Heart Disease Factors Analysis & Prediction

Group 4: Jiaye Li, Maohe Wang, Yuchen Jiang, Ziyi Pan, Shanghao Wu

Table of contents

01&02

Data Description & Wrangling

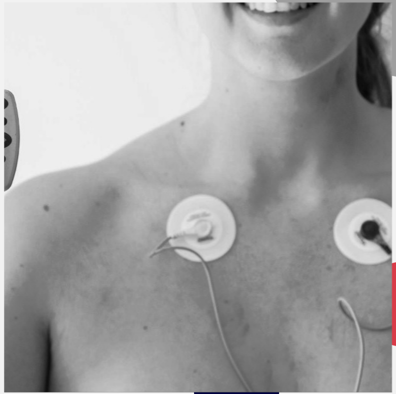
Source & Variables Description

Introduction

Cardiovascular diseases (CVDs) are the **number 1 cause of death** globally, taking an estimated 17.9 million lives each year, which accounts for **31% of all deaths** worldwide.

Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) **need early detection and management wherein a machine learning model can be of great help.**



Data description

- Use heart disease datasets from UCI Machine Learning Repository.
- Combine 5 independent heart disease datasets, including Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog Data Set.
- 1190 observations and 11 variables, 6 categorical and 5 continuous variables
- The response variable is whether the heart disease event happens.

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [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]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

```
> summary(cardio.data)
      Age             Sex          ChestPainType     RestingBP      Cholesterol     FastingBS       RestingECG
 Min.   :28.0   Min.   :0.000   Min.   :1.00   Min.   : 92   Min.   :110   Min.   :0.000   Min.   :0.000
 1st Qu.:46.0   1st Qu.:0.000   1st Qu.:1.00   1st Qu.:120   1st Qu.:206   1st Qu.:0.000   1st Qu.:0.000
 Median :54.0   Median :0.000   Median :2.00   Median :130   Median :234   Median :0.000   Median :0.000
 Mean   :52.7   Mean   :0.238   Mean   :1.86   Mean   :131   Mean   :239   Mean   :0.162   Mean   :0.631
 3rd Qu.:59.0   3rd Qu.:0.000   3rd Qu.:3.00   3rd Qu.:140   3rd Qu.:271   3rd Qu.:0.000   3rd Qu.:1.000
 Max.   :77.0   Max.   :1.000   Max.   :4.00   Max.   :170   Max.   :369   Max.   :1.000   Max.   :2.000
     MaxHR      ExerciseAngina     Oldpeak         ST_Slope     HeartDisease
 Min.   : 71   Min.   :0.000   Min.   :-0.10   Min.   :1.00   Min.   :0.000
 1st Qu.:122   1st Qu.:0.000   1st Qu.: 0.00   1st Qu.:2.00   1st Qu.:0.000
 Median :141   Median :0.000   Median : 0.40   Median :2.00   Median :0.000
 Mean   :141   Mean   :0.373   Mean   : 0.83   Mean   :2.44   Mean   :0.462
 3rd Qu.:160   3rd Qu.:1.000   3rd Qu.: 1.50   3rd Qu.:3.00   3rd Qu.:1.000
 Max.   :202   Max.   :1.000   Max.   : 3.60   Max.   :3.00   Max.   :1.000
```

Data wrangling

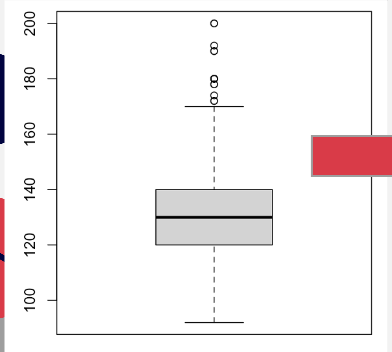
01 Sorting out the data

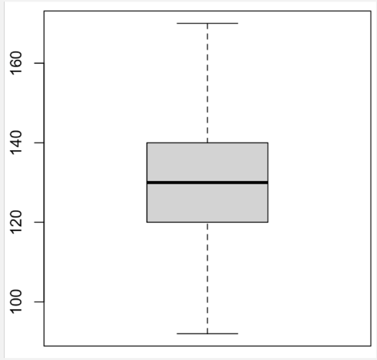
Build up categorical variables:
transfer characters levels into numbers based on their actual seriousness, such as 4,3,2,1 respectively.

02 Remove duplicate value & missing value

03 Remove outliers

Draw box plot of the continuous variables with R to find out the outliers





04 Feature selection

Test whether the variables are significantly related to heart failure

Categorical: Chi-square

Categorical variables	P-value
Gender	3e-15
Fasting BS	< 2e05
Resting ECG	0.001
Exercise Angina	< 2e05
ST_Slope	< 2e-16
ChestPainType	< 2e-16

