Heart Failure Disease Factor Analysis

Section 1. Summary

Heart diseases are the number one cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. People with heart disease or who are at high cardiovascular risk need early detection and management wherein a machine learning model can be of great help. In this project, we use the heart disease datasets, containing 11 medical, professional, and heart-disease related variables to predict the heart failure event. We first sort out and clean our data. Secondly, we make an exploratory data analysis to dig out the deep relationship between different variables. Thirdly, we build up logistic regression, KNN, decision trees, and random forest models to fit our data. Finally, by comparing diverse indexes of the results of four models, we find out that random forest has the best prediction result.

Section 2. Data Description

We use the heart disease datasets from UCI Machine Learning Repository *(see links in Reference & Data Source)*. We combine five independent heart disease datasets, including Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog Data Set. The combined dataset has a total of 1190 observations and 11 variables. The response variable is whether the heart disease event happens, which is binary.

The 11 variables are age, sex, chest pain type (ChestPainType: the variable name we use in the dataset), resting blood pressure (RestingBP), serum cholesterol (Cholesterol), fasting blood sugar (FastingBS), resting electrocardiogram result (RestingECG), oldpeak (Oldpeak), maximum heart rate achieved (MaxHR), exercise-induced angina (ExerciseAngina), the slope of the peak exercise ST segment (ST_slope) and heart disease. Specifically, serum cholesterol means the fat in people's blood. The slope of the peak exercise means the shift relative to exercise-induced increments in heart rate. Exercised-induced angina means a kind of pain around the heart.

Section 3. Data wrangling

3.1 Sorting out the data

In the combined dataset, some variables are demonstrated by characters. For instance, the chest pain type is shown as TA, ATA, NAP, and ASY four levels in characters. To fully utilize the

information in these variables, we replaced all character levels into numbers based on their actual seriousness, such as 4, 3, 2, and 1, respectively *(for full value information disclosure, please refer to "0. Data Cleaning Codes.R")*.

3.2 Remove duplicate value and missing value

To narrow down the redundant information, we delete 272 duplicated data in our dataset. Then, we further sort out our data and remove the missing value in the dataset.

3.3 Remove outliers

By drawing the box plot of the continuous variables with R, we could easily find that there are some outliers in variables. Then, we also delete outliers to prevent them from negatively affecting the inaccuracy of the model's prediction. Finally, we use 692 observations to do further exploration and fit in our models.

3.4 Feature selection

Since we can not guarantee that all of the variables are significantly related to heart failure, we make a feature selection analysis. For categorical variables, we use chi-square to test whether they have a statistically significant relationship with heart disease events. For continuous variables, we use one-way ANOVA to test whether they have a statistically significant relationship with heart disease events. Fortunately, all of the variables selected are significantly related to the response variable, heart disease event. The test results are shown in *Exhibit 1 & 2*.

Section 4. Exploratory Data Analysis

In the combined dataset, the minimum age is 28 and the maximum age is 77, the average is 53, meaning the dataset mainly describes the middle-elders. 76% of them are male and 46% of them have heart disease, *as shown in Chart 4.1*.

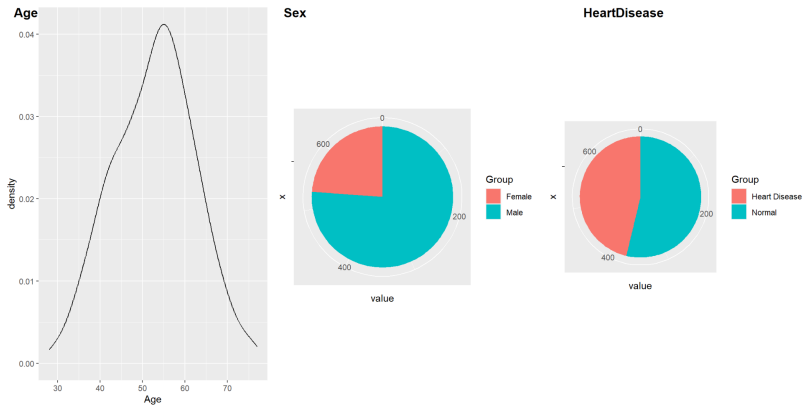


Chart 4.1

We find that the average age of those with heart disease was higher than those without heart disease, and that exercise-induced angina significantly increases the risk of heart disease. Those who do not have a tendency to depression probably have no heart disease, while those who get exercise-induced angina regularly have more tendency to depression *as shown in Chart 4.2*.

