## $ oldpeak <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0…

## $ slope <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1…

## $ ca <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0…

## $ thal <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3…

## $ target <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…

Since most of the predictor variable in the dataset do not have correct data type, there was the need to adjust the variable types.

heart\_data <- heart\_data %>%

mutate(cp = as.factor(cp),

restecg = as.factor(restecg),

slope = as.factor(slope),

ca = as.factor(ca),

thal = as.factor(thal),

sex = factor(sex, levels = c(0,1), labels = c("female", "male")),

fbs = factor(fbs, levels = c(0,1), labels = c("False", "True")),

exang = factor(exang, levels = c(0,1), labels = c("No", "Yes")),

target = factor(target, levels = c(0,1), labels = c("Health", "Not Health")))

glimpse(heart\_data)

## Rows: 303

## Columns: 14

## $ age <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5…

## $ sex <fct> male, male, female, male, female, male, female, male, male, m…

## $ cp <fct> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0…

## $ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1…

## $ chol <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2…

## $ fbs <fct> True, False, False, False, False, False, False, False, True, …

## $ restecg <fct> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1…

## $ thalach <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1…

## $ exang <fct> No, No, No, No, Yes, No, No, No, No, No, No, No, No, Yes, No,…

## $ oldpeak <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0…

## $ slope <fct> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1…

## $ ca <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0…

## $ thal <fct> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3…

## $ target <fct> Not Health, Not Health, Not Health, Not Health, Not Health, N…

Check for any missing values in the dataset for each variable.

colSums(is.na(heart\_data))

## age sex cp trestbps chol fbs restecg thalach

## 0 0 0 0 0 0 0 0

## exang oldpeak slope ca thal target

## 0 0 0 0 0 0

Check the target variable proportions.

table(heart\_data$target)

##

## Health Not Health

## 138 165

Methodology

Having the data pre-processed and identifying that the target variable is balanced, the next step involves splitting the heart\_data to train and test set. The train set will be for training the model, and the test set will validate the model.

set.seed(100)

index <- sample(nrow(heart\_data), nrow(heart\_data)\*0.7)

# Data train

heart\_train <- heart\_data[index,]

# Data test

heart\_test <- heart\_data[-index,]

Next, check the train data proportion of whether there is a balance for training the model to help minimize the problem of overfitting.

prop.table(table(heart\_train$target))

##

## Health Not Health

## 0.4575472 0.5424528

The model is created based on the train data. Variables like thal, cp, thal, fbs, and sex that may have a significant effect on the target variable have been used.

model\_heart <- glm(formula = target ~ sex + cp + fbs + thal, family = "binomial", data = heart\_train)

# Model summary

summary(model\_heart)

##

## Call:

## glm(formula = target ~ sex + cp + fbs + thal, family = "binomial",

## data = heart\_train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.5494 -0.6195 0.2813 0.6259 2.1699

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.42390 1.89429 -0.224 0.822931

## sexmale -1.00808 0.46155 -2.184 0.028955 \*

## cp1 2.63834 0.58159 4.536 5.72e-06 \*\*\*

## cp2 1.96708 0.43148 4.559 5.14e-06 \*\*\*

## cp3 2.65973 0.75439 3.526 0.000422 \*\*\*

## fbsTrue -0.11120 0.55873 -0.199 0.842244

## thal1 -0.01373 1.99362 -0.007 0.994505

## thal2 0.99582 1.89198 0.526 0.598653

## thal3 -0.71130 1.89910 -0.375 0.708000

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 292.36 on 211 degrees of freedom

## Residual deviance: 188.77 on 203 degrees of freedom

## AIC: 206.77

##

## Number of Fisher Scoring iterations: 5

Next, there needs to have a model where one model has to have the target variable without predictor and the other with all predictors.

# Create a model without predictor

model\_none <- glm(target ~ 1, family = "binomial", data = heart\_data)

# Create a model with all predictor

model\_all <- glm(target ~ ., family = "binomial", data = heart\_data)

# Stepwise regression backward

model\_back <- step(object = model\_all, direction = "backward", trace = F)

# Stepwise regression forward

model\_forw <- step(object = model\_all, scope = list(lower = model\_none, upper = model\_all), direction = "forward", trace = F)

# Stepwise regression both

model\_both <- step(object = model\_all, scope = list(lower = model\_none, upper = model\_all), direction = "both", trace = F)

Results

Having build the model the next step involve making predictions by the use

heart\_test$prediction <- predict(model\_both, type = "response", newdata = heart\_test)

pred <- predict(model\_both, type = "response", newdata = heart\_test)

result\_pred <- ifelse(pred >= 0.5, "Not Health", "Health")

