Report

Rob Meulenkamp

2022-11-08

1 Introduction

Cardiovascular diseases (CVD) is one of the most common death globally. This makes up for 31% of all deaths worldwide. Heart disease is provoked by atherosclerosis. This means the buildup of plaques or fatty deposits in the walls of the coronary arteries in several years. The coronary arteries enclose the outside of the heart and provide blood oxygen and nutrients to the heart muscle. If the plaque builds up in the arteries, there is fewer space for blood to flow naturally and deliver oxygen to the heart. It possibly cause angina (chest pain) or a heart attack. Four out of five CVD deaths are result of heart attacks and strokes. One-third of these deaths appear in people under the age of 70. The goal of the project is to produce an accurate (false negatives <= 5%) machine learning algorithm that predicts the possibility of developing a heart disease in a wide array of patients.

2 Material & Methods

2.1 Materials

The data used for this research can be found at: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?resource=download The code for making the exploratory data analysis and the application can be found at: https://github.com/RobMeulenkamp/thema9

2.1.1 Source data

The dataset was made by combining different datasets together. Several independently datasets are available but were not yet combined into one big dataset. In this dataset, five different heart datasets with eleven common features are merged into one of the biggest heart disease dataset for research goals. The origin of the five datasets are: - Cleveland: 303 observations - Hungarian: 294 observations - Switzerland: 123 observations - Long Beach VA: 200 observations - Stalog (Heart) Data Set: 270 observations

Merged dataset contained a total of 1190 observations. 272 observations were duplicated and removed, thus leaving the final dataset with 918 observations.

The separable datasets can be accessed at: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

The eleven features and the description of it can be seen in table 1.

Table 1: Attribute Information

Name	Description	Category
Age	age of patient	In 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	Normal: Normal, ST: having ST-T wave abnormality, LVH: showing left ventricular hypertrophy by Estes' criteria
MaxHR	Maximum heartrate	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

2.2 Methods

2.2.1 Exploring methods

The exploratory data analysis and data cleaning were done in Rstudio (R version 4.0.4). Machine learning experiments were done in Weka (version 3.8.5). Plenty of algorithms were compared to each other with the goal if they're significantly better in accuracy. To determine if an algorithm is significant better a t-test was used with p-value of 0.05. The base learner ZeroR was applied for this research thus every chosen algorithm were compared with ZeroR for significance. If there's any significant difference between the other algorithms a standard deviation error barplot was made. The standard deviation error bar (black stick) were used to check if there was any overlap between the different algorithms in order to determine the significance in accuracy. The chosen algorithms were: ZeroR, OneR, Random Forest Tree, J48, Naive Bayes, AdaBoostm1, SGD, IBK, and SMO with the thought that all categories were representative. For extra research, the ensemble learners exist of: Voting, Stacking, and Bagging. At last the cost sensitive classification were used to improve the best two performing algorithms. Different cost matrices were used to reduce the false negatives but still maintaining a high accuracy.

The table below shows the used R libraries with the corresponding versions.

Table 2: R libraries in their respective versions

Library	Version
ggplot2	3.3.5
pander	0.6.4
kableExtra	1.3.4
scales	1.1.1
factoextra	1.0.7
dplyr	1.0.7
tidyr	1.1.3
tibble	3.1.4
ggpubr	0.4.0
foreign	0.8.82
tools	4.0.4

2.2.2 Developed methods

For this research a wrapper was made in Intellij IDEA 2021.2.1 (Ultimate Edition) with Java version 16.0.2. The wrapper takes an attribute-relation file format (arff) file with unclassified instances and classify the instances with the created model in Weka. The program predicts if a patient has heart disease yes or no. Based on the wishes of the user, the output was written to the terminal or saved in an arff or comma separated values (csv) file.

3 Results

The results are divided in three different paragraphs. First paragraph is about the data exploration, second paragraph tells something about data cleaning and the last paragraph shows the results of the machine learning experiments.

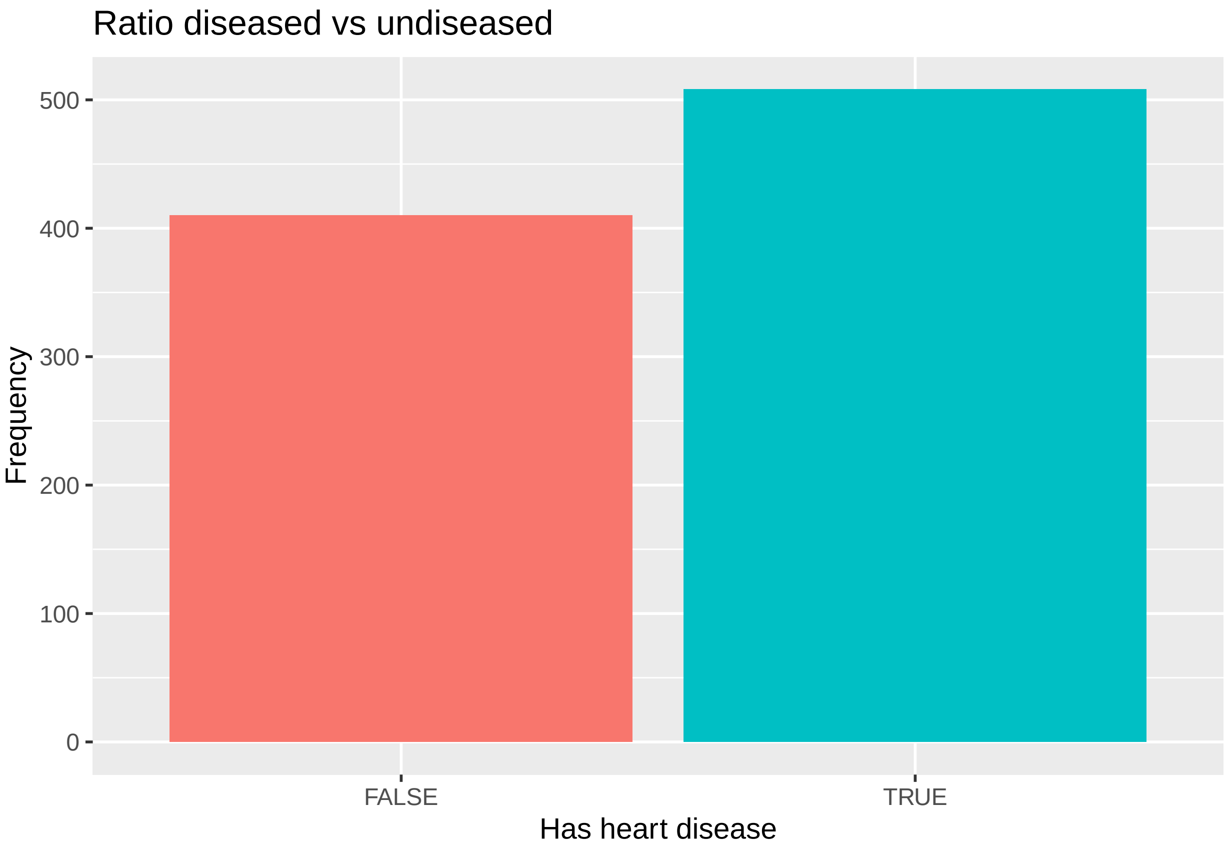
3.1 Data exploration

The goal of the data explorations was to get a better understanding about the data and to look for relations between different variables in the dataset. For example to look for correlation between variables. Another research question to answer was if the dataset was suitable for using in machine learning.

