ST494 Final Project
Classification of Heart Disease Risk

Senad Kokic
Melisa Hayalioglu
Mariam Lo
Kunal Vaid

April 12, 2023

Contents

Abstract

Heart disease is one of the deadliest killers in the modern era. While efforts are made day-to-day in order to mitigate travesty after the fact, it is equally as important to stop heart disease before it starts. There are many medical facets to this task, however, it is essential to accurately predict heart disease risk in order to take the right steps to stop development in its tracks. In this report, various deep and machine learning algorithms are used and compared so that accurate predictors of heart disease risk, and the most influential of these predictors, can be acknowledged and studied further. While suitable models seem to disagree with the literature, it is apparent that heart disease risk can be accurately predicted using known statistical models, commonly available medical devices, and patient insight.

1 Introduction

A report published by the World Health Organization in 2020 [1] tells a harrowing tale - pertaining to the leading causes of death in 2019. In this year, there were a recorded 55.4 million deaths, with the top 10 causes of death being directly responsible for 55% of them. These 10 causes are associated with 3 broad classifications: respiratory (chronic obstructive pulmonary disease, lower respiratory infections), neonatal (birth asphyxia, birth trauma, etc.), and, most pertinent to this report - cardiovascular conditions, such as stroke and ischaemic heart disease. 32% of deaths worldwide are caused by cardiovascular diseases, and this is purely in the sense of fatality [2]. If someone is to survive one of these cardiovascular events, their lives are permanently altered - with long term medical, musculoskeletal, and psychosocial complications [3].

According to the American Heart Association [4], 80% of heart diseases and strokes are preventable. However, a major component of prevention is the ability to recognize the risk of heart disease developing in the first place. Hence, the approach we have taken for this project - we are examining the efficacy of numerous covariates in order to accurately predict whether or not a certain patient is at risk for developing heart disease. Covariates including the patient's serum cholesterol level, resting electrocardiogram results, and the like. These results are of interest to many different groups - including, but not limited to:

1. **Patients**
   Of course, when it comes to recognizing a patient's risk for heart disease based on variety of different factors, it is of utmost importance for the patient to understand their levels of these factors in order to prevent disease occurrence. For instance, many factors that influence heart disease are manageable in the patient's life - such as their serum cholesterol levels. If a patient is made aware of any extremes in their cholesterol, they can make the appropriate changes to stop disease development in its tracks.

2. **Doctors**
   In a medical setting, time is of the essence. Performing extensive diagnostics may provide better insight into a patient's disease status, however, results may be delayed, leading to tragic consequences. Hence the need to research accurate predictors - harnessing the results of a couple of extremely strong predictors is considerably more time-efficient than prolonging diagnosis due to the sheer number of tests required.

3. **Medical Manufacturers**
   When it comes to the commercial and industrial availability of medical goods, research and development to produce the most efficacious product is a crucial step in the process. So, when it comes to manufacturers that produce medicine/equipment for diagnoses, it is in their best interest to know where to focus their efforts, and determine where the primary predictors of heart disease lie. If, say, serum cholesterol appears to be the best predictor, manufacturers can gear towards developing better tests for such a metric.

4. **Researchers**
   Finally, for many reasons similar to those outlined above, medical researchers will certainly be interested in determining the most effective predictors for heart disease. Not purely in terms of those identified - also when looking at the methods used to identify these predictors. Which models are used? What are the metrics used for comparison? As is commonality in medical research, new methods are frequently developed in evaluating collected data. Researchers can not only observe the results of these models, however they can also implement changes to further bolster the accuracy of the models being used.

To begin, we shall elaborate upon the data being investigated, as well as the methodology employed to examine such data.

2 Data

For this report, we used the Heart Failure Prediction data set, a combination of 5 independently created data sets, from Kaggle [5] The final data set contains 918 observations and over 11 common features related to a patients health status and medical history. We used the variable 'HeartDisease', a binary variable with a value of 1 to indicate if a patient has heart disease, or 0 if not. The variables in the data set include:

- Age - Age of the patient(years)
- Sex - Biological sex of the patient
- ChestPainType - The type of chest pain experienced (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
- RestingBP - Resting Blood Pressure (mm Hg)
- Cholesterol - Cholesterol level (mm/dl)
- FastingBS - Blood sugar level after fasting (if FastingBS ¿ 120 then 1, otherwise 0)
- RestingECG - Resting ECG levels (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)

- MaxHR - Maximum heart rate achieved (Numeric value between 60 to 202)
- ExerciseAngina - Exercise-induced angina (Y = Yes, N = No)
- Oldpeak - ST depression induced by exercise relative to rest (ST is a flat section of the ECG)
- STSlope - the slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping)
- HeartDisease - output class (1 = heart disease, 0 = Normal)

We used standardization before our data was used for the test and training model, to make sure all of the features in the data set were on the same scale. This was done by using the preProcess() function, which will make our model more interpretable and easier to compare.

