Harvard Capstone Project: Own Choice Project (edx HarvardX: PH125.9x)

Paul Daniel VLADU

2019-06-06

Project - Heart Disease Prediction

Executive Summary

This document proposes a classification solution for a dataset regarding different factors that could be used to identify the presence of heart disease. The dataset is from Kaggle web site : https://www.kaggle.com/ronitf/heart-disease-uci .

Though the dataset is provided on Kaggle, since Kaggle requests login credentials to get this dataset, the alternative solution was to upload this dataset on GitHub in order to have it accessible for download in the R scripts of this project. For this reason, the dataset can be found and downloaded in the file “heart.csv” in the GitHub folder : https://github.com/PaulVladu/HarvardProject/tree/master/OwnChoice

The data is about Heart Disease and it is a subset of the original database. The dataset has 303 records and 14 variables : 13 attributes (variables) describing different conditions determinant for identifying the presence of heart disease and the class label that is to be predicted, displaying if a heart disease is present or not, for the respective patient.

The variables (attributes) of the dataset are presented in the following table:

Index	Attributes	Description
1	age	the age of the patient
2	sex	the sex of the patient
3	cp	chest pain with 4 posible values (0,1,2,3)
4	trestbps	resting blood pressure
5	chol	serum cholestoral in mg/dl
6	fbs	fasting blood sugar bigger than 120 mg/dl
7	restecg	resting electrocardiographic results (values 0,1,2)
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina
10	oldpeak	oldpeak = ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels (0-3) colored by fluorosopy
13	thal	a blood disorder called thalassemia, with the following possible values:

- 1 = normal
- 2 = fixed defect
- 3 = reversable defect

14	target	the label to predict : 1 if heart disease is present or 0 if the disease is not present

The dataset will be split into a train dataset and a test dataset, the test dataset containing 20% of the records of the original data set. The objective of this project is to propose a classification model, to train it on the train dataset and then to use the model to predict the variable of interest on the test data set as if it were unknown.

The performance measure of the model will be the accuracy of the prediction. The accuracy of the prediction will be calculated by comparing the predicted labels versus the existing one in the test dataset.

Analysis section

Data Cleansing & Preparation

After running the scripts, we obtain two main data sets :

- a train dataset of 242 records , called dftrain
- a test dataset of 61 records , called dftest

The train dataset will be used to train a classification model. The performance of the model will be then measured by calculating the accuracy of the predictions on the test dataset. The model will predict [target], the variable of interest on the test dataset, as if it were unknown, and then the accuracy will be calculated for the predicted labels versus the real ones, in the test dataset.

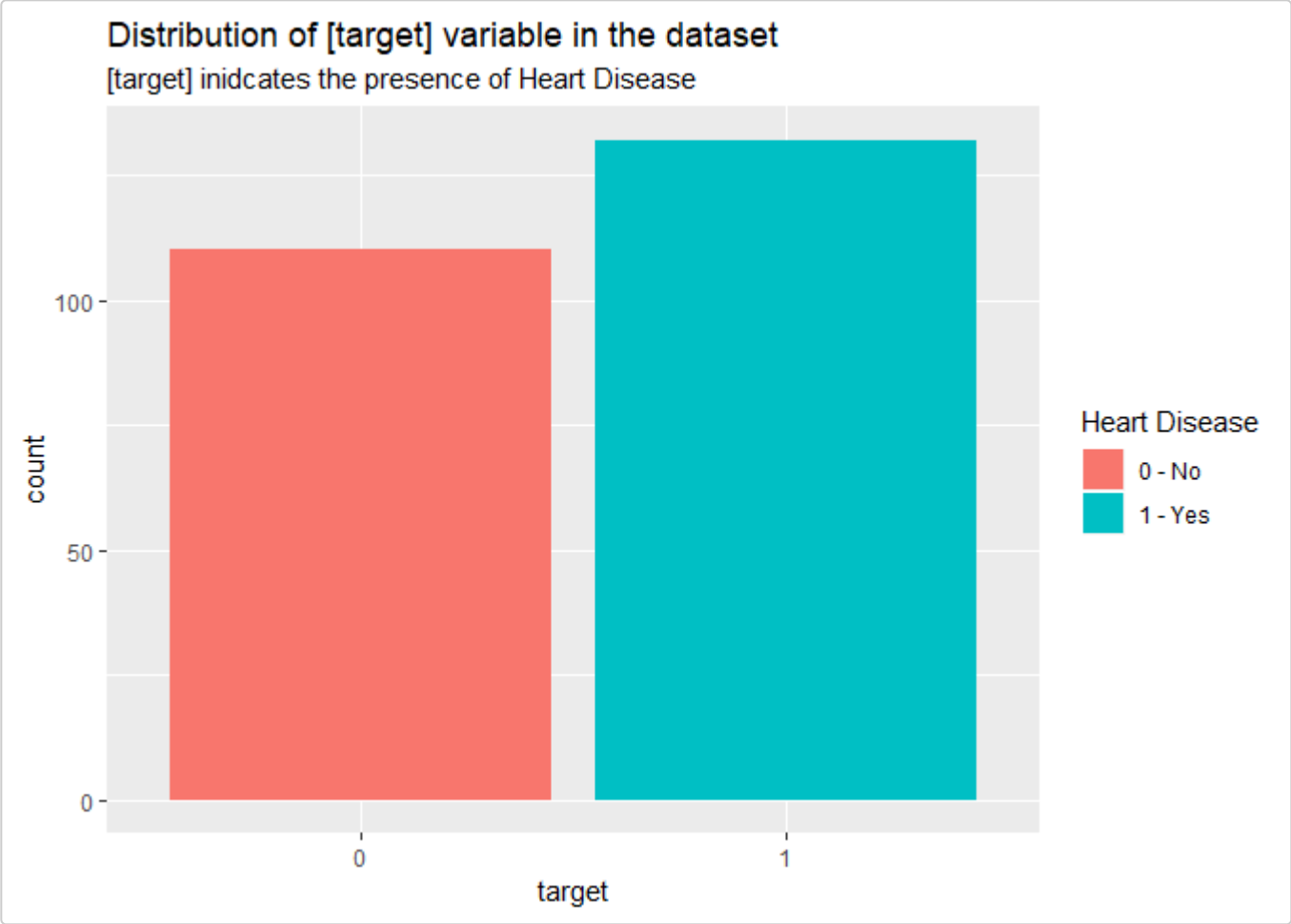
During the preparation phase we create a new feature called “age_range”, derived from the existing feature “age”, that replaces the variable “age”, and that organizes the dataset into 5 categories (or bins):

- [*_40] : describes patient up to 40 years old
- [41_50] : describes patient between 41 and 50 years old
- [51_60] : describes patient between 51 and 60 years old
- [61_70] : describes patient between 61 and 70 years old
- [71_*] : describes patient older than 71 years old

Initial Data Exploration

Variable of interest : [target]

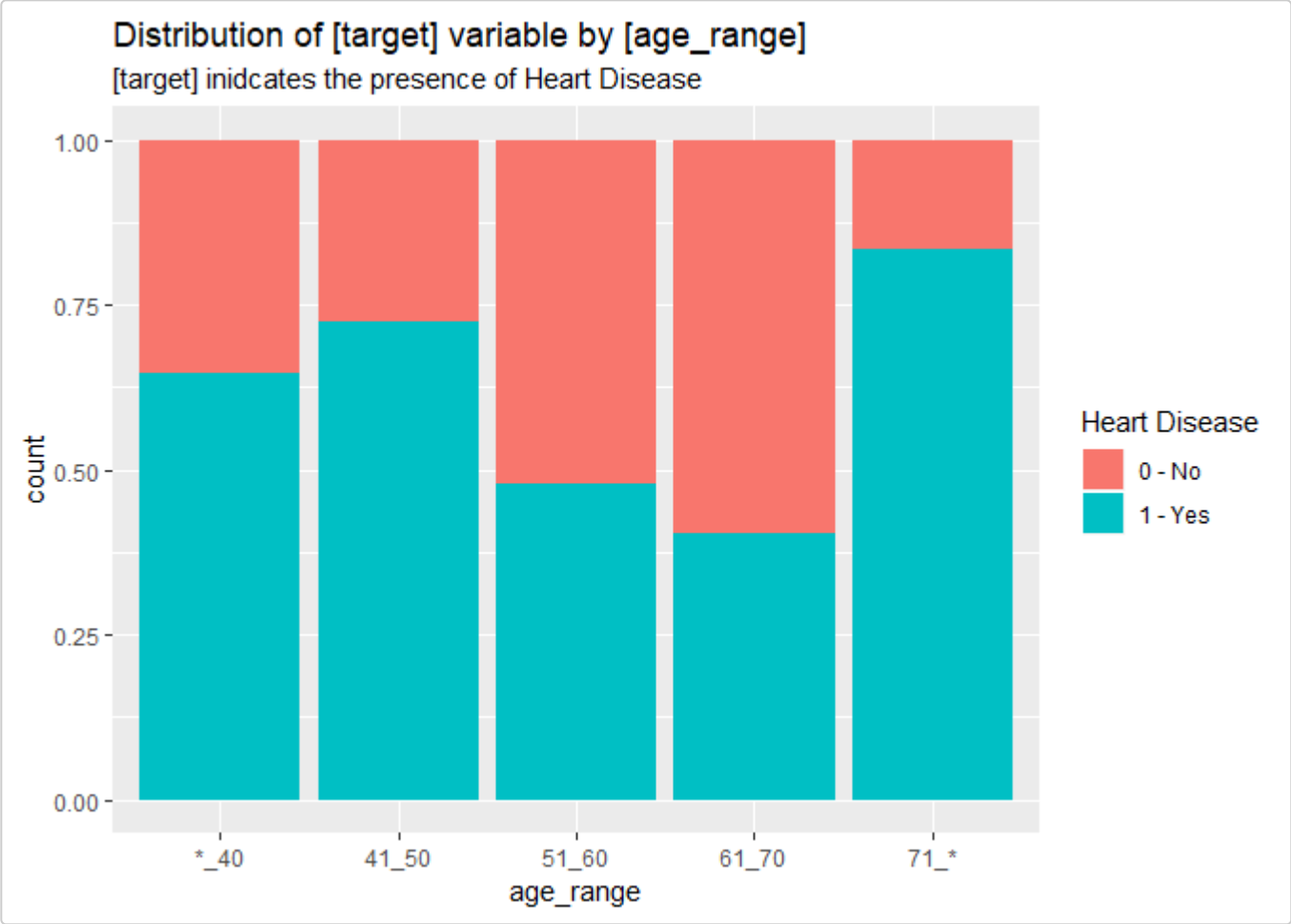
The variable of interest (to predict) is the attribute “target” in the dataset, and for this reason we start displaying a graph that presents the distribution of this variable in the train dataset.



As noticed in the graph, the label to predict is relatively well balanced in the train dataset and that is helpful in the classifying model towards a better accuracy.

