MA308 Statistical Computation and Software (Fall 2020)

Project Report

Report title:

Symptomatic factors on Heart Disease: a cross-group comparative study

Members: Zhai Yibo, Shen Hanpu, Luo Yiling

Student Number: 11812724, 11812718, 11812532

Major: Statistics

Content

1. Introduction

1.1 Background

1.2 Objectives of research

1.3 Case data and terminology

1.3.1 Case data

1.3.2 Quick-review of all variables

1.3.3 Terminology explanation

2. Data visualizing

3. Generalized Multiple Linear Regression Model

3.1 Model assumption

3.2 Variable selection

3.3 Model diagnosis

3.4 Estimation of reduced model

3.5 Summary

4. Advanced models

4.1 LDA (Linear Discrimination Analysis) model

4.1.1 Motivation

4.1.2 Model construction

4.1.3 Model diagnosis

4.2 KNN Classifier

4.2.1 Motivation

4.2.2 Model construction

4.2.3 Model diagnosis

4.3 Random Forest

4.3.1 Motivation

4.3.2 Model construction

4.3.3 Model diagnosis

4.4 ANN (Artificial Neural Network)

4.4.1 Motivation

4.4.2 Model construction

4.4.3 Model diagnosis

4.5 Model comparison

5. Conclusions and summary

5.1 Conclusions

5.2 Limitations and further study

6. References

Abstract

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. This research is aimed at identifying the significant explanatory variables and predicting the heart disease with given condition. In this report, generalized linear model and some advanced models in machine learning, such as LDA and ANN to explore the factors affecting the target variable. The report reveals that sex and asymptotical chest pain are the main factors. Finally, this report summary the limitations and provide some direction of further study.

.

1. Introduction

1.1 Background

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.**[1]**

1.2 Objectives of research

This research is aimed at identifying the significant explanatory variables and predicting the heart disease with given condition. Furthermore, providing the approaches to preventing heart disease is also a target in the research. To conquer this problem by part, the research analyzes the problem by following procedures:

Firstly, we use data visualization and display data.

Secondly, establish a generalized linear model to explore the factors affecting the target variable.

Thirdly, apply more advanced models in machine learning, such as LDA and ANN to predict better.

Lastly, conclude the research and summary the entire procedure.

1.3 Case Data and terminology

1.3.1 Case Data

The dataset called *Cleveland Heart Disease dataset* is from UCI repository, which contains information about 297 individuals, including 14 attributes.

1.3.2 Quick-review of all variables:

(Abbreviation: “int”: integer, “cat”: categorical, “con”: continuous, “dep”: dependent)

Predictors	**Explanation**	
Age	The person’s age in years	**int**
Sex	The person’s sex	**cat**

	1 = male 0 = female	
Chest-pain type	The chest pain experienced 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic	**int**
Resting Blood Pressure	Displays the resting blood pressure value of an individual in mmHg (unit)	**con**
Serum Cholesterol	displays the serum cholesterol in mg/dl (unit)	**con**
Fasting Blood Sugar	compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then: 1 (true) else: 0 (false)	**cat**
Resting ECG	displays resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hyperthrophy	**Int**
Max heart rate achieved	The person's maximum heart rate achieved	**con**
Exercise induced angina	1 = yes 0 = no	**cat**
Old-peak	ST depression induced by exercise relative to rest	**con**
Peak exercise ST segment	the slope of the peak exercise ST segment 1 = upsloping 2 = flat 3 = downsloping	**cat**
Number of major vessels	The number of major vessels colored by flourosopy (X 线透视检查) (0-3)	**int**

Thallium Test	displays in the thallium test: 0 = normal 1 = fixed defect 2 = reversible defect	**cat**
Diagnosis of heart disease	0 = no disease 1 = disease	**dep**

Table 1. Explanation for all variables

1.3.3 Terminology explanation **[2]**:

Angina (Chest Pain):

Angina is chest pain or discomfort caused when your heart muscle doesn’t get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.

Resting Blood Pressure:

Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.

Serum Cholesterol:

A high level of low-density lipoprotein (LDL) cholesterol (the “bad” cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the “good” cholesterol) lowers your risk of a heart attack.

Fasting Blood Sugar:

Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body’s blood sugar levels to rise, increasing your risk of a heart attack.

Resting ECG:

For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.

Max heart rate achieved:

The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.

Exercise induced angina:

The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands.

Thallium Test:

Thallium is injected during peak exercise and will be extracted by the hearts muscle tissue (myocardium). Ideally there will be a uniform (equal) distribution of thallium in all myocardial segments (‘0’). If a patient has heart disease, myocardial uptake is reduced and there will be a defect, which may be fixed (‘1’) and the thallium will not redistribute normally after some time or it may be reversible and the thallium will distribute normally after its initial defect (‘2’).

Old-peak:

The difference of peak values (T in Figure 1) of ECG between resting and after-exercising. This feature may be different between patients and common people.

Peak exercise ST segment:

Description for ST segment (J in Figure 1). Upsloping, flat and downsloping stand for the direction of wave. Flat slope may indicate ischemia of myocardium(心肌充血). Downsloping indicates angina caused by ischemia of myocardium. Upsloping means transmural ischemia(透壁型缺血), which may lead to myocardial infarction(心梗).

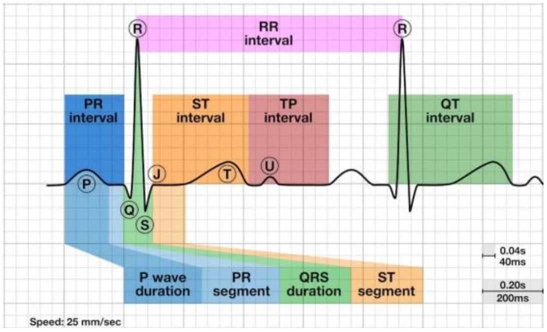


Figure 1. ECG diagram

2. Data visualizing

To know the entire data set, we explore some primary relation between heart disease with other variables, which are considered as high relevant factor with heart disease in common knowledge. Firstly, we examine the amounts of patients and normal people in different age groups. As shown in Figure 1, most patients concentrate in high age groups (over 55 years old). This is consistent with common knowledge.