3 Methodology

For this project, we used several different classification algorithms to train our models by using the 'train()' function from the Caret package. These algorithms include logistic regression, linear discriminant analysis, quadratic discriminant analysis, naive Bayes, and k-nearest neighbours.

First, we conducted exploratory data analysis to understand the relationships between our variables. Once this was done, we divided the data into training and testing sets to train and evaluate the performance of our models. We evaluated the performance of our models by calculating different metrics like accuracy, precision, sensitivity, specificity and F1 score. To further evaluate the performance of our models, we created ROC curves for the different models we created and calculated the area under the curves (AUC). The AUC will give us a measure of how well the givenmodel can classify between positive and negative cases. By using these different metrics, we can gain valuable insights into the performances of the models and help us in selecting which one will be best for our problem.

We used logistic regression since it is a simple and efficient algorithm that can be used for binary classification problems (see if a patient has heart failure or not). Linear discriminant analysis was used to separate the classes linearly by projecting the data onto a low dimensional space, and Quadratic discriminant analysis was used to allow more flexibility. Since our data has a lot of features, it was a good idea to use a naive Bayes Algorithm, since it works for high-dimensional data. Finally, the last algorithm was the k-nearest neighbours algorithm, which classifies a new observation based on the majority class of the k-nearest neighbours.

Other models could have been chosen for this problem, such as decision trees or support vector machines, but they may be too complex for our data set. In addition, creating a decision tree may not be the best idea because our data set has many features, and decision trees tend to over fit on high-dimensional data.

4 Results

As stated prior, the aim of the research is to classify heart disease risk based on the series of variables elaborated upon previously. To begin, it is important to see the summary statistics associated with each variable, with the below figure describing these statistics:

```
      Age             Sex          ChestPainType   RestingBP      Cholesterol
 Min.   :28.00   F:193   ASY:496   Min.   :  0.0   Min.   :  0.0
 1st Qu.:47.00   M:725   ATA:173   1st Qu.:120.0   1st Qu.:173.2
 Median :54.00           NAP:203   Median :130.0   Median :223.0
 Mean   :53.51           TA : 46   Mean   :132.4   Mean   :198.8
 3rd Qu.:60.00                     3rd Qu.:140.0   3rd Qu.:267.0
 Max.   :77.00                     Max.   :200.0   Max.   :603.0
   FastingBS       RestingECG      MaxHR       ExerciseAngina    Oldpeak
 Min.   :0.0000   LVH   :188   Min.   : 60.0   N:547          Min.   :-2.6000
 1st Qu.:0.0000   Normal:552   1st Qu.:120.0   Y:371          1st Qu.: 0.0000
 Median :0.0000   ST    :178   Median :138.0                  Median : 0.6000
 Mean   :0.2331                Mean   :136.8                  Mean   : 0.8874
 3rd Qu.:0.0000                3rd Qu.:156.0                  3rd Qu.: 1.5000
 Max.   :1.0000                Max.   :202.0                  Max.   : 6.2000
 ST_Slope   HeartDisease
 Down: 63   0:410
 Flat:460   1:508
 Up  :395
```

Figure 1: R output describing summary statistics of variables

From this output, it appears that some of the data may be erroneous. For instance, the minimum value of the RestingBP variable is 0. This is a very questionable result, as a resting blood pressure of 0 would practically indicate a deceased individual being evaluated. So, the entry containing this erroneous result was deleted in order to maintain logical results (which also matches analyses of the datasets conducted by other parties). In addition, there appears to be inconsistencies in the measurements across the Cholesterol variable: 172 of the entries have a value of 0. However, as mentioned prior, this is an aggregation of multiple datasets. It is entirely possible that some of these datasets do not measure the cholesterol level, or they are measuring the level in a different way. This will be drawn as a weakness in the dataset, and not considered further in terms of data cleaning.
Now, we will examine the correlation between variables in the dataset, using the corrgram function:

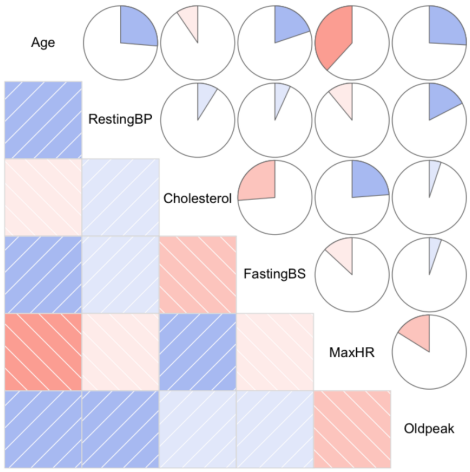


Figure 2: A correlation plot of all variables in the dataset

The correlation plot is divided between upper panel which indicates the magnitude of correlation based on the filled portion of pie, and lower panel which indicates the magnitude of correlation based on the depth of shading. As is apparent, there exists strong correlations between some of the variables. All of these relationships are certainly explainable, with medical phenomena outlining the correlation.
Next, we will visualize the relationship between variables by plotting and reflecting on box-plots below:

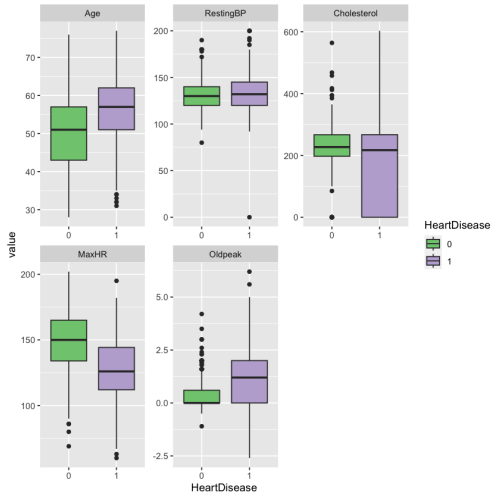


Figure 3: A boxplot representation between variables in the dataset

The image above includes a five different box plots that are labelled with heart disease on the horizontal axis, and value on the vertical axis for 5 different variables - Age, RestingBP, Cholesterol, MaxHR, Oldpeak. etc. It is evident from the box plots that the variables include outliers which are far off from the average results of heart disease. In addition, one specific variable of interest is "RestingBP" which has the same distribution for both classes of heart disease. It is evident that the box plot for "RestingBP" includes 2 classes which has quite similar characteristics. These include almost identical whisker plots placed within the range of 120 and 145 with very similar estimated median value of approximately 127, and possible outliers both above and below the whisker plots with identical placements. Thus, it makes sense to remove "RestingBP" since there does not appear to be a separable plane for the variable. As a result, this explains the output of "RestingBP" will be of no use to determine the value of heart disease.

Furthermore, we will look at the ROC curves for various models which shows the trade-off between sensitivity and specificity:

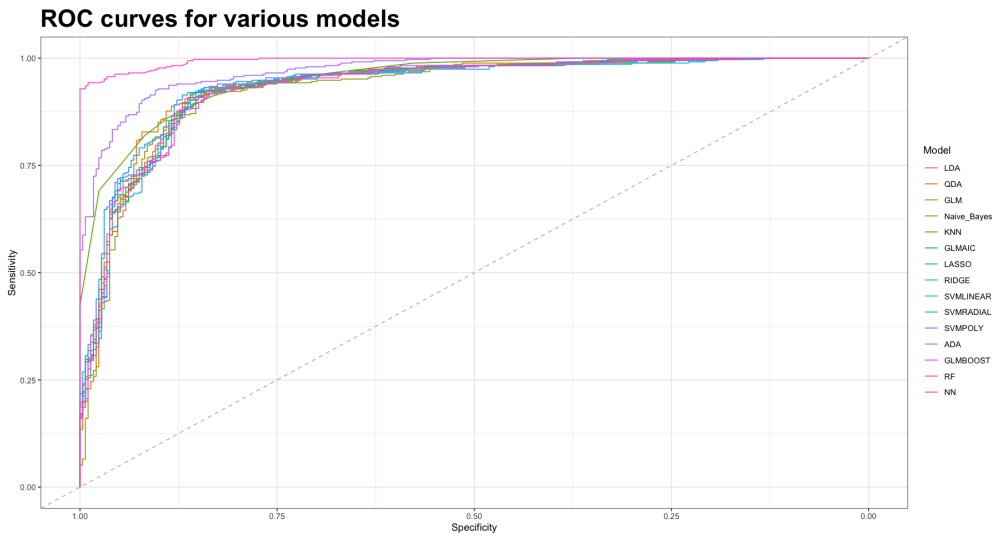
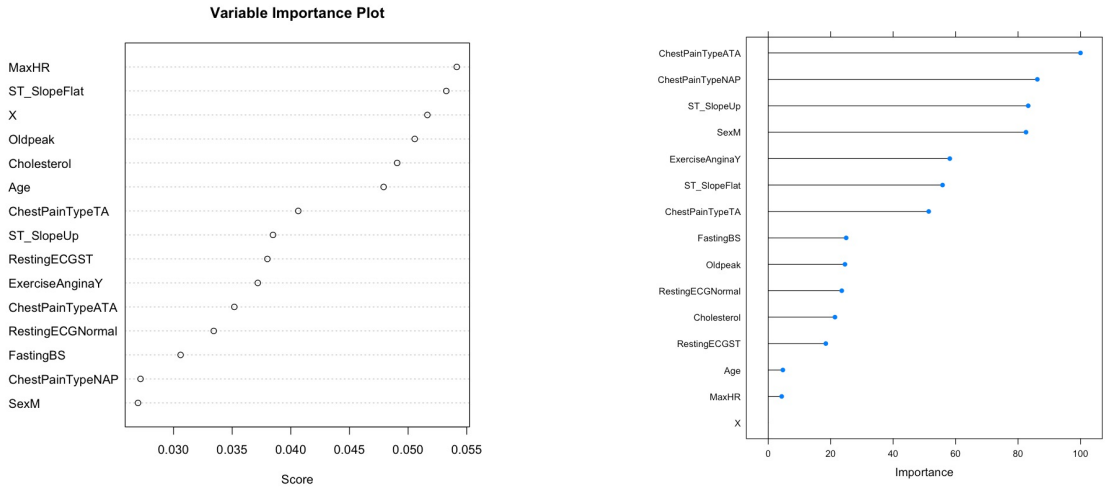


Figure 4: ROC curves for various models

To begin with, any classifier that produces a curve close to the top-left area indicates a good performance, whereas a classifier that produces a curve close to the 45-degree line is said to indicate a bad performance, resulting in a less accurate test. Similarly, the closer the curve is to the upper-left corner, the closer the AUC (Area under the curve) is to 1. For this matter, "ADA" and "NN" models are closest to the upper-left corner as compared to other models in the plot. This further explains that "ADA" and "NN" models have an AUC value close to 1. This indicates that these models have the highest area under the curve, and therefore are the best models for evaluating classifiers that predict factors contributing to heart disease. In contrast, the "SVMRADIAL" model is closer to the diagonal line and represents an AUC value of 0.5. This model is as good as any model making random classifications. Thus, "SVMRADIAL" is one of those models that is not good for evaluating classifiers for prediction.

In accordance with the discussion above, it appears (based on ROC) that the NN and ADA models are the most suitable models for the conducted classfication. Now, in alignment with the audience for this type of research, it is important to answer the question - which variables are the most influential in predicting heart disease risk? To answer this, importance plots were created for both of these models, seen below.



(a) ADA

(b) Neural Network

Figure 5: Variable importance plots for suitable models

As we can see, these models produce drastically different results when it comes to the importance of classification variables. Both of the leading variable results make sense - with respect to the ADA model, if someone suffers from tachycardia (being a heart rate of ≥ 100 BPM) [6] they are likely to have an issue with their heart function. With respect to the neural network, it follows quite naturally that the type of chest pain experienced would serve as a proper indicative of someone's heart condition.

Of more interest, however, is the disagreement between the rest of the variables used in the model. For instance, consider the ranking of serum cholesterol levels. They are ranked as the fifth *most* important variable in the ADA model, however they are ranked as the fourth *least* important in the neural network. This is both of considerable interest and concern.