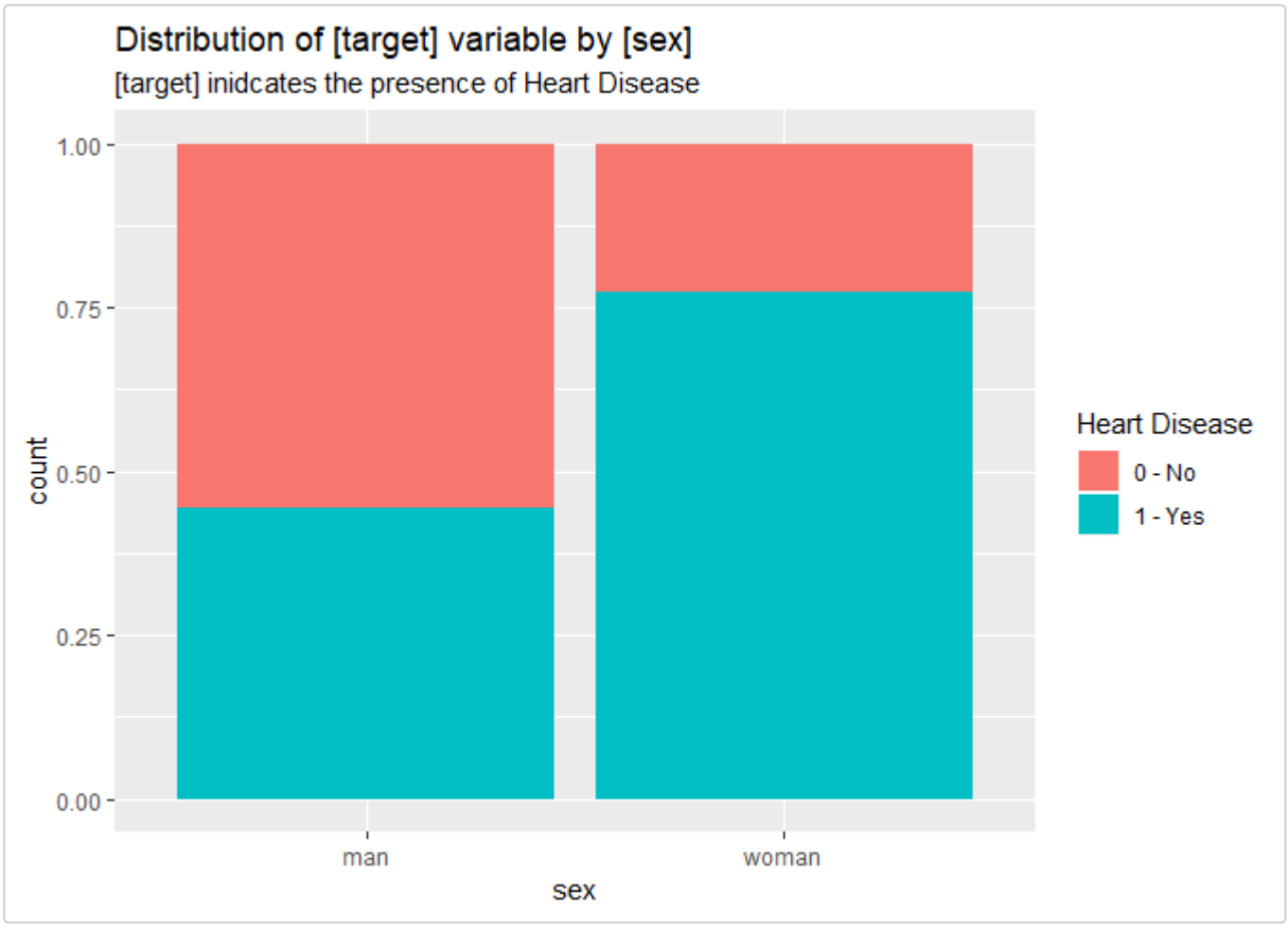
Variable : [age_range]

As one of the preprocessing operations was the binning of the age variable into 5 bins and thus engineering the new feature “age_range”, the following graph displays the distribution of the variable of interest by this new “age_range” variable.

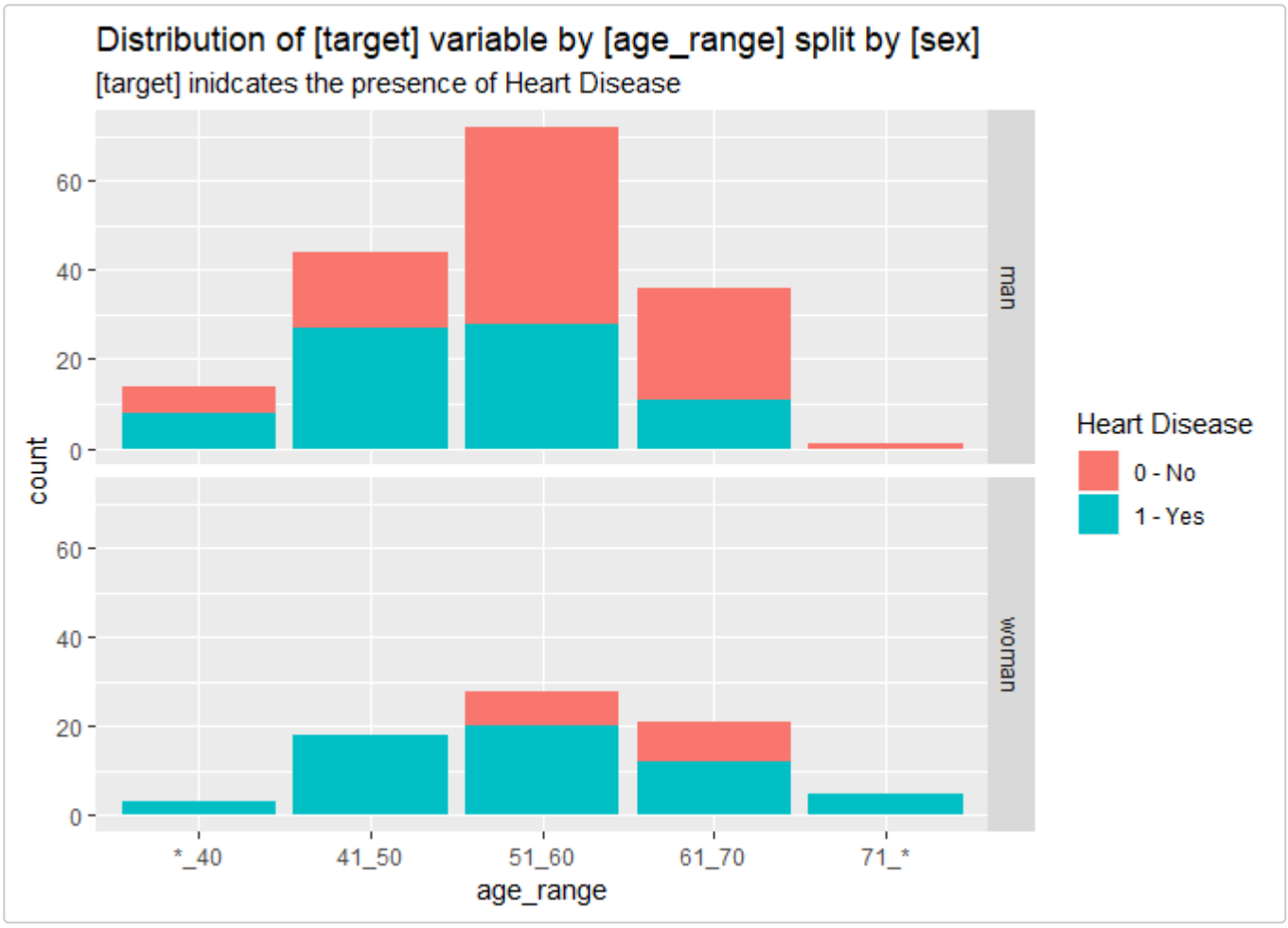


We notice that heart disease has prevalence in subjects up to 50 years old or older than 71 years old.

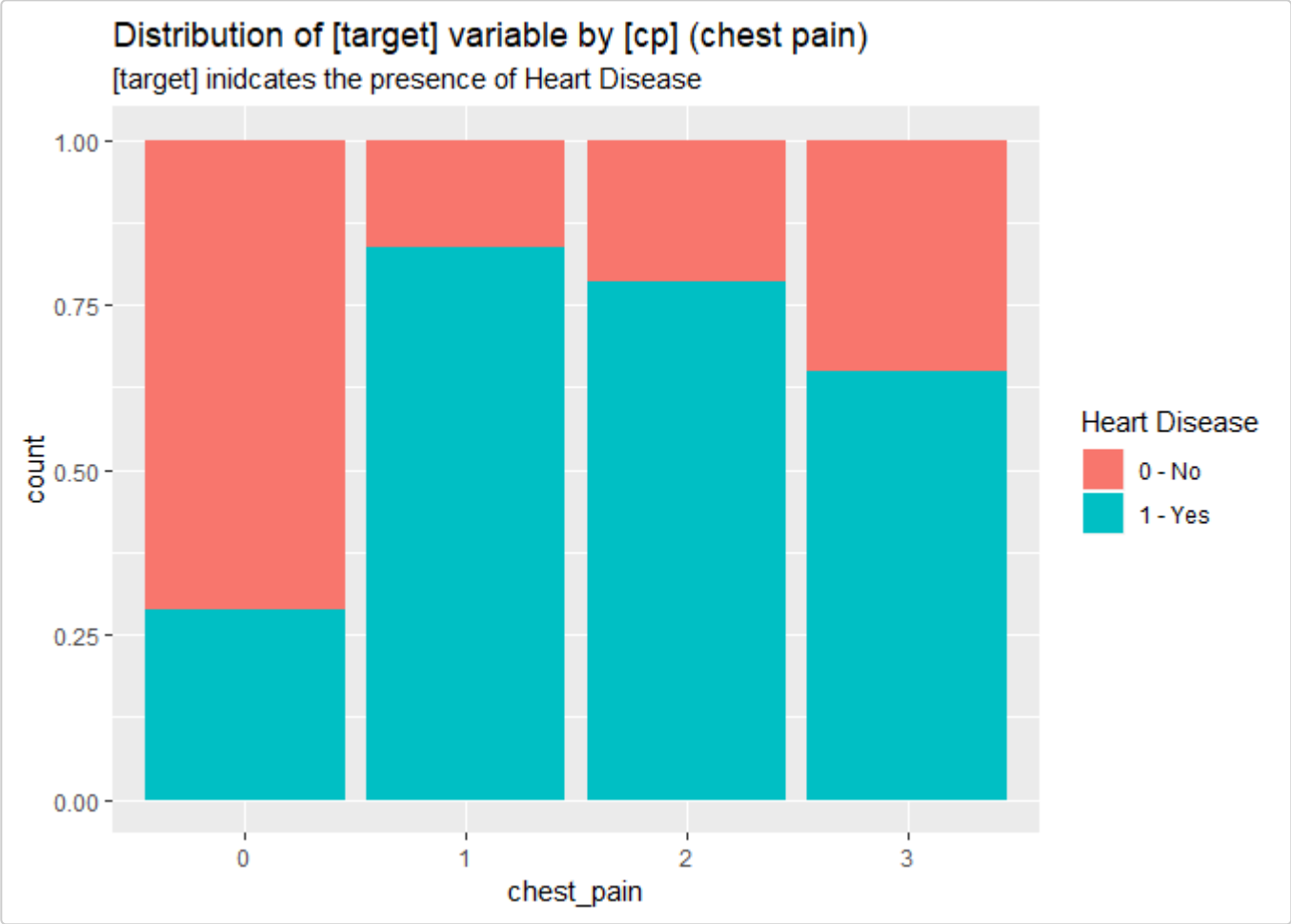
Variable : [sex]



The dataset displays a prevalence of heart disease for the women more than for men. Furthermore, in the graph below we notice that the dataset displays all women under 50 years old affected by a heart disease.

