Study of Factors for Heart Disease Prediction between Men and Women based on Multiple Regression Models

Jingyi Guo*

Department of Bioinformatics, Shanghai Jiaotong University, Shanghai, China

*Corresponding author: Victoriaalbert@sjtu.edu.cn

Abstract. Although previous studies have shown that there are differences in heart disease between men and women, the importance of some specific physical and chemical factors in the prediction of heart disease in different genders has not been clearly clarified. In this research, K-means clustering, multiple linear regression, logistic regression and random forest are adopted to analyze the UCI Heart Disease Data Set, which contains various physical and chemical indicators worth studying. The results demonstrate that exercise induced angina is more significant to the judgement of heart disease in women, while number of major vessels colored by fluoroscopy is more significant to the judgement of heart disease in men and type of chest pain is a statistically significant variable for both men and women. Thalassemia, ST depression induced by exercise relative to rest, greatest number of heartbeats per minute, age, resting blood pressure also have reference value for the judgment of heart disease. In terms of each model's fit to heart disease prediction, for women, the accuracy of random forest is the first, logistic regression is the second, and multiple linear regression is the third, while for men, the accuracy of random forest is the first, multiple linear regression is the second, and logistic regression is the third. These conclusions are an optimization of previous studies, and to a certain extent reflect that this study is of great significance to the prevention of heart disease in different groups of people.

Keywords: Heart disease; prediction; important factors; gender; regression models.

1. Introduction

In recent years, heart disease has always occupied a significant position in global diseases. World Heart Day which is held every September 29th can reflect how much each country values heart disease. Heart disease is of a relatively large proportion among global death toll. The number of patients and deaths of the disease are gradually on the increase, thus leading to a profound impact on human health. Meanwhile, the audience of this disease is expanding: on the one hand the number of young patients is increasing, on the other hand if population aging trend is taken into consideration, plus for a multitude of heart diseases such as valvular heart disease, its morbidity is positively related to people's age, the situation will get worse [1]. Only in China, the number of patients with cardiovascular disease in China has already reached up to 330 million. But a majority of people including potential patients of heart disease actually lack in-depth knowledge in terms of this disease and they don't pay much attention to physical and chemical indicators associated with it, which largely make them miss the best opportunity to prevent heart disease. In addition, there are differences between different groups of people, such as men and women, in the risk factors associated with heart disease prediction. So it is an effective measure to control the morbidity of heart disease and improve human health generally by using statistical methods like logistic model regression analysis to explore various human health indicators, predict or differentiate high-risk groups for heart disease among different groups of people, further emphasize susceptibility factors or risk factors. Then the potential patients are able to intervene early, cultivating healthy living habits in time, and delay or eliminate disease before it comes.

In the current clinical research on heart disease prediction, multiple linear regression and logistic regression are widely adopted [2]. In the physical and chemical indicators measurements and risk factors analysis, age, hypertension, hyperlipidemia, BMI, smoking, diabetes, cholesterol level, etc all have a certain research value. Ying Yang et al used logistic regression to draw the conclusion that age and hypertension were apparent and crucial discriminative markers for patients with valvular

heart disease in China [3]. Chen et al utilized logistic regression and figured out gender, resting blood pressure, number of major vessels colored through perspective were statistically significant for heart disease [4]. Nie et al also analyzed that age, systolic blood pressure, uric acid and serum total bilirubin were independent factors for coronary heart disease through logistic regression [5]. The results of these studies reveal the complexity and variety of heart disease risk factors. Although these researches are all based on logistic regression model, different samples, diverse data sets and processing methods may have some effect on the final conclusion. Zhao et al optimized the random forest algorithm and concluded that 7 physical and chemical indicators such as maximum heart rate, number of major vessels, chest pain type can be used as parameters of judging heart disease [6]. Li et al found that random forest was equipped with significant improvement over multiple linear regression and other machine learning models when predicted cardiovascular disease prevalence in eastern China [7]. Yu Liu et al figured out that age, cholesterol, maximum heart rate and post-exercise comparative heart pressure have the most significant effects on heart disease through XGBoost modeling [8]. These studies show that machine learning is widely applied to researches related to heart disease risk factors and it can improve the accuracy of various physical and chemical indicators for heart disease prediction to some extent. Low HDL cholesterol, diabetes and many sex-specific disorders, such as those associated with pregnancy, demonstrate differences in the importance of determinants and predictors of cardiovascular disease morbidity and mortality between women and men [9]. Women have more angina in study of ischemic heart disease [10]. Differences between men and women in heart disease prediction are worthy of attention and in-depth study.