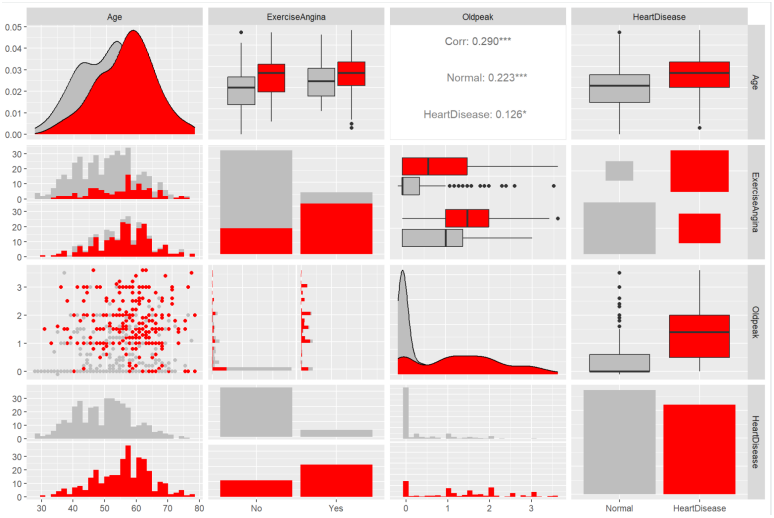


Chart 4.2

Physiologically, resting blood pressure level is higher in males than in females, but the maximum heart rate achieved and serum cholesterol are higher in females than in males *as shown in Chart 4.3*.

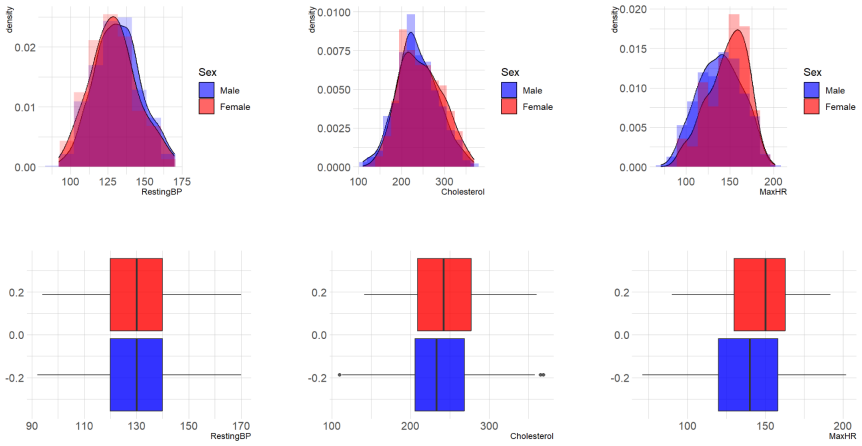


Chart 4.3

We perform correlation analysis on continuous data and we find a strong positive correlation between age and resting blood pressure, depression level, and a strong negative correlation with maximum heart rate achieved. Notably, there is a strong negative correlation between depression level and maximum heart rate achieved *as shown in Chart 4.4*.