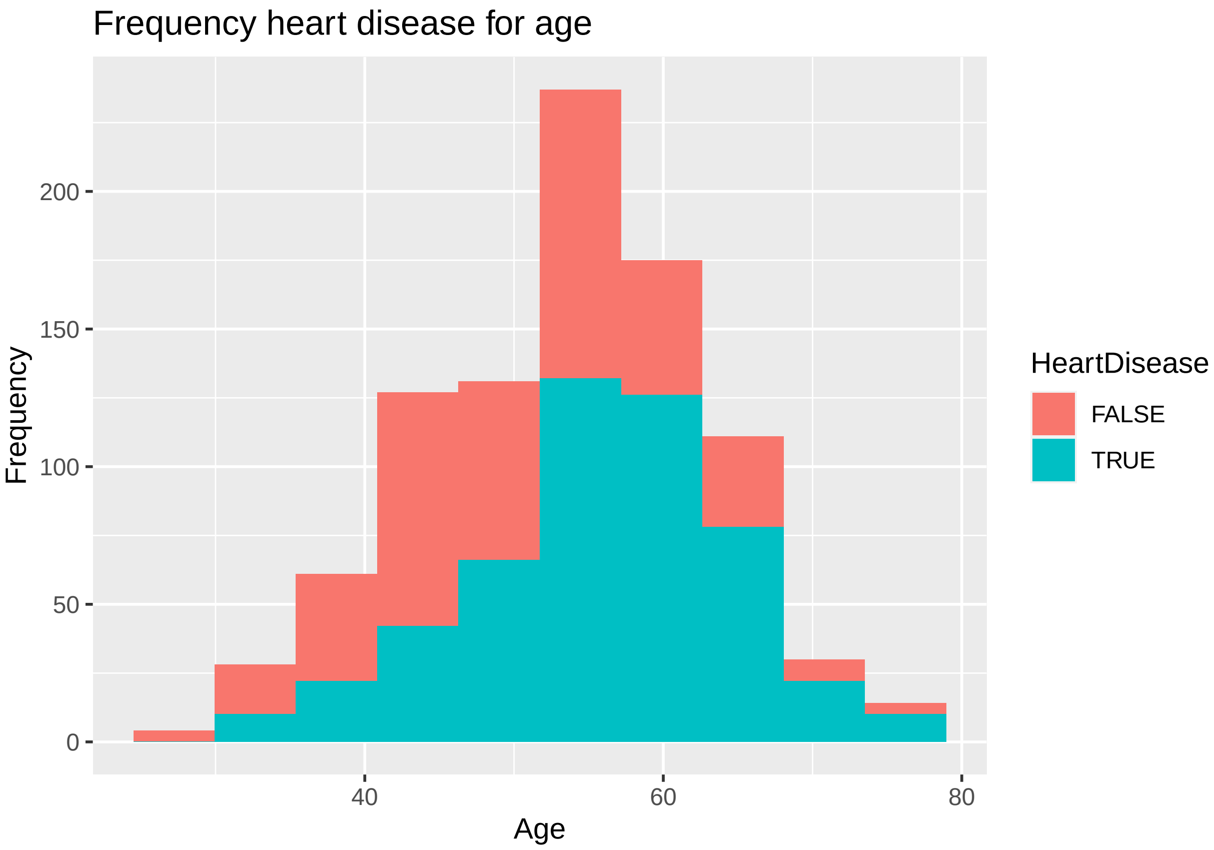
Table 3: 5 number summary of original dataset

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Min	28.00	0.0	0.0	60.0	-2.6000
1st. Qu	47.00	120.0	173.2	120.0	0.0000
Median	54.00	130.0	223.0	138.0	0.6000
Mean	53.51	132.4	198.8	136.8	0.8874
3rd. Qu	60.00	140.0	267.0	156.0	1.5000
Max	77.00	200.0	603.0	202.0	6.2000

The first thing that stands out is the 0 value as minimum value in the column RestingBP. If a person has a resting blood pressure of 0 it means that the person is dead. Column cholesterol has 0 values too as minimum value. A person have most of the time cholesterol level higher than 0.



The ratio between patients with or without the heart disease is almost even. Patients with heart disease contains 55% of the data and patients without heart disease has 45%. This is a slight difference but still tolerable.



The histogram shows seemly a normal distribution in age. It's observable that frequency of patients with HD is at the highest around the age of 50 till 65 years. After the year of 65 the frequency shrink again. This needs more investigation in order to find a relation/correlation between age and people who are diagnosed with HD.