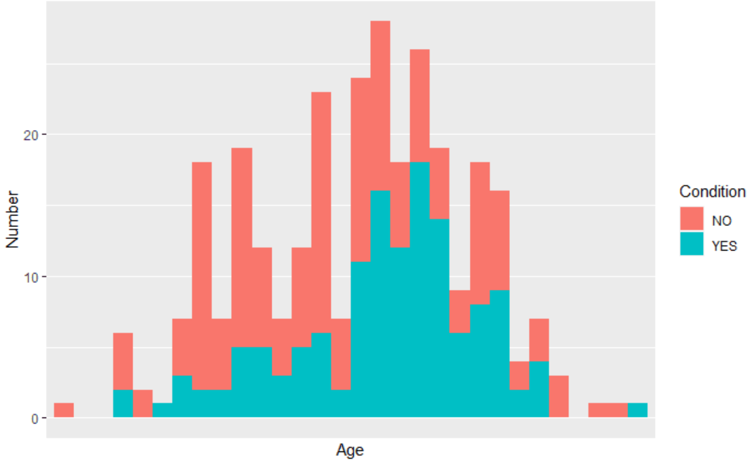


Figure 2. Age Vs Condition

Furthermore, the kind of chest pain also matters in the patients group. Most patients have asymptomatic chest pain, also called silent chest pain. From relative resource, silent pain is common cause in heart disease patients. **[3]**

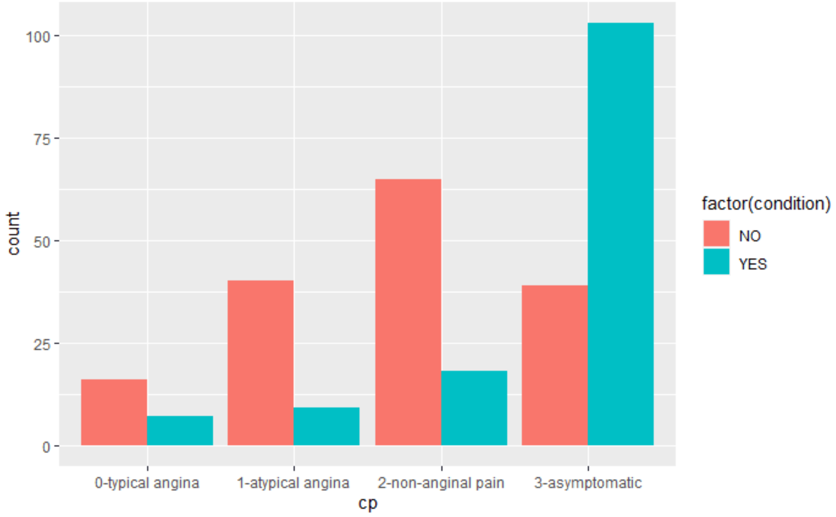


Figure 3. Chest pain Vs Condition

Additionally, we find that high maximum heart rate is a characteristic in the heart-disease group (Figure 4). In medical science, maximum heart rate reflects the ability of the cardiac to meet the body's oxygen needs.

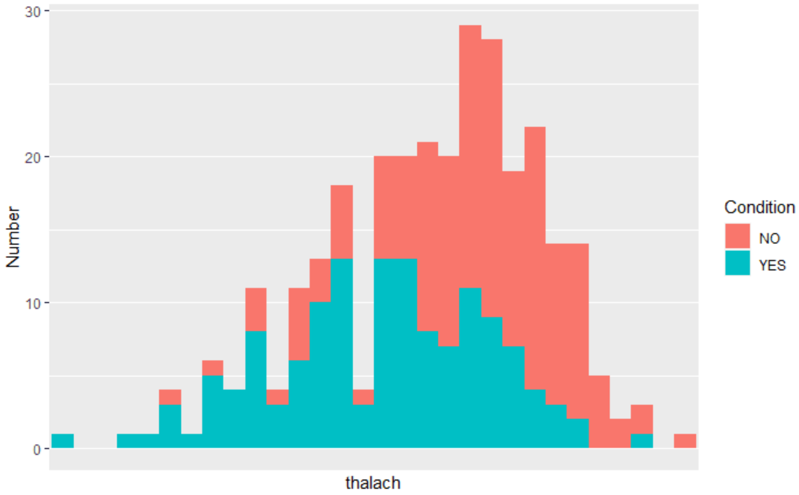


Figure 4. Maximum heart rate Vs condition

This heatmap (Figure 5) shows the correlations between the dataset features. From the heatmap, we can observe that the chest pain type (*cp*), exercise induced angina (*exang*), ST depression induced by exercise relative to rest (*oldpeak*), the slope of the peak exercise ST segment (*slope*), number of major vessels (0–3) colored by flourosopy (*ca*) and thalassemia (*thal*) are highly correlated with the heart disease (*condition*). We observe also that there is an inverse proportion between the heart disease and maximum heart rate (*thalch*).

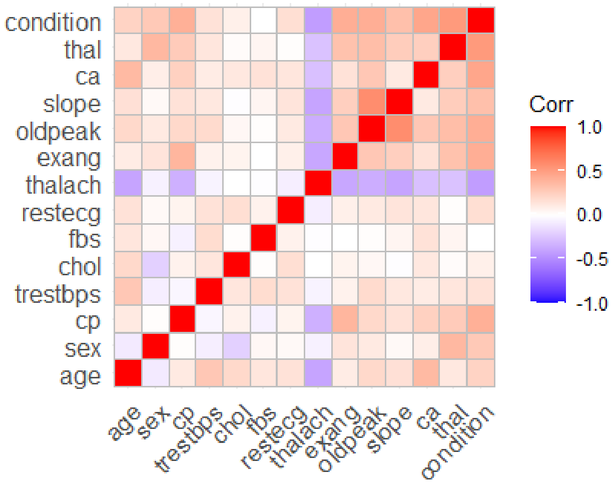


Figure 5. Attributes Correlation

Moreover, we can see that the age is correlated with number of major vessels (0–3) (*ca*) and maximum heart rate (*thalch*). There is also a relation between ST depression induced by exercise relative to rest (*old-peak*) and the slope of the peak exercise ST segment (slope). Moreover, there is a relation between the chest pain type (cp) and exercise induced angina (*exang*).

As shown in Figure 8, the people who have exercise induced angina. They usually suffer from asymptomatic chest pain, and they are more likely to have heart disease.

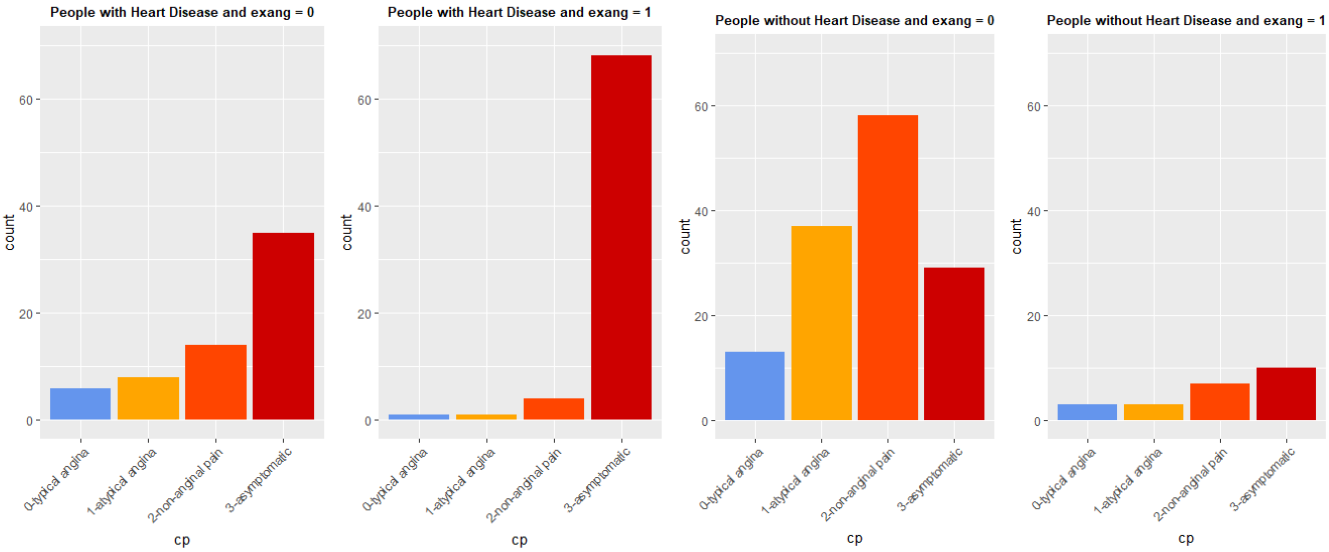


Figure 8. Chest Pain and Exercise Induced Angina

3. Generalized multiple linear regression

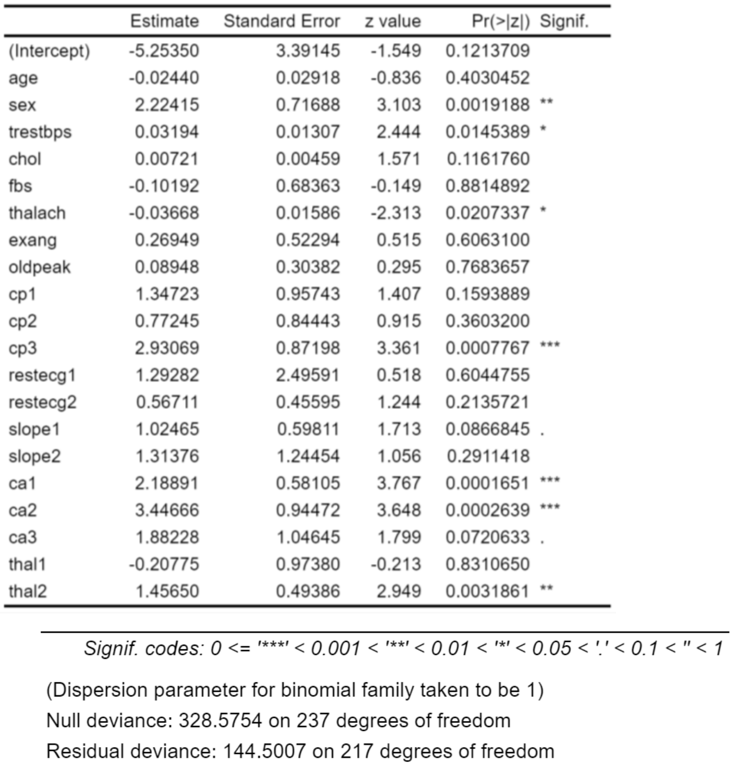
3.1 Model assumption & Data processing

Since the dependent variable Y_i, factor *condition,* is categorical (1 and 0), we assume $Y_i \sim \text{Binomial}(1, \text{p})$. Hence, we can use logistic regression model to predict the probability p.

About the data, we divide the datasets in to three parts, training, validation, testing, and they take 80%, 10% and 10% respectively.

3.2 Variable selection

Firstly, we take all variables into the model and get a primary view of model. As shown in table 1, 9 variables are significant with criterion α = 0.1. Then we test the multicollinearity among all variables by VIF (Table 2).



	Estimate	Standard Error	z value	Pr(>|z|)	Signif.
(Intercept)	-5.25350	3.39145	-1.549	0.1213709	
age	-0.02440	0.02918	-0.836	0.4030452	
sex	2.22415	0.71688	3.103	0.0019188	**
trestbps	0.03194	0.01307	2.444	0.0145389	*
chol	0.00721	0.00459	1.571	0.1161760	
fbs	-0.10192	0.68363	-0.149	0.8814892	
thalach	-0.03668	0.01586	-2.313	0.0207337	*
exang	0.26949	0.52294	0.515	0.6063100	
oldpeak	0.08948	0.30382	0.295	0.7683657	
cp1	1.34723	0.95743	1.407	0.1593889	
cp2	0.77245	0.84443	0.915	0.3603200	
cp3	2.93069	0.87198	3.361	0.0007767	***
restecg1	1.29282	2.49591	0.518	0.6044755	
restecg2	0.56711	0.45595	1.244	0.2135721	
slope1	1.02465	0.59811	1.713	0.0866845	.
slope2	1.31376	1.24454	1.056	0.2911418	
ca1	2.18891	0.58105	3.767	0.0001651	***
ca2	3.44666	0.94472	3.648	0.0002639	***
ca3	1.88228	1.04645	1.799	0.0720633	.
thal1	-0.20775	0.97380	-0.213	0.8310650	
thal2	1.45650	0.49386	2.949	0.0031861	**

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

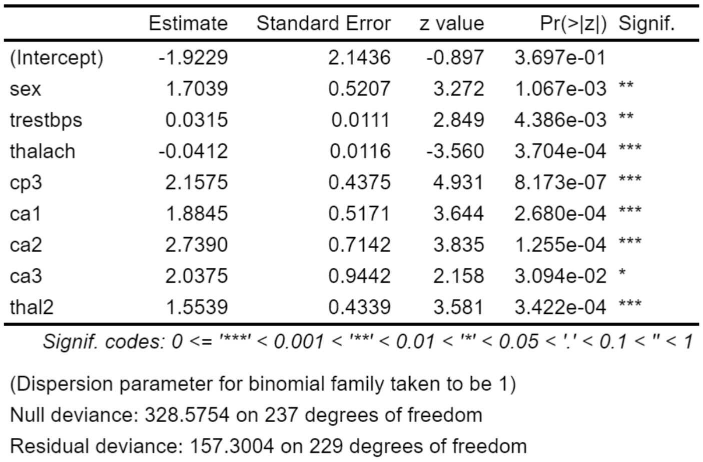
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 328.5754 on 237 degrees of freedom
Residual deviance: 144.5007 on 217 degrees of freedom

Table 1. Coefficients estimation

Factor	VIF
Age	1.54
Sex	2.18
Trestbps	1.37
Chol	1.45
Fbs	1.20
Thalach	1.91
Exang	1.21
Oldpeak	1.99
Cp1	2.54
Cp2	3.19
Cp3	**4.28**
Restecg1	1.07
Restecg2	1.17
Slope1	1.99
Slope2	1.84
Cal	1.34
Ca1	1.34
Ca2	1.55
Thal1	1.50
Thal2	1.34

Table 2. VIF of all variables

In Table 2, only VIF of *cp3* is over 4, which indicates there exists some multicollinearity among the explanatory variables. Therefore, we employ function *step()* to drop some insignificant variables. We get the reduced model. As shown in Table 3, the reduced model only contains 8 significant variables.



	Estimate	Standard Error	z value	Pr(>|z|)	Signif.
(Intercept)	-1.9229	2.1436	-0.897	3.697e-01	
sex	1.7039	0.5207	3.272	1.067e-03	**
trestbps	0.0315	0.0111	2.849	4.386e-03	**
thalach	-0.0412	0.0116	-3.560	3.704e-04	***
cp3	2.1575	0.4375	4.931	8.173e-07	***
ca1	1.8845	0.5171	3.644	2.680e-04	***
ca2	2.7390	0.7142	3.835	1.255e-04	***
ca3	2.0375	0.9442	2.158	3.094e-02	*
thal2	1.5539	0.4339	3.581	3.422e-04	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 328.5754 on 237 degrees of freedom
Residual deviance: 157.3004 on 229 degrees of freedom

Table 3. Estimation of coefficients in the reduced model

coefficient	
Sex	5.50
Trestbps	1.03
Thalach	0.96
Cp3	8.65
Ca1	6.58
Ca2	15.47
Ca3	7.67
Thal2	4.73

Table 7. Coefficients with exp-transform

3.4 Model diagnosis

Firstly, it's necessary to check the deviance and its distribution. In generalized linear model, the deviance should follow chi-square distribution, if n is large enough. As shown in Figure 9 and 10, the residuals roughly follow chi-square distribution, which indicates that the model captures all the features in the data.

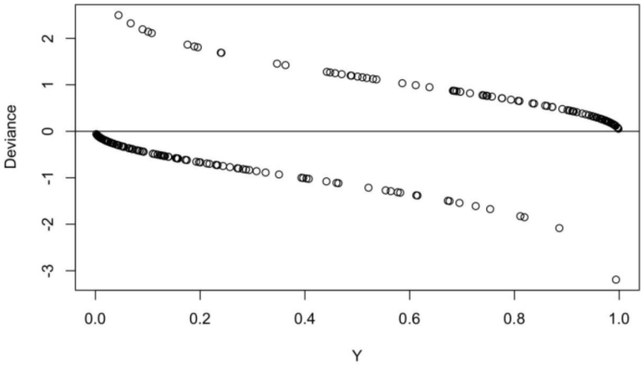


Figure 9. Residual plot

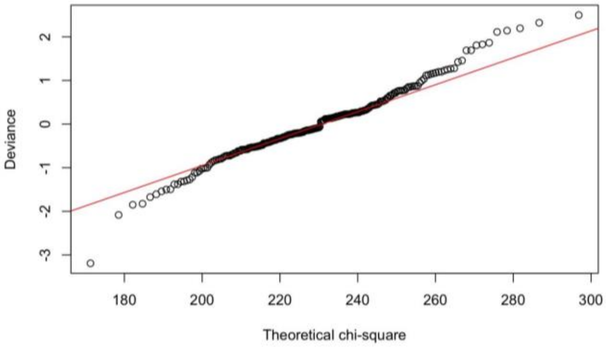


Figure 10. QQ-plot with chis-square

As shown in Figure 11, there are some influential points in the model. Particularly, patient 178 is an outlier (studentized residual > 3).

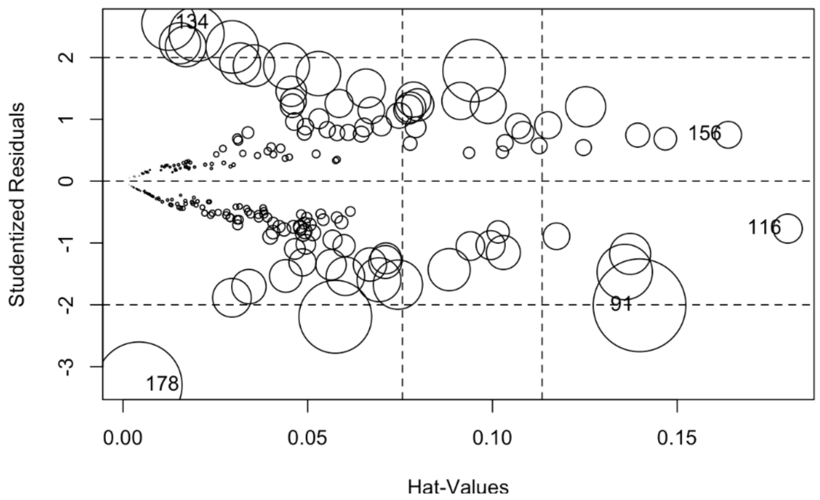


Figure 11. Influential points plot

Actually, all factors in the reduced model are significant (Table 4) and the multicollinearity among the explanatory variables can be acceptable with criterion VIF < 4 (Table 5).

Patients	StudRes	Hat	CookD
91	-2.01	0.14	0.09
116	-0.77	0.18	0.01
134	2.55	0.012	0.03
156	0.75	0.16	0.01
178	-3.30	0.004	0.08

Table 4. Influential points

Table 5. VIF of reduced model

Factor	VIF
Sex	1.32
Trestbps	1.18
Thalach	1.14
Cp3	1.17
Ca1	1.15
Ca2	1.14
Ca3	1.04
Thal2	1.12

To test the accuracy of model, we compute the confusion matrix of the reduced model (Table 6). The accuracy rate of model based on training data is 80% and we can accept this result. Furthermore, we'll employ more models to improve the efficiency and accuracy of prediction.

		Actual Response	
		Target_0	Target_1
Predicted Response	Target_0	14	4
	Target_1	2	10

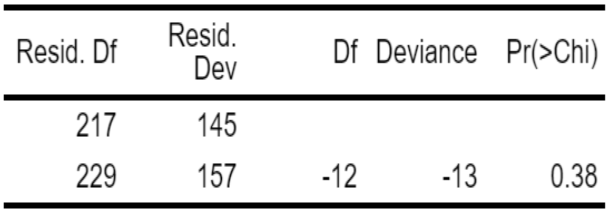
Table 6. Confusion matrix of generalized linear model

In summary, the reduced model is validated so we can estimate the model.

3.5 Estimation of reduced model

After model diagnosis, We employ function *AUC()* to rank the relative importance of explanatory variables. In Table 5, we can find that reversible defect of Thalassemia (Thal2) is the most important and asymptotical chest pain also plays a significant role in

	Factor	AUC
1	Thal2	0.81
2	Cp3	0.75
3	Thalach	0.70
4	Sex	0.67
5	Ca2	0.66
6	Trestbps	0.66
7	Ca3	0.54
8	Ca1	0.52



Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
217	145			
229	157	-12	-13	0.38

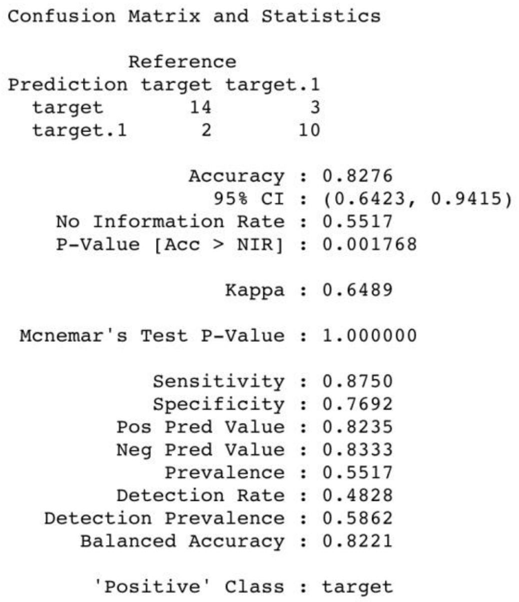
Table 5. Relative importance rank by *AUC()* Table 6. ANOVA of the reduced model

According to the ANOVA table, it suggests the final model is better, because the p-value > 0.05. Hence, we cannot reject the null hypothesis that the final model is as useful as the full model, which means the efficiency of model is improved.

3.6 Summary

After variable selection and model diagnosis, we find a reduced generalized linear model, where there are 8 significant variables, sex, resting blood pressure, maximum heart rate, chest pain type 3, number of major vessels (1,2,3) and thalassemia with fixed defect. Additionally, the multicollinearity is acceptable with VIF < 10 and the residuals follow chi-square distribution, which satisfies the assumption of logistic regression. Furthermore, the reduced model behaves barely satisfactory.

And the Table 3.6 shows the logistic model has a consistent performance at the test data, the accuracy is 82.76% and sensitivity is



```
Confusion Matrix and Statistics

          Reference
Prediction target target.1
  target       14        3
  target.1      2       10

               Accuracy : 0.8276
                 95% CI : (0.6423, 0.9415)
    No Information Rate : 0.5517
    P-Value [Acc > NIR] : 0.001768

                  Kappa : 0.6489

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.8750
            Specificity : 0.7692
         Pos Pred Value : 0.8235
         Neg Pred Value : 0.8333
             Prevalence : 0.5517
         Detection Rate : 0.4828
   Detection Prevalence : 0.5862
      Balanced Accuracy : 0.8221

       'Positive' Class : target
```

Table 3.6: Confusion matrix of logistic regression

4. Advanced models

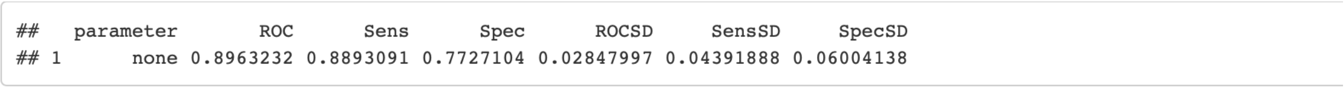
4.1 LDA (Linear Discrimination Analysis) model

4.1.1 Motivation

Since the dataset is quite small in the scale, so we want to improve the prediction accuracy by applying new models. In above concerns, we decided to try with the LDA model, because of the advantage of the LDA model is that it can perform well in small data case and doesn't require distribution assumption.

4.1.2 Model construction

Based on the Table 4.1.2, we can see the ROC value, sensitivity and specificity are good. And this supports our first belief that LDA worth considering.



```
##   parameter       ROC      Sens      Spec      ROCSD     SensSD     SpecSD
## 1      none 0.8963232 0.8893091 0.7727104 0.02847997 0.04391888 0.06004138
```

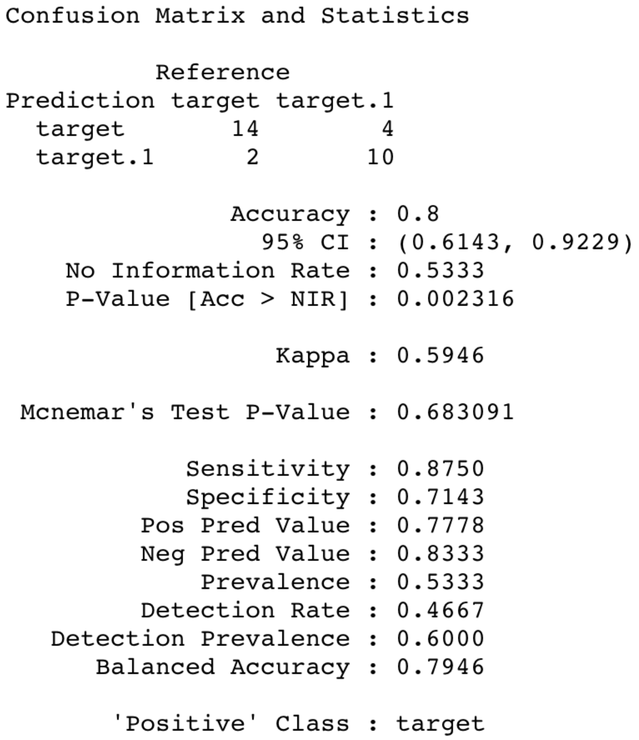
Table 4.1.2: LDA parameter on Train set

4.1.3 Model diagnosis

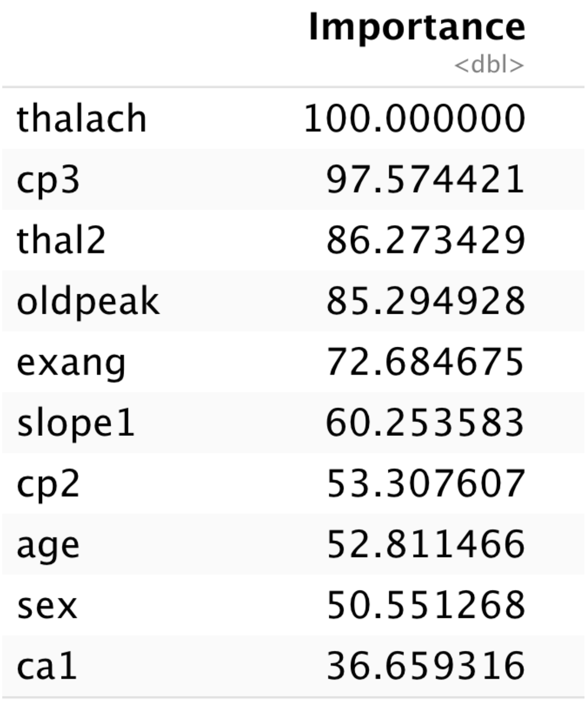
From the confusion matrix in Table 4.1.3.1 we can find that the LDA model has an accuracy of 80%

overall and sensitivity, specificity do not change much in the validation set.

