Prediction of Heart Disease Based on Logistic Regression and Random Forest Models

Xinyi Liu[1, †], Siyuan Su[2, †], Baoyi Wang[3, †], Xuewei Zhang[4, †, *]

[1] Henan Experimental High School, Zhengzhou, 451450, China

[2] Beijing Dongzhimen High School, Beijing, 100010, China

[3] Yichang Yiling Tianwen School, Yichang, 443100, China

[4] Department of Mathematics and Applied Mathematics, Guangdong University of Finance & Economics, Guangzhou, 510320, China

* Corresponding Author Email: 160706122@stu.cuz.edu.cn

[†]These authors contributed equally.

Abstract. Heart disease was a hot topic of research in the field of medicine in the 20th century. Heart disease is one of the more common diseases of the circulatory system. When they progress to a certain stage, they all have an impact on heart function. Severe heart disease can lead to heart failure, which can also affect a patient's quality of life and can be life-threatening. And using medical data to find appropriate ways to predict heart disease can also be effective in improving national health care. Our experiment is about predicting whether people have heart disease. This experiment was conducted to let people know for themselves if they have the disease and to treat it quickly. Some pattern recognition methods, such as logistic regression and random forest, were used in this study. Our attempt was to find the more important features using a large amount of experimental data, and then this study used different combinations of features and the two classification techniques mentioned above to build a model for predicting heart disease, in which we tried to identify the important features through a large number of experiments to improve the accuracy of cardiovascular disease prediction. The results of the heart disease model were related to the type of chest pain experienced and the maximum heart rate achieved by those with the highest odds of developing the disease. The prevalence of heart disease was higher in middle-aged and older adults, followed by a smaller proportion of younger people and those with genetic disorders. The number of men suffering from heart disease was also higher than that of women. The results of this experiment amply demonstrate the validity of our random forest model and logistic regression model in this dataset.

Keywords: Prediction; Heart disease; Logistic regression; Random forest.

1. Introduction

Heart disease is one of the most challenging medical conditions, despite the fact that physicians have a wide variety of instruments and techniques at their disposal. Heart disease is a relatively common disease of the circulatory system. The circulatory system consists of the heart, blood vessels and the neurohumoral tissues that regulate blood circulation. Diseases of the circulatory system, also known as cardiovascular disease, include diseases of all these tissues and organs and are common medical conditions, of which heart disease is the most common and can significantly affect a patient's work. In severe cases, it can lead to heart failure, which affects a patient's quality of life and can be life-threatening. According to the American Heart Association, nearly 20 million people worldwide die each year from heart disease, stroke, blood vessel blockage and other cardiovascular system disorders. In this context, it is essential to find an effective way to predict a patient's risk of heart attack.

In our study, we used several pattern recognition methods, such as logistic regression and random forest. Among them, we tried to identify the important features and improve the accuracy of cardiovascular disease prediction through a large number of experiments.

The earliest work on probabilistic algorithms for the prediction of cardiac diseases can be traced back to the 1989 [1]. The early articles, represented by this paper, mostly used small data samples with their own discriminant function models due to the lack of database resources and imperfect prediction models at that time. This made the early predictions take into account the small number of attributes, large prediction errors, and low prediction accuracy. With the enrichment of data samples, the problem of small number of attributes has been solved. Leslie Cho et al. in their paper considered about attributes including but not limited to baseline C-reactive protein, low-density lipoprotein cholesterol, etc [2]. The refinement of data models has allowed scholars to use the same set of data corresponding to different models and compare their prediction accuracy and other results. Zhang et al. utilized multiple models including but not limited to decision trees, neural networks, and other models alone as well as multiple models combined for prediction [3]. The preprocessing of data in this article also leads to further improvement in prediction accuracy. After the above three problems have been solved, scholars continue to search for new breakthroughs, and Talha et al. have tried the use of integrated methods in [4], which summarizes the results of the past thirty-three years. In addition, paper written by Dimitris Bertsimas, et al. and Rohit Bharti demonstrated their intention about using machine learning and deep learning approaches [5, 6]. Related works reported by Zhang et al. used the current rapidly developing artificial intelligence techniques [7].As well as the gene expression programming used by Dan Wenbin, et al have opened up new ideas for disease prediction [8].

In summary, research on the prediction of cardiac diseases has evolved and innovated over the past thirty-three years, from solving the problem of insufficient samples to trying new methods. However, comparisons between prediction models, such as those in the works of Zhang et al. Therefore, based on the existing research and literature, in this paper, we compare the results based on the Cleveland database of heart disease dataset, using logistic regression model and random forest model to predict the probability of heart disease for 14 out of 76 attributes and compare the results, then we based on the preliminary results of random forest The decision tree nodes were optimized by successive gradient changes. Then, study the fitness of model and data.