Continuous: Anova

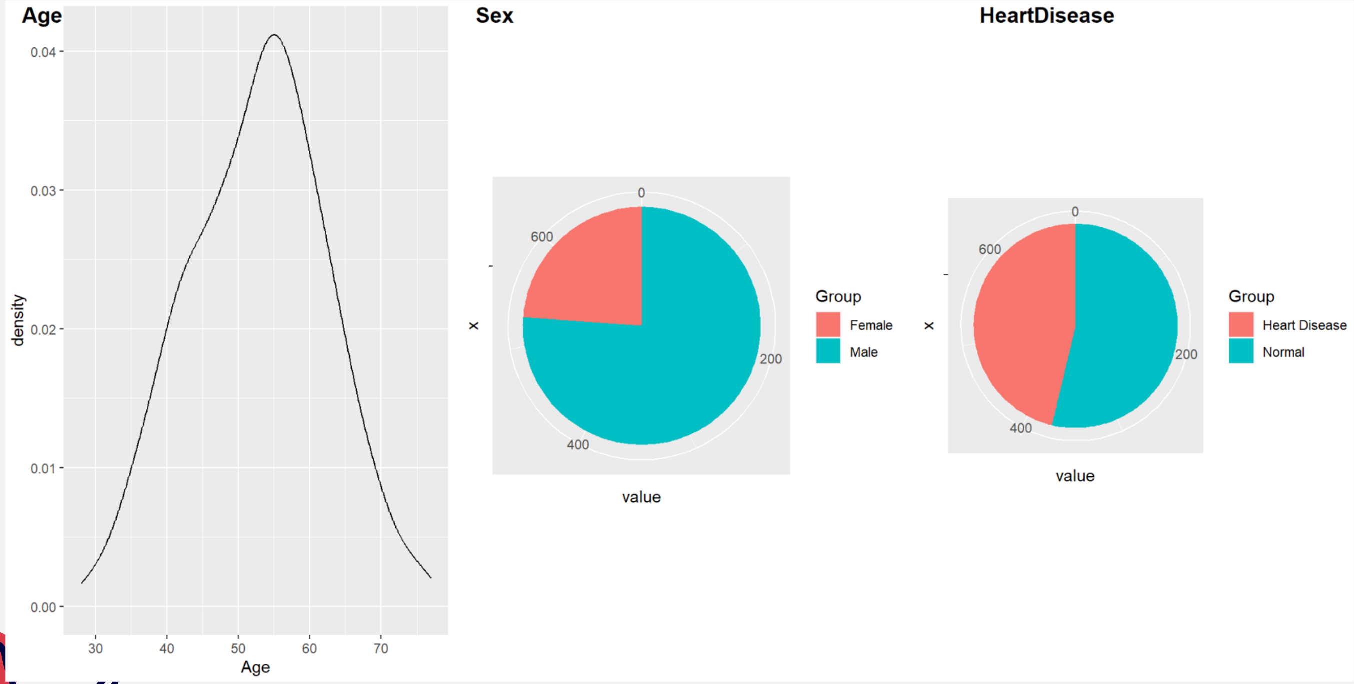
Continuous variables	P-value
Age	<2e-16
Resting BP	1.9e-06
Cholesterol	0.0045
MaxHR	<2e-16
Oldpeak	< 2e-16

03

Exploratory Data Analysis

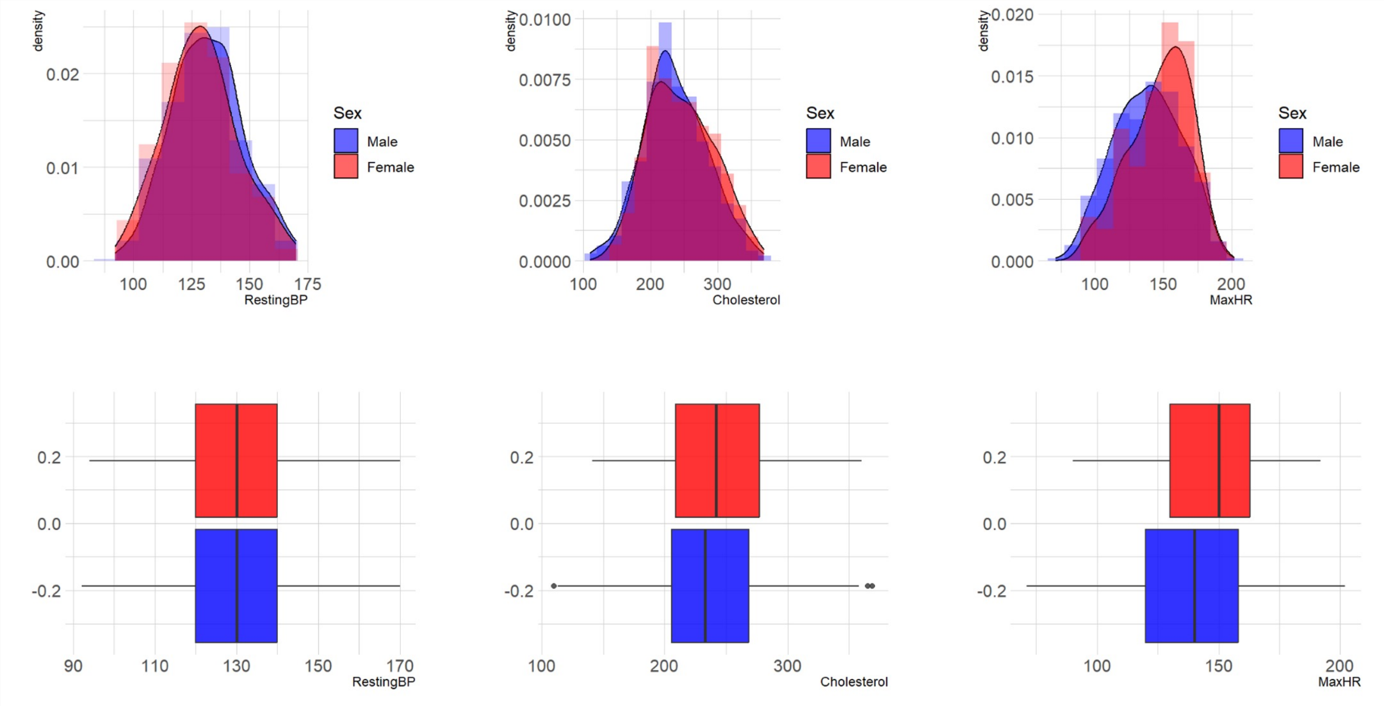
Distribution of variables & visualization of relationship between variables