Besides, we give the factors a rank of importance based on their ROC value. This could help us better understanding the data, because we can see this result is consistent with the previous logistic in Table 5, both of them shows thalach (The max heart rate), cp3 (Asymptomatic chest pain) and thal2 (reversable defect thalassemia) are ranked in the top 3. This motivates us to research further on those factors individually.



```
Confusion Matrix and Statistics

          Reference
Prediction target target.1
  target       14        4
  target.1      2       10

               Accuracy : 0.8
                 95% CI : (0.6143, 0.9229)
    No Information Rate : 0.5333
    P-Value [Acc > NIR] : 0.002316

                  Kappa : 0.5946

 Mcnemar's Test P-Value : 0.683091

            Sensitivity : 0.8750
            Specificity : 0.7143
         Pos Pred Value : 0.7778
         Neg Pred Value : 0.8333
             Prevalence : 0.5333
         Detection Rate : 0.4667
   Detection Prevalence : 0.6000
      Balanced Accuracy : 0.7946

       'Positive' Class : target
```



	Importance <dbl>
thalach	100.000000
cp3	97.574421
thal2	86.273429
oldpeak	85.294928
exang	72.684675
slope1	60.253583
cp2	53.307607
age	52.811466
sex	50.551268
ca1	36.659316

Table 4.1.3.1: Model diagnosis

Table 4.1.3.2: ROC importance

4.2 KNN (K-Nearest Neighbors)

4.2.1 Motivation

As KNN Algorithm is a non-parametric method which is easy to understand and operate, can be used both in classification and regression, and can be used for numerical and discrete data, so it’s suitable for us to use this method to learn this dataset and test.

4.2.2 Model construction

Based on the Figure 4.2.2.1, we choose the highest ROC and decide to use 19 as our neighbors' number, so that we construct a form such that while we put in a data, it could automatically retrieve the nearest 19 data and use them to predict the unknown one.

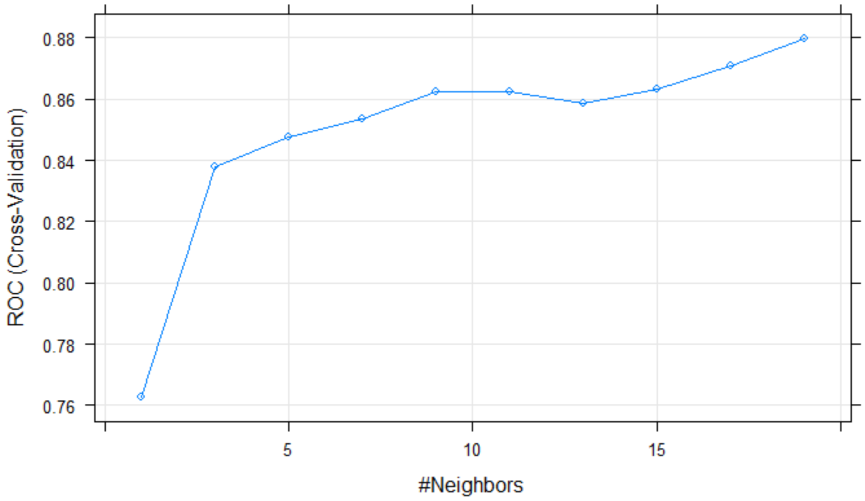


Figure 4.2.2: Plot of Neighbors' Number vs ROC

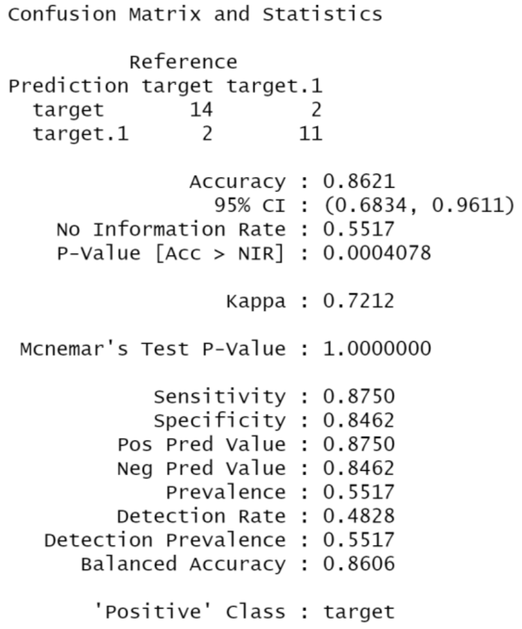
4.2.3 Model diagnosis

From the result of KNN, we can easily see that the ROC is a bit less than other method, but don't have significant difference, so we can also take into account this method.

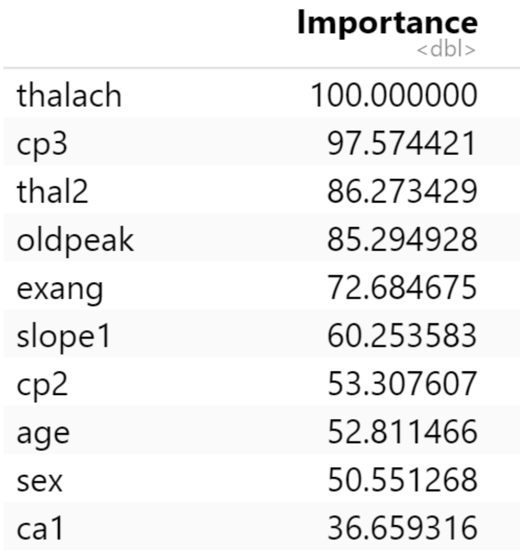
Neighbors' number	ROC	Sensitivity	Specificity
19	0.8795385	0.8910769	0.7090909

Table 4.2.3.1: KNN Model performance evaluation

Then, from the table of importance ranking, we can see that thalach, cp3 and thal2 are occupy the highest position, which is consistent with the previous results, showing that not only this method is useful, but also confirm our suspicions of some important factors.



```
Confusion Matrix and Statistics

          Reference
Prediction target target.1
  target       14        2
  target.1      2       11

               Accuracy : 0.8621
                 95% CI : (0.6834, 0.9611)
    No Information Rate : 0.5517
    P-Value [Acc > NIR] : 0.0004078

                  Kappa : 0.7212

 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 0.8750
            Specificity : 0.8462
         Pos Pred Value : 0.8750
         Neg Pred Value : 0.8462
             Prevalence : 0.5517
         Detection Rate : 0.4828
   Detection Prevalence : 0.5517
      Balanced Accuracy : 0.8606

       'Positive' Class : target
```



	Importance <dbl>
thalach	100.000000
cp3	97.574421
thal2	86.273429
oldpeak	85.294928
exang	72.684675
slope1	60.253583
cp2	53.307607
age	52.811466
sex	50.551268
ca1	36.659316

Table 4.2.3.2: Model Diagnosis

Table 4.2.3.3: Importance rating

4.3 Random Forest

4.3.1 Motivation

Random Forest is a basic machine learning method for the user to get feasible and effective results in a short period of time. It's easy to understand and can be adjusted through different parameters. It uses lots of decision tree so the result is more dependable, that's why we use this method.

4.3.2 Model construction

In order to get the best ROC result, we select 3 as the randomly selected predictors in our model and build 20 decision trees.

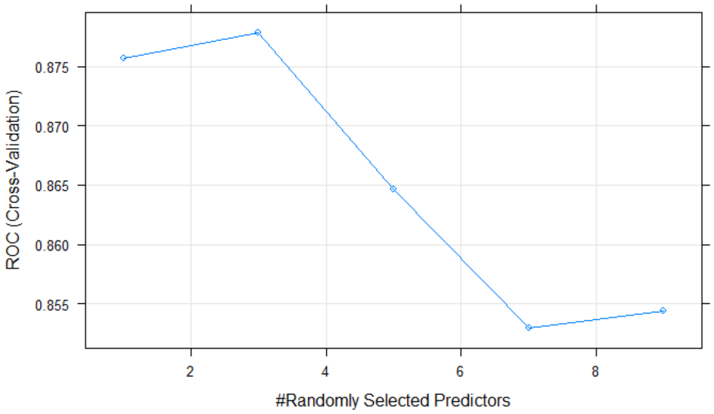


Figure 4.3.2: Plot of Predictors vs ROC

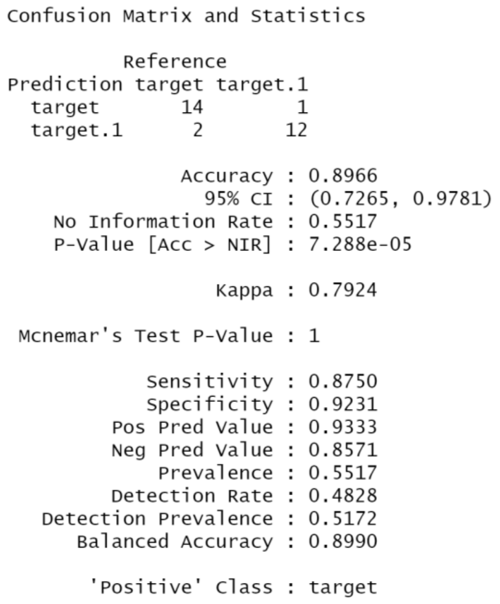
4.2.3 Model diagnosis

The final result shows that this model performance general, the reason may be that this model can easily get into over-fitting problem and we may ignore the correlation between attributes in data set.

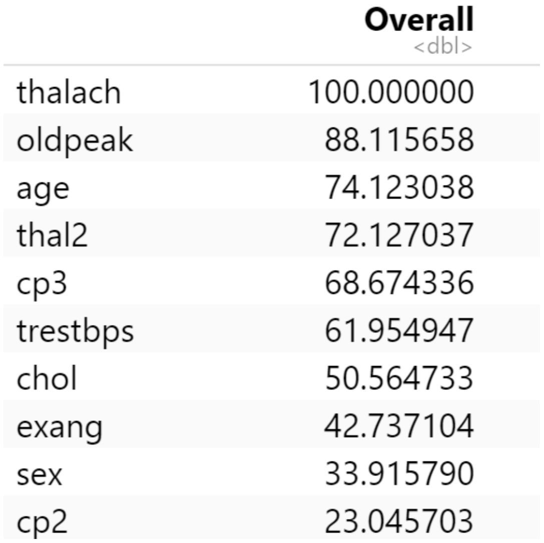
Randomly Selected Predictors	ROC	Sensitivity	Specificity
3	0.8777902	0.8126154	0.7636364

Table 4.2.3.1: Random Forest Model performance evaluation

From the relative importance table, we can find that thalach still take the first spot, while oldpeak and age rise to the second and third, it may due to the randomness this model has. But above all, the important factors don’t change much through these three models.



```
Confusion Matrix and Statistics

          Reference
Prediction target target.1
  target       14        1
  target.1      2       12
                                          
               Accuracy : 0.8966          
                 95% CI : (0.7265, 0.9781)
    No Information Rate : 0.5517          
    P-Value [Acc > NIR] : 7.288e-05       
                                          
                  Kappa : 0.7924          
                                          
 Mcnemar's Test P-Value : 1               
                                          
            Sensitivity : 0.8750          
            Specificity : 0.9231          
         Pos Pred Value : 0.9333          
         Neg Pred Value : 0.8571          
             Prevalence : 0.5517          
         Detection Rate : 0.4828          
   Detection Prevalence : 0.5172          
      Balanced Accuracy : 0.8990          
                                          
       'Positive' Class : target          
```



	Overall <dbl>
thalach	100.000000
oldpeak	88.115658
age	74.123038
thal2	72.127037
cp3	68.674336
trestbps	61.954947
chol	50.564733
exang	42.737104
sex	33.915790
cp2	23.045703

Table 4.3.3.2: Model Diagnosis

Table 4.3.3.3: Importance rating

4.4 ANN (Artificial Neural Network)

4.4.1 Motivation

The neural network is one of the most popular models used in machine learning area and we want to obtain some hands-on experience for future study, so we decided to use this model. As we all know, neural network is known as biologically inspired and it is capable of modeling extremely complex non-linear functions.

4.4.2 Model construction

Based on the cross validation result we formed a neural network with hidden layer 3, you may find the result in the Figure 4.4.2.

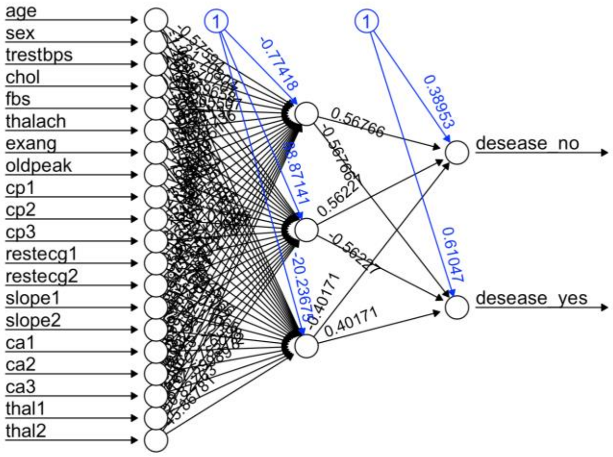


Figure 4.4.2

4.4.3 Model diagnosis

Base on the output from the neural network we can calculate the following result, firstly, the accuracy isn't high comparing with the above models and the specificity and sensitivity do not stand out either. After we dig into the neural network, we find that this might be as a consequence from the small dataset [7].