Nevertheless, existing researches can also be ameliorated and detailed, for instance, plenty of research methods like multiple linear regression, logistic regression, K-means clustering and random forest can combine with each other, which may help to explore the differences between different data regression fitting methods. Besides, the interrelationship between various influencing factors, the relationship between factors and heart disease are not clearly stated, for example, the difference in the degree of influence of different physical and chemical indicators on the prevalence of heart disease in men and women. By optimizing these aspects, people will have a clearer understanding of health indicators closely related to heart disease and the interrelationship between them.

2. Methods

2.1. Data Source and Description

This research is mainly based on Heart Disease data set from Kaggle platform, and the source is UCI Heart Disease Data Set: processed.cleveland.csv. Some of the existing research papers also use this dataset, indicating that this dataset has certain analytical value. The sample size of this dataset is 303, of which 139 suffer from heart disease and 164 do not, including 97 women and 206 men.

2.2. Variable Selection and Interpretation

This dataset contains 14 variables feasible for processing, which are physical and chemical indicators that this research wants to study. This research initially puts the results of categorical variables in this dataset in numbers: AHD stands for heart disease and is renamed to diseased_or_not (1=disease,0=no disease); 13 other variables are indicators to predict whether people suffer heart disease or not: ChestPain is renamed to ChestPainType (0,1,2,3), Thal is renamed to ThalType (0,1,2). Names and attributes of variables are listed in table 1.

Table 1. Names and attributes of variables

Variable name	Variable description	Variable type	Data type
Age	age (years)	numeric	int
Sex	gender (1=male,0=female)	categorical	int
RestBP	resting blood pressure (mmHg,upon admission to the hospital)	numeric	int
Chol	serum cholesterol (mg/dL)	numeric	int
Fbs	fasting blood sugar (>120 mg/dL,1=true,0=false)	categorical	int
RestECG	resting electrocardiogram results (0,1,2)	categorical	int
MaxHR	greatest number of beats per minute	numeric	int
ExAng	exercise induced angina (1=yes,0=no)	categorical	int
Oldpeak	ST depression induced by exercise relative to rest (mm)	numeric	dbl
Slope	the slope of the peak exercise ST segment (1,2,3)	categorical	int
Ca	number of major vessels (0-3)	numeric	int
ChestPainType	chest pain type (0,1,2,3)	categorical	dbl
ThalType	Thalassemia (0,1,2)	categorical	dbl
diseased_or_not	diseased or not (1=disease,0=no disease)	categorical	dbl

2.3. Research Protocol

The research methods include K-means clustering, multiple linear regression, logistic regression and random forest. The mice package in R is adopted to fill missing values.

On account of the many different types of heart disease such as congenital heart disease, hypertensive heart disease, coronary heart disease, valvular heart disease and rheumatic heart disease, there may be some differences between risk factors of different types of heart disease, different gender and age. Therefore, this research introduces K-means clustering algorithm to explore special taxa that may exist in this dataset. K-means clustering is to divide the input samples into k clusters through unsupervised learning, so that the loss function of the clustering results reaches the minimum value. The steps of the K-means algorithm are shown in Figure 1.

$$Y = J(c, \mu) = \min \sum_{i=1}^{M} ||X_i - \mu_{c_i}||^2 . \quad (1)$$

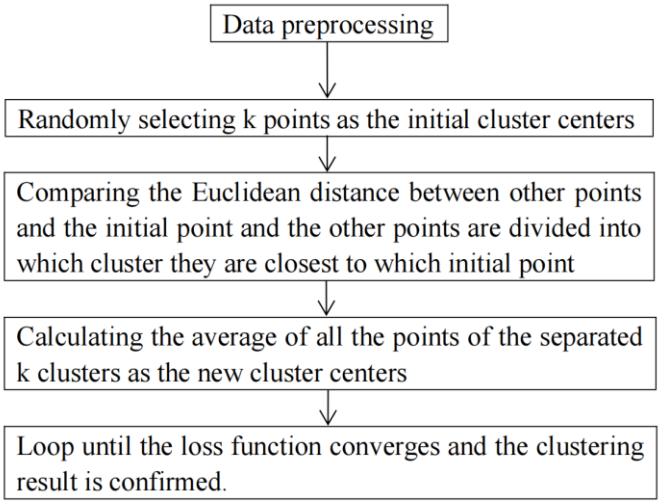

Fig. 1 K-means algorithm process

Multiple linear regression is used since each single variable doesn't fit well with diseased_or_not. For logistic regression, it takes advantage of the maximum likelihood estimation.

$$\log \frac{p_x}{1-p_x} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \quad (2)$$

X_i represents i-th predictor variable, β_i represents coefficient of the i-th predictor variable, when p_x is greater than 0.5, diseased_or_not of the test individual is defined as 1, otherwise it is defined as 0.

