Instructor: Prof. Dr. Katharina Jahn

Introduction to Focus Areas Data Science WS22/23 Project 1

—

Group 7 Maike, Jule, Carlos, Abhinav

Goals

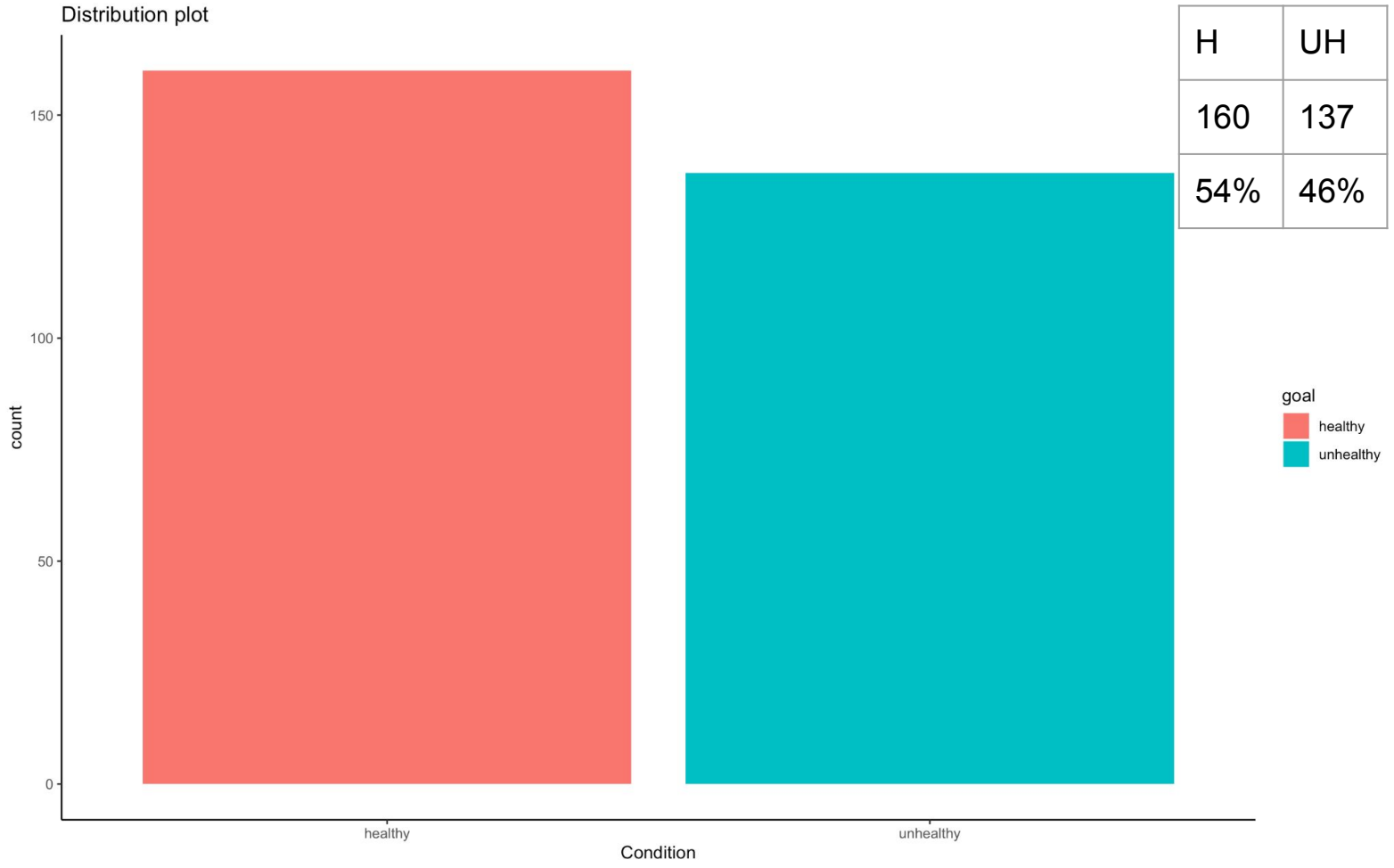
1. Exploratory data analysis
2. Classification
3. Performance metrics

Heart Disease (Cleveland)
Instances: 303
Attributes: 14 out of 76
Type: categorical, numeric, integer
Data Type: continuous, discrete

