Heart Disease Prediction

Develop a Prediction Model of Heart Diseases

Hu Xinbei, Liao Meng, Shi Wen, Tan Tian Le, Xu Pengtai

Outline

1. **Introduction**
   Motivation, Objectives and Hypothesis
2. **Data Description**
   Attributes and Exploratory Data Analysis (EDA)
3. **Feature Selection**
   Using Logistic Regression
4. **Model Comparison**
   Model Evaluation and Conclusion
5. **Further Exploration**
   Ensemble Model

Introduction

Motivation, Objectives and Hypothesis

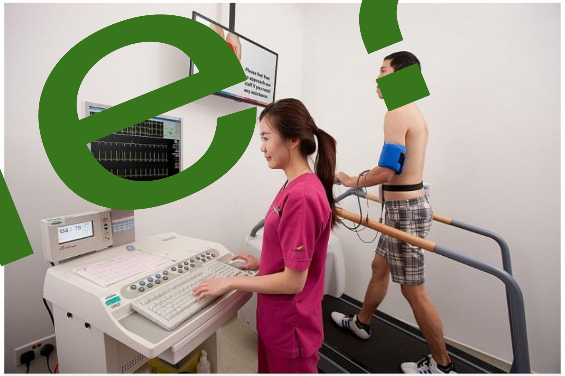
Motivation

True?

Activity trackers could help predict heart problems: Singapore researchers

Researchers put 233 volunteers through a series of clinical tests and used Fitbit activity trackers to monitor the number of steps they took, their heart rates and sleeping patterns over a week. PHOTO: REUTERS

Heart diseases kill more men at earlier age than women



A patient undergoing an exercise stress test at the National Heart Centre Singapore. Not only do more men develop cardiovascular diseases than women, but women here also develop the diseases about 10 years later than men. PHOTO: NATIONAL HEART CENTRE SINGAPORE

Objectives

Understand the **significance** of Risk Factors

Select the best set of predictors

Identify the **best** prediction method

Objectives

Early Detection of Heart Disease

Reduce Cost of Healthcare

Hypothesis

The most important predictors for Heart Disease:

Age

Sex

Resting Blood Pressure

Data Exploration

Attributes and Exploratory Data Analysis (EDA)

Attributes

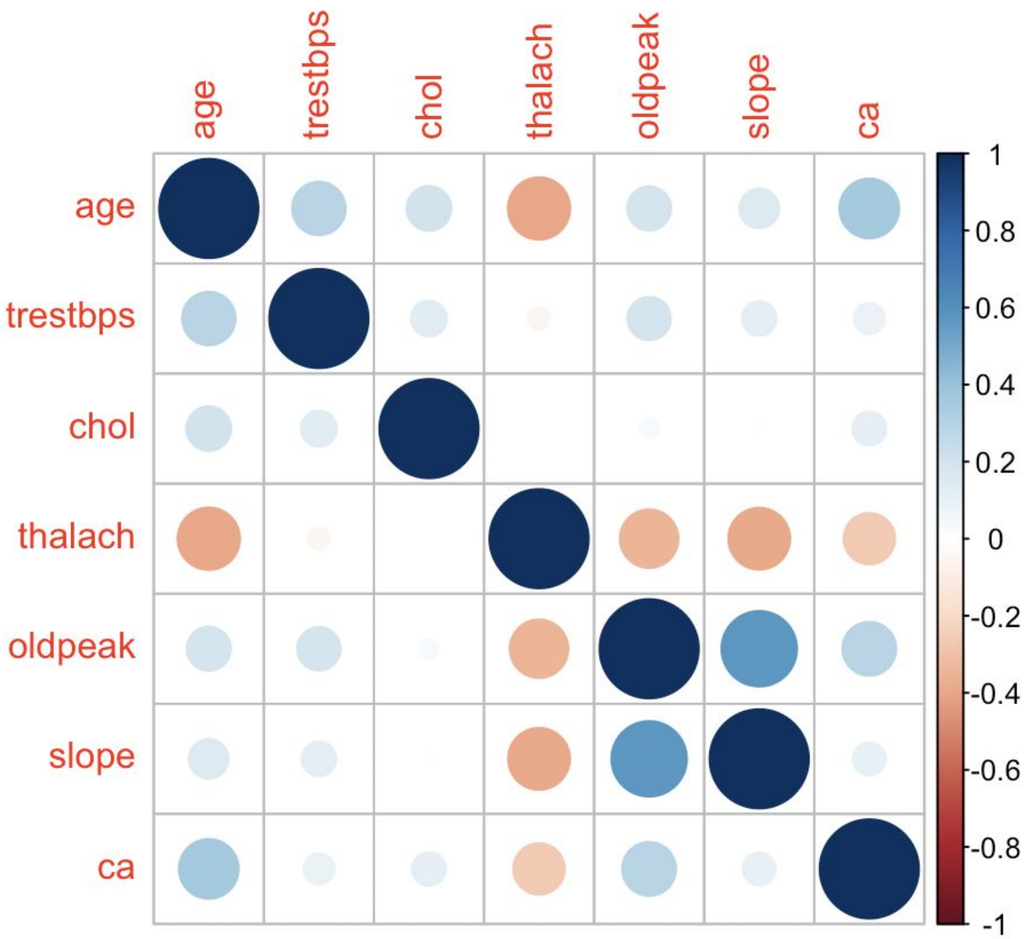
Categorical Variables

- Sex
- cp (Chest Pain Type)
- fbs (Fasting Blood Sugar) > 120 mg/dl
- restecg (Resting ECG Results)
- exang (Exercise-Induced Angina)
- thal (Thalassemia)
- target (Heart Disease or not)