For random forest, the parameters of the fitting models are adjusted, mainly the setting of ntree and mtry values.

2.4. Data Preprocessing

It can be seen from figure 2 that the dataset contains 6 missing values, which can be filled by proper data padding method. This research selects mice package in R to fill missing values, Bayesian linear regression, logit regression fitting, polynomial fitting, predictive mean matching are tried and predictive mean matching is chosen finally. Number of imputed datasets (m) is set to 15, number of iterations (maxit) is set to 60. This research also uses density plot to get the values that match best and fill them into original dataset.

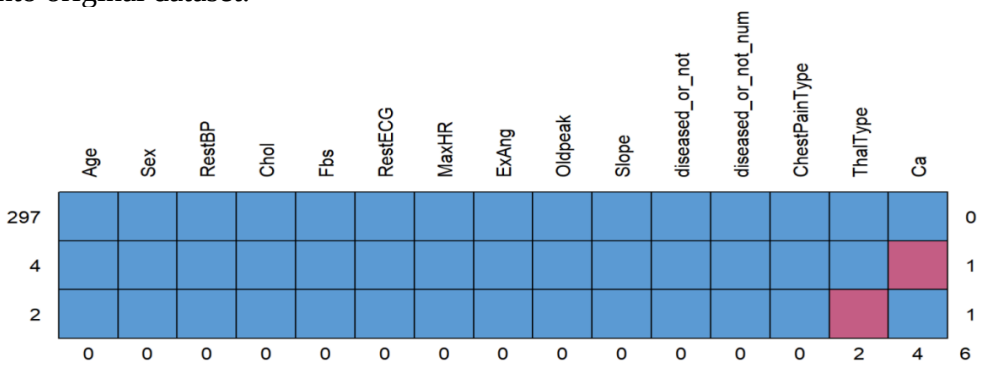

Fig. 2 Missing value

3. Results and Discussion

3.1. K-means Clustering

Clustering by all variables: Functions in R such as fviz_nbclust, clusGap, fviz_gap_stat are applied to determine the best number of clusters, which is 4 in this research according to figure 3. K-means function is utilized to get the result: number of observations in 4 clusters is 91,78,127,5. Among them, the 5 observations of the fourth cluster have obvious characteristics different from those of the other three clusters, including that the 5 observations are all female, the average age ranks first, values of Chol are significantly higher than other clusters.

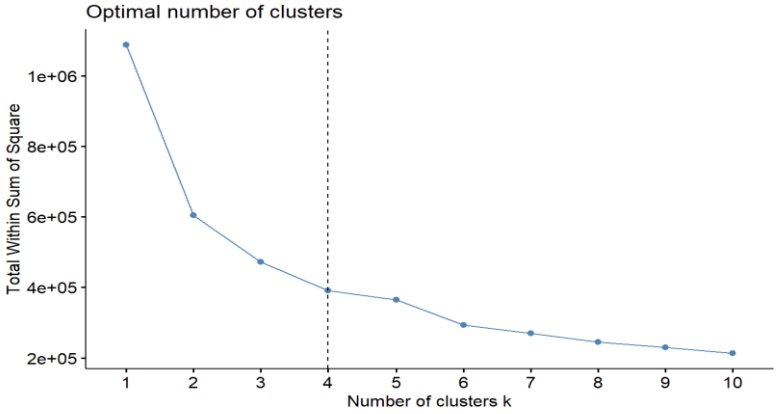

Fig. 3 Number of clusters for all variables

Clustering by Chol: The method is exactly the same as clustering by all variables. Revealed by figure 4, the best number of clusters is 4 as well, the number of observations in 4 clusters is 100,129,69,5 and these 5 observations are the same as the 5 observations obtained when all variables are clustered, showing the specificity of these 5 observations.