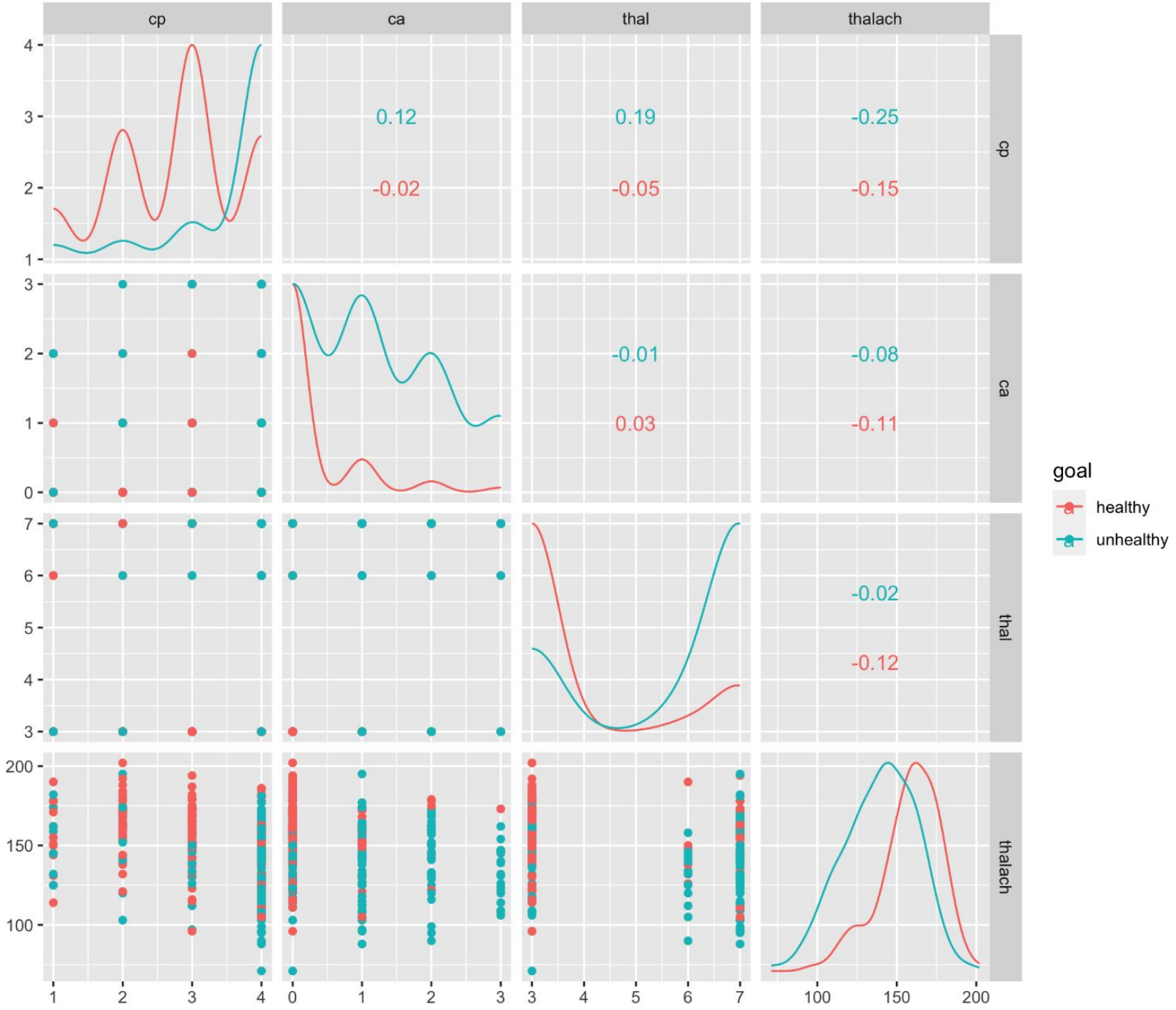
Exploratory Data Analysis

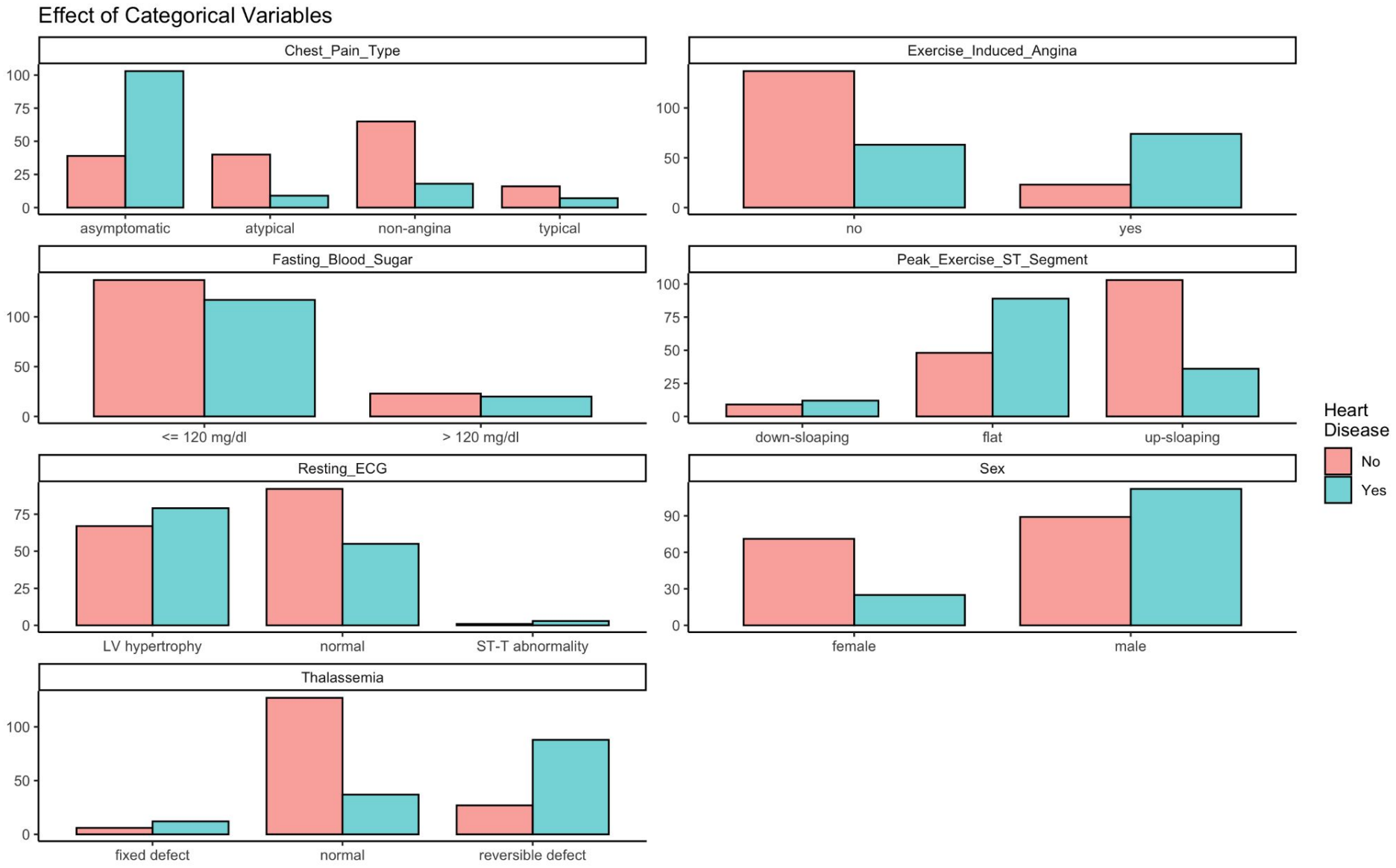
1. Barplot
2. Pairplot
3. Boxplot
4. Heatmap

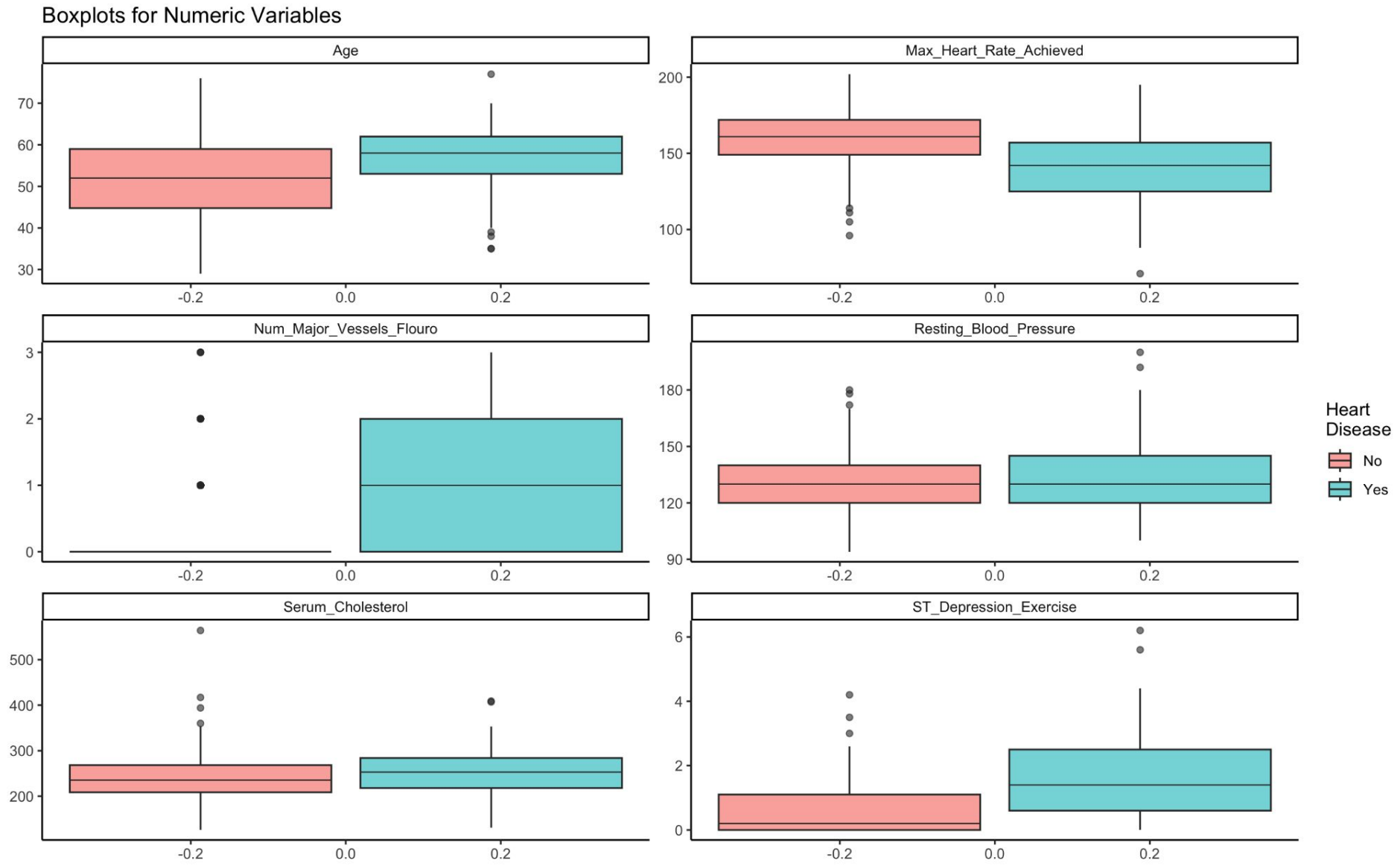
Results



H	UH
160	137
54%	46%

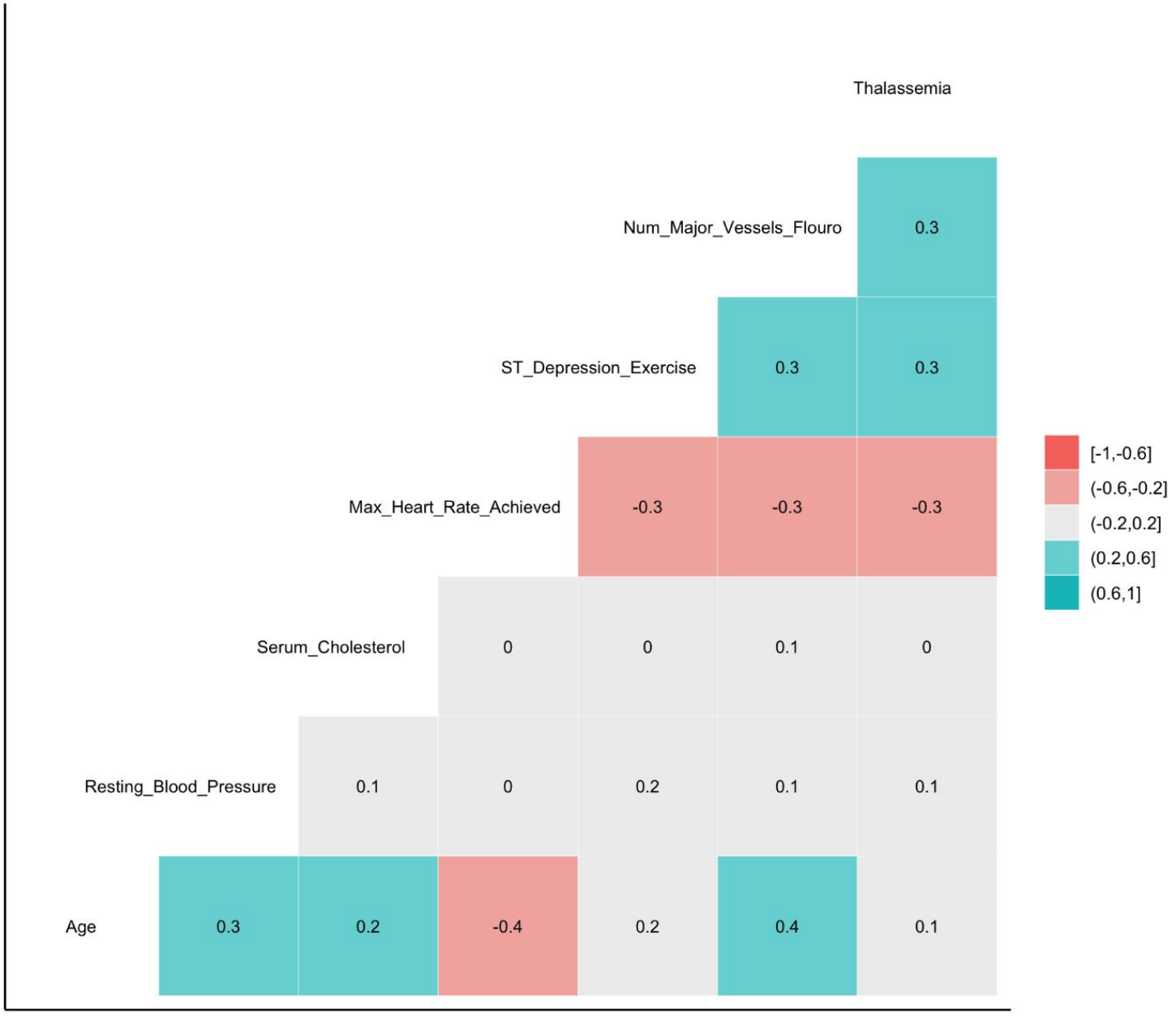

cp
ca
thal
thalach
0.12
0.19
-0.25
-0.02
-0.05
-0.15
-0.01
-0.08
0.03
-0.11
-0.02
-0.12
goal
healthy
unhealthy