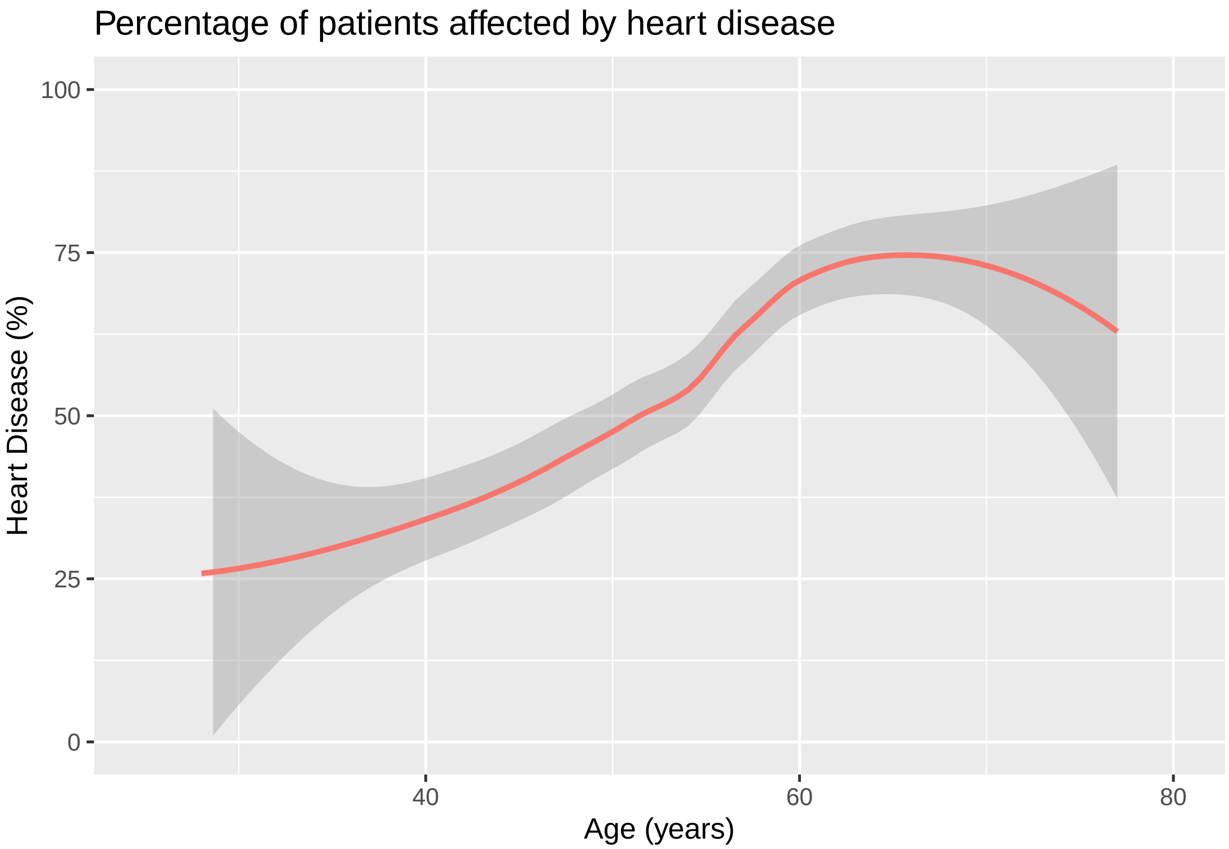
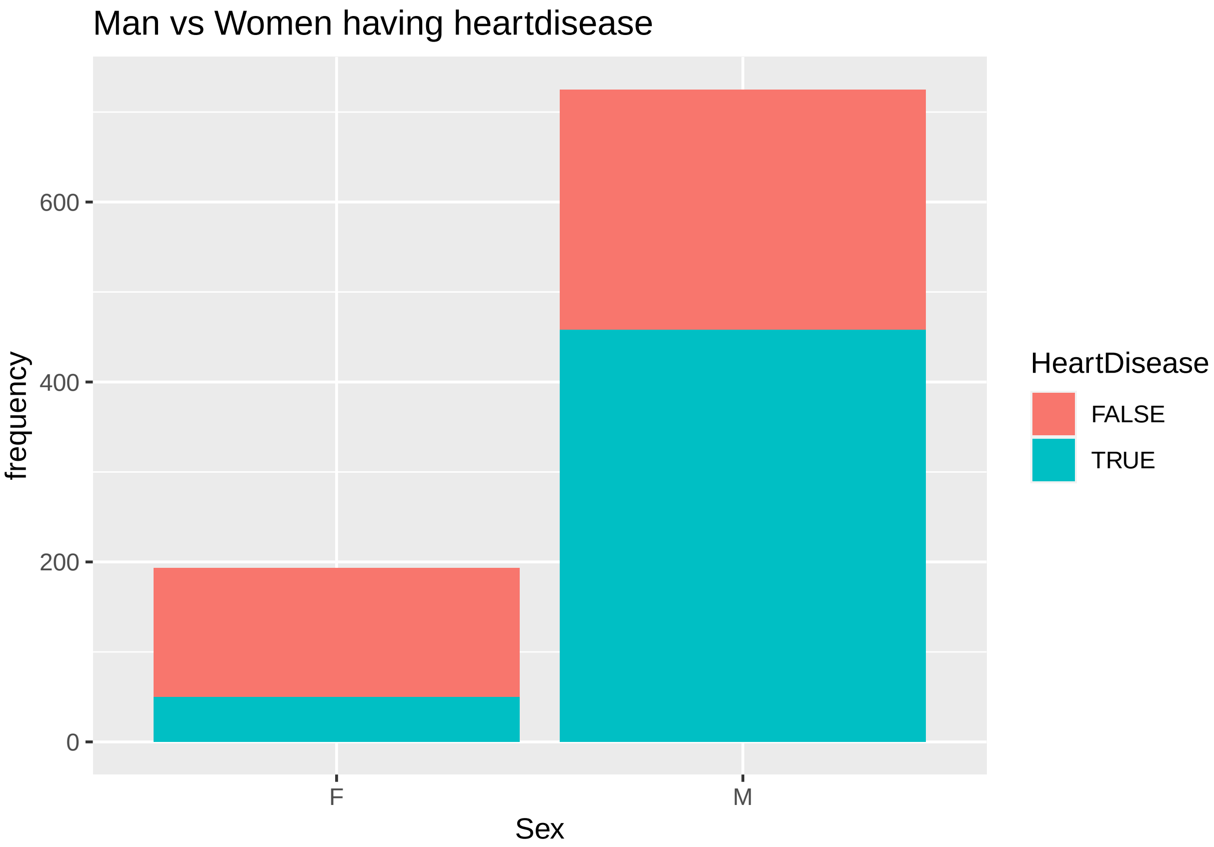
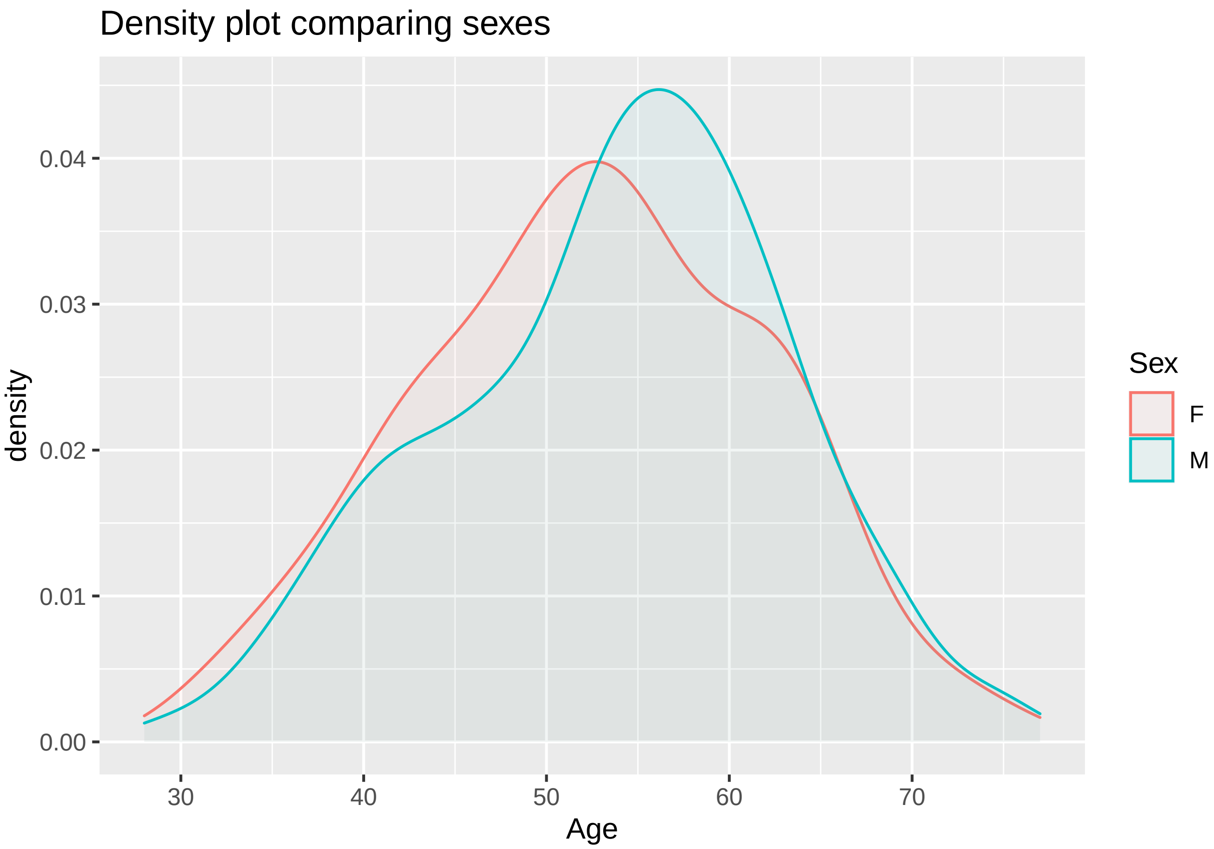


Figure above displays the correlation between age and patients with HD. The trend shows when age increase also the percentage of people with HD rise as well.

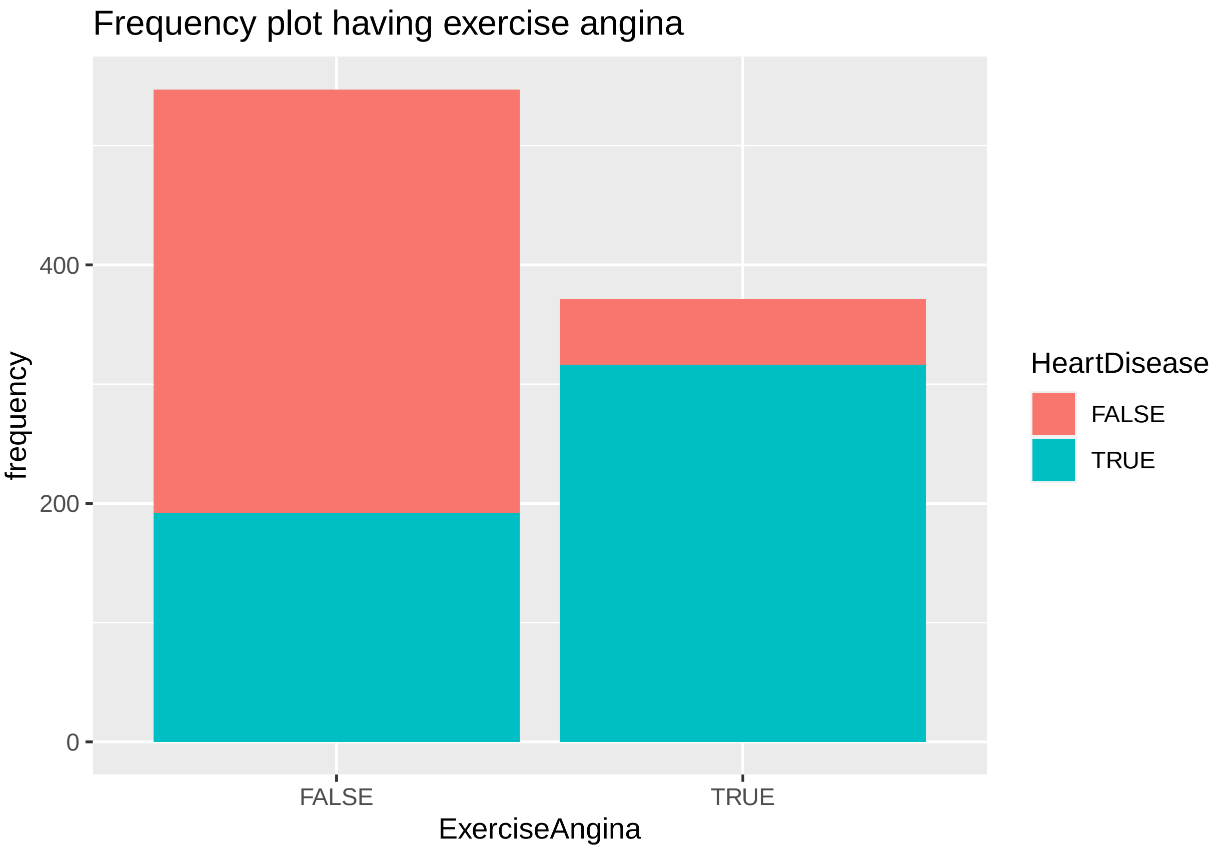
Patients who are diagnosed with HD at the age of 40 is around 35% although patients at 65 years have a 75% HD diagnosis. Last, it's noticeable that percentage people with HD decrease after the age of 65. For people above the age of 65 are more likely to develop a heart disease which can cause heart attacks, strokes or heart failures (Heart Health and Aging, n.d.-b). This is potentially the cause of the decline.



In the figure above the data is more skewed towards males instead of the females. The males ratio (78.98%) is tremendous higher in comparison with females (21.02%). The frequency of females with HD are lower in contrast with females without HD. For Males this is the other way around, the frequency for males with HD are higher in contrast with males without HD.



The figure above indicates that males and females have almost the same amount of age groups. There's a difference at the age of 55 for males. Males have a larger representation after 55 years of age. Females contain a larger representation before the age of 55.



It’s possible to see that a patient with exercise angina has an increase change of diagnosed positive HD compared to with someone who doesn’t have exercise angina.

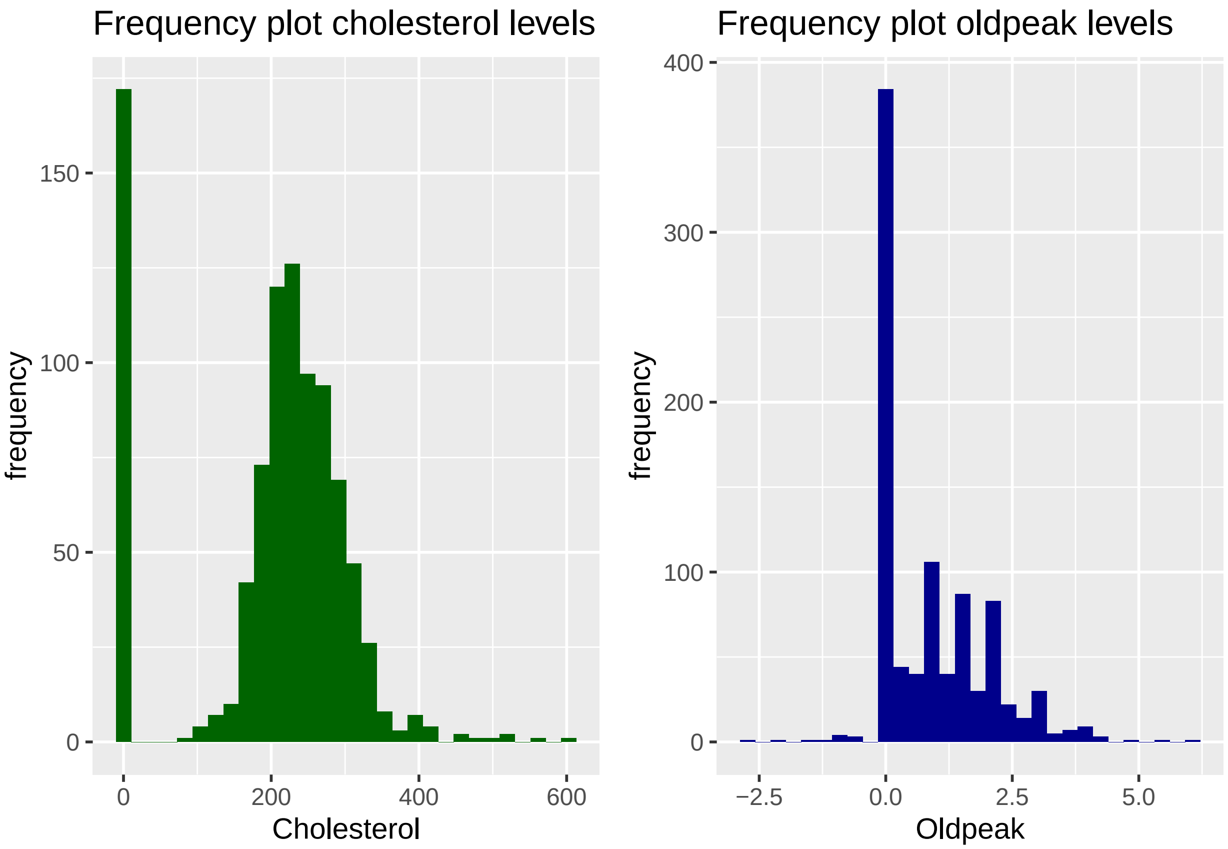


Figure about cholesterol levels show that cholesterol contains a high frequency with 0 values (172). The reason for this is maybe due to no measurements. This comes from the discussion of the dataset source, people discuss about the meaning of the 0 cholesterol values(Heart Failure Prediction Dataset, 2021).

At the end of the histogram there are some higher cholesterol values. The oldpeak frequency plot shows a high frequency of 0 values this states for no depression. The value 0 means no abnormalities and values usually revolve around the 0 (NCBI - WWW Error Blocked Diagnostic, n.d.).

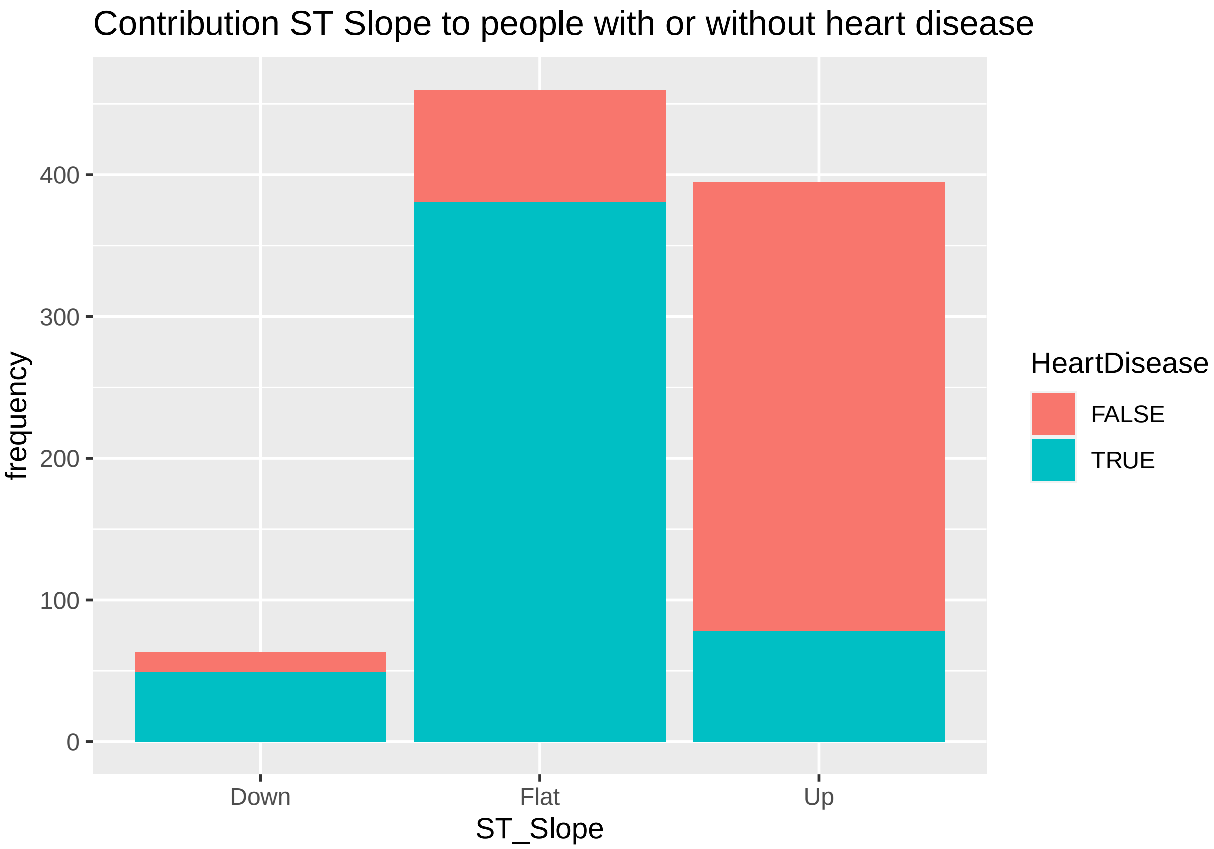


Figure above displays that patients with UP ST slope depression considerably lower the change of HD. For Flat is this the opposite. Patients with Flat ST slope have a higher change in developing HD. The last column Down has a low frequency of data points in comparison with the other two columns Flat and Up.