# Put our result prediction into our test data

heart\_test$prediction <- result\_pred

heart\_test %>%

select(target, prediction) %>%

head(5)

## target prediction

## 6 Not Health Not Health

## 10 Not Health Not Health

## 17 Not Health Not Health

## 18 Not Health Not Health

## 21 Not Health Health

#install.packages("proxy")

conf\_mat <- confusionMatrix(as.factor(result\_pred), reference = heart\_test$target, positive = "Not Health")

conf\_mat

## Confusion Matrix and Statistics

##

## Reference

## Prediction Health Not Health

## Health 32 2

## Not Health 9 48

##

## Accuracy : 0.8791

## 95% CI : (0.794, 0.9381)

## No Information Rate : 0.5495

## P-Value [Acc > NIR] : 1.373e-11

##

## Kappa : 0.752

##

## Mcnemar's Test P-Value : 0.07044

##

## Sensitivity : 0.9600

## Specificity : 0.7805

## Pos Pred Value : 0.8421

## Neg Pred Value : 0.9412

## Prevalence : 0.5495

## Detection Rate : 0.5275

## Detection Prevalence : 0.6264

## Balanced Accuracy : 0.8702

##

## 'Positive' Class: Not Health

##

recall <- round(46/(46+4),3)

specificity <- round(30/(30+11),3)

precision <- round(46/(46+11),3)

accuracy <- round((46+30)/(46+30+11+4),3)

matrix <- cbind.data.frame(accuracy, recall, specificity, precision)

matrix

## accuracy recall specificity precision

## 1 0.835 0.92 0.732 0.807

The prediction results illustrate that the probability of a male having heart disease is 16.4%. In addition, individuals with a higher level or severe type of pain, that is, cp=3, have a 93% chance of being diagnosed with heart disease.

# Return the probability value

model\_both$coefficients %>%

inv.logit() %>%

data.frame()

##.

## (Intercept) 0.96458414

## sexmale 0.16361982

## cp1 0.73702796

## cp2 0.90204468

## cp3 0.92820546

## trestbps 0.49447320

## exangYes 0.29894121

## oldpeak 0.38232308

## slope1 0.28766587

## slope2 0.66835966

## ca1 0.08665850

## ca2 0.04272171

## ca3 0.09384589

## ca4 0.77419824

## thal1 0.93239645

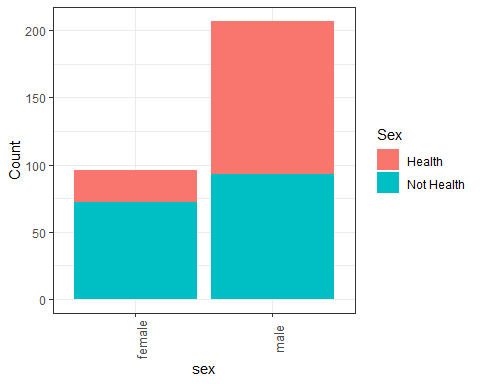
## thal2 0.91396258

## thal3 0.71437611

For sex and target, there are more male patients with the heart disease diagnosis than that of the female patients.

## `summarise()` has grouped output by 'sex'. You can override using the `.groups.`

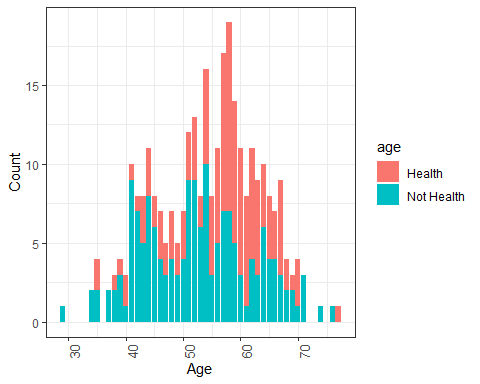
## argument.



From the analysis of age and diagnosis of heart disease, it is clear that

## `summarise()` has grouped output by 'age'. You can override using the `.groups.`

## argument.



Conclusion

The theory that more males than females are diagnosed with heart disease is true. More males are diagnosed with heart disease, as identified in the analysis. On the other hand, heart disease is not high in older patients, so the theory that Older people are more likely to be diagnosed with heart disease is not true. Heart disease is high in individuals aged 50 years, but beyond 60 years, the number of patients with heart disease is very few.

References

Atallah, R., & Al-Mousa, A. (2019, October). Heart disease detection using machine learning majority voting ensemble method. In 2019 2nd international conference on new trends in computing sciences (ictcs) (pp. 1-6). IEEE.

Kannan, R., & Vasanthi, V. (2019). Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. In Soft computing and medical bioinformatics (pp. 63-72). Springer, Singapore.

Reddy, N. S. C., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing, 9(1).