Effect of Categorical Variables
Chest_Pain_Type
asymptomatic
atypical
non-angina
typical
Exercise_Induced_Angina
no
yes
Fasting_Blood_Sugar
<= 120 mg/dl
> 120 mg/dl
Peak_Exercise_ST_Segment
down-sloaping
flat
up-sloaping
Resting_ECG
LV hypertrophy
normal
ST-T abnormality
Sex
female
male
Thalassemia
fixed defect
normal
reversible defect
Heart Disease
No
Yes


Boxplots for Numeric Variables
Age
Max_Heart_Rate_Achieved
Num_Major_Vessels_Flouro
Resting_Blood_Pressure
Serum_Cholesterol
ST_Depression_Exercise
Heart Disease
No
Yes

Heat Map

Pearson correlation

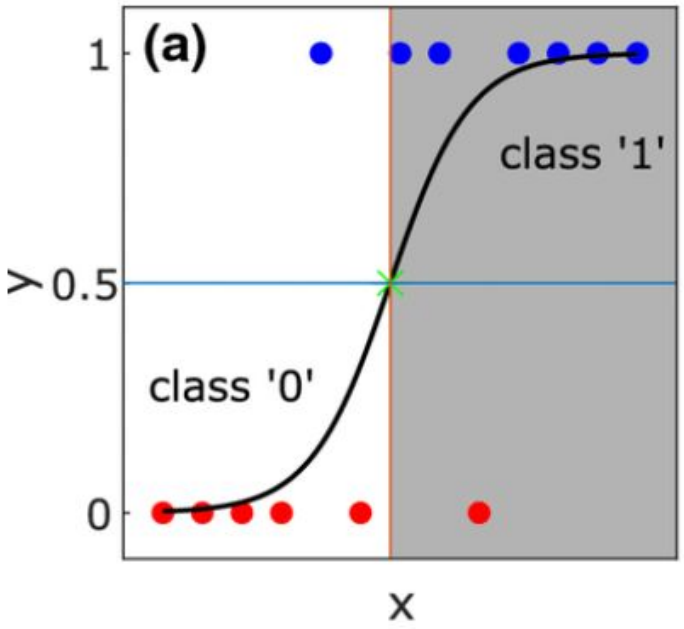


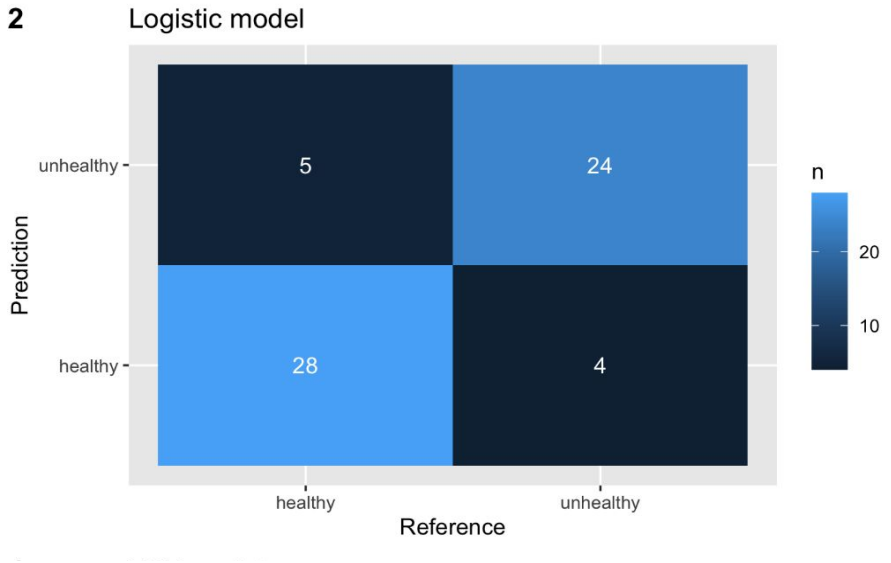
Classifiers

1. Logistic regression
2. Boosted Logistic regression
3. Random forest
4. K-nearest neighbour

Logistic regression

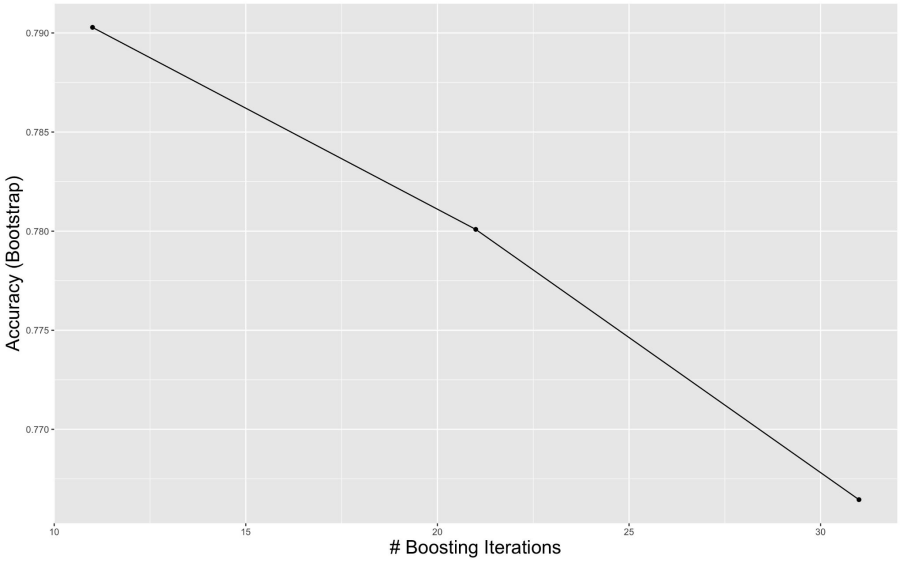
- A logistic regression model is used
- Model improvement by error minimization

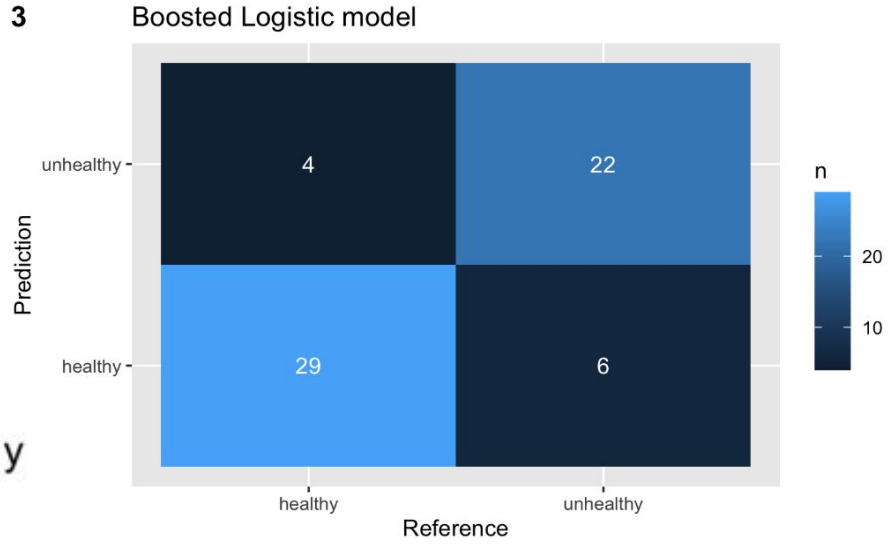


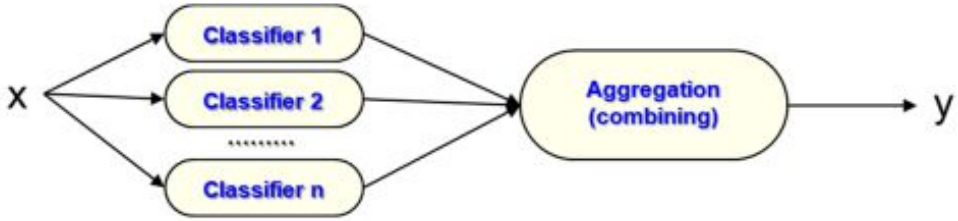


Boosted Logistic regression

- First classification based on each feature separately (decision stumps)
- Combining of the classifications by applying different weights and a logistic regression