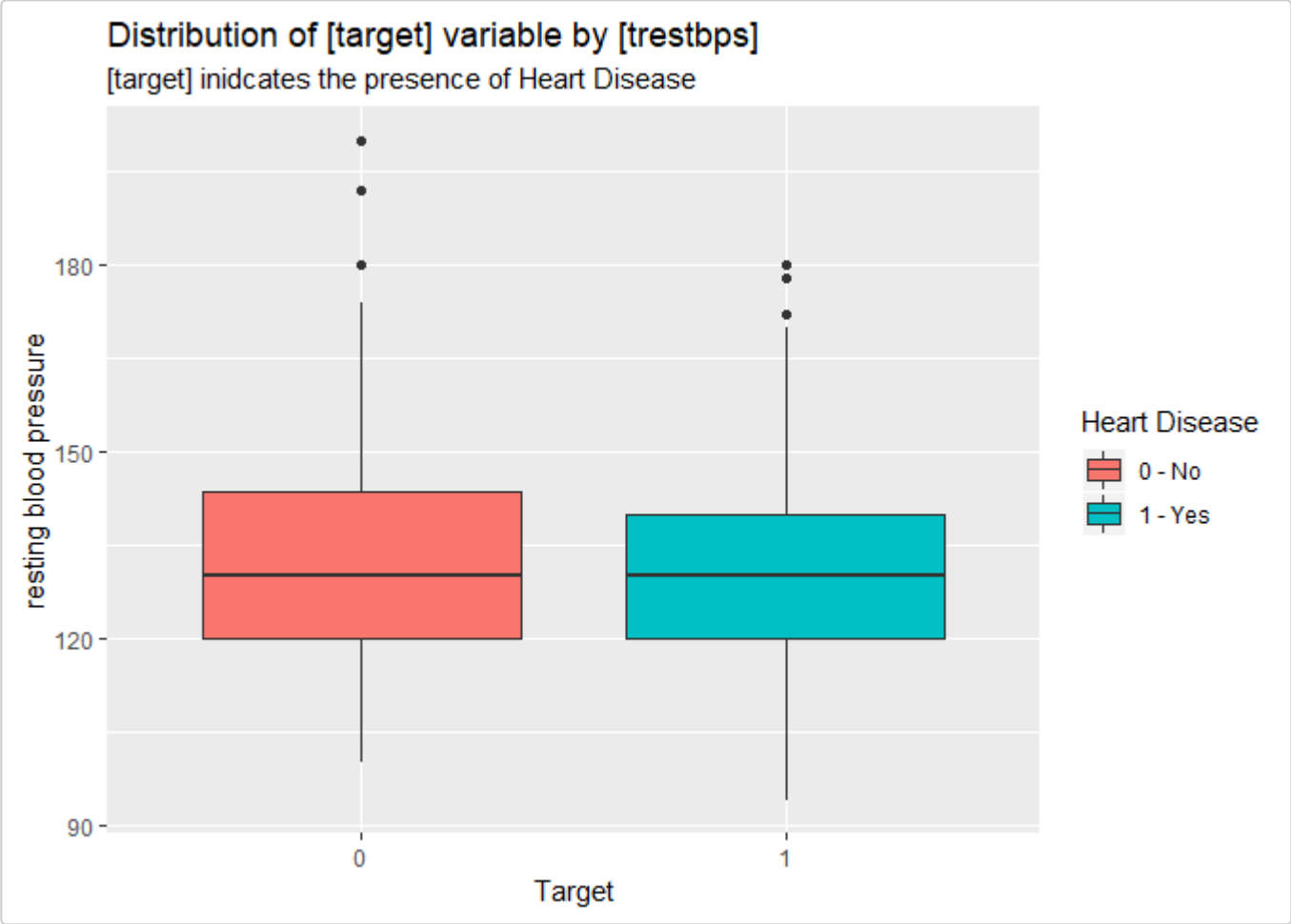


Variable : [cp] (chest pain)



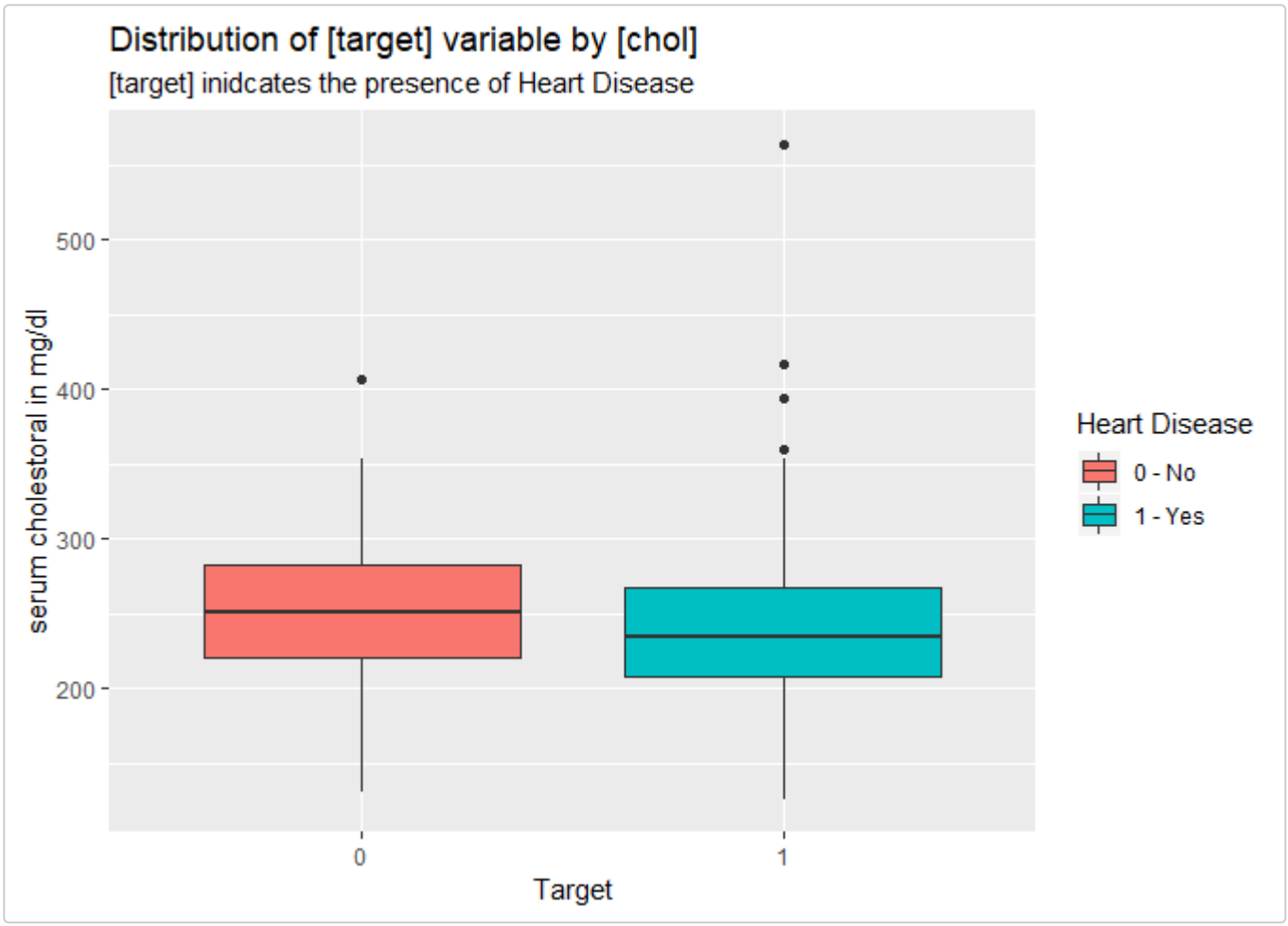
This feature seems to have quite some predicitcive power, aspect that will be presented in more details in the “Insights”" section.

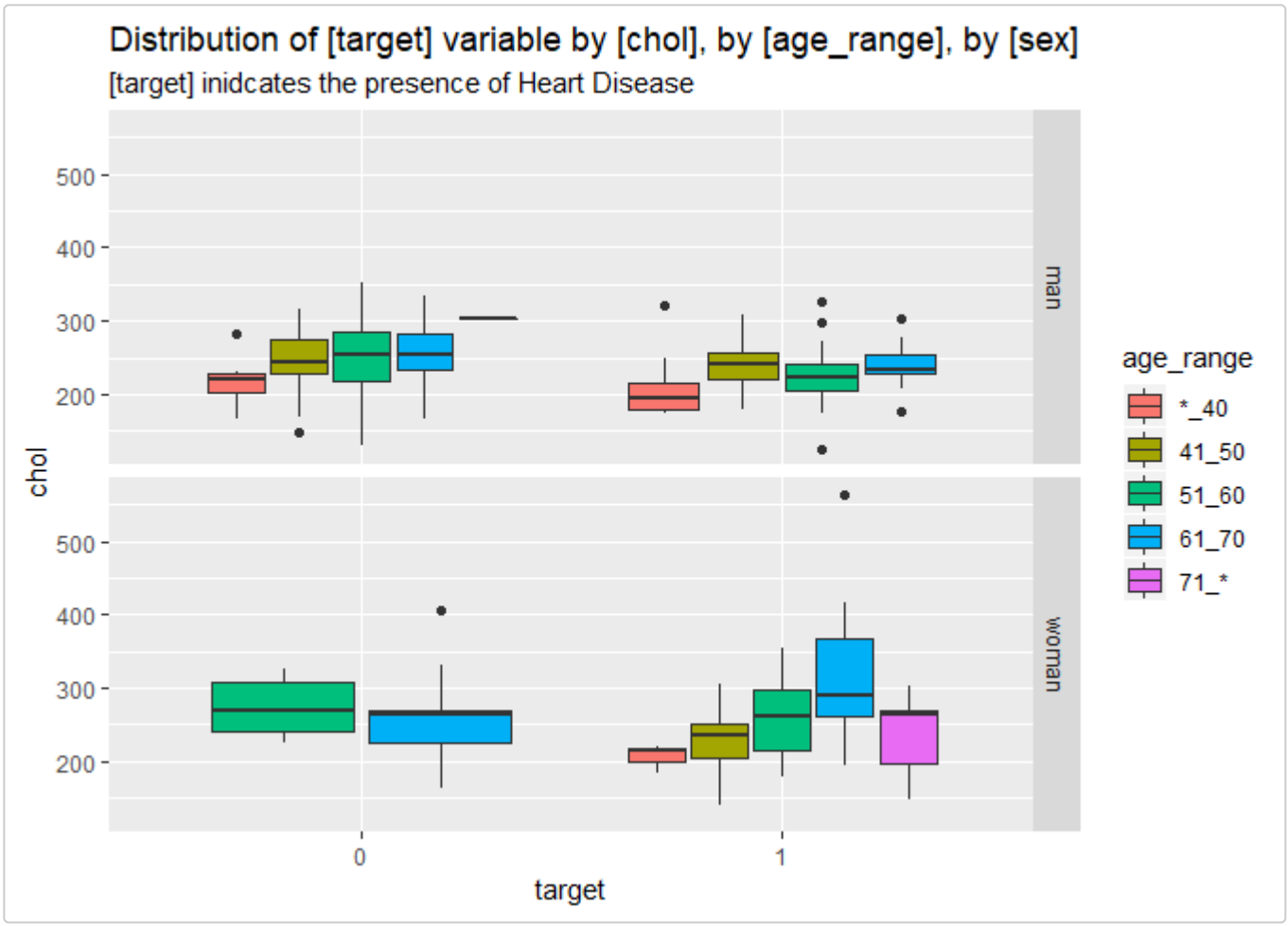
Variable : [trestbps] (resting blood pressure)



In this graph for this [trestbps] variables we do not see a much of a difference between those affected by a heart disease and those that are healthy.

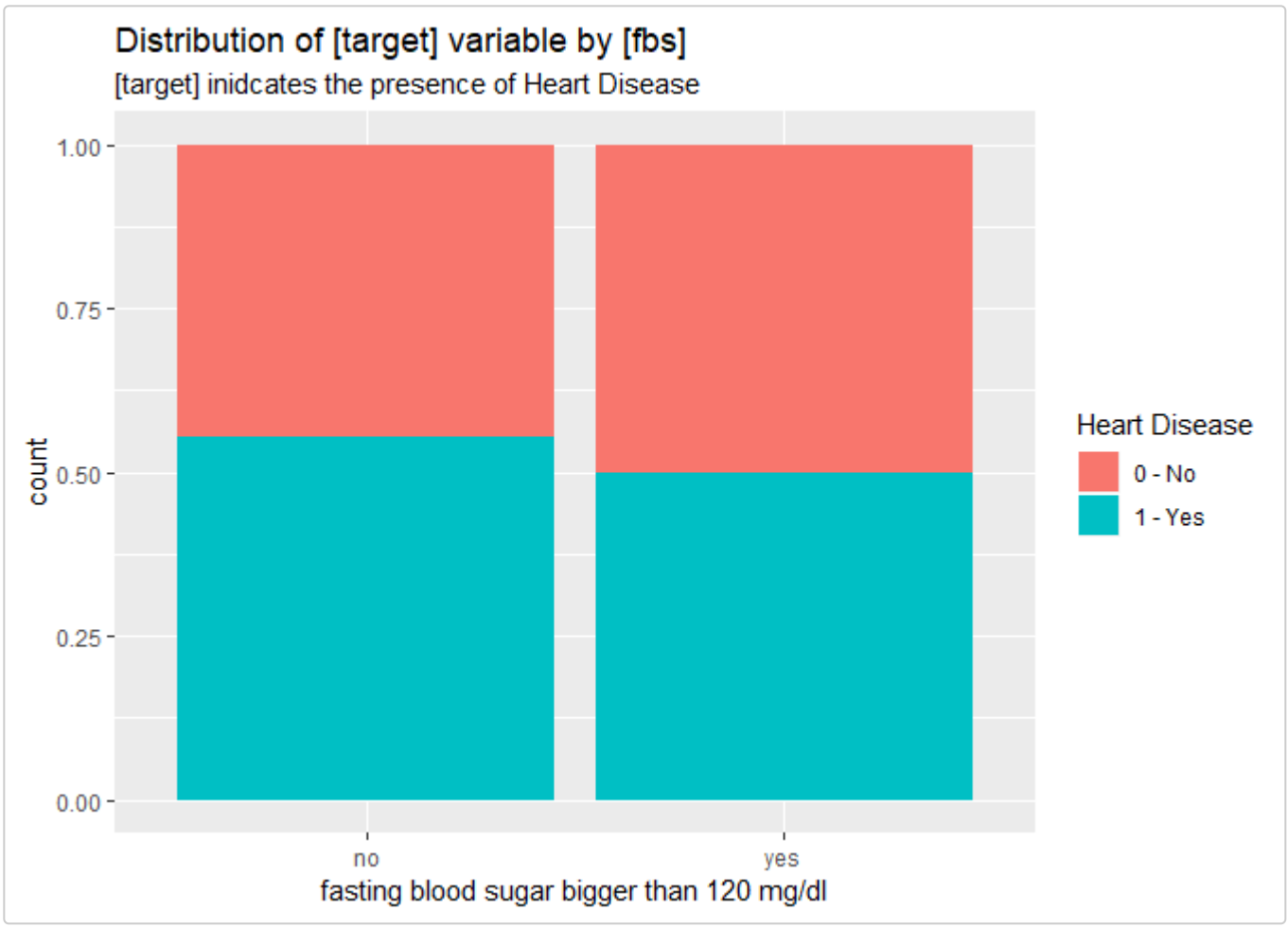
Variable : [chol] (serum cholestoral in mg/dl)


Distribution of [target] variable by [chol]
[target] inidcates the presence of Heart Disease
serum cholestoral in mg/dl
500
400
300
200
0
1
Target
Heart Disease
0 - No
1 - Yes


Distribution of [target] variable by [chol], by [age_range], by [sex]
[target] inidcates the presence of Heart Disease
chol
500
400
300
200
man
woman
0
1
target
age_range
*_40
41_50
51_60
61_70
71_*

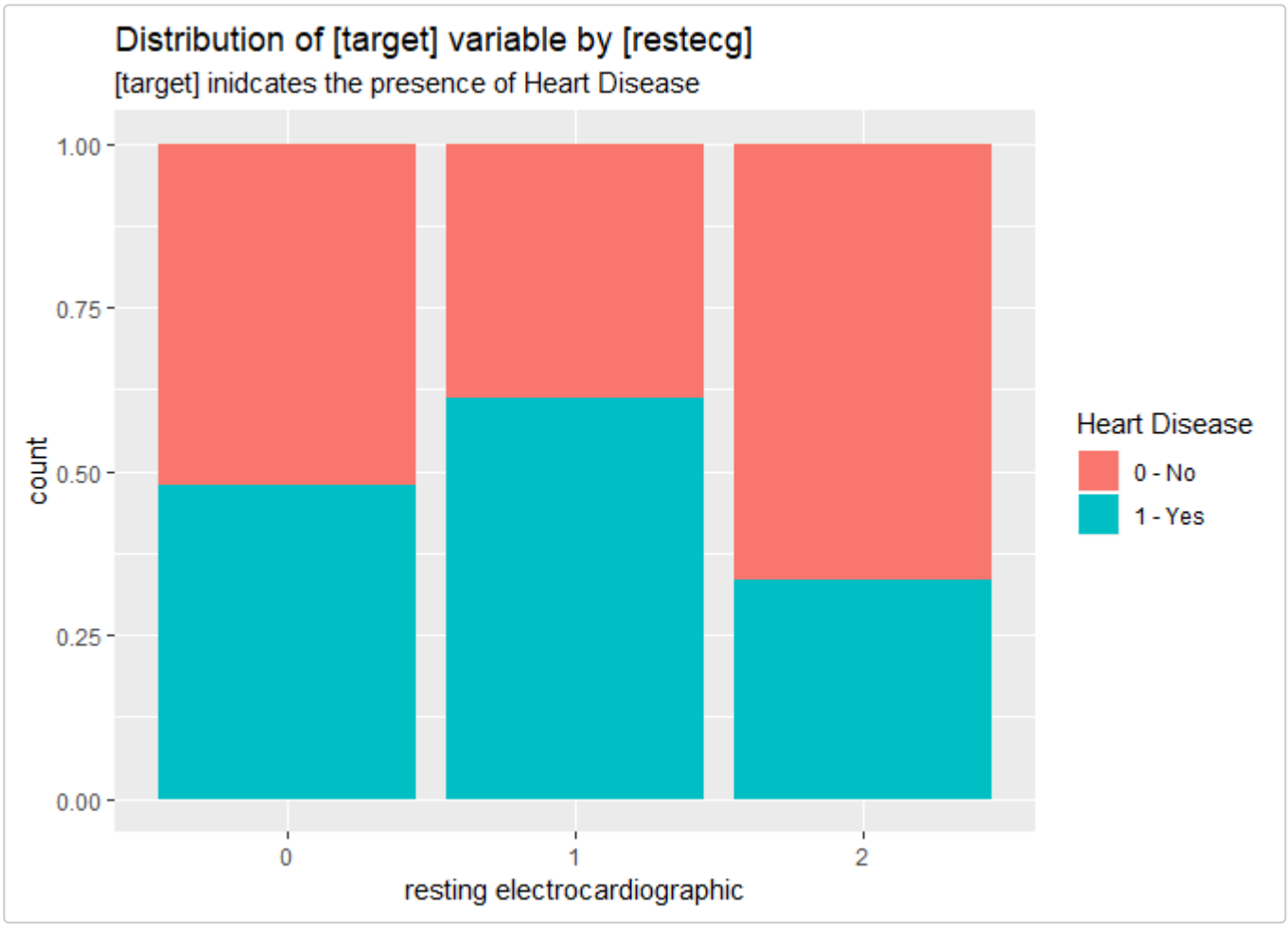
Surprisingly, this variable [chol] indicating the level of cholesterol does not offer a clear indication of a visible impact on the cause of a heart disease, even if broken down by [sex] and [age_range], though we might sense some incomplete data for healthy women.

Variable : [fbs] (fasting blood sugar bigger than 120 mg/dl)

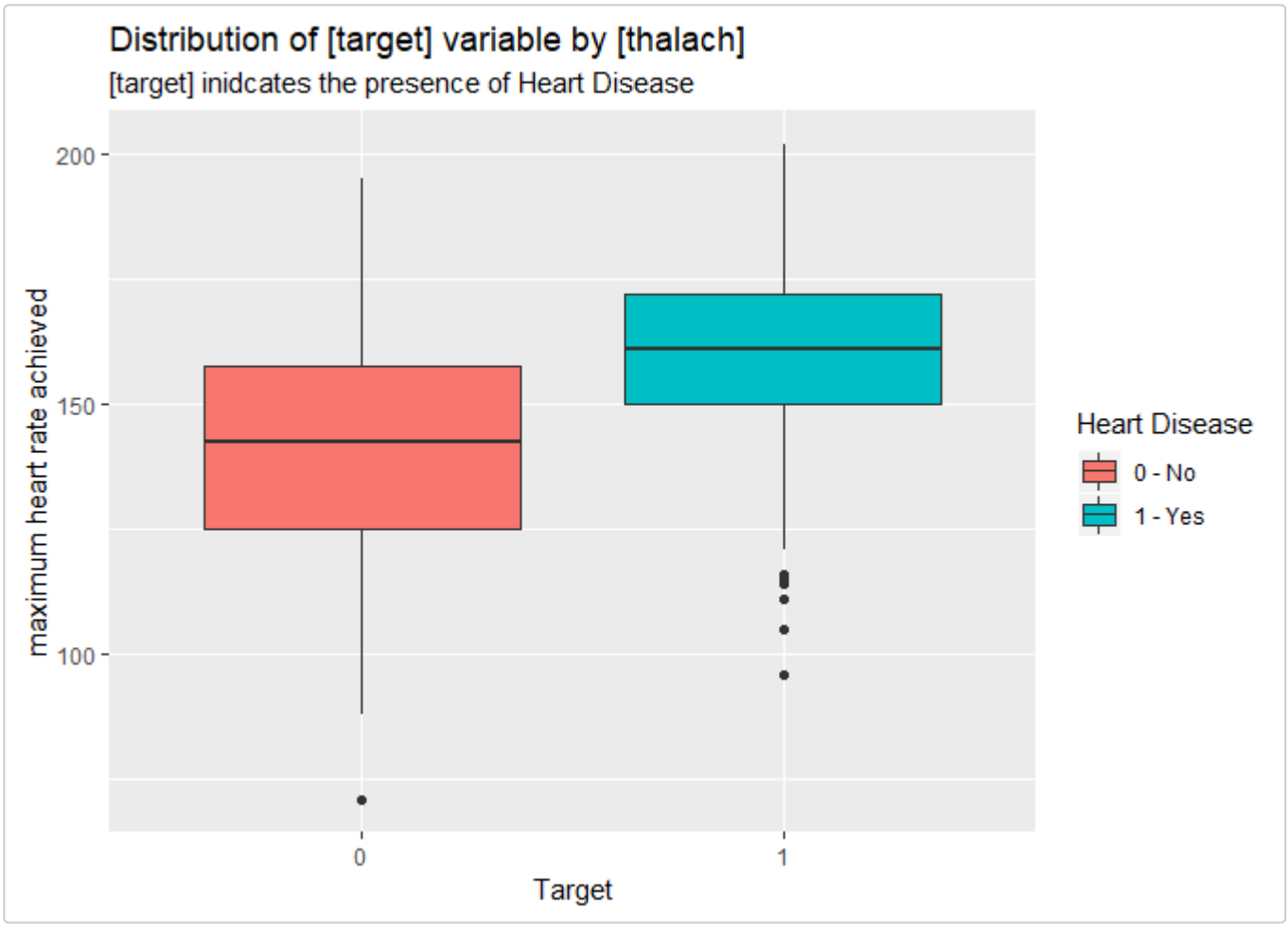


This variable [fbs] does not show too much of a difference between those with heart disease and those with no heart disease. usually, levels of fbs higher then 120mg/dl will indicate a pre-diabetes or diabetes conditions which somehow is considered to be a cause in aparition and evolution of a heart disease. However, very possible, once the diabetes condition aknowledged, it is very possible that the patientes are in constant monitoring of their health and that might be determinant in preventing a heart disease or its evolution.

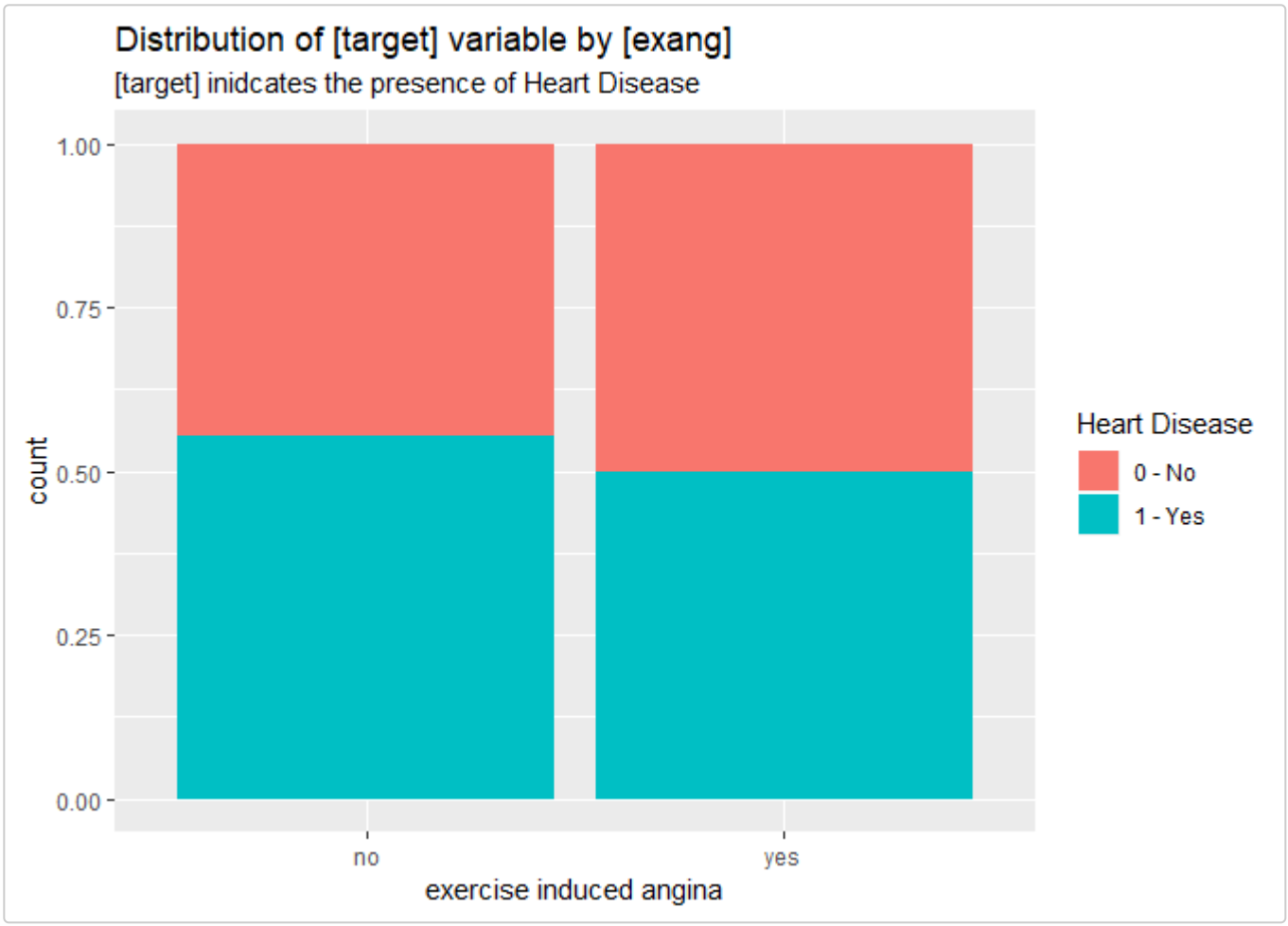
Variable : [restecg] (resting electrocardiographic)



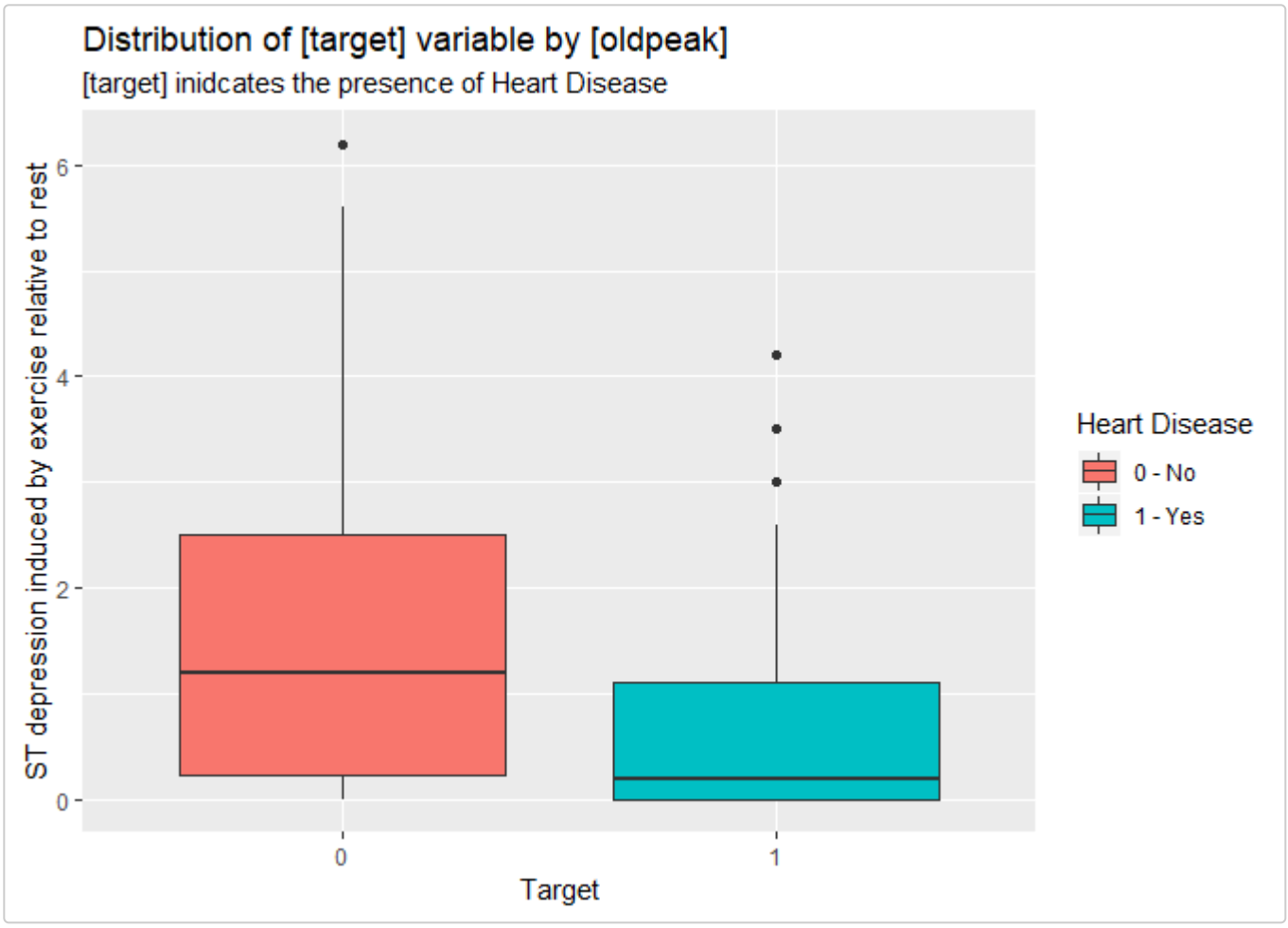
Variable : [thalach] (maximum heart rate achieved)



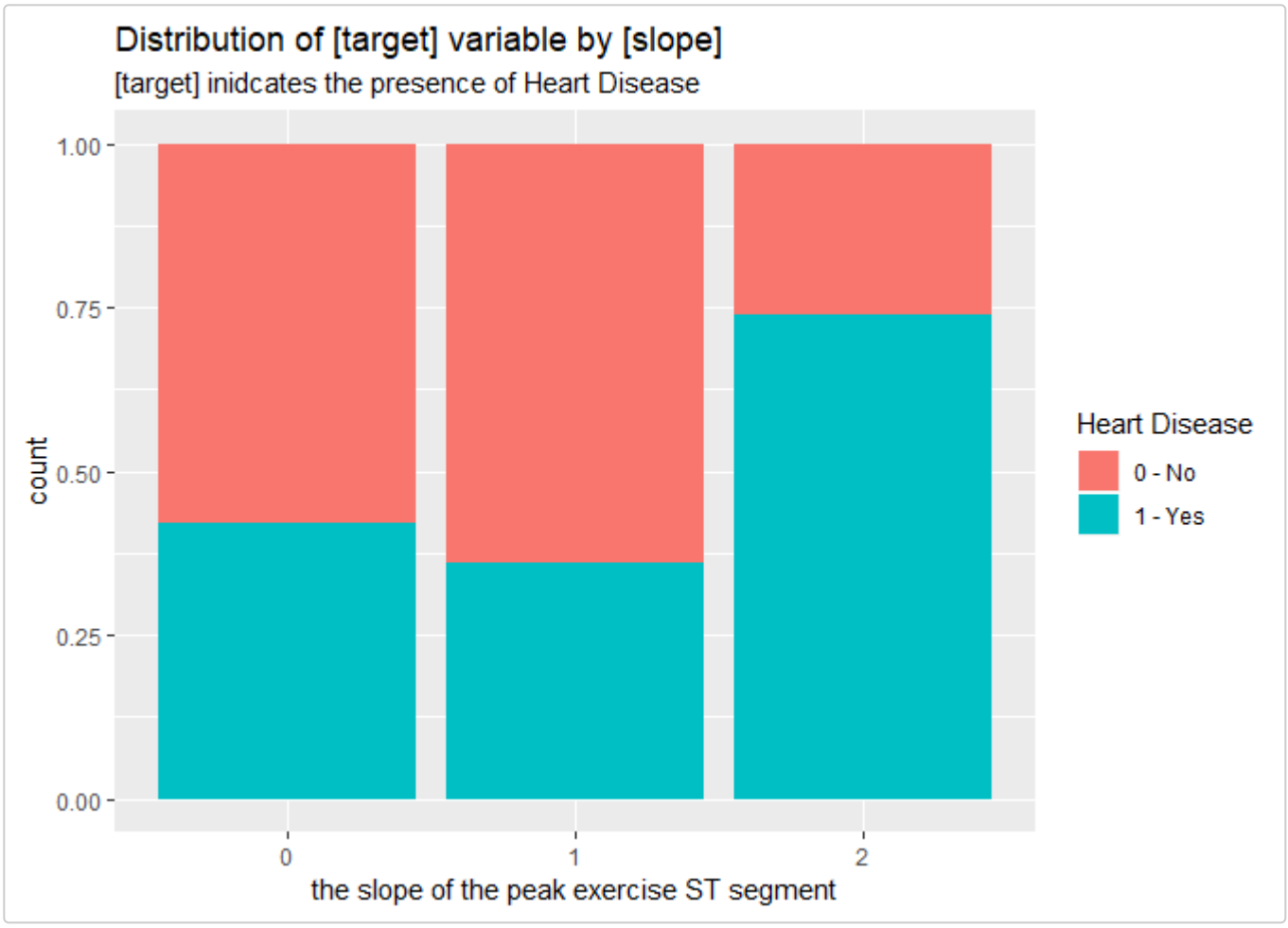
Variable : [exang] (exercise induced angina)



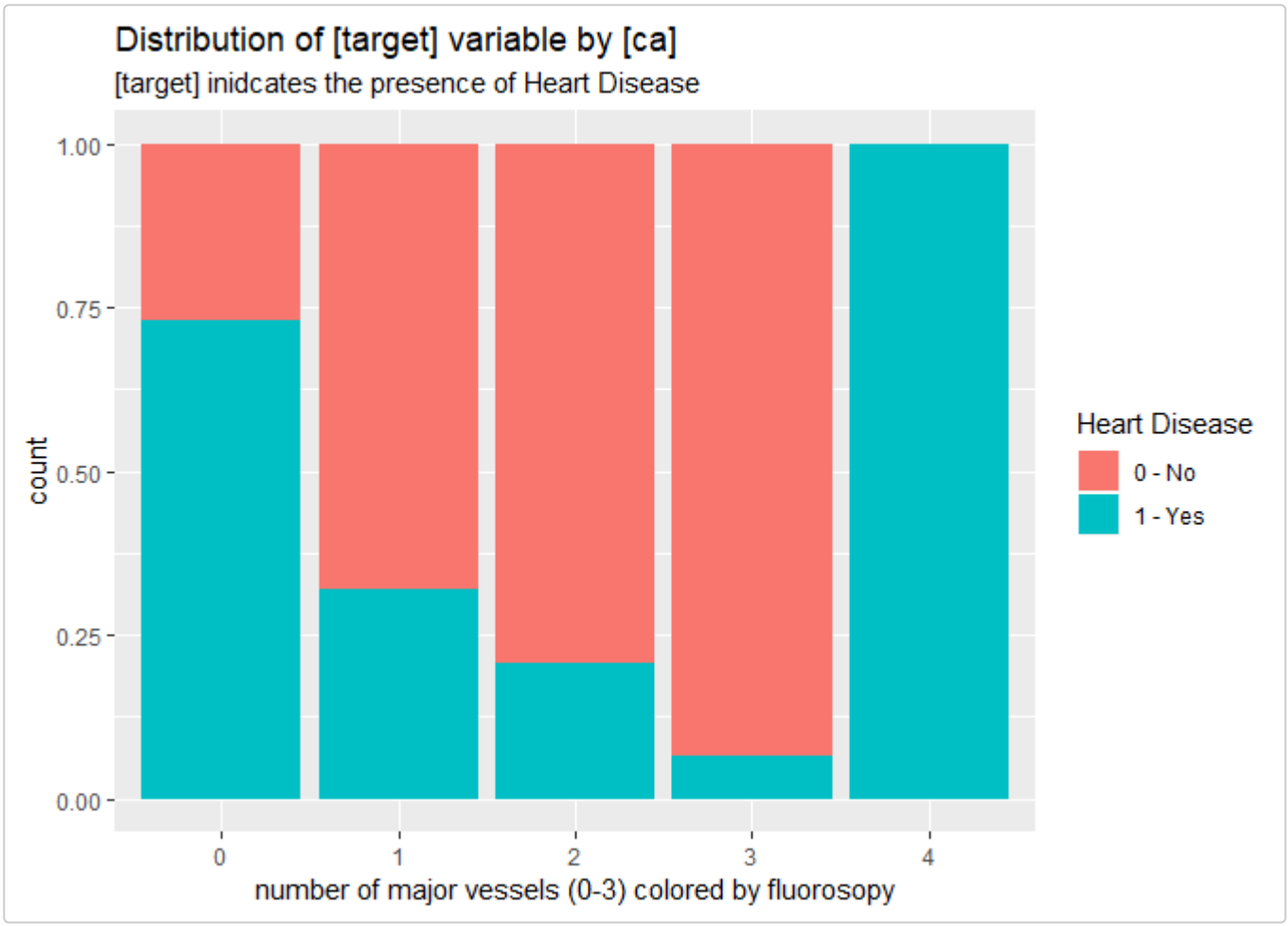
Variable : [oldpeak] (ST depression induced by exercise relative to rest)

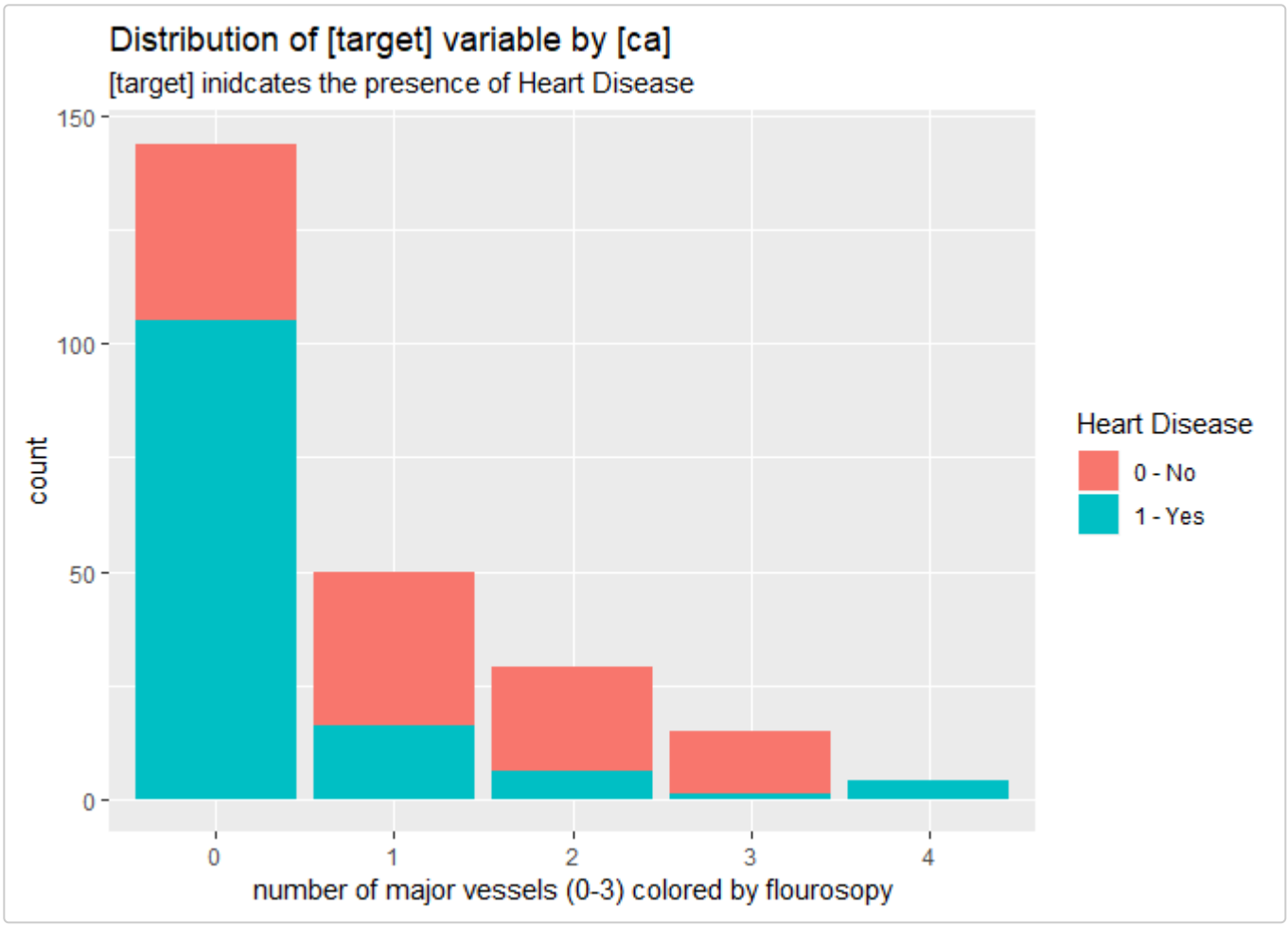


Variable : [slope] (the slope of the peak exercise ST segment)



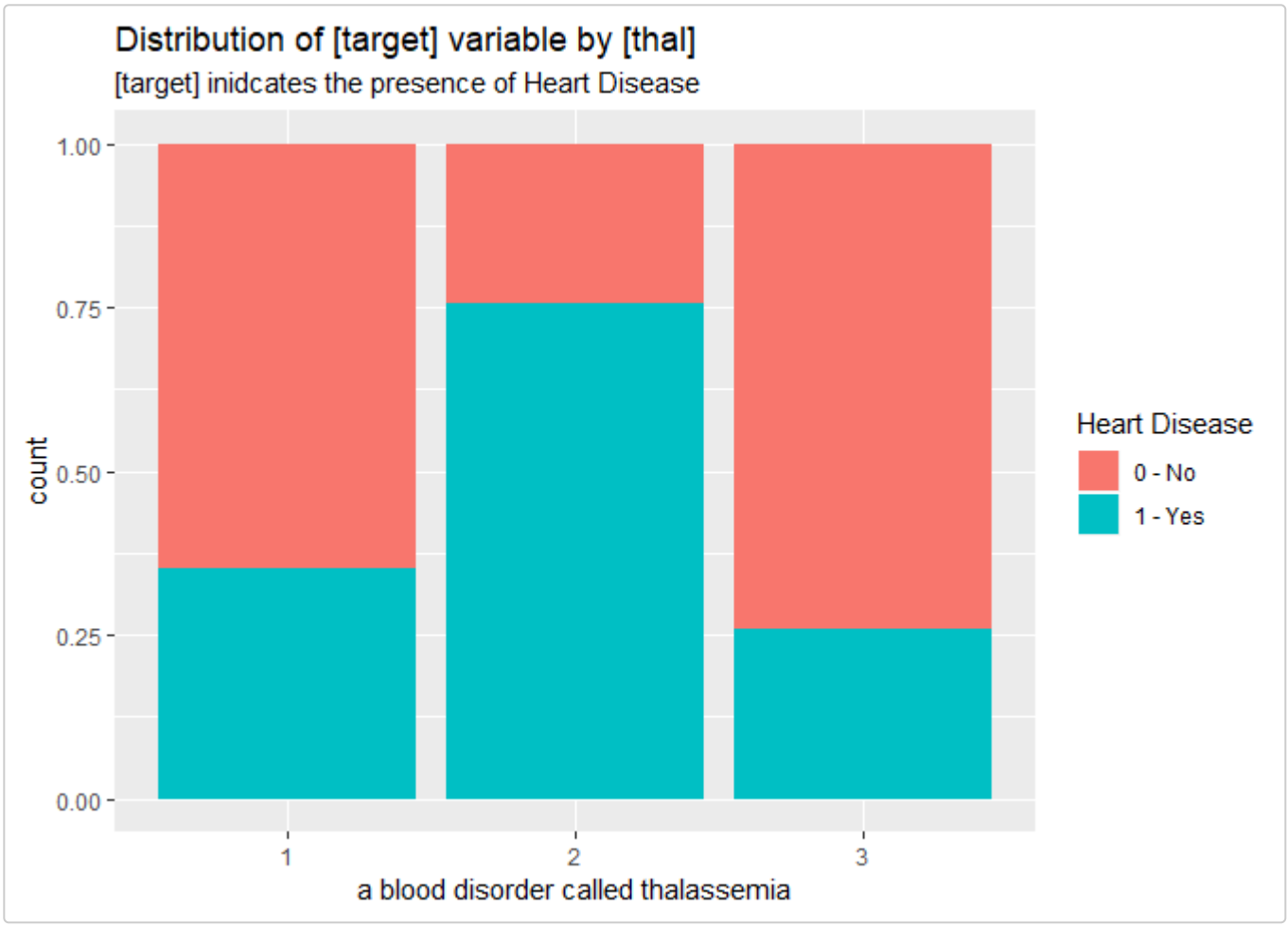
Variable : [ca] (number of major vessels (0-3) colored by fluorosopy)

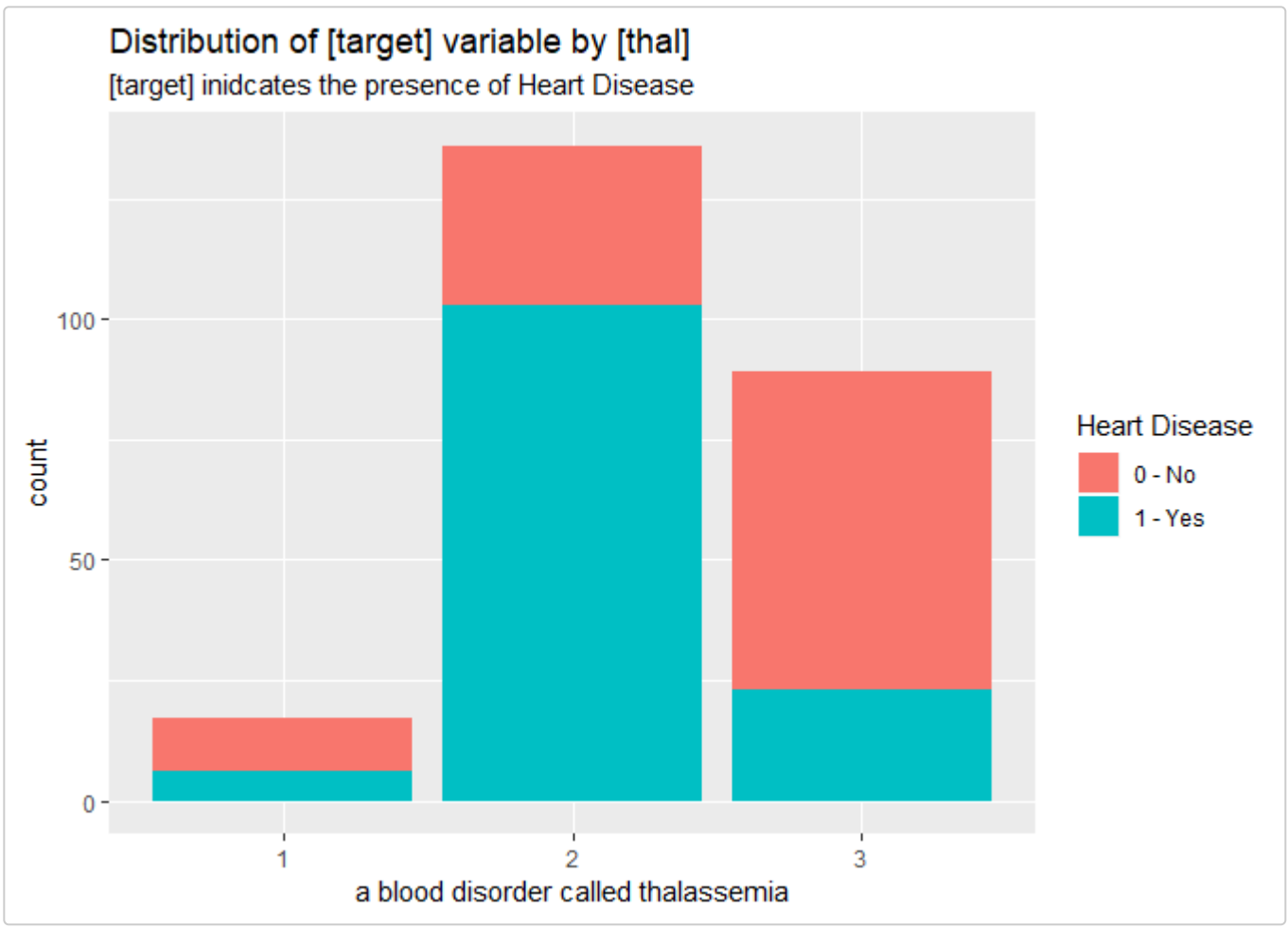

Distribution of [target] variable by [ca]
[target] inidcates the presence of Heart Disease
count
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
number of major vessels (0-3) colored by fluorosopy
Heart Disease
0 - No
1 - Yes


Distribution of [target] variable by [ca]
[target] inidcates the presence of Heart Disease
count
150
100
50
0
0
1
2
3
4
number of major vessels (0-3) colored by flourosopy
Heart Disease
0 - No
1 - Yes

From the graphs above we notice that those who have this variable [ca] = 0 are exposed to a 75% chance risk of having a heart disease.

Variable : [thal] (a blood disorder called thalassemia)


Distribution of [target] variable by [thal]
[target] inidcates the presence of Heart Disease
count
1.00
0.75
0.50
0.25
0.00
1
2
3
a blood disorder called thalassemia
Heart Disease
0 - No
1 - Yes

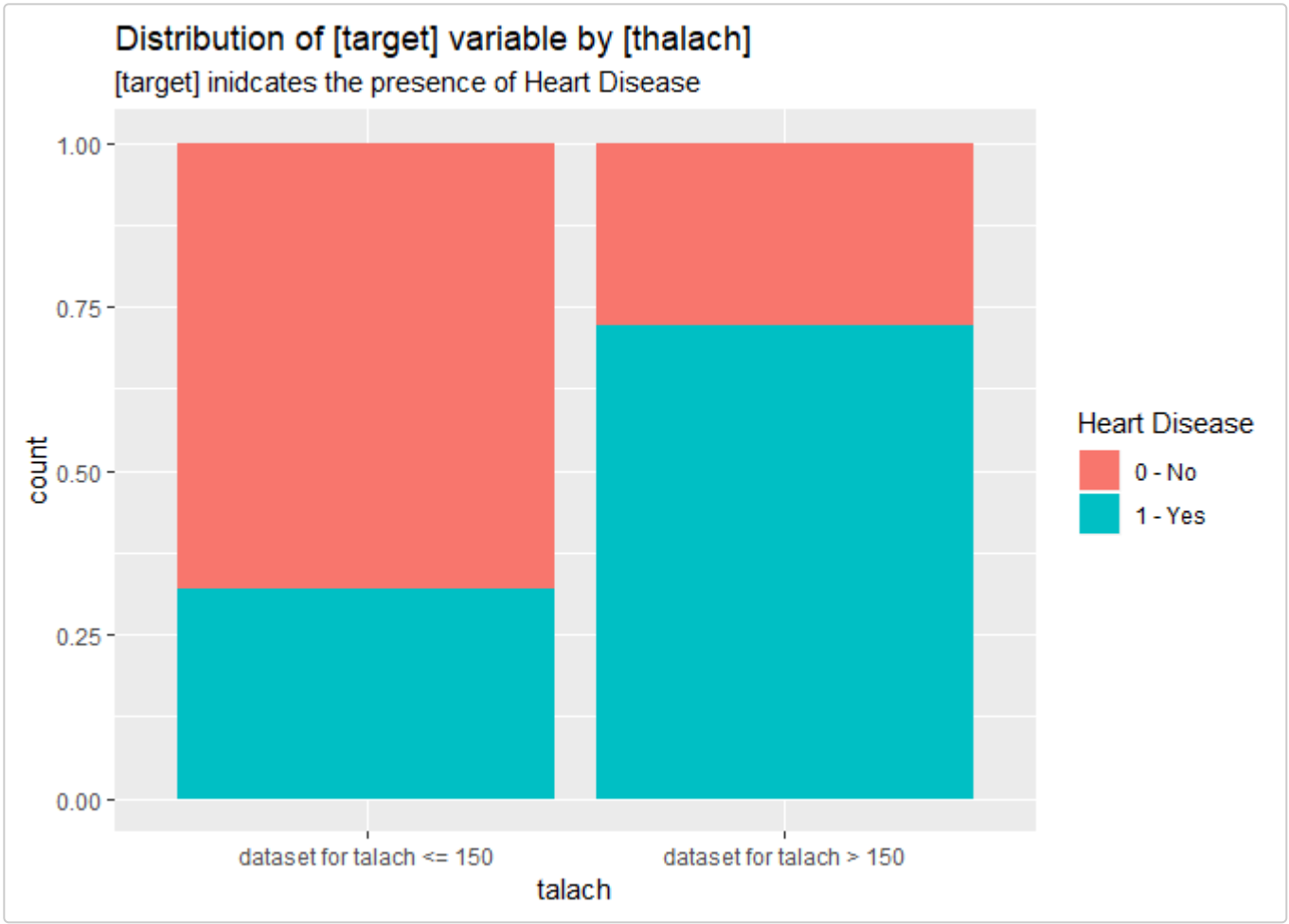

Distribution of [target] variable by [thal]
[target] inidcates the presence of Heart Disease
count
100
50
0
1
2
3
a blood disorder called thalassemia
Heart Disease
0 - No
1 - Yes