2. Methodology

2.1. Dataset description

In this project, we used the dataset provided by Cleveland database [9]. This dataset contains 76 attributes, but all published experiments were appropriate for using a subset of 14 of them.

The preprocessing is consisted of three parts. First, we categorized the dataset of 14 as nominal data, ordinal data, interval data and ratio data respectively. Because after checking the data types, we found that all of them were numerical types, which the computer compared the size of the figure for each data directly instead of distinguishing the different kinds of data representing. Second, the nominal data are converted from integer code to character string. For example, the number of 0,1,2,3 of chest pain types was changed into their actual meanings (typical angina pectoris, atypical angina pectoris, non-angina pectoris and asymptomatic); Third, the discrete nominal data and ordinal data were converted into one-hot single-heat coding, resulted in the dataset was expanded from the original 14 attributes to 26 attributes. For example, gender was converted into two attributes which used to judge whether it was male or female. And then, the numbers that represented the types of chest pain was converted into four attributes which were judging whether it was typical angina pectoris, atypical angina pectoris, non-angina pectoris and asymptomatic. It not only separates nominal data into the single hot vector coding, but also avoids the problem of comparing the size of the figure for each data.

2.2. Model

In this report, we used two models to study the heart disease dataset, in order to make up for a not accurate prediction of a single model. After building multiple models, we can compare and choose a

model which is more general and have the better potentiality to develop than the others. Furthermore, we can have more accurate predictions by comparison.

2.2.1. Logistic Regression

The first method we used was the logistic regression model. It is a common method used in machine learning to do classification tasks, which belongs to the generalized linear model. As a linear classifier called "regression", its essence is a generalized regression algorithm widely used in classification problems, which is changed by linear regression. It can be understood that logistic regression is to calculate the regression function first, and then transform the result through the logical function to obtain the final result. The logistic regression model has several advantages. First, the model is simple to understand, so the dataset will be trained at a fast speed. What's more, there is a good probability explanation for the output variables. Second, it can be applied to both continuous and discrete independent variables. Third, specific thresholds can be set according to actual needs by using the logistic regression model. However, this model also has some disadvantages. It can only deal with the second classification problem. Because maximum likelihood estimation performs poorly in small sample sizes, the logistic regression model is suitable for large sample sizes. If the data is unbalanced, this model will be difficult to handle.

2.2.2. Random Forest

In addition to the logistic regression model, we also used the random forest model, the integrated learning realization of decision tree. Its method is to train multiple decision trees at the same time and get the prediction results after comprehensively considering multiple outcomes. The randomness of the random forest is reflected in the fact that it is a sampling with return, and each tree is trained by randomly taking a subset from a fixed training set with putting back. Each time a forked feature is chosen, it is looked for in a randomly selected subset. Random forest model can be used to deal with high dimensional data without feature selection. After trained, it can give which features are important and detect the influence between features. For unbalanced data sets, random forest model can balance errors while logical regression model cannot. Meanwhile, this model has fairly good estimation performance in solving both classification and regression problems. Nevertheless, for sparse data with high dimensions, such as text data, random forest model often does not perform very well. For such data, it may be appropriate to use a linear model. If the number of decision trees in the random forest is large, it will take a lot of space and time to train. The random forest model is also easy to overfit for some noisy sample sets.

3. Result and Discussion

3.1. Research on Classification Report

We used both of the models to train and test on the training and testing dataset and got the following two classification reports as shown in Table 1 and Table 2. Then we calculated that the Mean Square Error (MSE) of the two were 0.17 and 0.09 respectively. For these data, it is obvious that the accuracy score, precision and recall of random forest are all higher than that of logistic regression, while the mean square error is smaller than that of logistic regression.

Table 1. Classification Report of Logistic Regression

	precision	recall	f1-score	support
Healthy	0.85	0.79	0.82	99
Disease	0.81	0.86	0.84	103
Accuracy	-	-	0.83	202
Macro avg	0.83	0.83	0.83	202
Weighted avg	0.83	0.83	0.83	202

Table 2. Classification Report of Random Forest

	precision	recall	f1-score	support
Healthy	0.94	0.86	0.90	99
Disease	0.88	0.95	0.91	103
Accuracy	-	-	0.91	202
Macro avg	0.91	0.91	0.91	202
Weighted avg	0.91	0.91	0.91	202

3.2. Research on Confusion Matrix

Then we drew the following two confusion matrices as shown in Figure 1 and Figure 2 of the models according to the report. As a visual tool, the confusion matrix can well evaluate the model results. Each column of the confusion matrix represents the prediction category, and each row represents the real category of data.

