Predicting Heart Disease*

Walter Guillioli

December 05, 2024

Introduction

Unfortunately, cardiovascular diseases are the world's most common cause of death. The motivation of this project is to understand if we can predict the presence or absence of heart disease in people given basic medical information. This is a very personal project for me due to the disease history in my family - including my father.

In this project, we will look at a relatively simple dataset of 270 patients. We will first import and explore the data and then we will prepare it to apply different algorithms that will allow us to understand what can help predict a heart disease.

This file is just a summary. To see all the code and all the work behind the scenes see this file.

Data and Methods

The dataset was obtained from a study of heart disease that has been open to the public for many years. The study collects various measurements of patient health and cardiovascular statistics. It is a relatively small dataset with data for 270 patients and 13 variables of information for each patient. There is an additional binary variable that indicates the absence or presence of heart disease and that is the variable we will attempt to predict.

Table 1
Summary of patient's attributes

Attribute	Summary
Patient's age	Numeric, from 29 to 77 years
Patient's gender	Binary, 68% males, 32% females
Chest pain type	Categorical, 4 possible values
Resting blood pressure	Numeric, from 94 to 200
Serum cholestorol in mg/dl	Numeric, from 126 to 564
Fasting blood sugar	Binary, indicating if blood sugar > 120 mg/dl
Resting electrocardiographic results	Categorical, 3 possible values
Maximum heart rate (beats per minute)	Numeric, from 71 to 202
Exercise-induced chest pain (angina)	Binary, indicating if exercise produces pain
ST depression induced by exercise relative to rest	Numeric, from 0 to 6.2 (most are zero)
Slope of the peak exercise ST segment	Categorical, 3 possible values
Number of major vessels colored by flouroroscopy	Categorical, 4 possible values
Results of thallium stress test	Categorical, 3 possible values

*The data and scripts used are posted here: https://github.com/wguillioli/heart_disease/.

In total, there are 270 patients. Of those, 120 (44%) have heart disease and 150 (56%) do not have heart disease. Note that no missing values are present in the data.

The 13 available attributes of each patient are summarized in table 1. For more details see the competition page on drivendata.org.

Since we have a manageable number of variables for each patient, a detailed univariate exploratory analysis is performed to understand the details of the variable and more importantly to understand the potential prediction power of the absence or presence of heart disease.

Specifically, for numerical variables I explore the distribution of the data. Where appropriate I cap the observations at a particular percentile in order the make the distribution of the data "more normal". In other cases, a transformation of scale (logarithmic for example) is performed.

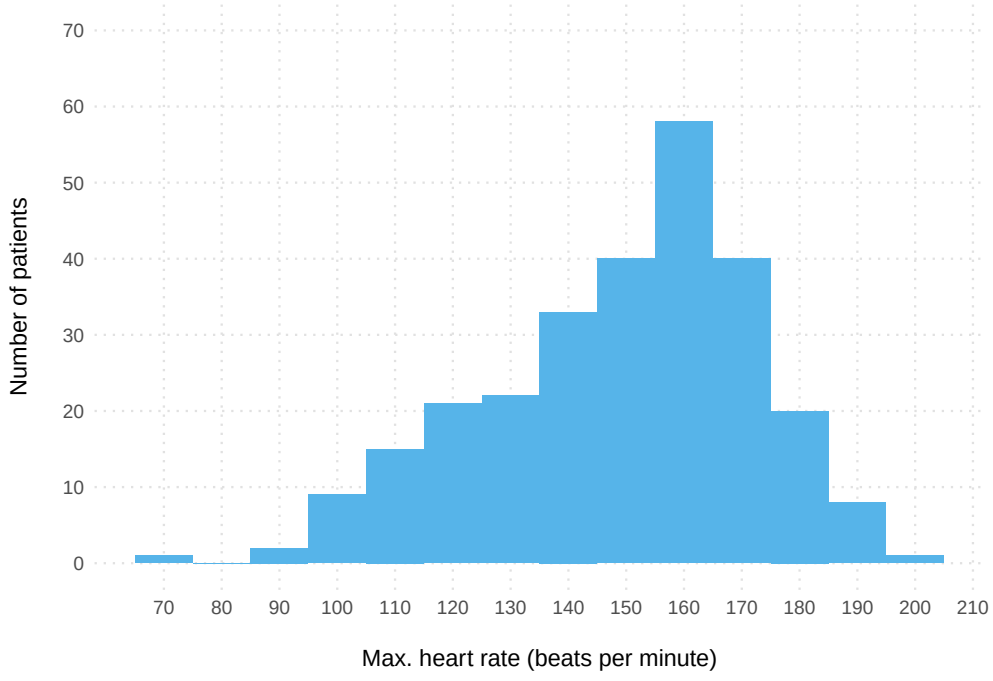
For discrete variables, I am mostly concerned about differences in the proportion of patients that have or don't have heart disease as it relates to the variable. If the counts of a particular value of these discrete variables are too low I perform groupings to help the ML algorithms we will use.

As an example let's look at one example of each variable here.

Maximum heart rate achieved (beats per minute)

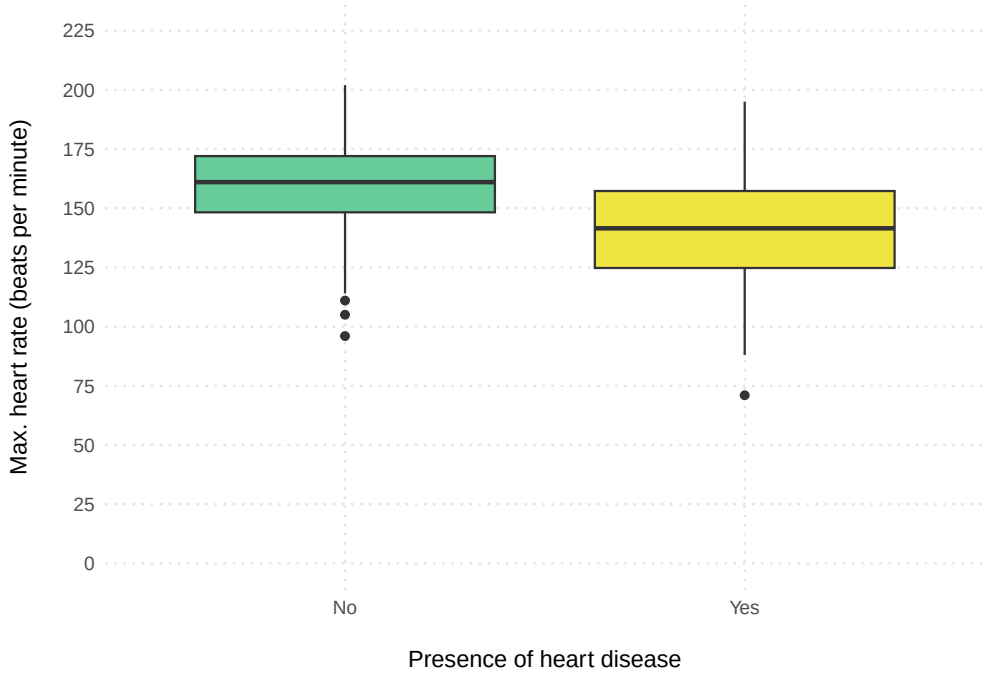
The number of beats per minute per patient ranges from 71 to 202 with a median of 153. The data is mostly normally distributed, as seen in figure 1.

Figure 1
Distribution of Patient's maximum heart rate achieved (beats per minute)



More importantly, we see a really good potential for prediction of maximum heart rate where patients with heart disease have a median of 141 vs 161 for those that don't have heart disease. It is also important to note that for patients with heart disease the middle half of the observations (percentile 25 to 75) range from 125 to 157, vs 148 to 174 for those without a heart disease. Figure 2 shows this distribution.

Figure 2
Patient's maximum heart rate achieved (beats per minute) by presence of heart disease



Results of thallium stress test

Test that measures the blood flow to the heart, with possible values normal, fixed_defect, reversible_defect.

At first, when looking at the data we observe that the counts for 6 (fixed) are low. We need to address this to help the ML algorithms we will use later. To do so, since the proportion of the presence of disease is similar to 3 (normal) we bin them together. See Table 2.

Table 2
Number of patients by results of thallium stress test

	3 (Normal)	6 (Fixed)	7 (Reversable)
No	119	6	25
Yes	33	8	79

After binning them together, we compute the proportion of patients with and without disease as it relates to the results of the thallium test. As we can see the proportions are inverted indicating that 80% of patients that have a heart disease have reversable results (Table 3). This is good news as far as the power of prediction goes, as we will see later.

Relationship within the attributes (Multicollinearity)

Another interesting thing is to see what variables are related to each other. This is important to understand the predictors and might need to be addressed. The chart below shows us the correlation values of the numerical variables of the patients. We are interested in big squares which means bigger correlations. Results are summarized on Figure 3.

For example, looking at the patient's age we see a relatively strong correlation between the number of vessels colored by fluoroscopy and the patient's blood pressure. In layman's terms for the latter, we could say that older patients tend to have higher blood pressure.

Table 3
Proportion of patients by results of thallium stress test

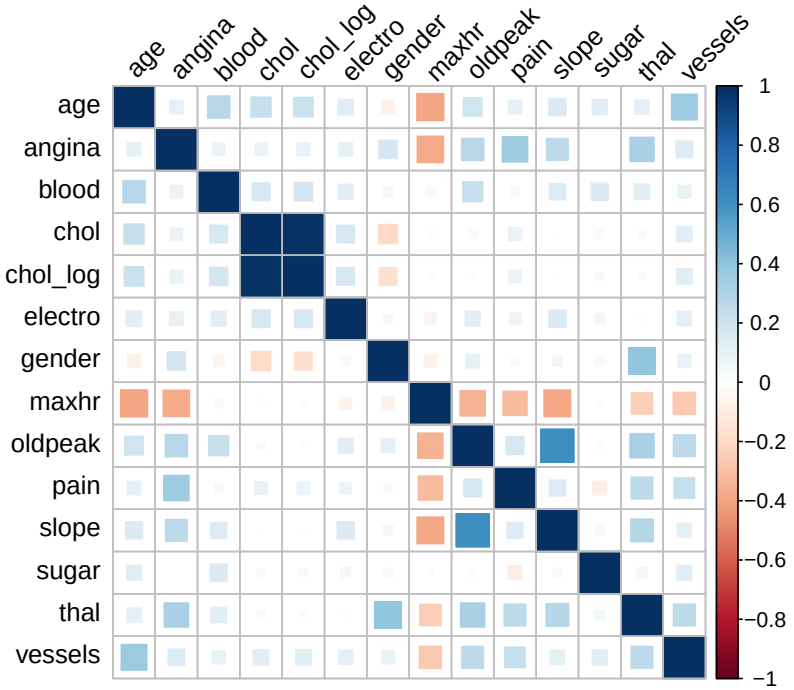
	Normal or Fixed	Reversable
No	0.8	0.2
Yes	0.2	0.8

In addition, we are interested in negative correlations (the red squares). For example, we can see that older patients tend to have lower maximum heart rates - which makes sense.

The slope of the peak exercise ST segment and ST depression induced by exercise relative to rest have a very high correlation of 0.6. The variables seem to measure similar information so one could be removed but will leave it for now.

Chest pain and exercise-induced pain (angina) are highly correlated. And as we will see they are very strong predictors of heart disease.

Figure 3
Relationship within numerical attributes



Results

Fortunately, after a careful univariate analysis one can see that several variables have a good potential of prediction of heart disease. Some to highlight are:

- **Chest pain type**: patients with higher pain types are more likely to have heart disease. For those with heart disease, 76% have the highest type of pain (4), while only 25% of patients without disease report that level of pain.

- **Results of thallium stress test**: similar to the previous attribute we see that for patients with disease, 66% report reservable type of pain while only 17% of those without disease report that result.
- **Maximum heart rate (beats per minute)**: see previous section for more details.
- **Patient's gender**: interesting to see that 55% of women have heart disease vs 83% of men.
- **Exercise-induced chest pain (angina)**: 55% of patients with heart disease report angina while only 15% of those without heart disease do.

Alternatively, there are a couple of unexpected results worth mentioning

- **Serum cholesterol in mg/dl**: there wasn't much difference in cholesterol values for patients with or without disease and this contradicts what I had expected.
- **Fasting blood sugar**: same as cholesterol there are no major differences across both groups of patients.

Now, let's compare our manual work with the automated work of ML algorithms and see what results are obtained.

Model results

In total, we performed three different ML models to predict the presence of heart disease.

1. Random Forest using the train/test validation approach to test the model on new data. I especially like this technique because it gives us a clear sense of what attributes are good predictors.
2. Logistic regression using train/test validation approach like in the previews model.
3. Random Forest using cross-validation. The idea here is to go beyond a simple random forest and let the algorithm explore different tuning parameters and in a way help us pick the best model. Interestingly it didn't change much.

The results are summarized in Table 4. We see all three are close to 80% in accuracy so the immediate question is, is that good or bad? Certainly, the results could be improved potentially if we tried different algorithms but that is not the goal here.

Considering that our dataset had 120 (44%) patients with heart disease and 150 (56%) without we see a very balanced dataset that was almost like flipping a coin. If we had said that all patients don't have heart disease we would be correct 56%. So these algorithms getting us to around 80% accuracy is a huge improvement.

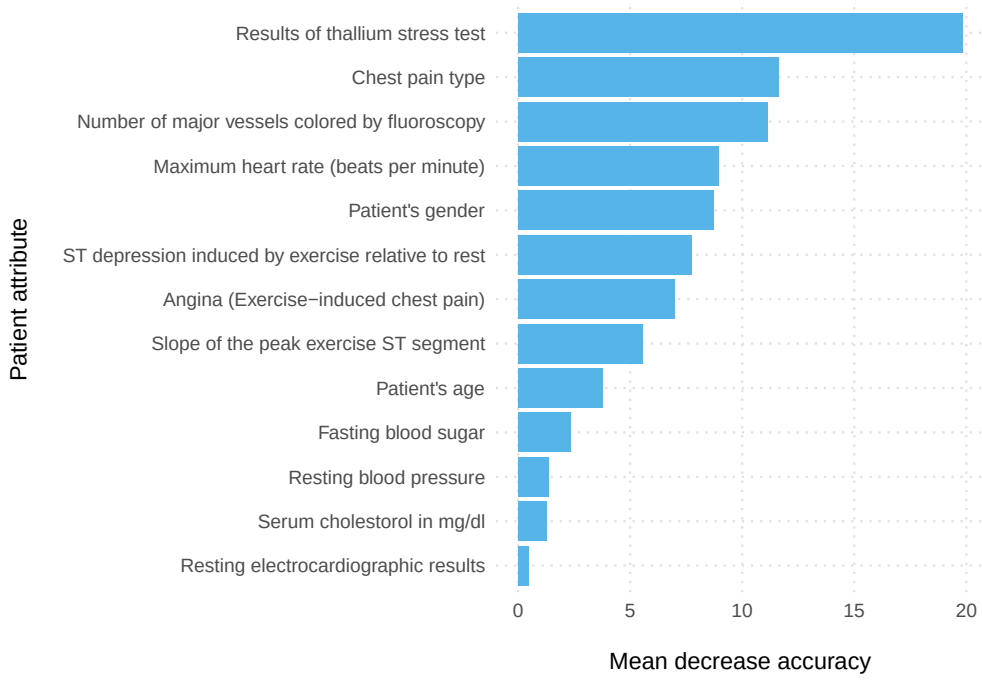
Table 4
Accuracy by Model

Modeling technique	Accuracy
Random Forest (train/test)	0.76
Logistic Regression (forward selection)	0.81
Random Forest (cross validation)	0.81

Important attributes predicting heart disease

One of the advantages of the Random Forest algorithm is that it ranks what variables are the most important in building all the trees. These results can be seen in Figure 4.

Figure 4
Attribute ranking



Another approach to explain the model

The previous results are great to get a sense of the importance of the variables. But the next question is how. For that, we used a very simple classification tree (Figure 6) that allows us to see some of the rules behind the trees to determine the presence of heart disease.

For example, let's consider the two scenarios where most of the patients are classified.

1. **Patients with heart disease**. If we follow the branch to the right we see first that patients with a reversible result of the thallium stress test will have a 90% of presence of disease. 26% of the patients fit this criteria.
2. **Patients without heart disease**. If we follow the branch to the left we see that if patients had a Slope of the peak exercise ST segment under 2 and a normal or fixed result on the Results of the thallium stress test they only have a 9% probability of heart disease. 36% of patients fit these criteria.

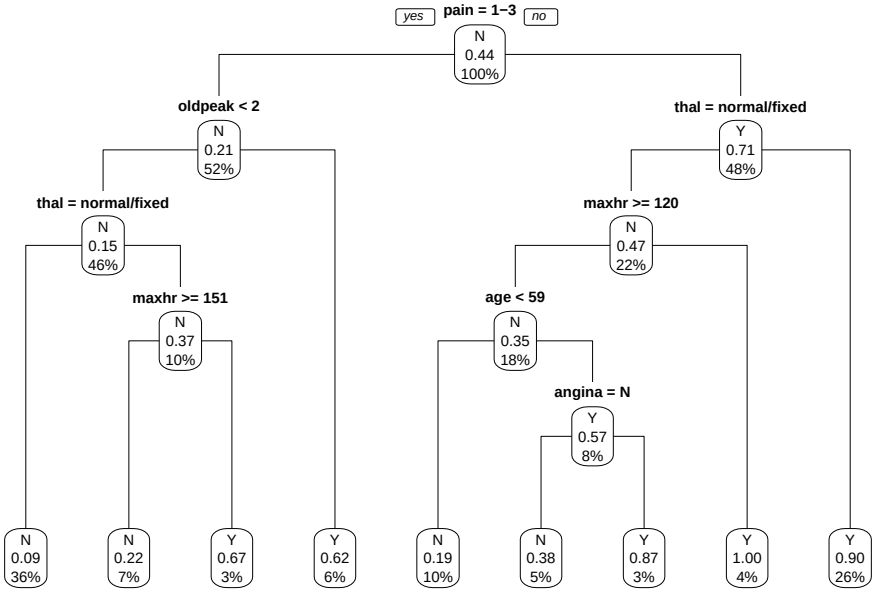
Conclusions

1. XXX add sobre modeling
2. Based on this simple dataset it seems that understanding what can cause the presence of heart disease is not that complex and there are simple measures we can take

3. Algorithms only provide partial solutions and we should not take them as the ultimate truth. For example, we know that many other variables are good potential predictors but the tree did not show them for example.

4. More data would be nice to explore different or complementary results. For example, it is a bit strange that Resting blood pressure and Serum cholesterol in mg/dl did not appear as strong predictors.

Figure 5
Top variables in Classification Tree



References

1. DrivenData. (n.d.). Warm-up: Machine learning with a heart. http://www.drivendata.org/competitions/54/machine-learning-with-a-heart

2. Cardiovascular diseases - global facts and figures. World Heart Federation. (2023, May 26). https://world-heart-federation.org/resource/cardiovascular-diseases-cvds-global-facts-figures/