It is interesting to see how the two most accurate models disagree in terms of their variable rankings. Even when looking at the most important variable in the ADA model (being MaxHR), it is ranked dead last in the neural network model. Similarly, the rankings for chest pain types are not consistent across the board. It goes without saying that comparable results driven by differing routes is of interest to those using the data for a similar purpose. However, there is also cause for concern - what is truly the best predictor for heart disease risk? Seeing as the models disagree, how can we determine what medical professionals should focus on when it comes to lessening the impact of heart disease on society? Setting aside which variable is ranked higher, it seems that the whole ordering is inconsistent. There is very little agreement between the models - so, what should the focus be? In addition, the CDC states that there are three health conditions that best indicate the risk for heart disease: high blood pressure, high cholesterol, and diabetes. Diabetes is not even considered in the model - while the blood sugar variable may serve as a related indicator, it would clearly be helpful to understand other medical conditions the patient may have. Additionally, the two best indicators (per the CDC [7]) are not listed as particularly important variables in the classification. In prior analysis, it was demonstrated that RestingBP should be omitted from the model, due to its statistical properties. So, simply due to its lack of statistical

significance, we potentially miss the best possible indicator for heart disease risk [6]. While that is of concern, it is a general recommendation to keep your blood pressure low through your lifestyle choices. A similar sentiment is echoed for cholesterol - simply adjust your lifestyle to lower your cholesterol, and get your lipids checked whenever possible or necessary. However, this does not explain the inconsistency. Further research is certainly required, but it is much outside the scope of our analysis.
Overall, the models appear to perform extremely well, however compare rather interestingly to the medically recommended measures for heart disease mitigation.

5 Conclusion

Through this report, multiple different models have been proposed for the classification of heart disease risk using the Heart Failure Prediction dataset. Models were compared using Receiver Operator Characteristic (ROC) curves. While the more complicated models fit better than the less complex, interpretable models, there may be some contexts in which a simpler model is preferred for medical research. However, there is a noticeable gap between the results of our research and the medical literature - features which are crucial in detecting heart disease risk are not deemed as "important" by the most suitable models.
This could be drawn up to a number of factors: the difference between statistical and real-world significance, the lack of context between the physical condition and what is simply a number, and so on and so forth. This is certainly of great cause for concern - how can we ensure that the statistical analysis matches the real-world analysis of disease prevention? Such an question is beyond the scope of this report - However, it is most certainly an area of interest to research further.
The advancement of the medical sciences has been amazing to see - some decades ago, humanity could not have even thought to cure the deadliest diseases that plagued our ancestors. Now, pandemic-associated vaccines are readily developed by year-end, and medical diagnostics have come farther than ever. This is all possible through a number of different research methods - however, it is truly ridiculous that even undergraduate students can produce insightful classification algorithms that can determine, with high testing accuracy, whether or not someone is at risk for developing heart disease.

A Delineation of Work

- **Senad Kokic** - Abstract, Introduction, Results, Conclusion
- **Melissa Hayalioglu** - Code (see GitHub)
- **Mariam Lo** - Data, Methodology
- **Kunal Vaid** - Results

References

[1] World Health Organization. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1, 2023.

[2] Centers for Disease Control and Prevention. Preventing 1 Million Heart Attacks and Strokes. https://www.cdc.gov/vitalsigns/million-hearts/index.html, 2018.

[3] Shakil Admed Chohan et. al. Long-term complications of stroke and secondary prevention: an overview for primary care physicians. 2019.

[4] Mercy Health. Hypotension (Low Blood Pressure). https://www.mercy.com/health-care-services/heart-vascular/conditions/low-blood-pressure, 2023.

[5] Heart Failure Prediction Dataset. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction, 2021.

[6] American Heart Association. Tachycardia: Fast Heart Rate. https://www.heart.org/en/health-topics/arrhythmia/about-arrhythmia/tachycardia--fast-heart-rate, 2022.

[7] Centers for Disease Control and Prevention. Know Your Risk for Heart Disease. https://www.cdc.gov/heartdisease/risk_factors.htm, 2023.

[8] American Heart Association. CDC Prevention Programs. https://www.heart.org/en/get-involved/advocate/federal-priorities/cdc-prevention-programs, 2018.