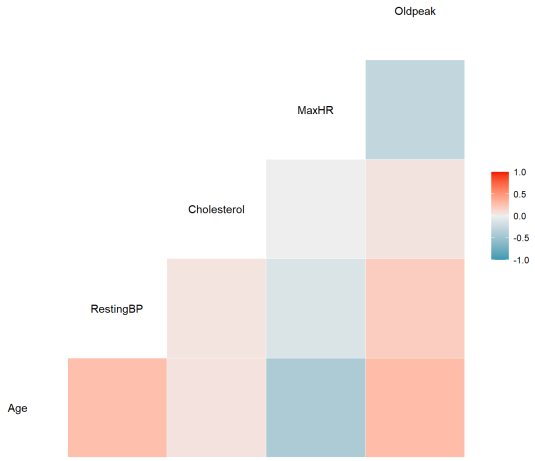


Chart 4.4

From analyzing serum cholesterol, we cannot find a clear pattern of how it influences heart disease. A lower maximum heart rate achieved and age above 60 males are more likely to get heart disease. We can also find that heart failure patients aged above 65 are mainly male *as shown in Chart 4.5*.

Chart 4.5

Section 5. Modeling

5.1 Logistic regression

We first define cardiovascular disease as HeartDisease as a response variable and hypothesize that not all of the variables are significant in predicting the response variable. Thus, to filter out the insignificant predictors, we carry out a logistic regression simulation that includes every variable and then we only keep those with a p-value less than 0.05 (5% significance level). After selecting relevant predictors, we begin setting our training data which includes 80% of the original data, and leave the rest for the testing purpose. Finally, we compare the prediction result

and the actual result of HeartDisease, and we achieve an 83.45% accuracy. The ROC plot is shown in *Exhibit 3* and the AUC value is 0.8311. The precision rate, recall rate, F1 score, and MSE are 0.8167, 0.8033, 0.8099, and 1.086, respectively.

In our model, the coefficient for variable Sex1(Female=1, Male=0) is -1.413, meaning that males are more likely to suffer from cardiovascular disease. For ExerciseAngina, the 1.2757 coefficient indicates that there is a positive relationship between Exercise-induced angina and cardiovascular issue. As shown in the model summary in *Exhibit 4*, having coefficients of 0.8741 and 0.047, oldpeak (level of depression) and age also contributes to the heart problem. These results are just as we have expected, yet what surprises us is the negative correlation between chest pain and cardiovascular disease. The negative coefficients tell us that patients with different levels of chest pain are less likely to have cardiovascular disease. To verify this result, we search for terms on websites and find that "more often than not, chest pain does not signal a heart attack". And a study of emergency room visits states that less than 6% of patients arriving with chest pain had a life-threatening heart issue. This discovery might partially explain why they are not positively related.

5.2 K-Nearest-Neighbor

We use KNN algorithm to find the best prediction model by performing different k values. In this model, the dependent variable can be categorical. 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 find out that k=9 is the best model based on the accuracy of predictions for the test data.

The accuracy of the final KNN model is 67.63%. The ROC plot is shown in *Exhibit 5* and the AUC value is 0.6775. The precision rate, recall rate, F1 score, and MSE are 0.7097, 0.6197, 0.6617, and 0.5690, respectively.

5.3 Decision Tree

We build a decision tree to predict the response variable. With defaults, the first decision tree contains 9 variables and 13 terminal nodes *(shown as Exhibit 6)*. Separating the training set and the test set, we could run predictions on the default tree with an accuracy of 86.96%. Then, we use cross-validation to find the number of nodes with the smallest deviance, which is eight *(shown as Exhibit 7)*. We prune the tree with eight nodes *(shown as Exhibit 8)* and repeated the

training and testing process, reaching an accuracy of 88.41%. The ROC plot is shown in *Exhibit 9* and the AUC value is 0.8801. The precision rate, recall rate, F1 score, and MSE are 0.8793, 0.85, 0.8644, and 0.3405, respectively.

5.4 Random Forest

To optimize the prediction power of the decision tree, we also use random forests to predict the response variable. We first tried all 11 predictors (mtry) to be considered for each split of trees, and 500 trees. After developing the training set and the test set, we predict the test dataset with above-mentioned settings with an 89.13% accuracy rate. Next, we want to find the number of variables tried at each split with the minimum out-of-bag (OOB) error. In our dataset, mtry = 2 has the least OOB *(shown as Exhibit 10)*. Using two predictors at each split, the prediction dataset reaches an accuracy of 91.3%. The ROC plot for the random forest is shown in *Exhibit 11* and the AUC value is 0.9135. The precision rate, recall rate, F1 score, and MSE are 0.8871, 0.9167, 0.9016, and 0.2949, respectively.

Section 6. Evaluation and Conclusion

The accuracy rate might be affected by the unbalance of the sample. For instance, if there are more people with heart disease (Y=1) than people without heart disease (Y=0) instead of Y=1:Y=0=1:1, the accuracy would be affected greatly. The problem of sample imbalance leads to high-accuracy results. That is, if the sample is unbalanced, the accuracy is invalidated. Thus, it also affects our decision to find the best model. To measure our models more precisely, we use several indexes including precision rate, recall rate, F1-score, ROC plot, AUC and MSE. Precision rate, recall rate, F1-score, and AUC are calculated based on the confusion matrix. Particularly, precision rate means the probability of the sample actually being positive out of all the samples predicted to be positive, meaning how certain we are that the prediction will be correct out of those predicted to be positive. In addition, according to the formulation of recall rate, the higher the recall rate, the higher the probability that an actual bad user will be predicted, meaning that it is better to kill a thousand wrongly than to spare one. This is critical in our project because we should not let go of any potential patients.

According to the prediction results of four models, we could find that the random forest model has the highest accuracy, precision rate, recall rate, and F1-score value. Also, it has the lowest

MSE value, meaning that it has the least error in prediction. And the important index recall rate is very high, about 0.9167. And this model could be universally used to predict heart disease. Based on the exploratory data analysis and models, we find that chest pain would not directly lead to heart disease. Also,a lower maximum heart rate achieved, higher resting blood pressure and higher exercise-induced angina events would result in heart disease. In addition, age and sex have a strong and positive relationship with heart disease. The evaluation result is shown in *Exhibition 12*.

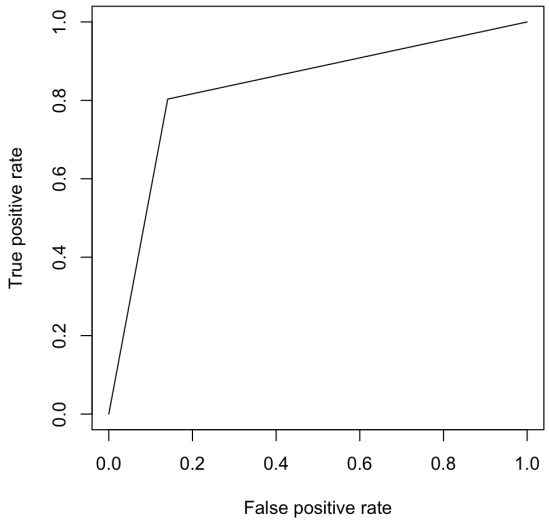
Appendix to Project Write-up

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

(Exhibit 1: Chi-square test results)

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

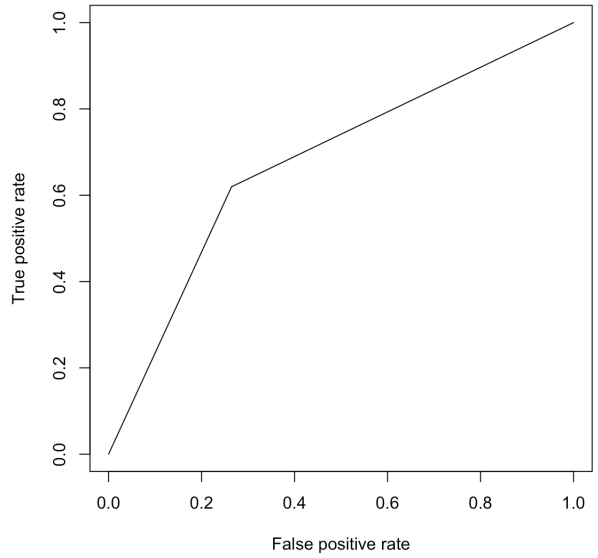
(Exhibit 2: one way ANOVA test results)



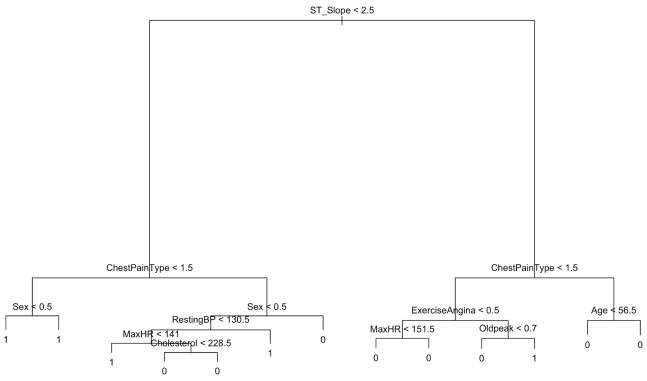
(Exhibit 3: ROC plot of Logistic Regression)

```
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 ‘ ’ 1
```

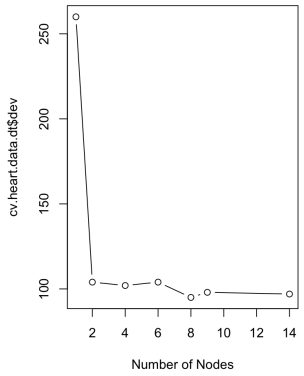
(Exhibit 4: Summary of Logistic Regression)



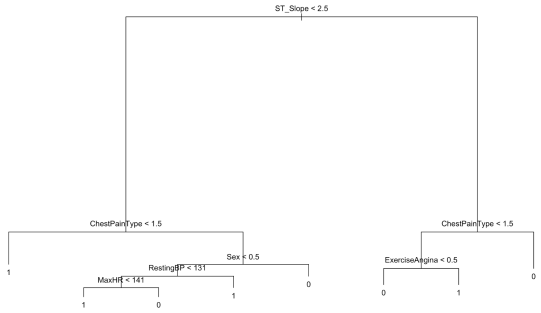
(Exhibit 5: ROC plot of KNN)



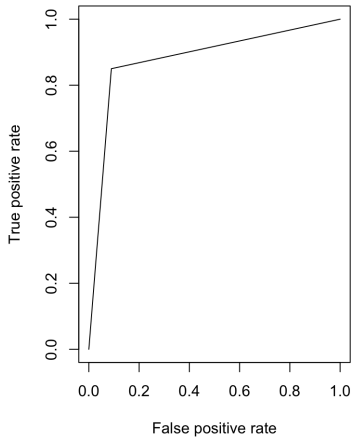
(Exhibit 6: Decision tree with default settings)



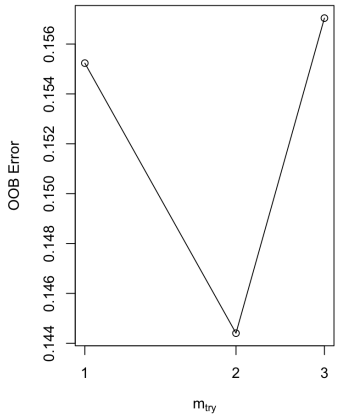
(Exhibit 7: Number of nodes with corresponding deviances)



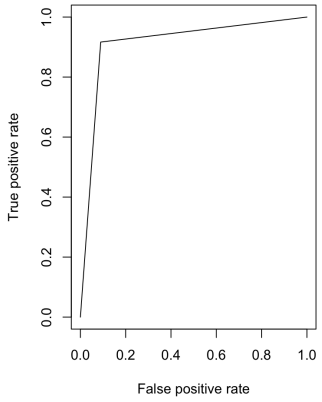
(Exhibit 8: Pruned tree with 8 terminal nodes)



(Exhibit 9: ROC plot of decision tree)



(Exhibit 10: Mtry with the corresponding OOB error)



(Exhibit 11: ROC plot of random forest)

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

(Exhibit 12: Table of evaluation results)

Reference & Data Source

Reference

https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates

Raw Data Source

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Combined Data Source

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction