What Causes Heart Disease?

Bram Mulder
Willem Blokland

January 9, 2022

Abstract

Now more than ever, an immense amount of data is being collected from and about people all over the world. With all this data it is now possible, for example, to predict certain diseases well, which is especially useful in the medical world. This report takes on the question of 'what causes people to have a heart condition' and tries to explain how this can be concluded. Using some machine learning methods, we look at what attributes have a lot, a little or no influence at all on getting such a disease. When looking at some previous research on this topic ,the random forest classifier seemed like a fine method to see how the different attributes in our data set affect the target. This report also covers looking at the attributes individually and seeing how the risk of a heart disease increases or decreases depending in how the values of the attributes change. With all this information the reader can inform themselves and of course others about this subject and thus try to be a little more aware and healthier with their body.

Contents

1 Introduction

Many people would like to live a long life, but a person does not always have that fate in their own hands. A common cause of death is heart disease. As of 2018, 30.3 million U.S. adults were diagnosed with heart disease(Thomas, 2020), looking at the whole world, this is of course much more. The reason this number is so high is because a heart disease is not always noticeable, some complaints of a heart disease can also belong to other things. It would be ideal to come up with something that could perfectly determine whether someone has heart disease. This makes the application domain of this report people with certain complaints associated with heart disease and the research problem determining when someone has heart disease.

There is some related previous work that attempted to address the same problem. A lot of different machine learning algorithms and techniques are used. Some examples of different techniques that are used are: random forest classifier, support vector machine, k-nearest neighbor algorithm and neural Network. The accuracy score is not the same everywhere. (S Anitha, 2019) study shows that the random forest has an accuracy of 80%, which is lower than the Neural Network (83.5%) and the Support Vector Classifier (84.0%). These numbers will differ per data set and per study, nevertheless this report will examine whether it is possible to increase the accuracy of the random forest classifier.

The hypothesis is as follows: The attributes chest pain and number of major vessels are thought to be important factors on the outcome of the model. Chest pain is very common just before a heart attack and the number of major vessels is very important if a vessel becomes blocked. The more vessels, the smaller the chance that all vessels are clogged. The attribute that is less thought of as important is sex, it doesn't seem to be that a particular gender is very much more likely to have a heart attack. The accuracy of the classification model should reach a value above the 80% mark if we use the right pre-processing and take a closer look at each of the attributes in the data set.

2 Data set And Data collection

The data set consist of 1025 samples and 14 columns. In the fourteenth column the information is given whether there is a heart disease, so there are 13 attributes. The data set is relatively simple to understand, but the meaning of some of the attributes are not that clear. For this reason a number of headers have been renamed, below is an overview of the new headers as well as the meaning and the possible values for every attribute:

	meaning	**new name**	**possible values**
age	person age in years		(29-77)
sex	persons sex		1=male, 0=female
cp	chest pain type	chest_pain	1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptomatic
trestbps	resting blood pressure (in mm Hg)	rest_bloodpressure	(94-200)
chol	serum cholesterol (mg/dl)	cholesterol	(126-564)
fbs	fasting blood sugar	fasting_bloodsugar	1=true, 0=false
restecg	resting electrocardiographic results	rest_elec_cardio_graphic	0=normal, 1=ST-T wave abnormality, 2=left ventricular hypertrophy
thalach	maximum heart rate achieved	max_heart_rate_achieved	(71-202)
exang	exercise induced angina	exercise_induced_angina	1=true, 0=false
oldpeak	ST depression induced by exercise relative to rest	st_depression	(0-6.2)
slope	the slope of the peak exercise ST segment	st_slope	0=upsloping, 1=flat, 2=downsloping
ca	the number of major vessels	number_major_vessels	(0-4)
thal	a blood disorder called thalassemia	thalassemia	1=normal, 2=fixed defect, 3=reversable defect

After observing the attributes, it was decided to remove the age attribute from the data set. This is because there are very few young people, the majority of whom have a heart disease. If this attribute were to be included, a distorted picture could arise that young people have a higher risk of heart disease. The reason that there are very few young people is probably due to selection bias in the study and due to the fact that the entire data set is relatively small.

3 Approach / Methods

To check the hypothesis, it will be necessary to look at the general accuracy of the model and to zoom in a little more on the attributes themselves to determine which have a large effect on the accuracy. After the data is pre-processed, it is divided into two variables: train and test. It was decided to make

the ratio between these two variables 80% and 20% because it is a quite commonly occurring ratio.