Data characteristics-overall



- Age range from 22-77, Mean = 53
- 76% Male
- 46% heart disease

Data characteristics-gender difference



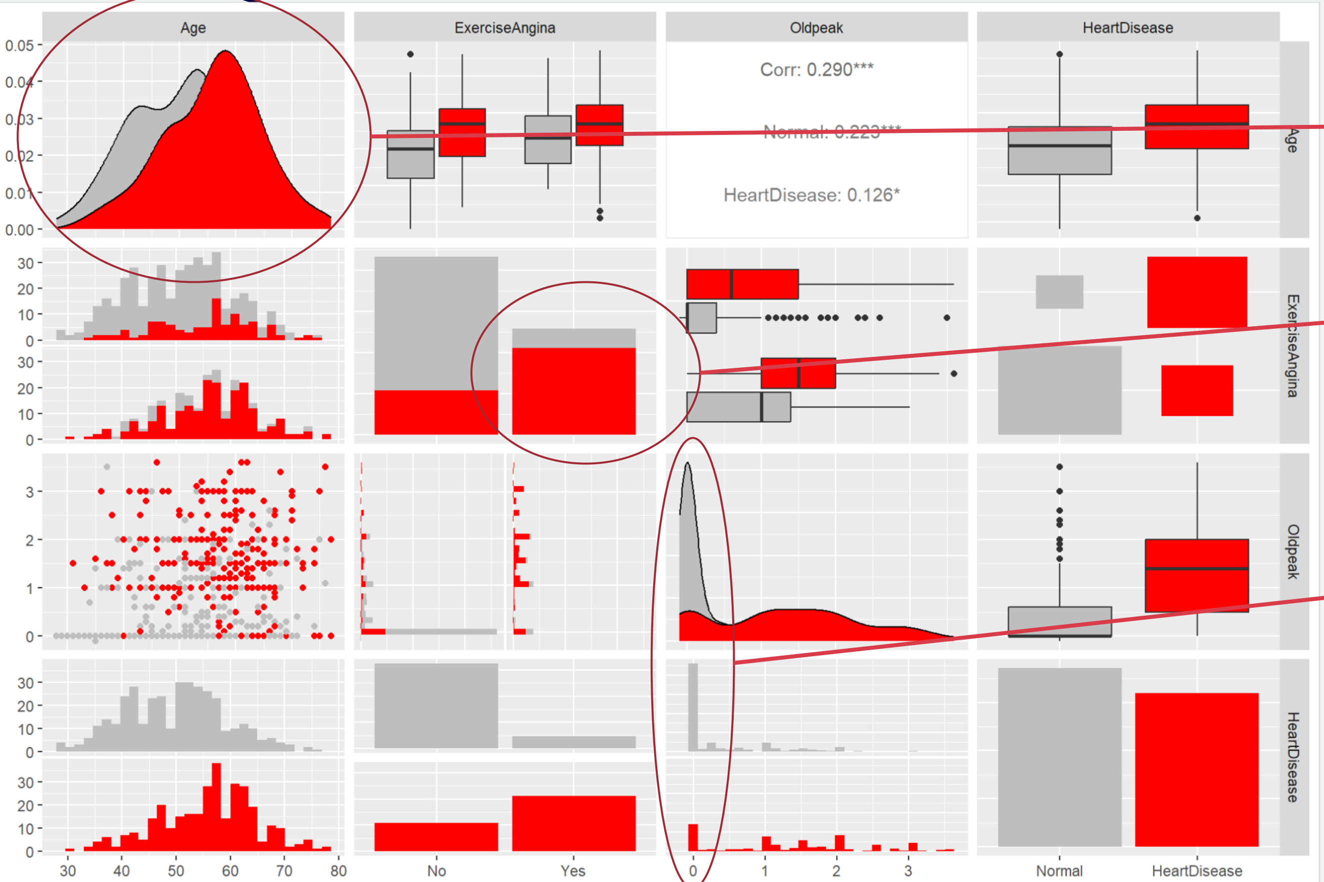
→ Male

- Resting blood pressure

→ Female

- **Maximum heart rate**
  Serum cholesterol

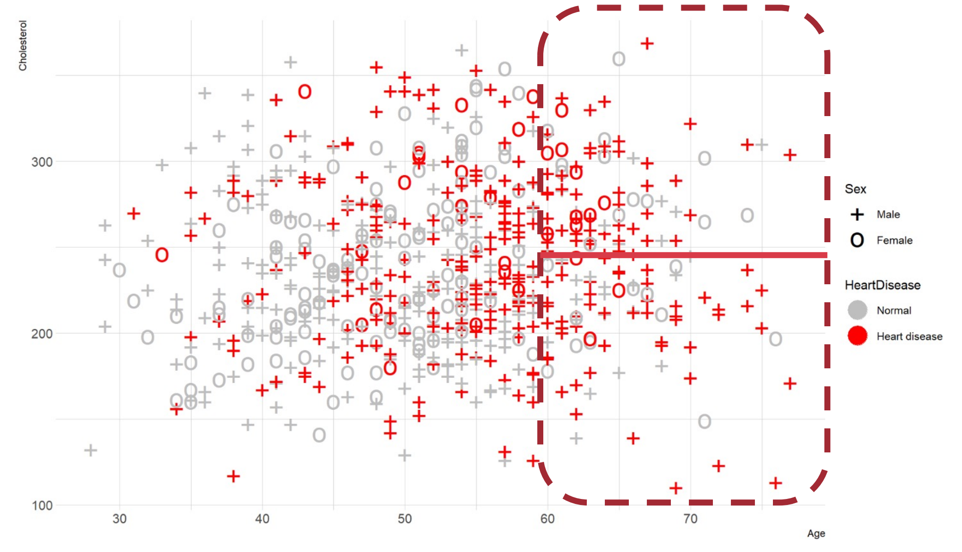
Patients analysis

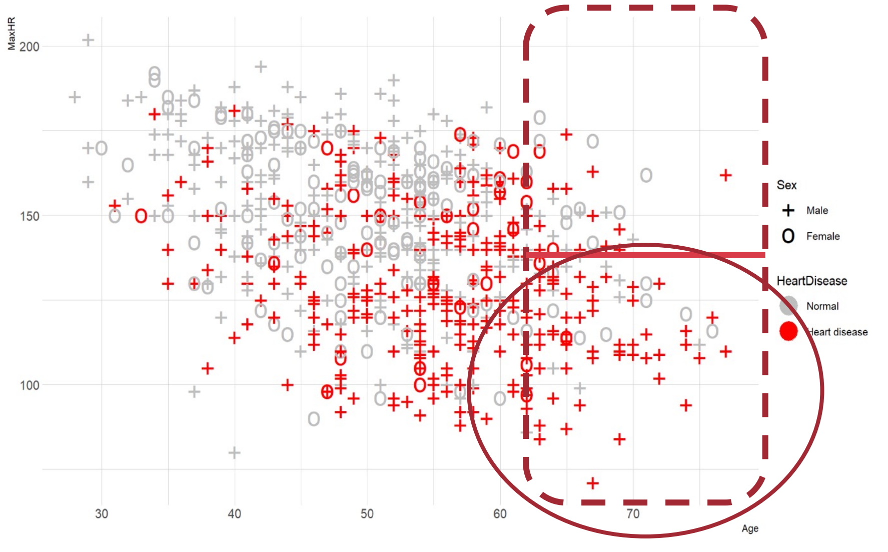


- Patients have higher age
- Exercise-induced angina significantly increases the risk of heart disease