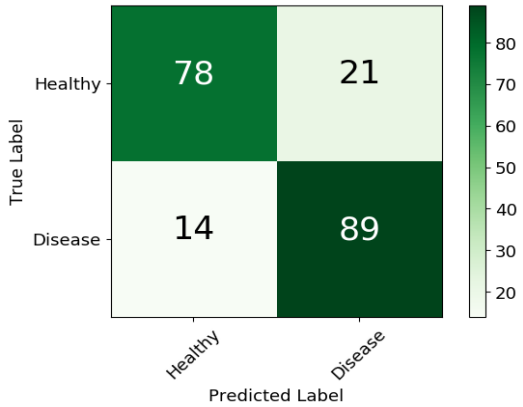

Fig. 1. Confusion Matrix of Logistic Regression

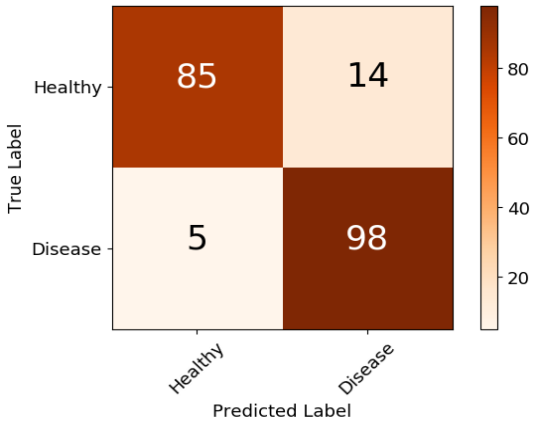

Fig. 2. Confusion Matrix of Random Forest

3.3. Research on Receiver Operator Characteristic Curve (ROC)

Finally, we draw ROC curve as shown in Figure 3 and Figure 4 to show the results of the comparison between the two models. In practice, the more convex to the upper left the ROC curve is, the better performance will be. The value of AUC represents the area under the ROC curve. When the Area Under the Curve (AUC) reaches 1, it is an ideal state. AUC is used to measure the performance of machine learning algorithms for binary classification problems. The ROC curve of random forest is more convex to the upper left than that of logical regression, and the AUC value (i.e. 0.97) of random forest model is closer to 1 than that of logical regression model (i.e. 0.91).

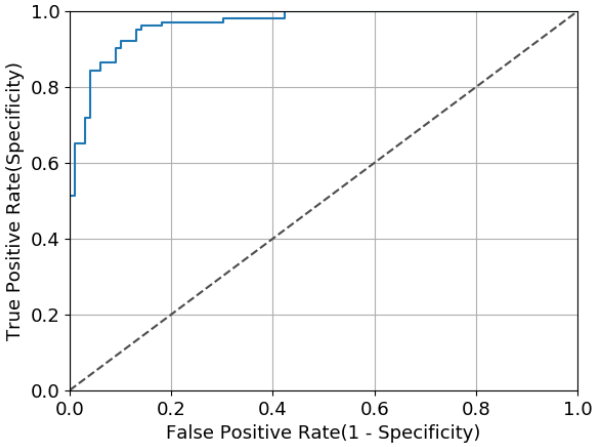

Fig. 3. ROC Curve of Logical Regression

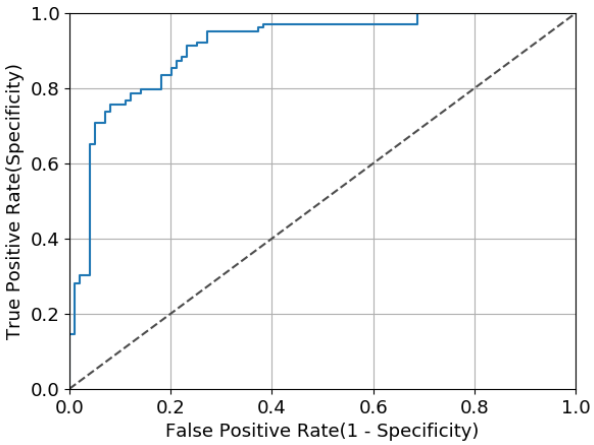

Fig. 4. ROC Curve of Random Forest

3.4. Comparative Study

In summary, by comparing the MSE, Accuracy Score and AUC of the two models as shown in Figure 5, it is showed that the fitting effect of the random forest model in this prediction of heart disease is significantly better than that of the logistic regression model.