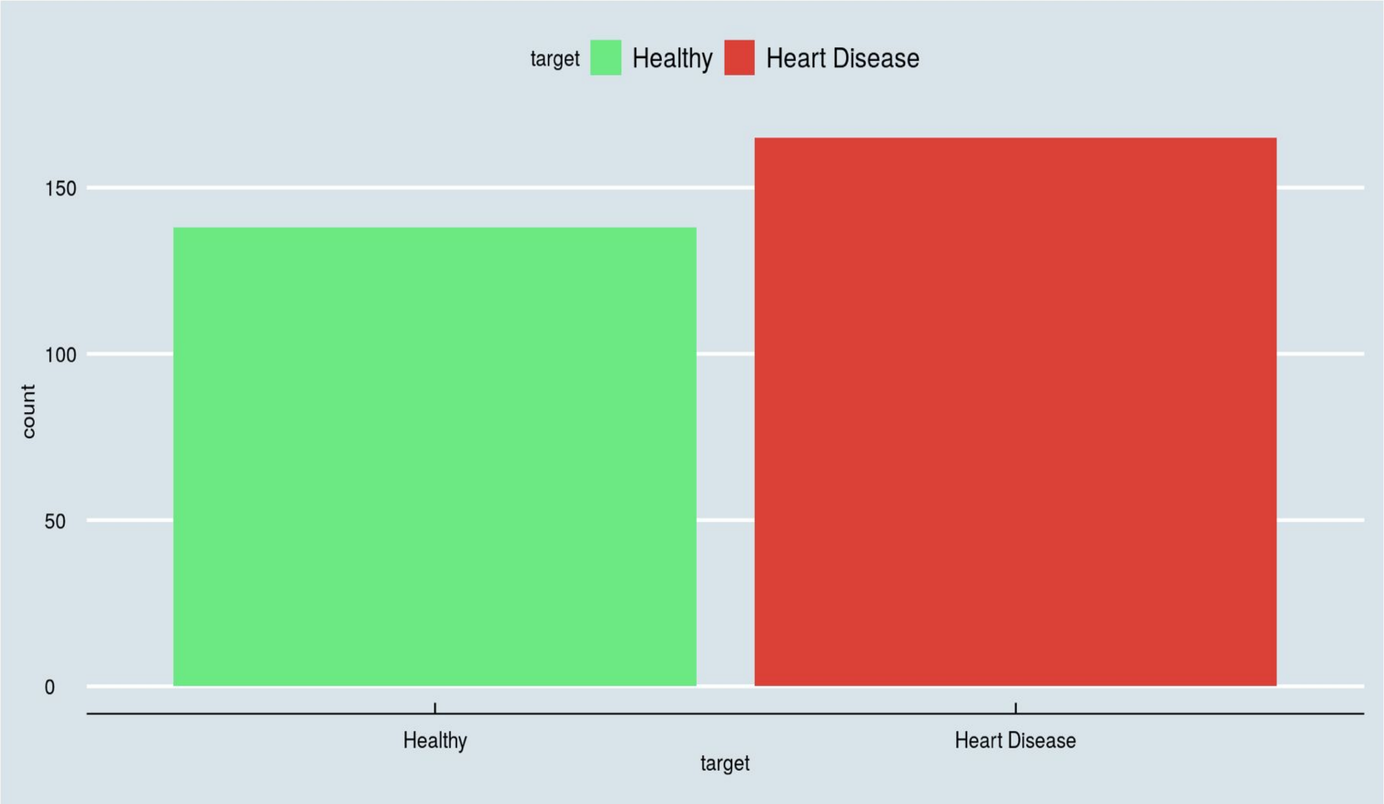
Continuous Variables

- Age
- trestbps (Resting Blood Pressure)
- chol (Serum Cholesterol)
- thalach (Maximum Heart Rate)
- oldpeak (ST Depression induced by Exercise relative to Rest)
- slope (Slope of the peak exercise ST segment)
- ca (Number of Major Vessels)

Descriptive Analytics



Descriptive Analytics



Hypothesis

The most important predictors for Heart Disease:

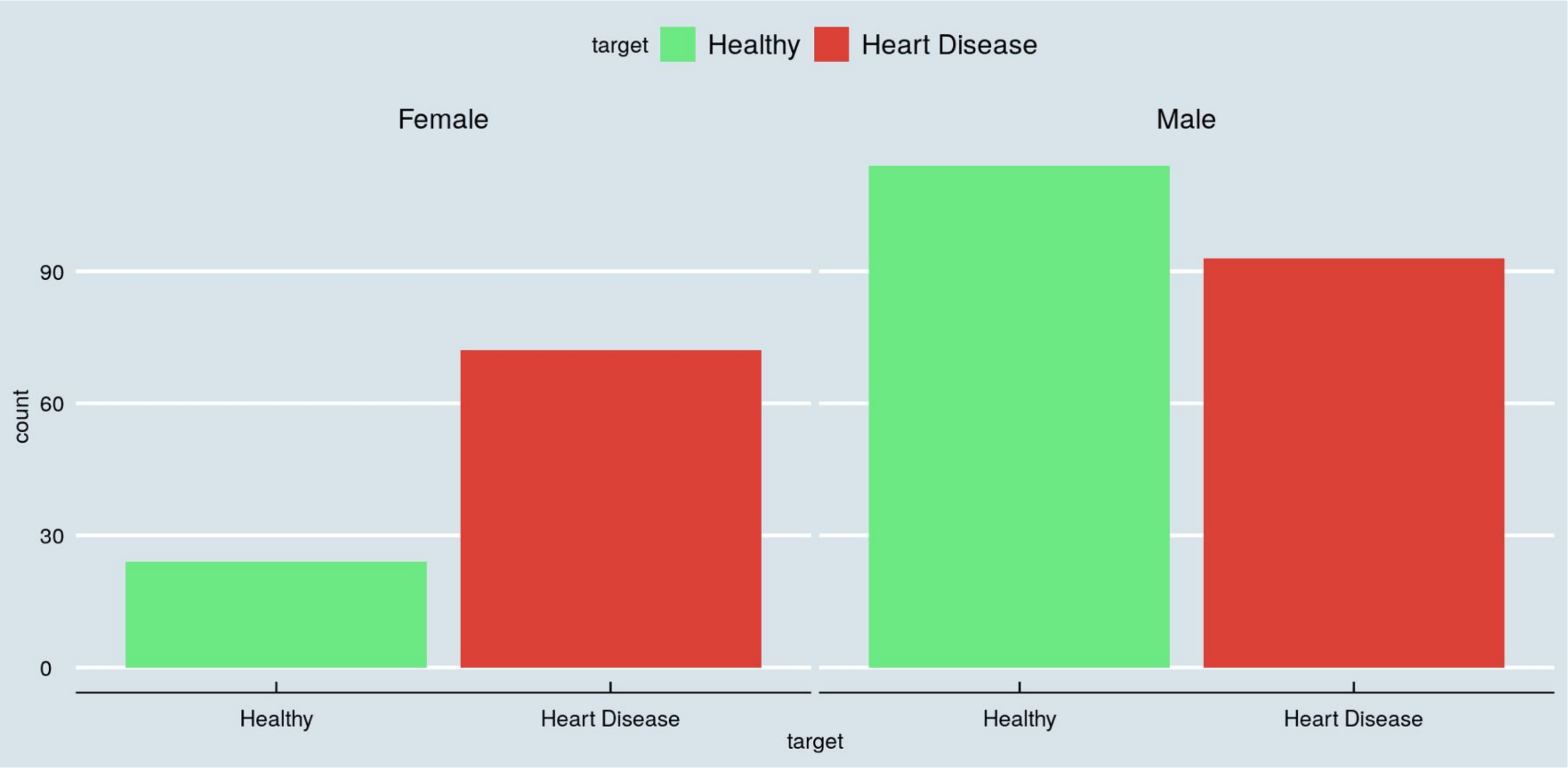
Age

Sex

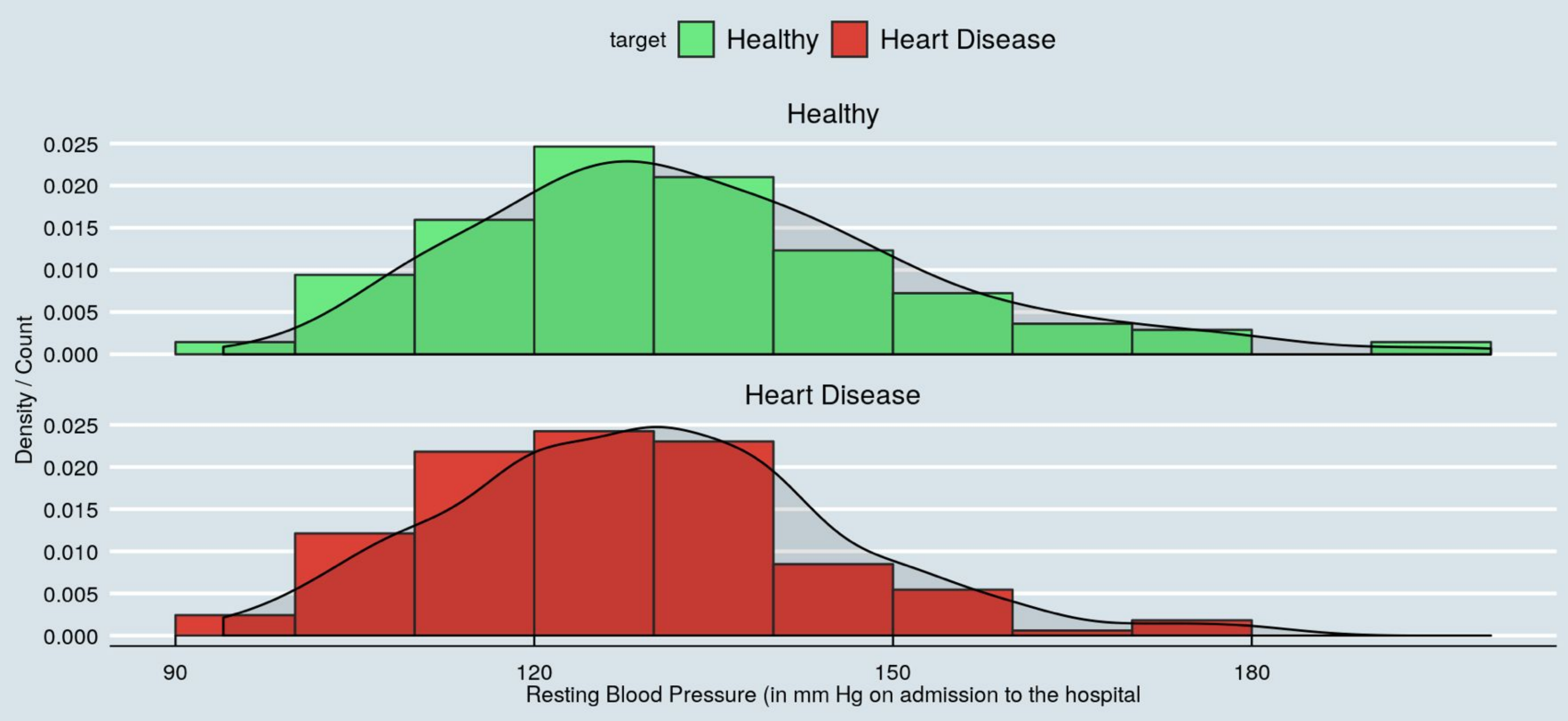
Resting Blood Pressure

Descriptive Analytics

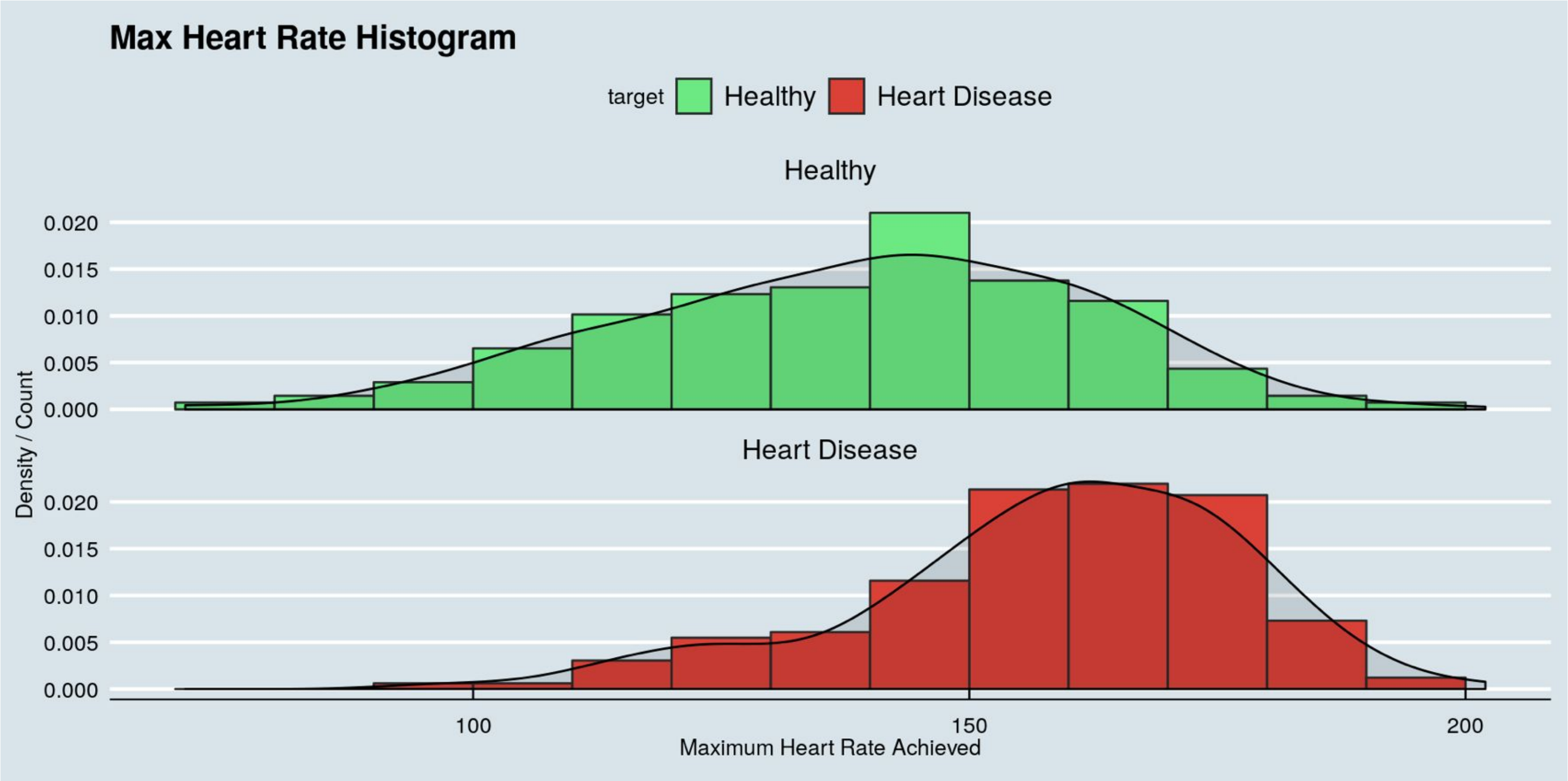
Descriptive Analytics



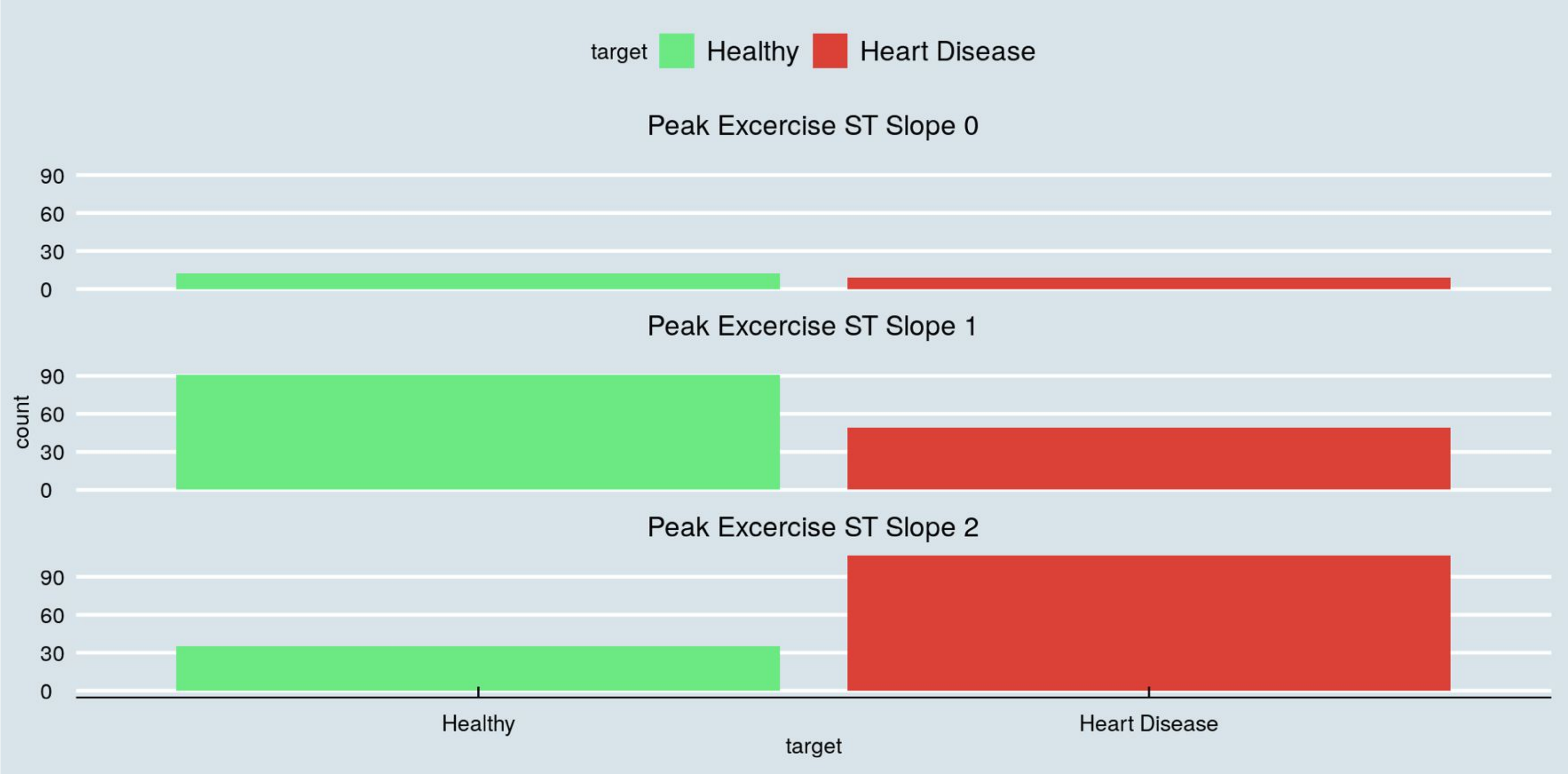
Descriptive Analytics



Descriptive Analytics



Descriptive Analytics



Feature Selection

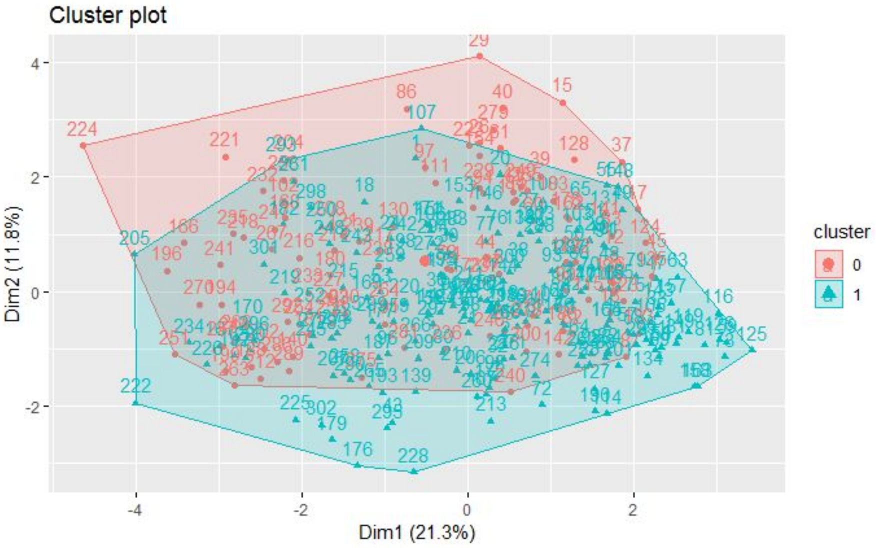
Using Logistic Regression

Results of Logistic Regression

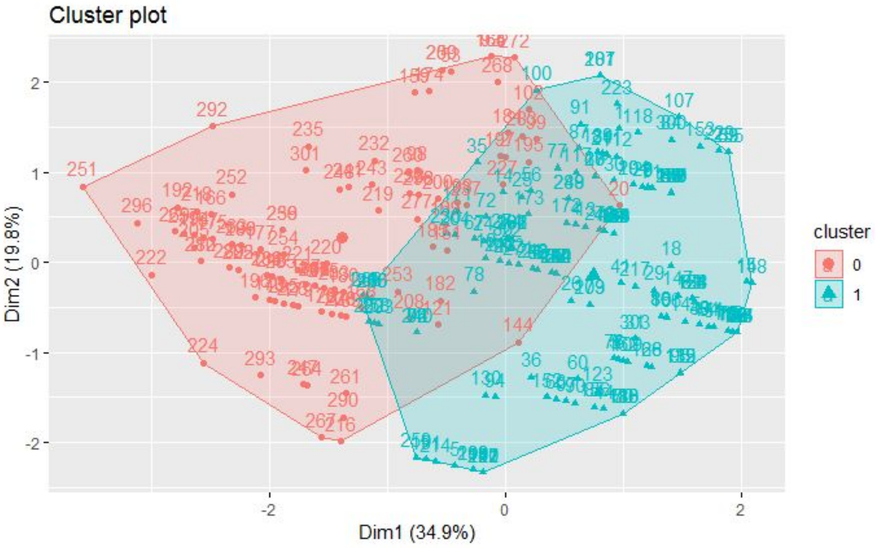
Variables	Coefficient (Std error)
Intercept	1.10 (3.36)
Age	-0.00057 (0.024)
Male (binary)	-1.51 (0.52) **
Chest Pain Type 1 (binary)	0.98 (0.56) *
Chest Pain Type 2 (binary)	1.95 (0.48) ***
Chest Pain Type 3 (binary)	2.02 (0.65) **
Resting Blood Pressure	-0.017 (0.011)
Fasting Blood Sugar > 120 mg/dl (binary)	0.18 (0.57)
Resting Electrocardiographic Results 1 (binary)	0.57 (0.37)
Resting Electrocardiographic Results 2 (binary)	-0.28 (2.27)
Maximum Heart Rate Achieved	0.017 (0.011)
Exercise Induced Angina	-0.76 (0.43) *
ST Depression Induced by Exercise Relative to Rest	-0.49 (0.23) **
Slope of the Peak Exercise ST Segment 1 (binary)	-0.72 (0.86)
Slope of the Peak Exercise ST Segment 2 (binary)	0.20 (0.94)
Number of Major Vessels (0-3) Colored by Flourosopy	-0.83 (0.20) ***
Normal Thalassemia (binary)	1.81 (2.38)
Fixed Defect Thalassemia (binary)	1.85 (2.29)

Significance indicator: *p <0.1, **p<0.05, ***p<0.001

Comparing Cluster Analyses

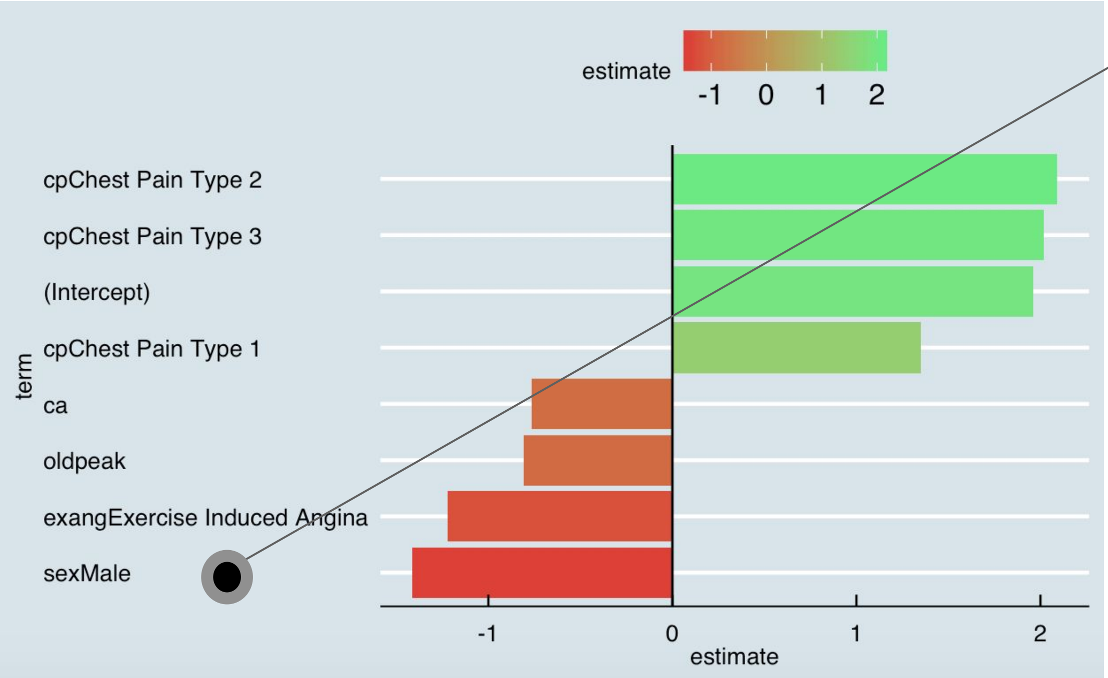


All Variables



Significant Variables

Plotting the Coefficients of the Logit Model



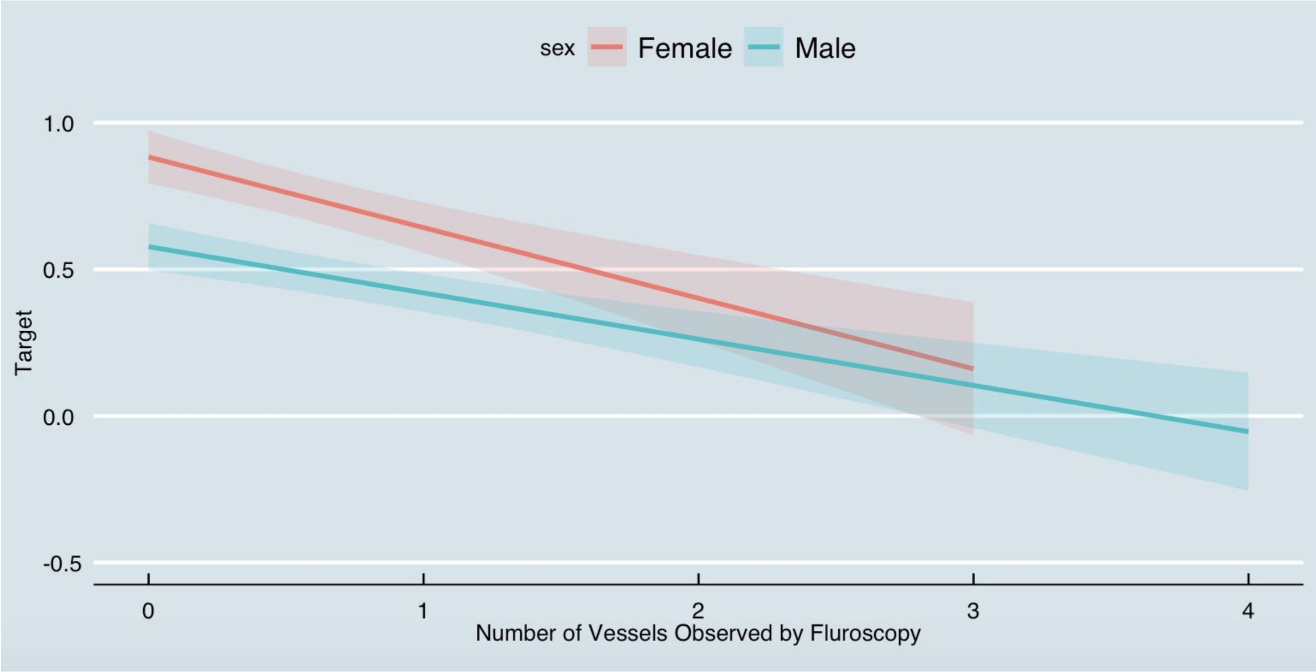
SexMale Coefficient : -1.4117

$$e^{-1.4117} = 0.24$$

$$1 - 0.24 = 0.68$$

68%

Understanding the Coefficients for Continuous Variables



As number of vessels observed increases, the risk of heart disease decreases.

Model Comparison

Model Evaluation and Conclusion

3 Classification Models Used

Naive Bayes

SVM

Random Forest

Why are these 3 machine learning models selected?
→ do not require linearity of independent variables

SVM Models

	Linearity		
Grid Search		Yes	No
	Yes	Linear with Grid Search	Radial Kernel with Grid Search
	No	Standard Linear	Radial Kernel

ROC of 4 SVM models

Model 1	SVM Linear with Grid Search	0.8723
Model 2	SVM Radial with Grid Search	**0.8739**
Model 3	SVM Standard Linear	0.8680
Model 4	SVM Radial	0.8722

Among the 4 SVM models, radical with grid search has the highest ROC (receiver operating characteristic) score. Hence, we chose this SVM model for comparison with other models.

Data Splitting

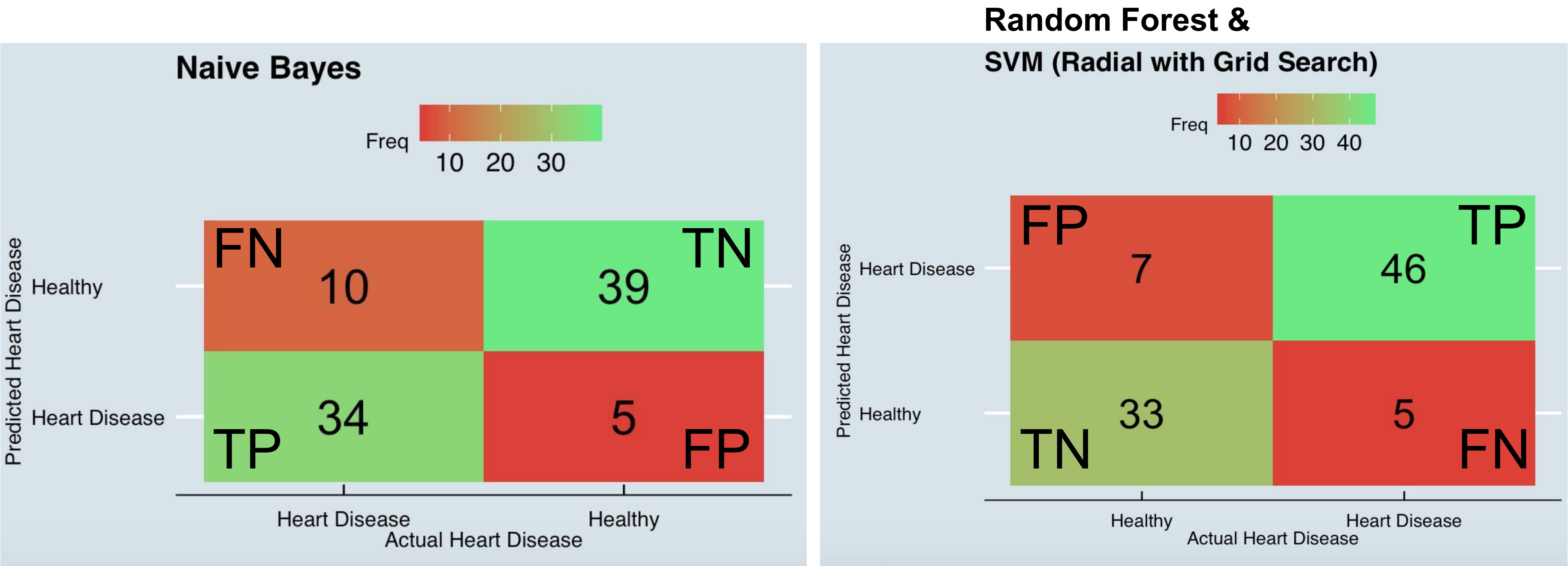
30%

Test Data

70%

Training Data

Comparing the Prediction Performance on Test Data



(two models have the same prediction result)

Prediction results are presented in confusion matrices

Conclusion

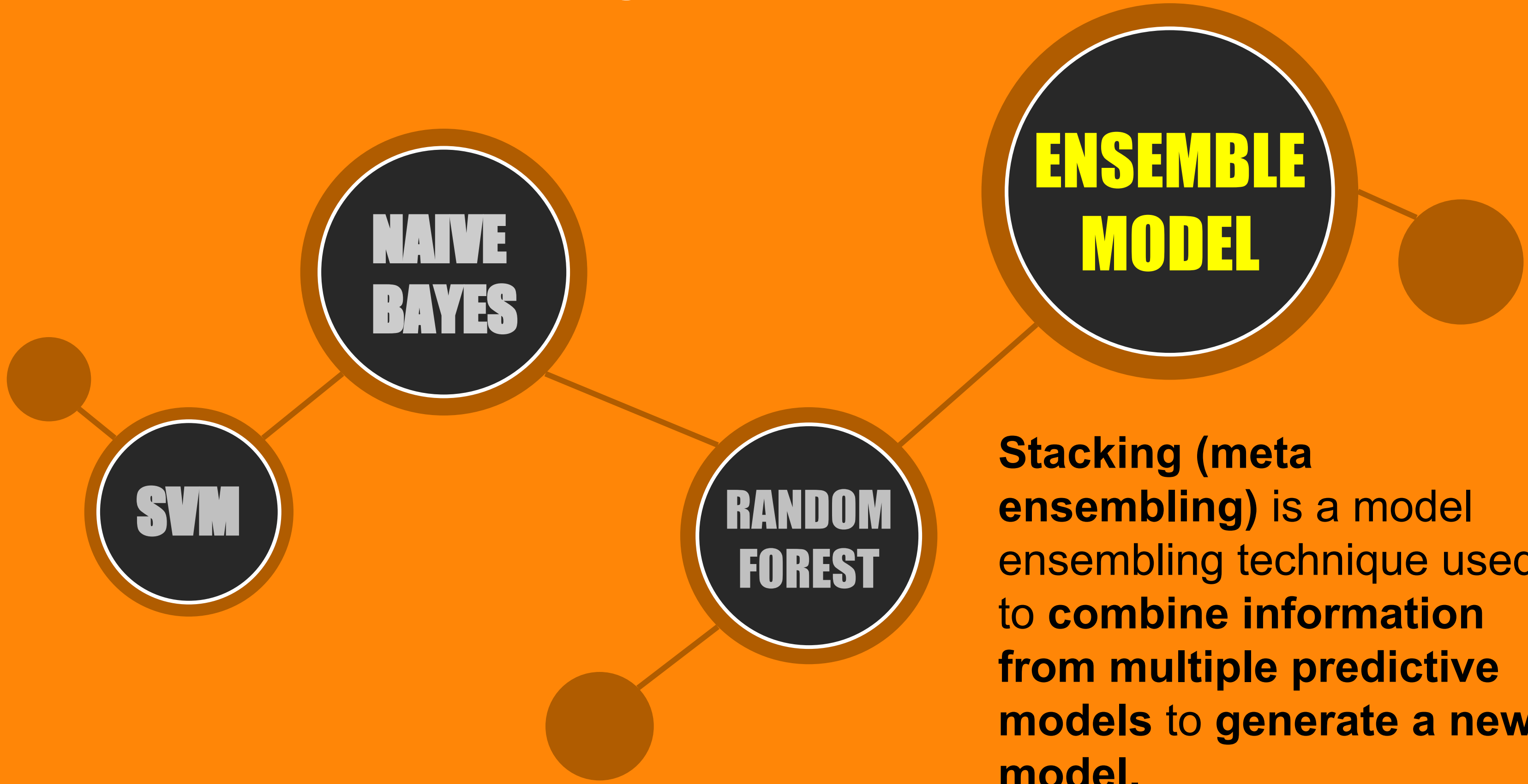
Models\Metrics	Accuracy	Sensitivity	Specificity	Precision
Naive Bayes	82.95%	77.27%	**88.64%**	**87.18%**
SVM Radial with Grid Search	**86.81%**	**90.20%**	82.50%	86.79%
Random Forest	**86.81%**	**90.20%**	82.50%	86.79%

SVM Radial with Grid Search and **Random Forest** are the more accurate classification models for this data set. **Sensitivity** is the most important metric in this case since false negative (predict the person to be healthy, when actually the person has heart disease) is more risky and undesirable than false positive.

Further Exploration

Ensemble Model

Ensemble Method: Stacking



Stacking (meta ensembling) is a model ensembling technique used to **combine information from multiple predictive models** to **generate a new model.**

Ensemble Model Performance

Low Sensitivity:

	SVM Radial	Naive Bayes	Random Forest	Ensemble
Sensitivity	90.20%	78.38%	90.20%	78.11%

Low AUC:

	SVM Radial	Naive Bayes	Random Forest	Ensemble
AUC (healthy vs disease)	0.8889463	0.8765496	0.9039256	0.8481405

Why does ensemble model perform worse than the base models?

Analysis

Obtaining correlation between models: `modelCor(resamples(models))`

	svmRadial	Random Forest	Naive Bayes
SVM Radial	1.0000000	0.9263731	0.9164968
Random Forest	0.9263731	1.0000000	0.9255797
Naive Bayes	0.9164968	0.9255797	1.0000000

The base models seem to be highly correlated
⟶ poor performance of the ensemble model

THANK YOU

THANK YOU