First, a tree will be created by means of a random forest classifier. In the model of the RandomForestClassifier, a maximum depth of 5 has been chosen. When determining the maximum depth, overfitting must be taken into account. Overfitting can lead to a model that does not generalize well from the training data to unseen data. After a number of tests with adjusting the maximum depth, it appeared that from a depth of 6 there is a clear sign of overfitting, for this reason a depth of 5 was chosen. After generating the tree, a confusion matrix will be used to determine the sensitivity and specificity. It was decided to measure the accuracy of the model partly in sensitivity and specificity because this provides more information than an accuracy score. Sensitivity gives insight in the proportion of actual positive cases that have gotten predicted as positive by our model and the specificity gives insight in the proportion of actual negative cases that have gotten predicted as negative by our model. After the sensitivity and the specificity scores, a ROC-curve was made, showing the performance of the classification model at all classification thresholds. This curve plots the True Positive Rate and the False Positive Rate. Associated with the ROC-curve is the AUC score, which indicates the size of the area under the curve. With the sensitivity, specificity and AUC score a complete indication is given about the accuracy of the model.

To get a better look at the individual attributes, a number of graphs are used. First, looking at partial dependence plots that show how the probability of heart disease changes when the value of the attribute changes, for example when the maximum heart rate varies. This is very useful in determining which variables have a positive or negative influence on the risk of a heart disease. Secondly, a graph is used that clearly states which attributes have the most impact on the output, so it is clear which attributes are the most and least decisive in determining whether someone suffers from a heart disease. This graph uses Shap values.

4 Main Results

The following results have been determined with a random state of 4. A random state ensures that the results can be reproduced. The first two results that can be seen are the decision tree of the random forest and the ROC-curve.

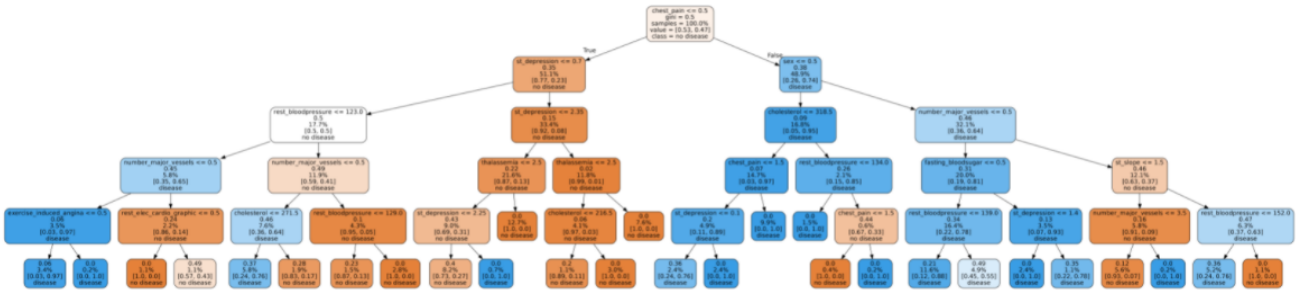


Figure 1: Decision tree

Below are the scores that determine the accuracy of the model:

	score
sensitivity	0.946236559139785
specificity	0.8303571428571429
Area Under the Curve (AUC)	0.9640473011634559

After the overall results are known, it is time to look at the results related to the individual attributes. First are twelve partial dependence plots, which tell how the risk of heart disease changes when the value of an attribute is adjusted.

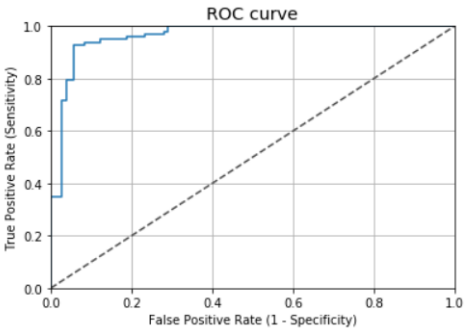
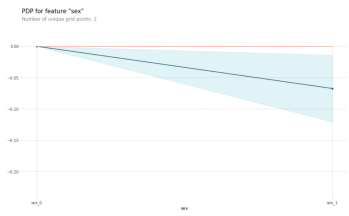
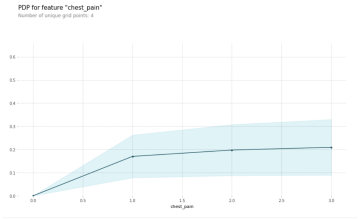


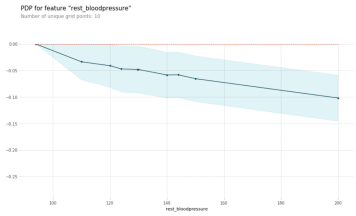
Figure 2: ROC-curve



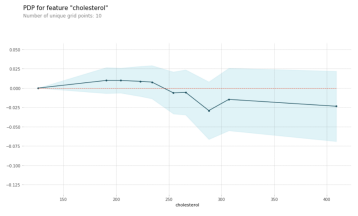
(a) Sex



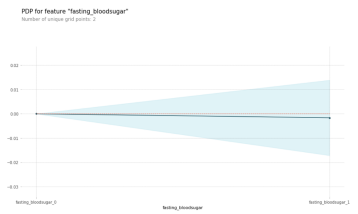
(b) Chest pain



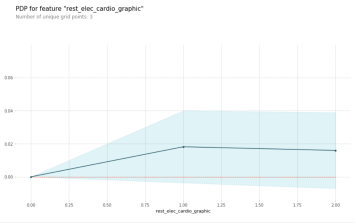
(c) Resting blood pressure



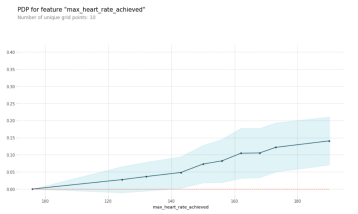
(d) Serum cholesterol



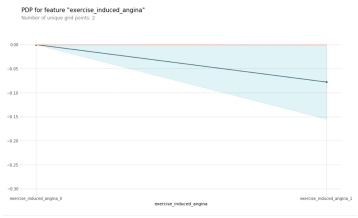
(e) Fasting blood sugar



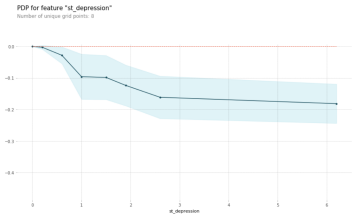
(f) Resting electrocardiographic results



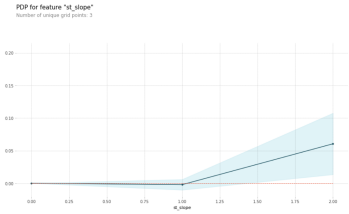
(g) Maximum heart rate achieved



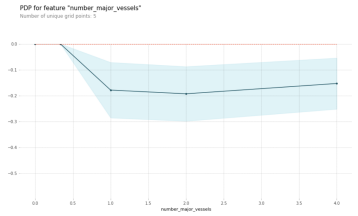
(h) Exercise induced angina



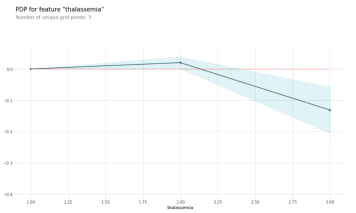
(i) ST depression induced by exercise relative to rest



(j) The slope of the peak exercise ST segment



(k) The number of major vessels



(l) A blood disorder called thalassemia

Figure 3: Partial dependence plots

And finally the graph is displayed showing the attributes that have the most impact on the output:

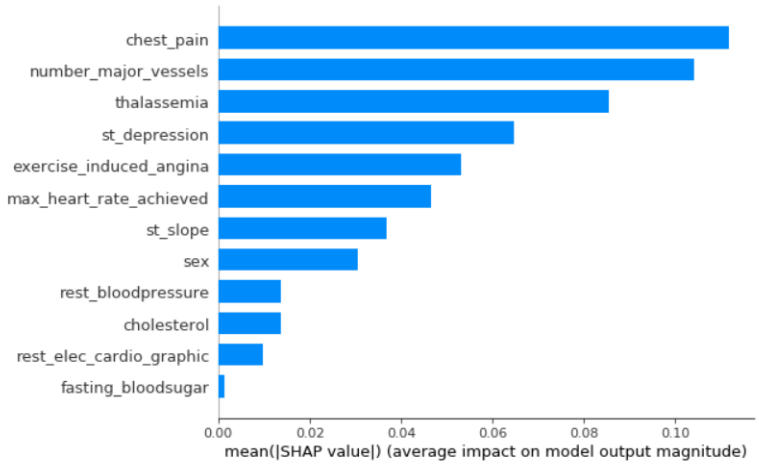


Figure 4: Impact values on output

5 Analysis and evaluation

5.1 Overall results

The overall results include the tree, ROC-curve, sensitivity, specificity and AUC score. The tree looks fine and the sensitivity and specificity scores are high. For the AUC score it applies that a model has a score of 0.0 if the predictions are 100% wrong, a model whose predictions are 100% correct has a score of 1.0. The ROC-curve looks good and has a good AUC score: 0.964. This means that the performance of the classification model at all classification thresholds is excellent.

5.2 Individual attributes

In order to keep it clear, it has been decided to treat the 12 partial dependence plots in list form. The first item in the list belongs to the first plot, and so on.

1. The sex attribute clearly shows that women have about 6 to 7 percent less chance of heart disease.
2. The chest pain attribute clearly shows that more chest pain leads to a higher risk of heart disease, which sounds quite logical.
3. For the resting blood pressure, it can clearly be seen that as the pressure increases, the risk of heart disease decreases. A higher blood pressure ensures that the blood can be pumped around the body better. This can also be seen in the plot.
4. The plot for cholesterol is slightly more variable. In the beginning, the risk of heart disease increases as the amount of cholesterol increases, but from a certain point the risk of heart disease decreases as the amount of cholesterol increases even more. According to the graph, the ideal amount is somewhere between 275 and 300 mm Hg.
5. According to the plot, it makes little difference to the risk of heart disease whether there is fasting blood sugar. If it is the case, the risk of heart disease is slightly smaller.
6. If the values of the resting electrocardiographic results deviate from normal (value=0), the risk of heart disease is greater.
7. According to a medical school(*Harvard medical school*, 2021), heart rates that are above 100 can sometimes be caused by an abnormal heart rhythm. A high heart rate can also mean the heart muscle is weakened by a virus or some other problem. The plot clearly shows that a maximum heart rate above 100 increases the risk of heart disease, which makes perfect sense.

8. Exercise-induced angina reduces the risk of heart disease. So if there's no exercise induced angina, there's something else causing the pain, which could be heart disease. So this plot also makes sense.

9. According to the plot, a person is less likely to have heart disease if that person suffers from depression. This is an interesting outcome.

10. If the slope is descending, the risk of heart disease is greater. A lower heart rate after exercise is very illogical, so it makes sense that something is wrong with the heart.

11. It is clear in the plot that as the number of major blood vessels increases, the risk of heart disease decreases. This makes sense, the more vessels the more blood can get to the heart.

12. In the case of a fixed defect, the probability of heart disease is, according to the plot, slightly higher than normal. When there is a reversible defect, the chance is much smaller.

The plot indicating which attributes have the most impact is fairly self-explanatory. The three attributes that have the most impact on the outcome of our model are: chest pain, number of major vessels and thalassemia. These three attributes also had a clear rising or falling line in the partial dependence plots. In the hypothesis it was predicted that the attributes chest pain and number of major vessels would be important on the outcome, this is correct. It was also predicted that the attribute sex would not be very important. The results show that sex is in eighth place, so this part of the hypothesis is also correct.

6 Future Directions

There are a number of things that could be done next time to improve research on this topic. The first is to find (or create) a better data set. The data set used in this report was not optimal because it contained too few samples and too many attributes, this could result in overfitting. Also, attributes such as age were not suitable to work with because there were too many older people in the data set compared to young people, which was probably due to selection bias (the inclusion of people in a study to systematically select participants for whom the intervention to be studied will have more effect, in this case older people).

In the future it would also be good to look at multiple different classification methods and see how they perform on their own and against each other. An example of some classifiers are K-nearest Neighbors or Support Vector Machines. A better and a more modern data set combined with multiple classification methods should provide a good approach for later research on this topic, which will become increasingly important as machine learning has and will play a greater and more significant role in health care.

7 Insight to code

More explanation to our code and results is given in the 'DataMiningReport.ipynb' file which is uploaded to our github. The link to to the github respitory:
https://github.com/WillemBlokland/datamining-report
The respitory contains the notebook, data set and some other files that are required to run the notebook.

References

Harvard medical school. (2021, Jun). Harvard Medical School. Retrieved from https://www.health.harvard.edu/heart-health/should-i-worry-about-my-fast-pulse

S Anitha, N. S. (2019, Jul). Retrieved from https://webcache.googleusercontent.com/search?q=cache:jP_qkwbOnEEJ:https://hal.archives-ouvertes.fr/hal-02196156/document&cd=17&hl=nl&ct=clnk&gl=nl

Thomas, J. (2020, Jul). *Facts and statistics on heart disease.* Healthline Media. Retrieved from https://www.healthline.com/health/heart-disease/statistics#Who-is-at-risk?