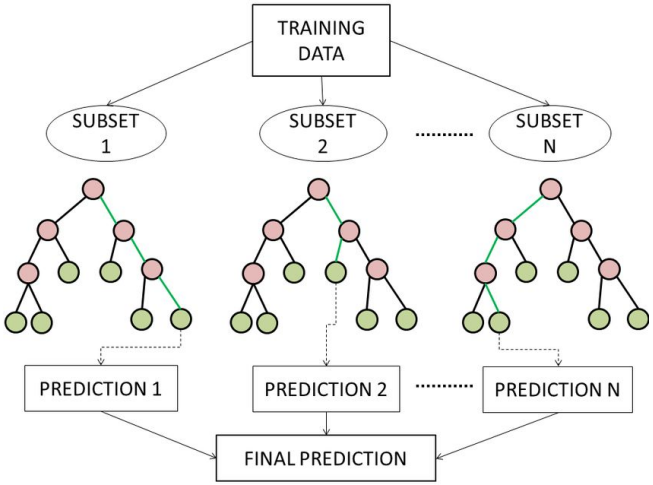


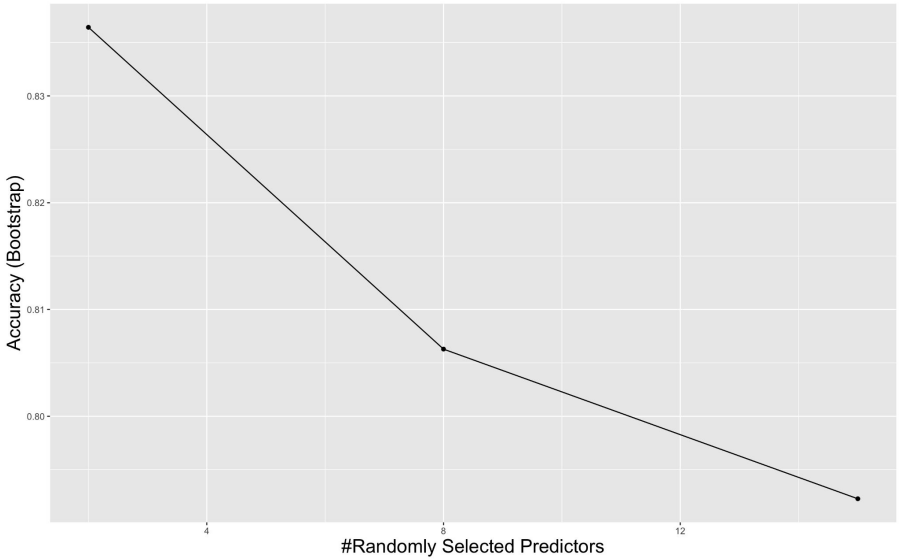


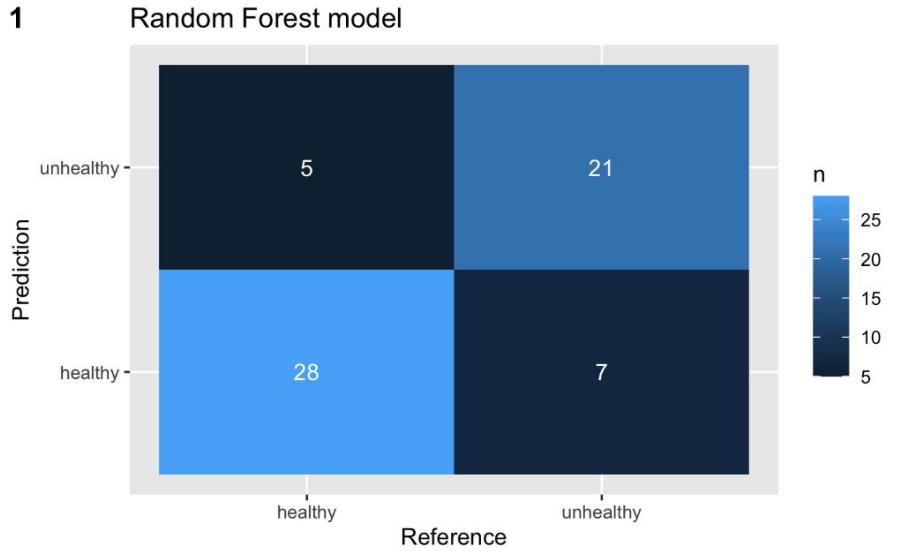


Random Forest

- Multiple Decision Trees are randomly created
- A data point is classified by each of them
- The class with the most ‘votes’ is the final classification

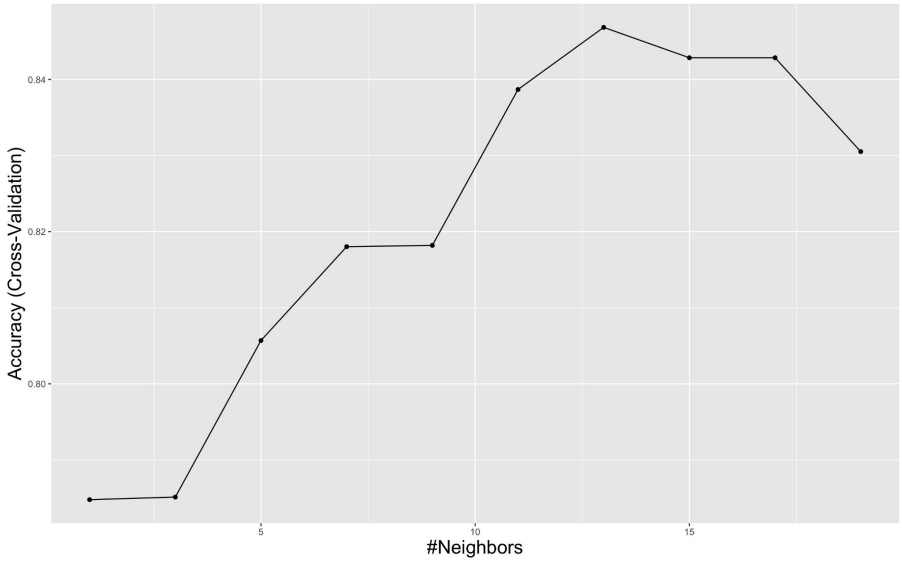


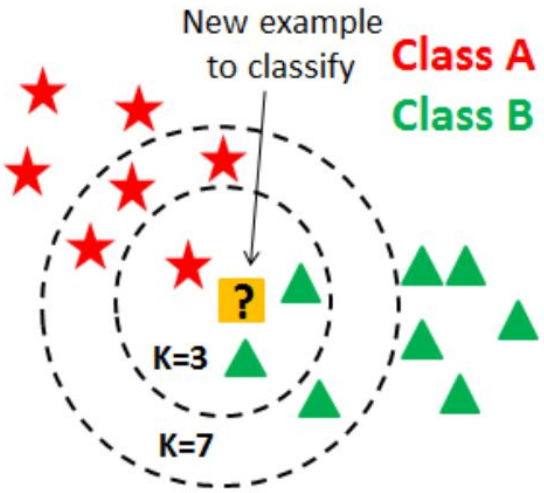


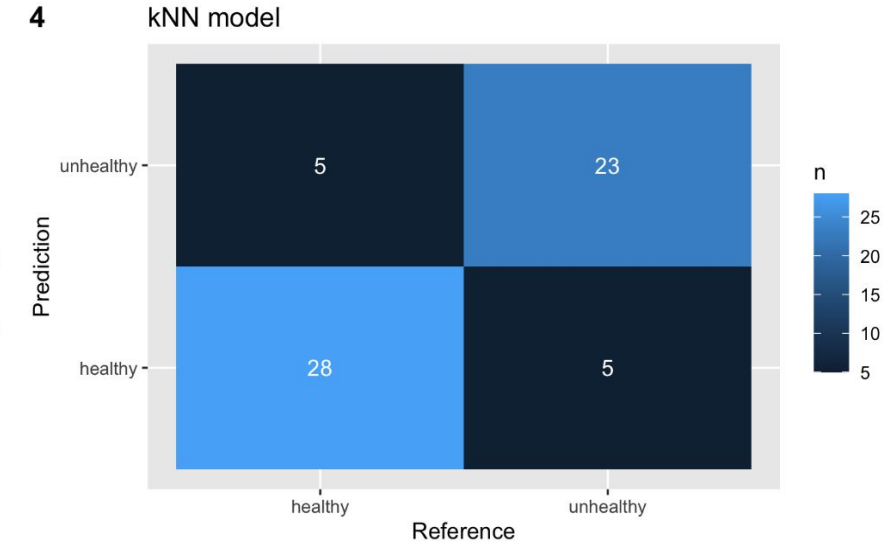


kNN

- A new datapoint is classified by looking at the k nearest neighbors
- Best tuned model → $\mathbf{k = 13}$

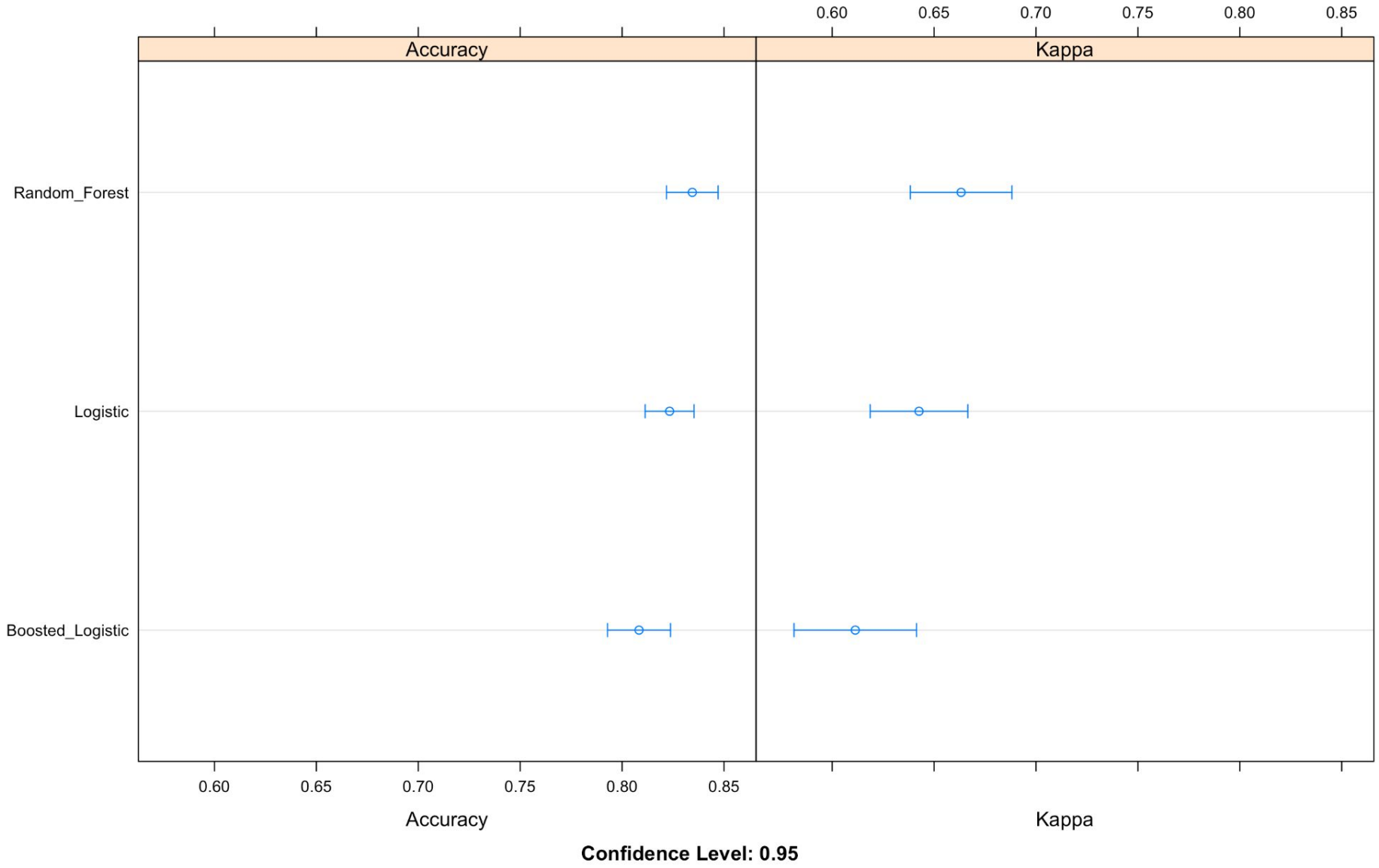




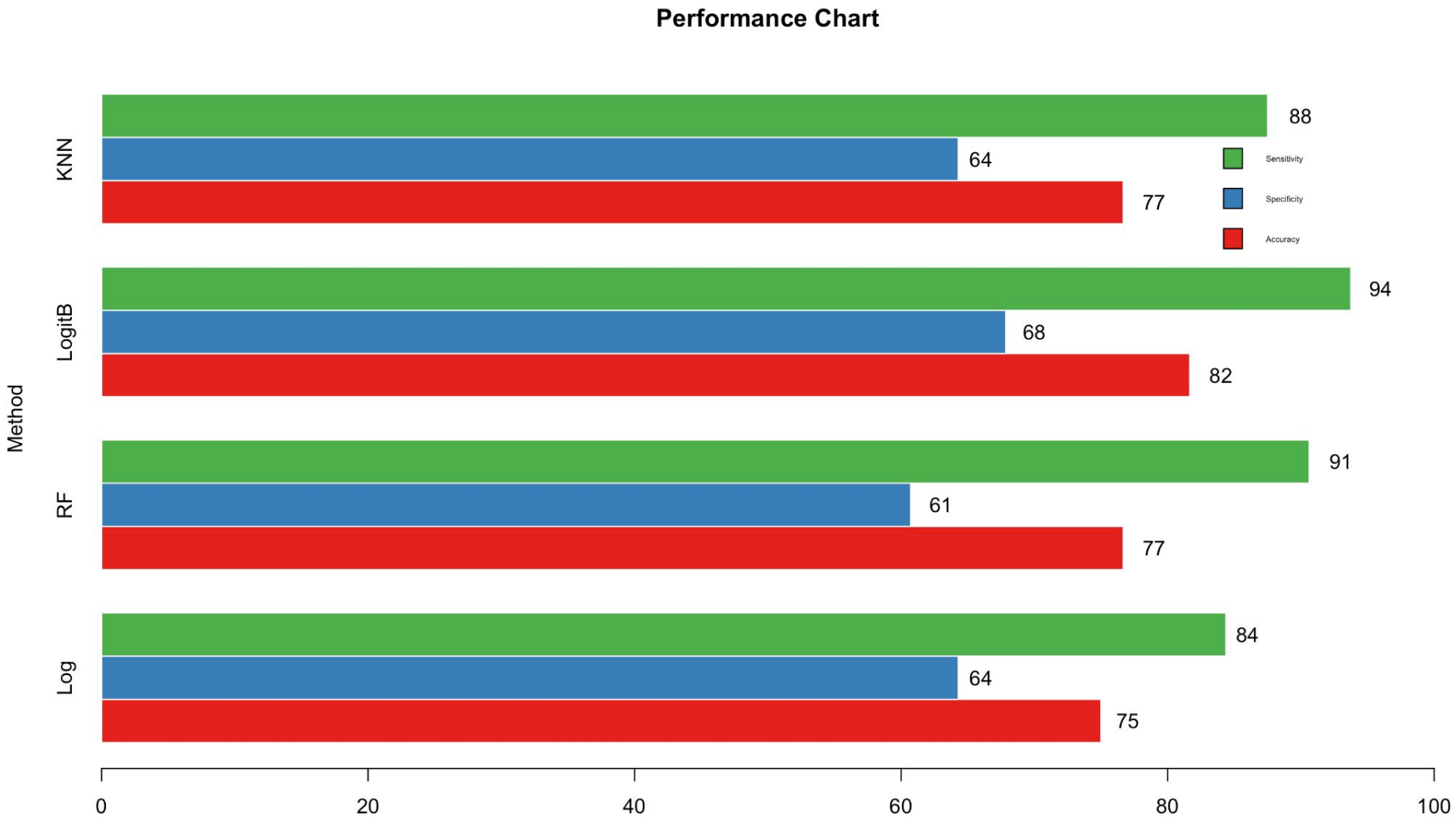


Performance Metrics

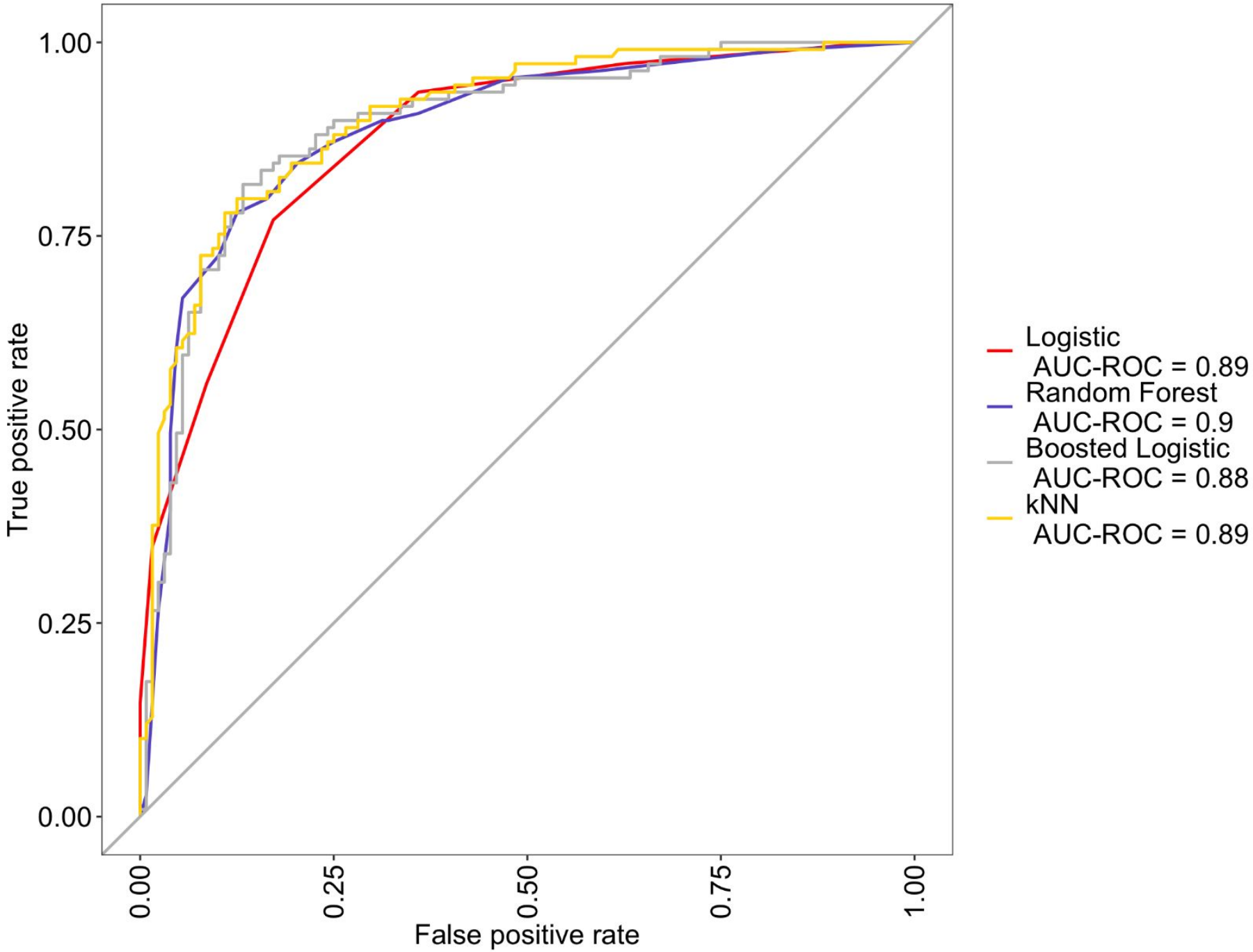
1. Accuracy dotplot
2. Comparison graph
3. PR curve
4. PRG curve
5. ROC curve
6. Calibration curve