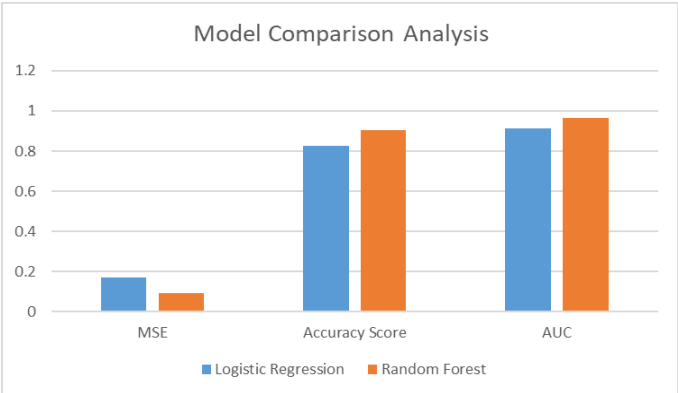

Fig. 5. Model Comparison Analysis

3.5. Feature Analysis

After the training, we used feature importance as an auxiliary tool to help analyze the prediction results as shown in Figure 6. Since the performance of the random forest model was better than that of the logistic regression model, we selected the feature importance ranking of the random forest model as a display. As we can see from the figure, “num _ major _ vessels” (number of main blood vessels around the heart) and “atypical angina” have a great influence on inducing heart disease.

Weight	Feature
0.0693 ± 0.0469	num_major_vessels
0.0376 ± 0.0204	chest_pain_type_typical angina
0.0307 ± 0.0146	age
0.0228 ± 0.0284	ST_depression
0.0208 ± 0.0202	cholesterol
0.0168 ± 0.0255	ST_slope_downsloping
0.0168 ± 0.0119	thalassemia_reversable defect
0.0129 ± 0.0148	exercise_induced_angina_yes
0.0129 ± 0.0255	thalassemia_fixed defect
0.0119 ± 0.0161	max_heart_rate
0.0109 ± 0.0211	ST_slope_flat
0.0089 ± 0.0115	sex_male
0.0079 ± 0.0134	sex_female
0.0030 ± 0.0049	rest_ecg_ST-T wave abnormality
0.0030 ± 0.0194	exercise_induced_angina_no
0 ± 0.0000	rest_ecg_left ventricular hypertrophy
0 ± 0.0000	chest_pain_type_asymptomatic
0 ± 0.0000	rest_ecg_normal
0 ± 0.0000	thalassemia_normal
0 ± 0.0000	fasting_blood_sugar_higher than 120mg/dl
	… 5 more …

Fig. 6. Feature Importance

For this result, we dug a little deeper as shown in Figure 7 and Figure 8. We performed data visualization of these two features on the initial dataset. From the figures, we can see that among all chest pain types, people with atypical angina have the highest proportion of heart disease. In addition, as the number of blood vessels around the heart increases, the rate of heart disease decreases.

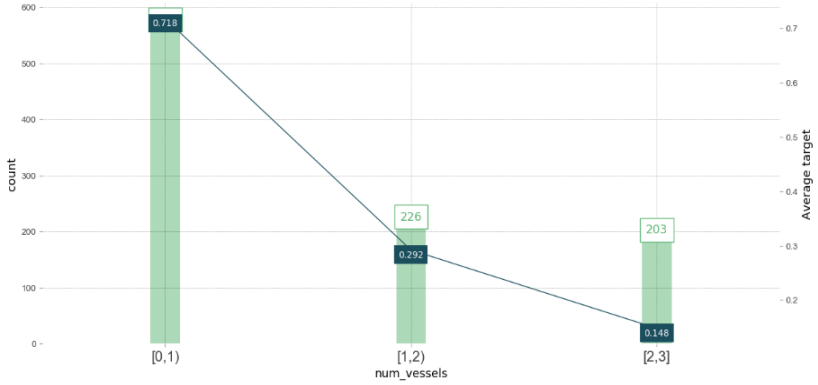

Fig. 7. Target Plot for Feature "chest_pain_type"