ECG type:

Code	Explanation
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.

Electrocardiogram (ECG or EKG) is a straightforward test that records and detects your heart's electrical activity. An ECG shows how fast your heart is beating and shows if your rhythm of your heartbeats is steady or irregular. But also indicates the strength and timing of the electical impulses passing through each part of your hart (Heart Tests, 2022).

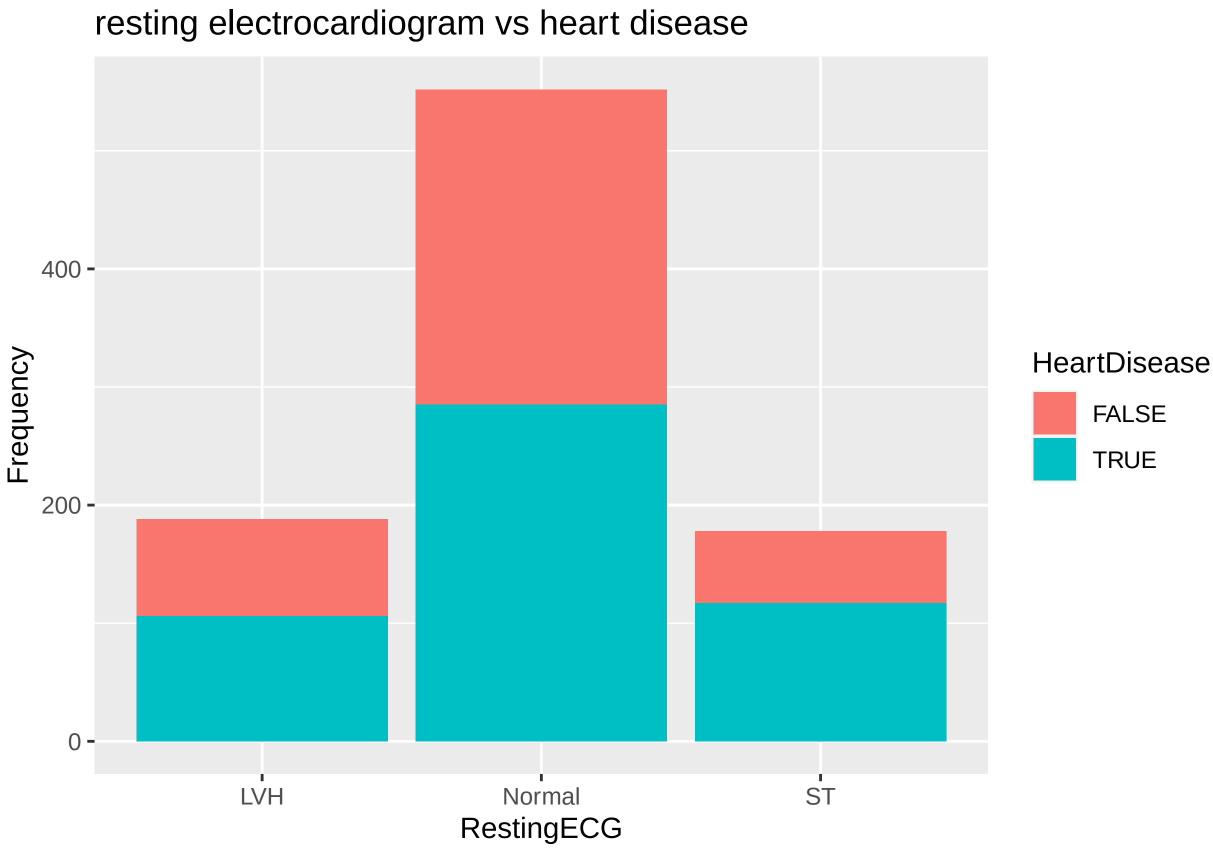


Figure 1: Frequency different ECG types against heart disease

The frequency for a normal (552) electrical activity is exceptionally high in comparison with ST (178) and LVH (188). The ratio between people with HD and without HD are almost even with a normal electrical activity.