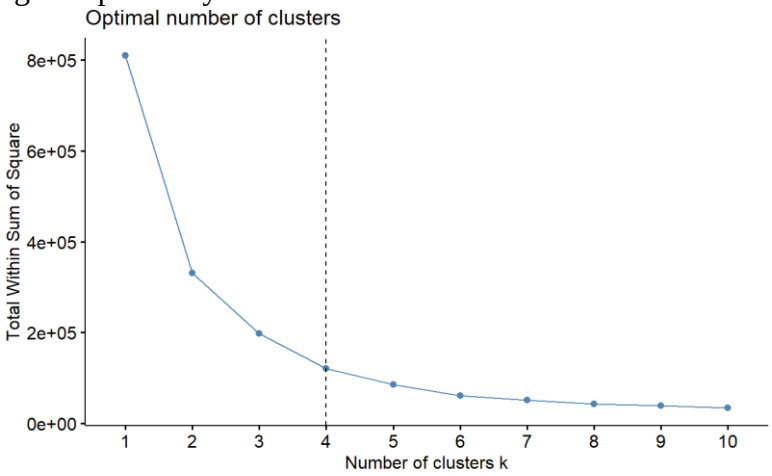

Fig. 4 Number of clusters for Chol

Due to the existence of detailed research focused on Chol, this research emphasizes differences caused by gender. These 5 observations are all from women, which may provide some evidence that there are differences in the importance of multiple physical and chemical indicators or risk factors in the prediction of heart disease in different genders.

So the data analysis is divided into 3 parts, including the mixed-gender data, the male-only data, and the female-only data because this research wants to figure out the preference of different gender prediction results for various physical and chemical indicators. Data for these 3 parts is divided into 70% training and 30% testing. The following regression fittings and predictive models all use the partitioned data so these data processing methods are comparable.

3.2. Multiple Linear Regression

Mixed-gender data: lm function in R is used to analyze the fitting prediction of all 13 variables except diseased_or_not to diseased_or_not in mixed-gender data. Adjusted R-squared of the fitting result is 0.513 and it is revealed that Sex, MaxHR, ExAng, Ca, ChestPainType, ThalType are statistically significant (p-value less than 0.05). ROC curve is drawn, AUC is 0.845 and RSE estimate is 0.349. Then this research analyzes fitting prediction of these 6 statistically significant variables to diseased_or_not. Adjusted R-squared of the fitting result is 0.492, ROC curve is drawn, AUC is 0.858 and RSE estimate is 0.357.

Male-only data: lm function is used to analyze the fitting prediction of 12 variables except diseased_or_not and Sex to diseased_or_not in male-only data. Adjusted R-squared of the fitting result is 0.456 and it is revealed that Oldpeak, Ca, ChestPainType, ThalType are statistically significant (p-value less than 0.05). ROC curve is drawn, AUC is 0.816 and RSE estimate is 0.370. Then this research analyzes fitting prediction of these 4 statistically significant variables to diseased_or_not. Adjusted R-squared of the fitting result is 0.444, ROC curve is drawn, AUC is 0.761 and RSE estimate is 0.374.

Female-only data: lm function is used to analyze the fitting prediction of 12 variables except diseased_or_not and Sex to diseased_or_not in female-only data. Adjusted R-squared of the fitting result is 0.500 and it is revealed that ExAng, ChestPainType, ThalType are statistically significant (p-value less than 0.05). ROC curve is drawn, AUC is 0.827 and RSE estimate is 0.309. Then this research analyzes fitting prediction of these 3 statistically significant variables to diseased_or_not. Adjusted R-squared of the fitting result is 0.475, ROC curve is drawn, AUC is 0.827 and RSE estimate is 0.316.

3.3. Logistic Regression

Mixed-gender data: glm function in R is used to analyze the fitting prediction of all 13 variables except diseased_or_not to diseased_or_not in mixed-gender data. It is revealed that Sex, RestBP, Ca, ChestPainType, ThalType are statistically significant (p-value less than 0.05). ROC curve is drawn and AUC is 0.848. Then this research analyzes fitting prediction of these 5 statistically significant variables to diseased_or_not. ROC curve is drawn and AUC is 0.833.

Male-only data: glm function in R is used to analyze the fitting prediction of 12 variables except diseased_or_not and Sex to diseased_or_not in male-only data. It is revealed that Oldpeak, Ca, ChestPainType are statistically significant (p-value less than 0.05). ROC curve is drawn and AUC is 0.787. Then this research analyzes fitting prediction of these 3 statistically significant variables to diseased_or_not. ROC curve is drawn and AUC is 0.726.

Female-only data: glm function in R is used to analyze the fitting prediction of 12 variables except diseased_or_not and Sex to diseased_or_not in female-only data. The p-values of the fitting results are all relatively large, probably due to the small amount of female-only data. ROC curve is drawn and AUC is 0.851.

3.4. Random Forest

Mixed-gender data: randomForest function in R is used to analyze the fitting prediction of all 13 variables except diseased_or_not to diseased_or_not in mixed-gender data. In terms of figure 5, it is revealed that Ca, ChestPainType, ThalType, MaxHR, Oldpeak are significant. ROC curve is drawn and AUC is 0.925.

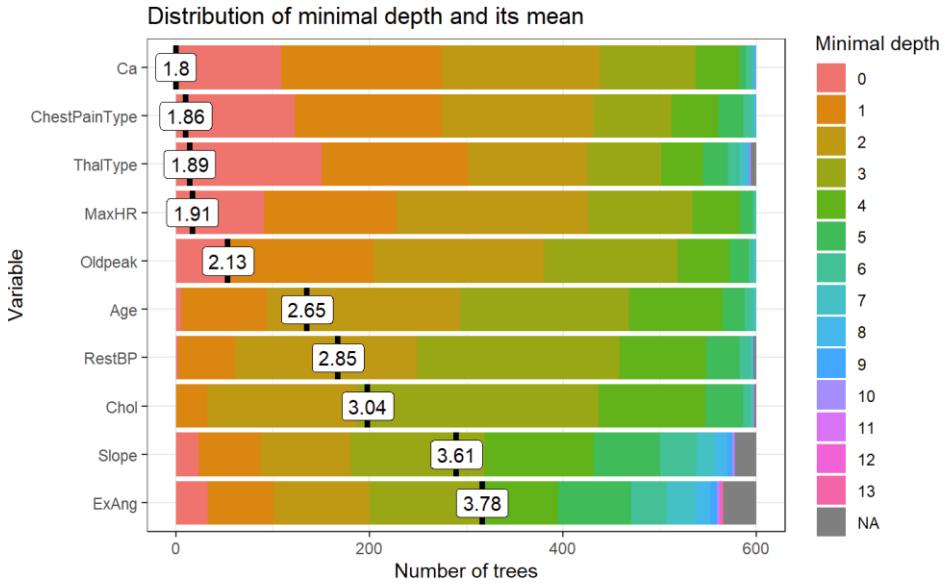

Fig. 5 Important variables for mixed-gender data

Male-only data: randomForest function is used to analyze the fitting prediction of 12 variables except diseased_or_not and Sex to diseased_or_not in male-only data. In terms of figure 6, it is revealed that Ca, Oldpeak, ChestPainType, MaxHR, Age are significant. ROC curve is drawn and AUC is 0.862.