- Who do not have a tendency to depression had almost no heart disease

Get exercise-induced angina regularly have more tendency to depression

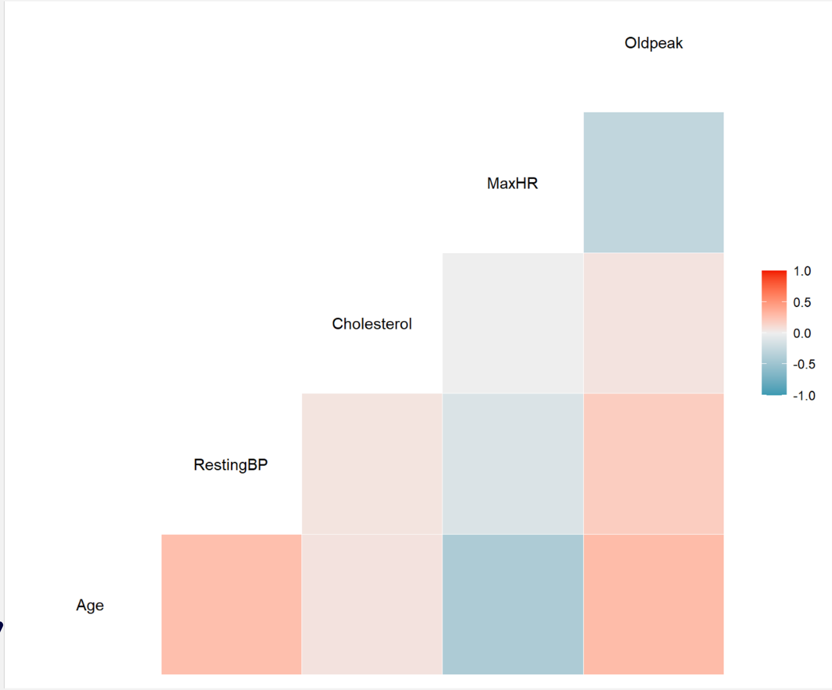
Patients analysis





- From analyzing serum cholesterol,we cannot find a clear pattern of how this variable influences the heart disease, but lower maximum heart rate achieved and age above 60 male have more tendency to get heart disease. We can also find that patients aged above 65 are mainly male.