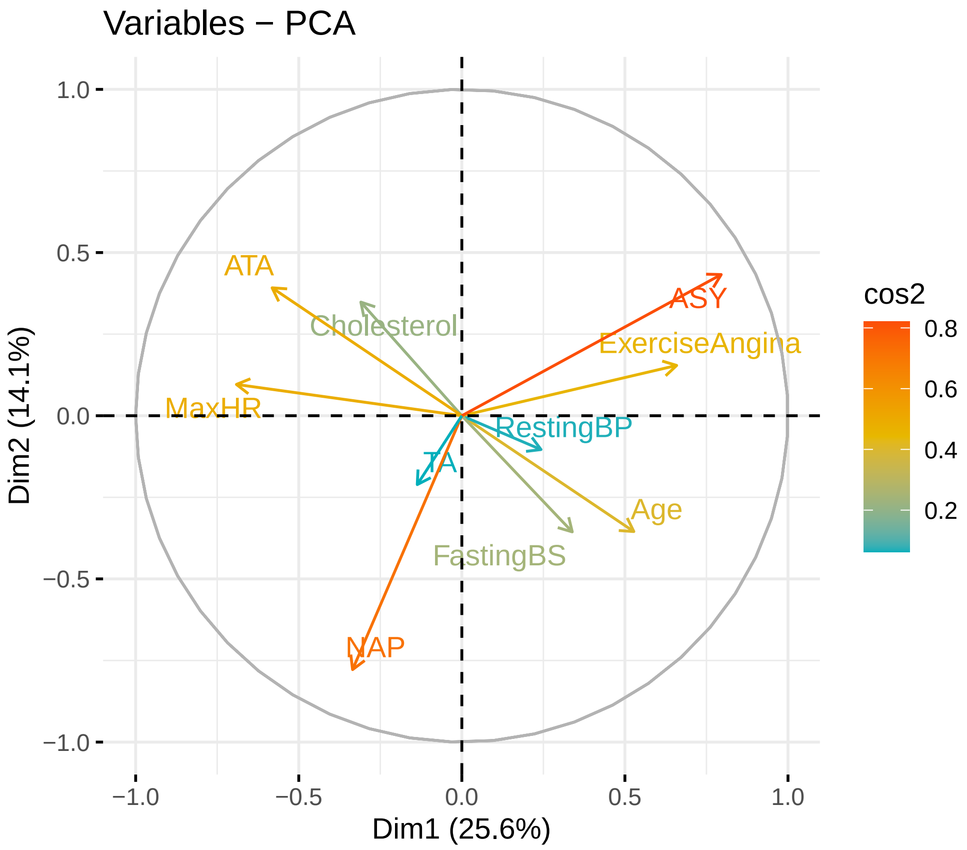
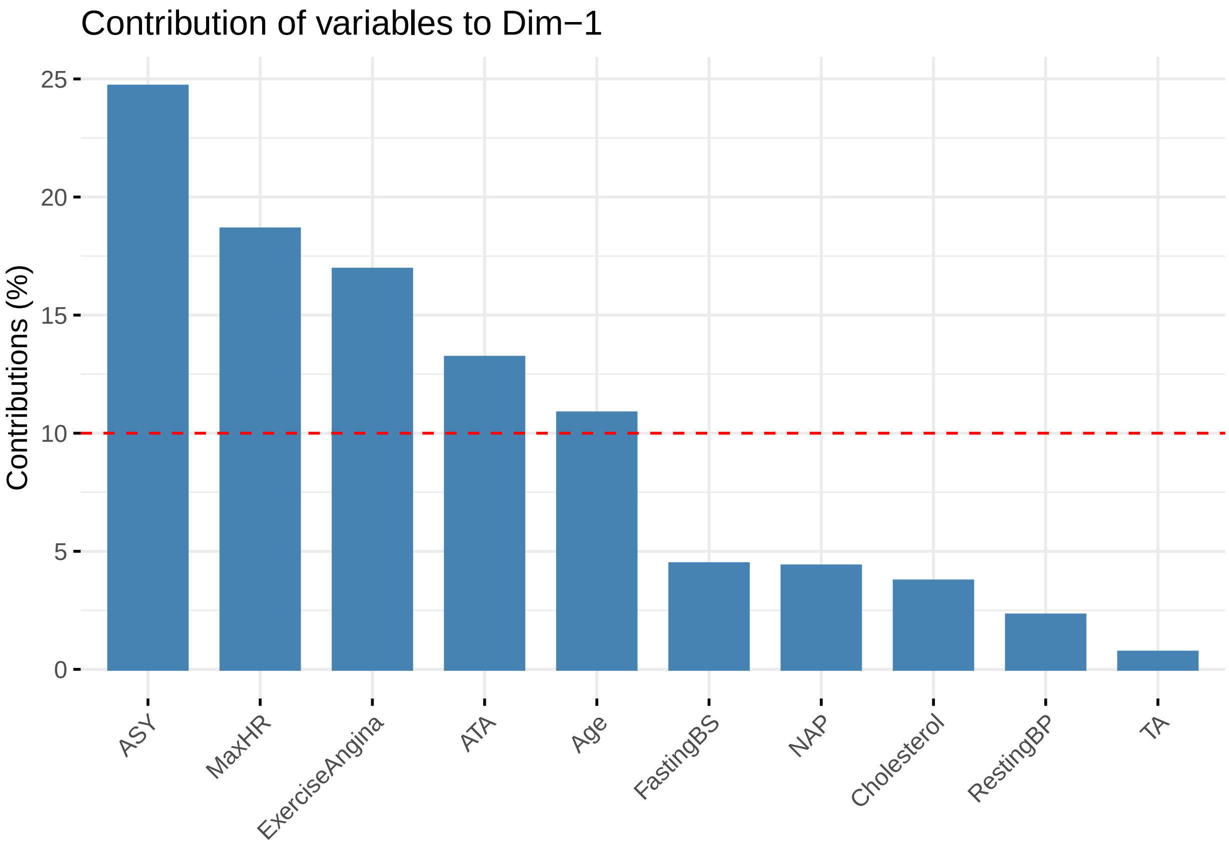
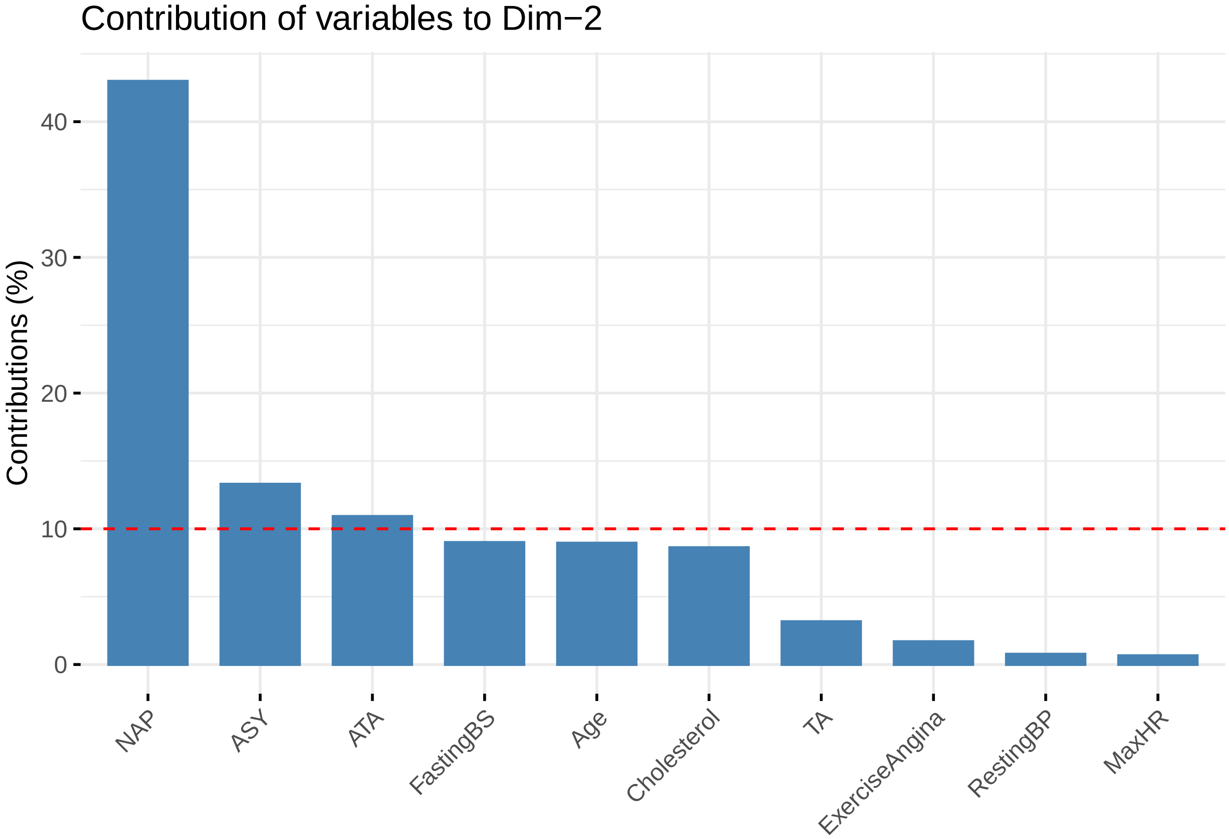


Figure 2: PCA diagram with contribution plot





In the circle diagram and in the contribution diagram the asymptomatic pain and non-Anginal pain has the biggest contribution for a positive or negative diagnosing the heart disease. One downside is that it’s hard to predict for patients who don’t have visible symptoms. Three other important contributors are MaxHR, Age, and ExerciseAngina. These three variables are important because it could be measured for every patient thus are viable.

