A Model Predicting Heart Disease

Machine Learning

Created By: Lukas van der Watt

Date: 12/12/2021

Contents

[Abstract 3](#_Toc90218098)

[Background 3](#_Toc90218099)

[Business Problem 3](#_Toc90218100)

[Data Description 4](#_Toc90218101)

[Approach 5](#_Toc90218102)

[Analysis and Model Implementation 6](#_Toc90218103)

[Model Interpretation: Logistical Regression 9](#_Toc90218104)

[Model Interpretation: Naïve Bayes 14](#_Toc90218105)

[Conclusion 14](#_Toc90218106)

[References 15](#_Toc90218107)

**Figures**

[Figure 1: Logistic Regression Prediction vs. Actual Health Status 10](file:///C:\Users\lukas\Desktop\Fundamental%20of%20Machine%20Learning\Heart_MLFinal.docx#_Toc90218112)

## Abstract

In this project paper a model is created with the purpose of predicting the health status of a patient relating to Heart Disease. This model addresses a business problem on behalf of a hospital. The data is sourced globally and is analyzed and processed in R studio. A better understanding of the importance in indicators available is obtained and a model is implanted to predict the health status of the patient.

## Background

Cardiovascular diseases (CVDs) are the leading cause of death globally. It causes the demise of an estimated 17.9 million lives each year. This accounts for 32% of all deaths worldwide. Different types of heart disease exist with coronary artery disease and heart attacks being the most common. CVDs represents 38% of premature deaths for people under the age of 70. In the US alone the financial impact of heart disease is astounding, amounting to a whopping $363 billion in costs ranging from medicines, health care services and lost productivity due to mortality. It is therefore important for hospitals and doctors to think about ways to reduce the death rate and the cost of this devastating illness. Most CVDs can be treated and prevented by addressing the patient’s behavioral risk factors. These factors include the use of tobacco and alcohol, unhealthy diet, obesity and physical activity. It is important to detect cardiovascular disease as early as possible so that treatment with counselling and medicines can begin. Herein lies the opportunity for a machine learning model to be instrumental in the fight against heart disease.

## Business Problem

A hospital has collected a set of records and information on patients with and without heart disease and would like to be able to predict and classify whether an individual is at high/low risk of developing heart disease. They would also like to know which of the indicators are more important to look at when trying to make this prediction. They have asked the data engineers to analyze the data and come up with a model that predicts the likelihood of patients contracting heart disease as early detection is critical in the management of CVD.

## Data Description

The following heart failure data set was sourced from Kaggle and contains 11 indicators that can be used to predict heart disease in patients. These 11 indicators are as follow:

1. Age: age of the patient in years
2. Sex: Indicating whether the patient is Male or Female
3. ChestPainType: The chest pain type the patient is experiencing [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: The resting blood pressure in mm Hg
5. Cholesterol: serum cholesterol in mm/dl
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

These indicators differ in data types like categorical variables, characters and numerical variables. This dataset was sourced from multiple countries around the world with 503 observation in the United States, 294 observations in Hungary, 123 observations in Switzerland and 270 observations from the Stalog Heart dataset. This makes 918 observations in total when duplicate observations are removed.

## Approach

To be able to predict whether a patient is likely to get heart disease, two techniques were used to attain the final model. Firstly, logistic regression was used to be able to understand the relationship between the dependent variable and one or more independent variables in the data. An estimation of probabilities was made using this technique. This type of analysis is used to help predict the likelihood of an event happening or a choice being made which works well for the prediction of sick or not sick. A more comprehensive understanding of the weights that each variable/indicator carries in the determination of heart disease was attained using the logistic regression technique. The most important variables were then identified and used to create the Naive Bayes model.

The Naïve Bayes model is used because a probability prediction of different classes needs to be made. Through this model the assumption of independence is being made. Meaning it is making the prediction that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. In this dataset it is a very rare occurrence that there would be predictors which are completely independent. Nevertheless, it is a technique that performs well with categorical variables where a sick/not-sick outcome is expected and despite the literal naive assumption of independence the technique does very well as it outperforms more sophisticated methods. To be able to validate the accuracy of the final model. The Naïve Bayes technique will be made from a 70% partition of the data. After this is done the prediction accuracy is validated with the remaining 30% of the data. Through this the final model accuracy can be attained.