Correlation analysis



- Strong Positive Correlation
  - ❑ Age and resting blood pressure
  - ❑ Age and depression level
- Strong Negative Correlation
  - ❑ Age and maximum heart rate achieved
  - ❑ Depression level and maximum heart rate achieved

04

Modeling

Logistic regression & KNN analysis &
decision tree & random forest

Logistic regression

✓	Age
✓	Sex
✓	ChestPainType
✓	ExerciseAngina
✓	Oldpeak

Variable selection (p≤0.05)

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -2.61952    0.66340  -3.949 7.86e-05 ***
Age              0.04716    0.01230   3.835 0.000125 ***
Sex1            -1.49550    0.28145  -5.314 1.07e-07 ***
ChestPainType2  -1.53680    0.26461  -5.808 6.33e-09 ***
ChestPainType3  -2.02776    0.31868  -6.363 1.98e-10 ***
ChestPainType4  -1.37807    0.43400  -3.175 0.001497 ** 
ExerciseAngina1  1.35065    0.24118   5.600 2.14e-08 ***
Oldpeak          0.79427    0.13225   6.006 1.91e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Model accuracy test

83.45%

	cardio.test.result	
lr.fit3.pred	0	1
0	67	12
1	11	49

Take 80% as training data, the rest for test purpose

```
#set train data
cardio.lol<-lr
set.seed(10)
train=sample(1:nrow(cardio.lol), nrow(cardio.lol)*0.8)
cardio.train=cardio.lol[train,]
cardio.test=cardio.lol[-train,]
```

And then build up a contingency table to see the accuracy of our prediction

```
table(lr.fit3.pred,cardio.test.result)
```

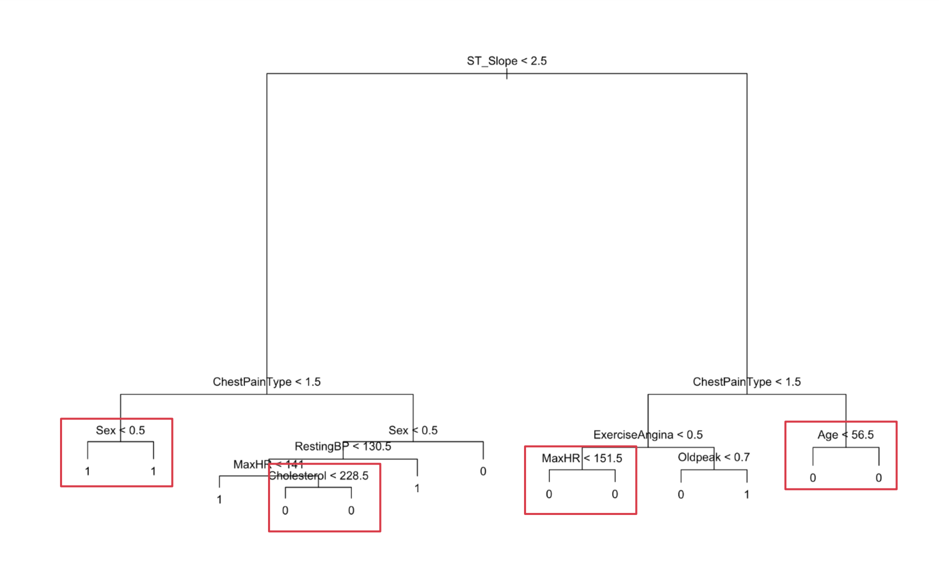
K-Nearest-Neighbor

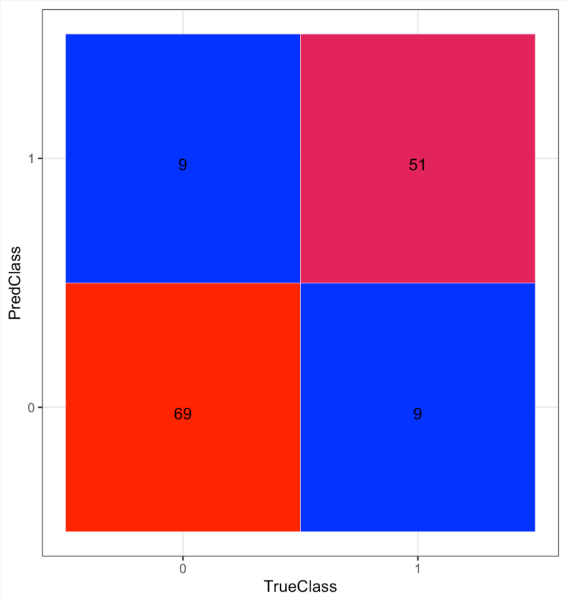
We use KNN algorithm to find the best prediction model by using different k value. In this model, dependent variable can be categorical.

```
> set.seed(8)
> predict.knn = knn(cardio.data.knn.
> table(predict.knn,number.test)
           number.test
predict.knn  0  1
          0 50 27
          1 18 44
> mean(predict.knn==number.test)
[1] 0.676259
```

- First, we set seeds and split data into two subsets, one training data and one test data.
- Second, we use 80% of our data to test out our training data for better accuracy. Then, we run the KNN with k=1,2,3,4,5,6,7,8,9,10,11 to find out under which value of k we have the highest accuracy.
- Finally, we found out that k=9 is the best model based on the accuracy of predictions for the test data.

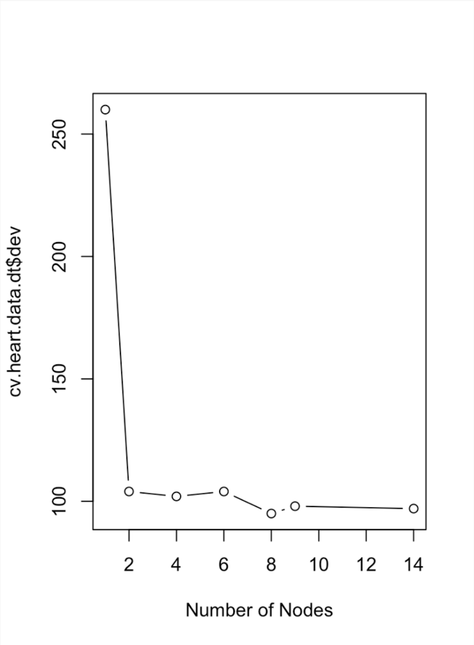
Decision tree with CART model

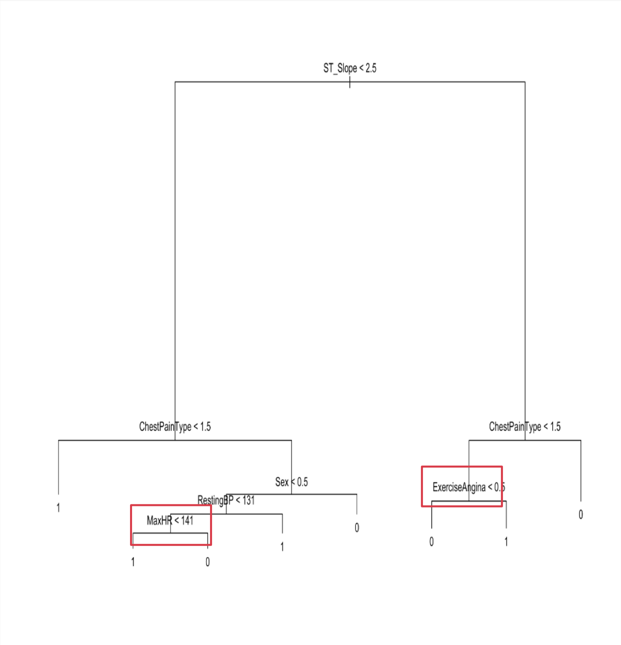


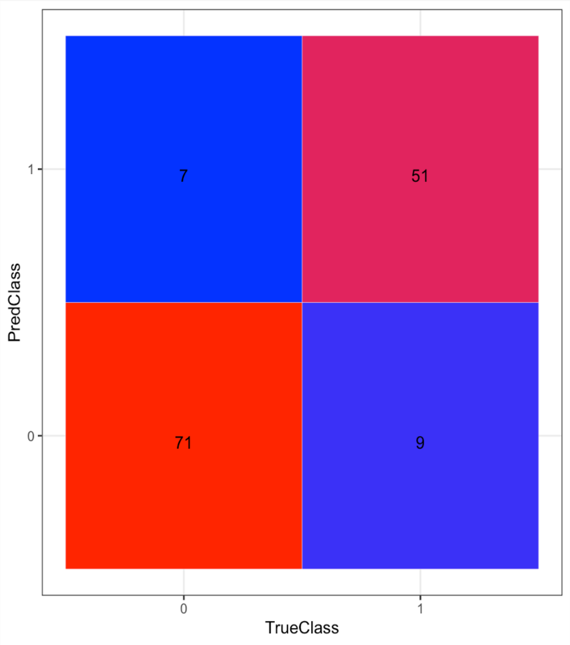


Accuracy: 86.96%

Pruned decision tree





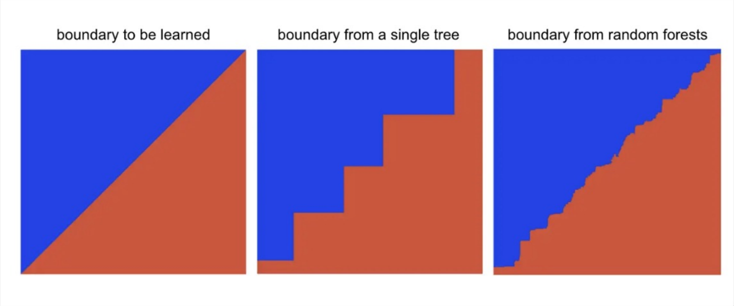


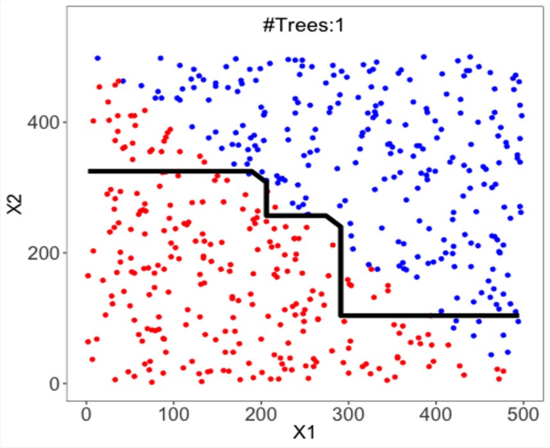
Using CV to avoid overfit, find number of nodes with least deviance

Accuracy: **88.41%**

86.96%

Random forest





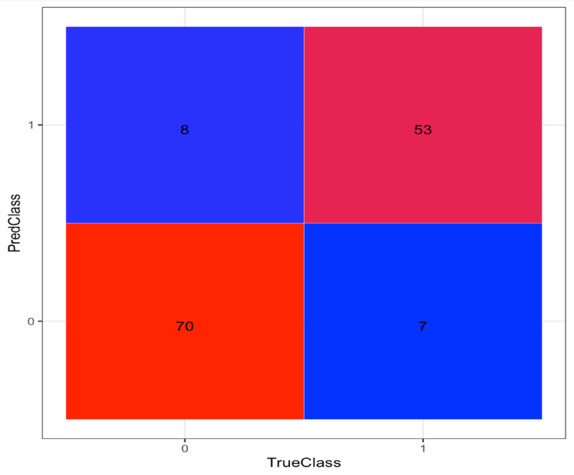
- Trees are unpruned
- Trees are diverse
- Handling overfitting

Random forest

```
Call:
 randomForest(formula = heartdisease ~ ., data = heart.data.rf,      mtry = 11, importance = T, subset
 = train.rf)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 11

        OOB estimate of  error rate: 16.43%
Confusion matrix:
    0   1 class.error
0 246  48      0.1633
1  43 217      0.1654
```

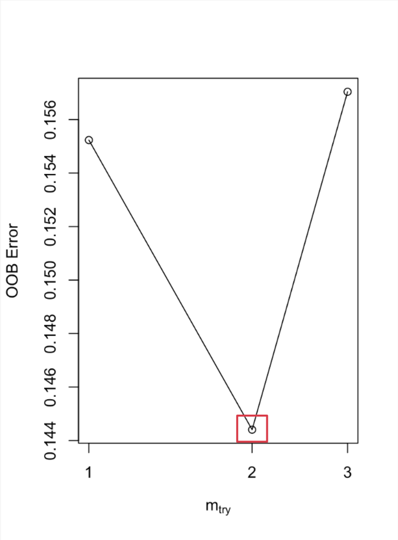
Considering all **11** variables
at each split



Accuracy: **89.13%**

88.41%

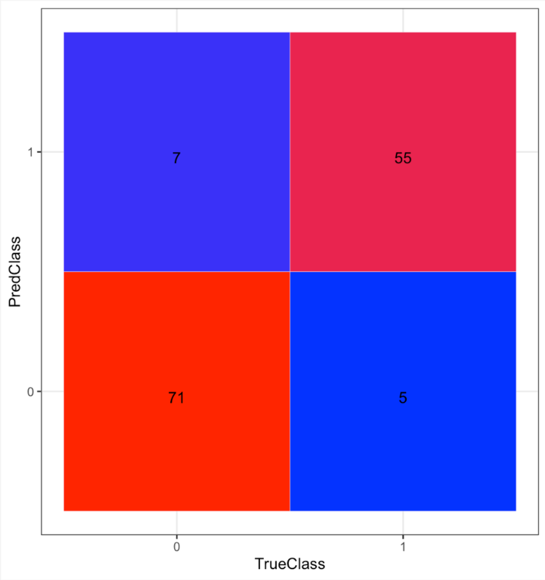
Random forest (cont'd)



Choosing 2 variables in each split with minimum out of bag error

```
Call:
 randomForest(formula = heartdisease ~ ., data = heart.data.rf,      mtry = 2, importance = T, subset
 = train.rf)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 14.44%
Confusion matrix:
    0   1 class.error
0 249  45      0.1531
1  35 225      0.1346
```



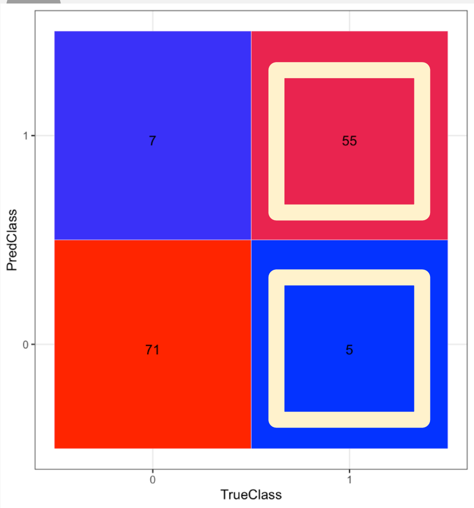
Accuracy: **91.30%**

83.13%

05

Evaluation & Conclusion

Evaluation



$$TPR = \frac{TP}{TP+FN}$$

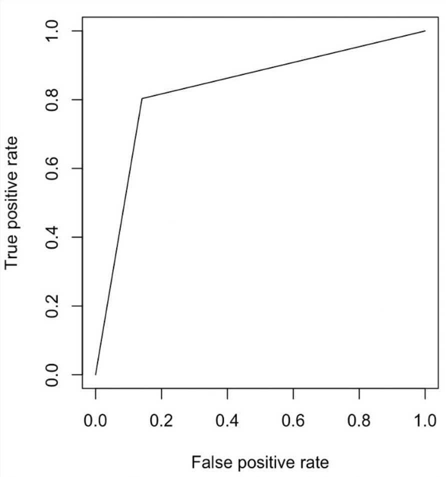
Model	Accuracy	Precision rate	Recall rate	F1-score	MSE
Logistic Regression	83.45%	0.8167	0.8033	0.8099	1.086
KNN	67.63%	0.7097	0.6197	0.6617	0.5690
Decision Tree	88.41%	0.8793	0.85	0.8644	0.3405
Random Forest	91.3%	0.8871	0.9167	0.9016	0.2949

1.The problem of sample imbalance leads to high accuracy results.

2.The higher the recall rate, the higher the probability that an actual bad user (patients) will be predicted, meaning that it is better to kill a thousand wrongly than to spare one.

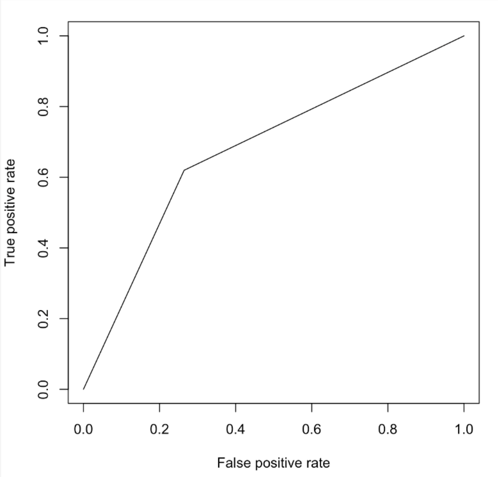
3.This is critical in our project because we should not let go of any potential patients!

Evaluation
Roc Plots & AUC



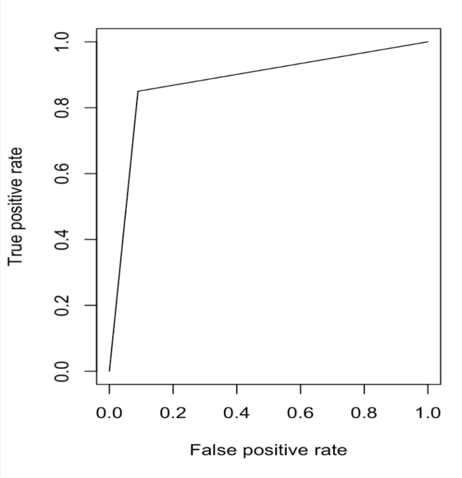
Logistic Regression

AUC: 0.8311



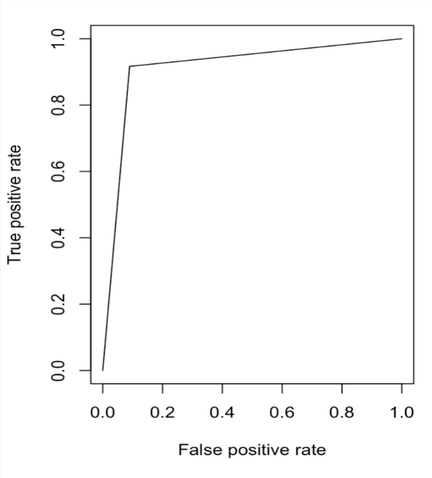
KNN

AUC: 0.6775



Decision Tree

AUC: 0.8801



Random Forest

AUC: 0.9135

Conclusions

Age	Target HR Zone 50-85%	Average Maximum Heart Rate, 100%
20 years	100-170 beats per minute (bpm)	200 bpm
30 years	95-162 bpm	190 bpm
35 years	93-157 bpm	185 bpm
40 years	90-153 bpm	180 bpm
45 years	88-149 bpm	175 bpm
50 years	85-145 bpm	170 bpm
55 years	83-140 bpm	165 bpm
60 years	80-136 bpm	160 bpm
65 years	78-132 bpm	155 bpm
70 years	75-128 bpm	150 bpm

- **Random Forest is the best model for this dataset**
- **Chest Pain would not directly lead to heart disease**
- **Age & Sex have a strong & positive relationship with heart disease**
  (Man, 65 ages above—more likely to get)
- **Lower maximum heart rate achieved, higher resting blood pressure & higher exercise-induced angina events would result in heart disease**

Thank You!