3.2 Data cleaning

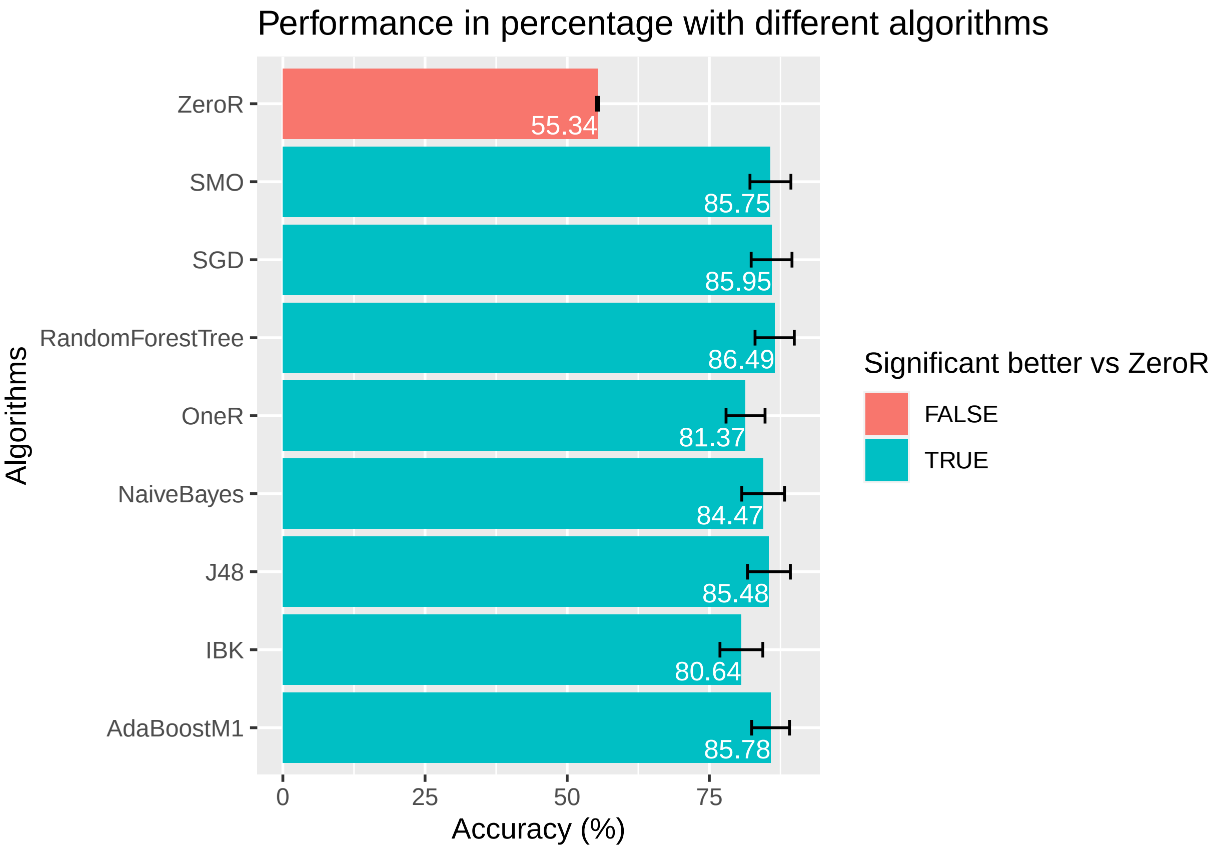
In order to make the dataset ready for machine learning it had to be cleaned up. Cholesterol 0-values are replaced with the median of the cholesterol column due to missing measurements. The column RestingBP has one measurement with 0 and is replaced with the median of their respective column. Median is less sensitive to outliers in comparison with the mean. That’s the reason why the median was used. Last, the heart disease column was moved to the last column. This is because the program Weka functions better if the class what needs to predict is at the end.

3.3 Machine learning

The machine learning experiments were all done with the same p-value of 0.05. All algorithms were chosen based on representation for every classifier category. The base learner was ZeroR and this means that every algorithm was compared for significant difference in accuracy against ZeroR. The plots contain standard deviation error bar. This is done for investigating the spread around the mean or maybe indication about statistically signification between different algorithms. The goal is to find the best performing algorithm and optimize the parameters of the best performing model.

Table 5: Used algorithms with respective Weka parameters

Algorithm	WekaParameter
ZeroR	rules.ZeroR ” 48055541465867954
OneR	rules.OneR ’-B 6’ -3459427003147861443
J48	trees.J48 ’-C 0.25 -M 2’ -217733168393644444
RandomForestTree	trees.RandomForest ’-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1’ 1116839470751428698
AdaBoostM1	meta.AdaBoostM1 ’-P 100 -S 1 -I 10 -W trees.DecisionStump’ -1178107808933117974
IBK	lazy.IBk ’-K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""’ -3080186098777067172
SMO	functions.SMO ’-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"’ -6585883636378691736
SGD	functions.SGD ’-F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1’ -3732968666673530290
NaiveBayes	bayes.NaiveBayes ” 5995231201785697655

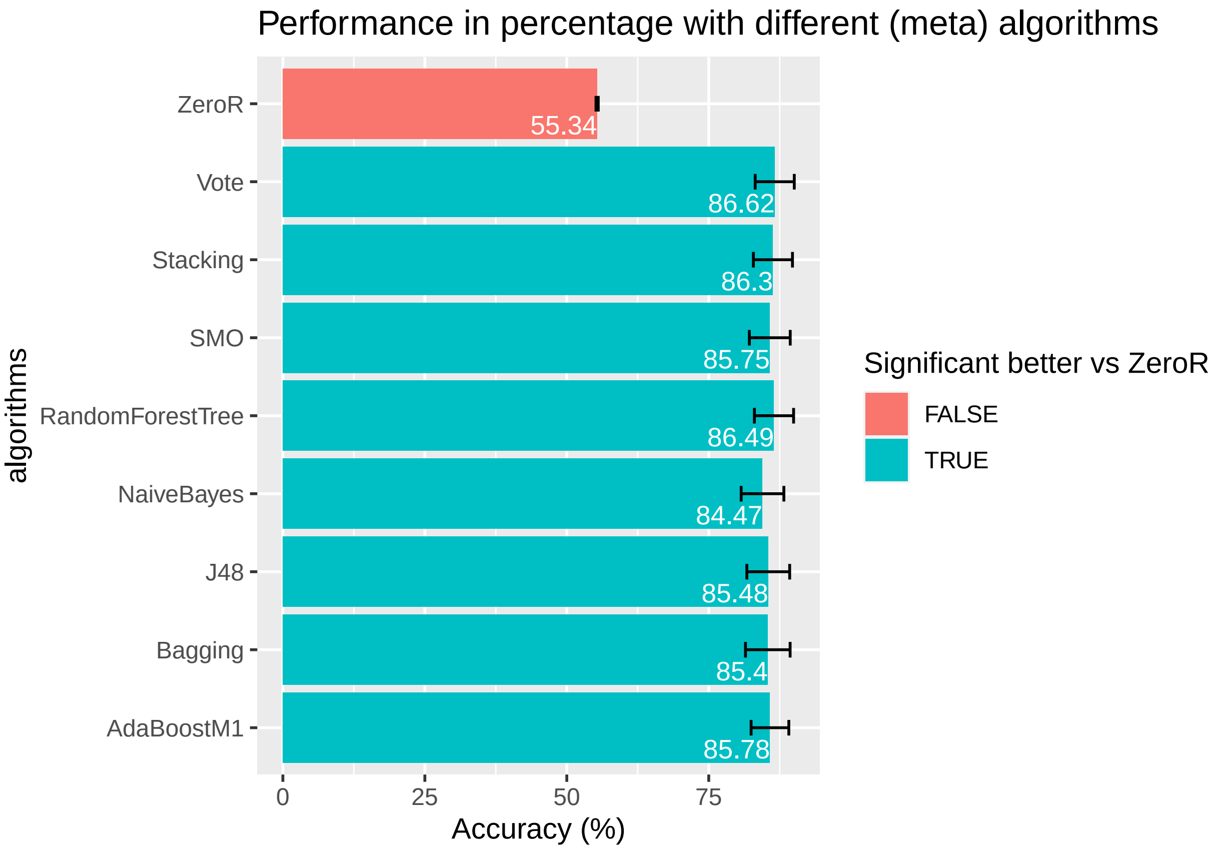


In the graph above, every algorithm is significant better in comparison with ZeroR. This indicates that the algorithm doesn't predict by chance. RandomForestTree and SGD are one of the best two performing algorithms in comparison with other algorithms. For every algorithm except ZeroR there is a certain overlap between the standard deviation error barplots (black sticks). This probably indicates that the difference is not statistically significant between the algorithms. Statistical test is needed to investigate the statistical significance. The standard deviation error bars are relative small and it suggest that the spread of the data are clumped around the mean.

In the context of investigating algorithms, the algorithms Stacking, Voting and Bagging were used for this research. Meta learners learns to solve several tasks and not simply one task. Meta learners doesn’t focus on training one model on one specific dataset.

