Heart Disease Analysis Report

MITCHELL MONTANIO

ERIN VASUT

JUSTIN JOSEPH

Table of Contents

[Deliverable 1 3](#_Toc75458954)

[Frame the Problem 3](#_Toc75458955)

[Explore and Wrangle the Data 3](#_Toc75458956)

[Training and Testing the Data 5](#_Toc75458957)

[Deliverable 2 6](#_Toc75458958)

[Identifying Promising Models 6](#_Toc75458959)

[Conclusion 7](#_Toc75458960)

[Works Cited 9](#_Toc75458961)

[Appendix 9](#_Toc75458962)

# Deliverable 1

## Frame the Problem

According to the CDC, cardiovascular diseases are one of the nation’s biggest causes of fatalities for both men and women, with 1-in-4 deaths attributed to heart disease3. Many factors come into play when determining the cause of heart disease and who is most likely to be affected by it. Determining factors include gender, age, health history (i.e., diabetic, cholesterol levels, if the person is a smoker, etc.), and more. The overall purpose of this report will be to provide insight into what factors are most likely to cause heart disease by making observations on the heart disease dataset. Because heart diseases can occur due to a multitude of varying factors, understanding the root causes are important. These causes could be traced back to social, economic, educational, or other underlying reasons. By studying the impact of each factor on cardiovascular diseases, better efforts to combat the illnesses can be implemented.

A few of the questions we will be looking to answer include: Does gender and/or age increase the likelihood of heart disease? Does health history play a significant role in causing heart disease? If so, what specific pieces of health history are most significant? And finally, what model(s) can create the best predictions of TenYearCHD? A training set and test set of data will be built to compare how efficient multiple machine learning models are at predicting heart disease. Success will be measured by comparing visualizations and statistics, such as coefficients, the mean squared error, the coefficient of determination, and hit rates. Ideally, the model will reflect the ability to correctly train a test dataset and provide accurate results using machine learning predictions.

## Explore and Wrangle the Data

This dataset contains some basic information and health history on a sampling of individuals ranging from ages 32 to 70. From a high-level perspective, this dataset consists of 16 base fields and a total count of 4,238 rows. All the original field types are either integer or float, which is appropriate for this dataset. Each field may potentially serve an important role in this report analysis, so no columns will initially be removed. Several fields—including male, currentSmoker, BPMeds, prevalentStroke, prevalentHyp, diabetes, and TenYearCHD—are already binary fields This is good for data quality purposes as the data does not need to be manipulated into bins/categories for easier analysis and more effective machine learning. We do, however, anticipate that certain columns will not be helpful because they are heavily weighted in one direction; these columns include: diabetes, prevalentStroke, and BPMeds. They are binary fields where in each case, less than 3.1% of the datapoints are considered true. Male and non-male value counts are somewhat skewed with value counts of male=1622 and non-male=2034; this may render some visualizations less accurate than they appear. It may be possible to mitigate this effect by using averages instead of raw counts.

From a data quality perspective, the lack of a date column makes it difficult to understand how datapoints relate to one another; this also prevents the creation of any timeseries visualizations and observations. Another potentially useful column missing from this dataset is the lack of racial or ethnic background field. While not completely necessary, it would be useful to look at trends amongst different ethnic groups and to see if geographical background has any impact on cardiovascular illnesses.

This dataset contains a significant number of nulls in various columns; dropping rows that contain nulls would delete about 13.7% of the total data in this dataset. We consider this to be a relatively small percentage of the data and will test how accurate this assumption is using multiple data frames; two data frames will contain filled nulls while two data frames will have dropped all nulls. Our assumption is that dropping nulls will minimize the number of inaccurately skewed plots and simplify handling nulls for multiple people working with one dataset across multiple local machines. The first data frame, df, is the dataset where the missing values are filled in using a combination of means and modes, depending on each individual data skew/count. The second data frame, df2, is the original dataset where all rows with null values are dropped entirely. The third data frame, df3, is df2 with all strings converted to numeric values. The fourth and final data frame, df4, is df, but all columns containing strings have been dropped so we can feed it into a machine learning model.

To better understand the data, we have created three new binned columns using the following existing columns: age, totChol, sysBP, and diaBP. For the age column, we create ageRanges (Figure 3) by binning age in increments of 10; for example, 30-39, 40-49, etc. For the totChol column, anything under 200 is binned as “good”, 200 to 239 is binned as “elevated”, and 240 and above is binned as “high”2 to create the cholRanges column. For the sysBP and diaBP columns, we created bins that classify the combination of blood pressure values by creating a new column, bpRanges (blood pressure ranges); values are binned into one of “normal”, “elevated”, “HBP\_s1” (high blood pressure – stage 1), “HBP\_s2” (high blood pressure – stage 2), or “HC” (hypertensive crisis); see *Figure 1* for the logic1 behind these bins. Utilizing the column created in *Figure 1*, *Figure 2* shows the comparison between males and non-males and their respective blood pressure ranges at any given age.

## Training and Testing the Data

Two different test sets of data are created to compare/contrast the methods of predicting accurately. To keep things consistent, all datasets are partitioned at 80% training and 20% test. An 80:20 ratio is ideal because this is a common industry standard. The higher the training to testing ratio, the better the chance of predicting all the datapoints into test dataset, which in turn produces a better model. The 20% test data provides a more ideal indication of the model’s performance on unseen data. A balance between training and test data is important because if poor values are chosen, the data could be overfitted or underfitted and create inaccurate data predictions.

For the first set of testing code, we utilized the train\_test\_split method on df2. Our chosen target variable (y) was sysBP and the predictor variable (x) was diaBP. By reshaping the training and testing sets and creating a linear regression model, we were able to receive the following values:

Coefficients: 1.44254454

Mean Squared Error: 163.7811278400134

Coefficient of Determination (COD): 0.6231587940104739

These values show that the model is about 62.3% accurate (from the COD) at predicting sysBP from diaBP. Unfortunately, when running this test set against other test and target variables, the COD was usually on the lower end, indicating this test method is not ideal for the dataset. The second test method utilizes the decision tree method with the training and test datasets. This test was run against df4 and compared against our stored correct data for TenYearCHD, as TenYearCHD is the response variable for this dataset. With this method, an output of about 85.% accuracy is achieved. While both methods of testing can be useful, the decision tree method is able to provide a significantly higher accuracy and therefore is our chosen method for testing the data.

# Deliverable 2

## Identifying Promising Models

With multiple data frames to utilize, the random forest model was run several times to help understand which model yields the best coefficient of determination (COD) and mean squared error (MSE). For each data frame, the depth of the decision trees was manipulated to see how it would affect the outputs. For this dataset, the numbers improved significantly until around a depth of 20, and then plateaued completely at a depth of 25. For this reason, we chose to use 25 as our default depth level to maximize the model’s COD and minimize the model’s MSE. The number of decision trees (n\_estimators) for this dataset is set to 100 because anything above this level provided negligible changes in accuracy. It is also worth mentioning that because the random forest function cannot ingest strings, our custom function excludes all data types outside of bool, int, and float; because of this, df and df4 are effectively the same dataset with this test, as is reflected by both the COD and MSE.

By default, train size is 80% and test size is 20%, which results in a COD of 65.3% when using df and df4. These train and test set parameters are chosen to keep consistent with previous models, however, we have observed that by increasing to a 90/10 ratio, this model improves COD to about 75.7% and MSE to about 0.0472.

In terms of chosen predictor variables, a function is created that helps determine the best predictor variables for each data frame. The resulting top predictors for each data frame are not surprising. Factors such as sysBP, age, diaBP, prevalentHyp, glucose, BMI, totChol, diabetes, BPMeds, and male are consistently the top variables the function finds. Using our default parameters, the COD values range from 63.8% to 65.5%, depending on which data frame is used. For this dataset, we expected ‘male’ (gender) to play a larger role in predicting accuracy, however it is ranked consistently on the lower end of the top 10 predictors in each test. This implies that, according to this data, gender does not play a large role in developing heart disease. This may be partially because the dataset contains around 400 more datapoints for non-males.

Number of total predictors has been set at a default of 12, which was chosen because it gives the highest R-squared (R2) value and lowest Mean Squared Error (MSE). However, the change in quality of prediction is not as significant as it may seem with fewer predictors, as the top five predictors for TenYearCHD on df provide an R2 and MSE of about 0.626 and 0.044, respectively; R2 and MSE on the same model with 12 predictors are about 0.655 and 0.047, respectively. In situations with significantly more data that would require longer waits and more processing power to run these functions, we would likely opt to use a lower default amount of predictors.

Also surprising are the MSE values for the random forest models. In each case, the MSE values are very low (<1) in comparison to the previous MSE values above 100. This implies this model is much more reliable than a linear regression model. In terms of COD values, this model ranks between the first two models with a slightly better COD than the linear regression model and significantly less accurate than the single decision tree model.

# Conclusion

It is noted that while several data frames were created, the ones with filled nulls instead of dropped nulls yielded best results. Also, while the binning was useful for creating visualizations to see trends within the data and understand the data set, the binned variables themselves were not as useful in predictions.

In response to the questions originally asked, several answers were found through data analysis. First, while our assumption was that gender would play a significant role in predicting heart disease, the data analysis does not reflect this. ‘Male’ was not a top predictor for any of the created data frames in any of the models run. However, as expected, age is indicative of heart disease likelihood based off this analysis. For all four data frames, ‘age’ was the second top predictor variable. Health history does play a role in TenYearCHD predictions. Factors such as blood pressure ranges, diabetes, glucose, etc. are all in the top predictors, which will tie back strongly to diet, exercise, and other lifestyle routines.

In terms of the models, the linear regressions were not effective at predicting TenYearCHD, but could recognize relationships between some variables. This indicates while the test itself is a valid test in some cases, it won’t be useful for the data we wanted to know. The random forests models were moderately effective, providing a slightly better COD value but a significant improvement in MSE values. Overall, the best model for our data set was the decision tree. It had the highest correct predictions rate with acceptable MSE values.

This data set has potential to provide useful information to health organizations looking to pinpoint the likelihood of cardiovascular diseases within individuals, and hopeful catch signs early on before too much damage can occur.

# Works Cited

1: American Heart Association. “Understanding Blood Pressure Readings.” Www.heart.org, 2021, www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

2: Goldman, Rena. “What Are the Recommended Cholesterol Levels by Age?” *Healthline*, Healthline Media, 16 June 2020, www.healthline.com/health/high-cholesterol/levels-by-age.

3: “Heart Disease in the United States.” Heart Disease Facts, Centers for Disease Control and Prevention, 8 Sept. 2020, www.cdc.gov/heartdisease/facts.htm.

# Appendix

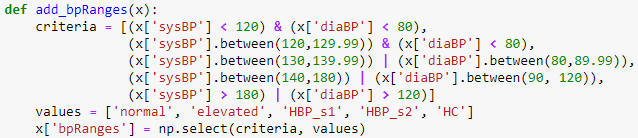


Figure : bpRanges Criteria

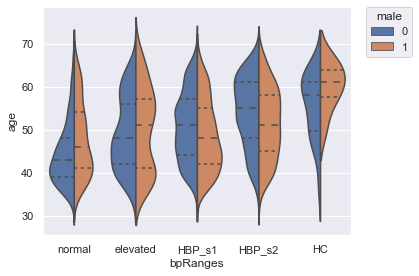


Figure : bpRanges by "male" and "age”

Chart, histogram

Description automatically generated

Figure 3: ageRanges bins