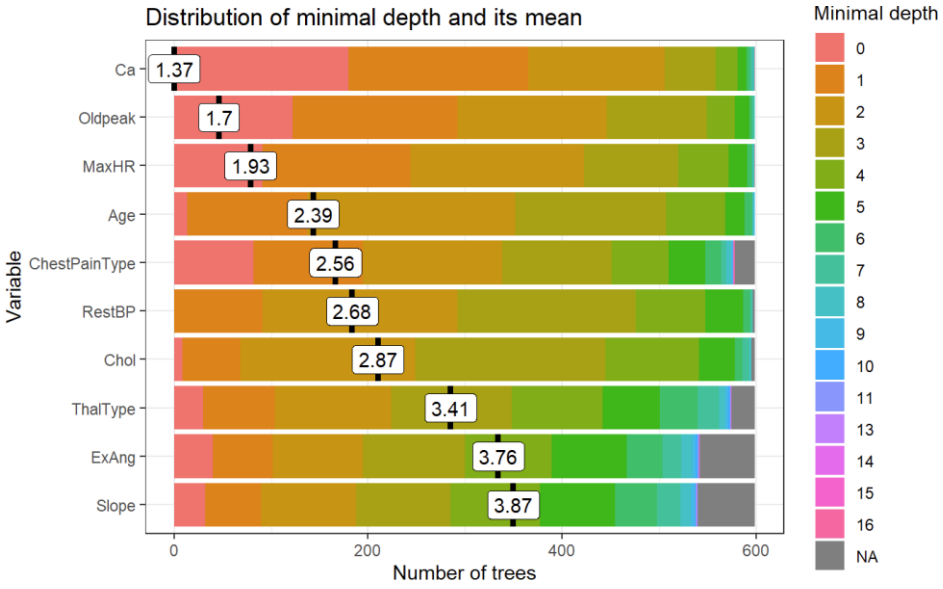

Fig. 6 Important variables for male-only data

Female-only data: randomForest function is used to analyze the fitting prediction of 12 variables except diseased_or_not and Sex to diseased_or_not in female-only data. In terms of figure 7, it is revealed that ChestPainType, ThalType, ExAng, Oldpeak,Age are significant. ROC curve is drawn and AUC is 0.946.

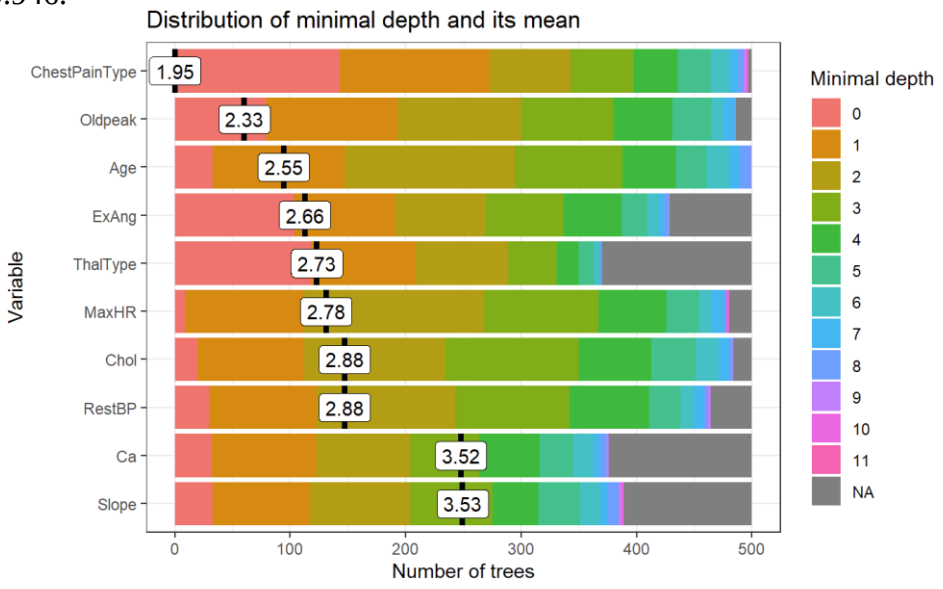

Fig. 7 Important variables for female-only data

3.5. Comparison of These Models

Table 2 shows a lot of useful information: Compared to men, ExAng is more significant to the judgement of diseased_or_not in women. Compared to women, Ca is more significant to the judgement of diseased_or_not in men. In all models, ChestPainType is a statistically significant variable or a variable with high importance whether it is for mixed-gender data,male-only data or female-only data. For other variables, ThalType and Oldpeak often appear in variables that are statistically significant or in the forefront of importance, followed by MaxHR, Age, RestBP, and they all have reference value for the judgment of diseased_or_not.

Table 2. Significant variables

Variable	All-m	Male-m	Female-m	All-l	Male-l	All-r	Male-r	Female-r
Sex	√	-	-	√	-		-	-
ExAng	√		√					√
Ca	√	√		√	√	√	√	
ChestPainType	√	√	√	√	√	√	√	√
ThalType	√	√	√	√		√		√
Oldpeak		√			√	√	√	√
MaxHR	√					√	√	
Age							√	√
RestBP				√				

All-m means mixed-gender data by multiple linear regression. Male-m means male-only data by multiple linear regression. Female-m means female-only data by multiple linear regression. All-l means mixed-gender data by logistic regression. Male-l means male-only data by logistic regression. All-r means mixed-gender data by random forest. Male-r means male-only data by random forest. Female-r means female-only data by random forest.

Concluded from Figures 8, 9, and 10, in terms of data processing in this research according to the comparison of AUC values, random forest is the most accurate, followed by logistic regression, and finally multiple linear regression whether it is for mixed-gender data or female-only data, while for male-only data, the order is random forest, multiple linear regression, logistic regression.

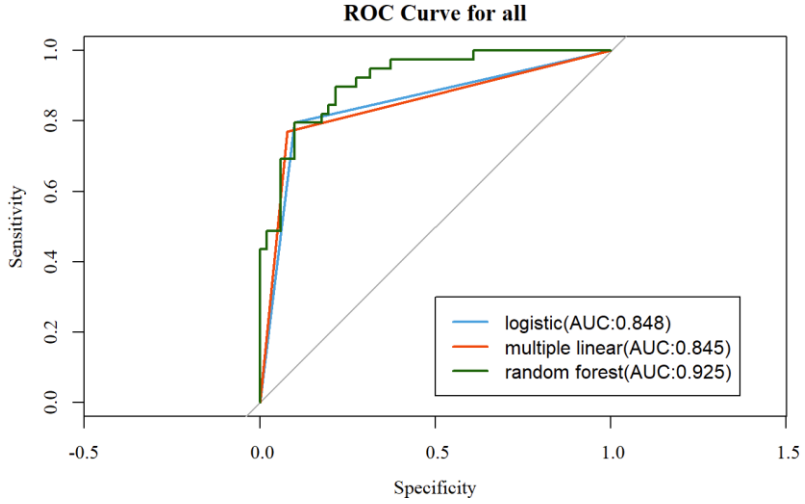

Fig. 8 ROC curves for mixed-gender data

The reasons for these results in this study are complex. For instance, ExAng is more significant to the judgement of diseased_or_not in women maybe partly because women have more angina than men in this study. Ca is the standard indicator to detect narrowing or blockage of the major arteries of the heart but women with heart disease such as coronary heart disease usually develop in small arteries that are not visible by fluoroscopy, which may be part of reasons for difference of the significance of Ca to the judgement of diseased_or_not between men and women, etc.

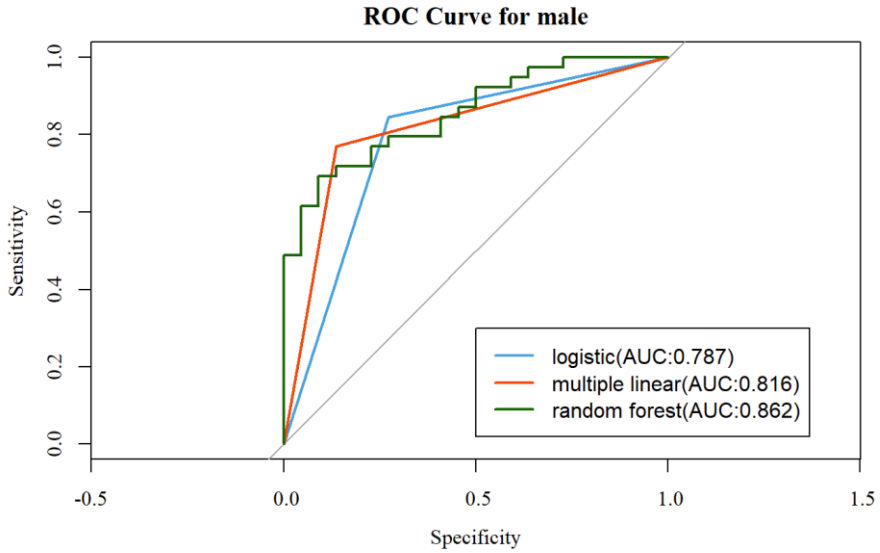

Fig. 9 ROC curves for male-only data

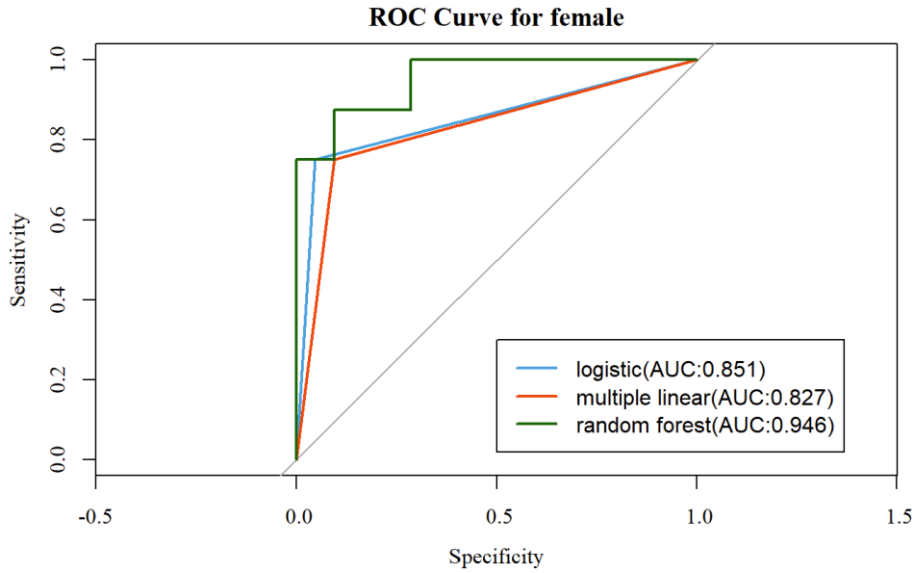

Fig. 10 ROC curves for female-only data

4. Conclusion

This research explores difference in the degree of influence of factors on the prevalence of heart disease in men and women by different data regression fitting methods. According to the above study, it can be concluded that exercise induced angina is more significant to heart disease prediction in women, number of major vessels colored by fluoroscopy is more significant to heart disease prediction in men, type of chest pain is a statistically significant variable for both men and women, Thalassemia, ST depression induced by exercise relative to rest, greatest number of heartbeats per minute, age, resting blood pressure also have reference value for the judgment of heart disease. For each model's fit to heart disease prediction, for women, the accuracy ranks are random forest, logistic regression, multiple linear regression, while for men, the accuracy ranks are random forest, multiple linear regression, logistic regression.

Therefore, even if there is no problem detected by fluoroscopy, women should not take it lightly. They should judge comprehensively through other indicators such as ExAng, while men should pay more attention to the angiography, in other words, the Ca values. At the same time, whether they are men or women, they should pay attention to ChestPainType, other indicators like ThalType, Oldpeak, MaxHR, Age and RestBP can also be used as reference factors. If the factors in this study are to be

utilized to predict heart disease, it would be better to use random forest and logistic regression for women and random forest and multiple linear regression for men.

This research also has some deficiencies and room for improvement: The total number of observations in this dataset is not large, which may lead to a certain deviation between obtained results and the real situation. Adding more machine learning algorithms may enhance the level of accuracy in heart disease prediction models. Some other methods can be added for imputation of missing values and data clustering such as k nearest neighbor algorithm and sklearn to enrich research content.

References

[1] Chenxi Xia, Xiang Wang, Xuyang Meng, et al. Risk factors analysis of coronary heart disease in Chinese elderly patients with severe valvular heart disease: a national multicenter cross-sectional study. Chinese Journal of Molecular Cardiology, 2022, 22(06): 5000-5004.

[2] Yin Ouyang, Ye Tan, Xin Deng. Application of regression analysis in clinical research of coronary atherosclerotic heart disease. China Medicine and Pharmacy, 2021, 11(18): 44-47.

[3] Ying Yang, et al. Current status and etiology of valvular heart disease in China: a population-based survey. BMC Cardiovasc Disord, 2021, 21(01): 339.

[4] Mengmeng Chen, Zhenhong Fang, Wenyi Tu, et al. Construction and effect analysis of heart disease prediction model based on logistic regression model. Hospital Management Forum, 2022, 39(02): 32-35.

[5] Xiaodan Nie, Saisai Cui, Bo Sun, et al. Analysis of risk factors of coronary atherosclerotic heart disease. Trauma and Critical Care Medicine, 2022, 10(01): 68-70.

[6] Jinchao Zhao, Yi Li, Dong Wang, et al. Heart disease prediction algorithm based on optimization random forest. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2021, 42(02): 112-118.

[7] Li Yang, Haibin Wu, Xiaoqing Jin, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports, 2020, 10(01): 5245.

[8] Yu Liu, Mu Qiao. Prediction of heart disease based on clustering and XGboost algorithm. Computer system application, 2019, 28(01): 228-232.

[9] Norris C M, et al. State of the science in women's cardiovascular disease: a Canadian perspective on the influence of sex and gender. J Am Heart Assoc, 2020, 9(04): e015634.

[10] Nussbaum S S, et al. Sex-specific considerations in the presentation, diagnosis, and management of ischemic heart disease: JACC focus seminar 2/7, Journal of the American College of Cardiology, 2021, 78(02): 189-192.