## Analysis and Model Implementation

#Calling packages required to run the various commands   
library(ISLR)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(e1071)  
library(ggplot2)  
library(cowplot)  
#Reading in the data  
Heart <- read.csv('heart.csv')  
str(Heart) #Looking at the structure of the data to see what kind of variables are present

## 'data.frame': 918 obs. of 13 variables:  
## $ Serial.No. : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : int 40 49 37 48 54 39 45 54 37 48 ...  
## $ Sex : chr "M" "F" "M" "F" ...  
## $ ChestPainType : chr "ATA" "NAP" "ATA" "ASY" ...  
## $ RestingBP : int 140 160 130 138 150 120 130 110 140 120 ...  
## $ Cholesterol : int 289 180 283 214 195 339 237 208 207 284 ...  
## $ FastingBS : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ RestingECG : chr "Normal" "Normal" "ST" "Normal" ...  
## $ MaxHR : int 172 156 98 108 122 170 170 142 130 120 ...  
## $ ExerciseAngina: chr "N" "N" "N" "Y" ...  
## $ Oldpeak : num 0 1 0 1.5 0 0 0 0 1.5 0 ...  
## $ ST\_Slope : chr "Up" "Flat" "Up" "Flat" ...  
## $ HeartDisease : int 0 1 0 1 0 0 0 0 1 0 ...

head(Heart) #Looking at the first part of the data.

## Serial.No. Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG  
## 1 1 40 M ATA 140 289 0 Normal  
## 2 2 49 F NAP 160 180 0 Normal  
## 3 3 37 M ATA 130 283 0 ST  
## 4 4 48 F ASY 138 214 0 Normal  
## 5 5 54 M NAP 150 195 0 Normal  
## 6 6 39 M NAP 120 339 0 Normal  
## MaxHR ExerciseAngina Oldpeak ST\_Slope HeartDisease  
## 1 172 N 0.0 Up 0  
## 2 156 N 1.0 Flat 1  
## 3 98 N 0.0 Up 0  
## 4 108 Y 1.5 Flat 1  
## 5 122 N 0.0 Up 0  
## 6 170 N 0.0 Up 0

#Converting the required variables to factors   
Heart$Sex <-as.factor(Heart$Sex)  
Heart$ChestPainType <-as.factor(Heart$ChestPainType)  
Heart$RestingECG <-as.factor(Heart$RestingECG)  
Heart$ExerciseAngina <-as.factor(Heart$ExerciseAngina)  
Heart$ST\_Slope<-as.factor(Heart$ST\_Slope)  
#Omitting any missing values from the data.  
Heart<- na.omit(Heart)  
str(Heart)

## 'data.frame': 918 obs. of 13 variables:  
## $ Serial.No. : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : int 40 49 37 48 54 39 45 54 37 48 ...  
## $ Sex : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 1 2 2 1 ...  
## $ ChestPainType : Factor w/ 4 levels "ASY","ATA","NAP",..: 2 3 2 1 3 3 2 2 1 2 ...  
## $ RestingBP : int 140 160 130 138 150 120 130 110 140 120 ...  
## $ Cholesterol : int 289 180 283 214 195 339 237 208 207 284 ...  
## $ FastingBS : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ RestingECG : Factor w/ 3 levels "LVH","Normal",..: 2 2 3 2 2 2 2 2 2 2 ...  
## $ MaxHR : int 172 156 98 108 122 170 170 142 130 120 ...  
## $ ExerciseAngina: Factor w/ 2 levels "N","Y": 1 1 1 2 1 1 1 1 2 1 ...  
## $ Oldpeak : num 0 1 0 1.5 0 0 0 0 1.5 0 ...  
## $ ST\_Slope : Factor w/ 3 levels "Down","Flat",..: 3 2 3 2 3 3 3 3 2 3 ...  
## $ HeartDisease : int 0 1 0 1 0 0 0 0 1 0 ...

# Creating a table to see if variables related to sex are distributed throughout the data set. If it is distributed throughout the data set there is no need to omit a certain variable because of uneven distribution.  
xtabs(~HeartDisease + Sex, data = Heart)

## Sex  
## HeartDisease F M  
## 0 143 267  
## 1 50 458

xtabs(~ChestPainType + Sex, data = Heart)

## Sex  
## ChestPainType F M  
## ASY 70 426  
## ATA 60 113  
## NAP 53 150  
## TA 10 36

xtabs(~RestingECG + Sex, data = Heart)

## Sex  
## RestingECG F M  
## LVH 47 141  
## Normal 118 434  
## ST 28 150

#Looking at the tables above it is clear that the variable data is distributed throughout the data.

**Logistic Regression Model**

#Creating a logistical regression model called Log\_mod  
  
Log\_mod <- glm(HeartDisease ~ . , data = Heart, family = "binomial") #Specifying that the binomial family of generalized linear models are being used insures that logistic regression is being preformed with the glm() command.  
summary(Log\_mod)

##   
## Call:  
## glm(formula = HeartDisease ~ ., family = "binomial", data = Heart)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6566 -0.3739 0.1774 0.4482 2.5777   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.1557913 1.4160841 -0.816 0.414392   
## Serial.No. 0.0002254 0.0004887 0.461 0.644644   
## Age 0.0145904 0.0138578 1.053 0.292403   
## SexM 1.4720611 0.2802828 5.252 1.50e-07 \*\*\*  
## ChestPainTypeATA -1.8148593 0.3279399 -5.534 3.13e-08 \*\*\*  
## ChestPainTypeNAP -1.6924335 0.2665782 -6.349 2.17e-10 \*\*\*  
## ChestPainTypeTA -1.4916447 0.4324012 -3.450 0.000561 \*\*\*  
## RestingBP 0.0043751 0.0060191 0.727 0.467303   
## Cholesterol -0.0041406 0.0010888 -3.803 0.000143 \*\*\*  
## FastingBS 1.1384075 0.2749570 4.140 3.47e-05 \*\*\*  
## RestingECGNormal -0.1279691 0.2916698 -0.439 0.660845   
## RestingECGST -0.2116033 0.3710958 -0.570 0.568534   
## MaxHR -0.0049068 0.0052031 -0.943 0.345647   
## ExerciseAnginaY 0.8982456 0.2446288 3.672 0.000241 \*\*\*  
## Oldpeak 0.3779859 0.1186547 3.186 0.001445 \*\*   
## ST\_SlopeFlat 1.4705611 0.4295853 3.423 0.000619 \*\*\*  
## ST\_SlopeUp -0.9816831 0.4494492 -2.184 0.028948 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1262.14 on 917 degrees of freedom  
## Residual deviance: 593.97 on 901 degrees of freedom  
## AIC: 627.97  
##   
## Number of Fisher Scoring iterations: 6

### Model Interpretation: Logistical Regression

In knowing whether a variable is statistically significant we must look at the P value as well as the effect size (Estimate) of the variable. A P-value smaller than 0.05 is likely to be a significant variable to the target variable. A small p-value together with a higher effect size indicates that the variable is important in the determination of the target variable which is the HeartDisease variable. For Example: From the above results we can see that Age p-value is at 0.292403 which is quite high and above 0.05 with an effect size of 0.0146 (smaller related to other estimates) indicating that this variable is not very useful. Sex is a good predictor because the p-value 1.50e-07 being far below 0.05. The effect size is also bigger when compared to other variables.

Based on the Logistical model output the following variables have been selected to use in the Naïve Bayes model:

1. Sex
2. ChestPainType
3. Cholesterol
4. FastingBS
5. ExerciseAngina
6. Oldpeak
7. ST\_Slope
8. HeartDisease

Calculation of the McFadden’s Pseudo R^2:

log\_liklihood\_null <- Log\_mod$null.deviance/-2  
log\_liklihood\_prop <- Log\_mod$deviance/-2  
#The calculation for the Pseudo R^2:   
(log\_liklihood\_null-log\_liklihood\_prop)/log\_liklihood\_null

## [1] 0.5293913

This R squared is also known as the over-all effect size of the model. For the model above we get 0.5293913. This means the model explains 53% variability to the target variable. This accuracy of 53% is not very good but explains why it can be difficult in the real world to predict the development of heart disease. A way to improve this model might be to look at the genetics of the patient and family history pertaining to heart problems as well as the diet type of the patient. Having these additional indicators could help to improve the accuracy or over-all effect size.

#Plotting a probability graph for the Logistical Regression Model  
P\_data <- data.frame(Prob\_HD = Log\_mod$fitted.values,HD = Heart$HeartDisease)  
#Sorting the data from low to high probability  
P\_data <- P\_data[order(P\_data$Prob\_HD,decreasing = FALSE),]  
P\_data$rank <- 1:nrow(P\_data)  
  
ggplot(data = P\_data, aes(x = rank, y = Prob\_HD))+geom\_point(aes(color=HD),alpha = 4,shape=1, stroke = 1)+xlab("Heart Disease Status")+ylab("Prediction Probability of Contracting Heart Disease")

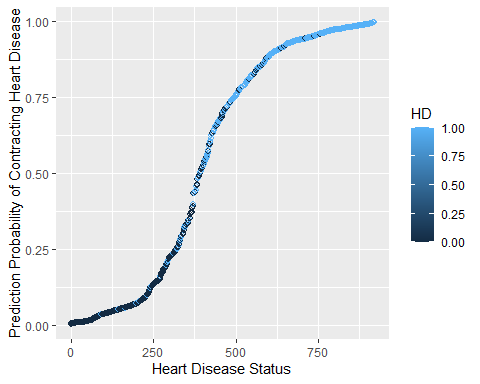


Figure : Logistic Regression Prediction vs. Actual Health Status

Figure 1 is plotting the prediction of each patient contracting heart disease against their actual heart disease status. The light blue indicates that most people who have heart disease have a high probability of contracting heart disease. Similarly, we can also see the low probability end indicated with the dark blue. These are patient that does not have heart disease and have a low probability of getting heart disease. We can see that there are a few cases where a patient without heart disease (dark blue markers) has a high probability of contracting heart disease at some point. This is what we want to be able to identify and predict.

**Naïve Bayes Model**

#Creating the Naive Bays Model  
#Partition the data into training(70) validation(30)  
selected.var <- Heart[,c(3,4,6,7,10,11,12,13)] #Selecting variables to be partitioned  
  
set.seed(123) #randomize  
train.in <- createDataPartition(selected.var$HeartDisease, p = 0.7, list = FALSE) #creating a training index with the stratification variable (HeartDisease) that contains 70% of the Universal bank data. The createdatapartition command has to have a target variable identified when the index is created. This has to be the target variable for which you want similar representation in the Training and Validation sets.  
Heart.train <- selected.var[train.in,] #Training set  
Heart.valid <- selected.var[-train.in,] #Validation set  
str(Heart.train) #structure of the training set

## 'data.frame': 643 obs. of 8 variables:  
## $ Sex : Factor w/ 2 levels "F","M": 1 2 2 2 1 1 2 2 1 2 ...  
## $ ChestPainType : Factor w/ 4 levels "ASY","ATA","NAP",..: 3 3 3 2 2 3 2 1 2 1 ...  
## $ Cholesterol : int 180 195 339 208 284 211 204 234 273 248 ...  
## $ FastingBS : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ExerciseAngina: Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 2 1 1 ...  
## $ Oldpeak : num 1 0 0 0 0 0 0 1 1.5 1 ...  
## $ ST\_Slope : Factor w/ 3 levels "Down","Flat",..: 2 3 3 3 3 3 3 2 2 2 ...  
## $ HeartDisease : int 1 0 0 0 0 0 0 1 0 1 ...

#Method 1: This shows the pivot table with row variables (Online and CreditCard) and the column variables(Personal.Loan) These table shows the count  
attach(Heart.train) # Attaching the training set to the following statements  
ftable(HeartDisease, Sex, ChestPainType)#Creating a pivot table with ChestPainType as a column variable and the Heart Disease and Sex as row variables

## ChestPainType ASY ATA NAP TA  
## HeartDisease Sex   
## 0 F 25 45 29 5  
## M 41 66 68 11  
## 1 F 30 2 5 1  
## M 245 13 41 16

ftable(HeartDisease,FastingBS,ExerciseAngina)#Creating a pivot table with ExerciseAngina as a column variable and the Fasting BS and ExerciseAngina as row variables

## ExerciseAngina N Y  
## HeartDisease FastingBS   
## 0 0 217 36  
## 1 30 7  
## 1 0 85 152  
## 1 56 60

The above pivot tables show the conditional probabilities (count) as they relate to heart disease.

#The Following table is a probability representation of the pivot tables previously formed  
prop.table(ftable(HeartDisease, Sex, ChestPainType))

## ChestPainType ASY ATA NAP TA  
## HeartDisease Sex   
## 0 F 0.03888025 0.06998445 0.04510109 0.00777605  
## M 0.06376361 0.10264386 0.10575428 0.01710731  
## 1 F 0.04665630 0.00311042 0.00777605 0.00155521  
## M 0.38102644 0.02021773 0.06376361 0.02488336

prop.table(ftable(HeartDisease,FastingBS,ExerciseAngina))

## ExerciseAngina N Y  
## HeartDisease FastingBS   
## 0 0 0.33748056 0.05598756  
## 1 0.04665630 0.01088647  
## 1 0 0.13219285 0.23639191  
## 1 0.08709176 0.09331260

detach(Heart.train)

In the above results we see that the conditional probabilities used for reference when the implementation of the Naïve Bayes model has been completed. The above table show the probability of getting heart disease given the patient’s Sex and Chest Pain Type. The 2nd table output shows the probabilities of getting heart disease given the FastingBS and the ExerciseAngina of the patient.

Heartdata.nb <- naiveBayes(HeartDisease~., data = Heart.train)# Creating the Naive Bayes Model on the training set. Using all variables as they relate to the target variable HeartDisease  
Heartdata.nb # Showcasing the model

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## 0 1   
## 0.4510109 0.5489891   
##   
## Conditional probabilities:  
## Sex  
## Y F M  
## 0 0.3586207 0.6413793  
## 1 0.1076487 0.8923513  
##   
## ChestPainType  
## Y ASY ATA NAP TA  
## 0 0.22758621 0.38275862 0.33448276 0.05517241  
## 1 0.77903683 0.04249292 0.13031161 0.04815864  
##   
## Cholesterol  
## Y [,1] [,2]  
## 0 223.2552 75.90595  
## 1 178.5637 125.01401  
##   
## FastingBS  
## Y [,1] [,2]  
## 0 0.1275862 0.3342052  
## 1 0.3286119 0.4703753  
##   
## ExerciseAngina  
## Y N Y  
## 0 0.8517241 0.1482759  
## 1 0.3994334 0.6005666  
##   
## Oldpeak  
## Y [,1] [,2]  
## 0 0.4134483 0.6698075  
## 1 1.3186969 1.1723818  
##   
## ST\_Slope  
## Y Down Flat Up  
## 0 0.03448276 0.19310345 0.77241379  
## 1 0.08781870 0.77337110 0.13881020

pre <- predict(Heartdata.nb,Heart.valid)#Class membership prediction of the target variable HeartDisease using the Naïve Bayes model  
pre.prob <- predict(Heartdata.nb,newdata = Heart.valid,type = "raw")#Probabilities

Heart.valid$HeartDisease<-as.factor(Heart.valid$HeartDisease)

cfm <- confusionMatrix(pre,Heart.valid$HeartDisease) )#Creating a confusion matrix to be able to see the accuracy of the model as well as see its efficiency  
cfm

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 105 19  
## 1 15 136  
##   
## Accuracy : 0.8764   
## 95% CI : (0.8315, 0.9128)  
## No Information Rate : 0.5636   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.7496   
##   
## Mcnemar's Test P-Value : 0.6069   
##   
## Sensitivity : 0.8750   
## Specificity : 0.8774   
## Pos Pred Value : 0.8468   
## Neg Pred Value : 0.9007   
## Prevalence : 0.4364   
## Detection Rate : 0.3818   
## Detection Prevalence : 0.4509   
## Balanced Accuracy : 0.8762   
##   
## 'Positive' Class : 0   
##

### Model Interpretation: Naïve Bayes

Bear in mind that a 0 indicates the patient is normal where a 1 indicates the presence of heart disease.

False Negatives: A Total of 19 cases of the 275 observations in the validation set.

False Positives: A total of 15 cases of the 275 observations in the validation set. Therefore, a total of 34 misclassification errors appears in the prediction. The Naïve Bayes Model Indicates a Prior probability of 45% that a patient is healthy and a 55% that a patient will be sick with Heart Disease.

Sensitivity is the proportion of positives correctly identified and in this Naive Bayes model we see that 87.5% of the positives are correctly identified.

Specificity is the True Negative Rate and, in this model, we see that 87.7% of the negatives are correctly identified. Finally, the Accuracy of the model can be seen to be 87.6%. This is not ideal in the real world as in has an error percentage of over 10%. However, there is potential for this model to be more accurate as more indicators could be added to the data set namely genetic sequence of the patient, family history and diet types. This will be able to have a relevant impact in the determination and likelihood of getting heart disease.

## Conclusion

Through the implantation of this model, we are able to accomplish the goal set out in the business problem. A prediction model is made that can determine whether a patient will contract heart disease or not. This model has an accuracy of about 88% and therefore could be deemed instrumental in the use of early detection for heart disease. In using the logistic regression technique, we are able to distinguish the importance of different indicators when trying to predict heart disease. From the 11 predictors, 8 were seen to be important. With more indicators and data, a more accurate model can be made. The assumption of independence was made when the Naïve Bayes model was made.

In predicting the health status of the patient, a sufficient management strategy can be implemented to prevent further deterioration in health and can even be reversed. This model has the potential of saving countless lives as well as money to the magnitude of millions. In the words of Dr. Caldwell Esselstyn, a researcher of coronary heart disease at Cleveland Clinic; “*If the truth be known coronary artery disease is a toothless paper tiger that need never, ever exist and if it does exist it need never, ever progress*”.

## References

1. fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [12/01/2021] from <https://www.kaggle.com/fedesoriano/heart-failure-prediction>
2. Heart\_MLFinal : GitHub R-code for Lukas van der Watt

<https://github.com/lvanderw/lvanderw_64060>