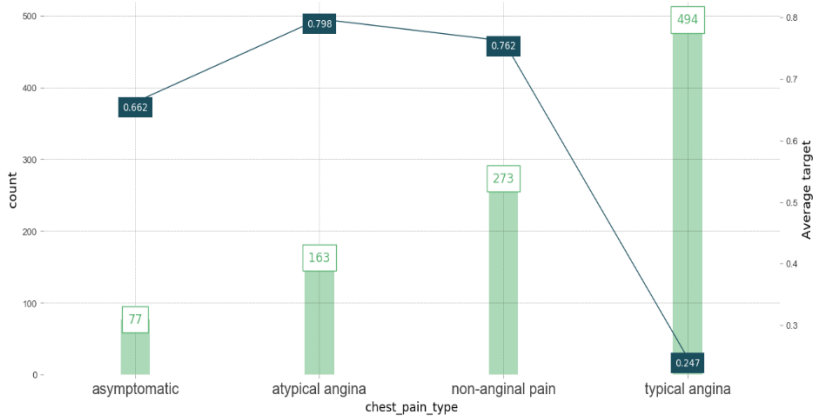

Fig. 8. Target Plot for Feature "num_vessels"

3.6. Additional Studies

Given the good performance of the random forest model, we wondered whether it can be optimized to improve its result. The optimization of random forest model generally takes two directions which are feature selection and parameter optimization. We adopted the method of adjusting the number of decision trees in parameter optimization to optimize the model. We set the number of trees from 0 to

300 lessons and trained and tested the model every 10 lessons while keeping the other variables constant. Finally, the number of optimal trees was 51, and the Accuracy Score was 0.92, which was 0.01 higher than the initial one. It was a very good optimization. The graph below shows the variation of classification accuracy with the number of trees as shown in Figure 9.

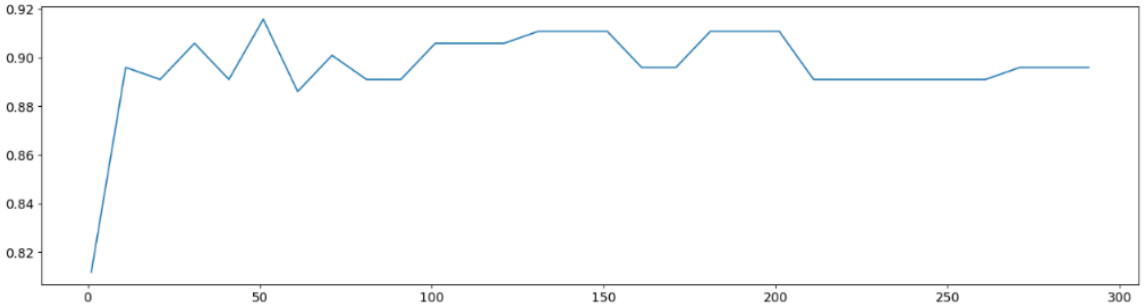

Fig. 9. Graph of the Variation

4. Conclusion

In this study, we predicted the rate of suffering from heart diseases based on logistic regression and random forest models, looking forward to getting effective disease factors to provide help for preventing cardiovascular diseases. Finally, it is concluded that “num _ major _vessels” (the number of main blood vessels around the heart), atypical angina are the main reasons that affect people's heart disease. Besides the performance of random forest model is better than that of logistic regression model in predicting processes. However, there are still many deficiencies in our research. As for other diseases factors, such as “max _heart _rate”, “exercise _induced _angina” and thalassemia, which are not in front of feature sorting, are the traits that linked to the causes of diseases. Even we did not work out a causal relationship between them. In the future, we will plan to do further studies about the changes of target attributes caused by pairwise feature attributes.

References

[1] Detrano R et al. 1989 International application of a new probability algorithm for the diagnosis of coronary artery disease The American journal of cardiology 64(5) 304-310

[2] Cho L et al. 2008 Gender differences in utilization of effective cardiovascular secondary prevention: a Cleveland clinic prevention database study. Journal of women's health 17(4) 515-521

[3] Zhang Z et al. 2022 Heart disease prediction based on feature selection and probabilistic neural network Modern Electronics Technology 45(1): 95**-**99

[4] Karadeniz T et al. 2021 Ensemble methods for heart disease prediction. New Generation Computing 39(3) 569-581

[5] Bertsimas D et al. 2021 Machine learning for real-time heart disease prediction. IEEE Journal of Biomedical and Health Informatics 25(9) 3627-3637

[6] Bharti R et al. 2021 Prediction of heart disease using a combination of machine learning and deep learning Computational intelligence and neuroscience

[7] Jian W et al. 2019 A novel method of prediction for heart disease based on convolution neural networks Journal of Natural Science of Heilongjiang University 36 115-120

[8] Dai W et al. 2009 The application of gene expression programming in the diagnosis of heart disease. Journal of Biomedical Engineering 26(1) 38-41

[9] Kaggle 2019 Heart Disease Cleveland UCI https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci