Classifying Broken Hearts

Abdullah Naimzadeh

[anaimzad@syr.edu](mailto:anaimzad@syr.edu)



# Abstract

Machine learning algorithms have proven that they can be extremely powerful tools to solve complex problems from recommending your next purchase to forecasting and predicting fraudulent accounts. Applying these data mining techniques to classify healthy patients from non-healthy patients will save doctors time and potentially mitigate issues that lead to serious health concerns. For this reason, researchers have utilized machine learning to help health-care organizations diagnose illnesses such as heart disease [1]. Using RStudio, the following investigation employs classification algorithms to try and predict whether or not a patient has heart disease. Training models on over 200 patient observations shows high accuracy, but still requires the need of experts, like doctors, to verify the results in the event of misclassifying non-healthy patients as healthy ones.

# Introduction

According to the CDC, the leading cause of death here in the US is heart disease, which a quarter of Americans die from. That is no surprise with the rife amount of processed food selections, fast-food chains, and lack of walkable cities. Although we cannot cure heart disease when it is diagnosed, we can mitigate it by making it better.

The Cleveland health clinic has found that blood pressure and cholesterol are some of the key indicators to monitor when aiming to mitigate these health issues. These concerns can lead to serious problems like coronary artery disease, cardiovascular disease, etc. Many, including my own family, are even born with a heart defect. With this in mind, researchers have turned to data mining and machine learning as a possible tool in analyzing patient data to predict different diseases. Data mining techniques can be employed to help doctors in the industry make better decisions based on patient records and potentially enable them to react more effectively.

#### Heart Disease Data Set

The data to be utilized for our analysis is provided by The Cleveland Clinical Foundation and downloaded from the University of California Irvine’s machine learning repository. The observations within the data set include results from 303 patient medical tests. The original results contained over 76 different attributes but most research studies have found 14 to be the most useful. These attributes include patient demographics like age and gender, and the remaining are lab results that include blood pressure, exercise induced chest pain, cholesterol levels, etc. The final column is the ultimate diagnosis of the patient having various levels of heart disease or none at all. The diagnosis will be used as the predicting feature in our models later in this investigation. A mutated version of the original dataset was made for interpretability and complexity reduction. Both the full numerical and mutated dataset will be used. The metadata for the mutated dataset can be seen in Table A of the appendix.

# Research questions

The priority of this investigation is simple, yet important.

*Can classification algorithms predict heart disease accurately?*

Topics to explore that stem from the question above are:

* How does one mitigate bias?
* Does discretizing the data optimize the models?

Answering part or all of these questions will rely on preliminary research into the field of epidemiology and cardiology. It is no surprise that understanding how the data is obtained and what it means is advantageous in transforming and sub-setting the data. With this in mind, meaningful and appropriate binning techniques can be implemented to handle errors and discretize certain features.

# Method

The data science process enables scientist to properly investigate any data-related problem and have their work be replicable. Outlined below are the steps taken to effectively preprocess and analyze the data in question.

The Cleveland dataset was previously processed, but the data was reviewed to ensure that there were no inconsistencies that would skew the analysis. Since the number of patient data was relatively small and the majority of the data was numerical, missing values were replaced with the median rather than the mean which is sensitive to outliers.

As mentioned before, a copy of the original data was created and transformed into categorical values. In particular, continuous variables like age, heart rate, cholesterol, etc. were binned in order to reduce the complexity and improve interpretability for the Apriori analysis, and classification models. Lab results were binned according to the thresholds outlines by reputable sources like the American Heart Association and Hopkins Medicine [2,3].

For machine learning, both numerical and categorical datasets were then split 3:1 into a training and testing set. A partitioning method was used to divide the data set in a way that sustained similar distributions in the original datasets. Recall and F1 score were the main results used to compare each model, before and after tuning via a search grid; explained later in the discussion.

All analysis was conducted in the RStudio environment.

# Results

### Data Exploration

#### Initial Findings

In regards to medical research, it is widely known that there is bias in the findings of many papers. Historically, women and minorities are often underrepresented in research groups, and thus generalization leads to sometimes fatally incorrect diagnosis' due to the differences in biology between underrepresented populations and cis-white males. Thus it was no surprise that after pre-processing, it was found that more than half of the patients in our data set were males. More specifically, 114 of 139 patients that were diagnosed as having heart disease were male. Does this mean that females are typically healthier than males? Not necessarily. Models are susceptible to the bias, or imbalance, from the data. The full distribution of patients sex versus diagnosis can be seen in Table 1.

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| No Heart Disease | 72 | 92 |
| Heart Disease | 25 | 114 |

Table 1: Cross table of patient gender and diagnosis.

Some other interesting things to point out:

- The distribution of age reveals patients tend to be older in the data set as seen in Figure 1.

- The majority of patients with heart disease are in there 50's-early 60's as seen in Figure 2.

- Very few patients indicated normal chest pain.

- Many patients whom are healthy have normal ECG results.

- Healthier patients can reach a higher max heart rate when exercising.

- Patients with a flat slope in their ST wave segment or reversible-defect typically have heart disease.

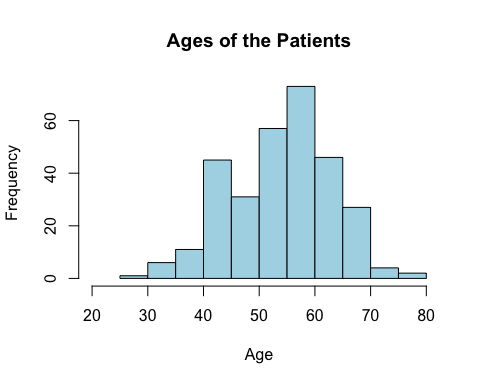


Figure 1. Distribution of patients age

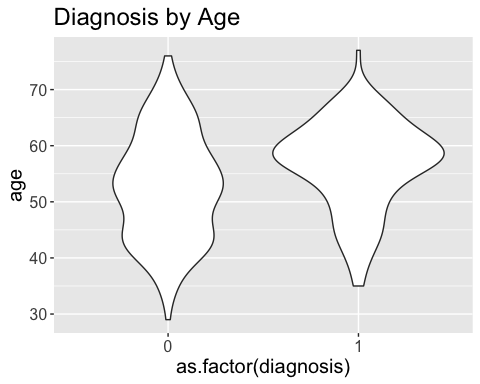


Figure 2. Violin plot of patient age and diagnosis.

## Models

#### Logistic Regression

A regression model was implemented to fit and identify significant attributes from the dataset. The best fit was with the categorical dataset, but unfortunately it could not explain 42% of the change in the data. The indicators that effectively influenced the models fit were chest pain, sex, number of blood vessels, thalassemia, cholesterol and oldpeak. These attributes were then used as a possible subset for other algorithms.

Information gain is another way of revealing how each attribute is significant in influencing the models. Similar to logistic regression, it was found that the top three indicators were the type of chest pain, the number of major vessels, and blood disorder.

#### Apriori

Apriori allows us to identify any indicators that occur more frequently together than individually with our target feature like heart disease. This *basket* of “combination : target label” is also known as rules.

Apriori rules revealed that for patients with heart disease, patients typically checked 2 or more of the following results:

* Asymptomatic chest pain
* Reversible thalassemia defect
* Is male
* Flat slope in ST segment

Apriori rules revealed that for patients with no heart disease, patients typically checked 2 or more of the following results:

* Is Female
* Zero major blood vessels
* Non-anginal chest pain
* No exercise induced angina

#### Classification Algorithms

As mentioned in the methods section, a search grid with the respective parameters were used to identify an optimal model. Table 3 shows how the recall and F1 score changes before and after tuning. Some scores are lower after tuning because of the prioritization of innate overall accuracy that the search grid uses.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ML Algorithm** | **Base Recall** | **Base F1 Accuracy** | **After Tuning Recall** | **After Tuning F1 Accuracy** |
| Decision Trees | 0.83 | 0.80 | 0.83 | 0.74 |
| Random Forest | 0.94 | 0.90 | 0.89 | 0.85 |
| SVM (Radial) | 0.93 | 0.88 | 0.93 | 0.88 |
| Naïve Bayes | 0.93 | 0.91 | 0.93 | 0.90 |

Table 3: Comparison of classification algorithm results before and after tuning.

SVM and naïve bayes were optimized using the numerical while decision trees and random forest were optimized with the categorical dataset.

# Discussion

Predicting heart disease is a type of classification problem where recall and F1 score make the most sense to focus on. Recall shows how selective the model is, and that there is high cost associated with false positives due to the fact that the consequences of misclassifying a patient as healthy can be dire. Due to the imbalance of the data, F1 score shows a more balanced accuracy since it incorporates coverage of the minority class in the dataset.

Each step in modeling the data reveal indicators that had the most weight in predicting heart disease. Apriori revealed how these indicators tended to be grouped together with the target feature. When comparing models, it was found that Naïve Bayes was the best classifier overall after tuning. This could be in due fact that the algorithm assumes that the variables are conditionally independent with respect to the target feature, and relies on the frequencies of the data.

The findings reveal fairly high accuracies for the positive class, no heart disease. Due to the high recall value, the model may incorrectly predict a healthy patient as having heart disease. Although this might be a scare, the opposite type of prediction would be worse. The indicator with the highest information gain, chest pain type and the number of major blood vessels, also reveal an important connection. When the number of major blood vessels available to supply blood, oxygen and nutrients to the heart decreases, the reduced blood flow will consequently induce angina. This relationship proves to cause other symptoms like shortness of breath, or even a heart attack. Moreover, having a preexisting condition like thalassemia would exacerbate the issue. This would make sense as to why the indicators mentioned have the highest influence on heart disease.

Training a classification model to predict heart disease has shown to be beneficial. However the caveat is when more and more patients become involved, then the possibility to misclassifying a patient as healthy becomes more consequential. For this reason, classification algorithms should not be relied on as the sole decision maker when deployed, but rather in tandem with a doctor to double check. With this in mind, there are some ways to try and improve the models in future investigations. Perhaps doing factor analysis to see if any of the indicators are significantly correlated can be helpful to reduce the complexity of the problem and further increase the accuracy of the model(s). Another important task would be to use a more balanced dataset in terms of demographics, or to use some technique to under sample male patients. Right now, the models generated in this experiment significantly benefit male patients more than female patients. To be ethical and just, a more representative dataset is needed to reflect the respective communities they affect.

# References

[1] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.

[2] “Understanding Blood Pressure Readings.” *Www.heart.org*, American Heart Association, www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

[3] *Understanding Your Target Heart Rate*, Johns Hopkins Medicine, www.hopkinsmedicine.org/health/wellness-and-prevention/understanding-your-target-heart-rate.

# Appendix

Table A. Mutation of original dataset to mostly factored level categorical variables.

|  |  |  |
| --- | --- | --- |
| Attribute | Data Type (all numeric) | Description |
| age | ordinal | Measured in decades: twenties to seventies |
| sex | dichotomous | 0 = F, 1 = M |
| cp | nominal | Chest pain  1 = typical, 2 = atypical, 3 = non-anginal, 4 = asymptomatic |
| rbp | binary | Resting blood pressure at admission  Normal : less than/equal to 120 mm Hg  Hypertensive: greater than 120 mm Hg |
| chol | ordinal | Cholesterol level  Normal : less than 200 mg/dl  At Risk : (200,240] mg/dl  High : greater than 240 mg/dl |
| fbs | binary | Fasting blood sugar > 120 mg/dl ?  1 = True, 0 = False |
| recg | nominal | Resting electrocardiographic results  0 = Normal  1 = ST-T wave abnormality  2 = Left ventricular hypertrophy |
| thalach | binary | Maximum heart rate during exercise (based on age)  Normal: less than threshold\*  High: greater than threshold\*  \*threshold = 0.85\*(220 – age) |
| ex\_angina | binary | Exercise induced chest pain/angina?  1 = True, 0 = False |
| oldpeak | continuous | ST-T depression induced by exercise |
| slope\_oldpeak | ordinal | ST segment slope during exercise  1 = upslope, 2 = flat, 3 = downslope |
| ca | interval | Number of major vessels ( 0-3 ) |
| thal | ordinal | Evidence of thalassemia, blood disorder  Normal = 3, fixed defect = 6, reversable defect = 7 |
| diagnosis | binary | Heart disease label  No heart disease = 0, heart disease = 1 |