Table 6: Meta-learners for extra exploration

Algorithm	WekaParameter
ZeroR	rules.ZeroR ” 48055541465867954
RandomForestTree	trees.RandomForest ’-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1’ 1116839470751428698
J48	trees.J48 ’-C 0.25 -M 2’ -217733168393644444
NaiveBayes	bayes.NaiveBayes ” 5995231201785697655
AdaBoostM1	meta.AdaBoostM1 ’-P 100 -S 1 -I 10 -W trees.DecisionStump’ -1178107808933117974
SMO	functions.SMO ’-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"’ -6585883636378691736
Vote	meta.Vote ’-S 1 -B "trees.J48 -C 0.25 -M 2" -B "trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1" -B "bayes.NaiveBayes " -B "meta.AdaBoostM1 -P 100 -S 1 -I 10 -W trees.DecisionStump" -B "functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -calibrator \"functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4\"" -R AVG’ -637891196294399624
Stacking	meta.Stacking ’-X 10 -M "meta.AdaBoostM1 -P 100 -S 1 -I 10 -W trees.DecisionStump" -S 1 -num-slots 1 -B "meta.AdaBoostM1 -P 100 -S 1 -I 10 -W trees.DecisionStump" -B "bayes.NaiveBayes " -B "functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -calibrator \"functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4\"" -B "trees.J48 -C 0.25 -M 2" -B "trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1"’ 5134738557155845452
Bagging	meta.Bagging ’-P 100 -S 1 -num-slots 1 -I 10 -W trees.J48 – -C 0.25 -M 2’ -115879962237199703



The barplot shows that every algorithm is significant better in comparison with ZeroR. The algorithms Voting, Stacking, and Bagging are added to the graph. Voting and Stacking makes the top three best algorithms with RandomForest. Bagging has a percentage of 85.4 for classifying. It matches the accuracy of AdamBoostm1, J48 and SMO. The black sticks are the standard deviation error bars and there is overlap between the algorithms except for ZeroR. The standard deviation error bars are small. This means that the spread of the data is clumped around the mean and the overlap indicate that the difference is not statistically significant between different algorithms. Statistical test needs to be performed to confirm if there is not a statistically significant difference between the algorithms.

3.3.1 Cost-sensitive classification

One of the most important things to consider are the resulting false negatives when doing classification for people diagnosing with heart disease. The amount of people who are diagnosed negative but are positive for heart disease must be low as possible. In order to achieve this, it's possible to use a cost matrix. This result in a penalty for the algorithm when the prediction is negative for heart disease, but the patient is positive for the heart disease. Applying a cost matrix could possibly lead to lowering in accuracy for the algorithm. It's crucial to find balance between keeping the accuracy high as possible but lowering the false negatives as well.

3.3.1.1 RandomForest:

Accuracy: 87.08%
Default cost matrix: [0.0,1.0;1.0,0.0]
Confusion matrix:

a	b	classified_as
338	72	a = FALSE
45	364	b = TRUE

Accuracy: 86.71%
Cost matrix: [0.0,1.0;2.0,0.0]
Confusion matrix:

a	b	classified_as
323	87	a = FALSE
28	480	b = TRUE

Accuracy: 85.33%
Cost matrix: [0.0,1.0;3.0,0.0]
Confusion matrix:

a	b	classified_as
297	113	a = FALSE
17	491	b = TRUE

3.3.1.2 AdaBoostm1

Accuracy: 83.44%
Cost matrix: [0.0,1.0;3.0,0.0]
Confusion matrix:

a	b	classified_as
293	117	a = FALSE
31	477	b = TRUE

Randomforest accuracy decrease 0.37 percent with cost matrix of [0 0 2 0] and lowering the false negatives with 19 instances. The accuracy decreases around 1.4% when using the cost matrix [0 0 3 0] and still maintaining an accuracy of 85.33%. The false negatives decrease with 11 instances, but the false positives increase with 26 instances compared with cost matrix [0 0 2 0]. Using the cost matrix of [0 0 3 0] decreases false negatives a little more than 50% in comparison with the default matrix. It's more viable to lower the false negatives instead of increasing the false positives. AdaBoostm1 performs worse in accuracy and the amount of false negatives in comparison with Randomforest with a cost matrix of [0 0 3 0].

4 Discussion & Conclusion

4.1 Discussion

The cholesterol column contains a lot of zero-value patients. The cholesterol value is perhaps incorrectly recorded or read according to the discussion(Heart Failure Prediction Dataset, 2021). The same could be done for the variable oldpeak because it has the problem as well. In the discussion of the data set it's explained that the zero means there are no abnormalities (Heart Failure Prediction Dataset, 2021b). fortunate, the ratio between patients with or without heart disease are almost even (0.45 to 0.55). Asymptomatic pain is the main contributor in diagnosing patients with heart disease. Unfortunate, asymptomatic pain isn't usable because not all patients experience chest pain when the patients are diagnosed positive for heart disease. Asymptomatic pain is the most important contributor (see figure pca) in diagnosing patients positive for heart disease. 79 percent of the positive diagnosis of heart disease are due to asymptomatic pain. Chest pain type Non-Anginal Pain (NAP) has a reverse effect. Figure @ref(fig:pca-var) affirms that NAP has a large contribution for a negative diagnosis. If you left the variable chest pain type out, the accuracy of the prediction would probably go down for this data set. Moreover, the data is more favoured towards males due to it's more skewed towards males. Males has a very large percentage(63%) who are diagnosed with heart disease in comparison with females(26%). The decrease of patients with heart disease after the age of 65 perhaps be interpreted by the conviction that patients who are older then 65 die because of the heart disease. Other three important variables in diagnosing heart disease are maximum heart rate, age, and exercise angina. These variables could be useful in creating a realistic algorithm for diagnosing patients with heart disease.

For future research, it could increase the quality of the data to add more female data to the dataset. This could improve the heart disease ratio for females. Try to train an algorithm that don't rely heavy on the type of chest pain but create the algorithm that relies more on the variables maximum heart rate, age and exercise angina. That's because patients who suffer heart disease doesn't always have chest pain symptoms.

4.2 Conclusion

The goal of the research is to produce an accurate (false negatives <= 5%) machine learning algorithm that predicts the possibility of developing a heart disease in a wide array of patients. For this project, the Randomforest tree is the best suitable algorithm to predict if patients have a heart disease with an accuracy of 85.33%. The accuracy decreased with 1.7% for choosing Randomforest with a cost matrix of [0 0 3 0], but the false negatives decreased from 45 instances to 17 according to confusion matrices. That's around 62% decrease of patients who are diagnosed negative for heart disease but are positive for the heart disease. Only 1.73% of all the instances are classified as false negatives reported by the confusion matrix. The decrease of accuracy is due the increase of the false positives but it's less important if a patient is diagnosed positive for heart disease but is negative for heart disease instead of other way around.

Maximum heart rate, age, and exercise angina are viable variables in diagnosing patients with heart disease. It's a relatively small contribution however a possible measurement for every patient. If the chest pain type is measured the accuracy of this model increases but the problem is that a patient with heart disease don't always have a chest pain in the real world. Although, it's possible to predict heart disease with an excellent accuracy on this dataset in despite of that it won't be easy to predict the heart disease if the chest pain type is missing like in the real word data.

5 Project proposal for minor

There are various ways to improve the research. This could apply for the minor Application Design.

5.1 Current issues:

The Java wrapper isn't user friendly because the user needs to have basic command line and GIT knowledge in order to use the application. To tackle the problem at its roots, a web application would solve this problem. Another solution for making the program more dynamic is to give the user more possibilities to choose from different input files. Now the wrapper only accepts an arff file as input. All this would make the wrapper more user friendly.

5.2 Goal

To create a more dynamic application, a web application along with a web API could do it. This is an outcome for people working in healthcare who wants to get a quick overview in the current risk for heart disease in a patient. For example, larger hospitals like UMCG would be capable to review many patients.

5.3 Target audience

The audience would be especially professional healthcare workers. The main instrument of this tool would be a quick evaluation in the possibility of developing heart disease. An effect would lessen the number of professionals in the hospital because fewer patients need more testing in the UMCG.

5.4 Design

As mentioned before, the input could only be an arff file and isn't dynamic. The output is somewhat more dynamic because the user could choose between a csv or arff file as output but still not user friendly. This output doesn't give additional information about the why or how the output is created. Adding extra information or details of the output would benefit the reliability of the created output. A simple example would be to add the created tree if a tree model is used. The professional healthcare worker could distract why a patient is diagnosed positive or negative for heart disease.

6 References

Heart Health and Aging. (n.d.-b). National Institute on Aging. https://www.nia.nih.gov/health/heart-health-and-aging

Heart Failure Prediction Dataset. (2021, September 10). Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/discussion/293759?resource=download

NCBI - WWW Error Blocked Diagnostic. (n.d.). https://www.ncbi.nlm.nih.gov/books/NBK459364/

Heart Tests. (2022, March 24). NHLBI, NIH. https://www.nhlbi.nih.gov/health/heart-tests

Heart Failure Prediction Dataset. (2021b, September 10). Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/discussion/279156?resource=download