Accuracy	Specificity	Sensitivity
0.7931034	0.8571429	0.7333333

Table 4.4.3: Model diagnosis of neural network

4.5 Model comparison

After constructing and modifying above 5 models, we want to choose two models for two purpose, one is to make inference and explain the importance feature, the other one is the predict the heart disease and detecting fraud. Based on the Table 4.5, in the sense of best prediction accuracy and robustness, we choose Random forest model as the prediction model. And due to the fact that GLM perform only little behind other models and logistic regression has its nature benefits of making clear and evidential inference, we decided to use logistic regression to get explanatory conclusion later.

	Accuracy	Specificity	Sensitivity	ROC
GLM	0.8276	0.7692	0.8750	0.9183
LDA	0.8966	0.7727	0.8893	0.8963
KNN	0.8621	0.7091	0.8911	0.8795
RF	0.8966	0.7636	0.8126	0.8780
ANN	0.7931	0.8571	0.7333	NAN

Table 4.5: Comparison on GLM, LDA, KNN, Random forest and ANN

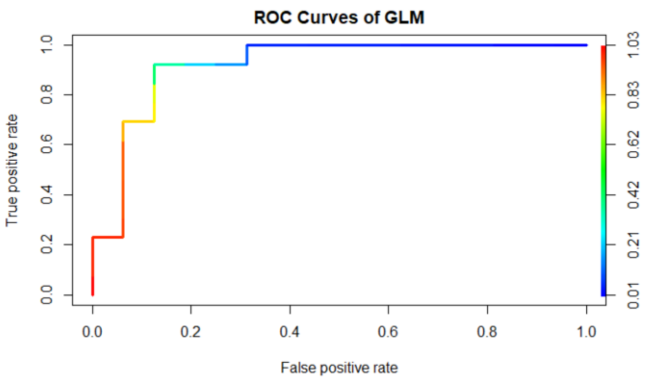


Figure 4.5.1: ROC Curve of GLM

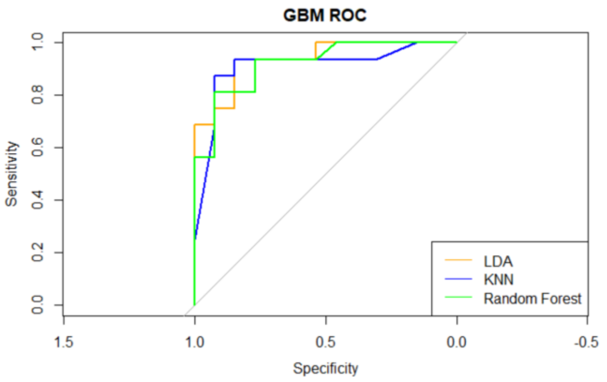


Table 4.5.2: ROC Curves of LDA, KNN and RF

5. Conclusions and summary

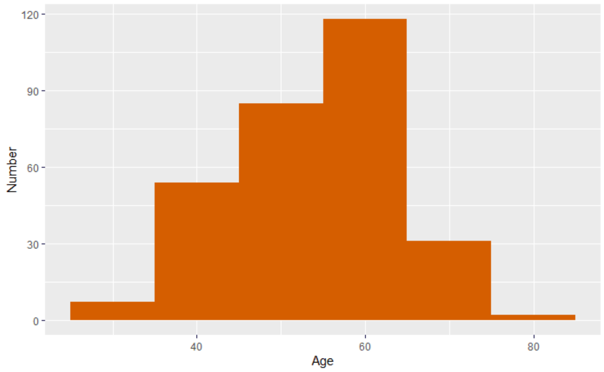
5.1 Conclusions

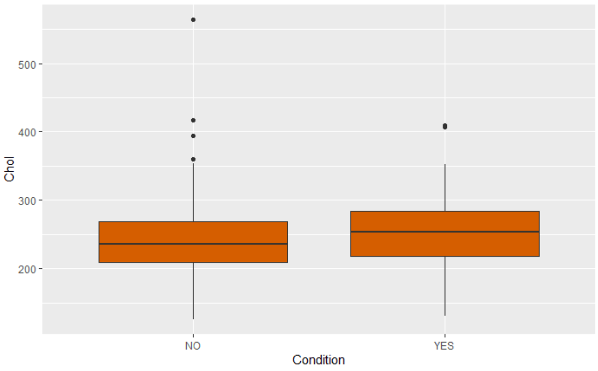
First of all, we find that sex and asymptotical chest pain are significant factors for heart disease. Asymptotical chest pain is also called silent pain. People who have a silent heart attack might later recall that they had indigestion, the flu, or a strained chest muscle. But a silent heart attack, like any heart attack, involves blockage of blood flow to your heart and possible damage to the heart muscle. [5] Studies suggest that men who have sex at least twice a week and women who report having satisfying sex lives are less likely to have a heart attack. [6]

Additionally, the main symptom of heart disease is insufficient blood supply. From the data, we find that many factors, more specifically, indicators, mainly reflect the change of heart performance before and after exercise such as Thallium Test and Resting Blood Pressure. Many indicators show that the performance of heart of patients group can't go back to the normal situation (contrast to rest) in time.

5.2 Limitations and further study

In this research, age and cholesterol are not significant. However, it's common knowledge that older and high-cholesterol people have more possibility to get heart disease. In fact, we find that the over 2/3 individuals in the data are older than 50 years old and have high cholesterol (normal level: 200). Hence, the data is biased. That may account for the reason that why age and cholesterol are not significant in the model.





For further study, to collect stratified sample in a more general group is necessary. And more models and methods can be applied to the study of heart disease, such as casual inference.

Reference

[1] https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease

[2] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3] https://elitecarehouston.com/silent-heart-attacks-what-do-asymptomatic-signs-of-a-heart-attack-mean/

[4] http://cooleysanemia.org/updates/Cardiac.pdf

[5] https://elitecarehouston.com/silent-heart-attacks-what-do-asymptomatic-signs-of-a-heart-attack-mean/

[6] https://www.hopkinsmedicine.org/health/wellness-and-prevention/is-sex-dangerous-if-you-have-heart-disease

[7] Ingrassia, Salvatore & Morlini, Isabella. (2005). Neural Network Modeling for Small Datasets. Technometrics. 47. 297-311. 10.1198/004017005000000058.