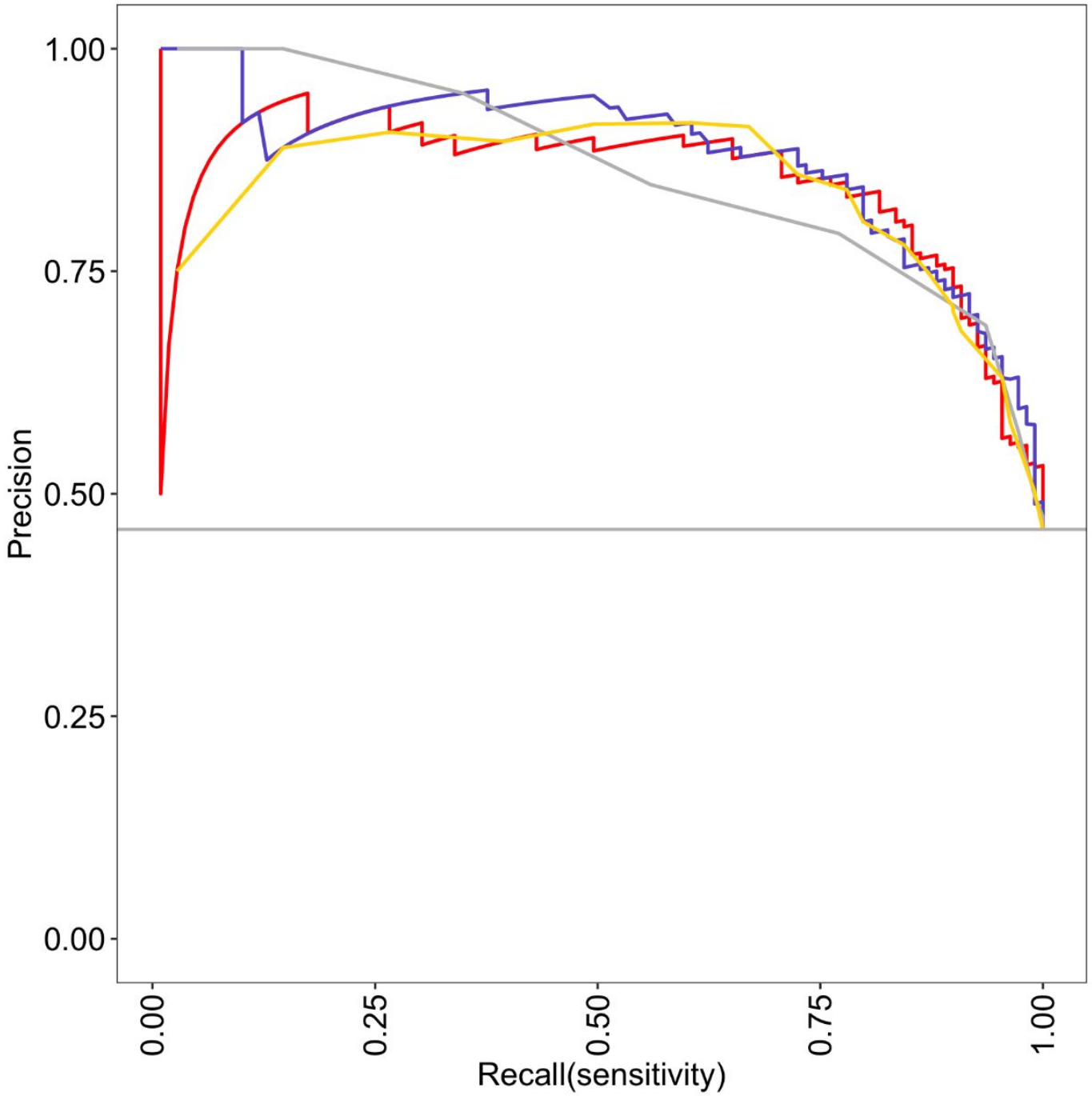


The amount of agreement correct by the agreement expected by chance is Cohen's **Kappa**




```
Logistic Optimal Informedness = 0.68370126146789
Random Forest Optimal Informedness = 0.673165137614679
Boosted Logistic Optimal Informedness = 0.598767201834862
kNN Optimal Informedness = 0.654816513761468
```


1.00
0.75
0.50
0.25
0.00
True positive rate
0.00
0.25
0.50
0.75
1.00
False positive rate
Logistic
AUC-ROC = 0.89
Random Forest
AUC-ROC = 0.9
Boosted Logistic
AUC-ROC = 0.88
kNN
AUC-ROC = 0.89



Precision:

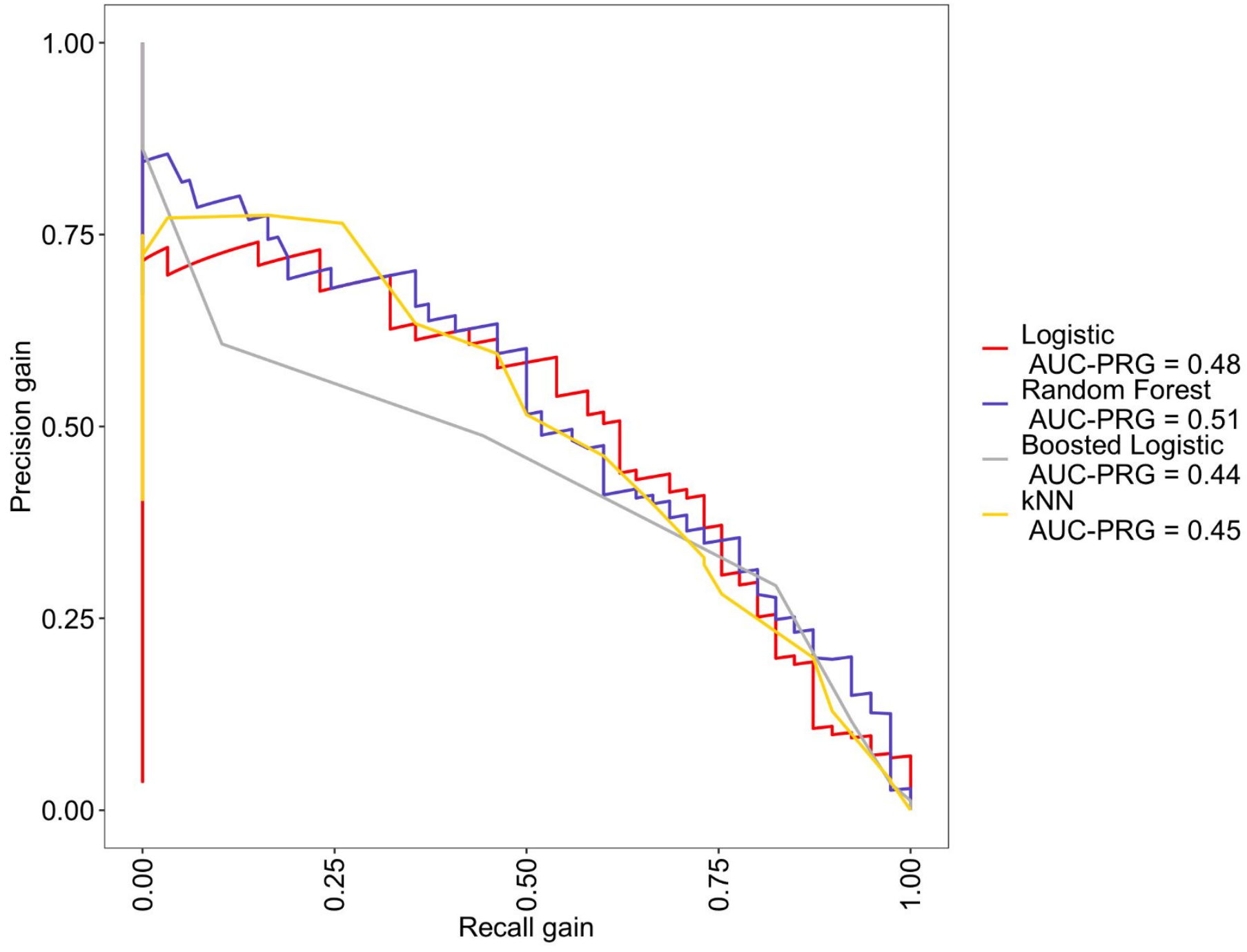
$$TP/(TP+FP)$$

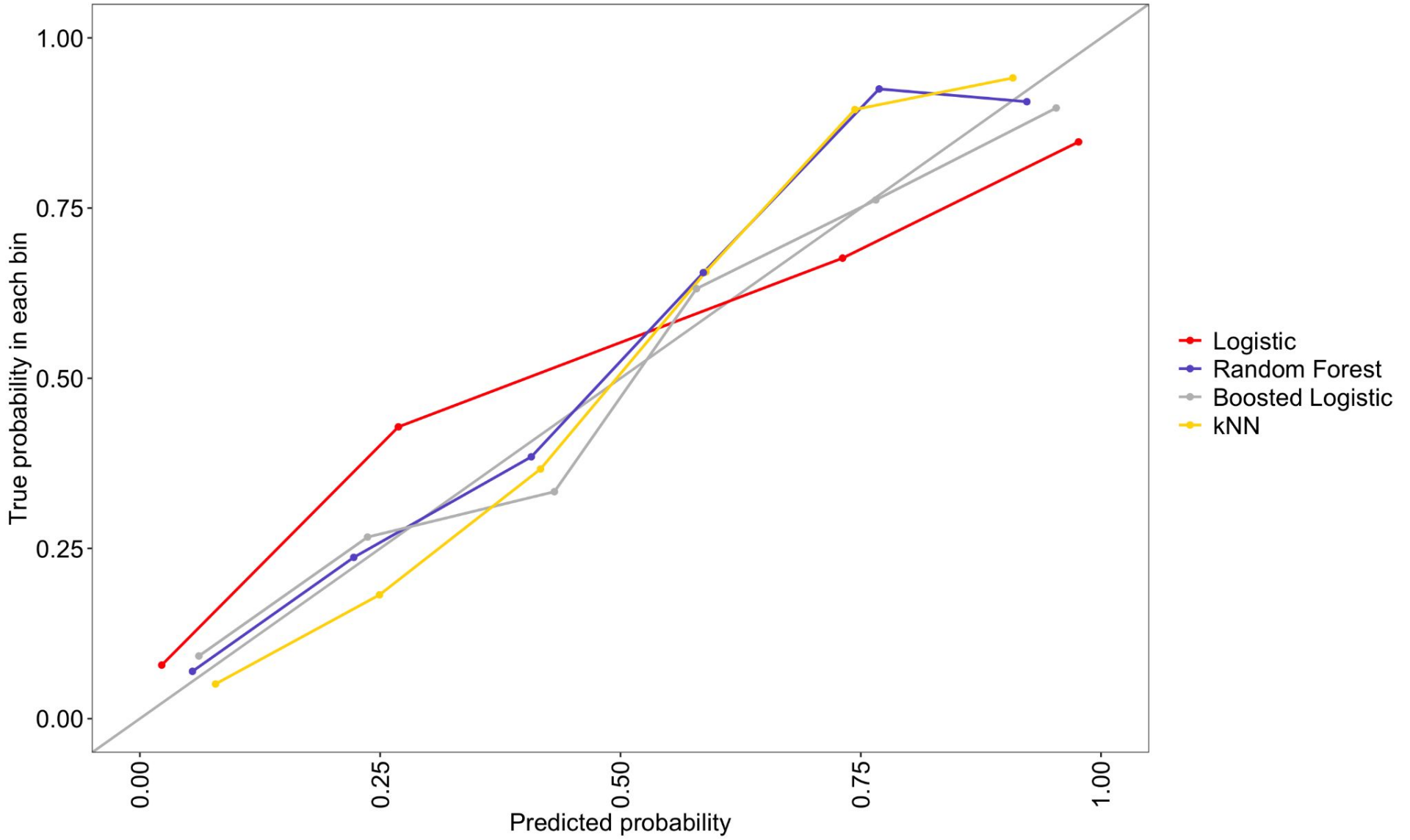
Recall:

$$TP/(TP+FN)$$

$$precG = \frac{prec - \pi}{(1-\pi)prec} = 1 - \frac{\pi}{1-\pi}\frac{FP}{TP}$$

$$recG = \frac{rec - \pi}{(1-\pi)rec} = 1 - \frac{\pi}{1-\pi}\frac{FN}{TP}$$


1.00
0.75
0.50
0.25
0.00
Precision gain
0.00
0.25
0.50
0.75
1.00
Recall gain
Logistic
AUC-PRG = 0.48
Random Forest
AUC-PRG = 0.51
Boosted Logistic
AUC-PRG = 0.44
kNN
AUC-PRG = 0.45


1.00
0.75
0.50
0.25
0.00
True probability in each bin
0.00
0.25
0.50
0.75
1.00
Predicted probability
Logistic
Random Forest
Boosted Logistic
kNN

References

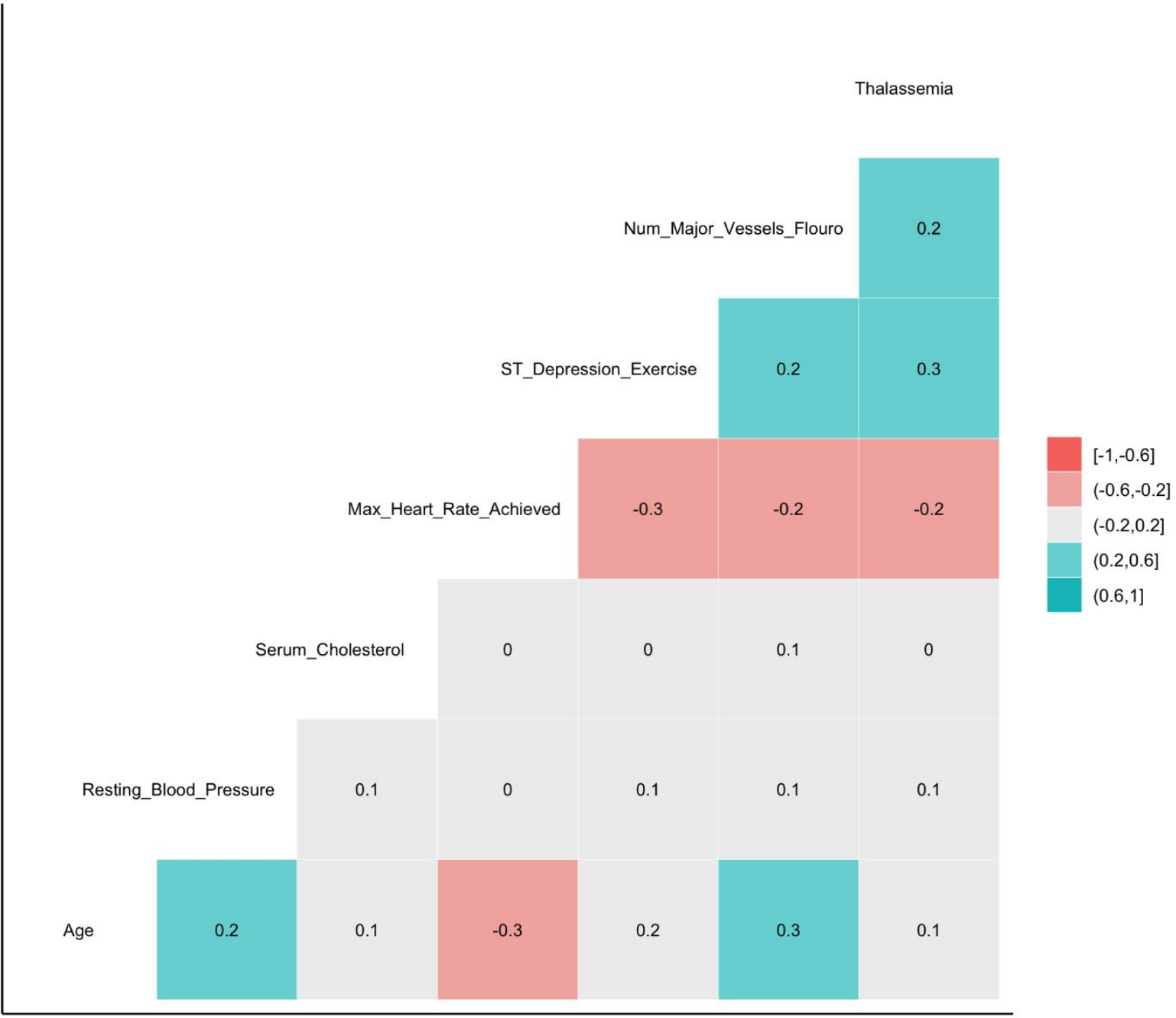
1. Pisula T. “An Ensemble Classifier-Based Scoring Model for Predicting Bankruptcy of Polish Companies in the Podkarpackie Voivodeship.” Journal of Risk and Financial Management. 2020; 13(2):37. https://doi.org/10.3390/jrfm13020037.
2. Detrano, R et al. “International application of a new probability algorithm for the diagnosis of coronary artery disease.” The American journal of cardiology vol. 64,5 (1989): 304-10. https://doi.org/10.1016/0002-9149(89)90524-9.
3. J. H. Gennari, P. Langley, and D. Fisher. 1989. “Models of incremental concept formation.” Artif. Intell. 40, 1–3 (Sep. 1989), 11–61. https://doi.org/10.1016/0004-3702(89)90046-5
4. Kibler, Dennis, David W. Aha, and Marc K. Albert. "Instance-based prediction of real-valued attributes." Computational Intelligence 5.2 (1989): 51-57.
5. R Core Team (2022). “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
6. Illustrations https://www.researchgate.net/figure/Data-classification-by-logistic-regression-aClassification-of-1D-data-showing-the-fitted_fig1_353913155
7. Random Forest sketch https://www.researchgate.net/figure/Example-of-a-Random-Forest-workflow_fig2_342028855

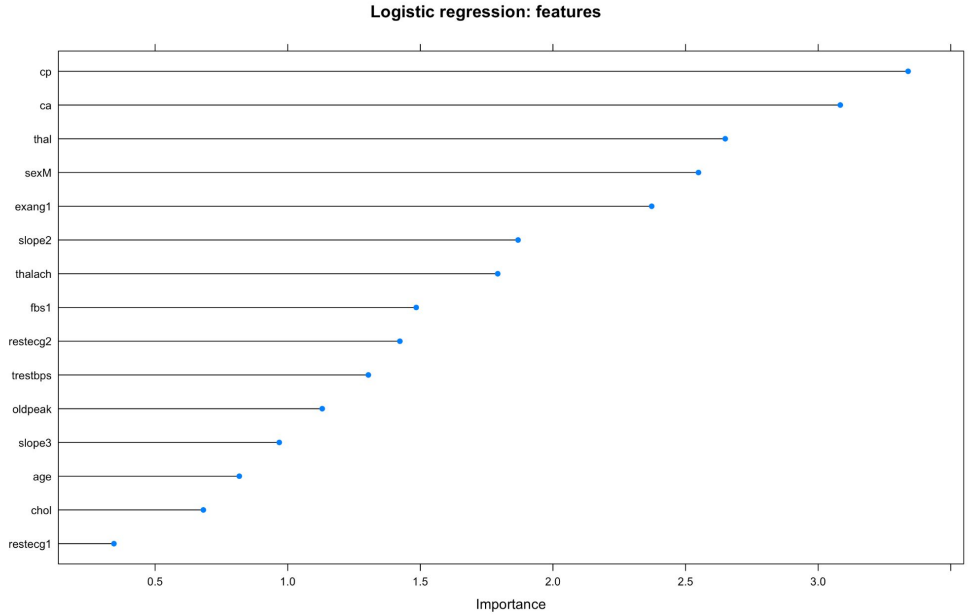
Dataset description can be found here , and data file here.

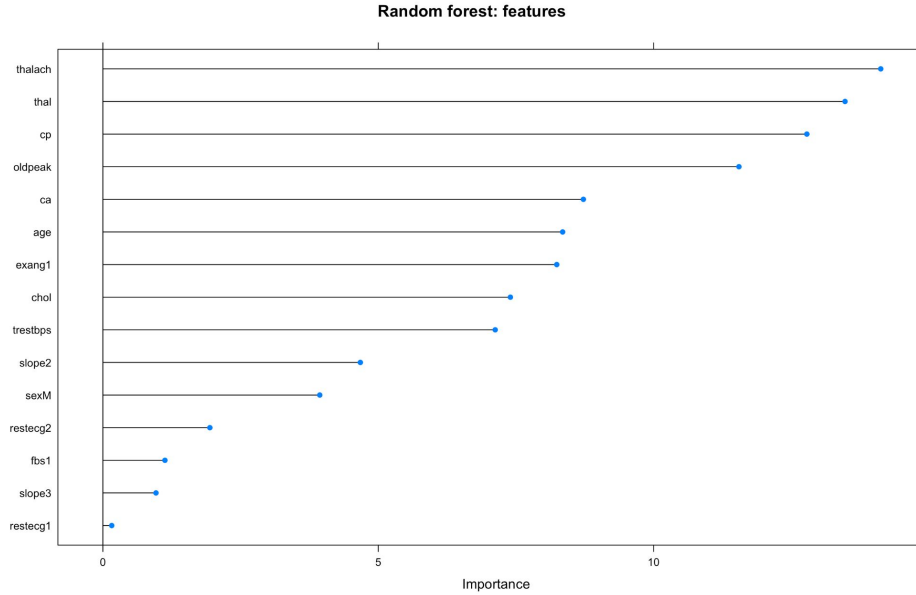
Thanks for listening!
Questions ?

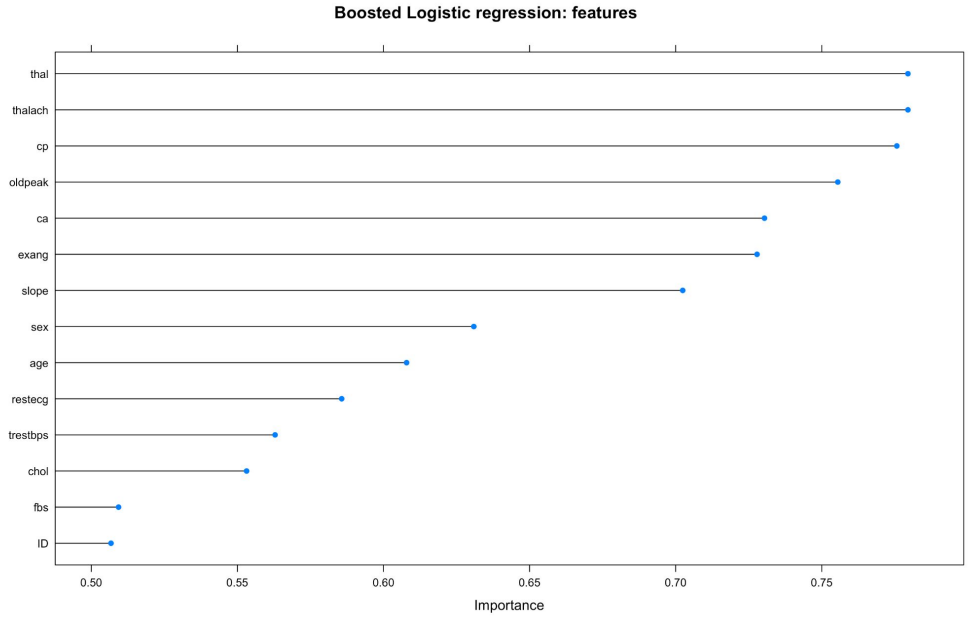
Heat Map

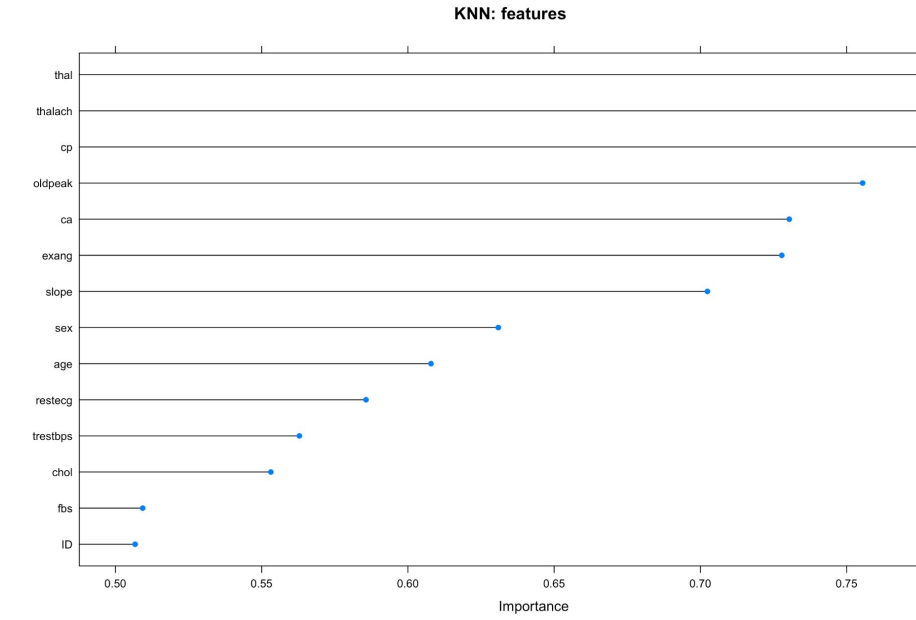
Kendall correlation



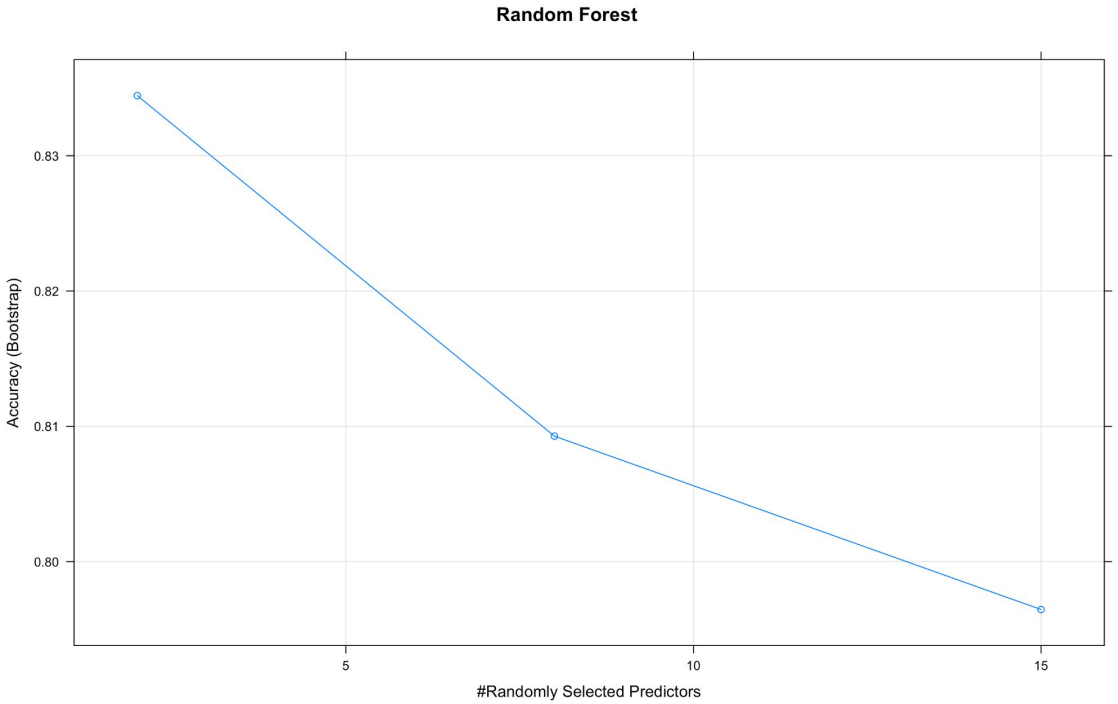








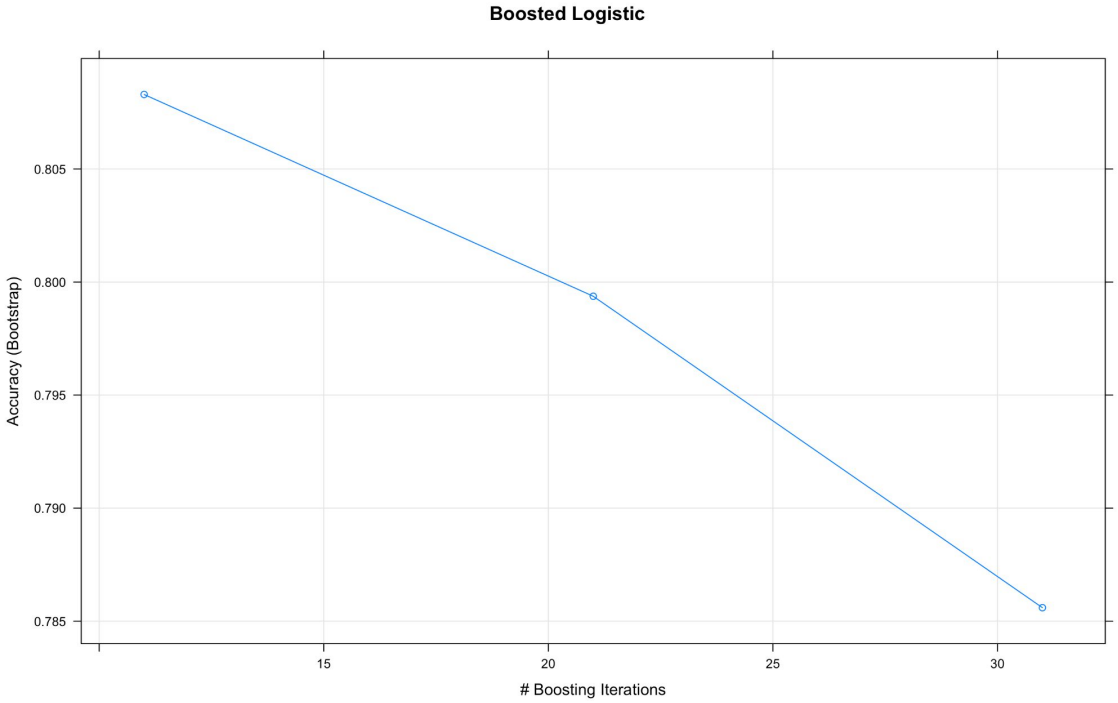
Top features: chest pain type (cp), number of major vessels (ca), maximum heart rate achieved (thalach), thallium scintigraphy (thal)



Accuracy ⇒ Downward trend

Possible reason: **Tuning parameters**

- *mtry* (by default) → bagging
- Fast
- Good variance-bias tradeoff



- *nIter* (by default) → decision stump
- Speedy
- Weak