Using machine learning to diagnose early heart disease

Report

Sander J. Bouwman

11/12/2021

Contents

1 Introduction

Currently, heart disease is the number one leading cause of death, with more than 17 million deaths annually (Finegold, Asaria, and Francis 2013). These deaths account for more than 30% of all deaths worldwide. An early detection of possible heart disease could prevent a lot of future deaths. Machine learning could play a significant role in the early detection of heart disease. Currently, heart disease is diagnosed by a combination of blood test, ECG, breathing tests and chest x-rays test. For mass screening, it would be best to provide an accurate prediction of developing heart disease with the least amount of tests. Preferable, tests that can be performed at a general practitioner, which prevents unnecessary hospital visits. The goal of this research was to create an accurate model that uses easy data which could easily be gathered at a general practitioner.

2 Materials & Methods

2.1 Materials

2.1.1 Original data

The data used is a public accessible dataset found at Kaggle. The used dataset can be found at https://www.kaggle.com/fedesoriano/heart-failure-prediction. The author of this dataset is (Fedesoriano 2021). This dataset consists of 5 combined datasets, which makes this dataset the largest heart disease dataset so far. The five datasets used are: -Cleveland: 303 observations
-Hungarian: 294 observations
-Switzerland: 123 observations
-Long Beach VA: 200 observations
-Stalog (Heart) Data Set: 270 observations
The individual datasets can be accessed at: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

The dataset has a total of 11 parameters, which can be explored further in the underlying table (table: 1).

Table 1: Attribute Information

Name	Description	Category
Age	age of patient	years
Sex	sex of patient	M: male, F: female
ChestPainType	chest pain type	TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic
RestingBP	resting blood pressure [mm/hg]	mm Hg
Cholesterol	serum cholesterol [mm/dl]	mm/dl
FastingBS	fasting blood sugar	1: if FastingBS > 120 mg/dl, 0: otherwise
RestingECG	resting electrocardiogram results	Normal: Normal, ST: having ST-T wave abnormality, LVH: showing left ventricular hypertrophy by Estes' criteria
MaxHR	maximum heartrate achieved	Numeric value between 60 and 202
ExcerciseAngina	excercised-induced angina	Y: Yes, N: No
Oldpeak	oldpeak = ST	Numeric value measured in depression
ST_Slope	the slope of the peak exercise ST	Up: upsloping, Flat: flat, Down: downsloping
HeartDisease	output class	1: heart disease, 0: Normal

2.1.2 Data cleaning

Data cleaning was performed using R (R Core Team 2021) with various libraries. An overview of the used libraries can be found in the underlying table (table: 2).

Table 2: Used R libraries

name	version	links
ggplot2	3.3.5	https://cran.r-project.org/web/packages/ggplot2
kableExtra	1.1.0	https://cran.r-project.org/web/packages/kableExtra
scales	1.1.1	https://cran.r-project.org/web/packages/scales
ggcorrplot	0.1.3	https://cran.r-project.org/web/packages/ggcorrplot
reshape2	1.4.4	https://cran.r-project.org/web/packages/reshape2
factoextra	1.0.7	https://cran.r-project.org/web/packages/factoextra
tibble	3.1.4	https://cran.r-project.org/web/packages/tibble
RWeka	0.4.43	https://cran.r-project.org/web/packages/RWeka
ggpubr	0.4.0	https://cran.r-project.org/web/packages/ggpubr

2.1.3 Model creation

Weka (version 3.9.5) (Hornik, Buchta, and Zeileis 2009) was used for the creation of the machine learning model.

2.1.4 Java wrapper

A Java library by Hornik, Buchta, and Zeileis (2009) was used for the creation of a Java application. Which makes it possible to use the model using an CLI.

2.2 Methods

Data cleaning was performed in R. Most categorical data was split out into logical attributes. For example; HeartPainType was split into 4 seperate columns. The logbook created for this research can be found on GitHub at https://github.com/devalk96/Thema09. PDF file can be found using this direct link: https://github.com/devalk96/Thema09/blob/main/log.pdf. The resulting Java implementation can also be found on GitHub at https://github.com/devalk96/Thema09-JavaWrapper.

3 Results

3.1 Cleaning

Datacleaning was done on multiple variables. Column ChestPainType originally categorical data was split out into 4 new columns containing logical data. This was also performed for the ST_Peak column which was split into 3 new colums containing logical data. There were no missing datafields.

3.2 Heart disease (HD)

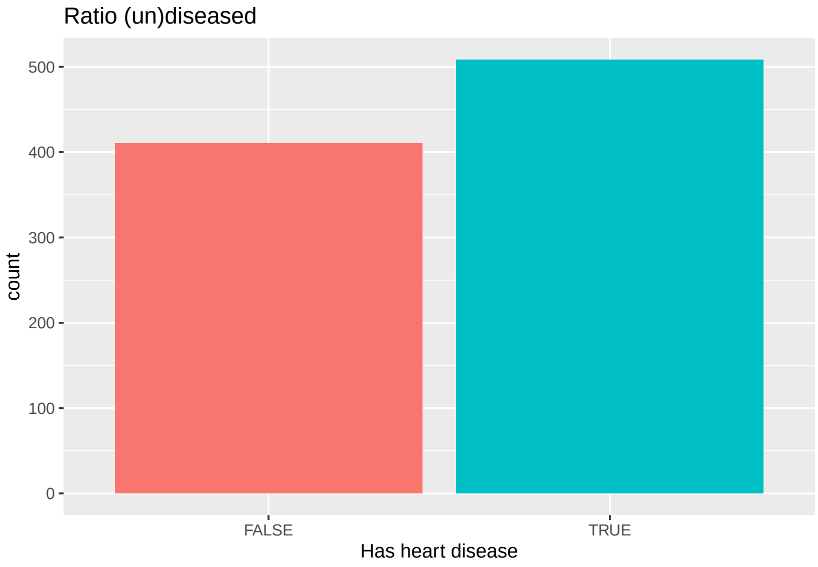

Figure 1: Count of heart disease

The ratio patients having HD is 55% (n=508) 1 while 45% (n=410) has no HD, altough there is a slight balance but is still acceptable.

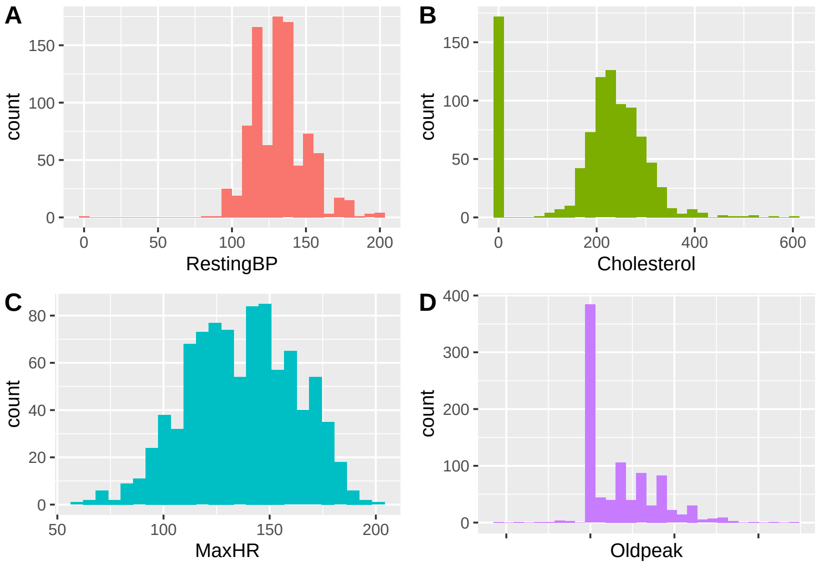

Figure 2: Four histograms of restingbp, cholesterol, maxhr and oldpeak

Figure 2 shows that data is quite normally distributed. Altough there are some values with very high counts. Notable cholesterol (B) with a very high count of 0 mm/dl (n = 172) and oldpeak (D) with a very high count of 0 ST.

3.3 Excercise angina

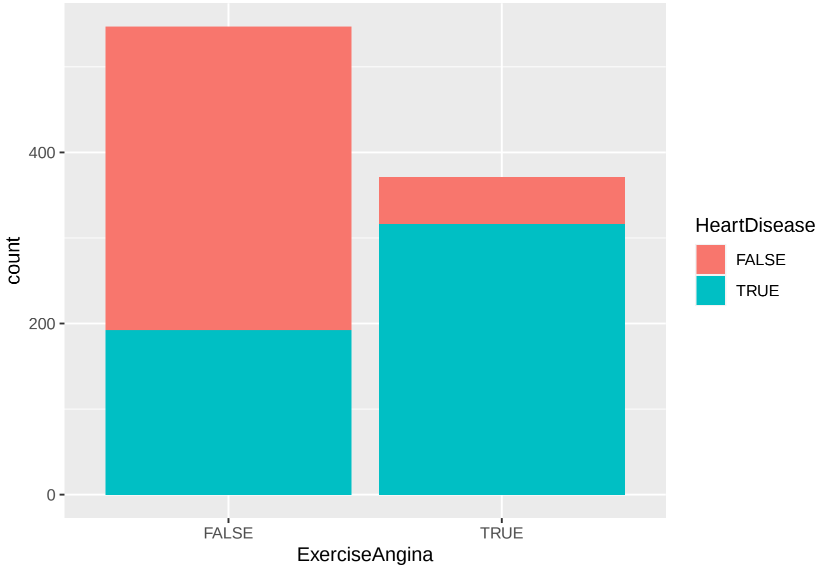

Figure 3: Effect of exercise angina on HD

In figure 3 it is visible that having excersise angina has a much higher chance of having HD in comparison to not having excercise angina.

3.4 ST-Slope

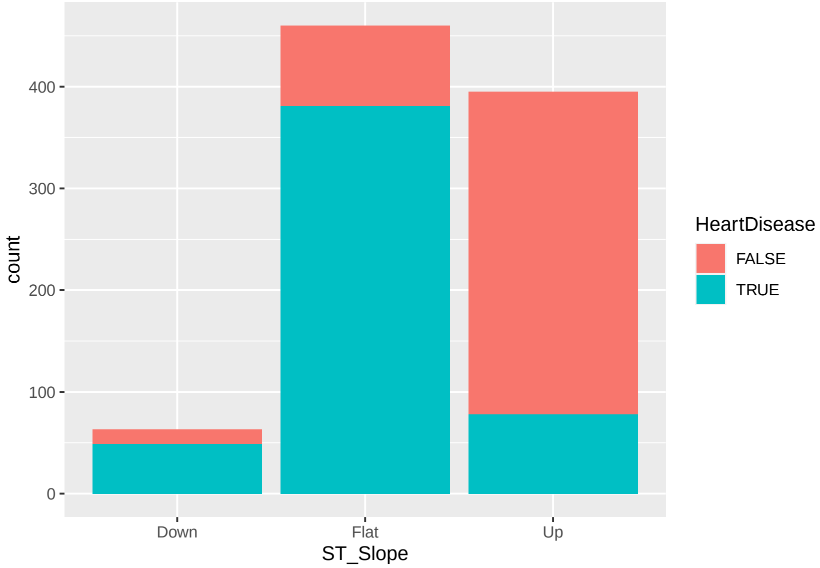

Figure 4: Relationship between heart slope and heart disease

Figure 4 shows that having an UP ST slope depression significantly lowers the chance of HD, when comparing this with a down or flat slope. Both down and flat ST slopes give a much heigher chance of having HD. ST slope down is fairly rare in this dataset.

3.5 Age

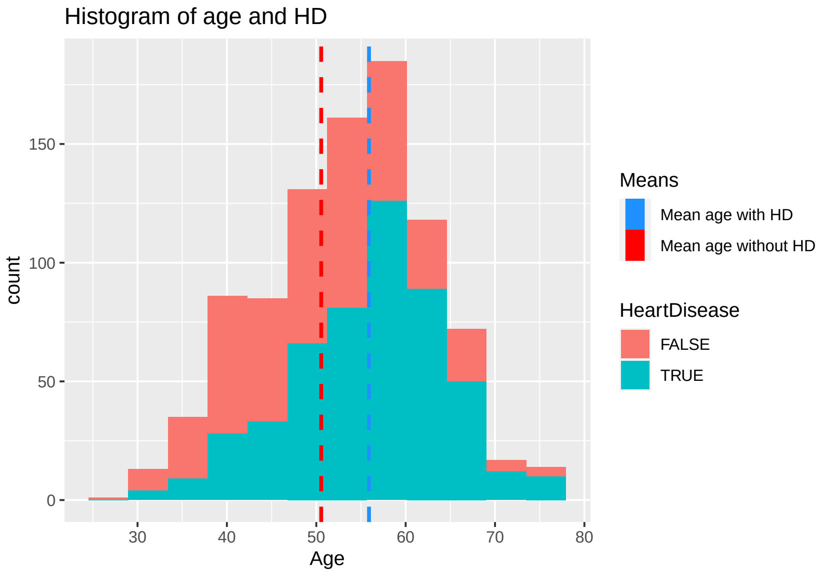

Figure 5: Histogram of age grouped by HD diagnoses

In the figure above (figure 5) it is visible that the percentage of patients with HD shifts to the majority with age. After the age of around 70, this percentage seems to shrink again. Using the dotted vertical lines, we can see that there is a difference in mean age in persons with or without HD. The mean age of persons with HD is ~56 years, whereas the mean age of persons without HD is ~51 years. The histogram is itself shows a decent normal distribution in age in the dataset. To further explore the possible relation/correlation between age and the prevalence of HD, another plot is used.

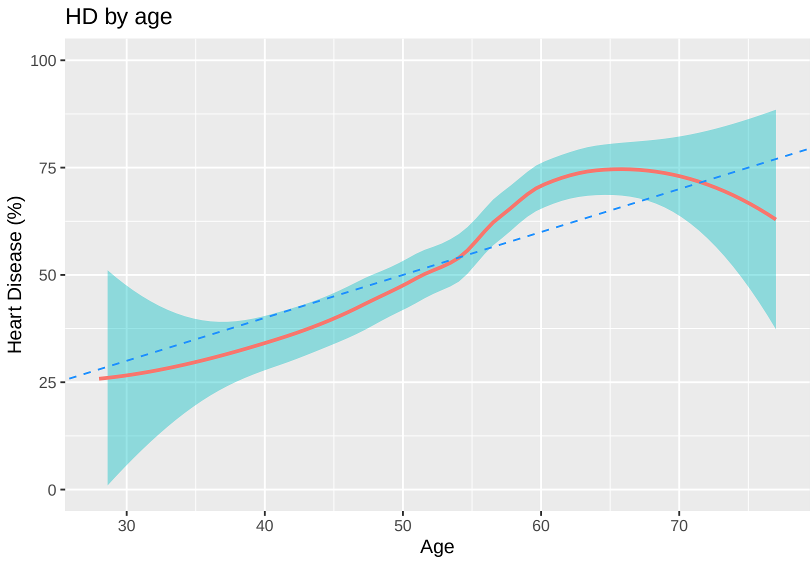

Figure 6: Percentage of patients affected by HD by age

Figure 6 shows the relation between age and prevalence of HD. Both ends show a major wide area of heart disease (percentage) while the overall trend shows a clear increase of HD with an increase of age. The percentage of patients with a positive HD diagnosis at age 30 is around 25%, whereas patients age 65 have a 75% positive diagnosis rate. As visible both figure 1 and in figure 2 it is visible that the percentage of positive diagnosis lowers after age 65.

3.6 Sex

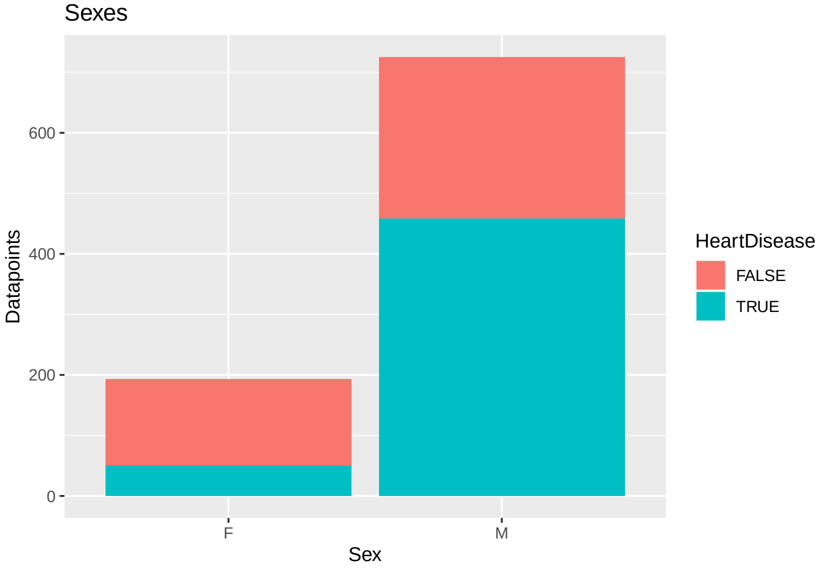

Figure 7: Amount of instances by sex grouped by HD diagnosis

As visible in the above figure (figure 7) the ratio of female(21.02%) to male(78.98%) is very skewed towards males. The ratio of a positive diagnosis is obviously different. Males tend to have on average a positive HD diagnosis, whereas females have a negative diagnosis.

To further visiualize if the density of age with sex is roughtly the same the underlying figure was generated.

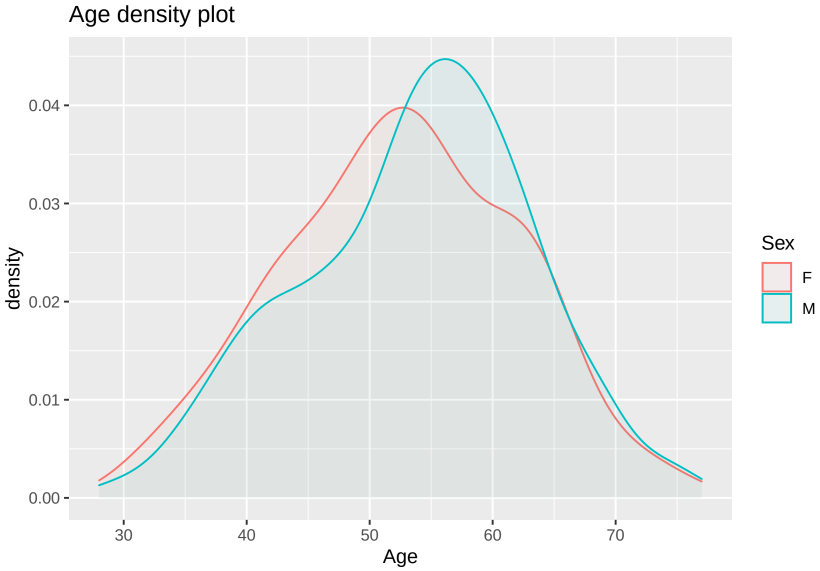

Figure 8: Density plot of ages grouped by sex

Figure 8 shows that both sexes have an about even amount of age groups. Where males have a higher representation in the ages 55+ years, and females have a larger (relative for females) representation lower than 55 years.

3.7 Pain types

To bring an easier understanding of different chestpain types (CPT), the underlying table can be used. Pain type:

Code	Type	Explanation
TA	Typical Angina	Meets all three of the following characteristics: 1. Substernal chest discomfort of characteristic quality and duration 2. Provoked by extortion or emotional stress 3. Relieved by rest and/or nitrates within minutes.
ATA	Atypical Angina	Meets two of the above characteristics
NAP	Non-Anginal Pain	Lacks or meets only one or none of the characteristics
ASY	Asymptomatic Pain	

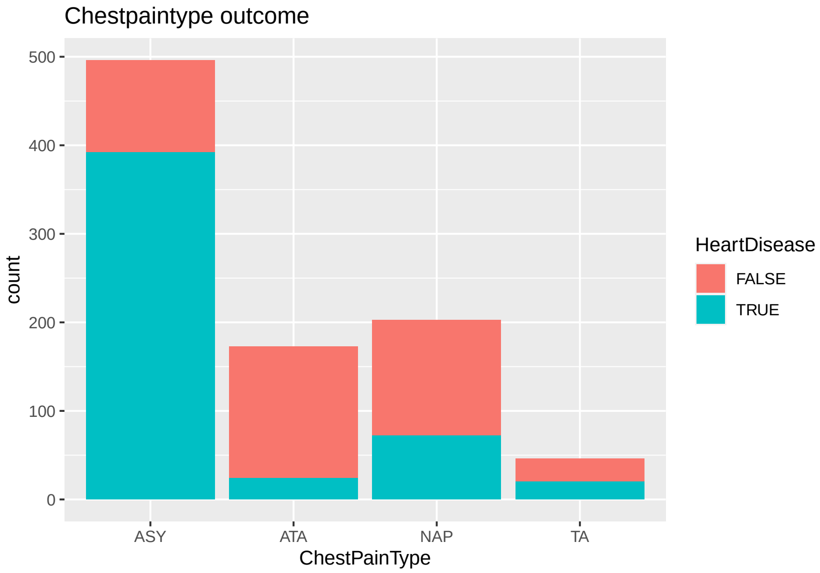

Figure 9: Ratio of diagnosis grouped by chestpain type

Note: For the creation of the PCA plots the categorical variable ChestPainType was split into multiple numerical values containting columns.

As visible see in the above figure (figure 9) the uttermost CPT with HD is asymptotic pain. After that Non-Angical Pain is the highest although more patients with NAP have no diagnosed HD, the same can be seen in TA where the ratio between having and not having HD is roughly the same as in NAP. ATA has by far the lowest ratio of diagnosed HD.

3.8 PCA plotting

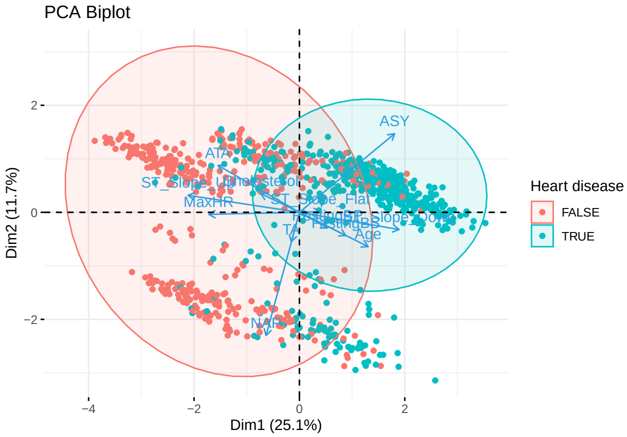

Figure 10: PCA Biplot of all non-categoral variables

In the above figure (figure 10) two clusters are visible. The postive (turqoise) cluster seems to be in correlation with ASY (asymptomic pain). Whereas there are two distinct negative clusters. The left upper cluster is affected by ATA, Max heartrate and cholesterol, while the left down clusters is effected by NAP.

To further see the contribution of each variable the underlying figure was generated.

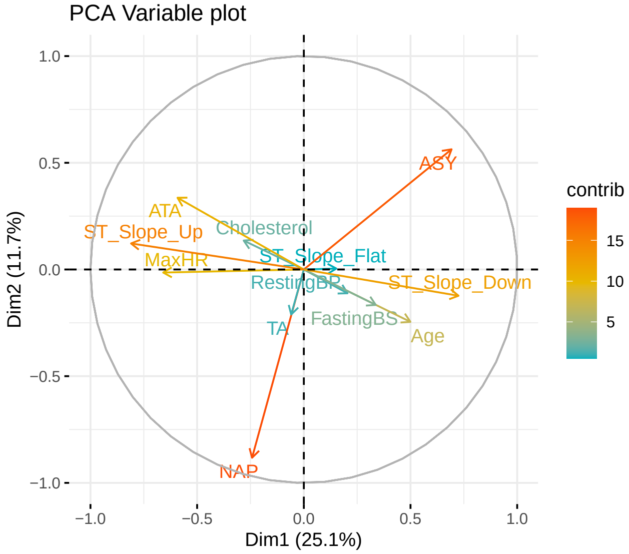

Figure 11: PCA Variable plot of all non-categoral variables

Figure 11 shows that both ASY and NAP are the biggest contributors between a positive or negative diagnoses. As previously mentioned, it is also visible here that an ST_Slope up lowers the chance of HD while ST_Slope down has a positive correlation with HD.

4 Discussion & Conclusion

4.1 Discussion

The undiseased to diseased ratio is not fully balanced (0.45 to 0.55 ratio), but this is not critical. There seems to be quite a lot of patients where cholesterol is not measured and defaulted to 0 mm/dl (n=172) (see figure 2). Currently, asymptomatic pain is the major point of differentiation. Asymptomatic pain is not something usable in itself, as it is not measurable by patient nor caregiver. If asymptomatic pain is removed, it might still be possible to create a reliable prediction of HD if an appropriate algorithm is used. But using asymptomatic pain in combination with other variables is might still be used to increase prediction accuracy. Furthermore the data is highly skewed towards males. Of these males a very large percentage (63%) have HD while females have a far lower ratio of heart disease (26%). In the real-world males have a slightly higher chance of heart disease but as big as this dataset implies. For further research, it would benefit if more female data would be added to the dataset with a better HD ratio then the current situation. Looking at figure 10, it seems that cholesterol has a positive effect on HD, a higher cholesterol generally increases the chance of heart disease, which seems counterintuitive with this dataset. The sudden lowering of patients with HD above the age of 65 might be explained by the fact that patients older than 65 simply die due to their HD.

The current dataset seems to be cleaned well enough, depending on the used algorithm some changes might be necessary, factorizing certain variables such as age, cholesterol, max hr and oldpeak might be needed. During EDA, it was found that certain variables such as cholesterol had a high count of the same values, for example cholesterol had a very large count of 0 mg/dl, the dataset source doesn't provide a reason, but it is suspected that this is due to no measurements available and defaulting to 0, the same could be said for the variable oldpeak where there are also a lot of instances with the value 0.

4.2 Conclusion

Age seems a decent indicator in the diagnosis of heart disease. This is confirmed by the PCA plots (figures 10 & 11) where it shows that it is a small but meaningful indicator. Asymptomatic pain seems to be the major indicator in the diagnosis of HD As visible in figure 9, the presence of asymptomic pain is a large indicator of a positive HD diagnosis as 79% percent of asymptomic pain patients have diagnosed HD The PCA variable plot (figure 11) also confirms this as NAP has a large contribution. NAP has a high contribution to a negative diagnosis according to figure 11. Figure 9 confirms this as a large amount of patients experiencing NAP (64.5%) has a negative HD diagnosis. ATA has a low (13.9%) positivity rate of HD, this also shows in the PCA plot (figure 11) as ATA has relative smaller overall contribution. Having a UP ST slope depression lessens the chance of HD significantly. It also seems that non-numerical variables have a higher contribution than numerical variables.

As visible by the data, it is probable that it is possible to predict heart disease with fair accuracy on this dataset, although it is not probable that it is accurate on real-world data as there are significant differences between the used dataset and real-world data.

5 Project proposal for minor

There are multiple possibilities in which this research can be improved. ## Current issues: The current Java application is not very user-friendly in the sense that it is a desktop CLI application, and demands some basic knowledge about command line interfaces and version control using GIT. A web application would remove this hurdle and make it more accessible. Another point in which the software could be improved is the input data. The current input demands hard coded variable string or a .ARFF file. Making the input more dynamic would be easier to work with.

5.1 Goal

There are several posibilities in whichh the user friendliness might be improved. A webapplication in combination with a web API would the model very dynamic and could be used in most enviroments. Health practitioners could use the tool to gain a quick insight in the current risk for heart disease in a patient. Larger hospitals would be able to use an API to check a large amount of patients fast. Altough a simple offline API would also suffice in the last mentioned situation.

5.2 Target audiance

The target audiance would be mainly first line health care providers such as general practionors. The main appliance of this tool would be a quick assessment in the possible change of (developing) heart disease. This would decrease second line healthcare demand as less patients would need further testing in hospital.

5.3 Design

As previously stated, the current input is very strict and not dynamic at all. The current output is also not very user friendly as a simple output is provided. This simple output might not be trusted by users, because there is no clear why/how to the output. An increase in details and info of the output would make the output more trustworthy. For example, if the algorithm is switched towards a tree model where a clear decision tree is constructed. This decision tree shows why a patient is flagged as positive or negative.

References

Fedesoriano. 2021. "Heart Failure Prediction Dataset." https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Finegold, Judith A., Perviz Asaria, and Darrel P. Francis. 2013. "Mortality from Ischaemic Heart Disease by Country, Region, and Age: Statistics from World Health Organisation and United Nations." *International Journal of Cardiology* 168 (2): 934–45. https://doi.org/https://doi.org/10.1016/j.ijcard.2012.10.046.

Hornik, Kurt, Christian Buchta, and Achim Zeileis. 2009. "Open-Source Machine Learning: R Meets Weka." *Computational Statistics* 24 (2): 225–32. https://doi.org/10.1007/s00180-008-0119-7.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.