Insights gained & modeling approach

Insights gained

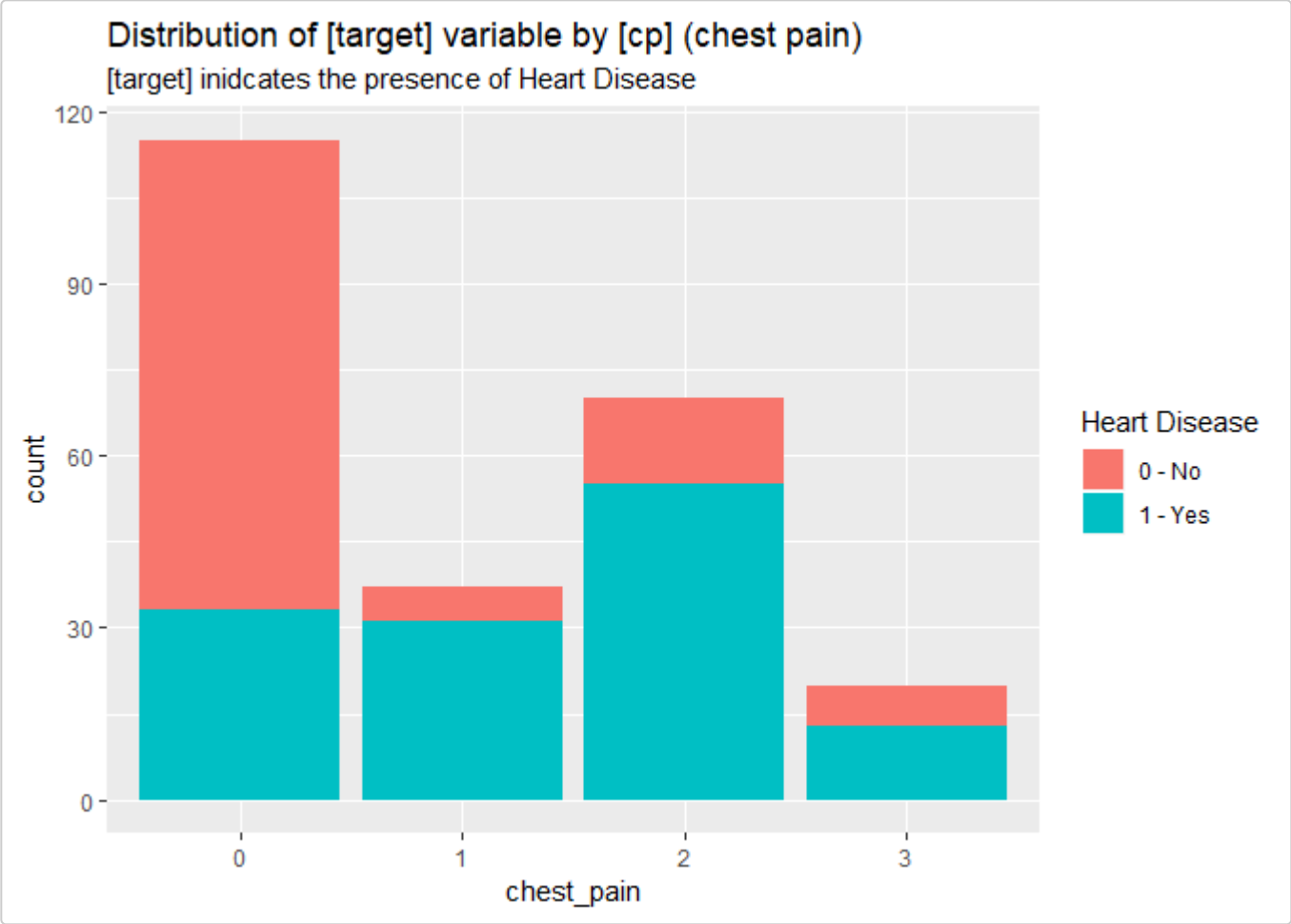
Following the graphs of the variables in the model, we can notice just at a first glance the power of prediction of some of the variables and how unimportant others could be. For example, looking at the boxplot graph for the variable [thalach] (maximum heart rate achieved), it is very clear that for people in the dataset for which this variable is bigger than 150, it is a very high chance they might have a heart disease. Also, circa 70% of those with this parameter thalach<=150 do not have heart disease.

Indeed, if we draw the graph for the train dataset, with the dataset split in two, those with thalach>150 and the rest, we can notice the above mentioned insights.



A similarly powerful variable in the model, easy to notice at first glance, it the variable [ca] (number of major vessels (0-3) colored by fluoroscopy) where those that have this parameter ca=0 could be 75% at risk of having a heart disease.

Another variable that visibly has power of prediction, is the variable [cp] (chest pain) - just looking at its graph in the section “Initial Data Exploration” we can conclude that people that proves chest pain of type 1 or 2 have at least 75% chance of having a heart disease meanwhile, circa 75% people with no chest pain (cp = 0) are actually healthy. Furthermore, if we plot the graph of counts distribution of the population by type of chest pain, we realize that the majority of the population in the dataset has the variable cp with the values 0,1 or 2 - and from here the visualization gives us a clue about the power of prediction for this variable [cp]



So far we just saw some example about how using visualization we can get some strong clues about what kind of power of prediction some variable might have in the final model.

At the opposite perspective, we can use the same visualization items in the section “Initial Data Exploration” to get an idea of how unimportant some other variable might be, thus getting a first impression about what variable could be left out since they do not bring too much benefits in the prediction process.

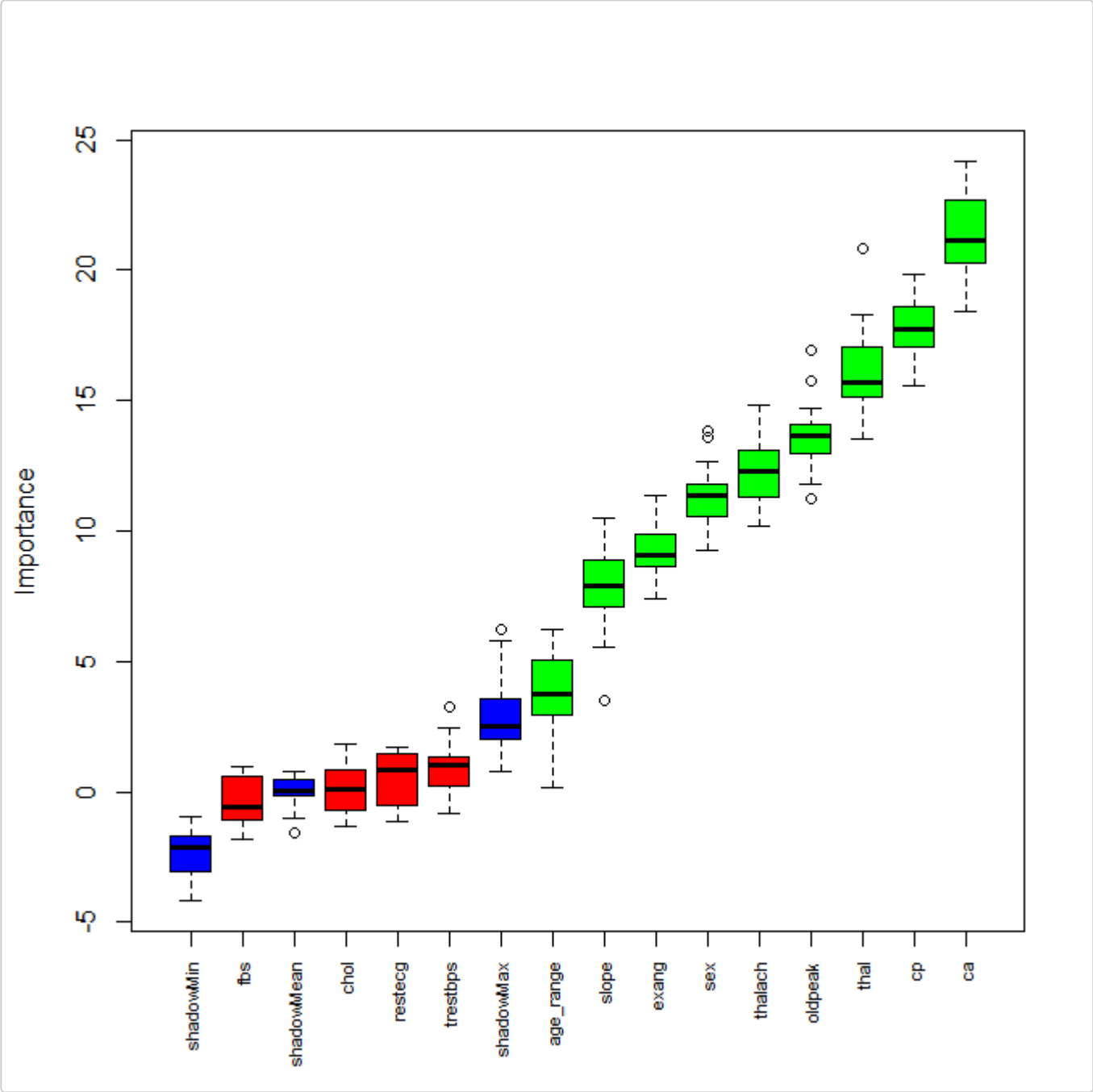
In this regard we can notice the graphs for the variable [trestbps] (resting blood pressure) , [chol] (serum cholestoral in mg/dl) and [fbs] (fasting blood sugar bigger than 120 mg/dl) and we notice that they do not make too much of a difference between those who have a heart disease and those who do not have it.

Features selection

In the effort to make this features selection more accurate, and thus to move from simple observation on the presented vizualizations to a more scientific way to select the features, the current project makes appeal to an R package developped specifically for features selection, the package “Boruta”. Using this package, we can calculate and then plot features importance in the model, thus obtaining the following visual:

```
#>  1. run of importance source...
#>  2. run of importance source...
#>  3. run of importance source...
#>  4. run of importance source...
#>  5. run of importance source...
#>  6. run of importance source...
#>  7. run of importance source...
#>  8. run of importance source...
#>  9. run of importance source...
```

```
#>  10. run of importance source...
#>  11. run of importance source...
#> After 11 iterations, +0.94 secs:
#>  confirmed 8 attributes: ca, cp, exang, oldpeak, sex and 3 more;
#>  rejected 3 attributes: chol, fbs, restecg;
#>  still have 2 attributes left.
#>  12. run of importance source...
#>  13. run of importance source...
#>  14. run of importance source...
#>  15. run of importance source...
#> After 15 iterations, +1.3 secs:
#>  rejected 1 attribute: trestbps;
#>  still have 1 attribute left.
#>  16. run of importance source...
#>  17. run of importance source...
#>  18. run of importance source...
#>  19. run of importance source...
#>  20. run of importance source...
#>  21. run of importance source...
#>  22. run of importance source...
#>  23. run of importance source...
#>  24. run of importance source...
#>  25. run of importance source...
#>  26. run of importance source...
#>  27. run of importance source...
#>  28. run of importance source...
#>  29. run of importance source...
#>  30. run of importance source...
#>  31. run of importance source...
#>  32. run of importance source...
#>  33. run of importance source...
#> After 33 iterations, +2.5 secs:
#>  confirmed 1 attribute: age_range;
#>  no more attributes left.
```



From the graph above - as obtained with the package “Boruta” - we are indicated with the boxplot in green what are the variables with the most impact in the prediction.

Not surprisingly, many of these variables were noticed as having power of prediction just scanning the graphs presented in the sections above, however, the package “Boruta” not only that confirmed the insights gained through visualization but also offers a ranking of these variables by their importance, in the model.

Modeling

For the modeling, using the insights gained through visualization of different variables as well as the results from applying “Boruta” package on train dataset, we retain the following variables with the most predictive power:

- ca
- cp
- thal

- oldpeak
- thalach
- sex
- exang
- slope
- age_range

The current project not only that will predict the labels in the test dataset but also will present three different algorithms used in this problem of classification, namely :

- naive Bayes
- logisitc regression
- random forest

naive Bayes

```
#>       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
#>    0.557377049    0.147956544    0.424459512    0.684535859    0.540983607
#> AccuracyPValue  McnemarPValue
#>    0.450335944    0.002075563
```

logisitc regression

```
#>       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
#>   8.360656e-01   6.716900e-01   7.191150e-01   9.184827e-01   5.409836e-01
#> AccuracyPValue  McnemarPValue
#>   1.184168e-06   7.518296e-01
```

random forest

```
#>       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
#>   8.688525e-01   7.373520e-01   7.578413e-01   9.416395e-01   5.409836e-01
#> AccuracyPValue  McnemarPValue
#>   5.049401e-08   7.236736e-01
```

Results

After executing the chosen classification algorithms (naive Bayes, logistic regression, random forest), we have the following table with the results :

Algorithm	Accuracy	Sensitivity	Specificity	Precision
naive Bayes	0.557377	0.8214286	0.3333333	0.5111111
logisitic regression	0.8360656	0.8571429	0.8181818	0.8
random forest	0.8688525	0.8928571	0.8484848	0.8333333

Conclusion

Among the algorithms chosen for this classification problem we notice in the “Results” section the poor performance of “naive Bayes” and thus the indication of such weak results coming from the assumption of independency that “naive Bayes” algorithm relies on.

It is very clear that random forest algorithm offers the best solution for this dataset and problem, scoring high not only for accuracy but also for the other indicators: sensitivity, specificity and precision.

We conclude that the model proposed, with the 9 features identified with the Boruta package and the dependent variable “target” - the variable of interest that is to be predicted - together with the “random forest” algorithm, makes the best solution in predicting the presence of the heart disease. With the model proposed we get an accuracy of 0.8688525.