DSCI 310: Diagnosing Heart Disease With Two Predictors

Group 03: Hanzhang Cao, Karlie Truong, Kaylie Nguyen, Ser Jie Ng

2023-04-10

Introduction

Cardiovascular disease is the leading cause of death worldwide (Ahmad and Anderson 2021). During 2020, the United States saw 690,882 deaths from this condition, compared to the 345,323 deaths from COVID-19 (Centers for Disease Control and Prevention 2022). There are several risk factors for developing heart disease, and the consequences of heart disease vary in severity. The risk of death increases when other comorbidities are present, including COVID-19 (Madjid et al. 2020). Because this disease affects a large amount of the population, it puts significant strain on the healthcare system.

We thus want to determine: Based on only two predictors gathered from health checkups performed by doctors, does somebody suffer from heart disease?

Data

Heart disease data were sourced from UCI Machine Learning Repository website (Detrano 2019). Our data has 302 rows and 15 column which contains variables that relates to physical health as well as their diagnosis of cardiovascular disease.

Exploratory Data Analysis

We want to get a first impression of the different predictors in the data set, thus we compare the mean values of the numerical variables (*see Table @ref(tab:training-set-averages-table)*) and the distribution of the factorial attributes (*see Figure @ref(fig:ratio-plot)*) for observations classified as sick and healthy.

Table 1: Average values of the numerical attributes

diagnosis	age	resting_bp	cholesterol	max_heart_rate	old_peak	no_vessels_colored
healthy	52.33	129.02	241.99	158.94	0.53	0.29
sick	56.52	135.06	245.42	138.68	1.67	1.13

As the mean values do not tell us much about the numerical variables, we also plot their histograms to visualize the distribution of the variables (*see Figure @ref(fig:var-hist)*). We will use this later to choose the predictors we want to use.

Building & Optimizing Our Model

We will use the k-nearest neighbor (KNN) algorithm for our data analysis (Timbers, Campell, and Lee 2022). KNN is a very well known classification algorithm, because it is very intuitive and does not make any major assumptions about the data. The only important requirement is that the distance between data points must represent their similarity. Furthermore, there are well-known ways to optimize a KNN model, which are implemented in the R library tidyverse, so it is easy to use.

Our goal is to find the two best predictors for a KNN classification. However, taking all variables into account takes too long, so we exclude some, based on the exploration above: Based on Figure @ref(fig:ratio-plot), `high blood sugar` did not show a significant increase between sick and healthy subjects. `Chest pain type` was removed from further analysis due to the initial study lacking documentation, as well as inconsistencies in the academic community for the terms and methods used to describe and categorize the types of chest pain. The academic community typically refers to only four types of angina: stable angina, unstable angina, microvascular angina, and vasospastic/variant angina (National Heart, Lung, and Blood Institute 2022). Therefore, it was challenging finding resources that properly described what abnormal angina was, as referenced in the original study. Additionally, `sex` as a predictor was excluded, as treating `sex` as binary is a topic we do not want to make a statement on.

KNN does not work with non-numerical parameters, so we need to convert our factorial parameters to numerical values. `Exercise pain` is a binary variable, so we convert it to 0 if no exercise pain was reported, and 1 otherwise. `Slope` can be easily replaced with -1 for "down", 0 for "flat", or 1 for "up". For better scaling of the parameters, `thal` and `resting ECG` values were replaced with non-standard numeric values; Normal was replaced with 0 and the non-normal values (i.e., symptomatic results) were either replaced with 4 or 5. We decided to use these values to represent that the abnormal states are versions of the same broader category.

To optimize the model, we need to ensure that there are no NAs values in the training data. We document the number of rows we omit, to make sure that the model still provides sufficient information: Only roughly 2% of our observations contain NA values, so we can safely ignore them.

To get a better grasp of the data set and be able to understand the following metrics, we also count the number of observations that are labeled as sick.

We can see that roughly 46% of our data points are sick. KNN sometimes struggles if the different classes are not balanced. Since this is not the case here, we do not need to artificially balance the classes (for example with oversampling).

Despite aiming for two predictors, we first want to build and optimize a model for all the predictors that have not been excluded so far. This allows us to have a reference for our two-predictor-models and lets us verify our model building.

The choice of the parameter k neighbors used has a lot of influence on the accuracy of the model. We apply cross-validation to optimize this parameter. Odd values of k ranging from 1 to 30 are tested, since odd values have the advantage of not needing a tie-breaker during the KNN algorithm. Looking at more than every 7th data point would probably lead to underfitting (Timbers, Campell, and Lee 2022). As noted above, KNN needs data where the euclidean distances describe the similarity between the data points, thus it is important that the data set is scaled. This way the euclidean distances mimic the distance of the data points on a graph with linear axes.

Table 2: The cross-validation results for all predictors

neighbors	.metric	.estimator	mean	n	std_err	.config
11	accuracy	binary	0.83	5	0.02	Preprocessor1_Model06

Table @ref(tab:cross-val-results) shows the optimal value of k for the `all predictors` model. We can see that $k = 11$ is the best choice, with an accuracy of 0.83.

However, we are not only interested in the accuracy, but also in the false negatives, as labeling someone as healthy who is actually sick is much more dangerous than the other way round. Since this information can not be extracted from the cross-validation, we train a KNN model with the chosen value of k on the whole training set, and let it process the training set. As we are still building our model, the testing data can not be used for this purpose. We could also split our training data once more, a validation set which we would use to determine the false negative rate. However, this would make our false negative rate very dependent on this one random split. Thus, we decided against this second approach.

Note that we can not and should not optimize our whole model for the lowest false negative rate, as always predicting "sick" would set this value to 0, and the resulting model would be worthless.

The resulting confusion matrix is shown in @ref(fig:prediction-confusion-plot). The first column shows the observations that are actually healthy, the second one the ones that are healthy. The first row shows the observations that are predicted as healthy, with those predicted as sick in the second row. Thus, there are 106 people who are correctly determined as healthy, 83 who are correctly determined as sick, 19 who are labeled as healthy despite being sick (which is the bad case), and 13 that are labeled as sick despite being healthy.

We now repeat the process described above for every subset of two predictors. We thereby track the predictors, the optimized value of k, the accuracy, and the number of false negatives. The resulting data are shown in Table @ref(tab:model-formula-results).

Table 3: The accuracy results for all tried formulas

formula	k	accuracy	false_healthy
diagnosis ~ .	11	0.83	19
diagnosis ~ old_peak + max_heart_rate	15	0.79	34
diagnosis ~ slope + old_peak	27	0.79	30
diagnosis ~ no_vessels_colored + old_peak	29	0.78	25
diagnosis ~ thal + exercise_pain	29	0.77	48
diagnosis ~ thal + age	9	0.77	23
diagnosis ~ thal + no_vessels_colored	25	0.76	49
diagnosis ~ old_peak + exercise_pain	3	0.76	41
diagnosis ~ thal + old_peak	21	0.76	31
diagnosis ~ thal + resting_bp	19	0.76	28
diagnosis ~ thal + max_heart_rate	13	0.76	24
diagnosis ~ no_vessels_colored + max_heart_rate	17	0.76	34
diagnosis ~ thal + cholesterol	15	0.76	28
diagnosis ~ thal + resting_ecg	29	0.76	28
diagnosis ~ thal + slope	25	0.76	28
diagnosis ~ exercise_pain + age	23	0.76	23
diagnosis ~ no_vessels_colored + exercise_pain	25	0.75	30
diagnosis ~ no_vessels_colored + slope	17	0.75	41
diagnosis ~ no_vessels_colored + age	15	0.75	37
diagnosis ~ max_heart_rate + age	11	0.75	31
diagnosis ~ slope + max_heart_rate	15	0.75	34
diagnosis ~ no_vessels_colored + resting_bp	23	0.74	35
diagnosis ~ exercise_pain + max_heart_rate	27	0.74	34
diagnosis ~ max_heart_rate + resting_ecg	9	0.73	33
diagnosis ~ no_vessels_colored + cholesterol	21	0.73	37
diagnosis ~ slope + age	27	0.73	28
diagnosis ~ slope + cholesterol	23	0.73	23
diagnosis ~ old_peak + cholesterol	7	0.73	32
diagnosis ~ old_peak + age	27	0.73	38
diagnosis ~ slope + resting_bp	25	0.73	30
diagnosis ~ max_heart_rate + resting_bp	23	0.72	39
diagnosis ~ old_peak + resting_ecg	11	0.72	37
diagnosis ~ no_vessels_colored + resting_ecg	17	0.72	48
diagnosis ~ exercise_pain + resting_bp	19	0.71	48
diagnosis ~ exercise_pain + cholesterol	29	0.71	48

formula	k	accuracy	false_healthy
diagnosis ~ exercise_pain + resting_ecg	15	0.71	75
diagnosis ~ slope + exercise_pain	11	0.71	57
diagnosis ~ slope + resting_ecg	29	0.71	62
diagnosis ~ old_peak + resting_bp	9	0.70	34
diagnosis ~ max_heart_rate + cholesterol	17	0.70	38
diagnosis ~ resting_ecg + age	23	0.66	43
diagnosis ~ resting_bp + age	21	0.65	38
diagnosis ~ cholesterol + age	27	0.64	32
diagnosis ~ resting_ecg + resting_bp	3	0.59	60
diagnosis ~ resting_ecg + cholesterol	27	0.57	56
diagnosis ~ cholesterol + resting_bp	9	0.54	48

We also visualize the data in Figure @ref(fig:predictors-results). The formulas on the y axis can be read as follows: The words like `diagnosis`, `thal`, `old_peak` reference the corresponding variables; `~` can be read as "is predicted by"; `+` as "and"; and `.` as "all predictors". Thus `diagnosis ~ .` means we take all the predictors (except the ones we excluded above, such as sex), and `diagnosis ~ thal + slope` means we use the two predictors `thal` and `slope` to predict the diagnosis.

Model Selection & Verification

We now have to select the formula we want to use.

For this, we mostly want high accuracy. If the accuracy is comparable, we want to make sure to choose the one with fewer false negatives. As we are aiming for a simple check-up, using all predictors (`diagnosis ~ .`) is not ideal, though yielding the highest accuracy. We pick `exercise_pain + age` because it has similar false negatives while having an accuracy that is similar to when using all predictors while being attributes that are easily examined.

Table 4: The accuracy results for the selected formula compared with the `all predictors` version

formula	k	accuracy	false_healthy
diagnosis ~ .	11	0.83	19
diagnosis ~ exercise_pain + age	23	0.76	23

The chosen model uses `exercise pain` and `age` as predictor. Exercise pain can be evaluated by the patient doing exercises suggested by the doctor and recording any pain that has emerged. Furthermore, there is a very useful tool that doctors use called the exercise electrocardiogram. This can be done to determine the stress on the heart caused by exercising (Heart and Stroke Foundation of Canada, n.d.).

We can now verify the model on the testing data.

Table 5: The accuracy results for `exercise_pain` and `age` when tested against `testing_set`

.metric	.estimator	.estimate
accuracy	binary	0.68

Overall, the accuracy for our model is 0.68, which is much lower than expected. This could have been due to only taking into account 2 predictor variables and randomization during the splitting process.

Using this version, the rate for diagnosing sick patients as healthy is ~34%. This means that 3 out of 10 cases of heart disease are misdiagnosed. However, as this is a simple method for doctors to use as a signal to further monitor patients for risk of developing disease, the accuracy for this version may be acceptable. We can also plot the whole data set and the areas where new data points will be classified as sick or healthy (*see Figure @ref(fig:classification-plot)*).

Figure @ref(fig:classification-plot) suggests that having any exercise-induced pain is more likely to be classified as sick regardless of age. However, the rigged areas in the graph suggests that there might have been outliers, confounding variables and/or over-fitting of the KNN model since we are not taking into considerations other factors that are highly interrelated to age or exercise pain.

A research paper shows that older adults have a higher incidence of musculoskeletal pain especially chronic pain and are generally less physically active than younger adults (Niederstrasser and Attridge 2022). However, the main reason is not due to aging but low physical activity and low wealth during their younger age. Another research (Ciolac, Brech, and Greve 2010) suggests that age does not affect the musculoskeletal and cardiovascular response to resistance and aerobic training in women. The elderly can exercise safely while having similar exercise intensity compared with the younger and the workload progression does not increase the risk of muscle incidence or injuries. In conclusion, both articles we found suggest that age does not directly affects exercise pain.

Discussion

This analysis challenges some of the initial research done for this project. For example, cholesterol is typically considered highly correlated with heart disease (Centers for Disease Control and Prevention 2022), however it appears not to be a strong predictor in this data set (*see Figure @ref(fig:predictors-results)*); there is not a significant distribution difference between the sick and healthy population (*see Figure @ref(fig:var-hist)*). On the other hand, exercise pain is a pretty good predictor for heart disease in this data set, which makes sense from a naive standpoint, as well as from the distribution of the data set (*see Figure @ref(fig:ratio-plot)*).

Beginning this project, we were optimistic about finding a model with high accuracy that could be used to predict heart disease. However, the highest accuracy found during this analysis was around 0.83. Ideally, this analysis would have revealed a set of predictors with an even higher accuracy. This might be due to confounding variables, the slight imbalance (*see @ref(fig:classification-plot)*) and outliers within our data set.

Our model offers a preliminary diagnosis, which helps healthcare professional to save time and medical costs when identifying high-risks population. Hence, age and exercise pain are appropriate predictors for this purpose as it is inexpensive, time-efficient and require minimal testing. Despite being a preliminary diagnosis tool for heart diseases, improvements such as taking confounding variables into account and/or using a bigger data set in future projects would help improve the model's accuracy.

There are several open questions: Can our model be improved to diagnose the severity of heart disease in patients? Are we able to improve our accuracy and false negative rate by taking more predictors? Are there predictors that are less expensive but still achieve a similar accuracy? In our data set, most formulas with high accuracy involve expensive tests (*see Figure @ref(fig:predictors-results)*). Lastly, is there a more effective and convenient predictor for diagnosing heart disease than the ones used in this analysis? One example for this might be a family history of heart diseases (American Heart Association 2015).

References

Ahmad, Farida B., and Robert N. Anderson. 2021. "The Leading Causes of Death in the US for 2020." *JAMA* 325 (March). https://doi.org/10.1001/jama.2021.5469.

American Heart Association. 2015. "Family History and Heart Disease, Stroke." www.heart.org. https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/family-history-and-heart-disease-stroke.

Centers for Disease Control and Prevention. 2022. "Heart Disease Facts." Centers for Disease Control; Prevention. https://www.cdc.gov/heartdisease/facts.htm.

Ciolac, Emmanuel G, Guilherme C Brech, and Júlia M D Greve. 2010. "Age Does Not Affect Exercise Intensity Progression Among Women." *Journal of Strength and Conditioning Research* 24 (November): 3023–31. https://doi.org/10.1519/jsc.0b013e3181d09ef6.

Detrano, R. 2019. "UCI Machine Learning Repository: Heart Disease Data Set." Uci.edu. https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Heart and Stroke Foundation of Canada. n.d. "Exercise Electrocardiogram." https://www.heartandstroke.ca/heart-disease/tests/exercise-electrocardiogram.

Madjid, Mohammad, Payam Safavi-Naeini, Scott D. Solomon, and Orly Vardeny. 2020. "Potential Effects of Coronaviruses on the Cardiovascular System: A Review." *JAMA Cardiology* 5 (March). https://doi.org/10.1001/jamacardio.2020.1286.

National Heart, Lung, and Blood Institute. 2022. "Angina (Chest Pain) - Types | NHLBI, NIH." www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/angina/types#:~:text=The%20types%20of%20angina%20are.

Niederstrasser, Nils Georg, and Nina Attridge. 2022. "Associations Between Pain and Physical Activity Among Older Adults." Edited by David Meyre. *PLOS ONE* 17 (January): e0263356. https://doi.org/10.1371/journal.pone.0263356.

Timbers, T., T. Campell, and M. Lee. 2022. *Data Science: A First Introduction*. datasciencebook.ca. https://datasciencebook.ca/.

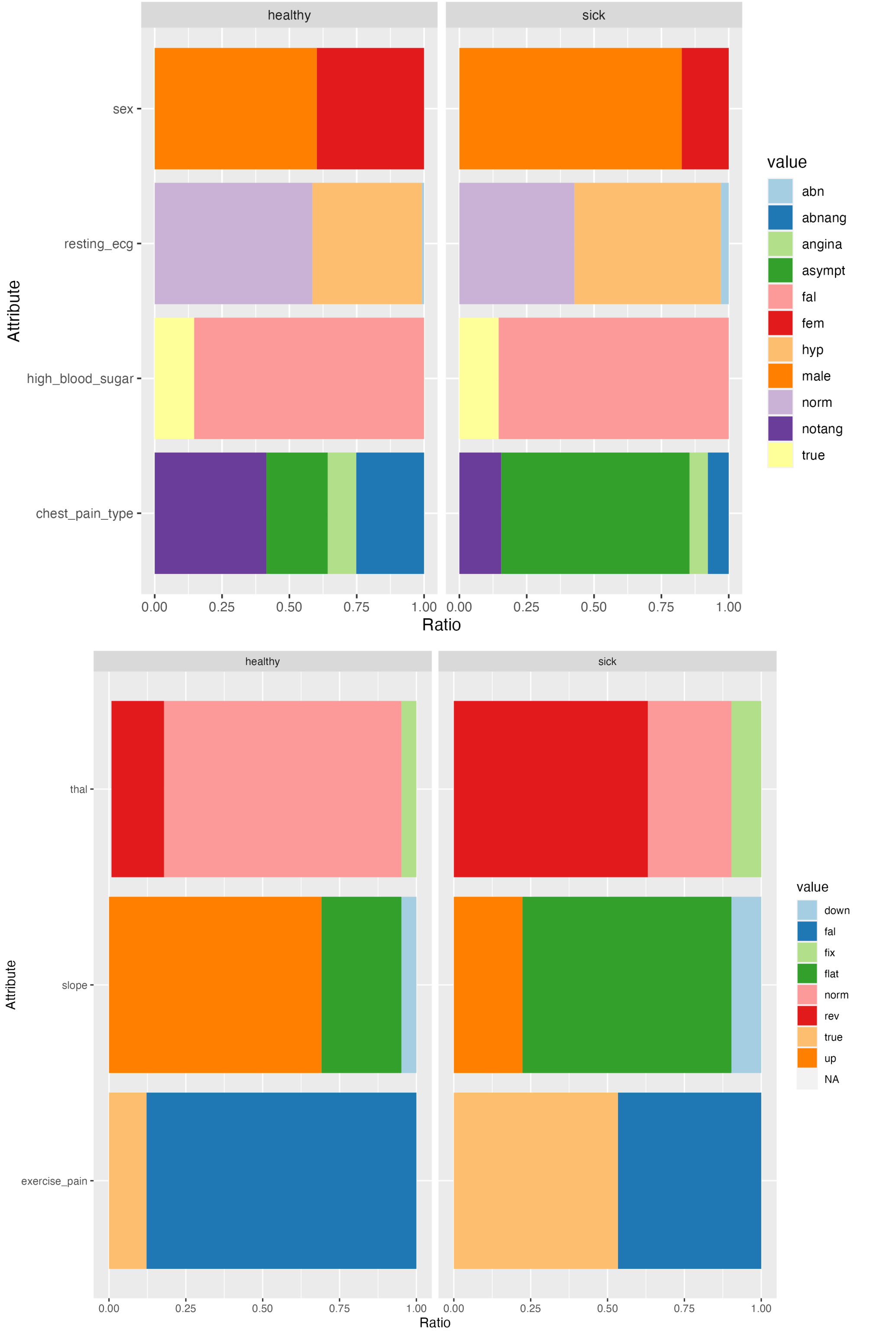

Figure 1: The proportion of different values in each factorial attribute

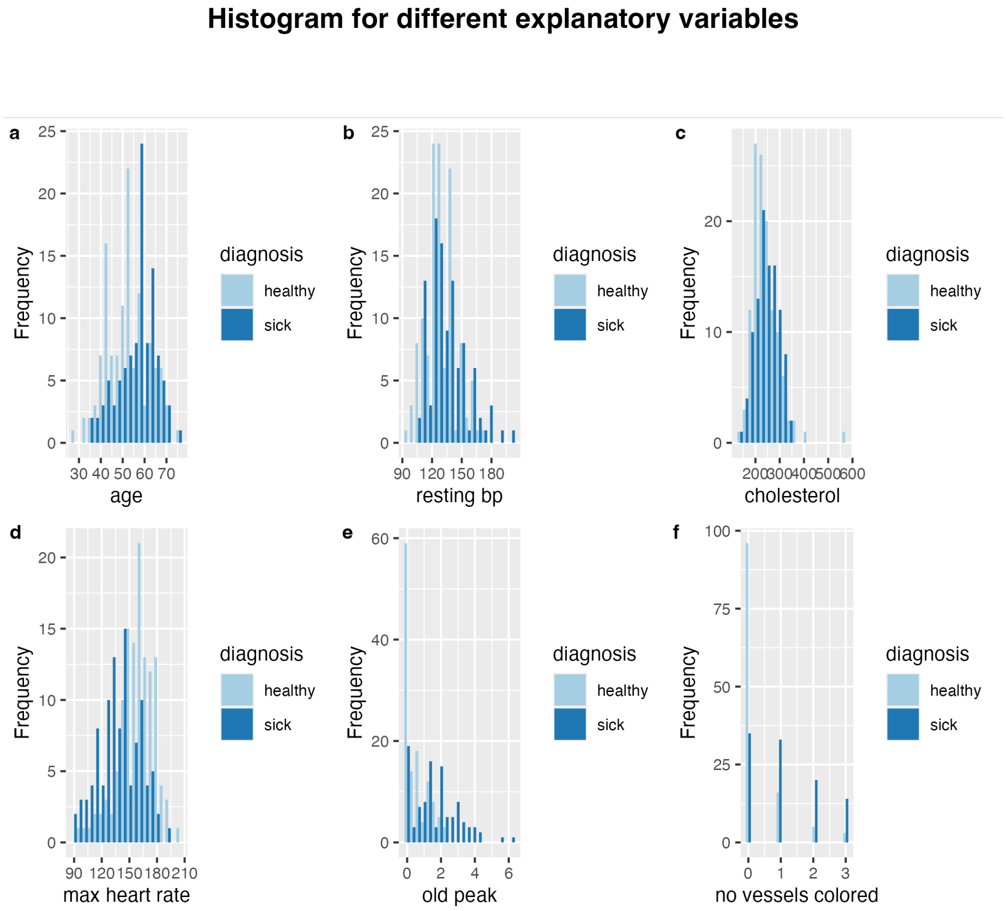

Figure 2: The distribution of value of each attribute colored by diagnosis: *(a)* Serium Cholesterol; *(b)* Age; *(c)* Resting Blood Pressure; *(d)* Max Heart Rate; *(e)* Oldpeak; *(f)* Number of Colored Vessels

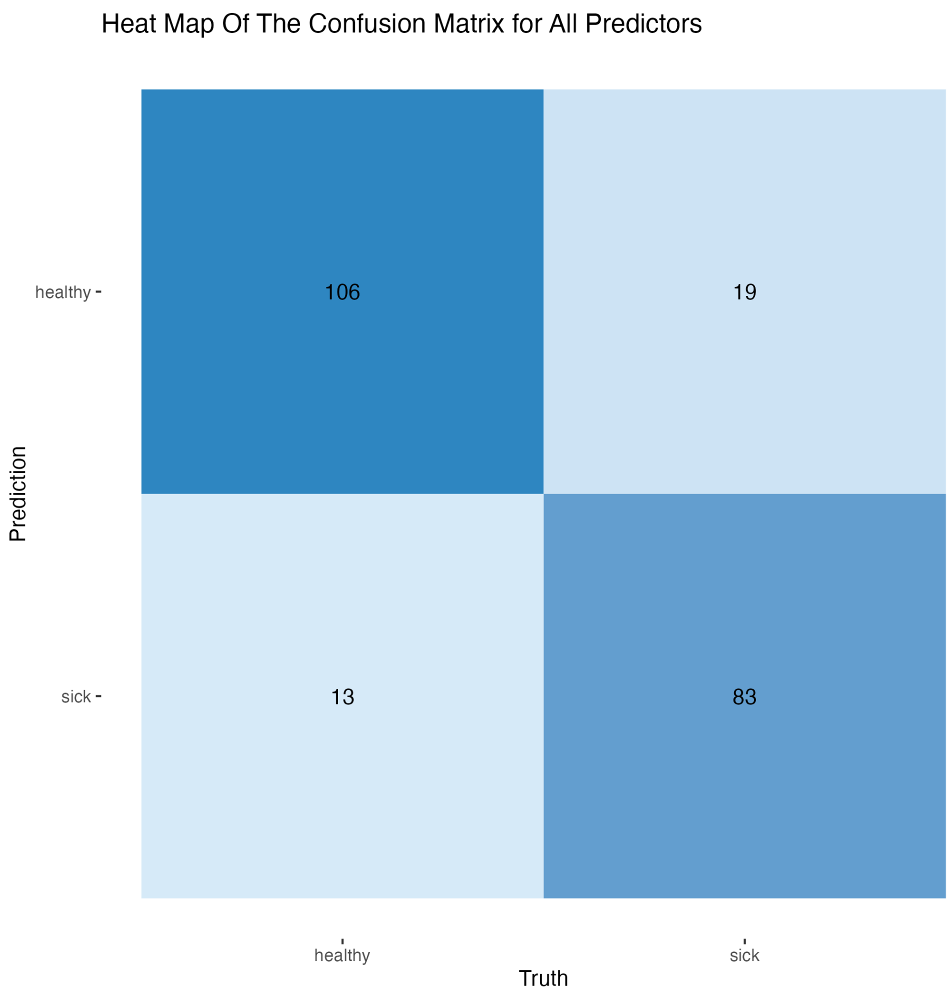

Figure 3: The Confusion Matrix for all predictors

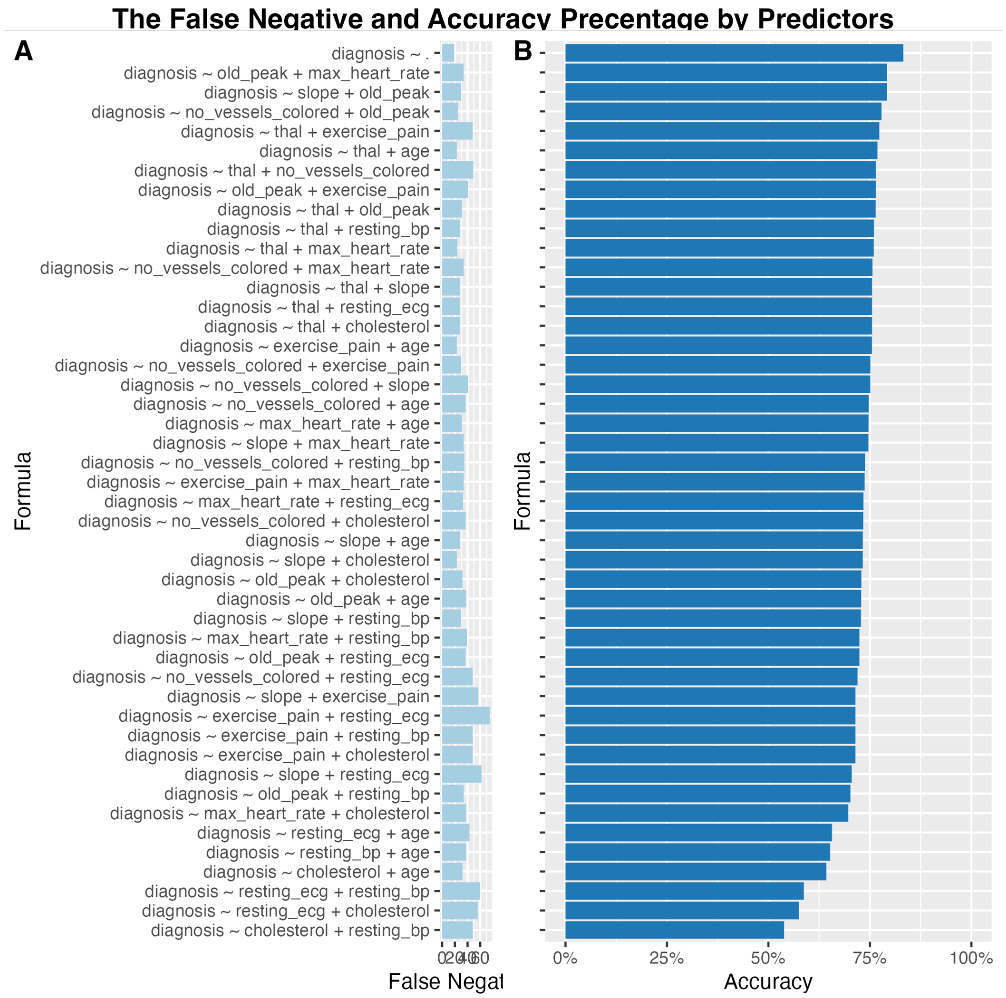

Figure 4: The *(a)* False Negatives value and *(b)* Accuracy percentage of different combinations of predictors

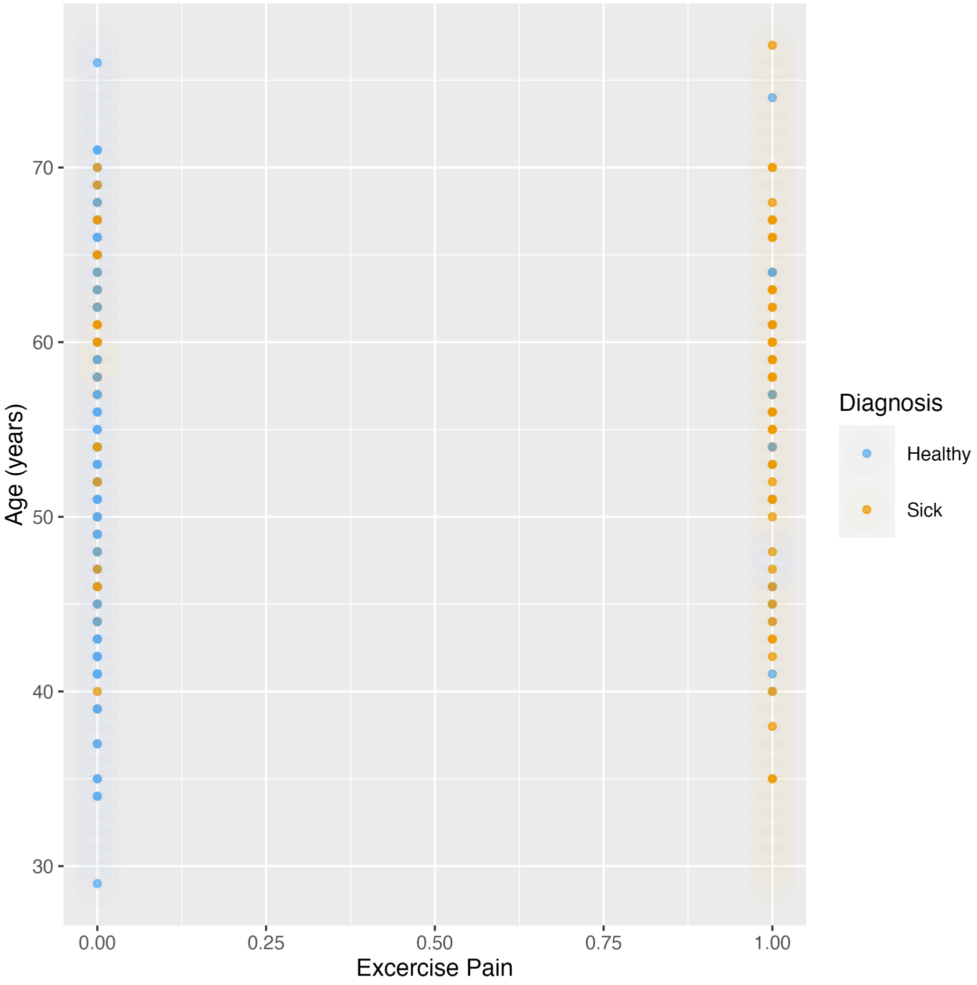

Figure 5: Fitting of data with ‘exercise pain‘ and ‘age‘ as predictors